ATTENTION GUIDED INVARIANCE SELECTION FOR LOCAL FEATURE DESCRIPTORS

Jiapeng Li[1], *Ge Li*[1], *Thomas H Li*[*2]

[1]School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School
[2]Advanced Institute of Information Technology, Peking University

ABSTRACT

To copy with the extreme variations of illumination and rotation in the real world, popular descriptors have captured more invariance recently, but more invariance makes descriptors less informative. So this paper designs a unique attention guided framework (named AISLFD) to select appropriate invariance for local feature descriptors, which boosts the performance of descriptors even in the scenes with extreme changes. Specifically, we first explore an efficient multi-scale feature extraction module that provides our local descriptors with more useful information. Besides, we propose a novel parallel self-attention module to get meta descriptors with the global receptive field, which guides the invariance selection more correctly. Compared with state-of-the-art methods, our method achieves competitive performance through sufficient experiments.

Index Terms— Local feature descriptors, Invariance, Self-attention

1. INTRODUCTION

Local feature detectors and descriptors are one of the essential computer vision problems in various applications[1, 2, 3]: Structure-from Motion, Visual Simultaneous Localization and Mapping, 3D reconstruction, Virtual and Augmented Reality. Given two images, classical methods[4, 5, 6]are based on a two-phase pipeline that first detects interest points from each image, then extracts local descriptors around the points.

In the real world exist large changes of light, rotation and so on, which limits the descriptors. Thus, handcraft descriptors capture as much invariance as possible, e.g. SIFT[4], ORB[6], similarly for learning-based descriptors[3, 7, 8, 9, 10]. LIFT[7] estimates keypoints and their orientation to get rotation invariance. Data augmentations enable Superpoint[8] and [9] to boost robustness to light and viewpoint changes.

However, more invariance makes descriptors less informative [11]. To keep the trade-off between them, BOLD[12] selects invariance for binary descriptors. For each image patch, the method chooses a subset of the binary tests to maximize the invariance for affine transformations. Besides, LISRD[11] selects invariance for local descriptors. The paper uses meta descriptors extracted by the NetVLAD[13] layers to weight the distance between the local descriptors. However, this method is limited by the very coarse spatial resolution of meta descriptors.

To keep the balance between invariance and information better, we design a unique attention guided framework to select appropriate invariance for local feature descriptors. Firstly, inspired by [14, 15], we explore an efficient multi-scale feature extraction module (called MSFE) that extracts local descriptors with several variance properties and multi-scale features. So each descriptor corresponds to a different kind of invariance. Furthermore, the multi-scale local patterns provide descriptors with more useful perception information than a single scale used in previous methods. After that, motivated by[16, 17], we design a novel parallel self-attention module (called PSA) to get meta descriptors with the global receptive field. For an image of size H×W×3, with the benefit of the self-attention layer, our meta descriptors have a finer spatial resolution(1/8 H, 1/8 W) than [11]. They get global context and guide the invariance selection more correctly. Finally, following LISRD[11], we use the similarities of the meta descriptors to weight the local descriptors distances between two keypoints (see Section 2.3).

To sum up, our major contributions are as follows:

- We explore an efficient multi-scale feature extraction module that extracts features across multiple scales for local descriptors.
- We come up with a novel parallel self-attention module to get meta descriptors with the global receptive field, which is able to guide the invariance selection better.
- Our method achieves competitive results, compared with SOTA methods.

2. METHOD

Fig. 1(a) shows our framework of AISLFD with two stages. At the first stage, the model learns multiple local descriptors. Secondly, it extracts meta descriptors which determine the suitable invariance when matching the local descriptors.

*Corresponding author: tli@aiit.org.cn

This project was supported by Hangzhou Science and Technology Development Program (No.20182014B09) and Key-Area Research and Development Program of Guangdong Province (No.2019B121204008).

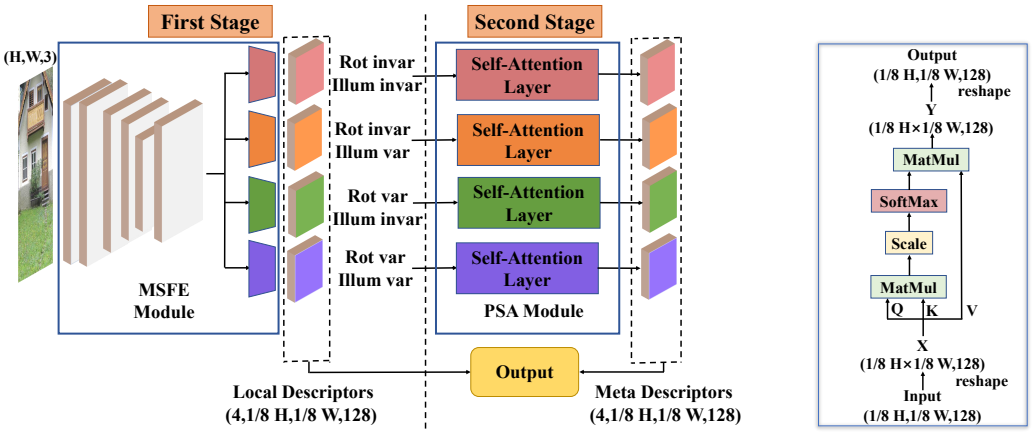

Fig. 1: (a) The framework of AISLFD. It includes the extraction for multiple local feature descriptors and the meta feature descriptors. (b) Architecture of the self-attention layer.

2.1. Local Descriptors with Different Invariances

In this paper, we only pay attention to rotation and illumination factors which heavily influence the performance of descriptors. Each factor can be invariant or variant, so we must produce four kinds of descriptors.

First, we design an efficient multi-scale feature extraction module (called MSFE, see Fig. 2) made up of MS module and 4-head Module. The MS module has a bottom-up pathway and a top-down pathway to get multiple scale features. The former computes a feature hierarchy while the latter unsamples higher pyramid levels. Different from Swin Transformer[18], we use convolutions to get multi-scale features rather than shifted windows based self-attention. So our method maintains more sufficient translation invariance than [18], which is important for invariance selection. Then the 4-head module predicts semi-dense local feature descriptors corresponding to diverse invariances individually.

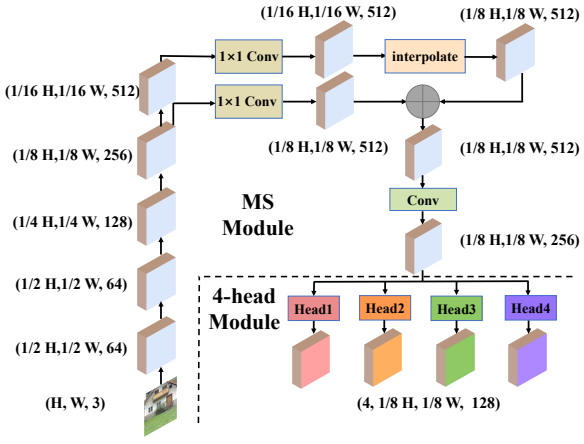

Fig. 2: Architecture of the MSFE module.

2.2. Meta Descriptors with the Global Receptive Field

Meta Descriptors need more context than local descriptors to guide the invariance selection. So we design a novel parallel self-attention module (called PSA) made up of 4 self-attention layers. Each self-attention layer (see Fig. 1(b)) provides meta descriptors with the global receptive field and semi-dense spatial resolution.

Following [11], we regard the local descriptors as input to extract meta ones. The attention layer first converts the input to matrix X which contains 1/8 H×1/8 W vectors of dimension 128. Inspired by[16], we adopt the scale dot-product attention to model global-range dependencies. The queries matrix Q, keys matrix K, and values matrix V are copies of X, regarded as the input of scale dot-product attention. Then the relation between each query vector in Q and each key vector in K will be established to weight the value vector paired with the key vector. Lastly, we reshape matrix Y to get the final output feature map as a meta descriptor.

Each meta descriptor, which has a finer spatial resolution(1/8 H, 1/8 W) than [11], makes the model more robust. The reason is our meta area has fewer matches than [11]. It means when our meta descriptor cannot get the right invariance in one meta area, it will affect fewer matches than [11].

2.3. Invariance Selection for Descriptors

Given two images $\mathbf{I}^q$ and $\mathbf{I}^k$, for each meta region, the match process is shown in Fig. 3.

First, the meta descriptors will compute four similarities and apply a softmax operation to get the weights corresponding to combinations of invariance. Then we use the weights to compute the final descriptor distance between two keypoints.

Formally, the keypoint $\mathbf{p}^q$ matches 4 local descriptors $\mathbf{l}^q$ and 4 meta ones $\mathbf{m}^q$ in image $\mathbf{I}^q$, similarly for the keypoint

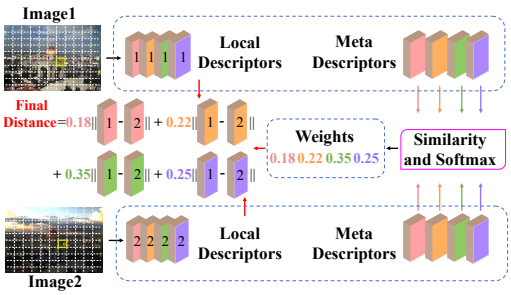

Fig. 3: The match process between two keypoints of images.

$\mathbf{p}^k$ in image $\mathbf{I}^k$. The final descriptor distance between the two keypoints is

$$\mathrm{dis} = \sum_{s=1}^{4} \frac{\exp\left(\left(\mathbf{m}_s^q\right)^\top \cdot \mathbf{m}_s^k\right)}{\sum_{t=1}^{4} \exp\left(\left(\mathbf{m}_t^q\right)^\top \cdot \mathbf{m}_t^k\right)} \left\|\mathbf{l}_s^q - \mathbf{l}_s^k\right\|_2 . \quad (1)$$

2.4. Training Loss

The total loss includes the loss of local and meta descriptors: $\mathcal{L} = \mathcal{L}_l + \mathcal{L}_m$. Following [3, 11], we utilize the triplet margin loss to compute the loss of the local and meta descriptors.

Given two images $\mathbf{I}^q$ and $\mathbf{I}^k$ connected by a homography $\mathcal{H}$, we warp every keypoint of images $\mathbf{I}^q$ to $\mathbf{I}^k$ by the $\mathcal{H}$, aiming at producing n correspondences between the two images. The right correspondence $\left(\mathbf{p}_x^q, \mathbf{p}_x^k\right)$ has a posititive distance $pd_x = \mathbf{dis}\left(\mathbf{p}_x^q, \mathbf{p}_x^k\right)$ in descriptor space. Besides, a negative distance exists between the wrong match points:

$$\mathrm{nd}_x = min\left(\mathbf{dis}\left(\mathbf{p}_x^q, \mathbf{p}_N^k\right), \mathbf{dis}\left(\mathbf{p}_N^q, \mathbf{p}_x^k\right)\right), \quad (2)$$

where $\mathbf{p}_N^k = \arg\min_{y\in[1,n]}\left(\mathbf{dis}\left(\mathbf{p}_x^q, \mathbf{p}_y^k\right)\right)$ s.t. $\left\|\mathbf{p}_x^q - \mathbf{p}_y^k\right\|_2 > T$, T is a threshold distance. And $\mathbf{p}_N^q$ is similar. Then a margin M is used to define the triplet margin loss:

$$\mathcal{L}_T\left(\mathbf{I^q}, \mathbf{I^k}, \mathbf{dis}\right) = \frac{1}{n}\sum_{x=1}^{n} \max\left((pd_x)^2 - (nd_x)^2 + M, 0\right). \quad (3)$$

So when we use the L2 distance for local descriptor $\mathbf{l}^a$ and $\mathbf{l}^i$, the loss of invariant descriptors $\mathcal{L}_i$ between the anchor image $\mathbf{I}^a$ and the invariant image $\mathbf{I}^i$ is defined as

$$\mathcal{L}_i = \mathcal{L}_T\left(\mathbf{I^a}, \mathbf{I^i}, \left\|\mathbf{l}^a - \mathbf{l}^i\right\|_2\right). \quad (4)$$

The loss of variant descriptors $\mathcal{L}_v$ makes variant descriptors of the anchor image differ from those of the invariant image while remaining similar to those of the variant image.

$$\mathcal{L}_v = \frac{1}{n}\sum_{x=1}^{n} \max\left(\left\|\mathbf{l}_x^a - \mathbf{l}_x^v\right\|_2^2 - \left\|\mathbf{l}_x^a - \mathbf{l}_x^i\right\|_2^2 + fM, 0\right). \quad (5)$$

Following [11], we use the same factor f to set the anchor images apart from the invariant images. Finally, the total loss of local descriptors is defined as:

$$\mathcal{L}_l = \frac{1}{|L|}\sum_{l\in L} \mathcal{L}_{i/v}(l), \quad (6)$$

where L is a set of local descriptors from images. If the descriptor is invariant to all the changes between $\mathbf{I}^a$ and $\mathbf{I}^i$, $\mathcal{L}_i$ is used. Otherwise, we use $\mathcal{L}_v$. And following[11], we adopt the same meta descriptor loss $\mathcal{L}_m$:

$$\mathcal{L}_m = \mathcal{L}_T\left(\mathbf{I^a}, \mathbf{I^i}, \mathbf{dis}\right). \quad (7)$$

3. EXPERIMENTAL RESULTS

3.1. Datasets and Metrics

Following the new version of LISRD[19], we train our model on the MS COCO dataset[20], Multi-Illumination Images in the Wild[21], and the Virtual Image Dataset for Illumination Transfer (VIDIT)[22], keeping the same settings as LISRD[19] except the learning rate of the Adam solver.

The benchmark dataset HPatches[23] and RDNIM[11] are used to test our descriptors. The former dataset is widely used but contains just a few rotations and medium illumination changes. The latter dataset contains different levels of rotation and day-night changes.

Following the new version of LISRD[19], we use the SuperPoint[8] to compute keypoints and only compare the descriptors' performance. Meanwhile, we also use Homography estimation, Precision (the proportion of correct matches across the total predicted matches) and Recall (the percentage of correctly matches predicted over the whole ground-truth matches) as our metrics.

3.2. Ablation Study

3.2.1. Ablation on the PSA module

To confirm the effectiveness of our parallel self-attention (PSA) module, we replace the NetVLAD layers of LISRD with our PSA module (called Ours-A method).

Seen from Table 1, Ours-A ranks first in HPatches view part and is close to the reference LISRD in illumination part. Table 2 also shows Ours-A ranks first in all benchmarks of RDNIM dataset. The results confirm that our parallel self-attention module makes the meta descriptors guide the invariance selection more correctly especially on the scenes of rotation variance and a wider range of illumination changes (e.g. in RDNIM).

Table 1: Comparison to the SOTA methods on HPatches.

HPatches		SOSNet	SuperPoint	D2-Ne	R2D2	GIFT	LISRD	Ours-A	Ours-B	Ours-C
Illum	HEstimation	0.933	0.912	0.905	0.937	0.944	0.947	0.933	**0.951**	**0.951**
	Precision	0.748	0.710	0.725	**0.771**	0.701	0.765	0.762	0.764	0.760
	Recall	0.821	0.811	0.775	0.814	0.681	0.920	0.918	**0.944**	0.936
View	HEstimation	**0.698**	0.671	0.617	0.620	**0.698**	0.688	**0.698**	0.654	0.675
	Precision	0.727	0.685	0.666	0.666	0.687	0.731	**0.733**	0.693	0.720
	Recall	0.760	0.750	0.664	0.639	0.660	0.757	**0.766**	0.700	0.745

Table 2: Comparison to the SOTA methods on RDNIM.

RDNIM		SOSNet	SuperPoint	D2-Net	R2D2	GIFT	LISRD	Ours-A	Ours-B	Ours-C
Day Ref	HEstimation	0.226	0.178	0.124	0.190	0.225	0.318	**0.365**	0.336	0.343
	Precision	0.218	0.191	0.196	0.173	0.155	0.406	**0.437**	0.431	0.425
	Recall	0.226	0.214	0.145	0.180	0.149	0.439	**0.541**	**0.541**	0.524
Night Ref	HEstimation	0.252	0.235	0.195	0.229	0.294	0.391	**0.448**	0.445	0.441
	Precision	0.288	0.259	0.265	0.237	0.240	0.487	**0.545**	0.534	0.536
	Recall	0.296	0.296	0.218	0.237	0.229	0.520	**0.634**	0.629	0.628

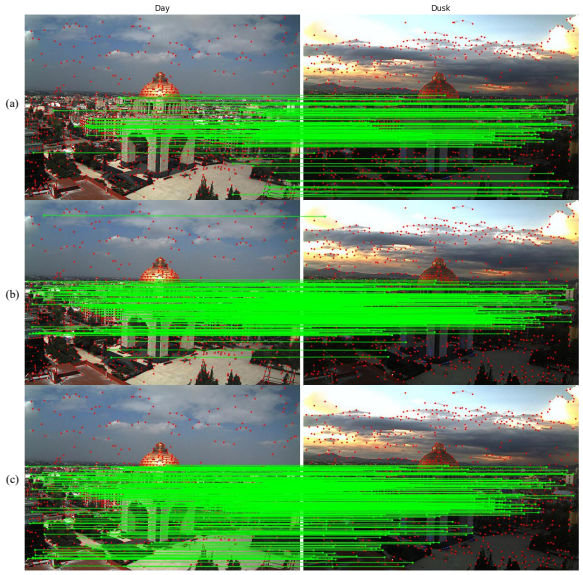

Fig. 4: The visual results of (a) LISRD, (b) Ours-B, (c) Ours-C. They show that our MSFE module improves the ability of our model to copy with the light change.

3.2.2. Ablation on the MSFE module

Aiming to indentify the effectiveness of our multi-scale feature extraction (MSFE) module, we replace the VGG-like backbone of LISRD with our MSFE module (called Ours-B).

Table 2 indicates Ours-B has better results than LISRD. Besides, Table 1 shows Ours-B almost ranks first in HPatches illumination part but indeed places lower than LISRD on HPatches view part. We explain the phenomenon as follows:

[24] shows convolutional neural networks (CNNs) are brittle when the input is translated using a small image transformation. Besides, the deeper network loses its translation invariance more easily [24]. Our MSFE module in Ours-B is deeper than LISRD's backbone (are both CNN). So Ours-B will lose more invariance on small changes, which induces its weaker results than LISRD on HPatches view part (only tiny translation and rotation). But in RDNIM (big transformation), Ours-B gets better than LISRD. And comparing Fig. 4(a) with Fig. 4(b), we can find Ours-B produces more matches than LISRD.

3.3. Comparison to State-of-the-Art Approaches

We compare our complete model (called Ours-C) with a number of state-of-the-art methods: SOSNet[25], SuperPoint[8], D2-Net[3], R2D2[9], GIFT[10], LISRD[11]. Table 1 shows Ours-C ranks the top in HPatches illumination part but has weaker results on view part. It seems that Ours-C obtains a balance between Ours-A and Ours-B. Compared with the SOTA methods, our complete model shows competitive performance in Table 2. The visualization result of Fig. 4(c) also reflects the effectiveness of Ours-C. To sum up, Ours-C, which is more robust to face various scenes, absorbs the characteristics of Ours-A and Ours-B.

4. CONCLUSION AND FUTURE WORK

We create an innovative attention guided deep framework to adopt the suitable invariance for local feature descriptors, which improves the performance of descriptors. To guide the invariance selection more correctly, we adopt a novel parallel self-attention module to get meta descriptors with the global receptive field. Besides, we explore an efficient MSFE module that extracts multi-scale features. The extensive ablation study identifies the validity of our parallel self-attention module and the MSFE module. Sufficient experiments show that our method has competitive performance. The future work will combine better the advantages of the self-attention module and the MSFE module to face more complex scenes.

5. REFERENCES

[1] Christopher B. Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Krishna Chandraker, "Universal correspondence network," in *Neural Information Processing Systems*, 2016, pp. 2406–2414.

[2] Johannes L. Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler, "Semantic visual localization," in *CVPR*, 2018, pp. 6896–6906.

[3] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler, "D2-net: A trainable CNN for joint description and detection of local features," in *CVPR*, 2019, pp. 8092–8101.

[4] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "SURF: speeded up robust features," in *ECCV*, 2006, vol. 3951, pp. 404–417.

[6] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision*, 2011, pp. 2564–2571.

[7] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua, "LIFT: learned invariant feature transform," in *ECCV*, 2016, vol. 9910, pp. 467–483.

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *CVPR*, 2018, pp. 224–236.

[9] Jérôme Revaud, César Roberto de Souza, Martin Humenberger, and Philippe Weinzaepfel, "R2D2: reliable and repeatable detector and descriptor," in *NeurIPS*, 2019, pp. 12405–12415.

[10] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou, "GIFT: learning transformation-invariant dense visual descriptors via group cnns," in *NeurIPS*, 2019, pp. 6990–7001.

[11] Rémi Pautrat, Viktor Larsson, Martin R. Oswald, and Marc Pollefeys, "Online invariance selection for local feature descriptors," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[12] V. Balntas, L. Tang, and K. Mikolajczyk, "Bold - binary online learned descriptor for efficient image matching," in *CVPR*, 2015.

[13] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic, "Netvlad: CNN architecture for weakly supervised place recognition," in *CVPR*, 2016.

[14] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2016.

[15] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou, "Loftr: Detector-free local feature matching with transformers," in *CVPR*, 2021, pp. 8922–8931.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Neural Information Processing Systems*, 2017.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and et al, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *ICCV*, 2021.

[19] Rémi Pautrat, Viktor Larsson, Martin R. Oswald, and Marc Pollefeys, "Lisrd," `https://github.com/rpautrat/LISRD`, 2020.

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, and et al, "Microsoft COCO: common objects in context," in *ECCV*, 2014, pp. 740–755.

[21] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand, "A multi-illumination dataset of indoor object appearance," in *ICCV*, 2019.

[22] Majed El Helou, Ruofan Zhou, Johan Barthas, and Sabine Süsstrunk, "VIDIT: Virtual image dataset for illumination transfer," *arXiv preprint arXiv:2005.05460*, 2020.

[23] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *CVPR*, 2017, pp. 3852–3861.

[24] Aharon Azulay and Yair Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?," *arXiv preprint arXiv:1805.12177*, 2018.

[25] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas, "Sosnet: Second order similarity regularization for local descriptor learning," in *CVPR*, 2019, pp. 11016–11025.