ContextDesc: Local Descriptor Augmentation with Cross-Modality Context

Zixin Luo[1] Tianwei Shen[1] Lei Zhou[1,*] Jiahui Zhang[2]
Yao Yao[1,*] Shiwei Li[1] Tian Fang[3,†] Long Quan[1]
[1]Hong Kong University of Science and Technology
[2]Tsinghua University [3]Shenzhen Zhuke Innovation Technology (Altizure)
{zluoag,tshenaa,lzhouai,yyaoag,slibc,quan}@cse.ust.hk
jiahui-z15@mails.tsinghua.edu.cn fangtian@altizure.com

Abstract

Most existing studies on learning local features focus on the patch-based descriptions of individual keypoints, whereas neglecting the spatial relations established from their keypoint locations. In this paper, we go beyond the local detail representation by introducing context awareness to augment off-the-shelf local feature descriptors. Specifically, we propose a unified learning framework that leverages and aggregates the cross-modality contextual information, including (i) visual context from high-level image representation, and (ii) geometric context from 2D keypoint distribution. Moreover, we propose an effective N-pair loss that eschews the empirical hyper-parameter search and improves the convergence. The proposed augmentation scheme is lightweight compared with the raw local feature description, meanwhile improves remarkably on several large-scale benchmarks with diversified scenes, which demonstrates both strong practicality and generalization ability in geometric matching applications. [code release]

1. Introduction

Designing powerful local feature descriptor is a fundamental problem in applications such as panorama stitching [21], wide-baseline matching [24, 54, 55], image retrieval [27] and structure-from-motion (SfM) [57, 39, 52, 56]. Despite the recent notable achievements, the performance of state-of-the-art learned descriptors is observed to be somewhat saturated on standard benchmarks. As shown in Fig. 1a, due to repetitive patterns, the matching algorithm often finds false matches as nearest neighbors that are visually indistinguishable from groundtruth, unless validated by geometry. Essentially, such visual ambiguity may not be easily resolved given only local information. In this spirit, we seek to enhance the local feature description with extra prior knowledge, which we refer to as introducing *context awareness* to augment local feature descriptors.

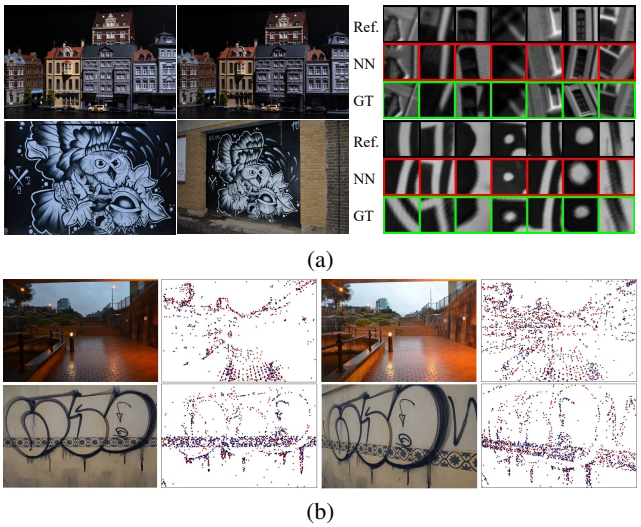


Figure 1: (a) Saturated results on standard benchmark [2] by a recent method [23]. The search of nearest neighbors (NN) returns false matches though visually similar to groundtruth (GT), indicating the limitation of relying on only local visual information. (b) 2D keypoints distribute structurally, on which we human beings are capable of establishing coarse matches even without color information.

As a common practice, a multi-scale-like architecture can help to capture *visual context* of different levels, which is referred to as multi-scale domain aggregation by DSP-SIFT [8] and adopted by recent learned descriptors [50, 19, 43]. Beside of the challenge on selecting proper domain sizes, a naïve multi-scale implementation may cost excessive computation such as doubled inference time and doubled feature dimensionality [50, 19, 43]. Seeking for more reasonable accuracy-efficiency trade-offs, we instead resort to well-studied high-level image representation, e.g., the regional representation used by image retrieval stud-

*Interns at Shenzhen Zhuke Innovation Technology (Altizure).
†Corresponding author.

ies [33, 38] which essentially incorporates rich image context. Thereby, we strive to effectively combine the local feature description and off-the-shelf visual understandings so as to go beyond the local detail representation.

In addition, it would be interesting to exploit context in other modality. In particular, as shown in Fig. 1b, since keypoint is principally designed to be repeatable in the same underlying scene, its distribution thus reveals comprehensive scene structure that allows we human beings to establish coarse matches even without color information, which further enables us to explore *geometric context* formed by the spatial relations of keypoints to help to alleviate the visual ambiguity of local descriptions.

Thus far, we have discussed two context candidates, referred to as *visual context* and *geometric context* that incorporate high-level visual representation over the image and geometric cues from 2D keypoint distribution, respectively. Instead of learning a completely new descriptor, in the present work, we target to flexibly leverage the above context awareness to augment off-the-shelf local descriptors without altering their dimensionality, in which process we consider the key challenges threefold:

- A proper integration of geometric local feature and semantic high-level representation. As keypoint description requires sub-pixel accuracy, the integration is not supposed to obscure the raw representation of local details.
- The instability of 2D keypoint distribution. Due to image appearance changes, keypoint distribution often suffers from substantial variations of sparsity, non-uniformity or perspective, which raises difficulties on acquiring strong invariance property of the feature encoder.
- An effective learning scheme. Input signals and features in different modalities are supposed to be efficiently processed and aggregated in a unified framework.

Finally, regarding practicability, the augmentation is not supposed to introduce excessive computational cost, as the local feature description is often regarded as part of preprocessing in practical pipelines.

Although contextual information has been widely explored in semantic-based tasks, the challenges faced by local feature learning are substantially different, posing many non-trivial technical and systematic issues to overcome. In this paper, we propose a unified augmentation scheme that effectively leverages and aggregates cross-modality context, of which the contributions are summarized threefold: 1) a novel *visual context encoder* that integrates high-level visual understandings from *regional image representation*, a technique often used by image retrieval [33, 38]. 2) A novel *geometric context encoder* that consumes unordered points and exploits geometric cues from 2D keypoint distribution, while being robust to complex variations. 3) A novel N-pair loss that requires no manual hyper-parameter search and has better convergence properties. To our best knowledge, it is the first work that emphasizes the importance of context awareness, and in particular addresses the usability of spatial relations of keypoints in local feature learning.

The proposed augmentation is extensively evaluated and achieves state-of-the-art results on several large-scale benchmarks, including patch-level homography dataset, image-level wild outdoor/indoor scenes and application-level 3D reconstruction image sets, while being lightweight compared with raw local description, demonstrating both strong generalization ability and practicability.

2. Related Work

Learned local descriptors. Initially, local descriptors are jointly learned with a new comparison metric [9, 50], which is later simplified as direct comparison in Euclidean space [40, 48, 3, 19, 1]. More recently, efforts are spent on efficient training data sampling [43, 25, 11], effective regularizations [43, 53], and geometric shape estimation of input patches [26, 7]. However, most of above methods take individual image patches as input, whereas in the present work, we aim to take advantage of contextual cues beyond the local detail and incorporate features in multiple modalities.

Context awareness. Although widely introduced in computer vision tasks, context awareness has received little attention in learning 2D local descriptors. In terms of visual context, the central-surround (CS) structure [50, 19, 43] leverages multi-scale information by additionally feeding the central part of patches to boost the performance, whereas sacrificing computational efficiency due to the doubled extraction time and feature dimensionality. To incorporate semantics, one previous practice [18] designs a new comparison metric and describes features from histogram of semantic labels. In contrast to geometric matching, a family of studies has focused on finding semantic correspondences [45, 34] across different objects of the same category. Beside of visual information, a recent study [49] explores to encode motion context for identifying outliers from keypoint matches, i.e., 4-d coordinate pairs, while we aim to exploit geometric context from single image without any reference. Overall, encoding proper context is non-trivial and still unclear in 2D local feature learning.

Point feature learning. In the present work, one of our goals is to explore geometric features from keypoint distribution, we thus resort to PointNet [31] and its variants [32, 5, 49] to consume unordered points. Although great success has been witnessed in learning tasks on 3D points, there are only few studies exploiting the potential outcome of 2D keypoint sets. In essence, keypoint structure is not intuitively meaningful and robust, as being highly dependent on the performance of interest point detectors and strongly affected by image variations. However, in descrip-

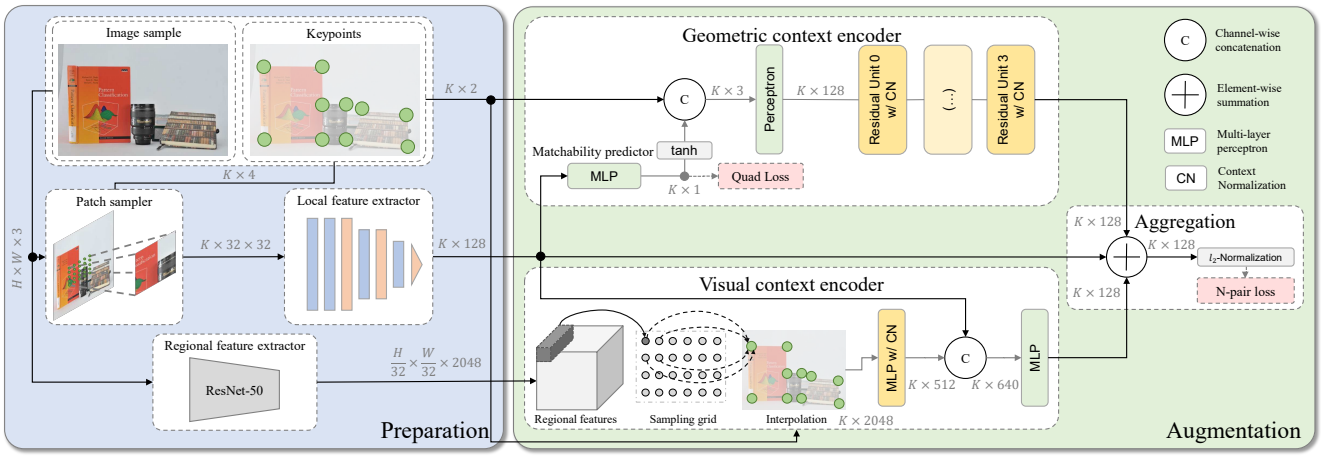


Figure 2: The proposed augmentation framework consumes a single image as input, from which 2D keypoints, local and regional features are extracted and encoded as geometric and visual context to improve the raw local feature description.

tor learning, we consider the keypoint location as an important cue that bridges each individual local feature that has potentials to alleviate the local visual ambiguity.

Loss formulation. Recent local descriptors are often evolved with advanced variants of N-pair losses. Initially, L2-Net [43] adopts a log-likelihood formulation, which is later extended by HardNet [25] with a subtractive hinge loss. Furthermore, GeoDesc [23] applies an adaptive margin to improve the convergence in terms of different hard negative mining strategies, while AffNet [7] approaches the same issue by fixing the distance to hardest negative sample during training. Meanwhile, on the other hand, DOAP [11] extends the N-pair loss to a list-wise ranking loss, while [17] points out and studies the scale effects in N-pair losses while introducing additional manual tuning of hyper-parameters. Principally, a good loss is supposed to encourage similar patches to be close while dissimilar ones to be distant in the descriptor space. In this spirit, we aim to further resolve the scale effects in [17] in an self-adaptive manner, without the need of complex heuristics or manual tuning.

3. Local Descriptor Augmentation

Overview. As illustrated in Fig. 2, the proposed framework consists of two main modules: *preparation* (left) and *augmentation* (right). The *preparation* module provides input signals in different modalities (raw local feature, high-level visual feature and keypoint location), which are then fed to the *augmentation* module and aggregated into compact feature descriptions. At test time, the augmentation needs to be performed once per image, resulting in K feature vectors for K corresponding keypoints.

3.1. Preparation

Patch sampler. This module takes images and their keypoints as input, producing 32×32 gray-scale patches. Akin to [48, 23], image patches are sampled by a spatial transformer [16], whose parameters are derived from keypoint attributes (coordinates, orientation and scale) from the SIFT detector. As a result, the sampled patch has the same support region size with the SIFT descriptor.

Local feature extractor. This module takes image patches as input, producing 128-d feature descriptions as output. We borrow the lightweight 7-layer convolutional networks as used in several recent works [43, 25, 23].

Regional feature extractor. In contrast to aggregating features of different domain sizes [50, 19, 43], in the present work, we fix the sampling scale of patches, and exploit contextual cues by inspiration of well-studied regional representation in image retrieval tasks [44, 33, 28]. Without the loss of generality, we reuse features from an off-the-shelf deep image retrieval model of ResNet-50 [12]. As in [44], feature maps are extracted from the last bottleneck block, across which each response is regarded as a regional feature vector effectively corresponding to a particular region in the image. As a result, we derive regional features of $\frac{H}{32} \times \frac{W}{32} \times 2048$, where H and W denote the original image height and width. The aggregation of regional and local features will be later discussed in Sec. 3.3.

3.2. Geometric context encoder

This module takes K unordered points as input, and outputs 128-d corresponding feature vectors. Each input point is represented as 2D keypoint coordinate, and can be associated with other attributes.

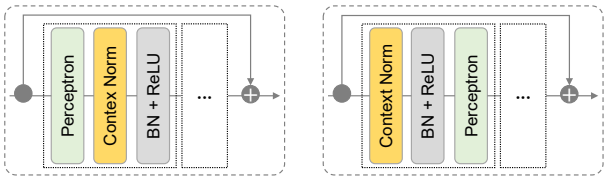
2D point processing. At first glance, 2D keypoints are inappropriate to serve as robust contextual cues, as its presence is heavily dependent on image appearance and thus affected by various image variations. As a result, keypoint distribution depicting the same scene may suffer from significant density or structure variations, as examples shown in Fig. 1b. Hence, acquiring strong invariance property is

the key challenge when designing the context encoder.

Initially, we attempt to approach the goal by PointNet [31] and its variants [32, 5]. Although having shown great success on processing 3D point clouds, those prevalent PointNet methods fails to achieve consistent improvement in terms of 2D points processing (Sec. 4.4.1). Instead, we resort to [49], in which context normalization (CN) is equipped in PointNet and consumes putative matches (4-d coordinate pairs) for outlier rejection in image matching. In this work, we aim to further explore the usability of CN for modeling 2D point distribution in single image.

Formally, CN is a non-parametric operation that simply normalizes feature maps according to their distribution, written as $\hat{\boldsymbol{o}}_i^l = \frac{(\boldsymbol{o}_i^l - \boldsymbol{\mu}^l)}{\boldsymbol{\sigma}^l}$, where $\boldsymbol{o}_i^l$ is the output of i-th point in layer l, and $\boldsymbol{\mu}^l, \boldsymbol{\sigma}^l$ are mean and standard deviation of the output in layer l. To equip the operation, we borrow the residual architecture in [49], where each residual unit is built with perceptrons followed by context and batch normalization, as illustrated in Fig. 3a.

However, the above design leads to a *non-negative* output from the residual branch that may impact the representational ability as investigated in [13] and witnessed in our experiments (Sec. 4.4.1). Following the teachings of [13], we re-arrange the operations in each residual unit with *pre-activation*, which is compatible with CN as presented in Fig. 3b. We then construct four such units for the encoder, as shown in Fig. 2. We will show that this simple revision plays an important role to ease the optimization.



(a) Original design (b) Re-arrange w/ pre-activation

Figure 3: Different designs of residual unit with context normalization, where re-arranging with *pre-activation* improves by a considerable margin than its counterpart.

Intuitively, the non-parametric CN suffices to model the keypoint distribution in our task, while high-level abstractions (e.g., in PointNet++ [32]) may not be necessary.

Matchability predictor. In 3D point cloud processing, low-level color and normal [31] information or complex geometric attributes [5] are often incorporated to enhance the representation. Similarly, associating 2D coordinate input with other meaningful attributes would be promising to boost the performance. However, due to the substantial variations, e.g. perspective change, it is non-trivial to define appropriate intermediate attributes on 2D points.

Although this issue has been merely discussed, we draw inspiration from [10], which poses a problem named *matchability prediction* that targets to decide *whether a keypoint descriptor is matchable before the matching stage*. In practice, the matchability serves as learned attenuation to diversify the keypoints, so that the feature encoder can implicitly focus on the points that are more robust, i.e., matchable, in order to improve the invariance property.

In the present work, we approach the matchability prediction with deep learning techniques instead of a random forest in [10], and constrain the prediction to be consistent between images. Inspired by learning-based keypoint detection methods [35, 51], we resort to an unsupervised learning scheme that aims to appropriately rank points by their matchability. Formally, given K correspondences (p_1^n, p_2^n), $n \in [1, K]$ from an image pair, we first extract their local features $(\boldsymbol{f}_1^n, \boldsymbol{f}_2^n)$, then construct *feature quadruples* as $(\boldsymbol{f}_1^i, \boldsymbol{f}_1^j, \boldsymbol{f}_2^i, \boldsymbol{f}_2^j)$, satisfying $i, j \in [1, K], i \neq j$ and holding that:

$$\begin{cases} H(\boldsymbol{f}_1^i) > H(\boldsymbol{f}_1^j) & \& \quad H(\boldsymbol{f}_2^i) > H(\boldsymbol{f}_2^j) \\ & \text{or} \\ H(\boldsymbol{f}_1^i) < H(\boldsymbol{f}_1^j) & \& \quad H(\boldsymbol{f}_2^i) < H(\boldsymbol{f}_2^j) \end{cases}, \tag{1}$$

where $H(:)$ absorbs the raw local feature into a single real-valued matchability, implemented as standard multi-layer perceptrons (MLPs). Here, Cond. 1 aims to preserve a ranking of each keypoint, hence improves the repeatability of prediction. The condition can be re-written as:

$$\begin{aligned} &R(\boldsymbol{f}_1^i, \boldsymbol{f}_1^j, \boldsymbol{f}_2^i, \boldsymbol{f}_2^j) = \\ &(H(\boldsymbol{f}_1^i) - H(\boldsymbol{f}_1^j))(H(\boldsymbol{f}_2^i) - H(\boldsymbol{f}_2^j)) > 0, \end{aligned} \tag{2}$$

the final objective can be obtained with a hinge loss:

$$\mathcal{L}_{quad} = \frac{1}{K(K-1)} \sum_{i,j,i \neq j} \max(0, 1 - R(\boldsymbol{f}_1^i, \boldsymbol{f}_1^j, \boldsymbol{f}_2^i, \boldsymbol{f}_2^j)). \tag{3}$$

In the proposed framework, the matchability is learned as an auxiliary task, which is then activated by $\tanh$ and associated with keypoint coordinates as the network input, as in Fig. 2. Beside of Eq. 3, the gradient from final augmented features will flow through the matchability predictor, allowing a joint optimization of the entire encoder. The visualization of predicted matchability is shown in Fig. 4.

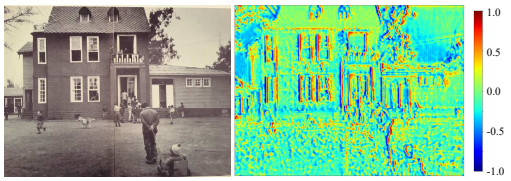


Figure 4: Visualization of matchability responding to the entire image (best viewed in color).

3.3. Visual context encoder

This module consumes regional features of $\frac{H}{32} \times \frac{W}{32} \times 2048$ in Sec. 3.1, K local features and their location, and produces K augmented features. To integrate visual information in different levels, a valid option as in [5] is to concatenate the global representation of entire image on raw local features. In our framework, the global feature can be derived by applying Maximum Activations of Convolutions (MAC) aggregation [33], which simply max-pools over all dimensions of regional features. However, such compact representation is shown to obscure the raw local description, due to the lack of spatial distinctions (Sec. 4.4.1). Hence, we stick to the regional representation, where the key issue is to handle the regional features and keypoints of different numbers ($\frac{H}{32} \times \frac{W}{32}$ and K).

To achieve the goal, we associate regional features to a regular sampling grid on the image, then interpolate $\frac{H}{32} \times \frac{W}{32}$ grid points at coordinates of the K keypoints. For interpolation, we use inverse distance weighted average based on k nearest neighbors (in default we use $k = 3$), formulated as:

$$\boldsymbol{f}(\hat{p}_i) = \frac{\sum_{j=1}^{k} w(p_j)\boldsymbol{f}(p_j)}{\sum_{j=1}^{k} w(p_j)}, \text{and } w(p_j) = \frac{1}{d(\hat{p}_i, p_j)}, \quad (4)$$

where $\boldsymbol{f}(:)$ is the regional feature located at a certain grid point. $\hat{p}_i$, $i \in [1, N]$ and p_j, $j \in [1, \frac{H}{32} \times \frac{W}{32}]$ indicate the interpolated and original grid point. Next, the dimensionality is reduced by applying point-wise MLPs, where we also insert CN after each perceptron in order to capture global context. Finally, raw local features are concatenated and further mapped by MLPs, forming the final 128-d features. The above process is illustrated in Fig. 2.

3.4. Feature aggregation with raw local feature

To aggregate the above two types of contextual features, similar to the CS structure, one option is to concatenate them together and forms features of, in our case, 384-d (128×3). However, the increased dimensionality will introduce excessive computational cost in the matching stage of $\mathcal{O}(n^2)$ complexity. Instead, as shown in Tab. 2, we propose to combine different feature streams into a single vector by element-wise summation and L2-normalization, i.e., without altering the feature dimensionality. Beside of the simplicity, such strategy allows flexible use of the proposed augmentation. For example, in situations where regional features are not available, one may aggregate with only geometric context without the need of retraining the model.

3.5. N-pair loss with softmax temperature

N-pair losses have been primarily used by recent works. Empirically, the subtractive hinge loss [25, 23, 7] has reported better performance, of which the main idea is to push similar samples away from dissimilar ones to a certain *margin* in the descriptor space. However, setting the appropriate margin is tricky, which does not always assure convergence as observed in [23, 7]. More generally, the criteria of making a good loss is studied in [17], from which guidelines are provided on tuning loss parameters on a particular dataset. In this spirit, we aim to further ease the pain of parameter search in [17], and obtain an adaptive loss that allows fast convergence regardless of the learning difficulty.

We use the log-likelihood form of N-pair loss [43] as a base, which originally does not involve any tunable parameter. Formally, given L2-normalized feature descriptors $\mathbf{F}_1 = [\boldsymbol{f}_1^1 \boldsymbol{f}_1^2 ... \boldsymbol{f}_1^N]^T, \mathbf{F}_2 = [\boldsymbol{f}_2^1 \boldsymbol{f}_2^2 ... \boldsymbol{f}_2^N]^T \in \mathbb{R}^{N \times 128}$, the distance matrix $\mathbf{D} = [d_{ij}]_{N \times N}$ can be obtained by $\mathbf{D} = \sqrt{2(1 - \mathbf{F}_1 \mathbf{F}_2^T)}$. By applying both row-wise (r) and column-wise (c) softmax, we derive the final loss as:

$$\begin{aligned} \mathcal{L}_{N\text{-}pair} &= -\frac{1}{2}\left(\sum_i \log s_{ii}^r + \sum_i \log s_{ii}^c\right), \\ \text{where } & [s_{ij}]_{N \times N} = \text{softmax}(2 - \mathbf{D}). \end{aligned} \quad (5)$$

Noted that since input features are L2-normalized, the resulting d_{ij} is bounded by $[0, 2]$, which causes convergence issues due to the scale sensitivity of softmax function [15]. Similarly, we introduce a single trainable parameter α, referred to as *softmax temperature*, to amend the inability of re-scaling the input. The loss now becomes:

$$[s_{ij}]_{N \times N} = \text{softmax}(\alpha(2 - \mathbf{D})), \quad (6)$$

where α is initialized to 1 and regularized with the same weight decay in the network, hence does not require any manual tuning or complex heuristics. In the experiments in Sec. 4.4.2, we show this simple technique improves drastically than its original form [43], whose performance we suspect is hindered due to the above-mentioned scale sensitivity. In the proposed framework, we compute the N-pair loss on augmented features, and obtain the total loss:

$$\mathcal{L}_{total} = \mathcal{L}_{N\text{-}pair} + \lambda \mathcal{L}_{quad}, \quad (7)$$

where we choose $\lambda = 1$ in the experiment.

4. Experiments

4.1. Implementation

Training details. Although the framework is end-to-end trainable, we *fix* the local and regional feature extractors in Sec. 3.1 during the training, in order to clearly demonstrate the efficacy of the proposed augmentation scheme. We train the networks using SGD with a base learning rate of 0.05, weight decay of 0.0001 and momentum at 0.9. The learning rate exponentially decays by 0.1 for every 100k steps. The batch size is set to 2, and each time 1024 keypoints are

randomly sampled including random numbers of matchable and noisy keypoints (see Appendix A.1). Input patches are standardized to have zero mean and unit norm, while input keypoint coordinates are normalized to $[-1, 1]$ regarding the image size.

Training dataset. Although UBC Phototour [4] is used as a common practice, this dataset consists of only three scenes with limited diversity of keypoint distribution. In order to achieve better generalization ability, we resort to large-scale photo-tourism [46, 33] and aerial datasets (GL3D) [38] as in [48, 23], and generate groundtruth matches from SfM. We manually exclude the data that is used in the evaluation.

Data augmentation. We randomly perturb input patches by affine transformations including rotation (90°), anisotropic scaling and translation w.r.t. the detection scale. For keypoint augmentation, we perturb the coordinate with random homography transformation as in [6] (see Appendix A.1).

4.2. Evaluation datasets

Homography dataset. HPatches [2] is a large-scale patch dataset for evaluating local features regarding illumination and viewpoint changes. As groundtruth homographies and raw images are provided, HPatches can also be used to evaluate image matching performance, which we accordingly refer to as HPSequences as in [20], consisting of 116 sequences and 580 image pairs.

Wild dataset. Similar to settings in [49], we also evaluate on outdoor YFCC100M [42] (1000 pairs) and indoor SUN3D [47] (539 pairs) datasets. Compared with HPSequences, the two datasets additionally introduce variations such as self-occlusions, and in particular, repetitive or feature-poor patterns in indoor scenes, which is generally considered challenging for sparse matching.

SfM dataset. Following [37], we evaluate on SfM dataset such as well-known *Fountain* and *Herzjesu* [41], or landmark collections [46]. We integrate the proposed framework into SfM pipeline, i.e., COLMAP [36], and use the keypoints provided in [37] to compute the local features.

4.3. Evaluation protocols

Patch level. For HPatches [2], we follow its evaluation protocols and use mean average precision (mAP) for three subtasks, including patch verification, matching, and retrieval.

Image level. For HPSequences, we use *Recall = # Correct Matches / # Correspondences* defined in [14], to quantify the image matching performance, where *# Correct matches* are matches found by nearest neighbor searching and verified by groundtruth geometry, e.g., homography, while *# Correspondences* are matches that should have been identified by the given keypoint locations. Following [14], a match point is determined to be correct if it is within 2.5 pixels from the wrapped keypoint in the reference image. We use a standard SIFT detector to localize the keypoints, of which the number is randomly sampled to 2048. For YFCC100M [42] and SUN3D [47], we follow the same setting in [49] and report the median number of inlier matches after RANSAC for each dataset.

Reconstruction level. For clarity, we report metrics in [37] that quantify the completeness of SfM, including the number of registered images (*# Registered*), sparse points (*# Sparse Points*) and image observations (*# Observations*).

4.4. Ablation study

4.4.1 Design of context encoder

In this section, we evaluate two splits of HPSequences [2]: *illumination (i)* and *viewpoint (v)*, regarding different image transformations. We report *Recall* as defined in Sec. 4.3. If not specified, we use GeoDesc [23] as a baseline model (*baseline (GeoDesc)*) to extract raw local features, whose parameters are *fixed* during the training of augmentation.

Visual context. We compare four designs, including i) *CS (256-d)*: the central-surround (CS) structure [50, 19, 43] as described in Sec. 2, which concatenates local features from different domain sizes. ii) *w/ global feature*: the integration with global features [5], which is originally designed for improving 3D local descriptors. iii) *w/ regional feature*: the proposed integration with interpolated regional features, and its variant iv) *w/ regional feature + CN*: with context normalization to incorporate global visual information.

As shown in Tab. 1 (left columns), the CS structure [50, 19, 43] delivers only marginal improvements despite the doubled dimensionality. Meanwhile, though being effective in 3D descriptor learning, the integration with global features [5] instead harms the performance, which we ascribe to the limited representation ability of a single global feature. Finally, the proposed integration with interpolated regional features shows clear improvements, as it better handles both spatial and visual distinctiveness. Moreover, to strengthen global context awareness, we show that the performance can be further boosted by equipping context normalization when encoding regional features.

Geometric context. We study five options: i) PointNet-like architecture, i.e., segmentation networks in [31] without the final classifier. ii) Pre-activated context normalization (CN) networks in Sec. 3.2 with 2D xy input, and its variants iii) with additional raw local feature input or iv) with matchability. We also compare the use of pre-activation of the residual unit in context normalization networks.

As presented in Tab. 1 (middle columns), though widely used in processing 3D points, PointNet [31] does not perform well in our task, while the similar phenomenon is also observed in [49] when processing 2D correspondences. Besides, it is noticed that input with raw local feature does

Visual context encoder			Geometric context encoder			Comparison with other methods		
Strategy	*Recall i/v*		*Network architecture*	*Recall i/v*		*Method*	*Recall i/v*	
baseline (GeoDesc [23])	59.46	71.24	baseline (GeoDesc [23])	59.46	71.24	SIFT [22]	47.36	53.06
CS (256-d) [50, 19, 43]	59.83	71.27	PointNet [31]	59.61	70.96	L2-Net [43]	47.58	53.96
w/ global feature [5]	59.11	71.02	w/ CN (pre.) + xy	61.67	72.63	HardNet [25]	57.63	63.36
w/ regional feature	63.64	73.37	w/ CN (pre.) + xy + raw local feature	60.91	72.99	GeoDesc [23]	59.46	71.24
w/ regional feature + CN	**63.98**	**73.63**	w/ CN (orig.) + xy + matchability	59.94	71.25	**ContextDesc**	**66.55**	**75.52**
			w/ CN (pre.) + xy + matchability	**62.82**	**73.40**	**ContextDesc+**	**67.14**	**76.42**

Table 1: Comparisons on HPSequences [2] of different designs of visual and geometric context encoder, and the performance of entire augmentation scheme. 'i/v' denotes two evaluations on *illumination* and *viewpoint* sequences, respectively.

not help to boost the performance, which we attribute to the weak relevance between local features as extracted from different orientations and levels of scale space pyramid. Instead, the incorporation with matchability is notably beneficial, as matchability is more comprehensive as a high-level abstraction of local feature. Finally, the pre-activation is clearly a preferable alternative than its original design.

Integration with cross-modality context. Finally, we evaluate the full augmentation with both visual and geometric context (*ContextDesc*). As shown in Tab. 1 (right columns), the simple summation aggregation in Sec. 3.4 effectively takes advantage of both context, delivering remarkable improvements over the state-of the-art.

4.4.2 Efficacy of softmax temperature in N-pair loss

To demonstrate the validity of proposed loss in Sec. 3.5, we train *only* the local base model without any context awareness, and compare different losses including: i) the plain N-pair loss in [43] without scale temperature, and ii) the scale-aware loss in [17] with its original parameters.

	GeoDesc [23]	w/ loss in [43]	w/ loss in [17]	Ours
HPatches, mAP [%]				
Verification	**91.1**	78.3	81.2	90.2
Matching	59.1	23.9	40.5	**59.2**
Retrieval	74.9	46.8	64.0	**76.0**
HPSequences, Recall				
Seq. i	59.5	32.2	50.0	**59.7**
Seq. v	71.2	48.5	64.8	**72.6**

Table 2: Evaluation results on 1) HPatches [2] of three complementary tasks: patch verification, matching and retrieval. 2) HPSequences of two sequence splits.

As shown in Tab. 2, the proposed loss improves the overall performance over the previous best-performing GeoDesc [23] under similar training settings, while GeoDesc requires additional geometric supervision. Besides, the proposed loss clearly shows better convergence compared with losses in [43] and [17]. Although we suspect that the loss in [17] may perform better with careful parameter searching, the proposed loss is advantageous due to its self-adaptivity without the need of complex heuristics or manual tuning.

Moreover, once replace GeoDesc with the above model as a base in the augmentation scheme, the final performance can be further improved by a significant margin, denoted as *ContextDesc+* in Tab. 1 (right columns), which again addresses the advance of improved base model. We will use this model to complete the following experiments.

4.5. Generalization

Wild dataset. The evaluation results on two challenging datasets (*outdoor* YFCC100M [42] and *indoor* SUN3D [47]) are presented in Tab. 3. The proposed cross-modality context augmentation delivers $\sim$35% and $\sim$125% improvements over the previous state of the art, which effectively demonstrates the strong generalization ability of the learned augmented features in practical scenes.

	SIFT [22]	L2-Net [43]	HardNet [25]	GeoDesc [23]	Ours
median number of inlier matches					
indoor	138	153	239	271	**365**
outdoor	168	173	219	214	**482**

Table 3: Evaluation results on wild datasets: *indoor* SUN3D [47] and *outdoor* YFCC100M [42] datasets.

SfM dataset. We further demonstrate the improvement in complex SfM pipeline. As shown in Tab. 4, the integration of augmented feature generalizes well among different scenes even in large-scale SfM tasks, meanwhile consistently boosts the completeness of sparse reconstruction. Some matching results are presented in Fig. 5, and more visualizations can be found in the appendix.

		# Images	# Registered	# Sparse Points	# Observations
Fountain	*SIFT [22]*	11	11	10,004	44K
	GeoDesc [23]		11	16,687	83K
	Ours		11	**16,965**	**84K**
Herzjesu	*SIFT*	8	8	4,916	19K
	GeoDesc		8	8,720	38K
	Ours		8	**9,429**	**40K**
South Building	*SIFT*	128	128	62,780	353K
	GeoDesc		128	170,306	887K
	Ours		128	**174,359**	**893K**
Roman Forum	*SIFT*	2,364	1,407	242,192	1,805K
	GeoDesc		1,566	770,363	5,051K
	Ours		**1,571**	**848,319**	**5,484K**
Alamo	*SIFT*	2,915	743	120,713	1,384K
	GeoDesc		893	353,329	3,159K
	Ours		**921**	**424,348**	**3,488K**

Table 4: Evaluation results on SfM dataset [37].

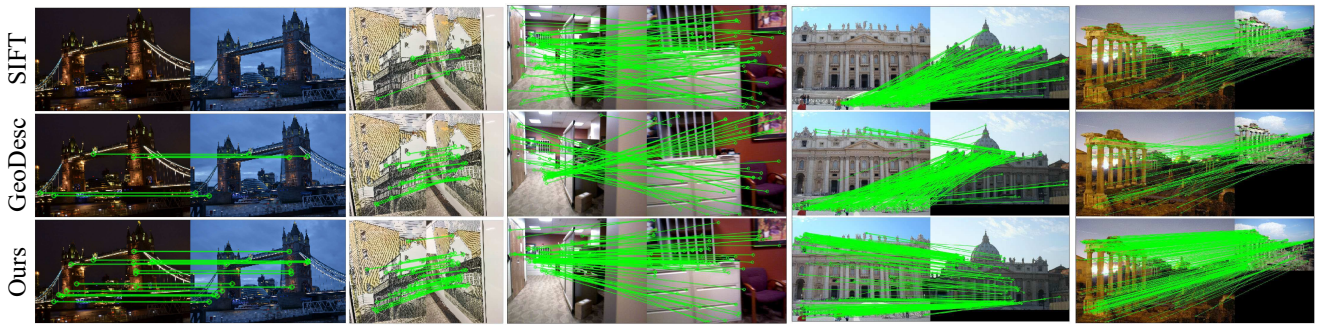


Figure 5: Matching results after RANSAC in different challenging scenarios. From top to bottom: SIFT, GeoDesc and ours. The augmented feature helps to find more inlier matches, and further allows a more accurate recovery of camera geometry.

4.6. Discussions on practicability

Invariance property. We again use *Recall* and evaluate on Heinly benchmark [14] to quantify the invariance property. As shown in Tab 5, the proposed method improves remarkably over the previous best-performing descriptor, except for some minor underperformance regarding *Rotation* change when images are rotated up to 180°, which may be caused by the inability of being fully rotation-invariant especially for the regional feature extractor.

	SIFT [22]	**GeoDesc [23]**	**Ours**
	Recall		
JPEG	60.7	66.1	**78.6**
Blur	41.0	47.7	**57.8**
Exposure	78.2	86.4	**88.2**
Day-Night	29.2	*39.6*	**43.3**
Scale	81.2	85.8	**88.1**
Rotation	82.4	**87.6**	86.3
Scale-Rotation	29.6	33.7	**38.0**
Planar	48.2	59.1	**61.7**

Table 5: Evaluation results regrading different transformations on Heinly benchmark [14].

Computational cost. Towards practicability, we only use shallow MLPs or non-parametric context normalization in the augmentation framework, which thus introduces only insignificant computation overhead. As reported in Tab. 6, suppose that regional features are readily extracted, e.g., from a retrieval model deployed in SfM pipeline for accelerating image matching, the full augmentation then requires only $\sim 5\%$ time cost compared with the raw local feature description. Virtually, the proposed framework allows flexible integration and reuse of other visual components to achieve system-level efficiency, such as saliency or segmentation masks, and thus has large rooms for future improvements.

End-to-end training. For ablation purposes, the parameters of base local and regional models are previously fixed in the training, and we here provide further studies about the efficacy of an end-to-end training scheme.

In the first setting, we freeze only the regional model and train from scratch with Eq. 7 on the augmented feature. As a result, the performance is notably improved from **67.14 to 67.53**, and **76.42 to 77.20** for *i/v* sequences of HPSequences, compared with *ContextDesc+* in Tab. 1.

	Preparation		*Augmentation*		
	local feat.	**regional feat.**	**geo. context**	**vis. context**	**multi-context**
Time (ms)	351	49	5	14	18
FLOPs (B)	802.9	123.4	1.7	13.9	15.7
Params (M)	2.4	24.5	<0.1	3.1	3.2

Table 6: The computational cost of proposed framework, evaluated on 10k keypoints from an 896×896 image. The inference time is estimated on an NVIDIA GTX 1080 GPU.

In the second setting, we further end-to-end train with the regional model, which is additionally optimized by a standard cross-entropy classification loss as in [28] for simplicity (see Appendix A.1 for details). Although several loss balancing strategies have been experimented, we did not observe a consistent improvement for final matching performance, which we ascribe to the substantial challenge posing by multi-task learning. Thus, we currently recommend a separate training for the regional model, and look forward to an improved solution in the future.

5. Conclusion

In contrast to current trends, we have addressed the importance of introducing *context awareness* to augment local feature descriptors. The proposed framework takes keypoint location, raw local and high-level regional feature as input, from which two types of context are encoded: *geometric* and *visual* context, while the training adopts a novel N-pair loss that is self-adaptive and parameter-tuning free. We have conducted extensive evaluations on diversified and large-scale datasets, and demonstrate remarkable improvements over the state of the art, meanwhile showing the strong generalization and practicability in real applications.

Acknowledgment. This work is supported by Hong Kong RGC GRF 16203518, T22-603/15N, ITC PSKL12EG02. We thank the support of Google Cloud Platform.

References

[1] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk. Pn-net: Conjoined triple deep network for learning local image descriptors. In *arXiv*, 2016. 2

[2] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017. 1, 6, 7

[3] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *BMVC*, 2016. 2

[4] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. In *IJCV*, 2007. 6

[5] H. Deng, T. Birdal, and S. Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *CVPR*, 2018. 2, 4, 5, 6, 7

[6] D. DeTone, T. Malisiewicz, and A. Rabinovich. Deep image homography estimation. In *arXiv*, 2016. 6

[7] J. M. Dmytro Mishkin, Filip Radenovic. Repeatability is not enough: learning discriminative affine regions via discriminability. In *ECCV*, 2018. 2, 3, 5

[8] J. Dong and S. Soatto. Domain-size pooling in local descriptors: Dsp-sift. In *CVPR*, 2015. 1

[9] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *CVPR*, 2015. 2

[10] W. Hartmann, M. Havlena, and K. Schindler. Predicting matchability. In *CVPR*, 2014. 4

[11] K. He, Y. Lu, and S. Sclaroff. Local descriptors optimized for average precision. In *CVPR*, 2018. 2, 3

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[13] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 4

[14] J. Heinly, E. Dunn, and J.-M. Frahm. Comparative evaluation of binary features. In *ECCV*. 2012. 6, 8

[15] E. Hoffer, I. Hubara, and D. Soudry. Fix your classifier: the marginal value of training the last weight layer. In *ICLR*, 2018. 5

[16] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 3

[17] M. Keller, Z. Chen, F. Maffra, P. Schmuck, and M. Chli. Learning deep descriptors with scale-aware triplet networks. In *CVPR*, 2018. 3, 5, 7

[18] N. Kobyshev, H. Riemenschneider, and L. Van Gool. Matching features correctly through semantic understanding. In *3DV*, 2014. 2

[19] B. Kumar, G. Carneiro, I. Reid, et al. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *CVPR*, 2016. 1, 2, 3, 6, 7

[20] K. Lenc and A. Vedaldi. Large scale evaluation of local image feature detectors on homography datasets. In *BMVC*, 2018. 6

[21] S. Li, L. Yuan, J. Sun, and L. Quan. Dual-feature warping-based motion model estimation. In *CVPR*, 2015. 1

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. 2004. 7, 8

[23] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *ECCV*, 2018. 1, 3, 5, 6, 7, 8

[24] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. In *Image and vision computing*, 2004. 1

[25] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *NIPS*, 2017. 2, 3, 5, 7

[26] K. Moo Yi, Y. Verdie, P. Fua, and V. Lepetit. Learning to assign orientations to feature points. In *CVPR*, 2016. 2

[27] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006. 1

[28] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017. 3, 8

[29] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[30] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[31] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 2, 4, 6, 7

[32] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017. 2, 4

[33] F. Radenović, G. Tolias, and O. Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016. 2, 3, 5, 6

[34] I. Rocco, R. Arandjelovic, and J. Sivic. End-to-end weakly-supervised semantic alignment. In *CVPR*, 2018. 2

[35] N. Savinov, A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys. Quad-networks: unsupervised learning to rank for interest point detection. In *CVPR*, 2017. 4

[36] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 6

[37] J. L. Schönberger, H. Hardmeier, T. Sattler, and M. Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *CVPR*, 2017. 6, 7

[38] T. Shen, Z. Luo, L. Zhou, R. Zhang, S. Zhu, T. Fang, and L. Quan. Matchable image retrieval by learning from surface reconstruction. In *ACCV*, 2018. 2, 6

[39] T. Shen, S. Zhu, T. Fang, R. Zhang, and L. Quan. Graph-based consistent matching for structure-from-motion. In *ECCV*, 2016. 1

[40] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *CVPR*, 2015. 2

[41] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008. 6

[42] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. In *CACM*, 2016. 6, 7

[43] Y. Tian, B. Fan, F. Wu, et al. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *CVPR*, 2017. 1, 2, 3, 5, 6, 7

[44] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *ICLR*, 2016. 3

[45] N. Ufer and B. Ommer. Deep semantic feature matching. In *CVPR*, 2017. 2

[46] K. Wilson and N. Snavely. Robust global translations with 1dsfm. In *ECCV*, 2014. 6

[47] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013. 6, 7

[48] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016. 2, 3, 6

[49] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua. Learning to find good correspondences. In *CVPR*, 2018. 2, 4, 6

[50] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015. 1, 2, 3, 6, 7

[51] L. Zhang and S. Rusinkiewicz. Learning to detect features in texture images. In *CVPR*, 2018. 4

[52] R. Zhang, S. Zhu, T. Fang, and L. Quan. Distributed very large scale bundle adjustment by global camera consensus. In *ICCV*, 2017. 1

[53] X. Zhang, X. Y. Felix, S. Kumar, and S.-F. Chang. Learning spread-out local feature descriptors. In *ICCV*, 2017. 2

[54] L. Zhou, S. Zhu, Z. Luo, T. Shen, R. Zhang, M. Zhen, T. Fang, and L. Quan. Learning and matching multi-view descriptors for registration of point clouds. In *ECCV*, 2018. 1

[55] L. Zhou, S. Zhu, T. Shen, J. Wang, T. Fang, and L. Quan. Progressive large scale-invariant image matching in scale space. In *ICCV*, 2017. 1

[56] S. Zhu, T. Fang, J. Xiao, and L. Quan. Local readjustment for high-resolution 3d reconstruction. In *CVPR*, 2014. 1

[57] S. Zhu, R. Zhang, L. Zhou, T. Shen, T. Fang, P. Tan, and L. Quan. Very large-scale global sfm by distributed motion averaging. In *CVPR*, 2018. 1