Learning Task-Aligned Local Features for Visual Localization

Chuanjin Liu, Hongmin Liu, Lixin Zhang, Hui Zeng, Lufeng Luo, and Bin Fan

***Abstract*—Visual localization plays a key role in various robot perception systems. Robust visual localization relies on reliable and repeatable local features to establish high quality point correspondences among images. This letter focuses on addressing two limitations of joint learning detector and descriptor. First, existing methods use independent structures and loss functions for keypoint detection and description separately, which poses difficulty in detecting keypoints corresponding to discriminative descriptors. Second, triplet samples are treated equally in most existing approaches, which limits the learning algorithm to obtain highly discriminative descriptors. In this letter, we propose Task-aligned SuperPoint (TaSP) to mitigate the above problems. First, we explicitly align descriptor and detector learning to improve the probability of being detected for those distinctive points. Second, we introduce a dynamic importance weighting module that calculates the weight of each triplet sample based on intrinsic and empirical importance, so as to make the network focus on the most informative triplets during the whole training process. In addition, we resort to 3D space to seek negative samples when forming triplets, which avoids the risk of selecting negatives from repetitive structures. State-of-the-art results on a variety of visual localization benchmarks demonstrate the superiority of our method.**

***Index Terms*—Localization, deep learning for visual perception, computer vision for automation.**

I. Introduction

VISUAL localization has attracted increasing attention in robot localization and navigation [1], [2], [3], [4], [5], [6] during the past few years. To estimate the 6DoF (degrees of freedom) camera pose of input image with respect to a reference 3D scene, visual localization requires establishing keypoint correspondences between a query image and database images, which is based on extracting local features that can be robustly matched across a series of changing scenes. In particular, illumination, season, and viewpoint changes as well as texture-less and repetitive structures [1], [7], [8] are widely presented in visual localization applications, which poses a great challenge for keypoint matching.

Motivated by the success of deep learning, joint learning of local feature detector and descriptor [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] has gained increasing popularity and achieved promising results over hand-crafted methods [20], [21] in visual localization. However, there are still some limitations that hinder further performance improvement from both detection and description perspective: 1) detector learning lacks the utilization of descriptor information, resulting in the mismatches of high detection scores and discriminative descriptors, and 2) descriptor learning lacks the exploration of sample importance, which limits the accuracy of matching.

To alleviate the above limitations, we propose Task-aligned SuperPoint (TaSP), which aligns keypoint detection and description, so that the distinguishness of descriptors can affect the learning of keypoints. In this way, it could significantly increase the probability that keypoints corresponding to discriminative descriptors are detected, so that more matches can be obtained to estimate pose. Meanwhile, we believe that the importance of a triplet in learning is related to both its intrinsic difficulty and loss response, so we introduce a Dynamic Importance Weighting (DIW) module to encourage the network to focus more on triplets that are simple in nature but still not sufficiently learned.

Essentially, learning with triplet loss aims to increase the relative distance between positive and negative pairs. For a pair of images with overlapping view, positive pairs can be obtained by pixel-level ground truth correspondences computed from the camera parameters and depth. To avoid the selection of too hard triples which make learning difficult to converge, existing methods tend to set a safe window centered on the correspondences, and take points within the image but outside the window as negative samples. However, due to image scale variation and complexity of scenes, it is difficult to set a uniform window size for all image pairs. There is a relatively large 3D distance between an object and its background, hence a suitable 3D distance could divide the object's outline (as shown in Fig. 4(b)). We set a 3D safe window to obtain negatives that are intrinsically different from positives.

Our contributions are summarized as follows:

- To link detector and descriptor learning, we introduce a task alignment factor to integrate the distinguishness of

Manuscript received 21 December 2022; accepted 5 April 2023. Date of publication 17 April 2023; date of current version 25 April 2023. This letter was recommended for publication by Associate Editor N. Radwan and Editor S. Behnke upon evaluation of the reviewers' comments. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1313002, in part by the National Natural Science Foundation of China under Grants 62222302, U22B2055, U2013202, and 62273034, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020B1515120050, in part by the Fundamental Research Funds for the Central Universities under Grant FRF-TP-22-003C1, and in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2021-007. *(Corresponding author: Bin Fan.)*

Chuanjin Liu, Hongmin Liu, Lixin Zhang, Hui Zeng, and Bin Fan are with the School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China, and also with the Key Laboratory of Intelligent Bionic Unmanned Systems, Ministry of Education, Beijing 100083, China (e-mail: 18751855987@163.com; hmliu_82@163.com; lxzhang@ustb.edu.cn; hzeng@ustb.edu.cn; bin.fan@ieee.org).

Lufeng Luo is with the School of Mechatronic Engineering and Automation, Foshan University, Foshan 528000, China (e-mail: luolufeng@fosu.edu.cn).

Digital Object Identifier 10.1109/LRA.2023.3268015

descriptor into detector learning, so as to extract keypoints that are easy to distinguish.
- By introducing a dynamic importance weighting module, which considers both of the intrinsic importance and the empirical importance of a triplet to calculate its weight during the learning procedure, our method can encourage the network learning to pay more attention to the samples that are intrinsically simple but are quite important.
- A triplet generation strategy based on 3D safe window is used to avoid selecting negatives that have similar structure to the positives. This window is scale-invariant and has an awareness of object outlines, so informative triplets can be obtained in learning. State-of-the-art localization performance is obtained on both outdoor and indoor localization benchmarks demonstrating the superiority of our method.

II. Related Work

A. Visual Localization

Motivated by the development of deep learning, some works directly regress the camera pose of input image by training an end-to-end neural network. PoseNet [22] is the first method to regress the 6DoF camera pose by CNN. However, it relies on accurate SfM annotation and the output pose is not accurate enough compared to other geometric methods [10], [11], [23], [24]. DSC-PoseNet [25] designs a self-supervised framework using only 2D bounding box annotations so that it no longer relies on 3D pose labels, bringing a new idea for end-to-end pose estimation.

In contrast to regression-based approaches, most visual localization [2] methods use 2D-3D matches between the query image and a 3D reference model to recover the absolute pose by Perspective-n-Point (PnP) [23]. The most classical approach to establish such matches is to perform local feature extraction followed by mutual nearest neighbor matching of descriptors. However, visual localization often encounters various complex scenes, which poses difficulties for feature extraction and description.

B. Local Features

The purpose of local feature extraction is to detect repeatable and reliable keypoints in an image, and characterize them by vector representation. local features determine the performance of downstream tasks such as SLAM [4], [5], [6], SfM [26], [27], and visual localization [1], [2]. In recent years, data-driven approaches exhibit strong robustness in challenging situations, and many deep learning-based detectors, descriptors, and joint learning methods have been proposed that are highly competitive compared to hand-crafted features [20], [21].

Detectors: These methods estimate a score map indicating keypoint probability. Recently, Key.Net [28] uses hand-crafted and learned features to estimate a score map and to use soft-argmax to extract keypoints. D2D [29] is a method to detect keypoints based on an off-the-shelf descriptor network, which extracts local features that can be reliably matched based on the information entropy and neighborhood saliency of dense descriptors. This inspires us to consider the distinguishness of descriptors in detection.

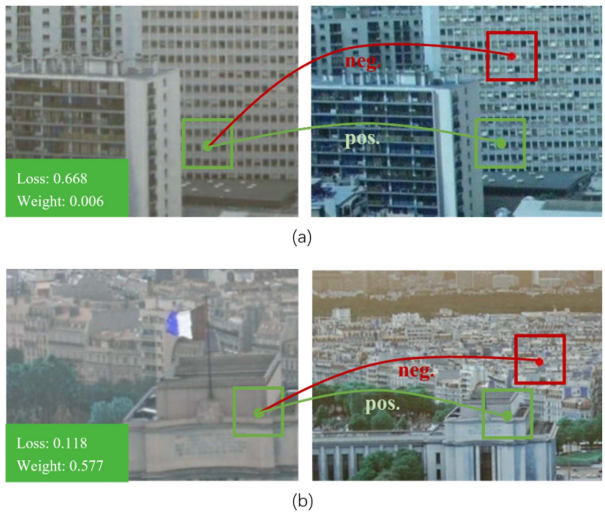


Fig. 1. The widely used triplet loss in descriptor learning will misguide the optimization of network when encountering repetitive structure as shown in (a), for which our method successfully decreases its weight to 0.006 by considering both intrinsic properties and triplet loss responses. In addition, as in (b), our method can pick out the triplet that is still helpful for network training and hence increase its weight in the training loss although its triplet loss is actually quite small according to existing works. More details can be found in Section III.

Descriptors: The robustness of deep learning-based descriptors compared to hand-crafted descriptors [20], [21] is impressive. TFeat [30] is an early approach to introduce triplet loss [31] for descriptor learning and applies a hard negative mining method called anchor swapping. HardNet [32] achieve advanced performance by finding the hardest triplet samples. CAPS [33] uses camera poses as weakly supervision and introduces an epipolar loss for descriptor learning. DSM [34] uses a cumulative distribution function (CDF) to measure the hardness of triplets and to assign larger weights to harder samples. Motivated by the good performance of hard negative mining, we further explore he combined effect of sample intrinsic importance and empirical importance.

Joint Learning of Detector and Descriptor: In the past few years, joint learning approaches have received more attention. SuperPoint [10] uses a self-supervised model to learn detector and descriptor on pre-labeled keypoints. D2-Net [11] models the keypoint score based on spatial properties of feature maps and implicitly adds it to the learning of descriptors. Based on this idea, ASLFeat [14] uses multi-level detection and deformable convolution [35] to obtain high-precision keypoints. DISK [15] uses reinforcement learning of policy gradients to learn a U-Net network for simultaneous keypoint detection and description. ALIKE [36] designs a differentiable keypoint detection by using softargmax operation on local fraction patches, and adopts neural reprojection error (NRE) [37] loss to train the entire descriptor map. PoSFeat [38] decouples the two parts and uses the Line-to-Window scheme to obtain negative samples, and then learns keypoints on the trained descriptor network. The success of PoSFeat inspires us to find a better strategy to obtain negative samples.

Recent methods usually obtain keypoints by detecting on feature maps of the neural network. SuperPoint [10] detects keypoint locations by carting a keypoint as one of 64 classes in

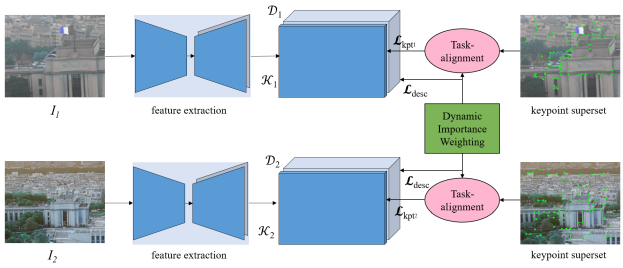


Fig. 2. The pipeline of our approach. A pre-labeled keypoint superset is required. Training aims to simultaneously minimize detector loss $\mathcal{L}_{kpt}$ and descriptor loss $\mathcal{L}_{desc}$.

a 8×8 grid. However, it does not take into account whether the keypoints can be reliably matched. D2-Net and ASLFeat learn keypoints implicitly through modeling the spatial properties of feature maps, which only rely on the relative size of points' neighbor and channel features, and so do not guarantee the distinguishness of corresponding descriptors.

Regarding the descriptor learning, contrastive loss [39] or triplet loss [31] are generally used to separate matching and non-matching keypoints by a margin and beyond. However, they don't concern about the importance of each sample, which limits further improvement of descriptor distinguishness. Dynamic Soft Margin (DSM) [34] applies a hard negative mining approach to calculate weights by triplet loss response, which allows the network learning to focus more on difficult samples. The aforementioned hard samples generally refer to those with large loss. However, a hard sample is not necessarily important. As Fig. 1(a) illustrates, local patches from texture-less or repetitive structures are inherently difficult to distinguish. They should not be excessively concerned as they are indistinguishable. In this view, simply focusing on samples with large loss responses limits further improvement in DSM performance.

III. Method

In this section, we introduce our proposed TaSP in detail. For an input image $I \in \mathbb{R}^{H \times W \times 3}$, we use EfficientNet-B0 [40] as the backbone. After the backbone network, we append a detector head to estimate a heatmap $\mathcal{K} \in \mathbb{R}^{H \times W}$, and a descriptor head to obtain a descriptor map $\mathcal{D} \in \mathbb{R}^{H_c \times W_c \times dim}$, where $H_c = H/4$ and $W_c = W/4$. The learning of our network is based on image pairs with ground truth correspondences, while using the homographic adaptation strategy in SuperPoint [10] to label pseudo-ground truth keypoints. For end-to-end training of the network, lower case use the *Dynamic Importance Weighting* module to assign weights to each training triplet. Based on this, the weighted margin-based hard triplet loss $\mathcal{L}_{desc}$ is used in training of the descriptor head. Meanwhile, the detector is learned from the distinguishness of the descriptors through the proposed *Task-aligned keypoint loss* $\mathcal{L}_{kpt}$. To sum up, our approach jointly learns the detector and descriptor, with the overall loss:

$$\mathcal{L} = \mathcal{L}_{desc} + \lambda \cdot \mathcal{L}_{kpt}, \tag{1}$$

where λ is the weight for balancing the detector and descriptor learning. The pipeline of our approach is shown in Fig. 2.

A. Dynamic Importance Weighting Module

The goal of descriptor learning is to separate same and different keypoints in an image pair, and it is usually trained using margin-based hard triplet loss [11], [32], [34]. We first briefly introduce this loss, then analyze some existing extensions, and finally elaborate our extended version for jointly considering empirical and intrinsic importance.

Given a pair of images (I_1, I_2), for a point p_1 on I_1, its corresponding point p_2 on I_2 can be obtained via warping I_1 to I_2 according to camera parameters and depths. Their corresponding descriptors are denoted as D_1, D_2 respectively, and we define N_1, N_2 as the set of points on I_1, I_2 that locate outside the safe window centered on p_1 and p_2. It is usually desired to minimize the distance d_{pos} between D_1 and D_2, which can be computed as:

$$d_{pos} = \|D_1 - D_2\|_2 , \tag{2}$$

and simultaneously maximize the negative distance d_{neg}, which is:

$$d_{neg} = \min \left(\|D_1 - D_{n_2}\|_2 , \|D_2 - D_{n_1}\|_2 \right), \tag{3}$$

where n_2 represents the hardest negative sample point on another image, calculated as:

$$n_2 = \arg\min_{n \in N_2} \|D_1 - D_n\|_2 ,$$

$$n_1 = \arg\min_{n \in N_1} \|D_2 - D_n\|_2 ,$$

The margin-based hard triplet loss can be defined as:

$$\mathcal{M} = \max \left(0, m + d_{pos} - d_{neg}\right). \tag{4}$$

With this loss, the difference between positive and negative samples is gradually increased until negative distance is larger than positive distance by the predefined margin *m*, in which case the positive is considered to be sufficiently distinguishable.

Motivated by the fact that harder samples are more important in descriptor learning, margin-baesd triplet loss selects the hardest negative samples. This strategy has been used by many methods [11], [32] and has yielded promising results. Based on a similar idea, assigning larger weights to harder samples [34] has also shown good performance. However, existing methods do not take into account the inherent distinguishness of samples, so we propose the Dynamic Importance Weighting (DIW) module to further explore the role of hard negative mining in descriptor learning.

In general, existing methods [32], [34] determine the importance of samples based on their empirical difficulty (i.e., loss response), which is not equal to their actual importance in descriptor learning. For example, repetitive structures like checkerboards are inherently indistinguishable (as illustrated in Fig. 1(a)), which leads to high empirical difficulties. Excessive attention on these regions is detrimental to the learning of descriptors. We believe that an important triplet should be relatively simple but still have positive contribution to the network training, and it is apparent that the empirical difficulty can only reflect the contribution for network optimization but is not able to measure the intrinsic difficulty. To distinguish whether a triplet is actually simple, we introduce intrinsic importance.

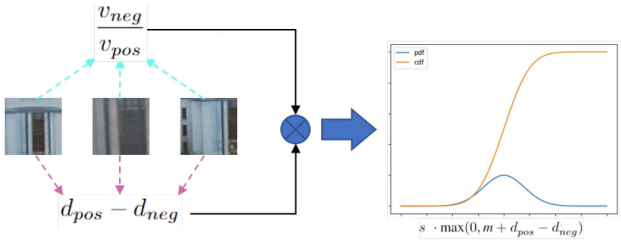


Fig. 3. Our dynamic importance weighting strategy. We count the product of the intrinsic importance and empirical importance of each triplet, then build a moving PDF histogram and integrate it into a CDF. The triplet weights (i.e. importance) can be found on the corresponding values of the CDF.

Intrinsic Importance: As discussed at the beginning of this section, for an anchor point p_1, a corresponding point p_2 and a hardest negative sample point p_{n_2} on another image can be obtained. Motivated by that humans look at regions around sample points when estimating their similarities, our method extracts $N \times N$ patches centered on sample points and uses the PHA (Perceptual Hash Algorithm) to obtain the similarity between patches for its simplicity. We generate the image hash codes by using DCT (Discrete Cosine Transform) based pHash [41]. After this we use Hamming distance to measure the difference between the hash codes. The process is shown as:

$$v_{pos} = Hamming\left(pHash\left(\mathcal{P}_1\right),\ pHash\left(\mathcal{P}_2\right)\right), \tag{5}$$

$$v_{neg} = Hamming\left(pHash\left(\mathcal{P}_1\right),\ pHash\left(\mathcal{P}_{n_2}\right)\right), \tag{6}$$

where $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_{n_2}$ represent patches centered on p_1, p_2, p_{n_2} respectively. A smaller v means the two patches are more similar and vice versa. Then the intrinsic importance score of a triplet can be defined as:

$$s = \frac{v_{neg}}{v_{pos}}. \tag{7}$$

Since a simple triplet is considered to have less difference in positive samples than in negative samples (i.e. $\mathcal{P}_1$ is more similar to $\mathcal{P}_2$ than $\mathcal{P}_{n_2}$), its intrinsic importance score s would be high. Conversely, when $\mathcal{P}_1$ is more similar to $\mathcal{P}_{n_2}$, the score s is low. Therefore, the intrinsic importance is inversely proportional to the difficulty of distinguishing two corresponding patches from other patches.

Empirical Importance: Given that the good performance of DSM [34], which measures the effect of different triplet in descriptor learning, we use a similar cumulative distribution function (CDF) to measure the empirical importance of triplets. We count the $d_{pos} - d_{neg}$ of each mini-batch into a histogram to obtain the discretized probability distribution function (PDF). We maintain it as an exponentially decaying moving histogram (with a weight of 0.1 on each new batch) to obtain a temporally stable PDF. The relative difficulty weight $k\%$ of the triplet over all samples can be obtained by integrating the PDF into the CDF, as shown in the Fig. 3. If the triplet corresponds to a CDF of 1.0 indicates that it is the hardest in recent batches, while the easiest triplet corresponds to a CDF ≈ 0. More generally, a CDF of $k\%$ for a triplet implies that it is empirically harder than $k\%$ of triplets within recent batches.

As explained before, high empirical difficulty of a triplet does not mean that it is important for descriptor learning. In this letter, we go a step further by taking into account the joint effect of intrinsic and empirical importance. We believe that if the relatively simple triplet according to intrinsic importance still has a large response, it is more important and should be focused on in the following training process. Meanwhile, triplets with large response but high intrinsic difficulty should be ignored due to their indistinguishable nature. Based on this idea, we propose the weighted triplet loss as follows:

$$\mathcal{L}_{desc} = \frac{1}{N}\sum_i w_i \cdot\ s_i \cdot \max\left(0, m + d^i_{pos} - d^i_{neg}\right), \tag{8}$$

$$w_i = \mathrm{CDF}\left(s_i \cdot \max\left(0, m + d^i_{pos} - d^i_{neg}\right)\right). \tag{9}$$

This loss dynamically assigns weights by jointly considering the intrinsic and empirical importance of triplets, so that the network can automatically focus on easy but not yet sufficiently learned triplets and ignore extremely difficult ones. When $d_{pos} - d_{neg} < -m$, the weight is set to 0, in which case the triplet is considered to be sufficiently easy for the network to distinguish regardless of its intrinsic importance.

B. Task-Aligned Learning of Detector

Since detection and description are implemented through two separate head branches, existing methods [10], [13], [16] suffer from misalignment between the two tasks, which means that some points corresponding to discriminative descriptors may not be detected. Inspired by TOOD [42], which explicitly aligns classification and localization in object detection to make good classification results and accurate detection boxes correspond to each other, we design a task alignment factor based on the margin-based hard triplet loss. This encourages discriminative descriptors to connect with higher detection scores.

Task-aligned Keypoint Loss: In the case of detector learning, in order to explicitly increase detection scores for discriminative descriptors, and decrease detection scores for unreliable descriptors, it is reasonable to desire the detector to be aware of the distinguishness of descriptors. Considering that empirical importance (i.e. Eq. (4)) can reflect the distinguishness of descriptors, for purpose of relating this to detector learning, a task alignment factor for each triplet can be written as:

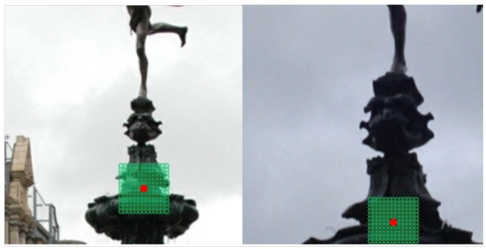
$$\delta = \exp\left(t \cdot (m - \mathcal{M})\right), \tag{10}$$

where t controls the sharpness of alignment intensity.

For each location (i, j) on feature map $\mathcal{X}$ obtained by the detector, we replace binary label of positive anchor with the task alignment factor δ, so that the adapted cross-entropy loss can be written as:

$$\mathcal{L}_{kpt} = -\sum_{(i,j)\in pos} \delta_{ij} \cdot \log\left(x_{ij}\right) - \sum_{(i,j)\in neg} \log\left(1 - x_{ij}\right). \tag{11}$$

When $\mathcal{M}$ is small (i.e. d_{pos} is much smaller than d_{neg}), the triplet can be well discriminated, and the alignment factor δ will be high. The above loss encourages the detector learning to pay more attention to the points whose descriptors are discriminative, which allows more reliable matches to be obtained for visual localization.



(a) Square Safe Window

(b) 3D Safe Window

Fig. 4. Comparison of the classic square safe window and the 3D safe window. The red dots are two corresponding points, and the green areas are the locations within their window respectively. When 3D safe window is used, the areas covered by the window are corresponded, and are adaptive to scale changes.

C. 3D Safe Window

In order to find an informative set of negative samples, D2-Net [11] and ASLFeat [14] define a safe radius K. The set of points on image I_2 whose distances to p_2 is greater than K, written as:

$$N_2 = \{p \mid p \in I_2, \|p - p_2\|_\infty > K\},$$

is considered as the negative set N_2 of p_1, and N_1 can be obtained in a similar way, as illustrated in Fig. 4(a). However, for objects of different scales and shapes, using just a simple window is sub-optimal. Since repetitive structures widely exist in the object itself, therefore, we choose to use the 3D distance to obtain negative set. The 3D coordinates of points can be obtained via ground truth depth and camera parameters, so the negative set $\hat{N}_2$ is calculated as follows:

$$\hat{N}_2 = \left\{p \mid p \in I_2, \|h(p) - h(p_2)\|_2 > \hat{K}\right\},$$

where $h(\cdot)$ represents the 3D coordinates of points. The difference between 3D safe window to square safe window is illustrated in Fig. 4. As can be seen, when 3D safe window is used, the areas covered by the window are corresponded, and are adaptive to scale changes. Since objects usually have large distances between them and their backgrounds, an appropriate 3D distance could avoid selecting negative points from the object itself so as to reduce the influence of repetitive structures.

D. Network Architecture

The lightweight network EfficientNet-B0 architecture [40], pretrained on ImageNet [43] and truncated after the $block\_13$, is used to initialize our feature extraction network. To maintain image resolution, stride sizes of the strided convolution layer of $stage\_4$, $stage\_5$, $stage\_7$ are modefied to 1. We branch out a strided convolution layer as the detector head before the last layer of the backbone. It consists of a convolution layer with 65 channels, which represents 64 positions on original resolution and an extra dustbin representing no keypoint in these positions. At the same time, descriptors are obtained by interpolating and normalizing the corresponding features of keypoints.

IV. Experiments

A. Training

The **MegaDepth** dataset [44] is used to train our model. It includes tourist photos of famous sites and 3D points reconstructed by COLMAP [26], [45]. It also estimates dense depth as well as camera parameters for each image, which allows us to establish dense correspondences between images. We adopt the image pairs generated in D2-Net [11] to train our model. There are 309 k image pairs for training from 118 scenes and 18 k pairs for validation from remaining 78 scenes.

Before training, we label pseudo-ground truth keypoints in one of the image pairs by homographic adaption strategy [10]. Meanwhile, we find that it is insufficient to use only labeled points for descriptors learning. To add more triplets, we sample grid points at 4×4 intervals on image I_1 randomly (no more than 256) and construct the triplets as in Section III.A.

Our approach finetunes all layers for 10 epochs using Adam [46] with an initial learning rate of $1e^{-3}$, and the learning rate linearly decaying to 0. For each scene, at most 100 pairs of images are taken to compensate the scene imbalance presented in the dataset. In our experiments, λ in (1) is set to $1e^{-3}$, while the temperature t in (10) for task alignment is set to 0.5. We spend 10 hours training on a single NVIDIA RTX 2080Ti GPU with a batch size of 4 and 256×256 images.

B. Inference

In our experiments, the detector outputs a tensor of 65 channels, with the first 64 dimensions corresponding to local, non-overlapping 8×8 grid regions of pixels and the last dimension representing the extra no keypoint dustbin. After a channel-wise softmax, the dustbin dimension is removed and the first 64 dimensions are reshaped into a heatmap of the same size as the original image. We detect on the heatmap with a threshold, while the Non-Maximum Suppression (NMS) with a fixed size of 3 is used to remove close keypoints. Descriptors are obtained by bilinear interpolation and L2-normalization at keypoint locations.

C. Evaluation Datasets

We evaluate our method on a wide range of benchmarks involving matching of images with illumination, viewpoint changes, texture-less areas, and repetitive structures.

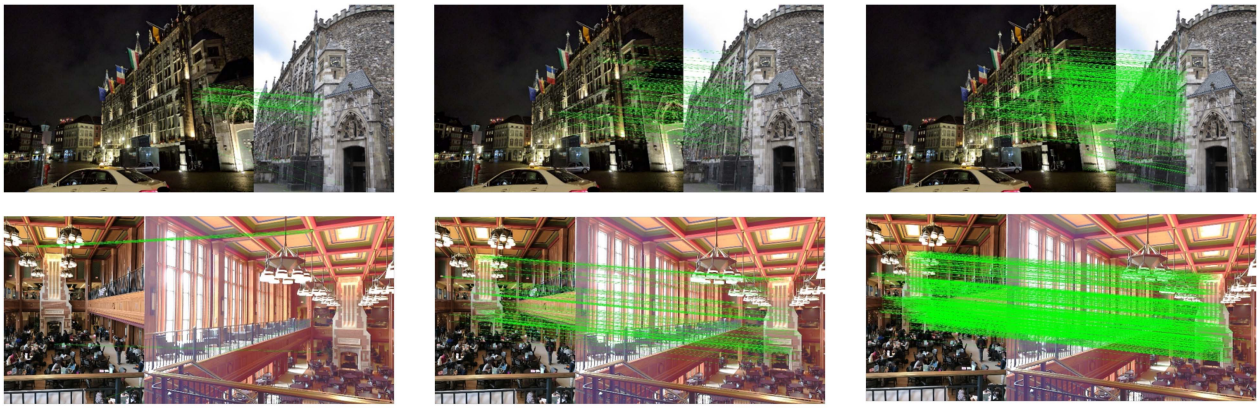
Long-Term Visual Localization: Visual localization rely on robust image matching results. The Aachen Day-Night v1.1 dataset [1] concerning outdoor scenes and the InLoc [47] dataset concerning indoor scenes are used in our experiments.

For Aachen Day-Night v1.1, we use the code and evaluation protocol from [1], and report the average results over three runs. For InLoc, we use the open-sourced hierarchical localization pipeline HLoc [24]. First, for the night query images of Aachen Day-Night v1.1, we use the officially provided matching image

TABLE I
RESULTS OF EXISTING METHODS AND OUR MODEL ARE SHOWN ON TWO DATASETS, AACHEN DAY-NIGHT V1.1 [2] AND INLOC [47]

	Aachen Day-Night v1.1						InLoc					
	day			night			duc1			duc2		
Method	0.25 m 2°	0.5 m 5°	5.0 m 10°	0.25 m 2°	0.5 m 5°	5.0 m 10°	0.25 m 10°	0.5 m 10°	5.0 m 10°	0.25 m 10°	0.5 m 10°	5.0 m 10°
SuperPoint [10]	85.2	92.8	96.8	68.0	85.1	95.4	39.9	55.6	67.2	37.4	57.3	70.2
D2-Net [11]	83.1	91.4	96.7	68.8	85.4	96.4	41.9	59.6	69.7	39.7	58.8	65.6
R2D2 [12]	84.8	90.9	94.9	64.1	77.7	87.7	36.9	53.0	65.7	34.4	52.7	59.5
ASLFeat [14]	86.8	94.2	97.5	70.4	85.2	96.3	37.4	54.5	64.1	38.2	58.0	62.6
ASLFeat-GAN [19]	88.2	94.5	97.8	71.7	85.9	96.9	-	-	-	-	-	-
TaSP (ours)	87.9	94.3	97.8	74.3	86.9	97.9	40.9	62.6	72.7	44.3	59.5	68.7

Red and blue represent the optimal and suboptimal methods.



(a) SuperPoint (b) ASLFeat-GAN (c) TaSP (ours)

Fig. 5. Examples of local feature matching after RANSAC from Aachen Day-Night v1.1 [2] and InLoc [47] benchmark. It can be seen that TaSP produces more correct matches. And TaSP clearly outperforms ASLFeat-GAN [19] and SuperPoint [10] for the indoor localization task. In addition, SuperPoint obtains incorrect matches due to the complexity of the indoor scenes. This is understandable because there is no such data in the SuperPoint training.

list, and for the day query images in Aachen and all query images in InLoc, we use NetVLAD [48] to find top 20 candidate images by matching the query with the database images. After that, we match local features between query image and the candidate images to compute the 6DoF camera pose. For comparison, we report the results of the state-of-the-art local feature extraction methods [10], [11], [12], [14], [19] from the benchmark website (www.visuallocalization.net/benchmark).

HPatches: The HPatches dataset [49] is a standard image matching benchmark. It contains 116 sets of sequences with ground truth homographies, each sequence contains 6 images, and the task requires matching the first one with the other five. We choose 52 sequences containing only illumination changes and 56 sequences containing only viewpoint changes to evaluate the performance of our method.

ETH: We evaluate our method on a popular downstream task of local features, i.e., 3D reconstruction, on the ETH local feature benchmark [50]. We choose the three medium-scale datasets in this benchmark to have comparison with reported results. All image pairs are exhaustively matched with both ratio test at 0.9 and mutual check for outlier rejection. Four metrics are used for comparison, including the number of registered images (#Reg. Images), the number of sparse points (#Sparse Points), the track length, and the reprojection error (Reproj. Err.).

V. RESULTS

A. Comparison to the State of the Art

Table I shows the results on Aachen Day-Night v1.1 [2] and InLoc [47]. The percentage of query images that are successfully localized within three error thresholds are reported respectively. Our method outperforms other methods in almost all cases, especially for indoor localization. This is due to the stronger robustness of our descriptors in difficult scenarios, which allows a large quantity of accurate matches to participate in visual localization. It is argued in ASLFeat-GAN [19] that for visual localization, a high number of correct matches helps to recover the camera pose. In other words, a high recall performance may correspond to better localization results. This is also demonstrated by our experiments. As shown in Fig. 5, our method obtains a large number of correct matches in both challenging day-night changing scenes and large viewpoint changing indoor scenes.

Fig. 6 shows the Recall of baselines and TaSP on HPacthes. TaSP outperforms existing methods in the case of illumination changing, and is second only to ASLFeat-GAN [19] in the case of viewpoint changing. But our localization performance is still better than ASLFeat-GAN. We speculate that this is because the keypoints of ASLFeat-GAN are not uniformly distributed and the total number is relatively small, making it struggle to obtain

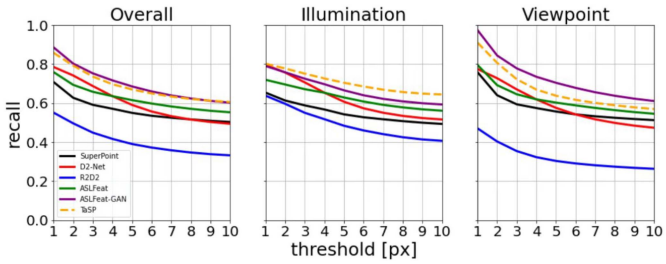


Fig. 6. Comparison of all baselines and TaSP on HPatches, recall curves at thresholds from 1 to 10 are reported.

TABLE II
EVALUATION RESULTS ON ETH BENCHMARK [50] FOR 3D RECONSTRUCTION

Datasets	Methods	#Reg. Images	#Sparse Points	Track Length	Reproj. Err.(px)
Madrid Metropolis (1344 imgs)	SuperPoint [10]	438	29k	9.03	1.02
	D2-Net [11]	495	144k	6.39	1.35
	ASLFeat [14]	649	129k	9.56	0.95
	PoSFeat [38]	419	72k	9.18	0.86
	TaSP (ours)	715	148k	9.49	1.21
Gendarmen-markt (1463 imgs)	SuperPoint [10]	967	93k	7.22	1.03
	D2-Net [11]	965	310k	5.55	1.28
	ASLFeat [14]	1061	320k	8.98	1.05
	PoSFeat [38]	956	240k	8.40	0.92
	TaSP (ours)	1082	360k	9.74	1.23
Tower of London (1576 imgs)	SuperPoint [10]	681	52k	8.67	0.96
	D2-Net [11]	708	287k	5.20	1.34
	ASLFeat [14]	846	252k	13.16	0.95
	PoSFeat [38]	778	262k	11.64	0.90
	TaSP (ours)	944	301k	11.22	1.27

enough matches after RANSAC (as Fig. 5 shows), which is also the reason why ASLFeat [14] is less effective on InLoc despite its higher precision in localization keypoints.

Table II shows the 3D reconstruction results of TaSP and other compared local features. Our method produces more number of registered images and sparse points than other methods, which demonstrates better reconstruction results obtained by our method. This is consistent with the result of TaSP that has more matches.

B. Ablation Studies

We conduct ablation studies on long-term visual localization benchmarks. To investigate the effectiveness of our method, we add Task alignment (TA), Dynamic Importance Weighting (DIW) and 3D Safe Window (3D SW) gradually to a base model (i.e., using EfficientNet-B0 as backbone, learning with the triplet hard loss combined with cross-entropy loss, and the square safe window is used). These results are presented in Table III.

The Task Alignment factor encourages the detector to pay more attention to triplets with lower descriptor loss responses. The results of using task alignment in Table III are better than those without it. Fig. 7 visually verifies that the models trained with task alignment have more keypoints. It is obvious that the extra keypoints, marked in red, are located in discriminative image regions.

The Dynamic Importance Weighting module enables descriptor learning to focus on triplets that still have positive contribution for descriptor learning, making our descriptors more discriminative. Table III shows that without DIW, TaSP degrades severely.

TABLE III
ABLATION STUDIES ON INLOC

TA	DIW	3D SW	DUC1	DUC2
			38.9 / 60.6 / 72.7	39.7 / 55.7 / 64.9
✓			39.9 / 60.6 / 72.7	41.2 / 55.7 / 67.9
	✓		38.9 / 59.1 / 72.7	42.0 / 60.3 / 68.7
		✓	37.4 / 56.6 / 70.2	41.2 / 57.3 / 64.9
✓	✓		41.4 / 59.1 / 72.7	43.5 / 59.5 / 65.6
✓		✓	40.4 / 58.1 / 69.2	41.2 / 55.7 / 65.6
	✓	✓	42.4 / 60.6 / 74.7	42.0 / 58.8 / 66.4
✓	✓	✓	40.9 / 62.6 / 72.7	44.3 / 59.5 / 68.7

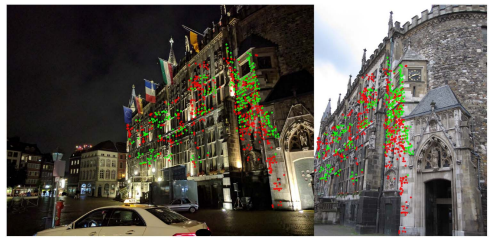


Fig. 7. A visual comparison of the matching keypoints after RANSAC for a pair of images, the network using task alignment gets more valid matching keypoints than the network without task alignment. The extra keypoints are shown in red. It can be seen that they are still in discriminative locations.

The 3D Safe Window is used to avoid selecting negatives on repetitive structures. The results in Table III show that 3D SW works well with DIW although using it alone is not quite effective. This is because that the triplets found by 3D SW are usually with high intrinsic importance, which are ignored by optimizing over the previous triplet loss and is conquered by the proposed dynamic importance weighting.

VI. CONCLUSION

In this letter, we present the TaSP, an end-to-end keypoint detection and descriptor extraction network. It uses a task alignment strategy to detect keypoints corresponding to discriminative descriptors. To train discriminative descriptors, it uses a dynamic importance weighting module, which concerns both intrinsic importance and empirical importance. Besides, it uses a 3D safe window to obtain more informative triplets. We have conducted extensive experiments to study the effect of each modification and demonstrated the superiority of our approach in various applications.

REFERENCES

[1] T. Sattler et al., "Benchmarking 6DoF outdoor visual localization in changing conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8601–8610.
[2] C. Toft et al., "Long-term visual localization revisited," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2074–2088, Apr. 2022.
[3] N. Piasco, D. Sidibé, V. Gouet-Brunet, and C. Demonceaux, "Learning scene geometry for visual localization in challenging conditions," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 9094–9100.
[4] Z. Liao, Y. Hu, J. Zhang, X. Qi, X. Zhang, and W. Wang, "So-SLAM: Semantic object SLAM with scale proportional and symmetrical texture constraints," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 4008–4015, Apr. 2022.

[5] R. Giubilato, W. Stürzl, A. Wedler, and R. Triebel, "Challenges of SLAM in extremely unstructured environments: The DLR planetary stereo, solid-state LiDAR, inertial dataset," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 8721–8728, Oct. 2022.

[6] H. Lim, J. Jeon, and H. Myung, "UV-SLAM: Unconstrained line-based SLAM using vanishing points for structural mapping," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 1518–1525, Apr. 2022.

[7] N. Sünderhauf et al., "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. Robot.: Sci. Syst.*, 2015, vol. XI, pp. 1–10.

[8] M. Venator, E. Bruns, and A. Maier, "Robust camera pose estimation for unordered road scene images in varying viewing conditions," *IEEE Trans. Intell. Veh.*, vol. 5, no. 1, pp. 165–174, Mar. 2020.

[9] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 467–483.

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 224–236.

[11] M. Dusmanu et al., "D2-Net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8092–8101.

[12] J. Revaud et al., "R2D2: Repeatable and reliable detector and descriptor," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[13] M. Venator, Y. E. Himer, S. Aklanoglu, E. Bruns, and A. Maier, "Self-supervised learning of domain-invariant local features for robust visual localization under challenging conditions," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 2753–2760, Apr. 2021.

[14] Z. Luo et al., "ASLFEAT: Learning local features of accurate shape and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6589–6598.

[15] M. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 14254–14265.

[16] A. Bhowmik, S. Gumhold, C. Rother, and E. Brachmann, "Reinforced feature points: Optimizing feature detection and description for a high-level task," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4948–4957.

[17] U. S. Parihar et al., "RoRD: Rotation-robust descriptors and orthographic views for local feature matching," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 1593–1600.

[18] F. Yuan, P. Neubert, S. Schubert, and P. Protzel, "SoftMP: Attentive feature pooling for joint local feature detection and description for place recognition in changing environments," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 5847–5853.

[19] B. Fan, Y. Yang, W. Feng, F. Wu, J. Lu, and H. Liu, "Seeing through darkness: Visual localization at night via weakly supervised learning of domain invariant features," *IEEE Trans. Multimedia*, early access, Feb. 24, 2022, doi: 10.1109/TMM.2022.3154165.

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

[21] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[22] A. Kendall, M. Grimes, and R. Cipolla, "POSENet: A convolutional network for real-time 6-DoF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2938–2946.

[23] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPNP: An accurate O (N) solution to the PNP problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.

[24] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12716–12725.

[25] Z. Yang, X. Yu, and Y. Yang, "DSC-POSENet: Learning 6DoF object pose estimation via dual-scale consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3907–3916.

[26] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4104–4113.

[27] G. Evangelidis and B. Micusik, "Revisiting visual-inertial structure-from-motion for odometry and SLAM initialization," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 1415–1422, Apr. 2021.

[28] A. Barroso-Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, " Key. Net: Keypoint detection by handcrafted and learned CNN filters," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5836–5844.

[29] Y. Tian, V. Balntas, T. Ng, A. Barroso-Laguna, Y. Demiris, and K. Mikolajczyk, "D2D: Keypoint extraction with describe to detect approach," in *Proc. Asian Conf. Comput. Vis.*, 2020.

[30] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016.

[31] F. Schroff, D. Kalenichenko, and J. Philbin, "FACENet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[32] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[33] Q. Wang, X. Zhou, B. Hariharan, and N. Snavely, "Learning feature descriptors using camera pose supervision," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 757–774.

[34] L. Zhang and S. Rusinkiewicz, "Learning local descriptors with a CDF-based dynamic soft margin," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2969–2978.

[35] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9308–9316.

[36] X. Zhao, X. Wu, J. Miao, W. Chen, P. C. Chen, and Z. Li, "ALIKE: Accurate and lightweight keypoint detection and descriptor extraction," *IEEE Trans. Multimedia*, early access, Mar. 03, 2022, doi: 10.1109/TMM.2022.3155927.

[37] H. Germain, V. Lepetit, and G. Bourmaud, "Neural reprojection error: Merging feature learning and camera pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 414–423.

[38] K. Li, L. Wang, L. Liu, Q. Ran, K. Xu, and Y. Guo, "Decoupling makes weakly supervised local feature better," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15838–15848.

[39] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1735–1742.

[40] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[41] C. Zauner, "Implementation and benchmarking of perceptual image hash functions," Master's thesis, Upper Austria Univ. Appl. Sci., Hagenberg Campus, Hagenberg, Austria, 2010.

[42] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3490–3499.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[44] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from internet photos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2041–2050.

[45] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 501–518.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[47] H. Taira et al., "INLOC: Indoor visual localization with dense matching and view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7199–7209.

[48] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NETVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[49] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "HPatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5173–5182.

[50] J. L. Schonberger, H. Hardmeier, T. Sattler, and M. Pollefeys, "Comparative evaluation of hand-crafted and learned local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1482–1491.