Aalto University

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

Digging Into Self-Supervised Learning of Feature Descriptors

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

*equal contribution

Current learned CNN-based descriptors

Fully- (weakly-) **_supervised_** methods

- ✓ Accurate and discriminative
- ✓ Good generalization
- ✓ Robust to illumination changes
- ✗ Require ground-truth data (SfM or relative camera poses)

Unsupervised methods

- ✗ Not very competitive
- ✓ Good generalization
- ✗ Poor illumination invariance
- ✓ Easy to get data

Goal

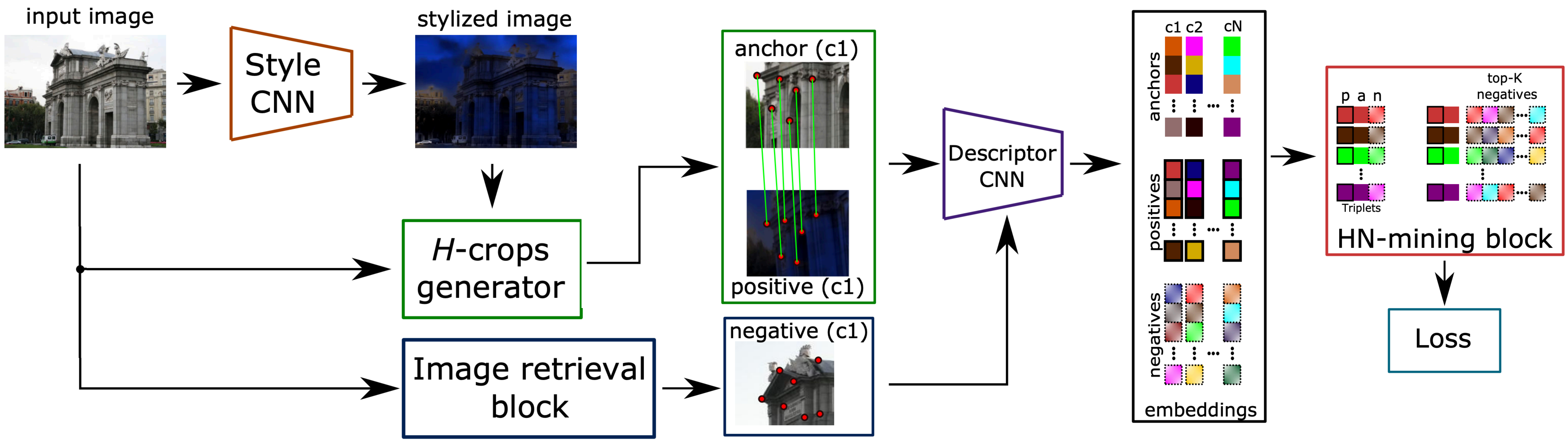
Credits: https://salesgravy.com/your-beliefs-bridge-the-gap-between-goals-and-results/

Goal

A method:

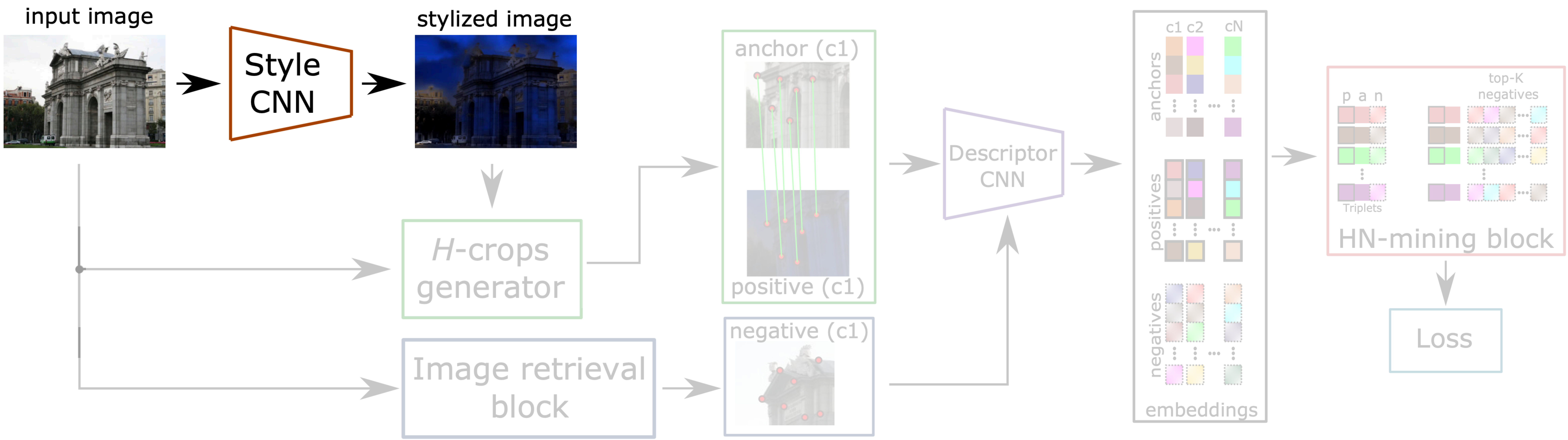
- ✓ Easy to get training data
- ✓ Good generalization
- ✓ Robust to illumination changes
- ✓ Competitive with fully- (weakly-) supervised methods

Our approach: HNDesc



- Robust to illumination changes
- Easy to get training data
- Good generalization
- Competitive with fully- (weakly-) supervised methods

HNDesc: Photorealistic Style Transfer



- ✔ Robust to illumination changes
- ✔ Easy to get training data
- ✔ Good generalization
- ✔ Competitive with fully- (weakly-) supervised methods

HNDesc: Photorealistic Style Transfer

- Following [1], we use the contributing views of AMOS Patches [2,3]

- The following 2 styles have been considered:

dusk

night

- The watermarks and timestamps have been removed

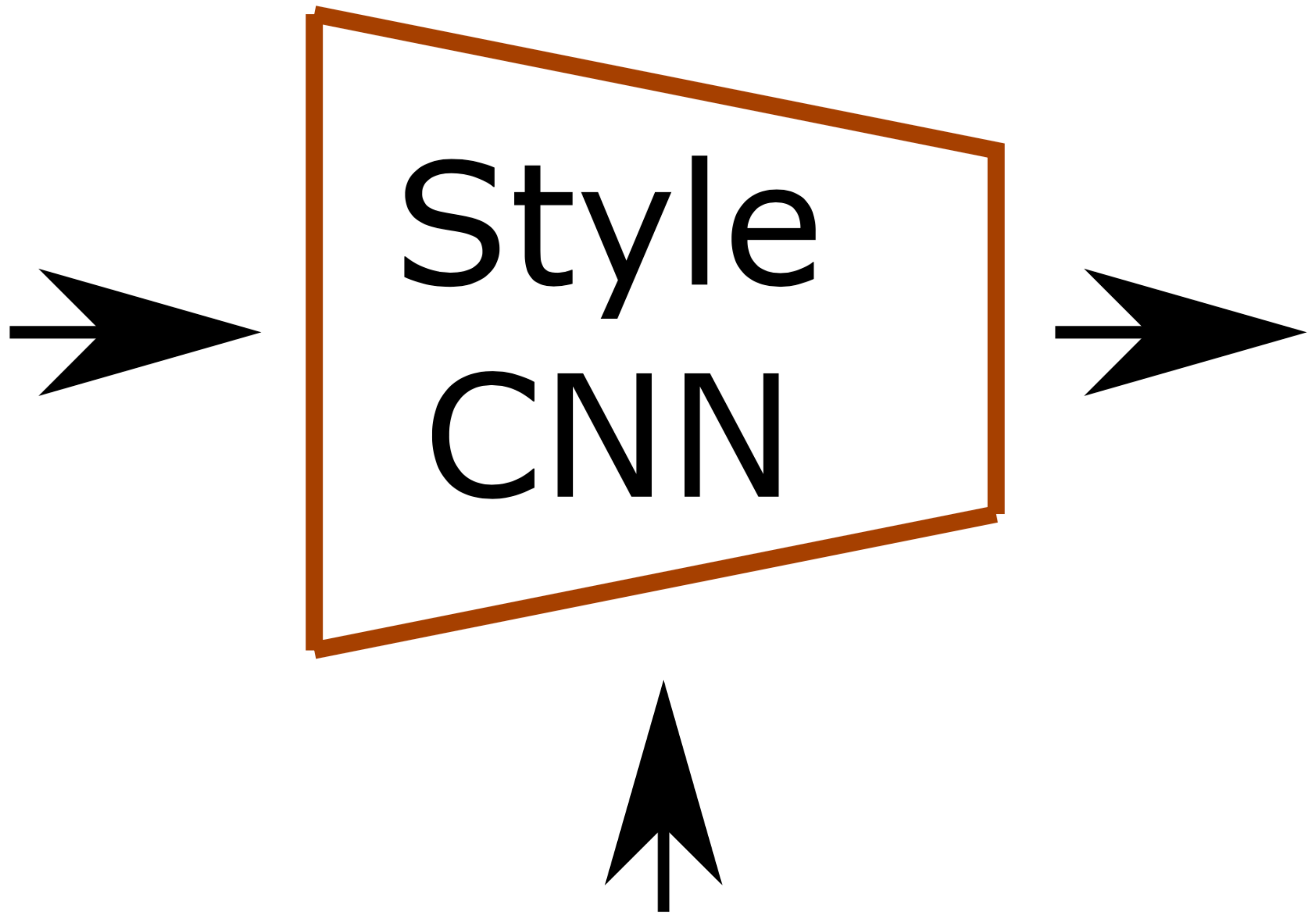
[1] Melekhov et al.: Image stylization for robust features. *ECCVW 2020*

[2] Jacobs et al.: Consistent temporal variations in many outdoor scenes. *CVPR* 2007

[3] Pultar et al.: Leveraging Outdoor Webcams for Local Descriptor Learning. *CVWW* 2019

HNDesc: Style CNN

content

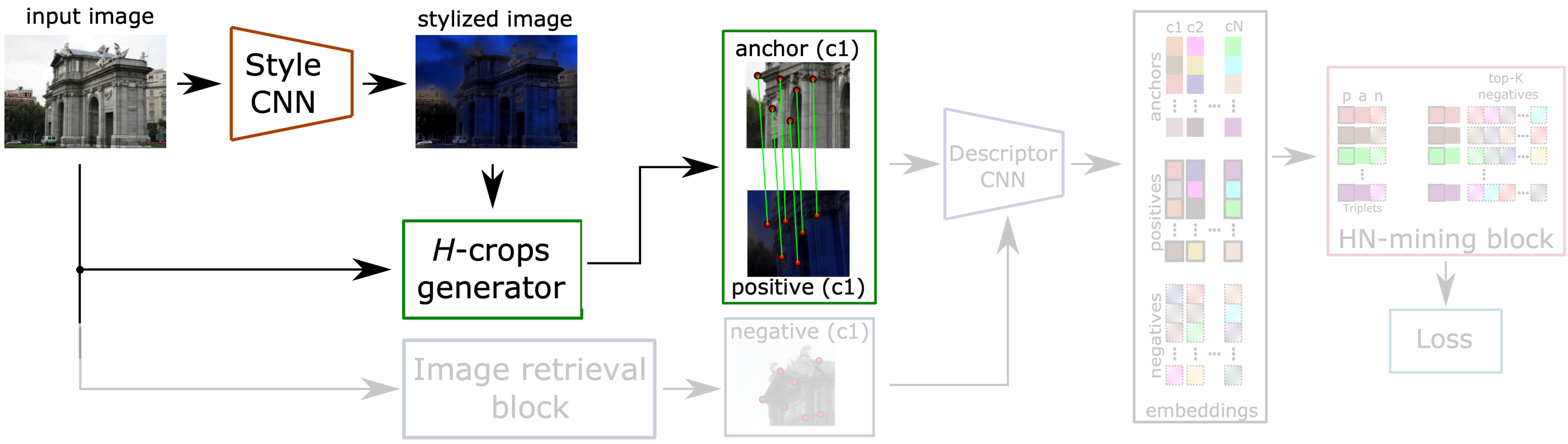


stylized image

style

[1] Li et al.: A Closed-form Solution to Photorealistic Image Stylization. *ECCV 2018*

HNDesc: H-crops generator



- ✓ Robust to illumination changes
- ✓ Easy to get training data
- ✓ Good generalization
- ✓ Competitive with fully- (weakly-) supervised methods

HNDesc: H-crops generator

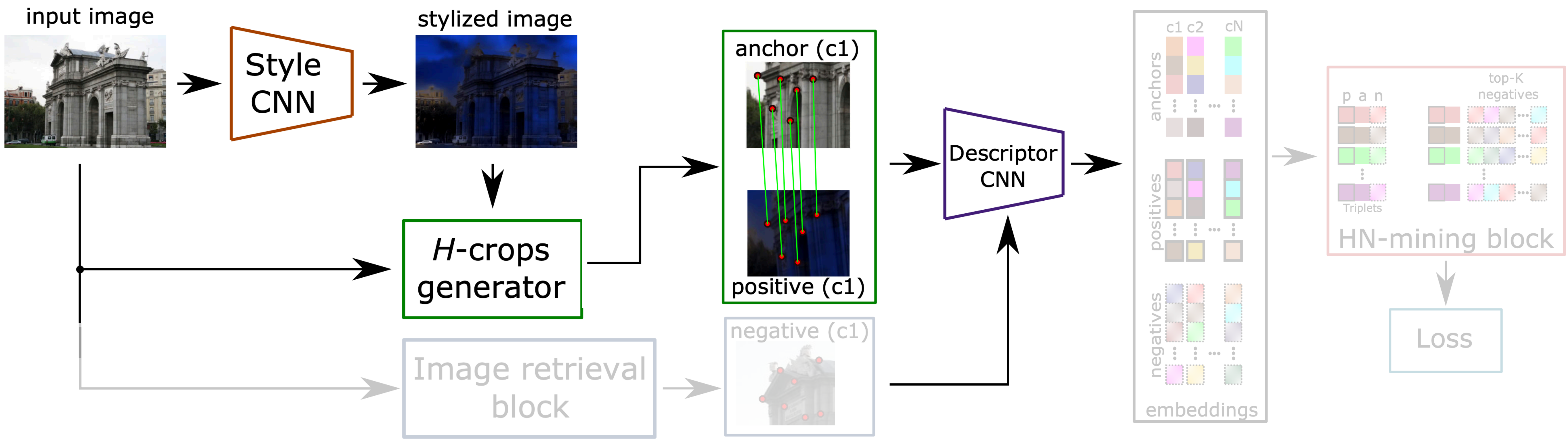
H-crops generator

crop 1

crop 2

$$\text{crop 1} = \mathscr{H}(\text{crop 2})$$

HNDesc: Descriptor CNN



- ✓ Robust to illumination changes
- ✓ Easy to get training data
- ✓ Good generalization
- ✓ Competitive with fully- (weakly-) supervised methods

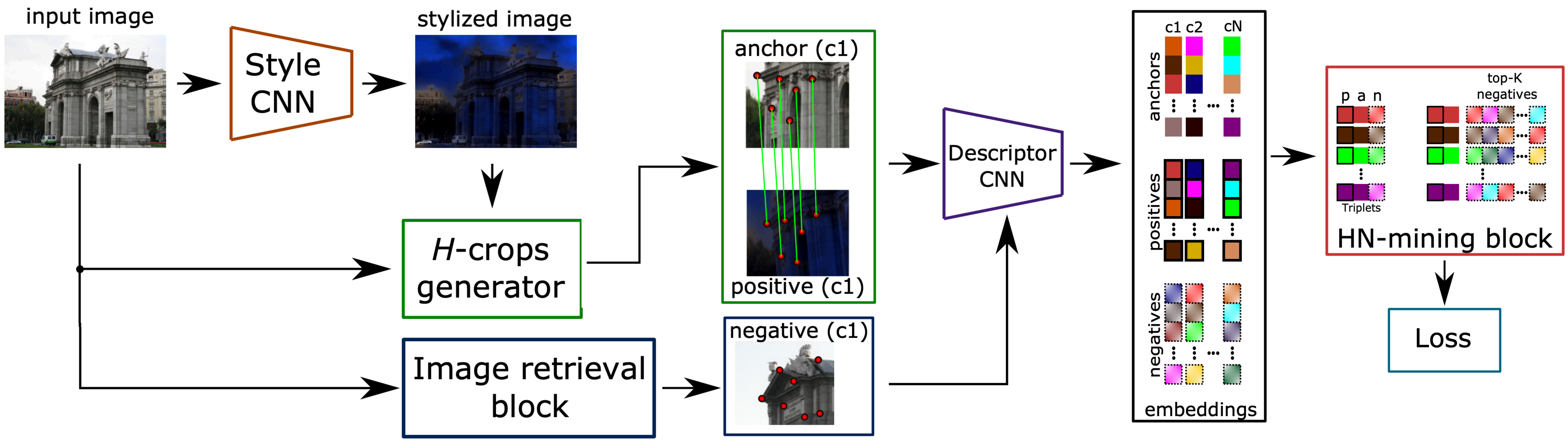
HNDesc: Descriptor CNN

- R2D2 architecture [1]
- CAPS (only fine descriptors are used) [2]

[1] Revaud et al.: R2D2: Reliable and repeatable detector and descriptor. NeurIPS 2019

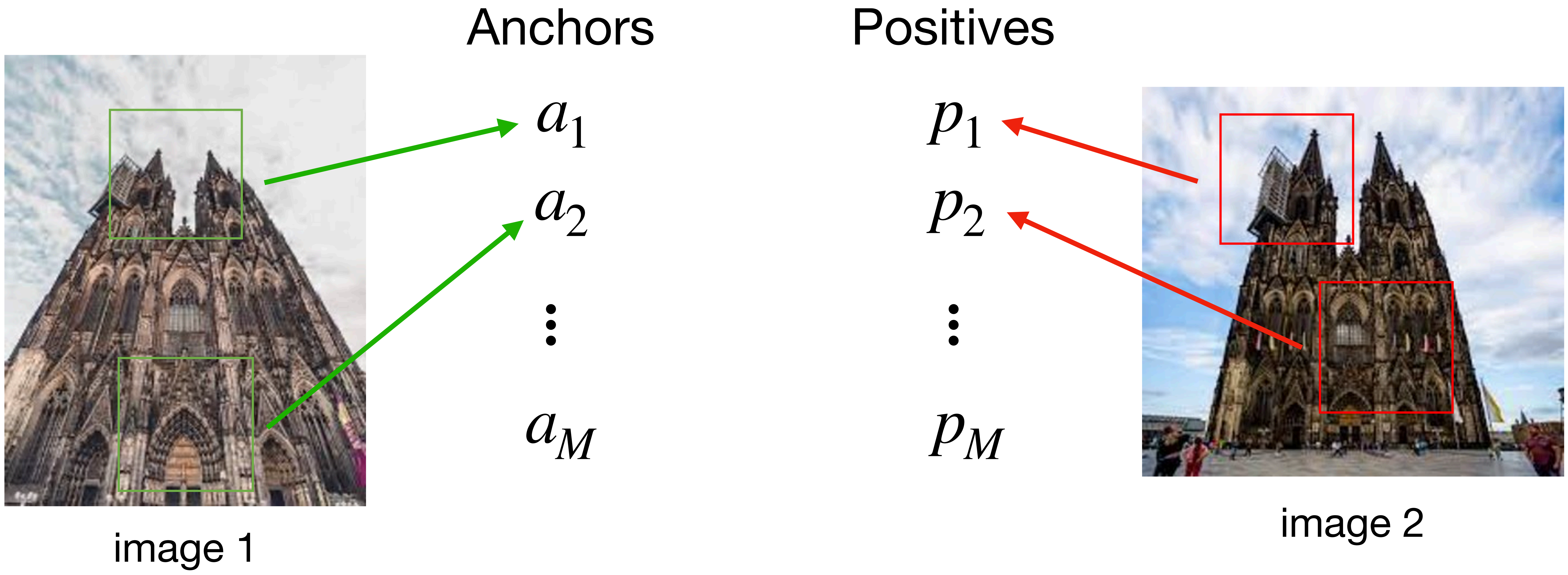
[2] Wang et. al: Learning feature descriptors using camera pose supervision. CVPR 2020

HNDesc: HN-mining block



- ✓ Robust to illumination changes
- ✓ Easy to get training data
- ✓ Good generalization
- ✓ Competitive with fully- (weakly-) supervised methods

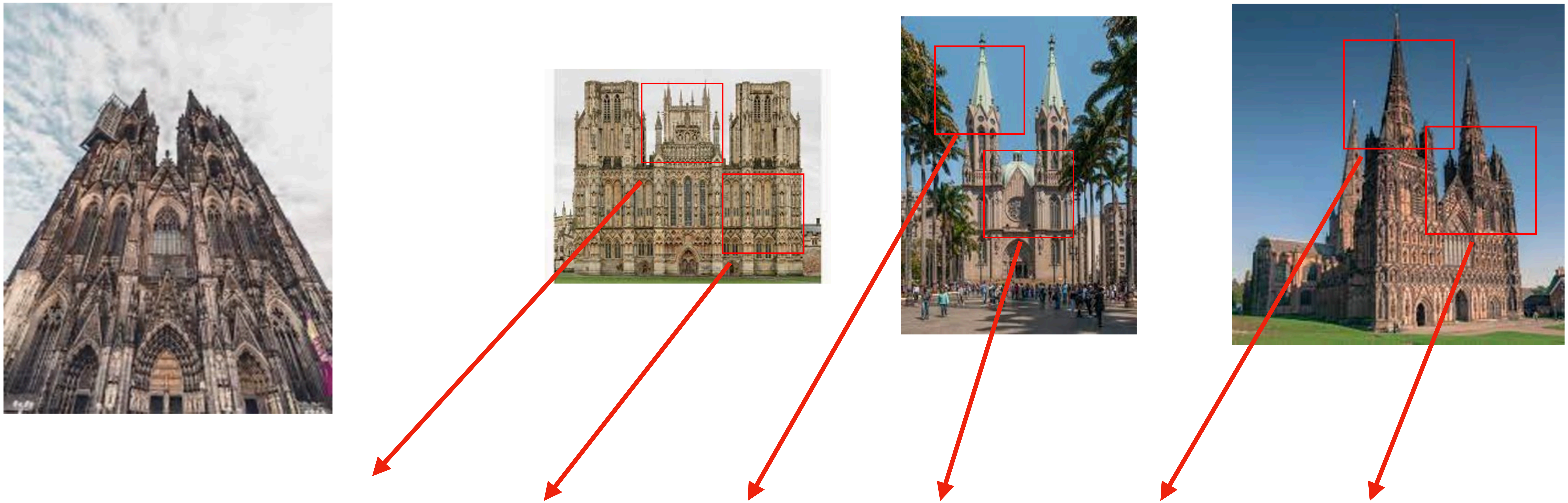
HNDesc: HN-mining block (in-pair sampling)



The index of non-matching descriptor p_n :

$$n = \underset{n=1..M, n \neq i}{\operatorname{argmax}} \; s\left(a_i, p_n\right)$$

HNDesc: HN-mining block (in-batch sampling)



$$\mathscr{P} = \{\, p_1 \ \cdots \ p_{12} \ \cdots \ p_{27} \ \cdots \ p_{39} \ \cdots \ p_{51} \ \cdots \ p_{73} \ \cdots \ p_M \,\}$$

The index of non-matching descriptor p_n :

$$n = \underset{n=1..M,\, n \neq i}{\operatorname{argmax}} \ s\left(a_i, p_n\right)$$

HNDesc: HN-mining block (in-pair vs. in-batch)

	Metric		Negative sampling type in-pair	in-batch
$\mathcal{R}$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**

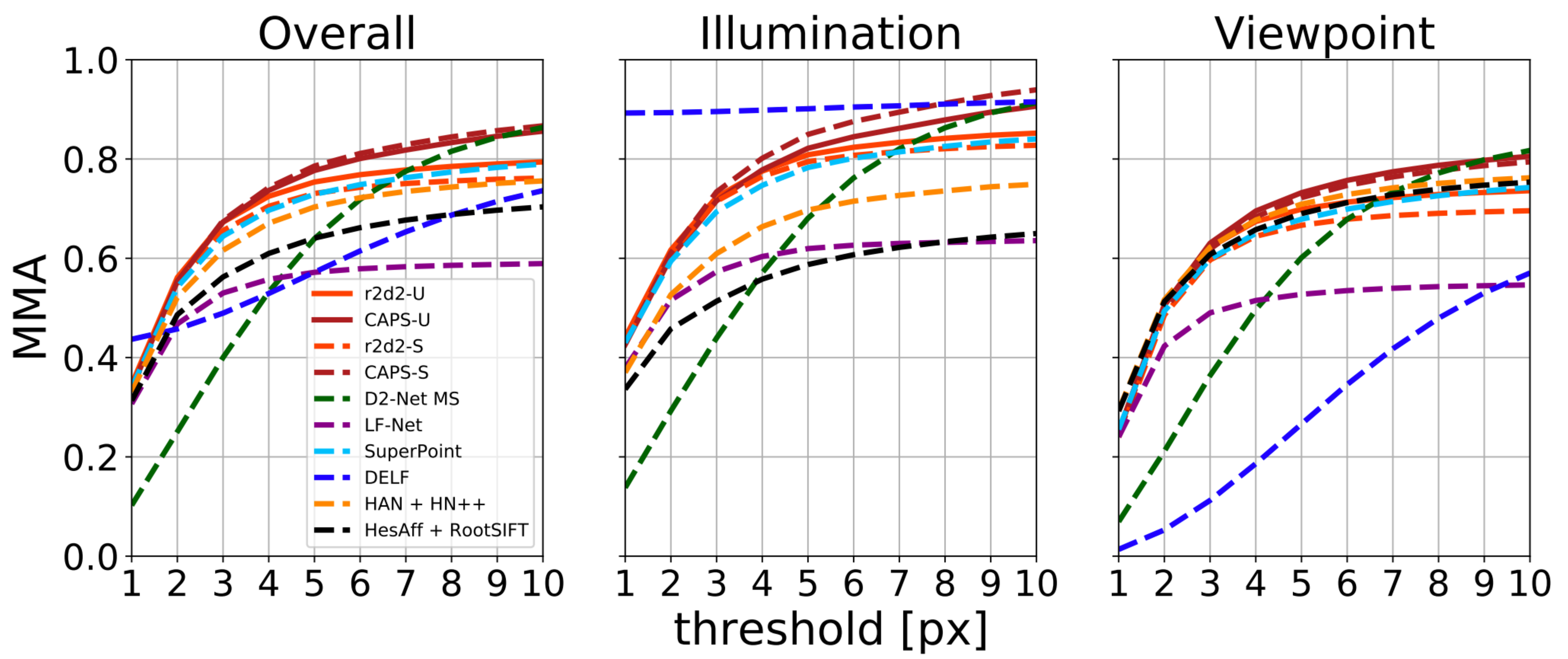
Training Pipeline

- We use Phototourism (P) and MegaDepth (M) datasets;
- For each image, we generate 12 stylized versions, i.e 6 for each of 2 styles;
- Adam optimizer, 1 RTX 2080Ti

Benchmarks

- Sparse feature matching (HPatches dataset);
- Image-based localization (Aachen Day-Night, Tokyo24/7 and InLoc datasets);
- Image retrieval (ROxford5k, RParis6k).

Benchmarks: Sparse Feature Matching



Color Augmentations (CA)

Style Transfer (ST)

, ST + CA

*-U == ,

Benchmarks: Sparse Feature Matching

Method	Illumination			Viewpoint		
	H	Precision	Recall	H	Precision	Recall
Root SIFT	0.933	**0.782**	0.799	0.566	0.651	0.527
HardNet [44]	0.940	0.702	0.731	0.664	0.701	0.734
SOSNet [70]	0.933	0.748	0.821	**0.698**	0.727	0.760
SuperPoint [17]	0.912	0.710	0.811	0.671	0.685	0.750
D2-Net [20]	0.905	0.725	0.775	0.617	0.666	0.664
LISRD [51]	**0.947**	0.766	0.920	0.688	0.731	0.757
R2D2 [57]	0.940	0.762	0.837	0.692	0.720	0.732
CAPS [76]	0.888	0.757	**0.938**	0.692	0.723	0.699
R2D2-(▦, ◨)	0.944	0.764	0.838	0.678	0.732	0.739
R2D2-(▦, ◨)-selfgd	0.933	0.761	0.817	0.678	0.715	0.705
R2D2-(▦, ◨)-gd	**0.947**	0.766	0.826	**0.698**	0.726	0.720
CAPS-(▦, ◨)	0.933	0.750	0.884	0.671	0.742	0.728
CAPS-(▦, ◨)-selfgd	0.937	0.756	0.893	0.661	0.740	0.752
CAPS-(▦, ◨)-gd	0.919	0.757	0.890	0.681	**0.747**	**0.762**

▦ Color Augmentations (CA)

◨ Style Transfer (ST)

▦, ◨ ST + CA

Benchmarks: Visual Localization

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

▦ Color Augmentations (CA)

◨ Style Transfer (ST)

▦, ◨ ST + CA

HNDesc - unsupervised local descriptor **imelekhov.com/hndesc**

HNDesc = synthetic homography + photorealistic style transfer + HN mining