Supplementary Material for Decoupling Makes Weakly Supervised Local Feature Better

Kunhong Li[1,2] Longguang Wang[3] Li Liu[3,4] Qing Ran[5] Kai Xu[3] Yulan Guo[1,2,3*]
[1]Sun Yat-Sen University [2]The Shenzhen Campus of Sun Yat-Sen University
[3]National University of Defense Technology [4]University of Oulu [5]Alibaba Group

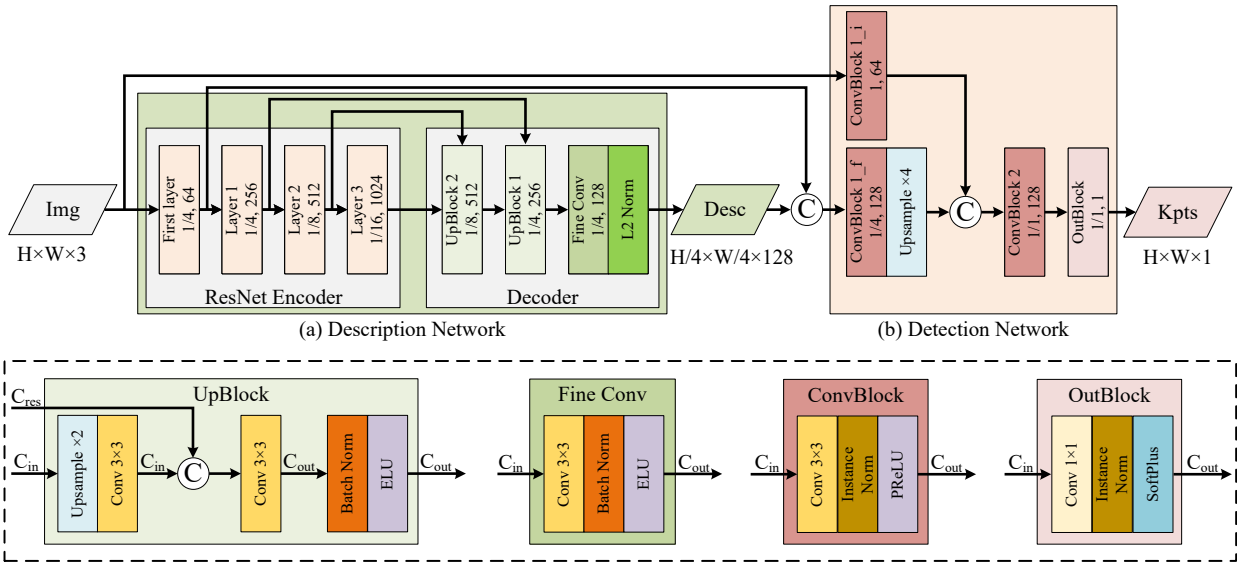


Figure 1. Model architecture. The description network (a) and detection network (b) consist of several blocks, we report the scale and output channels of each block, and illustrate the details of each block in the bottom box.

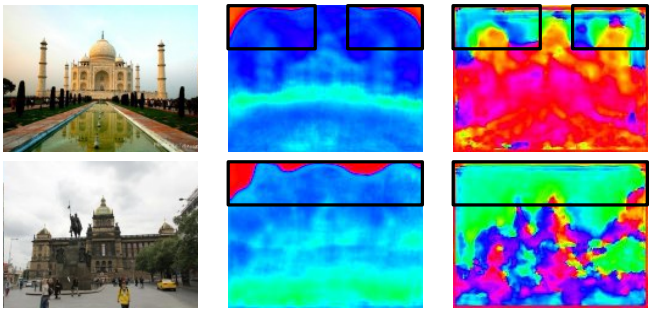
We first introduce the details of our model in Sec. 1 and discuss the query points generation during description network training in Sec. 2. Then, we expand the detection network training in Sec. 3. Next, we show detailed experimental settings in Sec. 4. After that, we give a discussion on the limitations and broader impact of our PoSFeat in Sec. 5. Finally, additional qualitative results are included in Sec. 6.

1. Model Architecture

Our model consists of two parts, *i.e*. the description network and the detection network, as illustrated in Fig. 1. For description network, we adopts the ResUNet used in [9], which follows a widely used encoder-decoder architecture. We use a truncated ResNet-50 [4] (pre-trained on ImageNet [2]) as the encoder, and use several 3×3 convolution layers combining with bilinear upsampling and residual connection to construct the decoder. For detection network, we use a simple three-layer architecture. The first layer takes the original image and two feature maps from description network as inputs, and aggregate the original image and feature maps from description network for detection. For better aggregation of original image and feature maps, we use the instance normalization [8] instead of batch normalization [5] in our detection network.

2. Query Points Generation in Description Network Training

We adopt grid-based random sampling to select query points for the training of description network to avoid the bias of pre-defined keypoints. When pre-defined keypoints (*e.g*. SIFT) are used to train the description network, the densities of SIFT keypoints in different areas vary a lot.



(a) original images (b) SIFT (c) random sample

Figure 2. An illustration of pre-defined keypoints bias. We adopt PCA [3] to visualize the descriptors of the original images (a). When the description network is trained with SIFT (b), there are insufficiently trained areas (black boxes), which leads to false keypoints detection. On the contrary. When the description network is trained with grid-based random sample (c), all the areas in the image will be sufficiently trained.

Consequently, areas with few SIFT keypoints are usually under optimized, as shown in Fig. 2. This bias limits the discriminativeness of the descriptors and leads to detection network produces considerable false keypoints detection. To address this problem, we use grid-based random sampling to generate query points. Specifically, we first split the image into N_g grids with the shape of $g \times g$. Then we uniformly select N_g points with with one point in a grid. With this grid-based random sample strategy, the description network will be sufficiently trained in all areas, and thus detection network can produce more accurate keypoints.

3. Detection Network Training

In this section, we present more details on the detection network training.

As described in the main paper, we first extract the feature maps F_1 and F_2 from a image pair I_1 and I_2 with the frozen description network. Then we feed F_1 and F_2 into the detection network to produce the keypoint heatmaps, and model the keypoint distributions based on the heatmaps. Specifically, we divide these heatmaps into grids and select at most one keypoint from each grid cell. For a pixel $\boldsymbol{x}$ in image I_1, the probability that $\boldsymbol{x}$ is a keypoint can be formulated as,

$$P_{kp}(\boldsymbol{x}|F_1) = \mathrm{Softmax}(F_1^{G_{\boldsymbol{x}}})_{\boldsymbol{x}} \cdot \mathrm{Sigmoid}(F_1)_{\boldsymbol{x}}, \quad (1)$$

in which $F_1^{G_{\boldsymbol{x}}}$ denotes the local heatmap of the grid cell that contains pixel $\boldsymbol{x}$, $\mathrm{Softmax}(F_1^{G_{\boldsymbol{x}}})_{\boldsymbol{x}}$ represents the local probability of $\boldsymbol{x}$ to be a keypoint, and $\mathrm{Sigmoid}(F_1)_{\boldsymbol{x}}$ represents the global probability of $\boldsymbol{x}$ to be a keypoint.

According to the keypoint probability distribution, we then select the candidate sets $Q_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots | \boldsymbol{x}_i \in I_1\}$ and $Q_2 = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots | \boldsymbol{y}_i \in I_2\}$ to compute the similarity matrix S, whose elements are defined as,

$$S_{i,j} = F_1(\boldsymbol{x}_i) \times F_2(\boldsymbol{y}_j)^{\mathrm{T}}, \boldsymbol{x}_i \in Q_1\ \boldsymbol{y}_j \in Q_2. \quad (2)$$

Afterwards, we can compute the matching probability P_m according to the similarity matrix,

$$P_m = \mathrm{Softmax}(S)_1 \cdot \mathrm{Softmax}(S)_2, \quad (3)$$

in which $\mathrm{Softmax}(S)_1$ and $\mathrm{Softmax}(S)_2$ denotes the softmax operation along the row and column, respectively.

Next, we compute the rewards according to the epipolar constraints,

$$R(\boldsymbol{x}_i, \boldsymbol{y}_j) = \begin{cases} \lambda_p, & \text{if distance}(\boldsymbol{y}_j, L_{\boldsymbol{x}_i}) \leq \epsilon \\ \lambda_n, & \text{if distance}(\boldsymbol{y}_j, L_{\boldsymbol{x}_i}) > \epsilon \end{cases}, \quad (4)$$

where the reward threshold ϵ is empirically set to 2. Since the description network is frozen and reliable, the matches with low matching probability are unreliable, and thus we further truncate the matching probability P_m according to the reward to omit the false positive rewards for the unreliable matches. Specifically, we manually set $P_m(\boldsymbol{x}_i, \boldsymbol{y}_j) = 0$ for the pairs $(\boldsymbol{x}_i, \boldsymbol{y}_j)$ whose reward $R(\boldsymbol{x}_i, \boldsymbol{y}_j) = \lambda_p$ and matching probability $P_m(\boldsymbol{x}_i, \boldsymbol{y}_j) < 0.9$.

Finally, we compute the loss for detection network,

$$\begin{aligned} \mathcal{L}_{kp} = -\frac{1}{|Q_1| + |Q_2|} \Big(& \sum_{\boldsymbol{x}_i, \boldsymbol{y}_j} \mathcal{L}_{rew}(\boldsymbol{x}_i, \boldsymbol{y}_j) \\ & + \lambda_{reg} \big(\sum_{\boldsymbol{x}_i} \log P_{kp}(\boldsymbol{x}_i) + \sum_{\boldsymbol{y}_j} \log P_{kp}(\boldsymbol{y}_j) \big) \Big), \end{aligned} \quad (5)$$

where λ_{reg} is a regularization penalty and the reward loss $\mathcal{L}_{rew}(\boldsymbol{x}_i, \boldsymbol{y}_j)$ is defined as:

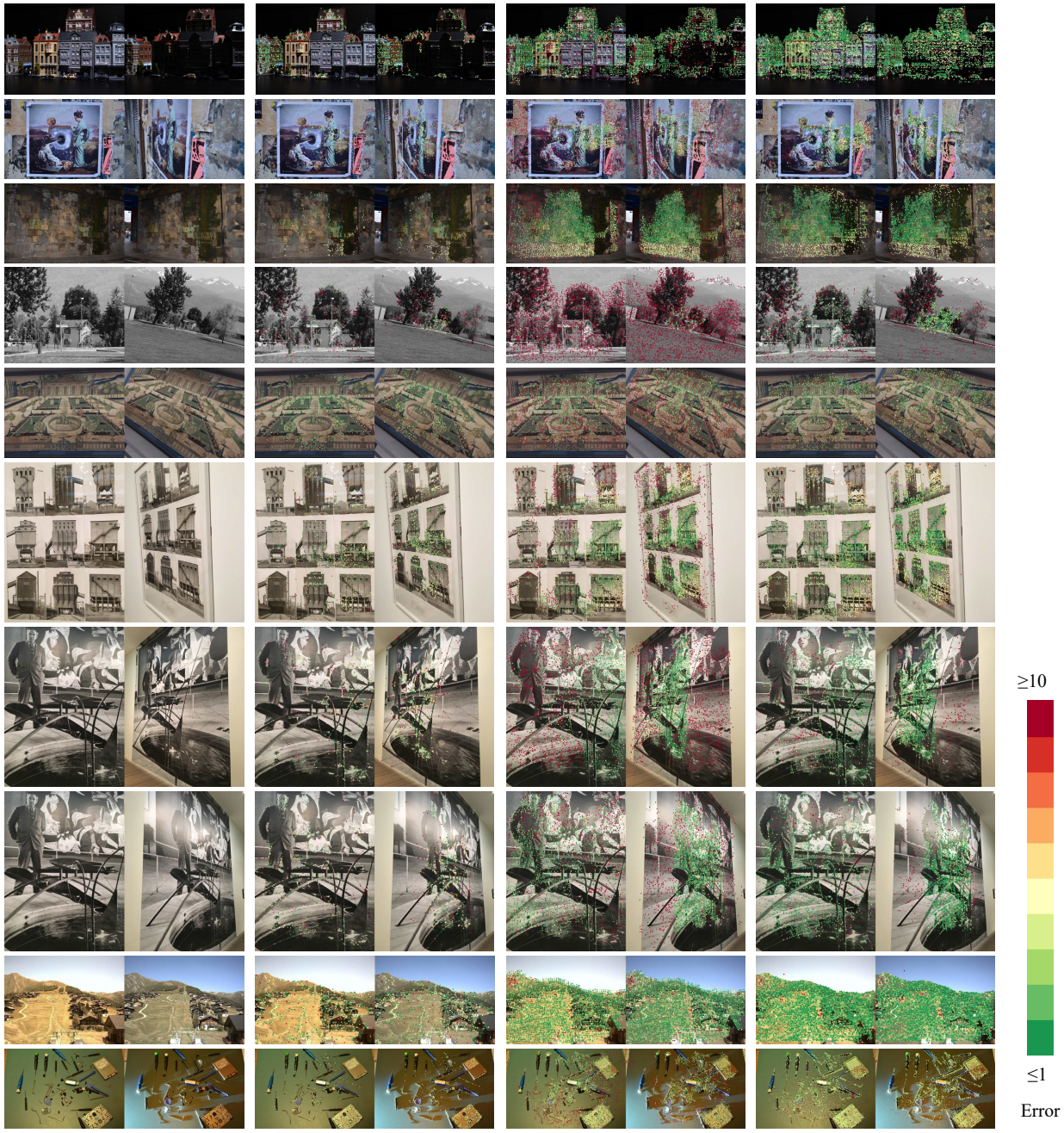
$$\mathcal{L}_{rew}(\boldsymbol{x}_i, \boldsymbol{y}_j) = P_m(\boldsymbol{x}_i, \boldsymbol{y}_j) \cdot R(\boldsymbol{x}_i, \boldsymbol{y}_j) \cdot \log(P_{kp}(\boldsymbol{x}_i) P_{kp}(\boldsymbol{y}_i)). \quad (6)$$

Note that, the P_m is truncated according to the rewards and thus the match pairs with positive rewards but low matching probability are left neutral.

4. Experimental Settings

In this section, we present the hyper-parameters of our method on different datasets. During inference, we apply non-maximum suppression (NMS) to detect keypoints, and use a mutual nearest neighbour matcher for matching. Instead of resizing the images, we crop the images from the top-left side to guarantee both the height and width of the images are divisible by 16.

HPatches Dataset [1]. The NMS size is set to be 3×3 due to the existence of low-resolution images, and the maximum keypoint numbers are limited to be 8192.



(a) Image pairs (b) SIFT + CAPS (c) DISK-W (d) PoSFeat (Ours)

Figure 3. Qualitative results on HPatches. The same as figures in main paper, only successfully matched keypoints are shown and colored according to their match errors. The colorbar is shown on the right. Best viewed in color.

Aachen Day-Night Dataset [10]. Because of the high image resolutions, the NMS size is set to be 7×7 on the Aachen Day-Night dataset, and the maximum keypoint numbers are limited to be $16k$. Note that, keypoints with scores smaller than 0.9 in the heatmaps are filtered out.

ETH Local Feature Benchmark [7]. The NMS size is set to be 7×7, and the maximum keypoints numbers are limited to be $20k$. Keypoints with scores smaller than 0.9 in the heatmaps are also filtered out. We additionally applying ratio test during matching with a threshold 0.8 to achieve robust reconstruction.

(a) Aachen Day-Night v1

(b) Aachen Day-Night v1.1

Figure 4. The sparse 3D models of Aachen. These models are reconstructed using Colmap [6] with features extracted by PoSFeat, and are further used to do night-time images localization. Note that these models are reconstructed based on the camera poses provided by the author of the dataset.

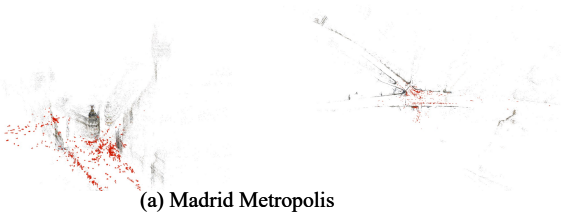
5. Limitations and Broader Impact

The PoSFeat suffers limited capability to deal with large rotation and scale changes. On the HPatches dataset, our PoSFeat produces limited performance on several scenes with pure rotation. On the ETH local feature benchmark, our PoSFeat cannot well handle the scenes with extreme scale changes thus has limited performance in #Imgs (*e.g.*, only 419 images are registered in Mardrid Metropolis).

The PoSFeat is a general local feature method, although we only apply it with image matching, visual localization and 3D reconstruction in our paper, it can be easily extended to recognize or reconstruct human faces. Therefore, the researches and the applications about the recognition or reconstruction of human faces must strictly respect the personality rights and privacy regulations.

6. Visualization

We present some qualitative results in this section. The image matching results on HPatches are shown in Fig. 3. The 3D models of Aachen are illustrated in Fig. 4, which is reconstructed with the features extracted by PoSFeat, and is used to do visual localization on Aachen Day-Night dataset . And the 3D reconstruction results on ETH local feature benchmark are shown in Fig. 5.



(a) Madrid Metropolis

(b) Gendarmenmarkt

(c) Tower of London

Figure 5. The sparse 3D reconstruction results on ETH local feature benchmark.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[3] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM review*, 53(2):217–288, 2011.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[5] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[6] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016.

[7] Johannes L Schonberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys. Comparative Evaluation of Hand-Crafted and Learned Local Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1482–1491, 2017.

[8] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[9] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning Feature Descriptors Using Camera Pose Supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12346, pages 757–774, 2020.

[10] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference Pose Generation for Long-term Visual Localization via Learned Features and View Synthesis. *International Journal of Computer Vision (IJCV)*, 129(4):821–844, 2021.