# Self-Supervised Equivariant Learning for Oriented Keypoint Detection
## *- Suppelementary material -*

In this supplementary material, we explain the reason for the periodic results under synthetic rotations, the effect of the number of keypoints in IMC2021 [7], and the details of the outlier filtering algorithm in section 1. We show additional results on the Extreme Rotation dataset [8], the ablation studies, and the separated results of the HPatches viewpoint/illumination in section 2. We compare the qualitative results of the predicted matches and orientation estimation in section 3.

## 1. Additional analysis

In section 1.1, we explain the performance variation cycles in Figs.4-5 of the main paper. In section 1.2, we explain why the performance of IMC2021 [7] largely drops from 2,000 points to 8,000 points. In section 1.3, we explain the detailed description of the outlier filtering algorithm.

### 1.1. Performance variation cycles in Figs.4-5

The periodic patterns in Figs.4-5 of the main paper are caused by input variations due to the grid structure of pixels and the square shape of convolution filters. (1) Since an image is a grid structure of pixels, a rotation of the image induces an interpolation artifact for the corresponding position, being minimal for a multiple of 90°and maximal in between. Fig. 1 plots the average errors from the original pixel values, which clearly show the same cycle. (2) Since convolution filters take a square grid of pixels as input, a rotation of the image makes the filters take a different set of pixels, being the same set again for a multiple of 90°. Therefore, compared to the reference image, the rotated input to the model varies most at 45°, 135°, 225°, 315°rotations, which induces the degrading cycle. The similar pattern can also be found in Fig.7 of ORB [14].

### 1.2. The effect of the number of keypoints in IMC2021 [7]

Fig.13 and Sec.5.4 in [7] show that the pose estimation accuracy increases until the number of keypoints reaches 8,000 and converges, so [7] adopt the 2,048 and 8,000 numbers of keypoints as standard evaluation protocols. A scene of IMC2021 consists of the exhaustive pairs of 100 images, and the accuracy increases at 8,000 keypoints than 2,048 keypoints as a keypoint in one image is likely to exist in the other images.

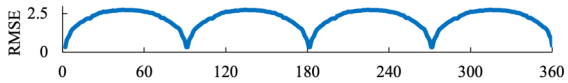


Figure 1. RMSE of the corresponding pixel values in rotating image.

### 1.3. Detailed descriptions of the outlier filtering

To show the effectiveness of the estimated orientations in Table 3 of the main paper, we use an outlier filtering algorithm. We filter the outlier matches through the global consensus of the orientation values assigned in keypoints of the tentative matches. We compute the orientation difference of two keypoints for each tentative match and then select the most frequent difference from all those tentative matches. This most frequent orientation difference is used to define outlier matches by measuring how large each tentative match deviates from it. Let $\mathbf{m} \in \mathbb{N}^{K\times 2}$ is a set of the tentative matches about the pair of keypoint indices, which is obtained using the mutual nearest neighbour matcher. The inlierness $p$ is defined for a tentative match of two keypoints with orientations $o^{\mathrm{a}}$ and $o^{\mathrm{b}}$:

$$p(o^{\mathrm{a}}, o^{\mathrm{b}}, t) = \begin{cases} 1, & \text{if } |\texttt{mode}(\vec{d}) - d| \leq t, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$
$$\text{and } d = (o^{\mathrm{b}} - o^{\mathrm{a}} + 360) \bmod 360,$$

where $\vec{d} \in \mathbb{R}^K$ is a vector of the orientation differences, `mode` function returns the most frequent value on the input vector, $t$ is a threshold to accept how far from the frequent orientation difference, and $K$ is the number of tentative matches. We use the outlier threshold $t = 30°$ for Table 3 in the main paper. Note that $o^{\mathrm{a}}$ and $o^{\mathrm{b}}$ denote two orientation values of a tentative match. We obtain the orientation vector $\vec{o} \in \mathbb{R}^K$ of our keypoints as follows:

$$\vec{o}^{\mathrm{a}} = \arg\max_g \delta(\mathbf{O}^{\mathrm{a}}; \mathbf{m}_{:,0})_g, \vec{o}^{\mathrm{b}} = \arg\max_g \delta(\mathbf{O}^{\mathrm{b}}; \mathbf{m}_{:,1})_g, \quad (2)$$

where $\mathbf{O} \in \mathbb{R}^{|G|\times H\times W}$ is the rotation-equivariant orientation tensor, $\delta : \mathbb{R}^{|G|\times H\times W} \rightarrow \mathbb{N}^{|G|\times K}$ selects the orientation values from the keypoint coordinates using the keypoint indices in tentative matches $\mathbf{m}_{:,i}$, and $g \in G$.

## 2. Additional results

In section 2.1, we demonstrate the results of keypoint matching on the Extreme Rotation (ER) benchmark [8].

| Det. | Des. | PCK@5 | PCK@2 | PCK@1 |
|---|---|---|---|---|
| SuperPoint [5] | SuperPoint [5] | 0.255 | 0.194 | 0.112 |
| SuperPoint [5] | GIFT [8] | 0.435 | 0.328 | 0.186 |
| ours | GIFT [8] | **0.476** | **0.353** | **0.212** |

Table 1. PCK on the Extreme Rotation dataset [8]. The numbers next to the PCK represent the pixel threshold to measure the correctness of the correspondence.

| *Loss.* | rep. | w/o out. filter. MMA @3px | w/o out. filter. MMA @5px | w/o out. filter. match. | out. filter. MMA @3px | out. filter. MMA @5px | out. filter. match. |
|---|---|---|---|---|---|---|---|
| $\mathcal{L}^{\text{ori}}$+$\mathcal{L}^{\text{kpts}}$ | **57.6** | **73.1** | **79.6** | **505.8** | **76.7** | **82.3** | **440.1** |
| $\mathcal{L}^{\text{ori}}$ | 30.0 | 44.4 | 56.6 | 403.3 | 49.6 | 61.9 | 291.6 |
| $\mathcal{L}^{\text{kpts}}$ | 50.8 | 69.8 | 76.8 | 358.7 | 75.2 | 81.2 | 226.7 |

Table 2. Ablation experiment about the dense orientation alignment loss $\mathcal{L}^{\text{ori}}$ and the window-based keypoint detection loss $\mathcal{L}^{\text{kpts}}$ in HPatches [1]. We use 1,000 keypoints and the HardNet descriptor [10]. 'out. filter.' denotes the results with outlier filtering, and 'match' denotes the number of predicted matches.

In section 2.2, we show the results of ablation studies. In section 2.3, we show the separated results of the HPatches viewpoint/illumination.

## 2.1. Evaluation on the ER dataset [8]

Table 1 shows the Percentage of Correctly Matched Keypoints (PCK) in the ER dataset proposed in [8]. The ER dataset contains image pairs with large rotations produced by artificially transforming the images of HPatches [1] and SUN3D [17]. We only use our keypoints without outlier filtering by the orientations in this experiment. Our rotation-invariant keypoint detector improves PCKs by finding the more reliable keypoints within the extreme rotation setting than SuperPoint [5]. In addition, the integration with ours and GIFT [8] achieves the best PCKs compared to the previous best, SuperPoint [5] with GIFT [8], in the ER dataset.

## 2.2. Ablation studies

**Ablations of the loss functions.** Table 2 shows the results without each loss function. Without $\mathcal{L}^{\text{kpts}}$ in the second row, the repeatability score is decreased because the model cannot obtain the keypoint at a reliable location, so the performances of matching are also decreased. Although without $\mathcal{L}^{\text{ori}}$ in the third row, outlier filtering is working because the rotation-equivariant representation $\mathbf{O}$ groups the rotation information of local patterns by the rotation-equivariant networks. However, using both loss functions as in the first row yields higher MMA with more matches, which shows both loss function contributes to generating reliable oriented keypoints in an image.

**Different pooling operators in networks.** Table 3 shows the results with different pooling operators to verify the design choice of our networks. We use max pooling, average pooling, and bilinear pooling [8] for the keypoint detection branch when collapsing the group, and $1 \times 1$ convolution, average pooling, and max pooling for the orientation estimation branch when collapsing the channel. We experiment with all possible exhaustive pairs of these combinations. As a result, the first row proposed in the main paper is best to use max pooling for keypoint detection and $1 \times 1$ convolution for orientation estimation. Collapsing the channel with $1 \times 1$ convolution in orientation estimation operates as a weighted sum with the learned kernel, giving richer information than max pooling and average pooling. Max pooling on the orientation branch yields compatible MMAs, but filters the excessive number of the predicted matches. We guess the poor performance of bilinear pooling is overfitting due to the excessive number of model parameters, although the loss converges during training, and the repeatability score of the validation set increases. Note that the bilinear pooling [8] takes a very long time because our keypoint map should compute all regions while GIFT generates only features of the extracted keypoints. Hence, the bilinear pooling is not appropriate to collapse the group of our method.

| **K** | **O** | rep. | w/o out. filter. MMA @3px | w/o out. filter. MMA @5px | w/o out. filter. match. | out. filter. MMA @3px | out. filter. MMA @5px | out. filter. match. |
|---|---|---|---|---|---|---|---|---|
| Max | 1x1Conv | **57.6** | **73.1** | **79.6** | **505.8** | **76.7** | **82.3** | **440.1** |
| Max | Avg | 54.6 | 70.6 | 77.4 | 483.1 | 76.2 | 81.5 | 397.8 |
| Max | Max | 56.0 | 71.8 | 78.7 | 500.3 | 76.9 | 82.7 | 352.1 |
| Avg | 1x1Monv | 55.7 | 72.3 | 78.6 | 480.6 | 75.9 | 81.7 | 339.6 |
| Avg | Avg | 51.0 | 67.2 | 75.5 | 459.5 | 70.0 | 78.2 | 315.4 |
| Avg | Max | 51.8 | 66.8 | 76.6 | 495.2 | 72.3 | 80.5 | 292.2 |
| Bilinear | 1x1Conv | 27.6 | 42.2 | 51.2 | 374.7 | 50.3 | 57.8 | 243.5 |
| Bilinear | Avg | 26.0 | 39.7 | 48.6 | 370.1 | 48.3 | 55.7 | 182.3 |
| Bilinear | Max | 29.6 | 43.4 | 52.6 | 381.4 | 53.1 | 60.5 | 178.3 |

Table 3. Results using different pooling operators. Column $\mathbf{K}$ denotes the operators of the keypoint detection branch, and column $\mathbf{O}$ denotes the operators of the orientation estimation branch. We use the same configuration of Table 2.

## 2.3. Separate results on HPatches

Tab. 4 shows the results of viewpoint/illumination on HPatches. Our rotation-equivariant detector with the group-invariant descriptor, GIFT [8], achieves the highest mean matching accuracy (MMA) overall on both variations. Although ORB [14] shows a higher repeatability under viewpoint changes, the results with our keypoint detector consistently show the better MMAs compared to ORB [14]. The repeatability score of our model is either the best or the second-best for each variation.

# 3. Qualitative results

Figure 2 qualitatively compares the orientation maps of SIFT [9], LF-Net [12], and ours. We use the synthetic images in Section 4.2 of the main paper. For obtaining the SIFT orientation, we partition an entire image into patches and estimate the dominant orientation of each patch except

| Detector | Descriptor | Illumination | | | | Viewpoint | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Rep. | MMA | | pred. match. | Rep. | MMA | | pred. match. |
| | | | @3px | @5px | | | @3px | @5px | |
| SIFT [9] | SIFT [9] | 42.3 | 48.0 | 51.0 | 405.4 | 41.8 | 51.0 | 53.9 | 406.5 |
| ORB [14] | ORB [14] | 54.6 | 48.1 | 52.0 | 378.2 | **60.0** | 45.1 | 48.1 | 346.3 |
| D2-Net [6] | D2-Net [6] | 26.9 | 47.7 | 61.8 | 411 | 12.9 | 23.2 | 35.8 | 333.9 |
| LF-Net [12] | LF-Net [12] | 48.9 | 56.1 | 61.3 | 337.8 | 38.8 | 48.1 | 52.6 | 322.9 |
| R2D2 [13] | R2D2 [13] | 48.5 | 70.0 | 80.6 | 399.3 | 42.6 | 59.3 | 69.2 | 320.0 |
| SuperPoint [5] | SuperPoint [5] | 51.7 | 68.6 | 76.2 | 469.9 | 42.4 | 58.5 | 63.9 | 467.6 |
| SuperPoint [5] | GIFT [8] | 51.7 | 69.5 | 77.5 | 484.2 | 42.4 | 68.2 | 74.6 | **508.3** |
| Key.Net [2] | HardNet [10] | 54.1 | 70.8 | 78.3 | 497.4 | 57.7 | 74.1 | 80.4 | 452.2 |
| Key.Net [2] | SOSNet [16] | 54.1 | 70.8 | 78.3 | 487.9 | 57.7 | 74.5 | 80.9 | 442.2 |
| Key.Net [2] | HyNet [15] | 54.1 | 69.8 | 77.3 | 499.9 | 57.7 | 74.1 | 80.5 | 451.5 |
| ours | HardNet [10] | **57.1** | 74.0 | 81.1 | **556.2** | 58.1 | 72.2 | 78.1 | 457.1 |
| ours | SOSNet [16] | **57.1** | 74.5 | **81.6** | 550.9 | 58.1 | 72.4 | 78.4 | 449.8 |
| ours | HyNet [15] | **57.1** | 73.5 | 80.6 | 555.6 | 58.1 | 72.3 | 78.4 | 452.8 |
| ours | GIFT [8] | **57.1** | **75.4** | 81.1 | 443.6 | 58.1 | **75.4** | **81.1** | 388.6 |

Table 4. Separated results on HPatches illumination/viewpoint variations. We evaluate the Key.Net [2] results using the re-trained model with the code provided by the authors. Results in bold indicate the best result, and underlined results indicate the second best results.

the boundary regions. Each result consists of three rows. The first rows show the source image and the estimated source orientation maps, and the second rows show the target image and the estimated target orientation maps spatially aligned to the source image with GT homography. In the third rows, we first compute the difference of orientation maps, and then compute the correctness by thresholding the error 15° using the ground-truth angle. Our correctness maps consistently keep more pixels as correct, implying that our model produce a more accurate relative orientation of each pixel than SIFT [9] and LF-Net [12].

Figure 3 and 4 show the qualitative results for the HPatches illumination and viewpoint, respectively. We use HardNet descriptor [10] for Key.Net [2] and ours and use their own descriptor for SIFT [9] and LF-Net [12]. We use mutual nearest matcher for all cases. Our model consistently finds the larger number of correct matches (green) and the smaller number of incorrect matches (red) compared to the baselines in the viewpoint and illumination examples.

Figure 5 visualizes the predicted matches on the validation set of Phototourism in IMC2021 [7]. We draw the inliers produced by DEGENSAC [4]. We color the correct matches from green (0 pixel off) to yellow (5 pixels off), and the incorrect matches in red (more than 5 pixels off). Matches with occluding keypoints by changing the camera pose are drawn in blue. In this unconstrained urban scene, our model generates a larger number of correct matches with a smaller number of false positives than the previous keypoint detectors, SIFT+AN [9, 11] and Key.Net [2], in the same image matching pipeline.

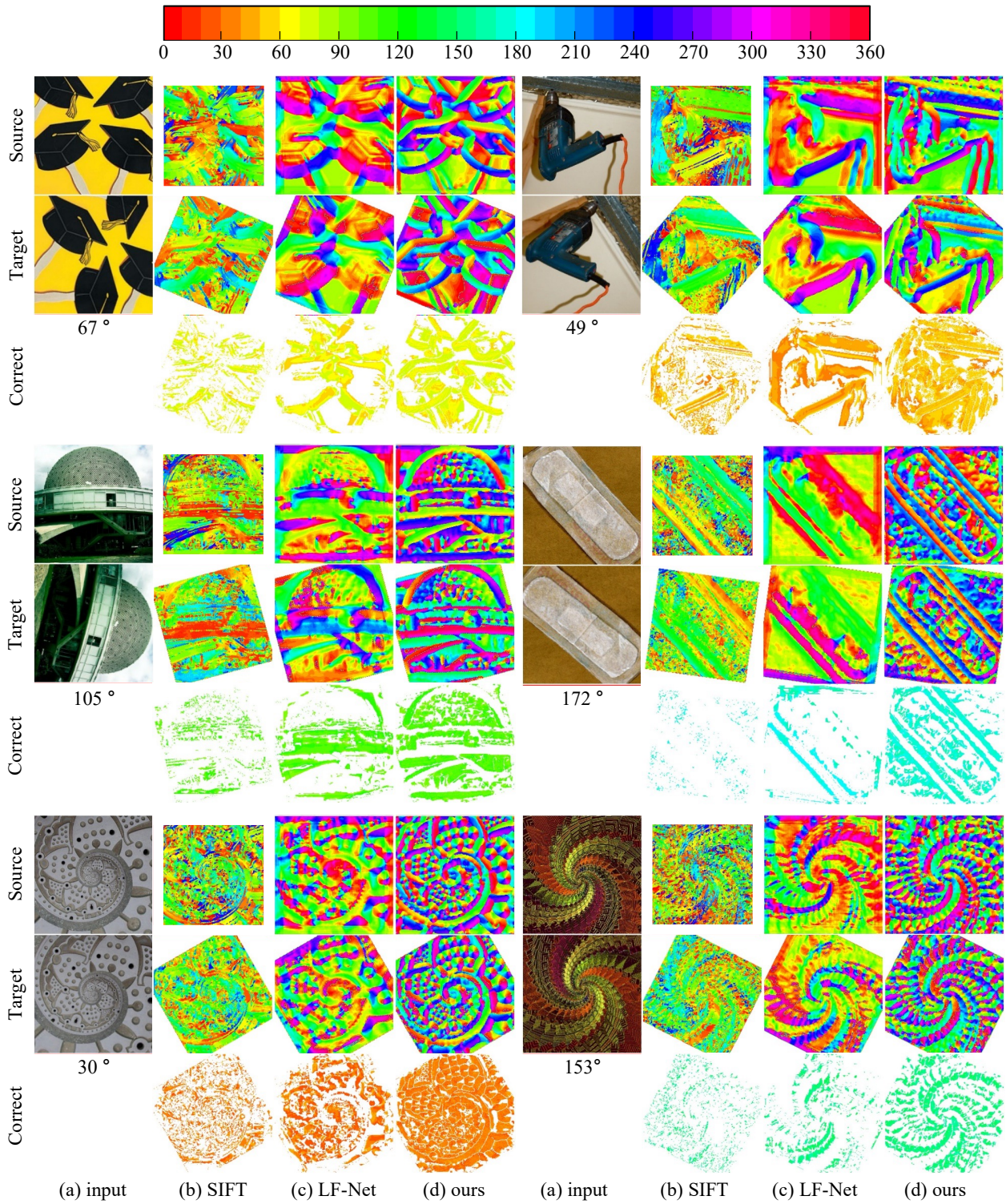
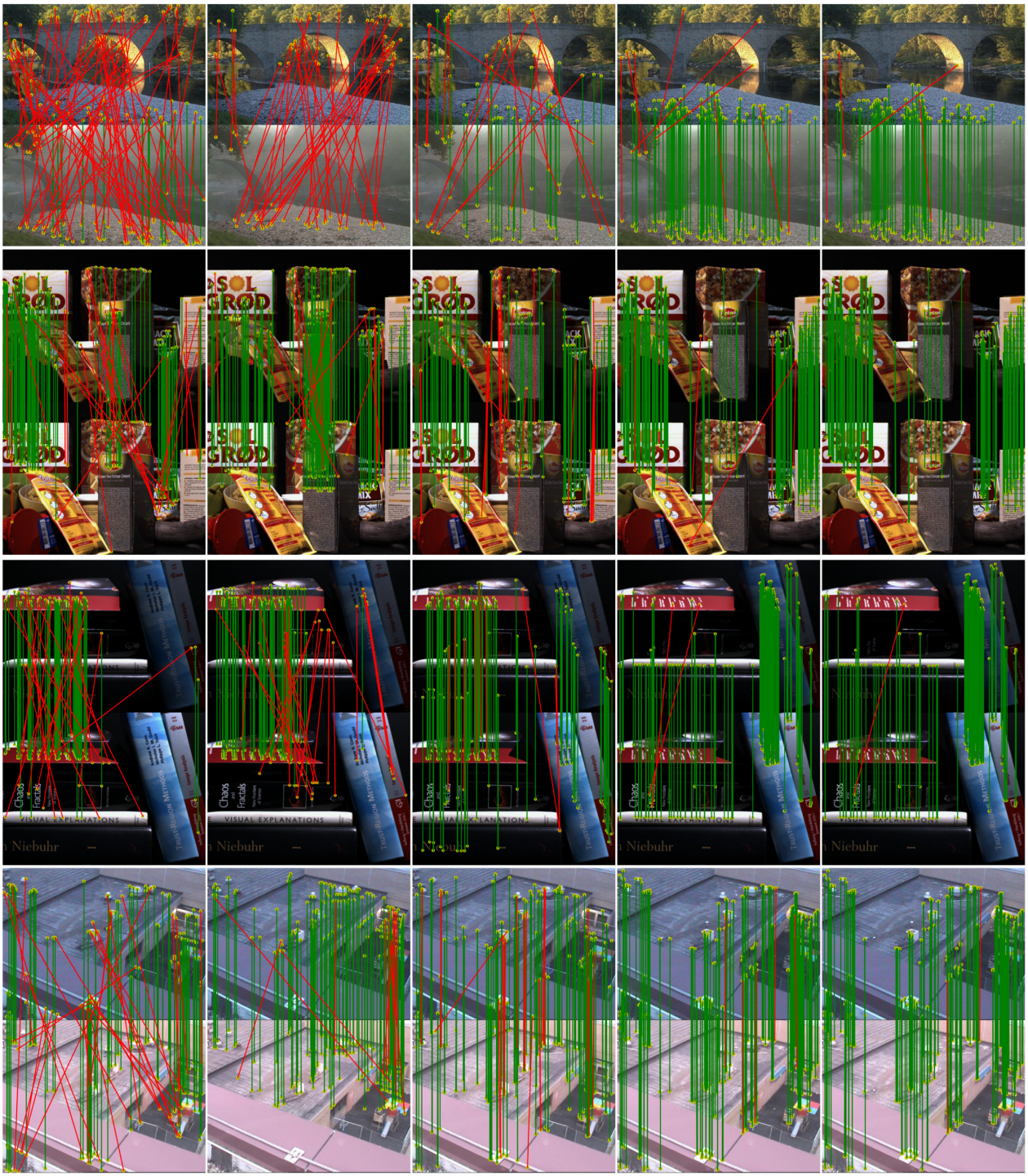
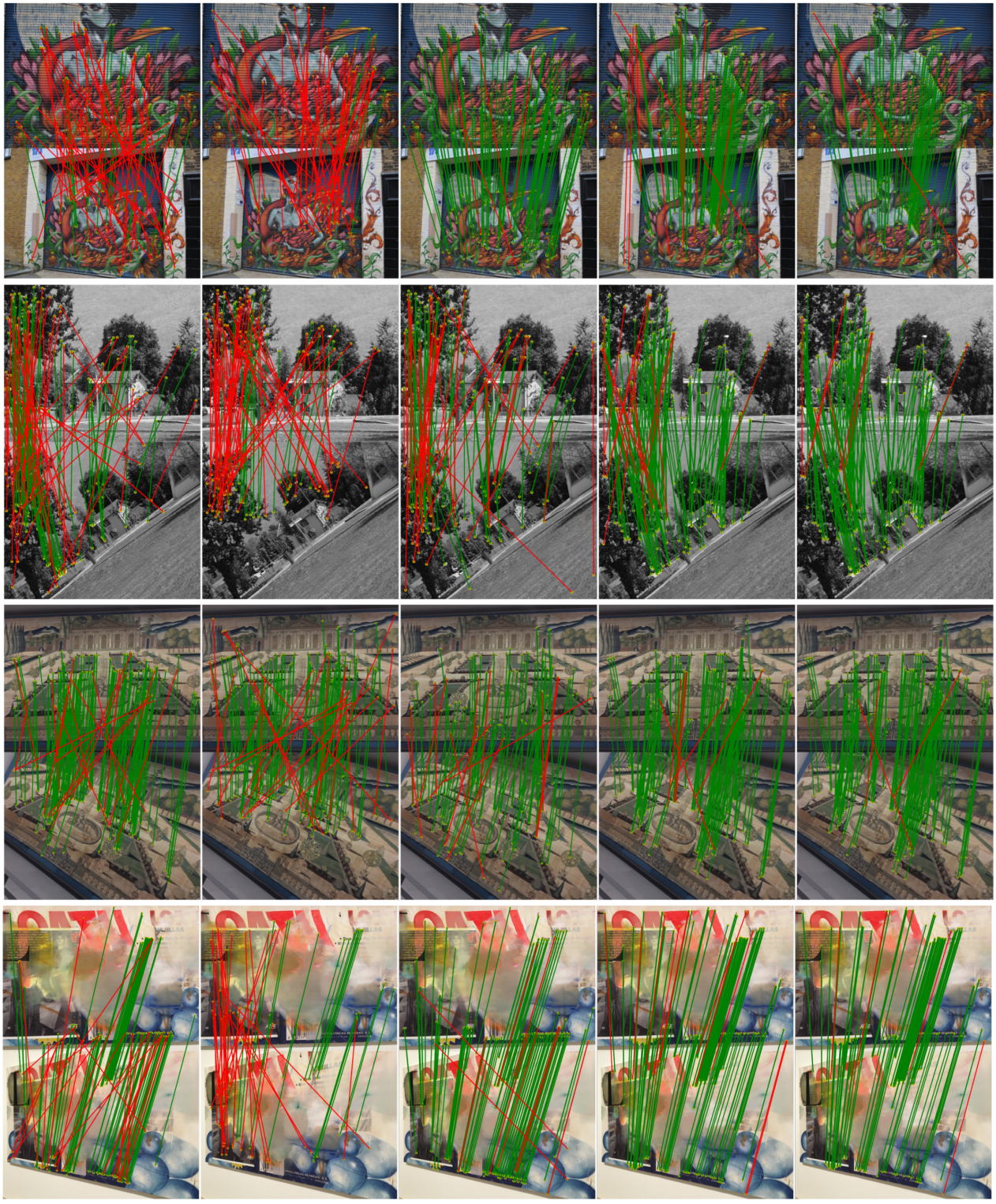



Figure 2. Visualization of the color-coded orientation maps under synthetic rotations with Gaussian noise. We spatially align the orientation map of the target image to the source image using a ground-truth homography for a better view. We create a correctness map between the source orientation map and the aligned orientation map by computing a 15° threshold in the third row. The numbers below the target images indicate the ground-truth angles between the source and target images. We use the HSV color representation to visualize the color map of the orientations. The color bar located at the top denotes the corresponding color of orientation degree. Two examples with complicated patterns at the bottom show that our orientation estimator derives more accurate orientations than the existing orientation estimators [9, 12].



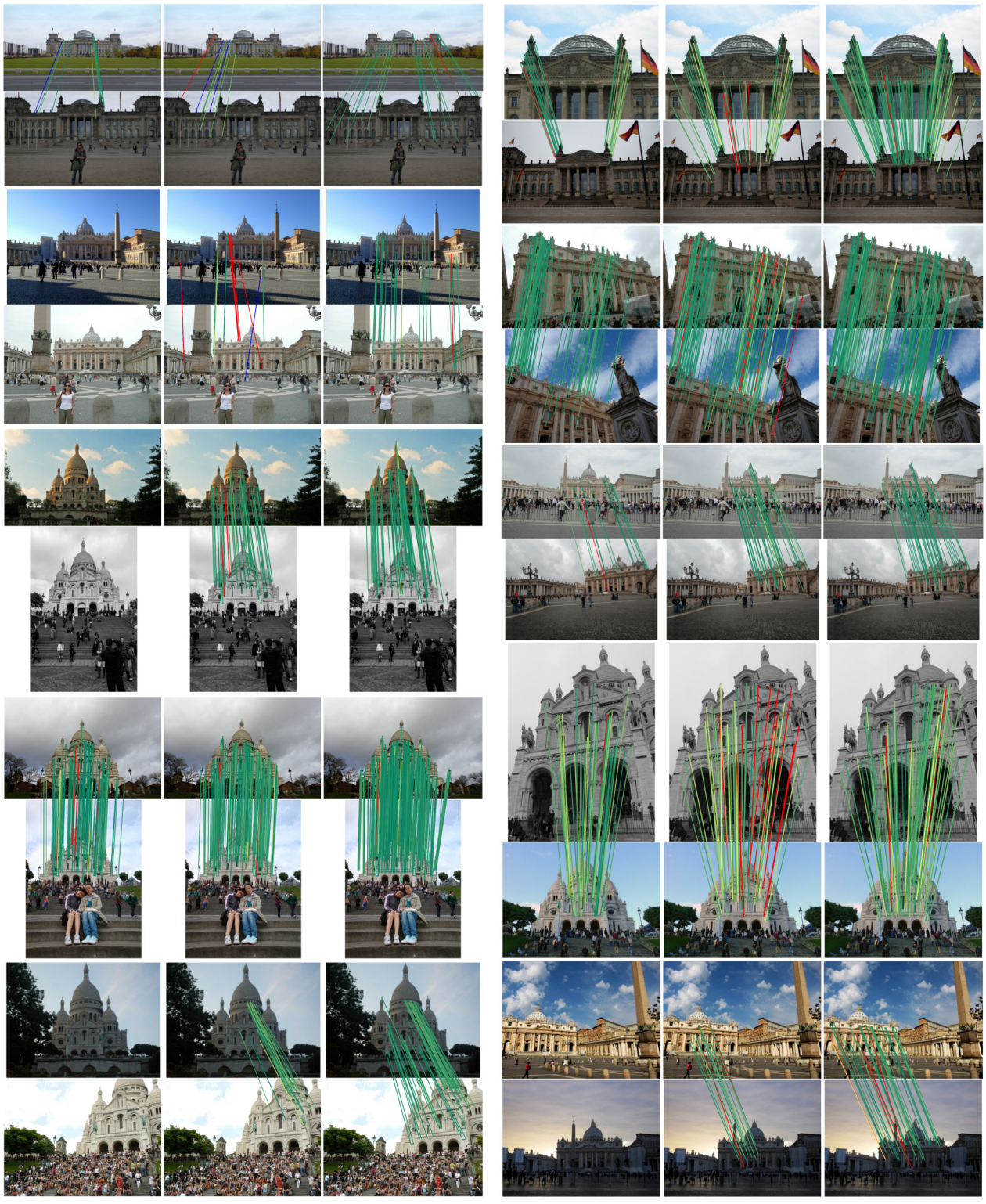
(a) SIFT (b) LF-Net (c) Key.Net (d) ours (e) ours (fltr)

Figure 3. Visualization of the predicted matches in HPatches illumination variations [7]. We detect 300 keypoints for each image and match them by mutual nearest neighbors. The green and red lines indicate correct and incorrect matches, respectively, by a three-pixel threshold. Our rotation-invariant keypoints in column (d) derives smaller number of false-positive matches than columns (a), (b), and (c). Column (e) using the outlier filtering shows that the characteristic orientations effectively filter a set of the correct matches in illumination variations.



(a) SIFT (b) LF-Net (c) Key.Net (d) ours (e) ours (fltr)

Figure 4. Visualization of the predicted matches in HPatches viewpoint variations [7]. We draw the correct matches (green) and the incorrect matches (red) by a three-pixel threshold. Our oriented keypoint detector with HardNet [10] (column (d), (e)) produce a larger number of matches with a smaller number of false positives in extreme rotation case (row 2) and 3D viewpoint changes (row 1, 3, 4) compared to the columns (a) SIFT [9], (b) LF-Net [12] and (c) Key.Net+HardNet [2, 10].



(a) SIFT+AN (b) Key.Net (c) ours (a) SIFT+AN (b) Key.Net (c) ours

Figure 5. Visualization of the predicted matches in IMC2021 [7], with HardNet descriptor [10], DEGENSAC [4] with AdaLAM [3]. Matches above the 5-pixel error threshold are displayed in red, and matches below are color-coded according to errors between 0 (green) to 5 pixels (yellow). Matches without a depth estimate are displayed in blue. We use 1,024 keypoints to compare with SIFT+AN [9, 11], Key.Net [2], and ours.

## References


[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5173–5182, 2017. 2

[2] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5836–5844, 2019. 3, 6, 7

[3] Luca Cavalli, Viktor Larsson, Martin Ralf Oswald, Torsten Sattler, and Marc Pollefeys. Handcrafted outlier detection revisited. In *European Conference on Computer Vision*, pages 770–787. Springer, 2020. 7

[4] Ondrej Chum, Tomas Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 772–779. IEEE, 2005. 3, 7

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Deep Learning for Visual SLAM Workshop*, 2018. 2, 3

[6] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 3

[7] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021. 1, 3, 5, 6, 7

[8] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Gift: Learning transformation-invariant dense visual descriptors via group cnns. *Advances in Neural Information Processing Systems*, 32:6992–7003, 2019. 1, 2, 3

[9] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2, 3, 4, 6, 7

[10] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, pages 4826–4837, 2017. 2, 3, 6, 7

[11] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–300, 2018. 3, 7

[12] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: learning local features from images. In *Advances in neural information processing systems*, pages 6234–6244, 2018. 2, 3, 4, 6

[13] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32:12405–12415, 2019. 3

[14] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 1, 2, 3

[15] Yurun Tian, Axel Barroso Laguna, Tony Ng, Vassileios Balntas, and Krystian Mikolajczyk. Hynet: Learning local descriptor with hybrid similarity measure and triplet loss. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[16] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11016–11025, 2019. 3

[17] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, pages 1625–1632, 2013. 2