Learning Rotation-Equivariant Features for Visual Correspondence
- supplementary material -

Jongmin Lee Byungjin Kim Seungwook Kim Minsu Cho
Pohang University of Science and Technology (POSTECH), South Korea
http://cvlab.postech.ac.kr/research/RELF

In this supplementary material, we present a formal introduction of group equivariance briefly, an additional explanation of multiple descriptor extraction, results on the ERDNIM dataset, and additional qualitative results. Section 1 explains a formal definition of equivariance and group equivariant networks. Section 2 evaluates the matching quality of our proposed method under rotation and illumination variations on the day/night image pairs, with details about the benchmark generation. Section 3 shows the results of realistic downstream task on the IMC2021 [8] dataset. Section 4 shows the comparisons of computational overhead and the number of parameters. Section 5 shows different strategies of multiple descriptor extraction using dominant orientation candidates. Section 6 evaluates the existing feature matching methods in the Roto-360 dataset. Section 7 shows the re-training results of GIFT with cyclic rotation augmentation. Section 8 shows the matching results with increasing the number of samples of the Roto-360 dataset. Section 9 presents additional qualitative results to visualize the consistency of dominant orientation estimation, the similarity maps under in-plane rotations of images, and predicted matches on the HPatches and extreme rotation (ER) datasets [1, 14].

1. Group equivariance

A feature extractor Φ is said to be equivariant to a geometric transformation T_g if transforming an input $x \in X$ by T_g and then passing it through Φ gives the same result as first passing x through Φ and then transforming the resulting feature map by T'_g. Formally, the equivariance can be expressed for transformation group G and $\Phi : X \rightarrow Y$ as

$$\Phi[T_g(x)] = T'_g[\Phi(x)], \tag{1}$$

where T_g and T'_g represent transformations on each space of a group action $g \in G$. If T_t is a translation group $(\mathbb{R}^2, +)$, and f is a feature mapping function $\mathbb{Z}^2 \rightarrow \mathbb{R}^K$ given convolution filter weights $\psi \in \mathbb{R}^{2\times K}$, the translation equivariance of a convolutional operation can be expressed as follows:

$$[T_t f] * \psi(x) = [T_t[f * \psi]](x), \tag{2}$$

where $*$ indicates the convolution operation.

Recent studies [2–4, 26, 27] propose convolutional neural networks that are equivariant to symmetry groups of translation, rotation, and reflection. Let H be a rotation group. The group G can be defined by $G \cong (\mathbb{R}^2, +) \rtimes H$ as the semidirect product of the translation group $(\mathbb{R}^2, +)$ with the rotation group H. Then, the rotation-equivariant convolution on group G can be defined as:

$$[T_g f] * \psi(g) = [T_g[f * \psi]](g), \tag{3}$$

by replacing $t \in (\mathbb{R}^2, +)$ with $g \in G$ in Eq. 2. This operation can be applied to an input tensor to produce a translation and rotation-equivariant output. Extending this, a network equivariant to both translation and rotation can be constructed by stacking translation and rotation-equivariant layers instead of conventional translation-equivariant layers. Formally, let $\Phi = \{L_i | i \in \{1, 2, 3, ..., M\}\}$, which consists of M rotation-equivariant layers under group G. For one layer $L_i \in \Phi$, the transformation T_g is defined as

$$L_i[T_g(g)] = T_g[L_i(g)], \tag{4}$$

which indicates that the output is preserved after L_i about T_g. This can be extended to apply T_g to input I and then pass it through the network ϕ to preserve the transformation T_g for the whole network.

$$[\Pi_{i=1}^{M} L_i](T_g I) = T_g[\Pi_{i=1}^{M} L_i](I). \tag{5}$$

2. Experiments in *extreme* rotated day-night image matching (ERDNIM)

To show the robustness of our method under both geometric and illumination changes, we evaluate the matching performance of our method in the *extreme* rotated Day-Night Image Matching (ERDNIM) dataset, which rotates the reference images of the RDNIM dataset [18], which is originally from the DNIM dataset [28].

		SIFT	SuperPoint	D2Net	R2D2	KeyNet+ HyNet	GIFT	LISRD	ours	ours*
Day	HEstimation	0.064	0.073	0.001	0.044	0.085	0.108	0.228	<u>0.232</u>	**0.272**
	MMA	0.049	0.082	0.024	0.054	0.068	0.123	<u>0.270</u>	0.245	**0.277**
Night	HEstimation	0.108	0.092	0.002	0.062	0.097	0.151	0.291	<u>0.316</u>	**0.364**
	MMA	0.082	0.111	0.033	0.076	0.093	0.177	0.358	<u>0.362</u>	**0.404**

Table 1. **Comparison of matching quality on the ERDNIM dataset.** We use two evaluation metrics: homography estimation accuracy (HEstimation), and mean matching accuracy (MMA) at 3 pixel thresholds. Results in **bold** indicate the best score and <u>underlined</u> results indicate the second best scores.

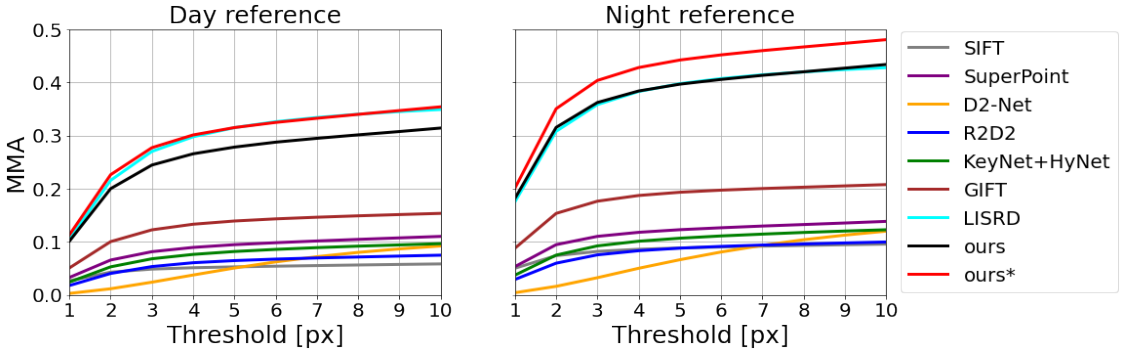


Figure 1. **Results of MMA with different pixel thresholds on the ERDNIM dataset.** 'ours*' uses k differently group-aligned features based on top-k selection. We use $k = 4$ in this experiment.

2.1. Data generation

The source dataset DNIM [28] consists of 1722 images from 17 sequences of a fixed webcam taking pictures at regular time spans over 48 hours. They construct the pairs of images to match by choosing a day and a night reference image for each sequence as follows: we first select the image with the closest timestamp to noon as the day reference image, and the image with the closest timestamp to midnight as the night reference image. Next, we pair all the images within a sequence to both the day reference image and the night reference image. Therefore, 1,722 image pairs are obtained for each of the day benchmark and night benchmark, where the day benchmark is composed of day-day and day-night image pairs, and the night benchmark is composed of night-day and night-night image pairs. To evaluate the robustness under geometric transformation, the RDNIM [18] dataset is generated by warping the target image of each pair with homographies as in SuperPoint [5] generated with random translations, rotations, scales, and perspective distortions. Finally, we add rotation augmentation to the reference image of each pair to evaluate the rotational robustness, and call this dataset *extreme* rotated Day-Night Image Matching (ERDNIM). We randomly rotate the reference images in the range $[0°, 360°)$. The number of image pairs for evaluation remains the same as RDNIM [18]. Figure 2 shows some examples of ERDNIM image pairs.

2.2. Examples of ERDNIM image pairs

2.3. Evaluation metrics

We use two evaluation metrics, HEstimation and mean matching accuracy (MMA), following LISRD [18]. We measure the homography estimation score [5] using RANSAC [7] to fit the homography using the predicted matches. To measure the estimation score, we first warp the four corners of the reference image using the predicted homography, and measure the distance between the warped corners and the corners warped using the ground-truth homography. The predicted homography is considered to be correct if the average distance between the four corners is less than a threshold: HEstimation$= \frac{1}{4}\sum_{i=1}^{4} ||\hat{c}_i - c_i||_2 \leq \epsilon$, where we use $\epsilon = 3$. MMA [6, 19] is the percentage of the correct matches over all the predicted matches, where we also use 3 pixels as the threshold to determine the correctness of matches.

2.4. Results

Table 1 shows the evaluation results on the ERDNIM dataset. We compare the descriptor baselines SIFT [15], SuperPoint [5], D2-Net [6], R2D2 [19], KeyNet+HyNet [9, 24], GIFT [14], and LISRD [18]. Our proposed model with the rotation-equivariant network (ReResNet-18) achieves state-of-the-art performance in terms of homography estimation. GIFT [14], an existing rotation-invariant descriptor,

shows a comparatively lower performance on this extremely rotated benchmark with varying illumination. Note that we use the same dataset generation scheme with the same source dataset [13] to GIFT [14]. LISRD [18], which selects viewpoint and illumination invariance online, demonstrates better MMA than ours on the *Day* benchmark, but ours* which extracts top-k candidate descriptors shows the best MMA and homography estimation on both *Day* and *Night* benchmarks.

Figure 1 shows the results of mean matching accuracy with different pixel thresholds on the ERDNIM dataset. Our descriptor with top-k candidate selection denoted by ours* achieves the state-of-the-art MMA at all pixel thresholds on both the day and night benchmarks. The results show our local descriptors achieve not only rotational invariance, but also robustness to geometric changes with perspective distortions and day/night illumination changes.

3. Results of the realistic downstream task

desc.	Stereo track			
	# kpts	mAA 5°	mAA 10°	# inliers
SuperPoint	1024	0.259	0.348	61.9
GIFT	1024	0.292	0.394	70.8
ours	1024	**0.305**	**0.404**	**99.8**
SuperPoint	2048	0.263	0.358	73.9
GIFT	2048	**0.313**	**0.420**	98.6
ours	2048	0.296	0.403	**118.5**

Table 2. **Results of the downstream task in IMC2021 [8].** We use the SuperPoint keypoint detector for all methods.

In Table 2, we evaluate on IMC 2021 stereo track [8] using the validation set of PhotoTourism and PragueParks to show the results on a realistic downstream task. Our descriptor consistently performs better than SuperPoint [5] descriptors under varying number of keypoints, and obtains comparable results with GIFT [14] descriptors. This shows that our method performs similarly for the general and non-planar transformations, while it significantly outperforms existing methods on Roto-360 and RDNIM datasets with extreme rotation transformations. Note that it is also possible to use image pairs with GT annotations of intrinsic and extrinsic parameters by *approximating* the 2D relative orientation for our training[1], and we leave this for future.

4. Computational overhead and the number of parameters

4.1. Computational overhead

Table 3 compares an average of the inference time and GPU usage with other descriptor extraction methods above.

[1]The details of obtaining the rotation from a homography can be found in Section 2 of "Deeper understanding of the homography decomposition (Malis and Vargas, 2007)".

method	speed (ms)	GPU usage (GB)
ours	147.4	5.21 GB
ours†	206.4	4.83 GB
SuperPoint [5]	**66.0**	**2.35 GB**
GIFT [14]	198.8	2.93 GB
LISRD [18]	781.0	2.85 GB
PosFeat [12]	208.8	4.67 GB

Table 3. **Comparison of computational overhead.** We compare inference time (milliseconds) and GPU memory usage (gigabytes) while fixing the number of keypoints.

While achieving strong rotational invariance, speed and GPU usage of ours are similar to those of existing local descriptor methods. Note that, our group aligning has a time complexity of $O(1)$ on the GPU with the predicted orientation because it is a transposition operation and does not take up extra memory. In addition, the time complexity of our group-equivaraint feature extractor is the same to the conventional CNNs on GPU since the steerable CNNs multiply the basis kernels and the learnable parameters in test time. (Section 2.8 of [26])

4.2. The number of parameters

The right table shows the number of parameters in millions, where the first group (top) are descriptor-only models and the second group (bottom) are joint detection and description models. Our model in the first row has a second smallest model size among those of descriptor-only models. When using our model with the deeper backbone denoted denoted by ours†, the number of model parameters increases, but it does not increase significantly compared to other comparison groups, where is still similar to that of LF-Net [17].

method	# params
ours	0.6M
ours†	2.6M
GIFT [14]	**0.4M**
LISRD [18]	3.7M
PosFeat [12]	21.1M
HardNet [16]	9.0M
HyNet [25]	1.3M
SuperPoint [5]	1.3M
LF-Net [17]	2.6M
RF-Net [22]	1.4M
D2-Net [6]	7.6M
R2D2 [19]	**0.5M**

5. Elaboration of multiple descriptor extraction

In this section, we show the results of different configurations of the multiple descriptor extraction scheme which was mentioned in Section 4.3, Table 3, Table 4, and Table 6 of the main paper.

Table 4 shows the results with different strategies for multiple descriptor extraction on the Roto-360 dataset. It can be seen that using a score ratio of 0.6 selects multiple candidates dynamically, where the total number of candidates is similar to using top-2 candidates, but the MMA@5px is as high as using top-3 candidates which

cand.	Roto-360			
	@5px	@3px	pred.	total.
top1	91.35	90.18	688	1161
top2	92.31	91.19	1315	2322
top3	**92.82**	**91.69**	2012	3483
0.8	92.25	91.13	951	1660
0.6	**92.82**	**91.69**	1333	2340

Table 4. **Results with different multiple descriptor extraction strategies.** The first group uses a static candidate selection strategy *i.e.,* the number of candidate orientations is fixed. The second group uses the dynamic candidate selection strategy, where only the score threshold is determined, and the number of orientation candidates may vary.

uses a higher number of candidates. Note that this multiple descriptor extraction scheme is largely inspired by the classical method based on an orientation histogram such as SIFT [15]. Owing to the parallel computation of GPUs for mutual nearest neighbor matching, the time complexity of constructing a correlation matrix to find matches is $O(1)$ regardless of the number of candidates.

6. Comparison with feature matching methods

method	Roto-360		
	@5px	@3px	pred.
ours+NN	**91.4**	**90.2**	688.3
SP+SG [5, 21]	30.1	29.8	874.1
LoFTR [23]	18.8	15.9	509.4

Table 5. **Comparison with keypoint matching methods on the Roto-360 dataset.**

Table 5 compares the feature matching methods to our descriptors with simple nearest neighbour matching (NN) algorithm. We evaluate our local feature with nearest neighbour matching (ours+NN) and compare it with SuperGlue [21] (*i.e.*, SuperPoint+SuperGlue [5, 21]) and LoFTR [23]. The results with the simple matching algorithm of ours+NN clearly outperforms the two other methods on the extremely rotated examples of the Roto-360 dataset. Note, however, that both SuperGlue [21] and LoFTR [23] are for feature *matching* and thus are not directly comparable to our method for feature *extraction*.

7. Changing the rotation range of the GIFT

Table 6 shows that GIFT* does not improve performance on the Roto-360 dataset because the bilinear pooling of GIFT does not guarantee invariance for rotation. This is because our group aligning computes invariant features without breaking any equivariance, in contrast to GIFT [14] whose bilinear pooling violates group equivariance due to their inter-group interaction from the 3×3 convolution across the group dimension, which makes invariance not guaranteed either. While GIFT and ours both use rotation-equivariant CNNs to finally yield an invariant descriptor, our architecture based on equivariant *kernels* guarantees cyclic rotation-equivariance *by construction*, unlike GIFT which relies on rotation augmentations to approximate equivariance.

method	Roto-360		
	5px	3px	pred.
ours	**91.35**	**90.18**	688.3
GIFT	42.05	41.59	589.2
GIFT*	40.71	40.27	564.2

Table 6. **The result of re-training the GIFT [14] model by replacing the rotation group with 360-degree cyclic.** GIFT* denotes a retrained model by extending the rotation sampling interval from -180° to 180°.

8. The number of sampled images for Roto-360

# sample	10	100	1K
Align	**91.4**	**80.0**	**89.9**
Avg	82.1	72.3	80.7
Max	78.0	69.3	79.2
None	18.8	16.4	20.5
Bilinear	41.0	28.5	43.7

Table 7. **Results on Roto-360 constructed using a different number of source images.**

Table 7 shows the mean matching accuracy (MMA) at 5 pixels threshold when increasing the number of source images to 100 images (3,600 pairs) and 1,000 images (36,000 pairs). The tendency of the matching results is maintained under increased diversity and complexity of the dataset, and group aligning consistently achieves state-of-the-art results. Therefore, we use 10 samples as they are sufficient to measure the relative rotation robustness of the local features.

9. Additional qualitative results

9.1. Visualization of the consistency of orientation estimation

We provide more examples for Figure 5 of the main paper, which visualize the consistency of orientation estimation. Additionally, we show the similarity map *w.r.t.* a keypoint under varying rotations. To visualize Figure 3, we create a sequence of 480×640 images augmented by random in-plane rotation with Gaussian noise sourced by ILSVRC2012 [20]. Figure 3 shows the qualitative comparison of the estimated orientation consistency. Given the dominant orientations estimated from the image pair, we

calculate the relative angle between the corresponding keypoint orientations and measure the difference between the relative angle and the ground-truth rotation. We evaluate the relative angle to be correct *i.e.,* the dominant orientation estimation is consistent if the difference with the ground-truth rotation is within a 30° threshold. Our rotation-equivariant model trained with the orientation alignment loss inspired by [10, 11] consistently estimates more correct keypoint orientations than LF-Net [17] and RF-Net [22].

9.2. Visualization of the similarity maps of a keypoint under varying rotations

Figure 4 shows the similarity maps with respect to a keypoint under varying rotations of images with a resolution of 180×180, with uniform rotation intervals of 45°. We compare one descriptor of a red keypoint from the source image at 0° to all other descriptors extracted across the rotated image using cosine similarity to compute the similarity maps. Yellow circles in the rotated images show the correct locations of the keypoint correspondences. We visualize 5 locations with the highest similarity scores with the query keypoint for better visibility. Our descriptor localizes the correct keypoint locations more precisely compared to GIFT [14] and LF-Net [17]. Specifically, although GIFT [14] uses group-equivariant features constructed using rotation augmentation, their descriptor fails to locate the corresponding keypoints accurately in rotated images - which shows that the explicit rotation-equivariant networks [26] yield better rotation-invariant features than constructing the group-equivariance features with image augmentation [14].

9.3. Visualization of the predicted matches on the extreme rotation

Figure 5 visualize the predicted matches on the ER dataset [14]. We extract a maximum of 1,500 keypoints from each image and find matches using the mutual nearest neighbor algorithm. The results show that our method consistently finds matches more accurately compared to GIFT [14] and LF-Net [17].

9.4. Visualization of the predicted matches on the HPatches viewpoint

Figure 6 visualize the predicted matches on the HPatches [1] viewpoint variations We extract a maximum of 1,500 keypoints from each image and find matches using the mutual nearest neighbor algorithm. The results show that our method consistently finds matches more accurately compared to GIFT [14] and LF-Net [17].

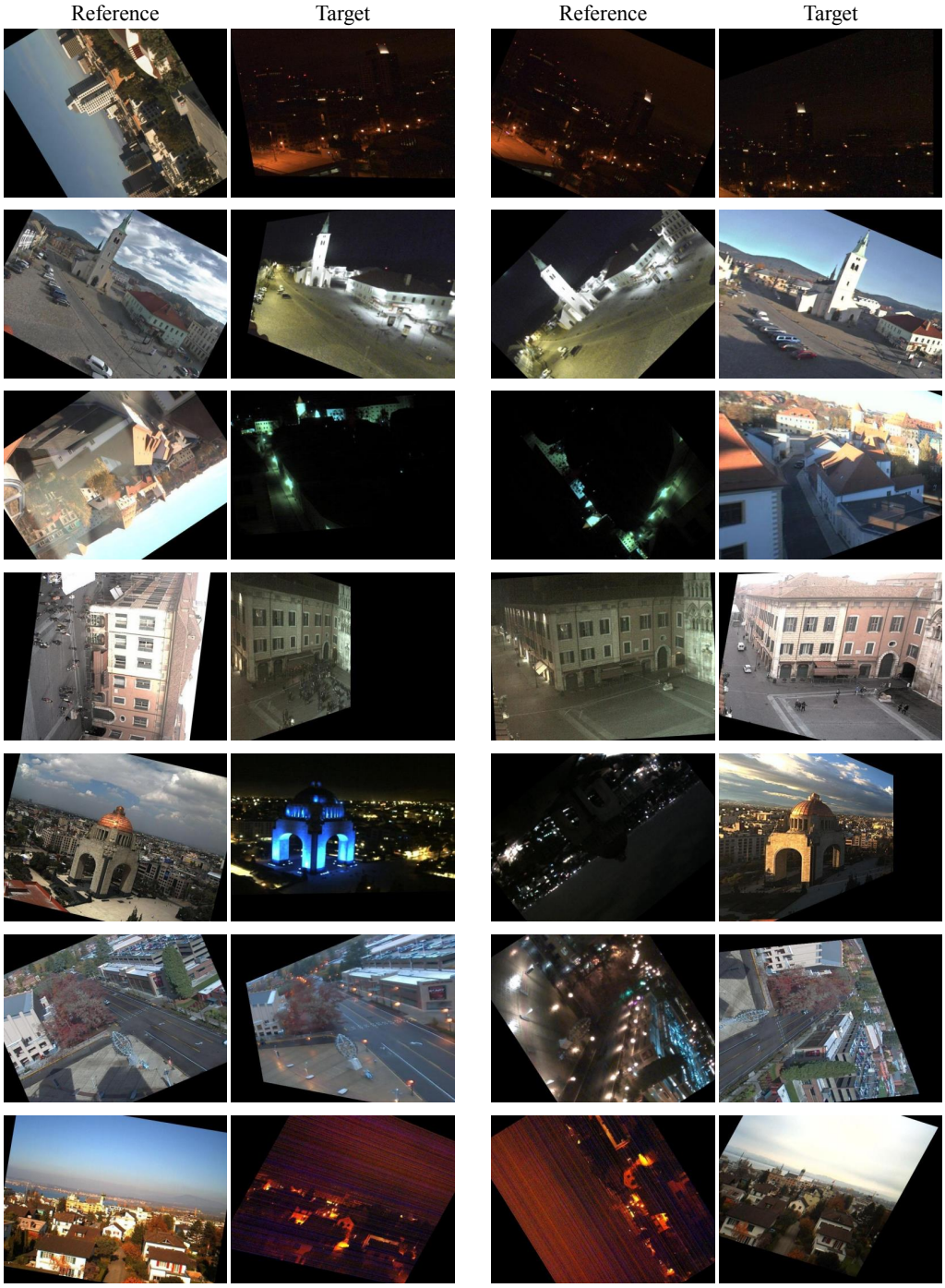


Figure 2. **Example of ERDNIM image pairs augmented from [18, 28].** The left two columns show the day reference benchmark with day-day and day-night image pairs. The right two columns show the night reference benchmark with night-day and night-night image pairs. The reference image of a pair is augmented with random rotation in the range $[0°, 360°)$, and the target image is augmented by homographies generated with random translation, rotation, scale, perspective distortion. The regions with black artifacts by homographies are masked out to measure the correctness of matching.

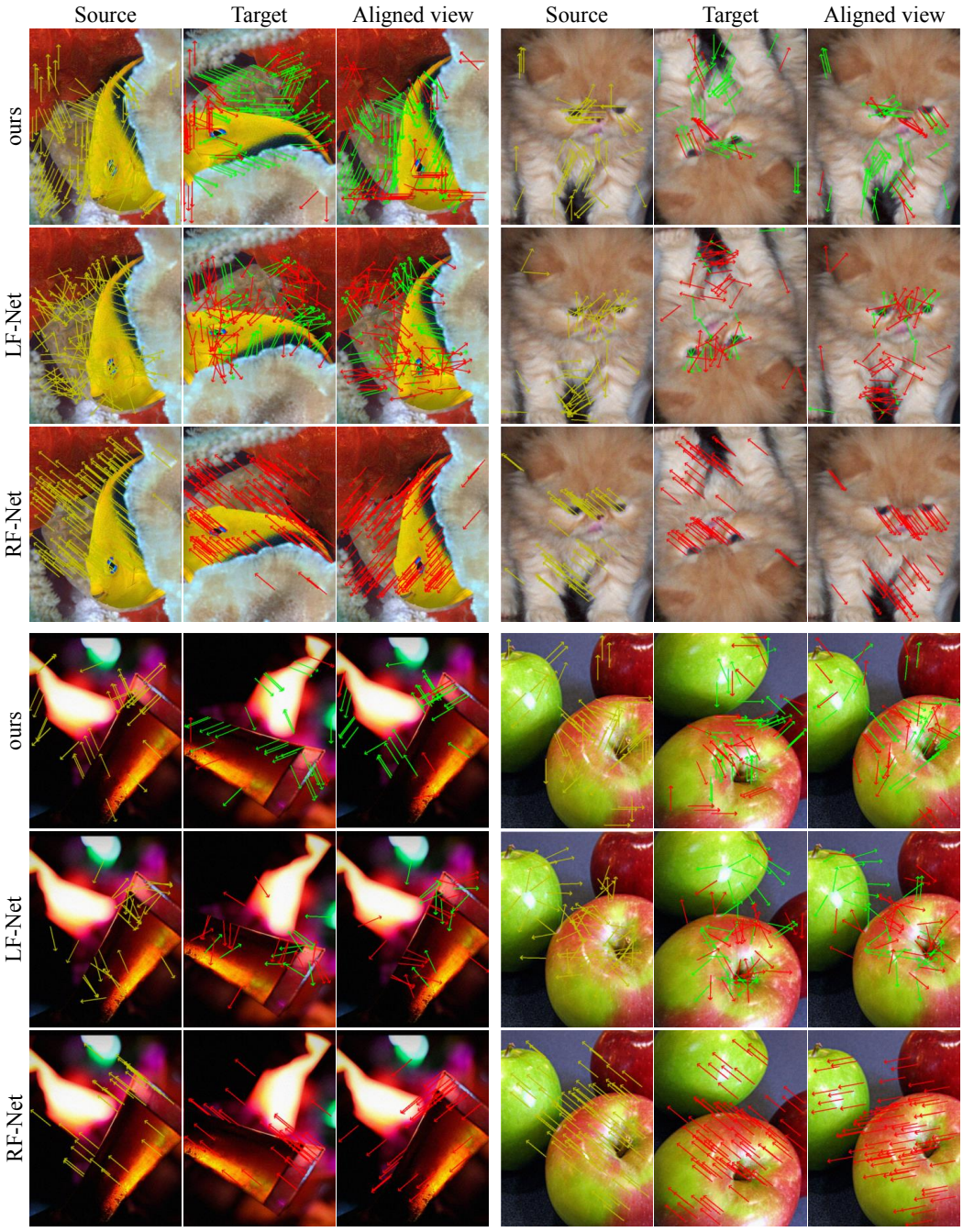


Figure 3. **Visualization of consistency of dominant orientation estimation.** We extract the source keypoints using SuperPoint [5] and obtain the target keypoints using GT homography. We evaluate the consistency of orientation estimation by comparing the relative angle difference and the ground-truth angle at a threshold of 30°. The green and red arrows represent consistent and inconsistent orientation estimations, respectively. We spatially align the target images and its' orientations to the source images for better visibility. Our method predicts more consistent orientations of keypoints compared to LF-Net [17] and RF-Net [22].

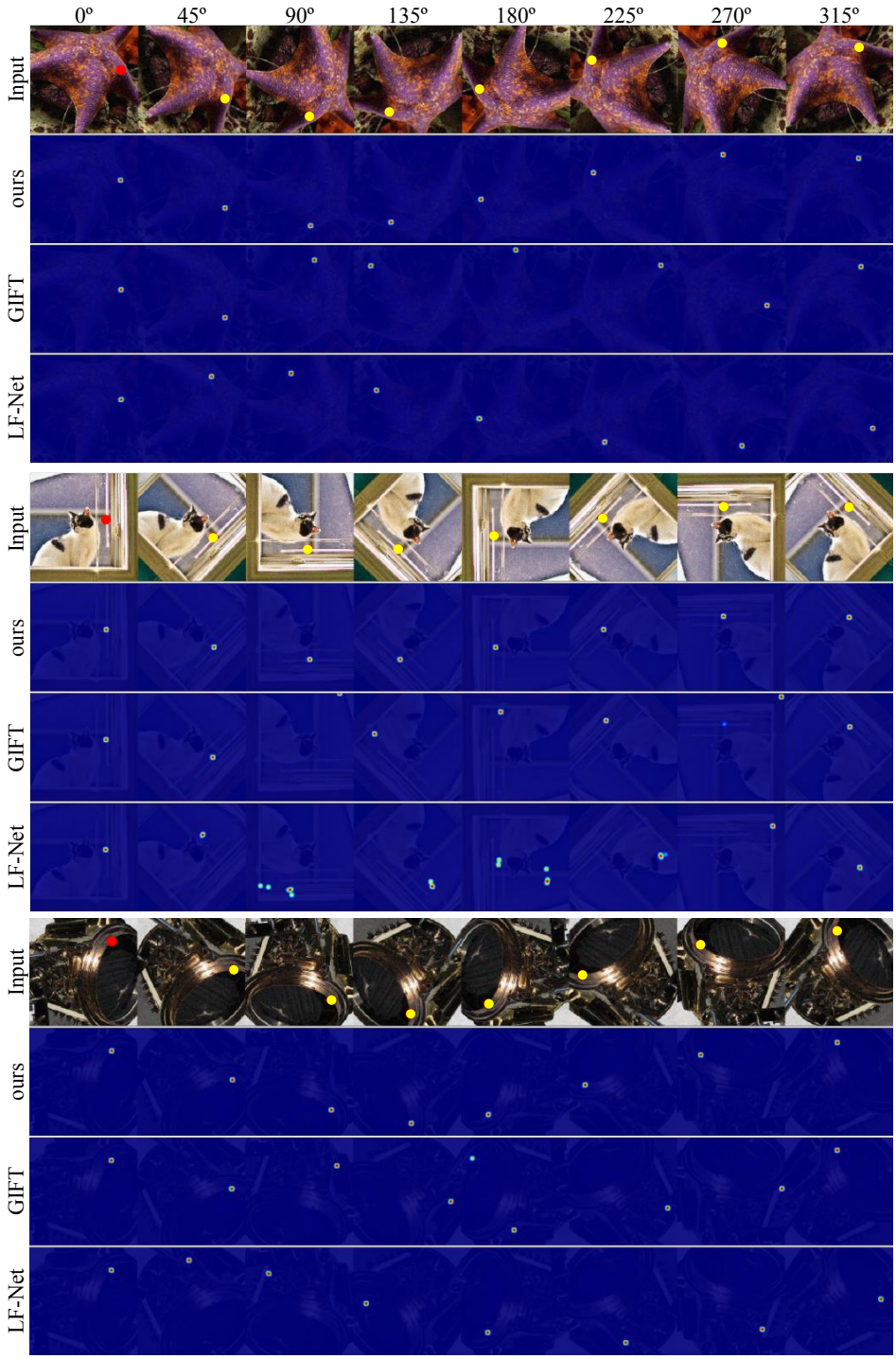
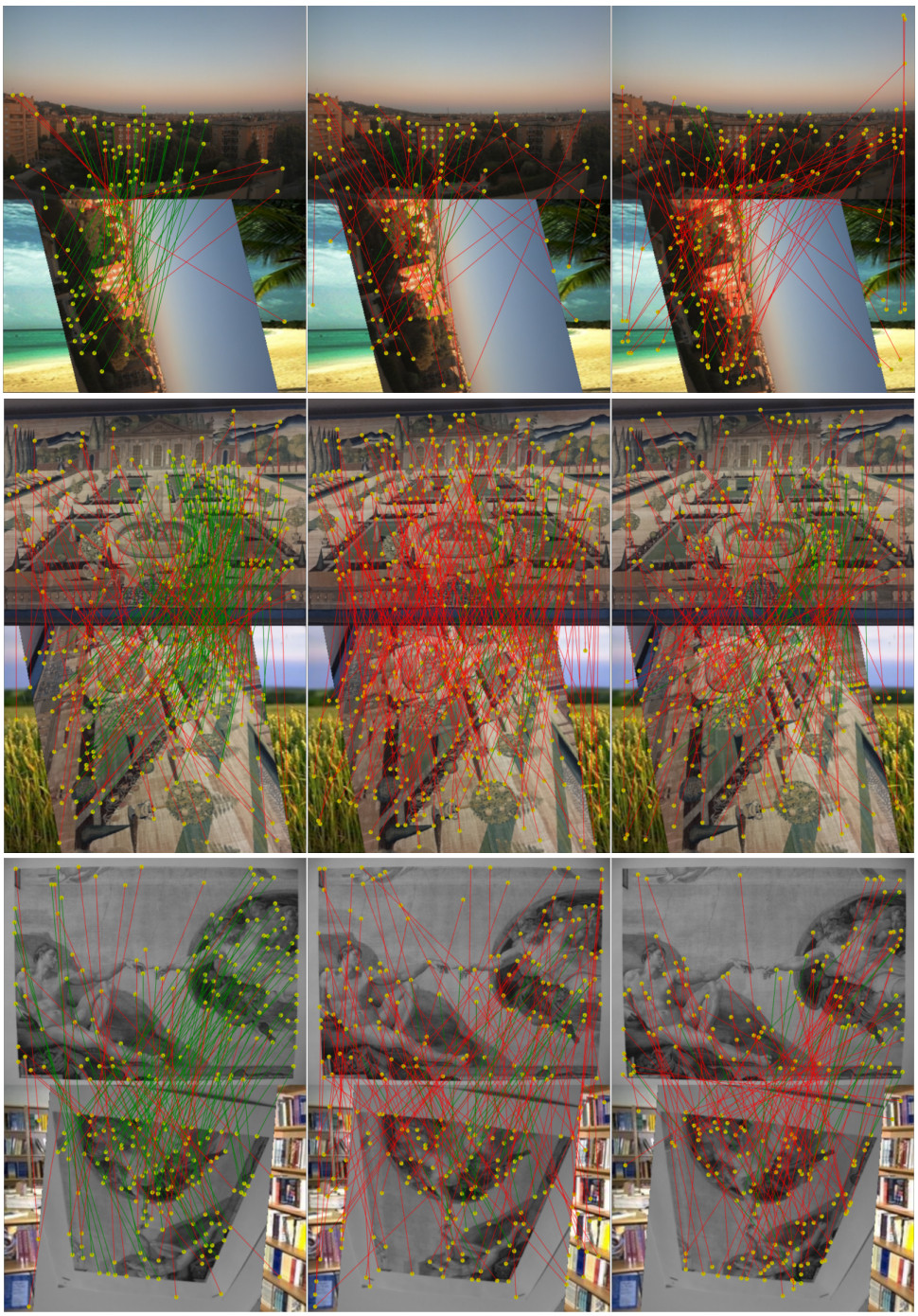
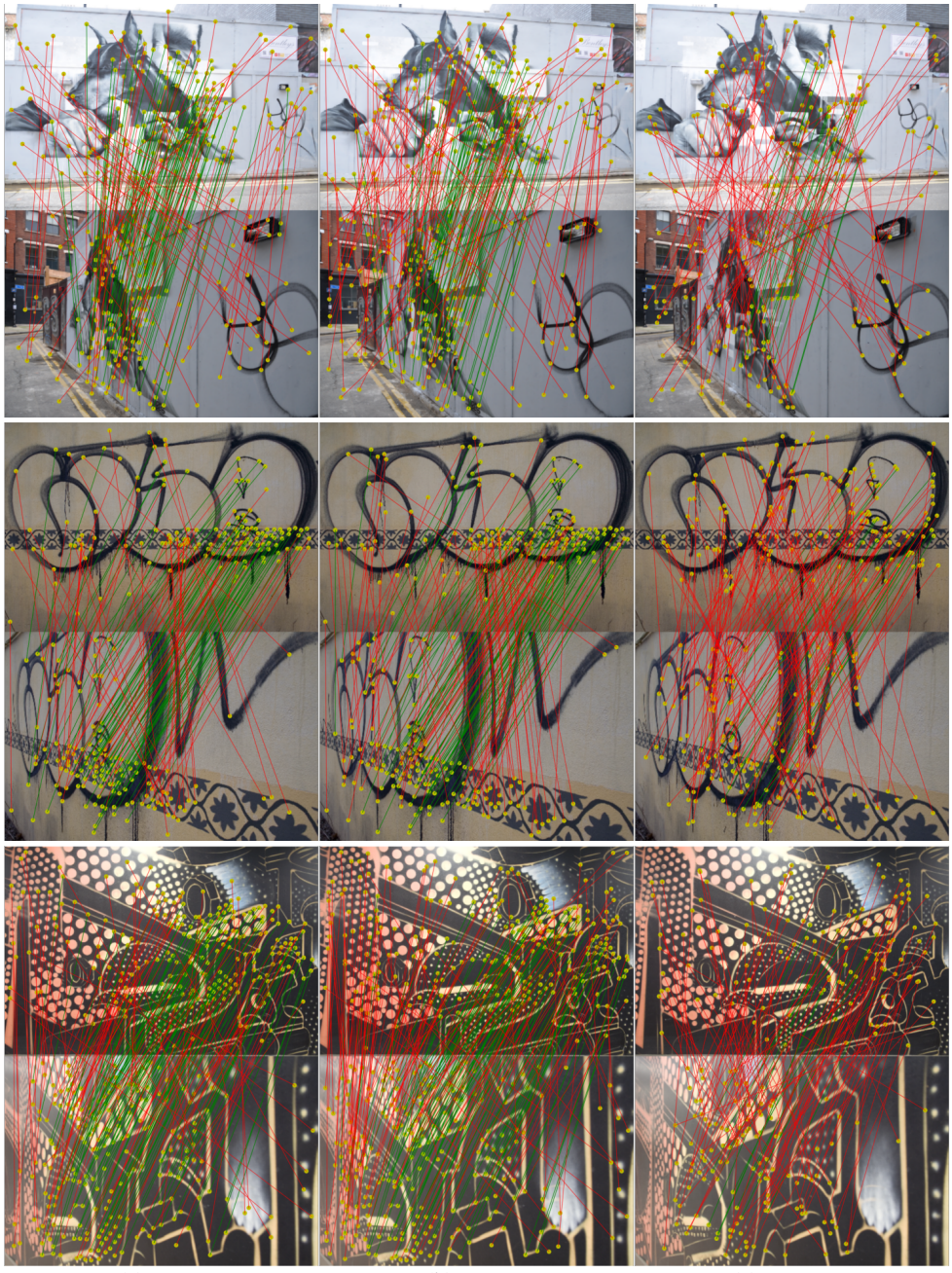


Figure 4. **Similarity maps with respect to a keypoint under rotation.** We compare one descriptor about the red keypoint from the source image at 0° to all other descriptors extracted across the rotated images, with yellow circles representing corresponding keypoints. For better visibility, we visualize the top 5 pixels with the highest similarity to the keypoints.



(a) ours (b) GIFT (c) LF-Net

Figure 5. **Visualization of predicted matches in the ER dataset [14].** We use a maximum of 1,500 keypoints for matching by the mutual nearest neighbor algorithm. We measure the correctness at a three-pixel threshold. The green lines denote the correct matches, and the red lines denote the incorrect matches.



(a) ours (b) GIFT (c) LF-Net

Figure 6. **Visualization of the predicted matches in HPatches viewpoint variations.** We use a maximum of 1,500 keypoints, the mutual nearest neighbor matcher, and a three-pixel threshold for correctness. In this experiment, we use the rotation-equivariant WideResNet16-8 (ReWRN) backbone, which is 'ours†' in table 4 of the main paper.

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. 1, 5

[2] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 1

[3] Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 9145–9156, 2019. 1

[4] Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016. 1

[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, 4, 7

[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. 2, 3

[7] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[8] 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

[9] Axel Barroso Laguna and Krystian Mikolajczyk. Key. net: Keypoint detection by handcrafted and learned cnn filters revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[10] Jongmin Lee, Yoonwoo Jeong, and Minsu Cho. Self-supervised learning of image scale and orientation. In *31st British Machine Vision Conference 2021, BMVC 2021, Virtual Event, UK*. BMVA Press, 2021. 5

[11] Jongmin Lee, Byungjin Kim, and Minsu Cho. Self-supervised equivariant learning for oriented keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4847–4857, 2022. 5

[12] Kunhong Li, Longguang Wang, Li Liu, Qing Ran, Kai Xu, and Yulan Guo. Decoupling makes weakly supervised local feature better. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15838–15848, 2022. 3

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[14] 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, 4, 5, 9

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

[16] 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. 3

[17] 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. 3, 5, 7

[18] Rémi Pautrat, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *European Conference on Computer Vision*, pages 707–724. Springer, 2020. 1, 2, 3, 6

[19] 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. 2, 3

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 4

[21] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 4

[22] Xuelun Shen, Cheng Wang, Xin Li, Zenglei Yu, Jonathan Li, Chenglu Wen, Ming Cheng, and Zijian He. Rf-net: An end-to-end image matching network based on receptive field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8132–8140, 2019. 3, 5, 7

[23] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 4

[24] 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*, 33:7401–7412, 2020. 2

[25] Yurun Tian, Axel Barroso Laguna, Tony Ng, Vassileios Balntas, and Krystian Mikolajczyk. Hynet: Learning local descriptor with hybrid similarity measure and triplet loss. In *NeurIPS*, 2020. 3

[26] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 32:14334–14345, 2019. 1, 3, 5

[27] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 849–858, 2018. 1

[28] Hao Zhou, Torsten Sattler, and David W Jacobs. Evaluating local features for day-night matching. In *European Conference on Computer Vision*, pages 724–736. Springer, 2016. 1, 2, 6