Digging into Self-Supervised Learning of Local Descriptors

Iaroslav Melekhov* Zakaria Laskar* Xiaotian Li Shuzhe Wang Juho Kannala

*equal contribution

Aalto University

International Conference on 3D Vision
December 1-3, 2021 Online
3DV 2021

Problem statement

Fully-supervised CNN-based approaches for learning local image descriptors require per-pixel ground-truth keypoint correspondence data which is difficult to acquire at scale. In this work, we focus on understanding the limitations of existing self-supervised approaches and propose a set of improvements that combined lead to powerful feature descriptors.

Challenges

Existing self-supervised approaches employ in-pair negative mining. However:

1. Supervised local descriptor learning methods show superior performance with in-batch negative mining. ***In self-supervised setting, it does not guarantee selection of true negatives. In randomized batch training does it have any impact?***
2. Global mining has been accessible only to global representation learning methods. Local mining cos***t grows exponentially with dataset size. Since global descriptors are functions of local descriptors, can local mining across the dataset be approximated with global mining?***

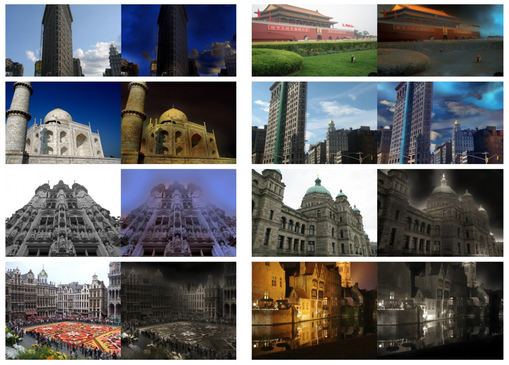

Figure 1: Stylization creates realistic illumination variation.

Method

To address the challenges, we propose a system shown in Fig. 2. All components are self-supervised and do not require human or costly machine guided labelling.

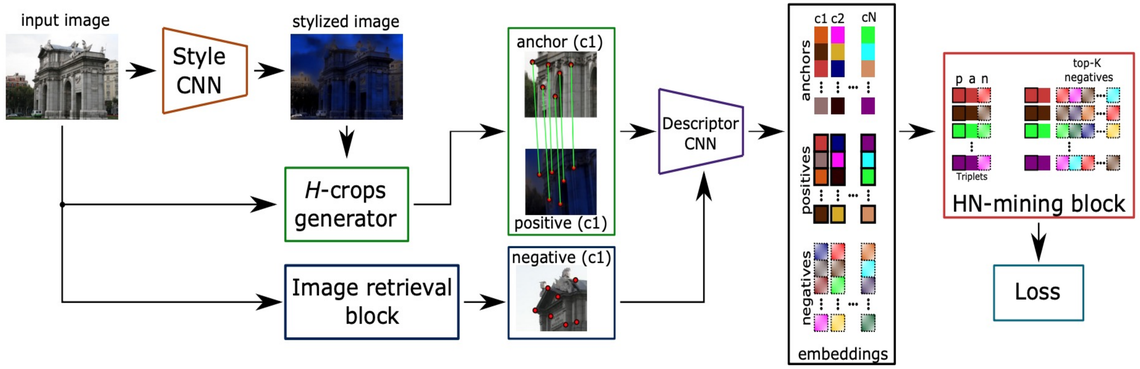

Figure 2: Proposed self-supervised local descriptor learning system

1. **HN Mining block**. This block selects the negatives for a given descriptor set. We use both in-pair and in-batch mining. The number of positives and negatives is balanced by selecting only the top-k negatives in terms of cosine similarity.
2. **Image Retrieval.** To enable global mining of local hard negative descriptors we use advances in global image retrieval. For each image in a training batch, in a coarse-to-fine manner, we first find hard-negative images in terms of global similarity. From the pool of global hard-negative images, the top-k hard-negative local descriptors are selected.
3. **Stylization.** Following Li et al. [37] we use a closed-form solution to image stylization (Fig. 1) based on a deep learning model trained in a self-supervised fashion by minimizing the sum of the reconstruction loss and perceptual loss. The method photorealistically transfers style of a reference photo to a content photo preserving local scene geometry.

imelekhov.com/hndesc

Results

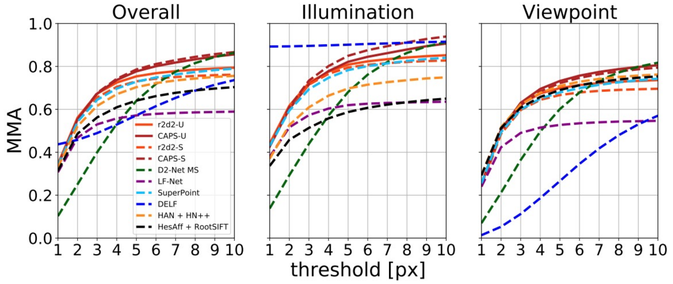

Figure 3: Mean Matching accuracy on HPatches dataset. Proposed self-supervised methods suffixed with –U perform comparable or better than supervised counterparts suffixed with -S

	Metric		Negative sampling type: in-pair	Negative sampling type: in-batch
*R*Oxford5k	mAP	M	55.38	**58.69**
		H	29.67	**33.17**
	mP@k [1, 5, 10]	M	[94.29, 86.57, 78.43]	**[95.71, 89.71, 83.29]**
		H	[82.86, 54.57, 42.29]	**[85.71, 60.29, 45.71]**
HPatches	MMA	1px	0.239 / 0.425 / 0.332	**0.254 / 0.439 / 0.346**
		3px	0.585 / 0.677 / 0.631	**0.630 / 0.707 / 0.669**
		5px	0.648 / 0.742 / 0.695	**0.706 / 0.784 / 0.745**
Aachen		day	87.9 / 94.2 / 97.9	**88.2 / 95.5 / 98.7**
		night	66.5 / 79.1 / 91.6	**68.1 / 83.8 / 94.8**

Table 1: **(Up)** Results show superior performance of in-batch over in-pair negative mining. **(Bottom)** Self-supervised methods perform at par with supervised counterparts on Aachen vislocalization benchmark

	Method	Supervision	Training data	Aachen v1.1 % localized queries: Day (824 images) 0.25*m*, 2°	Day 0.5*m*, 5°	Day 5*m*, 10°	Night (191 images) 0.25*m*, 2°	Night 0.5*m*, 5°	Night 5*m*, 10°
Super	R2D2 [57]	OF	A+R	**88.6**	**95.4**	**98.9**	72.8	89.0	97.4
	R2D2*	OF	A	87.7	94.7	98.7	69.6	86.4	95.3
	CAPS [76]	SL+RP	M	85.3	**93.8**	97.9	**75.9**	88.5	**97.9**
Self-supervised	R2D2-(▦)	-	A	87.4	94.9	98.3	63.9	80.1	92.1
	R2D2-(▦, ◨)	-	A	88.0	94.8	98.2	70.2	86.4	95.8
	R2D2-(▦, ◨)	-	M	87.4	94.7	98.3	72.3	88.5	97.4
	R2D2-(▦, ◨)-gd	-	M	87.5	94.9	98.3	71.7	86.4	96.9
	R2D2-(▦, ◨)-selfgd	-	M	88.1	94.8	98.1	71.2	88.0	95.8
	R2D2-(▦, ◨)	-	M+P	88.2	95.1	98.5	**73.3**	**90.1**	**97.4**
	CAPS-(▦)	-	M	85.8	93.8	98.2	67.0	82.2	96.9
	CAPS-(▦, ◨)	-	M	85.1	93.2	97.8	71.7	87.4	**97.9**
	CAPS-(▦, ◨)-gd	-	M	**87.0**	**93.8**	**98.3**	73.8	**89.0**	97.4
	CAPS-(▦, ◨)-selfgd	-	M	86.9	**93.8**	98.1	71.7	**89.0**	97.4
	CAPS-(▦, ◨)	-	M+P	85.4	93.2	97.9	72.3	88.5	**97.9**