TCDesc: Learning Topology Consistent Descriptors for Image Matching
Honghu Pan, Fanyang Meng, Nana Fan, Zhenyu He*, Senior Member, IEEE,

Abstract—The constraint of neighborhood consistency or local consistency is widely used for robust image matching. In this paper, we focus on learning neighborhood topology consistent descriptors (TCDesc), while former works of learning descriptors, such as HardNet and DSM, only consider point-to-point Euclidean distance among descriptors and totally neglect neighborhood information of descriptors. To learn topology consistent descriptors, first we propose the linear combination weights to depict the topological relationship between center descriptor and its kNN descriptors, where the difference between center descriptor and the linear combination of its kNN descriptors is minimized. Then we propose the global mapping function which maps the local linear combination weights to the global topology vector and define the topology distance of matching descriptors as 1 distance between their topology vectors. Last we employ adaptive weighting strategy to jointly minimize topology distance and Euclidean distance, which automatically adjust the weight or attention of two distances in triplet loss. Our method has the following two advantages: (1) We are the first to consider neighborhood information of descriptors, while former works mainly focus on neighborhood consistency of feature points; (2) Our method can be applied in any former work of learning descriptors by triplet loss. Experimental results verify the generalization of our method: We can improve the performances of both HardNet and DSM on several benchmarks.

Index Terms—learning descriptors, neighborhood consistency, image matching, triplet loss.

I. INTRODUCTION

IMAGE matching [1], [2], [3] is a fundamental computer vision problem and the crucial step in augmented reality (AR) [4], [5] and simultaneous localization and mapping (SLAM) [6], [7], which usually consists of two steps: detecting the feature points and matching feature descriptors. The robust and discriminative descriptors are essential for accurate image matching. Early works mainly focus on the handcrafted descriptors. SIFT [1] maybe is the most successful handcrafted descriptor and has been proven effective in various areas [8], [9], [10], which is scale-invariant and orientation-invariant benefited from the Difference-of-Gaussian (DoG) scale space and the assigned main orientation respectively. Meanwhile, the binary descriptors [11] are proposed to reduce storage and accelerate matching, where Hamming distance is employed to compare two binary descriptors. However, handcrafted descriptors are not robust enough since they only consider pixel-level information and lack high-level semantic information.

Recently with the successful application of convolutional neural networks (CNNs) in multiple fields [12], [13], [14], researchers [15], [16], [17], [18], [19] try to learn descriptors directly from image patch by using CNNs. Specifically, CNNs take image patches cropped around feature points as input and take the representation vector of last layer as the learned descriptors. Recent works [19], [20], [21] mainly focus on learning descriptors using triplet loss [22] to encourage Euclidean distance of negative samples is a margin larger than that of positive samples, where negative samples and positive samples denote the non-matching descriptors and matching descriptors respectively. During CNNs’ training, Euclidean distance of matching descriptors is minimized and that of non-matching descriptors is maximized.

However, as shown in Fig. 1(a) triplet loss of former works only considers Euclidean distance between descriptors and completely neglects the neighborhood information of descriptors, which results in the topology difference between matching descriptors. Neighborhood consistency is wide adopted by former works [23], [2] for more robust image matching, which assume the local neighborhood structures of two matching feature points should be as similar as possible. Motivated by above idea, we try to learn the neighborhood topology consistent descriptors as Fig. 1(b) by imposing the penalty to the topological difference of matching descriptors.

To this end, we first propose some assumptions about the distribution of descriptors in two learned descriptor sets.
Specifically, for matching descriptors $a_i$ and $p_i$, we assume (1) $k$NN descriptors of $a_i$ match that of $p_i$, (2) topological relationship between $a_i$ and its $k$NN descriptors should be similar with that between $p_i$ and $k$NN descriptors of $p_i$. Former works [23], [25] usually employ the hard weight and heat kernel similarity to indicate the topological relationship between two samples, however, we figure out that they both have their own disadvantages: The hard weight is not differential and the heat kernel similarity consists of tunable hyper-parameter. We then propose the linear combination weights to measure the topological relationship between the center descriptor and its $k$-nearest neighbor ($k$NN) descriptors. Specifically, the difference between center descriptor $a_i$ or $p_i$ and the linear combination of its $k$NN descriptors is minimized, where the linear combination weights have the closed-form solutions because above optimization question is the Least Squares problem.

The linear combination weights are defined in a small local region, then we propose the global mapping function which maps the linear combination weights to the global topology vector. The length of topology vector is equal to $n$, the training batch size, and it is the sparse vector with only $k$ non-zero elements. In this paper, topology distance between two matching descriptors is defined as the $l_1$ distance of their topology vectors, then we can learn topology consistent descriptors by minimizing topology distance of matching descriptors.

To learn more robust and discriminative descriptors, we jointly minimize the topology distance and Euclidean distance of matching descriptors by modifying the distance of positive sample in triplet loss to the weighted sum of topology distance and Euclidean distance. Otherwise, we propose the adaptive weighting strategy to adjust their respective weight: the more stable neighborhood set of descriptor is, the larger weight we assign to the topology distance, where the stability of neighborhood set is proportional to the matching pairs within two neighborhood sets. Our method modifies and consumes the distance measure of positive samples for triplet loss, which means our method can be applied in any other algorithms [19], [21], [20] of learning-based descriptors by minimizing topology distance of matching descriptors. The generalization of our method is verified in several benchmarks in Section IV-C.

The contributions of this paper are four-fold:

- We are the first to consider neighborhood information of learning-based descriptors, where the neighborhood topological relationship between center descriptor and its $k$NN descriptors is depicted by our linear combination weights;
- We propose the global mapping function to map the local linear combination weights to the global topology vector, and then define the topology distance of matching descriptors as $l_1$ distance between their topology vector;
- We propose the adaptive weighting strategy to jointly minimize topology distance and Euclidean distance, which automatically adjust the weight or attention of two distances in triplet loss;
- The experimental results verify the generalization of our method. We test our method on the basis of HardNet [19] and DSM [21], and experimental results show our method can improve their performances in several benchmarks.

The rest of the paper is organized as follows. Section II reviews some related works about the learning-based descriptors and works considering neighborhood information in unsupervised or supervised learning and image matching. Section III presents our proposed method, including our novel linear combination weights, the topology distance and the adaptive weighting strategy. Section IV shows the experimental results, including ablation experiments and extensive experiments in several benchmarks. Last Section V draws the brief conclusions.

II. Related Works

In this section, we first briefly introduce several algorithms on learning descriptors in Section II-A, and then illustrate the effectiveness and wide application of neighborhood consistency in Section II-B.

A. Learning-based Descriptors

Perhaps SIFT [1] is the most successful and widely used handcrafted descriptor, however, all handcrafted descriptors, including SIFT [1], LIOP [26], GLOHP [27], DAISYP [28], DSP-SIFT [29] and BRIEF [11] are not robust enough as they only consider the pixel-level information and neglect the high-level semantic information. With the successful application of deep learning on various fields [30], [31], researchers try to learn descriptors using CNNs directly from the image patch around feature points. DeepCompare [17] and MatchNet [15] learn the pairwise matching probabilities directly from image patches instead of the discriminant descriptors, in which MatchNet uses Siamese CNN with three convolutional layers and DeepCompare explores several CNN architectures. L2Net [18] proposes a deeper CNN with seven convolutional layers to extract the semantic information and a Local Response Normalization layer to normalize descriptors, and this architecture is employed by many works [19], [21], [32] including ours due to its effectiveness.

HardNet [19] first introduces triplet loss [22] to learn descriptors which encourages Euclidean distance of nearest non-matching descriptors is a margin larger than that of matching pairs. CDbin [33] combines triplet loss and other three losses to learn more robust descriptors and explores the performance of descriptors with different lengths. Followed by focal loss [34], Exp-TLoss [20] proposes the exponential triplet loss to focus more on hard positive samples to accelerate CNNs’ training. SOSNet [32] proposes the second-order regularization to learn more robust descriptors by minimizing the edge similarity between matching descriptors. DSM [21] replaces the hard margin in triplet loss by Cumulative Distribution Function-based soft margin which uses more negative samples to update CNNs.

However, above methods which use triplet loss to learn descriptors all focus on the point-to-point Euclidean distance measure and totally neglect the neighborhood information of descriptors. In this paper, we can learn neighborhood consistent descriptors by minimizing topology difference of matching descriptors.
B. Neighborhood Consistency

Neighborhood consistency or local consistency is widely applied in both unsupervised learning and supervised learning. As the classical unsupervised and non-linear data dimensionality reduction methods, manifold learning \cite{37, 38, 24, 39} tries to preserve the local relative relationship of high-dimensional data in the low-dimensional data. In Locally Linear Embedding (LLE) \cite{38}, the local relative relationship indicates neighborhood reconstruction coefficients, which are calculated by minimizing the neighborhood reconstruction error. In Laplacian Eigenmaps (LE) \cite{24}, the local relative relationship indicates the hard weights or heat kernel similarity between the center sample and its neighborhood samples. Recently, Sabokrou \cite{40} proposes the Neighborhood-Relational Encoding for unsupervised representation learning, which determines the loss of Encoder-Decoder structure by the neighborhood relational information. Similarly, Li \cite{25} employs the neighborhood information as the constraint to learn hash representation for large scale image retrieval.

Besides the effective constraints of local or neighborhood consistency in unsupervised learning, neighborhood consistency also shows its privilege in supervised or semi-supervised learning. Belkin \cite{41} proposes an universal semi-supervised framework with the manifold consistency regularization, which is proven effective in Least Squares and Support Vector Machine (SVM). NPNN \cite{42} proposes a non-linear method for data-driven fault detection, which considers the local geometrical structure of training data in neural networks.

PointWeb \cite{43} learns 3D point cloud representation with integrating neighborhood information of each points.

Neighborhood consistency is proven effective in image matching as well. Meng \cite{23} proposes the spatial order constraints bilateral-neighbor vote (SOCBV) to remove outliers with considering the $k$NN feature points of matching pairs. GMS \cite{2} depicts the neighborhood matching by statistical likelihood of the matching number in a region to enable ultra-robust matching. LPM \cite{44} attempts to remove mismatches with local neighborhood structures of potential true matches maintained. Above works mainly focus on the neighborhood topology of matching descriptors. As shown in Fig. 2, compared with DSM \cite{21}, our TCDesc is more robust to the illumination change with less mismatches.

III. METHODOLOGY

In this section, we first propose some assumptions about the distribution of descriptors in Section III-A, which serve as some basic neighborhood constraints to guide our model. Then we present our linear combination weights in Section III-B. In Section III-C, we define the topology vector and topology distance between matching descriptors. Last we propose the adaptive weighting strategy to fuse topology distance and Euclidean distance in Section III-D.
In order to illustrate our model and method more clearly, here we present implications or definitions of some notations:

- $a_i$, $p_j$: The learned descriptors, where descriptors with the same subscript are matching descriptors, such as $a_i$ and $p_i$;
- $t_{i,j}^a$, $t_{i,j}^p$: The topology weights which depict topological relationship between $a_i$ and $a_j$ and that between $p_i$ and $p_j$ respectively.
- $a_{ij}$, $p_{ij}$: The neighborhood descriptors of $a_i$ and $p_i$ respectively.
- $N(a_i)$, $N(p_i)$: The sets of neighborhood descriptors of $a_i$ and $p_i$, which consist of kNN descriptors of $a_i$ and $p_i$ respectively.

### A. Basic Neighborhood Constraints

Recent works [19, 27, 32] adopt L2-Net [18] to learn discriminative descriptors directly from image patches. Specifically, a batch of training data generates the corresponding descriptors \( \chi = \{ A; P \} \), where \( A = \{ a_1, a_2, \ldots, a_n \} \), \( P = \{ p_1, p_2, \ldots, p_n \} \) and \( n \) is the batch size. Normally descriptor vectors are unit-length and 128-dimensional as SIFT [1]. Note that descriptors from two sets with the same subscripts are a matching pair, such as $a_i$ and $p_i$, while descriptors with the different subscripts form the non-matching pairs. During CNNs’ training, Euclidean distance of matching descriptors is the smaller distance contributes the larger weight. For two descriptors $a_i$ and $a_j$, the hard weight $h(a_i, a_j)$ is equal to 1 if $a_j$ is in the neighborhood set of $a_i$ and equal to 0 if not:

\[
h(a_i, a_j) = \begin{cases} 1, & a_j \in N(a_i) \\ 0, & \text{otherwise} \end{cases}
\]  

(2)

The definition of hard weight is very simple without consuming large amount of computation. However, the hard weight only depicts whether $a_j$ is in the neighborhood set of $a_i$ and ignores the relative position relationship between $a_j$ and $a_i$. Otherwise, Eq. 2 is a discrete function, whose gradient is undefined at the boundary of neighborhood set and equal to 0 everywhere else. So a differentiable proxy is required for training purpose if we choose hard weight or binary weight as our topological weight.

#### Heat Kernel Similarity: The heat kernel similarity $s$ is widely adopted by former works [24, 45], which is defined as the exponential value of minus distance between center descriptor $a_i$ and its kNN descriptor $a_j$:

\[
s(a_i, a_j) = \begin{cases} \exp\left[\frac{-||a_i - a_j||^2}{t}\right], & a_j \in N(a_i) \\ 0, & \text{otherwise} \end{cases}
\]  

(3)

In above equation, the smaller distance contributes the larger similarity or topology weight. However, the heat kernel similarity has the following two disadvantages: Firstly, the hyper-parameter $t$ is hard to determined, for example, $s(a_i, a_j)$ would be close to 1 if $t$ is a very large number and close to 0 if $t$ is very small. Secondly, the heat kernel similarity only considers the distance between two descriptors and ignores the relative position. Specifically, $s(a_i, a_j)$ would be equal to $s(a_i, a_k)$ if $||a_i - a_j||_2 = ||a_i - a_k||_2$, even though $a_j$ and $a_k$ lie on the different positions of the hyper-sphere whose center is $a_i$ and radius is $||a_i - a_j||_2$ or $||a_i - a_k||_2$.

In this paper, we propose the linear combination weights to measure the topological relationship between $a_i$ and its kNN descriptors. Obviously our first step is to solve $k$NN descriptors for $a_i$, which are noted as $a_{ij}$ for $j = 1, 2, \ldots, k$. Then we try to linearly fit $a_i$ using $a_{ij}$ so that we minimize difference between $a_i$ and the linear combination of $a_{ij}$:

\[
w_{ij}^a = \arg \min_{w_{ij}^a} \| a_i - \sum_{j=1}^{k} w_{ij}^a a_{ij} \|^2
\]  

(4)

where linear combination weight $w_{ij}^a$ indicates topological relationship between $a_i$ and $a_{ij}$.
When \( w_{ij} \) can be arbitrary real number, \( w_{ij}^p \) denotes a linear space marked as \( \text{span}(a_{i1}, a_{i2}, \ldots, a_{ik}) \), which is the subspace of \( d \)-dimensional Euclidean space, where \( d \) is the length of learned descriptors and equal to 128 in our paper. Our purpose is to find the weights corresponding to the minimum distance between \( a_i \) and \( \text{span}(a_{i1}, a_{i2}, \ldots, a_{ik}) \). However, we may have countless solutions for weights \( w_{ij} \) when \( a_i \) is in \( \text{span}(a_{i1}, a_{i2}, \ldots, a_{ik}) \). To avoid this situation, we stipulate \( k \) much smaller than \( d \).

Assume \( W_i^a = [w_{i1}^a, w_{i2}^a, \ldots, w_{ik}^a]^T \in \mathbb{R}^{k \times 1} \) and \( N_i^a = \{a_{i1}, a_{i2}, \ldots, a_{ik}\} \in \mathbb{R}^{k \times d} \). Now transform Eq. (4) into matrix form:

\[
W_i^a = \arg \min_{W_i^a} \| a_i - N_i^a W_i^a \|^2
\]  

(5)

We note that above equation is the standard Least Squares problem so that \( W_i^a \) has the following closed-form solution:

\[
W_i^a = (N_i^a^T N_i^a)^{-1} N_i^a^T a_i
\]  

(6)

When \( a_{ij} \) is linearly independent with each other and \( k \) is smaller than \( d \), we have \( \text{rank}(N_i^a^T N_i^a) = \text{rank}(N_i^a) = k \), so that \( N_i^a^T N_i^a \in \mathbb{R}^{k \times k} \) is invertible. There is no doubt that we can solve topology weights \( W_i^a \) for descriptor \( p_i \) by the same steps.

Compared with hard weight and heat kernel similarity, our linear combination weights have the following advantages: First, the gradient of Equation (6) is defined everywhere, which contributes to the differential loss function for CNNs’ training. Second, no tunable parameter is required when solving our linear combination weights, which avoids the inappropriate experimental setup by human labor. Last, every linear combination weight \( w_{ij}^a \) retains the information of the whole neighborhood set, specifically, all weights would be affected and changed when a sole descriptor changes.

### C. Topology Measure

In previous section, we present our linear combination weights to depict the topological relationship between the center descriptor and its kNN descriptors. However, the linear combination weights \( W_i^a \) and \( W_i^p \) are defined in the small neighborhood sets of \( a_i \) and \( p_i \), which can not be used to compare the neighborhood difference between \( a_i \) and \( p_i \) as a result from the misalignment of their kNN descriptors.

In this paper, we propose the global mapping which maps the local weights \( W_i^a \) and \( W_i^p \) to the global topology vector \( T_i^a \) and \( T_i^p \). \( T_i^a = [t_{i1}^a, t_{i2}^a, \ldots, t_{in}^a] \) and \( T_i^p = [t_{i1}^p, t_{i2}^p, \ldots, t_{in}^p] \), where \( n \) is the batch size. The \( j \)-th element of \( T_i^a \) can be determined by following equation:

\[
t_{ij}^a = \begin{cases} 
  w_{ij}^a, & a_j \in N(a_i) \\
  0, & \text{otherwise}
\end{cases}
\]  

(7)

Where \( w_{ij}^a \) is one of the elements in \( W_i^a \) and equal to linear combination weight between \( a_i \) and \( a_j \). Note that \( T_i^p \), the topology vector for descriptor \( p_i \), can be established by the same way: the \( j \)-th element of \( T_i^p \) is equal to \( w_{ij}^p \), if descriptor \( p_j \) is one of kNN descriptors of \( p_i \) and equal to 0 if not. By above mapping, the linear combination weights with \( n \) elements. The topology vectors are sparse vectors because they only consist of \( k \) non-zero numbers and \( k \) is much smaller than \( n \).

The global topology vector \( T_i^a \) or \( T_i^p \) indicates the topological relationship between center descriptor \( a_i \) or \( p_i \) and other descriptors within a training batch, then we define topology distance between \( a_i \) and \( p_i \) as the \( l_1 \) distance of their topology vectors:

\[
d_T(a_i, p_i) = \frac{1}{k} \| T_i^a - T_i^p \|_1
\]  

(8)

To learn the topology consistent descriptors, the topology distance between matching descriptors should be minimized during CNNs’ training.

In Eq. (8) \( l_1 \) distance between topology vector \( T_i^a \) and \( T_i^p \) is equal to the sum of element-wise difference in \( T_i^a \) and \( T_i^p \) divided by \( k \). Firstly, when \( a_j \notin N(a_i) \) and \( p_j \notin N(p_i) \), \( t_{ij}^a \) and \( t_{ij}^p \) are both equal to 0. Secondly, when \( a_j \notin N(a_i) \) and \( p_j \in N(p_i) \), \( t_{ij}^a \) is equal to 0 and \( t_{ij}^p \) not. Under this case, absolute value of \( t_{ij}^p \) is encouraged to be closed to 0 with \( d_T(a_i, p_i) \) minimized, which means \( p_j \) is encouraged to be far away with \( p_i \) until \( p_j \) is not in the neighborhood set of \( p_i \). Thirdly, when \( a_j \in N(a_i) \) and \( p_j \in N(p_i) \), \( t_{ij}^a \) and \( t_{ij}^p \) are both non-zero values and \( | t_{ij}^a - t_{ij}^p | \) is minimized, which satisfies our Assumption 2 of Section III-A.

Above analyses show that our topology vector depicts the neighborhood information of descriptor and we can learn the neighborhood topology consistent descriptors by minimizing the distance of topology vectors of matching descriptors. The detailed steps to solve topology distance of the whole training batch are summarized in Algorithm 1.

### D. Adaptive Weighting Strategy

Triplet loss is widely used for learning descriptors by former works [19], [21], [22], [20], which encourages the distance of negative samples is a margin larger than that of positive samples. In descriptors learning, the positive samples and negative samples indicate the matching descriptors and non-

---

**Algorithm 1: Topology Distance**

**Input:** \( A = \{a_1, a_2, \ldots, a_n\} \) and \( P = \{p_1, p_2, \ldots, p_n\} \)

**Output:** Topology distance \( d_T(a_i, p_i) \) for \( i = 1, 2, \ldots, n \)

1. for \( i = 1 : n \) do
2. \( N_i^a \leftarrow \{a_1, a_2, \ldots, a_k\} \); // kNN descriptors of \( a_i \)
3. \( N_i^p \leftarrow \{p_1, p_2, \ldots, p_k\} \); // kNN descriptors of \( p_i \)
4. \( W_i^a \leftarrow (N_i^a^T N_i^a)^{-1} N_i^a^T a_i \)
5. \( W_i^p \leftarrow (N_i^p^T N_i^p)^{-1} N_i^p^T p_i \)
6. \( T_i^a \leftarrow W_i^a \); // global mapping by Equation (7)
7. \( T_i^p \leftarrow W_i^p \).
8. \( d_T(a_i, p_i) \leftarrow \frac{1}{k} \| T_i^a - T_i^p \|_1 \)
9. end
matching descriptors respectively. The triplet loss has the following uniform format:

\[ L_{triplet} = \frac{1}{n} \sum_{i=1}^{n} \max(0, \text{margin} + d_i^+ - d_i^-) \]  

(9)

Former works define \(d_i^+\) and \(d_i^-\) as the Euclidean distance of matching pairs and non-matching pairs respectively, which neglect the neighborhood information of descriptors.

In this paper, we jointly minimize the topology distance and Euclidean distance of matching descriptors to learn topology consistent descriptors:

\[ d_i^+ = \lambda d_T(a_i, p_i) + (1 - \lambda)d_E(a_i, p_i) \]  

(10)

where \(d_E(a_i, p_i)\) denotes Euclidean distance between \(a_i\) and \(p_i\) and \(\lambda\) is the hyper-parameter to balance our topology distance and Euclidean distance.

In Eq. (10) the weight \(\lambda\) is an important parameter that directly affects the performance of descriptors. Note that \(d_i^+\) with smaller \(\lambda\) focuses more on the distance between individual descriptors and contributes to the more discriminative descriptors, while \(d_i^+\) with a larger \(\lambda\) focuses more on the neighborhood information between descriptors and contributes to the more robust descriptors.

In this paper, we employ an adaptive strategy to adjust \(\lambda\) automatically during CNNs' training. In the early training stage, the learned descriptors are not stable enough, so we should focus less on the neighborhood information. In the later training stage, we focus more on neighborhood information to learn more robust descriptors since training CNNs using only point-to-point Euclidean distance will result in overfitting.

Assumption 1 of Section III-A constrain neighborhood descriptors of \(a_i\) match neighborhood descriptors of \(p_i\), however, this perfect matching rarely appears as a result from the inconsistent and unstable neighborhood information of descriptors. In this paper, we employ the number of matching pairs within two neighborhood sets \(N(a_i)\) and \(N(p_i)\) to measure the stability of learned descriptors: The more matching pairs denote the more stable neighborhood information of descriptors. \(N(a_i)\) and \(N(p_i)\) consist of \(k\)NN descriptors of \(a_i\) and \(k\)NN descriptors of \(p_i\) respectively, so the maximum number of matching pairs within \(N(a_i)\) and \(N(p_i)\) is \(k\).

We determine the value of \(\lambda\) by the following formula:

\[ \lambda = \min\{\frac{m(N(a_i), N(p_i))}{k}, 0.5\} \]  

(11)

where \(m(N(a_i), N(p_i))\) denotes the number of matching pairs and \(\gamma\) is the tunable parameter. By this method can we enable the adaptive and automatic adjustment of \(\lambda\) in Eq. (10) during CNNs’ training.

IV. EXPERIMENTS

The main contribution of our work is to propose the topology measure besides Euclidean distance to encourage the similar topology of matching descriptors. In the ablation studies of Section IV-B we choose DSM [21] as our baseline, which is the state-of-the-art method of learning descriptors by triplet loss. And in the extensive experiments of IV-C we test our method on the basis of both HardNet [19] and DSM [21] to verify the generalization of our method, where HardNet first introduces triplet loss into learning descriptors.

To validate the performance of our topology consistence descriptors TCDesc, we conduct our experiments in four benchmarks: UBC PhotoTourism [46], HPatches [36], WIBS dataset [47] and Oxford dataset [48]. UBC PhotoTourism [46] is currently the largest and the most widely used local image patches matching dataset, which consists of three subsets (Liberty, Notre dame and Yosemite) with more than 400k image patches. HPatches [36] presents the more complicated and more comprehensive three tasks to evaluate descriptors: Patch Verification, Image Matching, and Patch Retrieval. WIBS dataset [47] consists of 40 image pairs and provides more challenging tasks with several nuisance factors to explore the performance of descriptors in extreme conditions. Oxford dataset [48] presents the real image matching scenarios, in which matching score is taken to evaluate the performance of learned descriptors.

A. Implementations

We use the same configuration as former works to guarantee the improvement of experimental results attributes to our novel topology measure. We use the CNN architecture proposed in L2-Net [18] with seven convolutional layers and a Local Response Normalization layer. We only train our network on benchmark UBC PhotoTourism and then test other three benchmarks using the trained model. The size of image patches in UBC PhotoTourism is \(64 \times 64\), then we downsample each patch to size of \(32 \times 32\), which is required by of L2-Net. We conduct data augmentation as DSM [21] to flip or rotate image patches randomly. To accord with HardNet [19] and DSM [21], we set the training batch size \(n\) to be 1024. We train our network for 150k iterations using Stochastic Gradient Descent(SGD) with momentum 0.9 and weight decay \(10^{-4}\), and the learning rate is decayed linearly from 0.1 to 0.

B. Ablation Studies

In this section, three ablation experiments are conducted to verify the effectiveness of our method: We first explore the performances of learned descriptors under different \(k\) and \(\gamma\) in Section IV-B1 then compare three categories of topology weights (hard weights, heat kernel similarity and our linear combination weights) in Section IV-B2 last verify the validity of adaptive weighting strategy in Section IV-B3.

1) Impact of \(k\) and \(\gamma\): The hyper-parameter \(k\) denotes the number of descriptors in neighborhood sets of center descriptor \(a_i\) or \(p_i\), while the larger \(k\) means we take more neighborhood information for CNNs’ training, however, the larger \(k\) also results in lager computation and countless solutions for our linear combination weights. In this paper, we explore three values of \(k\): 16, 32 and 64. In Eq. (10) parameter \(\lambda\) denotes the weighting coefficient to fuse topology distance and Euclidean distance, which is decided by \(\gamma\) of Eq. (11) In this paper, we explore three values of \(\gamma\) as well: 0.5, 1 and 2.

We first conduct our experiment on UBC PhotoTourism benchmark [46]. UBC PhotoTourism [46] is the first large
benchmark of learning descriptors from image patches which consists of more than 400k image patches extracted from large 3D reconstruction scenes. UBC PhotoTourism consists of three subsets: Liberty, Notredame and Yosemite. Usually we train on one subset and test on another two subsets. The false positive rate at 95% recall (FPR95) is employed to evaluate the performance of learned descriptors, where the lower FPR95 indicates the better performance.

We choose DSM [21] as our baseline, which is the current state-of-the-art model of learning descriptors by triplet loss. In this experiment, we choose to train our model using Liberty subset and test on another two subsets, and the experimental results are presented in Table. We can see that descriptors learned by our method under different $\gamma$ and $k$ all outperform or perform equally than DSM [21], and we get our lowest FPR95 when $\gamma$ is equal to 1.0 and $k$ is equal to 32.

In Table. we found that descriptors learned under different $k$ or $\gamma$ may result the similar or even the same performance, such as case $k = 16$, $\gamma = 1.0$ and case $k = 64$, $\gamma = 1.0$. We then conduct our experiment on HPatches benchmark [36]. HPatches benchmark [36] consists of 116 sequences where the main nuisance factor of 57 sequences is illumination and that of 59 sequences is viewpoint. Compared with UBC PhotoTourism benchmark, HPatches benchmark [36] provides more diverse data samples and more sophisticated tasks. HPatches [36] defines three tasks to evaluate descriptors: Patch Verification, Image Matching, and Patch Retrieval, and each task is categorized as “Easy”, “Hard” or “Tough” according to the amount of geometric noise or changes in viewpoint and light illumination. The mean average precision(mAP) is adopted to evaluate descriptors and the higher mAP indicates the better performance.

We use model trained on subsets Liberty of UBC PhotoTourism benchmark to generate descriptors from image patches of HPatches. The experimental results are presented in Fig. as can be seen, the larger $\lambda$ contributes better performance of descriptors on task Patch Verification, while the smaller $\lambda$ contributes better performance of descriptors on task Image Matching. Otherwise, performance of descriptors is not relevant to the value of $k$ in three tasks of HPatches benchmark.

With the comprehensive consideration on both UBC PhotoTourism benchmark and HPatches benchmark, we determine the optimal value of $k$ and $\gamma$: $k = 16$ and $\gamma = 1.0$. Under this setup, the mean FPR95 on UBC PhotoTourism benchmark is 0.85% and mAPs on three tasks of HPatches benchmark are 88.23%, 51.72% and 70.49% respectively.

| $\gamma$ | $k$ | train | Liberty | Notre dame | Yosemite | mean |
|---|---|---|---|---|---|---|
| DSM [21] | 0.39 | 0.31 | 0.95 |
| 0.5 | 32 | 0.40 | 0.36 | 0.89 |
| 64 | 0.39 | 0.47 | 0.93 |
| 1.0 | 32 | 0.37 | 1.33 | 0.85 |
| 64 | 0.32 | 1.37 | 0.85 |
| 2.0 | 32 | 0.36 | 1.54 | 0.95 |
| 64 | 0.35 | 1.37 | 0.86 |

2) Comparison of three categories of topology weights: In Section III-B we mentioned that the hard weights and heat kernel similarity are widely adopted to measure the topological relationship between two samples and then we figure out their drawbacks. In this section we compare our linear combination weights and the former two topology weights on both UBC PhotoTourism benchmark and HPatches benchmark. For the fair comparison, we set $k$ to be 16 and set $\gamma$ to be 1.0 in all experiments of this section.

The hard weights of Equation 2 is not differential, which brings the obstacle to optimize the loss function. So we choose a differential proxy to approximate the hard weights:

$$h(a_i, a_j) = \begin{cases} 
\exp(-\frac{||a_i - a_j||_2}{10^3}), & a_j \in N(a_i) \\
0, & \text{otherwise}
\end{cases} \tag{12}$$

In above equation, $a_i$ and $a_j$ are both unit-length vector so that $||a_i - a_j||_2$ is a very small number and $\exp(-\frac{||a_i - a_j||_2}{10^3})$ is approximately equal to 1.

Otherwise, heat kernel similarity of Eq. 3 consists of a hyper-parameter $t$, where the larger $t$ results the lager similarity score. In this section, we test five values for $t$: 0.1, 0.5, 1.0, 5.0 and 10.0. We present the experimental results in Table III. We conclude that our linear combination weights work well on
adaptive weighting strategy and fixed λ and compare the performance of descriptors learned by our weight we propose the adaptive weighting strategy to adjust the weight λ of Eq 10 automatically. In this section, we explore and compare the performance of descriptors learned by our adaptive weighting strategy and fixed λ on UBC PhotoTourism benchmark [10]. We test six values for λ: 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5.

For the fair comparison, we conduct our experiments using the same settings as Section IV-B1 except the value of λ. As shown in Table V descriptors learned by our adaptive weighting strategy outperforms than that learned by any fixed λ: we achieve the lowest mean FPR95 0.85%.

C. Extensive Experiments

1) UBC PhotoTourism benchmark: As illustrated above, UBC PhotoTourism [10] consists of Liberty, Notredame and Yosemite three subsets, while we train on one subset and test on other two. In the former section, we only train on subset Liberty. And in this section, we conduct our experiments on all tasks.

We test our method on the basis of HardNet [19] and DSM [21], which are noted as TCDesc-HN and TCDesc-DSM respectively. Specifically, we modify the distance of positive sample in their triplet losses as the adaptive weighting strategy outperforms than that learned by any fixed λ. As shown in Table V descriptors learned by our adaptive weighting strategy outperforms than that learned by any fixed λ: we achieve the lowest mean FPR95 0.85%.

We test our method on the basis of HardNet [19] and DSM [21], which are noted as TCDesc-HN and TCDesc-DSM respectively. Specifically, we modify the distance of positive sample in their triplet losses as the adaptive weighting strategy outperforms than that learned by any fixed λ. As shown in Table V descriptors learned by our adaptive weighting strategy outperforms than that learned by any fixed λ: we achieve the lowest mean FPR95 0.85%.

C. Extensive Experiments

1) UBC PhotoTourism benchmark: As illustrated above, UBC PhotoTourism [10] consists of Liberty, Notredame and Yosemite three subsets, while we train on one subset and test on other two. In the former section, we only train on subset Liberty. And in this section, we conduct our experiments on all tasks.

We test our method on the basis of HardNet [19] and DSM [21], which are noted as TCDesc-HN and TCDesc-DSM respectively. Specifically, we modify the distance of positive sample in their triplet losses as the adaptive weighting strategy outperforms than that learned by any fixed λ: we achieve the lowest mean FPR95 0.85%.

C. Extensive Experiments

1) UBC PhotoTourism benchmark: As illustrated above, UBC PhotoTourism [10] consists of Liberty, Notredame and Yosemite three subsets, while we train on one subset and test on other two. In the former section, we only train on subset Liberty. And in this section, we conduct our experiments on all tasks.

We test our method on the basis of HardNet [19] and DSM [21], which are noted as TCDesc-HN and TCDesc-DSM respectively. Specifically, we modify the distance of positive sample in their triplet losses as the adaptive weighting strategy outperforms than that learned by any fixed λ: we achieve the lowest mean FPR95 0.85%.
As can be seen, our novel topology measure improves performance of both descriptors learned by HardNet and DSM. Specifically, mean FPR95 of HardNet declines from 1.51 to 1.22 and that of DSM declines from 1.18 to 1.08 after introducing our topology measure. Furthermore, our method reduces the FPR95 of HardNet and DSM on every testing task. Otherwise, as presented in Table.\[11\] our TCDesc on the basis of DSM leads the state-of-the-art result with the lowest FPR95 1.08. The experimental results on UBC PhotoTourism benchmark validate the generalization of our method: We can improve performances of several descriptors learned by former triplet loss.

2) HPatches benchmark: In this section, we test the performance of our TCDesc-HN and TCDesc-DSM on HPatches benchmark \[36\]. We use model trained on subsets Liberty of UBC PhotoTourism benchmark to generate descriptors from image patches of HPatches. We compare our topology consistent descriptors TCDesc-HN and TCDesc-DSM with SIFT \[1\], HardNet \[19\], DOAP \[49\], SOSNet \[32\], Exp-TLoss \[20\] and DSM \[21\], where our descriptors TCDesc-HN and TCDesc-DSM are trained on the basis of HardNet \[19\] and DSM \[21\] respectively.

As can be seen in Fig.\[3\] there only exists a small margin among mAP of various learning-based descriptors in three tasks. In task Patch Verification, our TCDesc-HN performs a little worse than DSM, and TCDesc-HN performs better than HardNet. In task Image Matching, our TCDesc-DSM and TCDesc-HN lead the state-of-the-art results and perform much better than DSM and TCDesc-HN, which proves the effectiveness of our topology consistent descriptors in image...
matching. In task Patch Retrieval, our TCDesc-DSM and TCDesc-HN both outperform than DSM and TCDesc-HN, and the TCDesc-DSM achieves the highest mAP (70.50) in this task.

3) Wide baseline stereo: Wide baseline stereo matching aims to find correspondences of two images in wide baseline setups, i.e., cameras with distant focal centers. So it is more challenging than normal image matching. To verify generalization of our TCDesc and prove its advantages in extreme conditions, we conduct our experiments on WIBS benchmark. WIBS dataset consists of 40 image pairs divided into 5 parts by the nuisance factors: Appearance, Geometry, Illumination, Sensor, and Map to photo.

WIBS database uses multi detectors MSER, Hessian-Affine and FOCI to detect affine-covariant regions and normalize the regions to size 41 × 41. The average recall on ground truth correspondences of image pairs are employed to evaluate the performance of descriptors.

We compare our TCDesc-HN and TCDesc-DSM with SIFT, HardNet, SOSNet, Exp-TLoss and DSM. Like the former experiment, we use the model trained on subsets Liberty of UBC PhotoTourism benchmark to generate descriptors. The experimental results are presented in Fig. where the larger mAUC indicates the better performance. The mean mAUC of our TCDesc-HN is 8.78%, which denotes the state-of-the-art performance. And the mean mAUC of our TCDesc-DSM is 8.30%, which is larger than that of DSM 8.12%. Conclusion could be drawn that our method can also improve performance of descriptors in extreme condition.

4) Image Matching on Oxford Dataset: Oxford dataset presents the real image matching scenarios, which takes in many nuisance factors including blur, viewpoint change, light change and compression. Oxford dataset only consists of 64 images but it is widely used to evaluate the robustness of image matching. In this paper, we detect feature points by Harris-Affine detector, and we extract no more than 500 feature points for each image. So we define the matching score as the right matches divided by 500, where the larger matching score denotes the better performance. We compare our TCDesc-HN and TCDesc-DSM with HardNet, SOSNet, Exp-TLoss and DSM. We did not compare our method with SIFT because SIFT can not extract a fixed number of feature points as Harris-Affine.

We present our experimental results in Table. As can be seen, our TCDesc-HN leads the state-of-the-art result (36.23% matching score) and our TCDesc-DSM+ (35.91% matching score) outperform than DSM+ (35.96% matching score).

V. CONCLUSIONS

We observe the former triplet loss fails to maintain the similar topology between two descriptor sets since it takes the point-to-point Euclidean distance among descriptors as the only measure. Inspired by the idea of neighborhood consistency of feature points in image matching, we try to learn neighborhood topology consistent descriptors by introducing a novel topology measure.

We first propose the linear combination weight to depict the topological relationship between center descriptor and its kNN descriptors, which is taken as the local topology weights. Then we propose the global mapping function which maps the local topology weights to the global topology vector. Topology distance between two matching descriptors is defined as the l1 distance between their topology vector. Last we propose the adaptive weighting strategy to jointly minimize topology distance and Euclidean distance of matching descriptors.

Experimental results on several benchmarks validate the generalization of our method since our method can improve performance of several algorithms using triplet loss.

However, our method is not appropriate for learning binary descriptors because the binary descriptor can not be linearly fitted by its kNN descriptors with float fitting weights. We note that the idea of our method, local or neighborhood consistency can be extended to many other fields like cross-modal retrieval and etc.

ACKNOWLEDGMENT

This study was supported by the National Natural Science Foundation of China (Grant No.61672183), by the Shenzhen Research Council (Grant No. JCYJ201704131045522946, JCYJ20170815113552036), and in part by funding from Shenzhen Institute of Artificial Intelligence and Robotics for Society (Grant No.2019-INT021).

REFERENCES

[1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.
[2] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, “Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4181–4190.
[3] B. Jiang, J. Tang, and B. Luo, “Feature matching with intra-group sparse model,” IEEE Transactions on Multimedia, 2019.
Honghu Pan received a master degree from Chongqing University in 2019. He is currently pursuing a Ph.D degree with the Department of Computer Science, Harbin Institute of Technology (Shenzhen). His research interests include computer vision, image matching and metric learning.

Fanyang Meng received the Ph.D. degree from the School of EE&CS, Shenzhen University, China, in 2016, under the supervision of Prof. X. Li. He currently serves as a assistant research fellow in Peng Cheng Laboratory. His research interests include video coding, computer vision, human action recognition, and abnormal detection using RGB, depth, and skeleton data.

Nana Fan received the MS degree from the Harbin Institute of Technology Shenzhen Graduate School. She is currently pursuing the PhD degree with the Department of Computer Science, Harbin Institute of Technology (Shenzhen). Her research interests include visual tracking, deep learning, and machine learning.

Zhenyu He received his Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2007. From 2007 to 2009, he worked as a postdoctoral researcher in the department of Computer Science and Engineering, Hong Kong University of Science and Technology. He is currently a full professor in the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His research interests include machine learning, computer vision, image processing and pattern recognition.