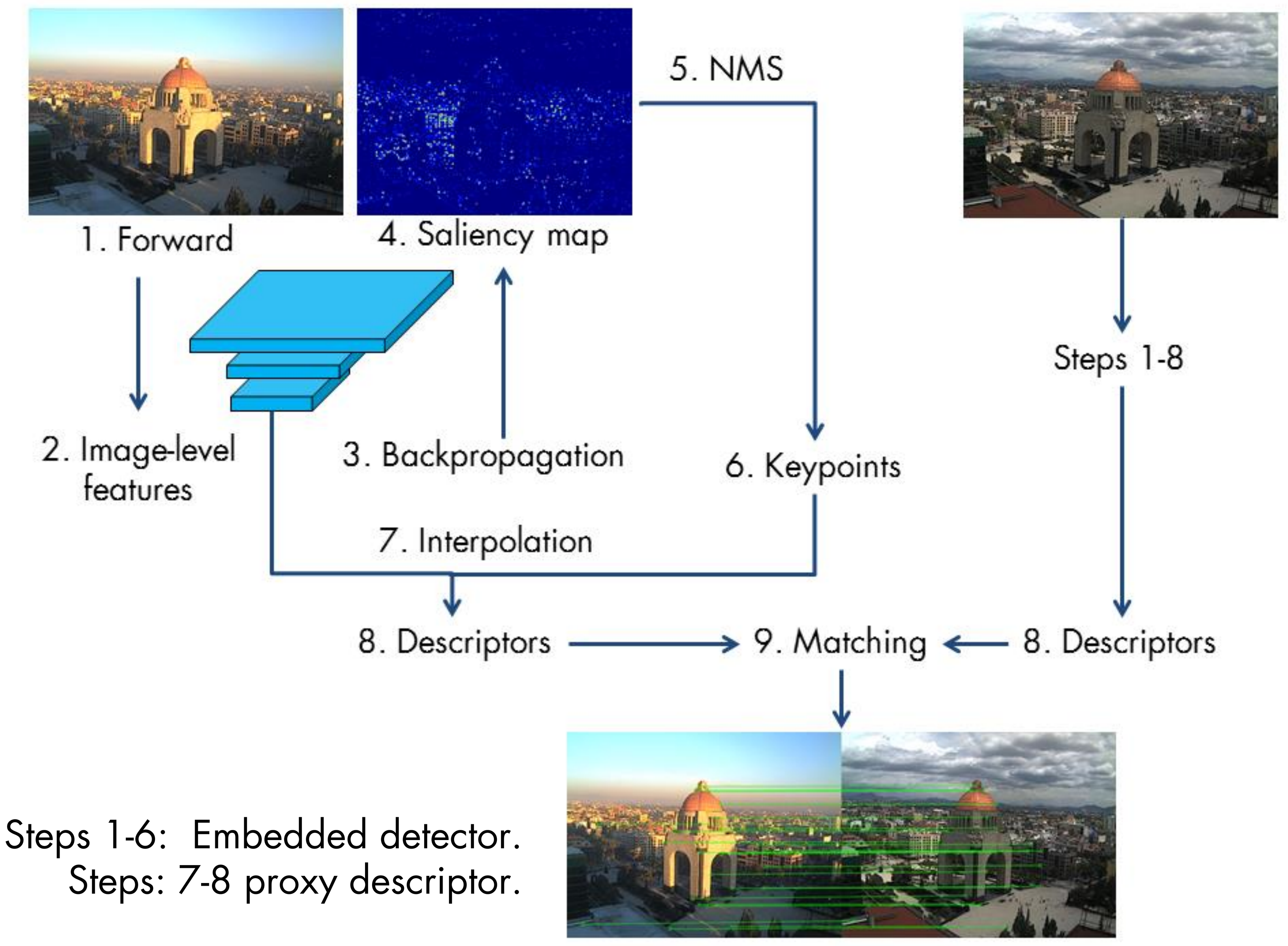
ELF: Embedded Localisation of Features in pre-trained CNN

Assia Benbihi, Matthieu Geist, Cédric Pradalier
abenbihi@georgiatech-metz.fr
Georgia Tech | cnrs — Unité Mixte Internationale 2958
CentraleSupélec
UNIVERSITÉ DE LORRAINE

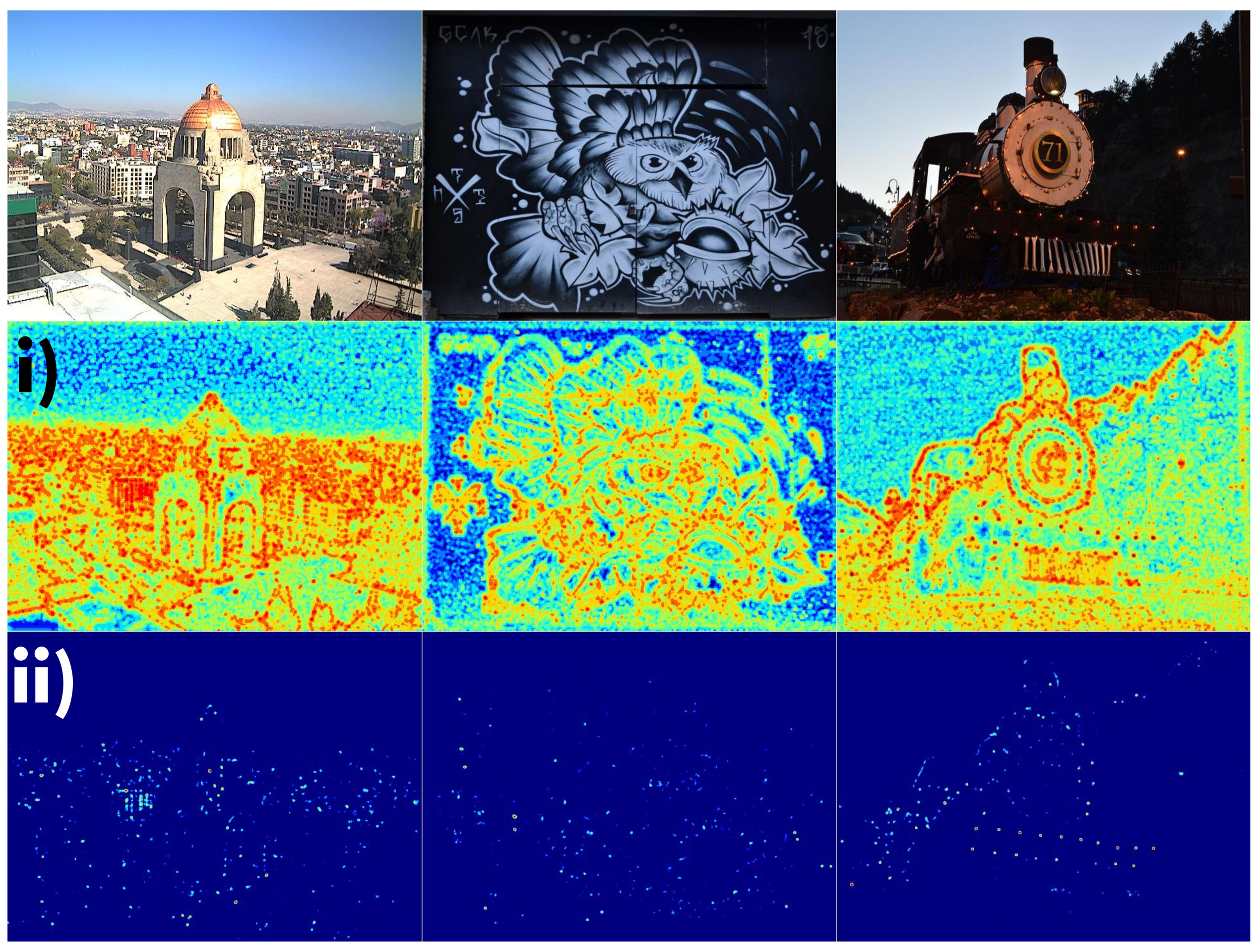
Abstract: ELF is a novel feature detector based only on information embedded inside a CNN already trained on a standard learning task (*e.g.* classification). This information is extracted from the gradient of the feature map with respect to the input image. It provides a saliency map with local maxima on the relevant keypoint locations. We compare our method to hand-crafted and learned feature matching pipelines and reach comparable performances although our method requires neither supervised training nor finetuning.



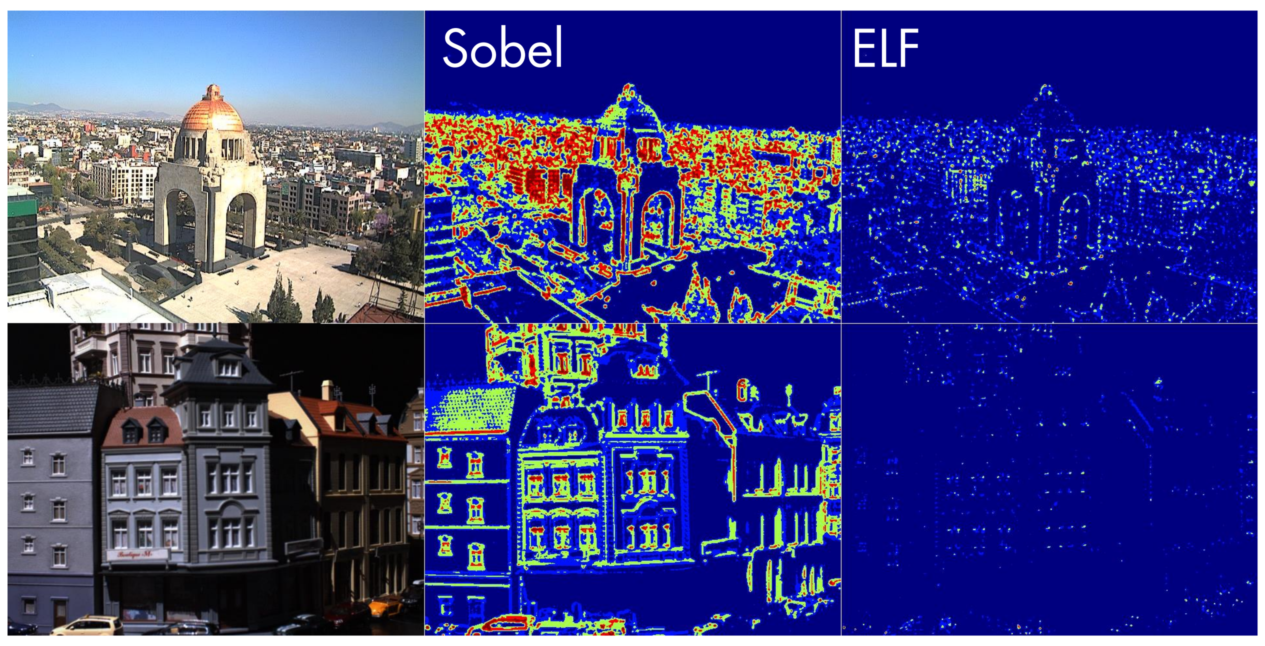
Steps 1-6: Embedded detector.
Steps: 7-8 proxy descriptor.

Method: *Feature detection*
i) Saliency map $S(I) = {}^tF^l(I) \cdot \nabla_I F^l$.
ii) Adaptive threshold (Kapur).
iii) Non-Maxima Suppression (NMS).

Feature description
i) Interpolate the feature map on detected keypoints.



ELF saliency (right) is distinct from the image gradient (middle): the saliency still activates on intensity gradients but only keeps the most informative ones based on their contribution to the CNN feature maps, hence the sparser and more informative signal.



Metrics [4]

1. *Repeatability*: Percentage of keypoints common to both images
2. *Matching Score*: Percentage of keypoints that are nearest neighbours in both image space and descriptor space.

Test datasets

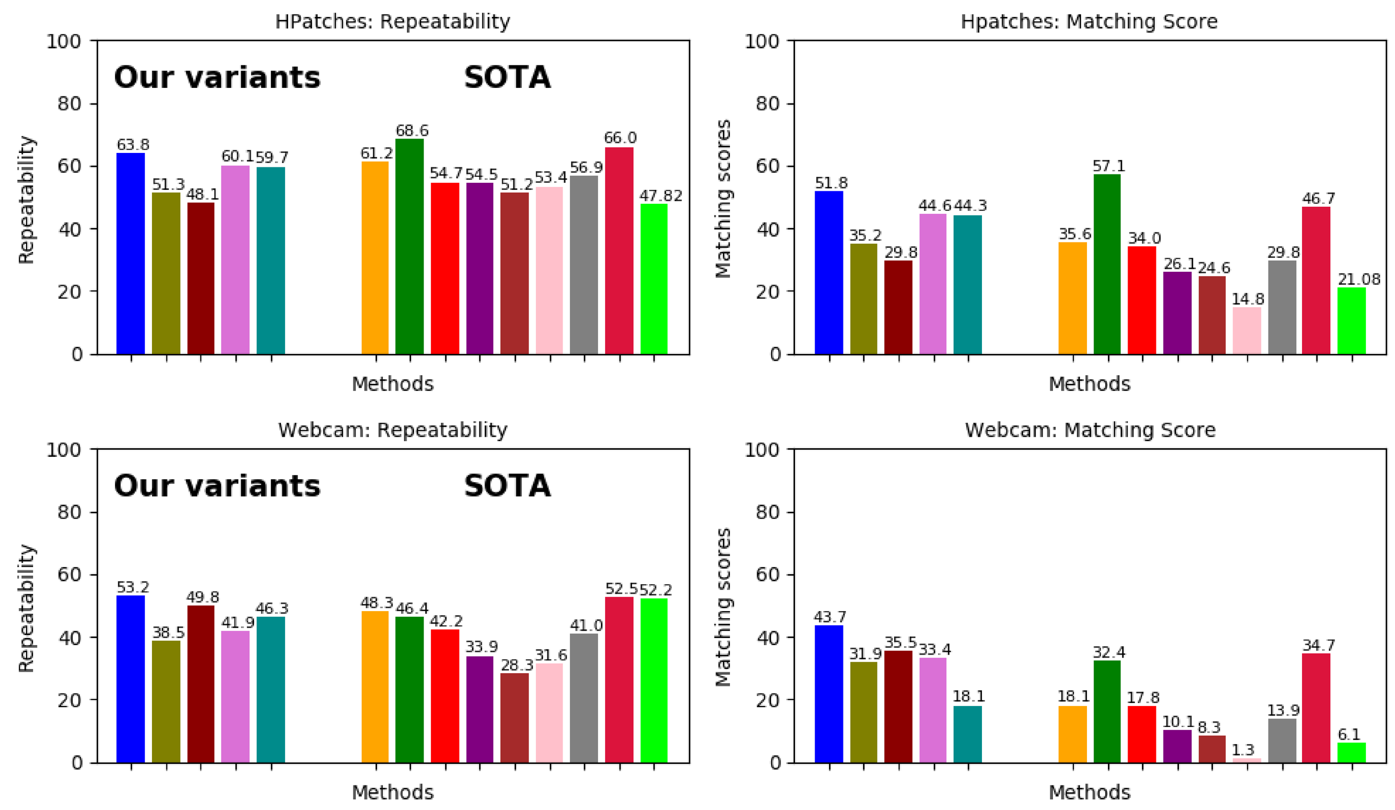
State-of-the-Art

	Detector	Descriptor	Hand-crafted	Learned
ELF	X	X		Semi-supervised
LF-Net	X	X		Supervised
SuperPoint	X	X		Supervised
LIFT	X	X		Supervised
SIFT	X	X	X	
SURF	X	X	X	
ORB	X		X	
KAZE	X	X	X	
TILDE	X			Supervised
MSER	X		X	

Full supervision is the standard training method for recent detector-descriptor. It requires corresponding keypoints generated with either an existing detector or with Structure from Motion.
Our method is **semi-supervised**: the CNN may require full supervision when trained on the standard task but it does not require corresponding keypoints.

LFNet, SURF, MSER, ELF-LFNet, SuperPoint (SP), ORB, ELF-VGG, ELF-SP, Lift, KAZE, ELF-AlexNet, Sobel, SIFT, Tilde, ELF-Xception, Laplacian

Results



General performance

We derive ELF on three classification networks as well as SuperPoint's and LF-Net's descriptor networks. Overall, VGG provides the best variation: we assume that this is because it has the biggest feature space, hence better discriminative properties.

ELF compares with state-of-the-art on HPatches (SuperPoint) and slightly outperforms it on Webcam (TILDE). LIFT and LF-Net curious underperformance may come from a poor data generalisation from their training data.

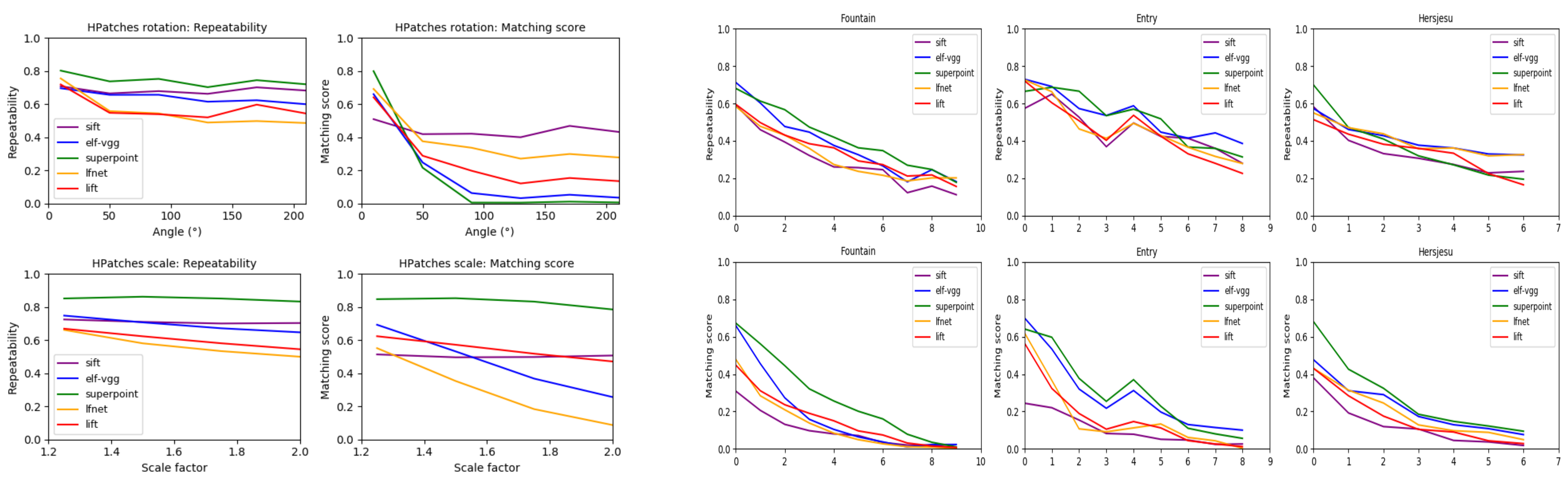
The repeatability variance across methods is low which justify the matching score as a more discriminative metric of the detectors.

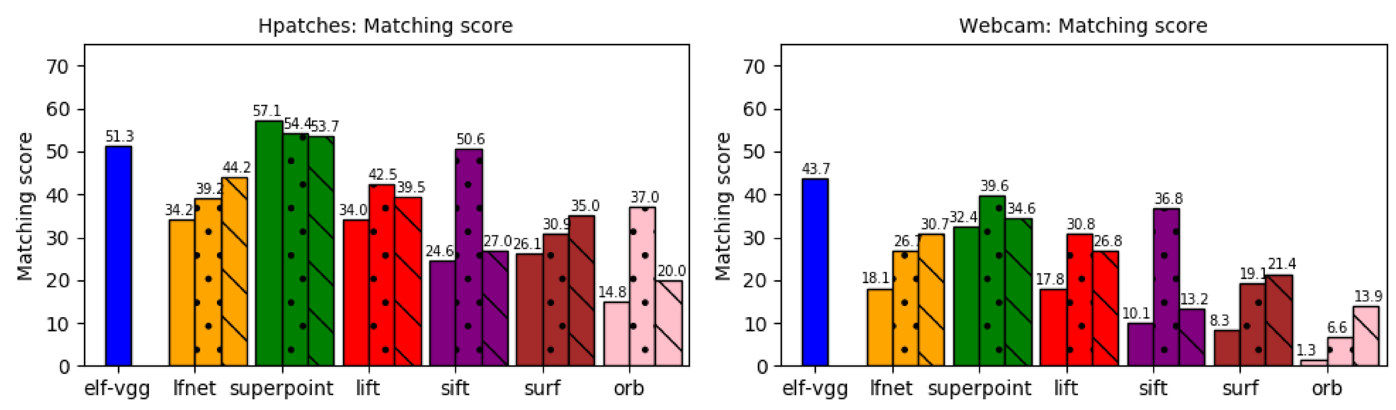
Robustness performance

Scale: Methods that process multiple scale of the same image (e.g. LIFT, LF-Net) can get outperformed by the one that delegate the multi-scale processing to the network (SuperPoint, ELF).

Orientation: All methods without explicit orientation estimation degrade (SuperPoint, ELF).

3D Viewpoint: All methods degrades similarly when the change increases.





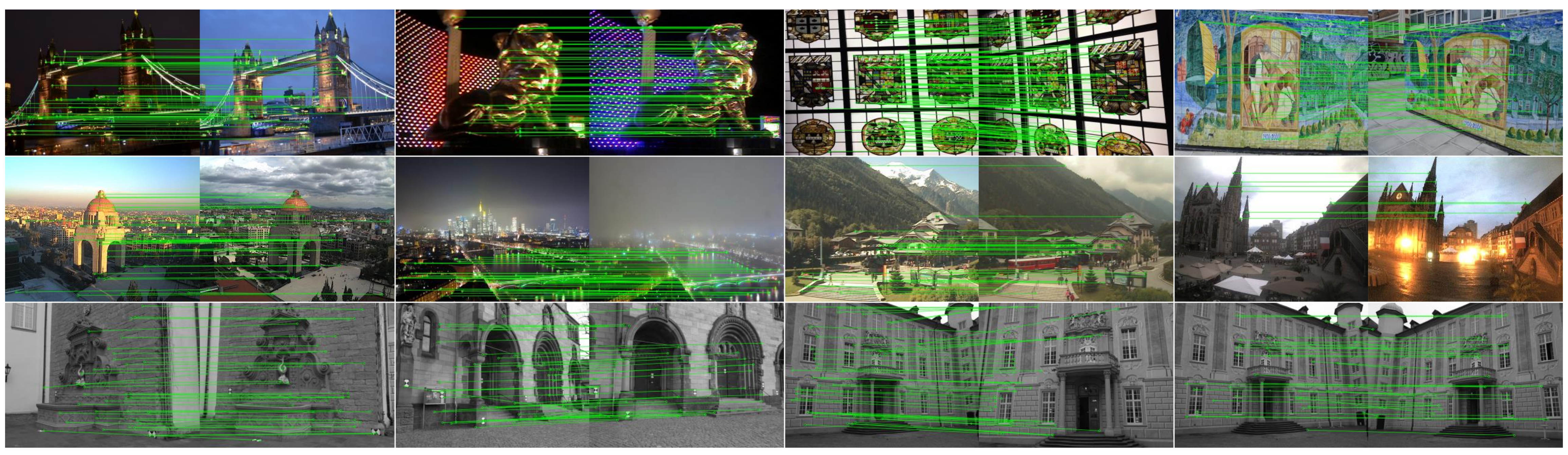
Integration performance

ELF detection (dots): When integrated with other descriptors, ELF boosts the matching score.

Simple description (hashes): Even integrating the interpolated descriptors boosts the performance.

These results show that the feature representation and localisation information learnt by a CNN to complete a task are as relevant as when the CNN is trained specifically for feature matching.

Qualitative results (before RANSAC-based homography estimation)



Bibliography

[1] Y. Ono, E. Trulls, P. Fua, and K.M.Yi. Lf-net: Learning local features from images. NIPS, 2018.
[2] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In CVPR Deep Learning for Visual SLAM Workshop, 2018.
[3] Yi, Kwang Moo, et al. "Lift: Learned invariant feature transform." ECCV. 2016.
[4] Mikolajczyk, Krystian, and Cordelia Schmid. "A performance evaluation of local descriptors." TPAMI 2005
[5] D. G. Lowe. Distinctive image features from scale invariant keypoints. International JCV 2004.
[6] Y. Verdie, K. Yi, P. Fua, and V. Lepetit. Tilde: A temporally invariant learned detector. CVPR 2015
[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb:An efficient alternative to sift or surf. ICCV, 2011.
[8] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. ECCV 2006