Supplementary material for ACCV 2020 paper D2D: Keypoint Extraction with Describe to Detect Approach

Yurun Tian[1], Vassileios Balntas[2], Tony Ng[1], Axel Barroso-Laguna[1], Yiannis Demiris[1], and Krystian Mikolajczyk

[1] Imperial College London
[2] Scape Technologies
{y.tian, tony.ng14, axel.barroso17, y.demiris, k.mikolajczyk}@imperial.ac.uk
vassileios@scape.io

Computation details of $\boldsymbol{S}_{\mathbf{RS}}(x, y)$. We rewrite the equitation here:

$$\boldsymbol{S}_{\mathrm{RS}}(x, y) = \sum_{u} \sum_{v} \boldsymbol{W}(u, v) ||\boldsymbol{F}(x, y) - \boldsymbol{F}(x + u, y + v)||_2, \tag{1}$$

where $\boldsymbol{S}_{\mathrm{RS}}(x, y)$ is computed on a ring region weighted by $\boldsymbol{W}(u, v)$. $\boldsymbol{W}(u, v)$ is a ring with radius $r_{\mathrm{RS}} = 2$ and step 2:

$$\frac{1}{8} \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{2}$$

Impact of K. In Figure 2, we report the mean matching accuracy (MMA) for different number of top K salient keypoints considered for matching.

Comparison with different detection methods. In Figure 1, we compare with three top performing hand-crafted detectors, namely HessianAffine(VLFeat), BRISK(opencv) and AKAZE(opencv), and the hard detection from D2-Net. As shown, the results further validate the advantage of D2D. Worth noting that D2D performs better than the D2-Net hard detection. We argue that, in contrast to our local-global saliency, the magnitude of the feature response may depend on the strength of a gradient pattern rather than point to a location of a discriminative descriptor.

Qualitative results. We visualise matching results in Figure 3, 4, 5, and 6 on four sequences with viewpoint and illumination changes. With the ground truth homography, we consider two keypoints correctly matched if the projected pixel error is smaller than 2. We can see that despite not using a learning based detector, HardNet/SOSNet+D2D provide more matches than other methods. Also note that the training set of HardNet/SOSNet+D2D does not include illumination changes [1], while D2-Net is trained on MegaDepth [2] and SuperPoint is trained on MS COCO [3]. In contrast to SuperPoint and D2-Net, our D2D provides keypoint locations without interpolation. For *Illumination* sequences D2D

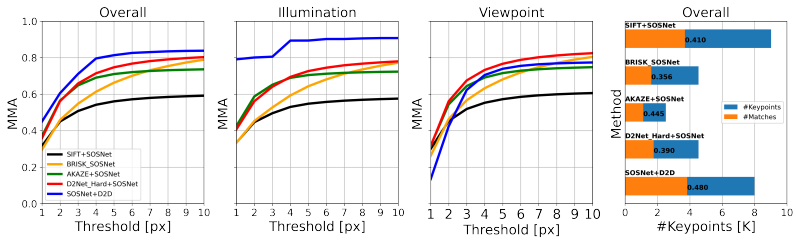


Fig. 1: Comparison with hand crafted methods

detects accurate corresponding point locations between images, resulting in much higher MMA at low pixel thresholds. [4] also noted this observation when compared with DELF [5] (which also excludes interpolations), further supporting our claim that the absence of interpolation is crucial to MMA for *Illumination* sequences at low pixel threshold.

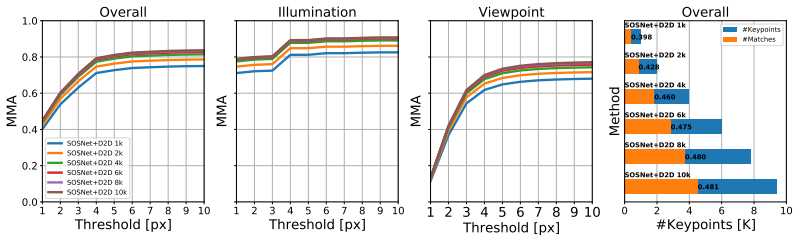
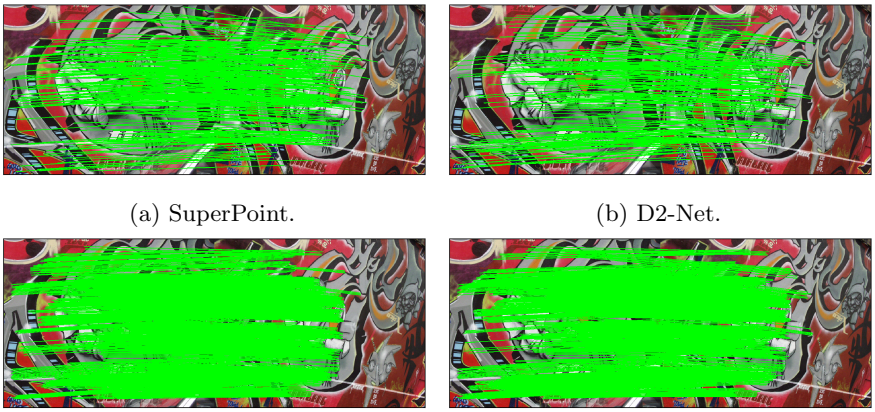


Fig. 2: MMA for top K salient keypoints on Hpatches [6]. The results are reported for #Keypoints$\times 10^3$.

The sequences with viewpoint changes in Hpatches [6] are limited to planar scenes. We therefore further test D2D on more challenging 3D scenes presented in Figure 7 and 8. We show matches that pass the RANSAC process. Consistently high performance of D2D on 2D and 3D scenes indicates that our saliency based approach effectively identifies keypoints with high matching potentials in HardNet and SOSNet feature maps. Consistently high performance of D2D on 2D and 3D scenes indicates that our saliency based approach effectively identifies keypoints with high matching potentials in HardNet and SOSNet feature maps.

References

1. Brown, M., Hua, G., Winder, S.: Discriminative learning of local image descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011) 43–57



(a) SuperPoint. (b) D2-Net.

(c) HardNet+D2D. (d) SOSNet+D2D.

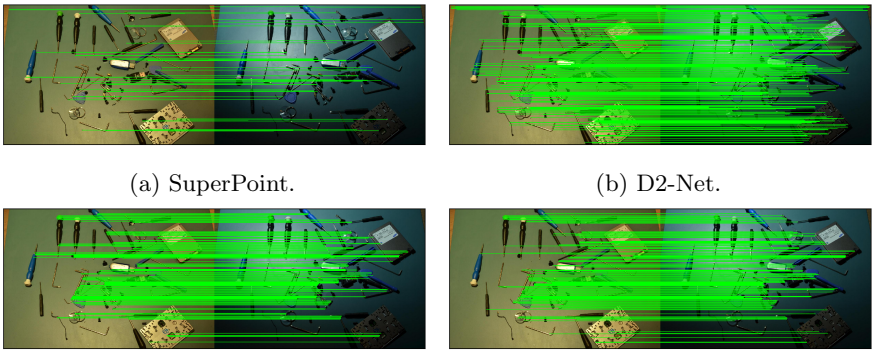
Fig. 3: Matching results on *v_graffiti* sequence from HPatches [6]

2. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2041–2050
3. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV), Springer (2014) 740–755
4. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
5. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 3456–3465
6. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 4. (2017) 6

(a) SuperPoint. (b) D2-Net.

(c) HardNet+D2D. (d) SOSNet+D2D.

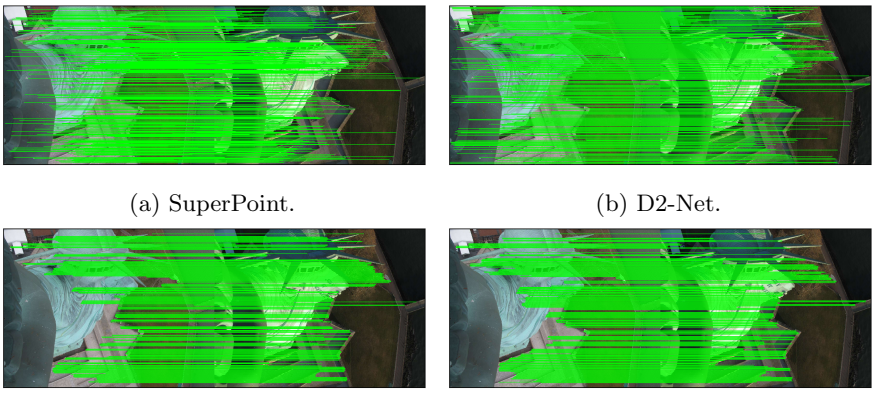
Fig. 4: Matching results on *v_bird* sequence from HPatches [6]



(a) SuperPoint. (b) D2-Net.

(c) HardNet+D2D. (d) SOSNet+D2D.

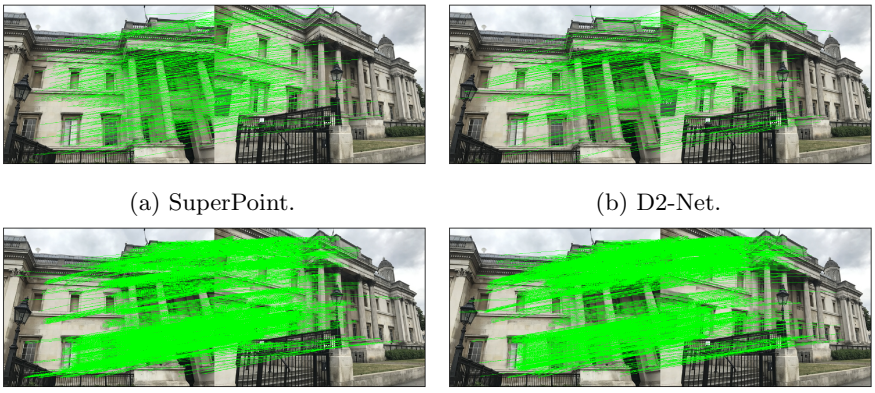
Fig. 5: Matching results on *i_tools* sequence from HPatches [6]



(a) SuperPoint. (b) D2-Net.

(c) HardNet+D2D. (d) SOSNet+D2D.

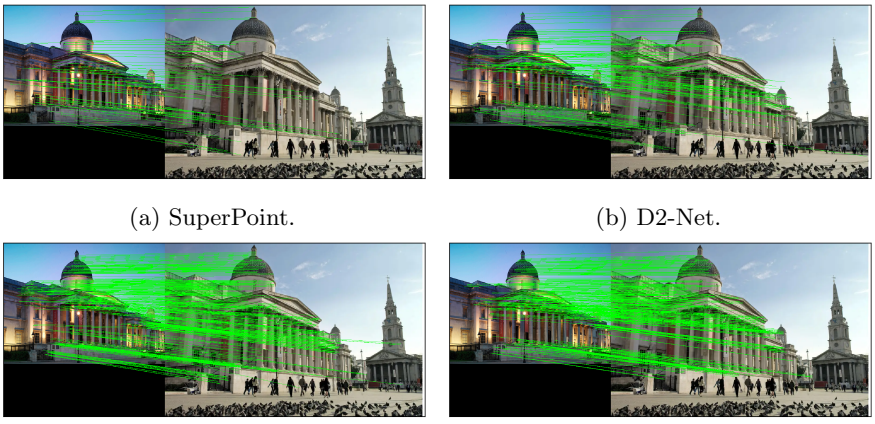
Fig. 6: Matching results on *i_crownday* sequence from HPatches [6]



(a) SuperPoint. (b) D2-Net.

(c) HardNet+D2D. (d) SOSNet+D2D.

Fig. 7: Viewpoint change.



(a) SuperPoint. (b) D2-Net.

(c) HardNet+D2D. (d) SOSNet+D2D.

Fig. 8: Viewpoint and illumination changes.