SOLAR: Second-Order Loss and Attention for Image Retrieval

Supplementary Material

Tony Ng[1], Vassileios Balntas[2], Yurun Tian[1], and Krystian Mikolajczyk[1]

[1] MatchLab, Imperial College London
[2] Facebook Reality Labs
{tony.ng14, y.tian, k.mikolajczyk}@imperial.ac.uk
vassileios@fb.com

1 Results Reported in FPR@95 on UBC Patches

Table 1: **FPR@95** on the UBC dataset. We compare original SOSNet results [4], our re-implemention with data augmentation – SOSNet+ (*reimpl.*) and SOSNET+ with the layer numbers after which SOA are inserted. We performed each experiments three times and report the mean and standard deviation. Note that results from $SOA_{4,5,6}$ are not reported as the network did not converge except when trained on the *liberty* subset.

Train		Liberty		Notredame		Yosemite		
Test	Extra Params	Notredame	Yosemite	Liberty	Yosemite	Liberty	Notredame	Mean
SOSNet	–	1.95	0.58	1.25	1.25	2.84	0.87	1.46
SOSNet+ (*reimpl.*)	–	1.31 ±0.01	0.46 ±0.04	1.21 ±0.06	1.07 ±0.03	2.25 ±0.03	0.80 ± 0.04	1.138 ±0.034
SOSNet+, SOA_3	12,288	1.22 ± 0.04	0.45 ±0.03	1.20 ± 0.03	**0.96** ± 0.05	**2.01** ±0.10	0.73 ±0.04	**1.065** ±0.050
SOSNet+, SOA_4	12,288	1.27 ±0.04	0.44 ±0.03	1.23 ±0.06	0.99 ±0.09	2.08 ±0.01	0.78 ±0.01	1.130 ±0.040
SOSNet+, SOA_5	22,528	1.26 ±0.03	0.44 ±0.02	1.17 ±0.02	0.99 ±0.06	2.06 ±0.04	0.73 ±0.02	1.108 ±0.031
SOSNet+, SOA_6	22,528	1.30 ±0.02	0.56 ±0.05	1.23 ±0.07	1.59 ±0.08	2.80 ±0.05	0.93 ±0.01	1.380 ±0.046
SOSNet+, $SOA_{3,4}$	24,576	**1.21** ±0.02	**0.42** ±0.03	**1.15** ±0.02	1.00 ±0.01	2.07 ±0.01	0.75 ±0.06	1.101 ±0.040
SOSNet+, $SOA_{3,5}$	24,576	1.27 ±0.06	0.45 ±0.05	1.30 ±0.02	**0.96** ±0.03	2.19 ±0.02	0.82 ±0.01	1.139 ±0.030
SOSNet+, $SOA_{4,5}$	34,816	1.22 ±0.03	0.48 ±0.01	1.29 ±0.01	0.97 ±0.01	2.22 ±0.04	0.75 ±0.04	1.130 ±0.021
SOSNet+, $SOA_{4,6}$	34,816	1.39 ±0.02	0.59 ±0.02	1.59 ±0.19	1.32 ±0.03	2.72 ±0.18	0.89 ±0.03	1.416 ±0.077
SOSNet+, $SOA_{3,4,5}$	47,104	1.32 ±0.03	0.46 ±0.02	1.36 ±0.04	1.02 ±0.10	2.10 ±0.06	**0.71** ±0.02	1.147 ±0.047

The results reported in **FPR@95** on UBC-Patches [5] is shown in Table 1. We present results on each of the six test runs with various configurations of SOA insertions. We did not perform experiments involving SOA_1 and SOA_2 as explained in Section 5 in the paper. The layers after which SOAs are inserted are based on the L2-Net architecture in Table 2. We performed experiments on SOA insertion of one to three blocks from between Layers-3 to 7, giving the set of results {SOA_3, SOA_4, SOA_5, SOA_6, $SOA_{3,4}$, $SOA_{3,5}$, $SOA_{4,5}$, $SOA_{4,6}$, $SOA_{3,4,5}$, $SOA_{4,5,6}$}. To resolve potential noise, we follow the practice by Mukundan *et al.* [2] in performing three separate runs for each experiment and reporting the mean value and standard deviation.

Comparing the results of SOSNet with various SOAs inserted in Table 1, we can see that in general the SOA blocks increase the results slightly with few extra parameters. Agreeing with HPatches results from Fig. 6 in the paper, configurations with SOA_6 inserted perform noticeably worse when compared to the baseline. We suspect this also due to the same reason of optimisation constraints for low-resolutions at very higher-level feature maps, as discussed in Section 5.3 in the paper. By comparing $\text{SOA}_{3,4}$ with $\text{SOA}_{3,5}$ and $\text{SOA}_{4,5}$ with $\text{SOA}_{4,6}$, we observe that SOAs inserted at consecutive feature levels performs noticeably better. One potential explanation would be the immediate sharing of information across consecutive feature maps, allowing for better gradients into the SOA blocks to optimise for feature re-weighting. This also agrees with the improved performance of $\text{SOA}_{4,5}$ over single SOA block insertion for ReseNet101 in Section 5.2 of the paper, and HPatches results in the paper.

2 L2-Net Architecture

Table 2: L2-Net [3] architecture. Note that we only show the convolutional kernel's parameters and intermediate feature map dimension to assist discussion of non-local block insertions. Refer to Tian *et al.* [3] for complete details of the architecture including normalisation and activation layers, and different variations of the model.

Layer	Kernel	Stride	Output shape $[h, w, c]$	Cumulative # Params.
1	3×3	1	32, 32, 32	288
2	3×3	1	32, 32, 32	9,216
3	3×3	2	16, 16, 64	18,432
4	3×3	1	16, 16, 64	36,864
5	3×3	2	8, 8, 128	73,728
6	3×3	1	8, 8, 128	147,756
7	8×8	1	1, 1, 128	285,984

Table 2 shows the L2-Net [3] architecture, which is used by SOSNet [4] and the ablation study from Section 5.3 in the paper. In our implementation of SOSNet and subsequent SOSNet+, SOAs experiments, the patch first passes through an InstanceNorm layer, then each convolution layer is followed by BatchNorm and ReLU (except for after Layer-7 which has no ReLU). Lastly, ℓ_2-norm is applied to the final 128-dimensional descriptor after Layer-7. During training, dropout of rate 0.1 is added between Layer-6 and Layer-7 to prevent over-fitting.

3 Second-order attention maps on patches

Fig. 1 on the next page visualises the second-order attention maps (similar to Figure 4 in the paper) on two example patch correspondences from HPatches [1].

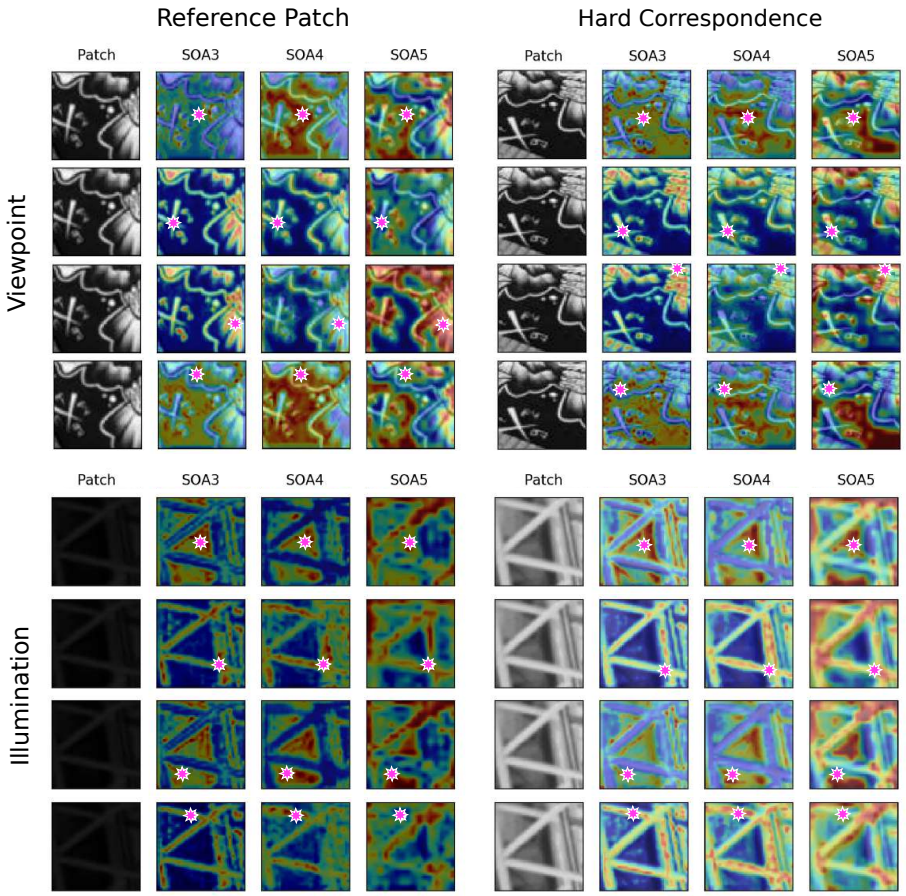

Fig. 1: Second-order attention maps for HPatches [1]. **Left:** reference patch. **Right:** hard correspondence. **Top:** viewpoint changes to reference patch. **Bottom:** illumination changed to reference patch. For each case, we select four pixel locations (pink star) to display the attention maps of SOSNet [4]+, $SOA_{3,4,5}$. which has the best results in HPatches evaluation.

We show two example reference patches and each a 'hard' corresponding patch from a sequence with viewpoint (top) and illumination changes (bottom). Firstly we observe that in contrast to large images, the second-order attention at a given spatial location focuses on similar / connected structures within the patch. This is due to much less semantic (and colour) information and lack of distinctive textures in patches compared to large images. Secondly we also observe that the attention maps are invariant to both viewpoint and illumination changes. As we compare the reference patch to the hard correspondence, the attentions between are consistent across all three levels in $SOA_{3,4,5}$.

References

1. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: CVPR (2017)
2. Mukundan, A., Tolias, G., Chum, O.: Explicit spatial encoding for deep local descriptors. In: CVPR (2019)
3. Tian, Y., Fan, B., Wu, F.: L2-Net: Deep learning of discriminative patch descriptor in euclidean space. In: CVPR (2017)
4. Tian, Y., Yu, X., Fan, B., Fuchao, W., Heijnen, H., Balntas, V.: SOSNet: Second order similarity regularization for local descriptor learning. In: CVPR (2019)
5. Winder, S.A., Brown, M.: Learning local image descriptors. In: CVPR (2007)