Learning Rotation-Equivariant Features for Visual Correspondence

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

Abstract

Extracting discriminative local features that are invariant to imaging variations is an integral part of establishing correspondences between images. In this work, we introduce a self-supervised learning framework to extract discriminative rotation-invariant descriptors using group-equivariant CNNs. Thanks to employing group-equivariant CNNs, our method effectively learns to obtain rotation-equivariant features and their orientations explicitly, without having to perform sophisticated data augmentations. The resultant features and their orientations are further processed by group aligning, a novel invariant mapping technique that shifts the group-equivariant features by their orientations along the group dimension. Our group aligning technique achieves rotation-invariance without any collapse of the group dimension and thus eschews loss of discriminability. The proposed method is trained end-to-end in a self-supervised manner, where we use an orientation alignment loss for the orientation estimation and a contrastive descriptor loss for robust local descriptors to geometric/photometric variations. Our method demonstrates state-of-the-art matching accuracy among existing rotation-invariant descriptors under varying rotation and also shows competitive results when transferred to the task of keypoint matching and camera pose estimation.

1. Introduction

Extracting local descriptors is an essential step for visual correspondence across images, which is used for a wide range of computer vision problems such as visual localization [29, 47, 48], simultaneous localization and mapping [7, 8, 39], and 3D reconstruction [1, 16, 17, 49, 66]. To establish reliable visual correspondences, the properties of invariance and discriminativeness are required for local descriptors; the descriptors need to be invariant to geometric/photometric variations of images while being discriminative enough to distinguish true matches from false ones. Since the remarkable success of deep learning for visual recognition, deep neural networks have also been adopted to learn local descriptors, showing enhanced performances on visual correspondence [44, 45, 64]. Learning *rotation*-invariant local descriptors, however, remains challenging; the classical techiniques [11, 27, 46] for rotation-invariant descriptors, which are used for shallow gradient-based feature maps, cannot be applied to feature maps from standard deep neural networks, in which rotation of input induces unpredictable feature variations. Achieving rotation invariance without sacrificing disriminativeness is particularly important for local descriptors as rotation is one of the most frequent imaging variations in reality.

In this work, we propose a self-supervised approach to obtain rotation-invariant and discriminative local descriptors by leveraging rotation-equivariant CNNs. First, we use group-equivariant CNNs [60] to jointly extract rotation-equivariant local features and their orientations from an image. To extract reliable orientations, we use an orientation alignment loss [21, 23, 63], which trains the network to predict the dominant orientation robustly against other imaging variations, including illumination or viewpoint changes. Using group-equivariant CNNs enables the local features to be empowered with explicitly encoded rotation equivariance without having to perform rigorous data augmentations [58, 60]. Second, to obtain discriminative rotation-invariant descriptors from rotation-equivariant features, we propose group-aligning that *shifts* the group-equivariant features by their dominant orientation along their group dimension. Conventional methods to yield invariant features from group-equivariant features collapse the group dimension by group-pooling, *e.g.,* max-pooling or bilinear-pooling [26], resulting in a drop in feature discriminability and quality. In contrast, our group-aligning preserves the group dimension, achieving rotation-invariance while eschewing loss of discriminability. Furthermore, by preserving the group dimension, we can obtain multiple descriptors by performing group-aligning using multiple orientation candidates, which improves the matching performance by compensating for potential errors in dominant orientation prediction. Finally, we evaluate our rotation-invariant descriptors against existing local descriptors, and

our group-aligning scheme against group-pooling methods on various image matching benchmarks to demonstrate the efficacy of our method.

The contribution of our paper is fourfold:

- We propose to extract discriminative rotation-invariant local descriptors to tackle the task of visual correspondence by utilizing rotation-equivariant CNNs.
- We propose group-aligning, a method to shift a group-equivariant descriptor in the group dimension by its dominant orientation to obtain a rotation-invariant descriptor without having to collapse the group information to preserve feature discriminability.
- We use self-supervisory losses of orientation alignment loss for orientation estimation, and a contrastive descriptor loss for robust local descriptor extraction.
- We demonstrate state-of-the-art performances under varying rotations on the Roto-360 dataset and show competitive transferability on the HPatches dataset [2] and the MVS dataset [53].

2. Related work

Classical invariant local descriptors. Classical methods to extract invariant local descriptors first aggregate image gradients to obtain a rotation-equivariant representation, *i.e.*, histogram, from which the estimated dominant orientation is subtracted to obtain rotation-invariant features [27, 46]. Several studies [4, 11, 59] suggest extracting local descriptors by invariant mapping of the order-based gradient histogram of a patch. However, these classical methods for shallow gradient-based feature maps cannot be applied to deep feature maps from standard neural networks, in which rotation induces unpredictable feature variations. Therefore, we propose a deep end-to-end pipeline to obtain orientation-normalized local descriptors by utilizing rotation-equivariant CNNs [60] with additional losses.

Learning-based invariant local descriptors. A branch of learning-based methods learns to obtain invariant local descriptors in an explicit manner. GIFT [26] constructs group-equivariant features by rotating or rescaling the images, and then collapses the group dimension using bilinear pooling to obtain invariant local descriptors. However, their groups are limited to non-cyclic discrete rotations ranging from $-90°$ to $90°$. Furthermore, their reliance on data augmentation implies a lower sampling efficiency compared to group-equivariant networks. LISRD [42] jointly learns meta descriptors with different levels of regional variations and selects the most appropriate level of invariance given the context. Another branch of learning methods aims to learn the invariance implicitly using descriptor similarity losses from the image pair using camera pose or homography supervision. These methods are either patch-based [10, 36, 54, 56] or image-based [8, 9, 22, 24, 28, 37, 40, 44, 50, 57]. While these methods may be robust to rotation, they cannot be said to be equivariant or invariant to rotation. We construct group-equivariant local features using the steerable networks [60], which explicitly encodes cyclic rotational equivariance to the features without having to rely on data augmentation. We can then yield rotation-invariant features by group-aligning that shifts the group-equivariant features along the group dimension by their dominant orientations, preserving feature discriminability.

Equivariant representation learning. There has been a constant pursuit to learn equivariant representations by explicitly incorporating group equivariance into the model architecture design [30–32, 51, 60, 65]. For example, G-CNNs [6] use group equivariant convolutions that reduce sample complexity by exploiting symmetries on discrete isometric groups; SFCNNs [61] and H-Nets [62] extract features from more diverse groups and continuous domains by using harmonics as filters. There are also studies that focus on scale-equivariant representation learning [3, 21, 52]. [12, 18, 23, 38, 43] leverage equivariant neural networks to tackle vision tasks *e.g.,* keypoint detection. In this work, we also propose to use equivariant neural networks to facilitate the learning of discriminative rotation-invariant descriptors. We guide the readers to section 1 of the supplementary material for a brief introduction to group equivariance.

3. Rotation-equivariant features, Rotation-invariant descriptors

In this section, we first draw the line between the terms *feature* and *descriptor* which will be used throughout this paper. The goal of our work is to learn to extract rotation-equivariant local *features* from our rotation-equivariant backbone network, and then to align them by their dominant orientation to finally yield rotation-invariant *descriptors*. In the subsequent subsections, we elaborate on the process of rotation-equivariant feature extraction from steerable CNNs (Sec. 3.1), assignment of equivariant features to keypoints (Sec. 3.2), how group-aligning is performed to yield rotation-invariant yet discriminative descriptors (Sec. 3.3), how we formulate our orientation alignment loss (Sec.3.4) and contrastive descriptor loss (Sec.3.5) to train our network to extract descriptors which are robust to not only rotation but also other imaging transformations, and finally how we obtain scale-invariant descriptors at test time using image pyramids (Sec.3.6). Figure 1 shows the overall architecture of our method.

3.1. Rotation-equivariant feature extraction

As the feature extractor, we use ReResNet18 [12], which has the same structure as ResNet18 [15] but is constructed using rotation-equivariant convolutional layers [60]. The

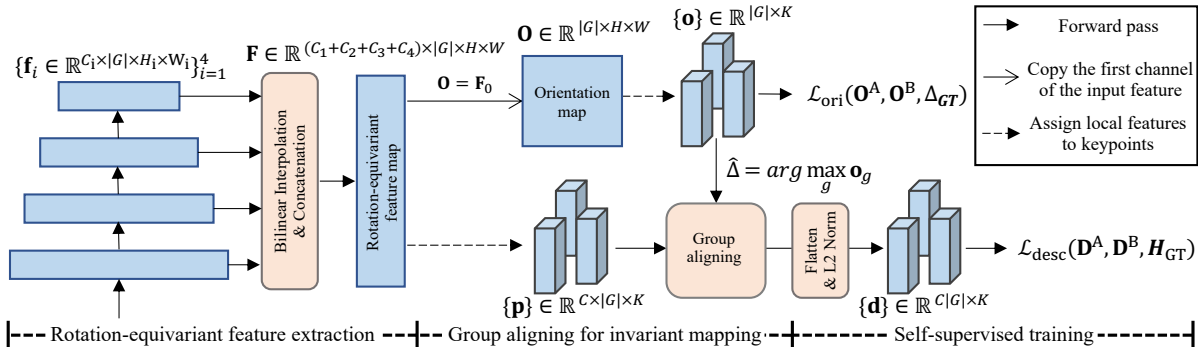


Figure 1. **Overview of the proposed pipeline.** An input image is forwarded through the equivariant networks to yield equivariant feature maps from multiple intermediate layers, encoding both low-level geometry and high-level semantic information. The feature maps are bilinearly interpolated to have equal spatial dimensions to be concatenated together. We use the first channel of the feature map $\mathbf{F}$ as the orientation histogram map $\mathbf{O}$ to predict the dominant orientations, which are used to shift the group-equivariant representation along the group dimension to yield discriminative rotation-invariant descriptors. To learn to extract accurate dominant orientation $\hat{\theta}$, we use the orientation alignment loss $\mathcal{L}^{\text{ori}}$. To obtain descriptors robust to illumination and geometric changes, we use a contrastive descriptor loss $\mathcal{L}^{\text{desc}}$ using the ground-truth homography $\mathcal{H}_{\text{GT}}$.

layer acts on a cyclic group G_N and is equivariant for all translations and N discrete rotations. At the first layer, the scalar field of the input image is lifted to the vector field of the group representation [60]. We leverage feature pyramids from the intermediate layers of the ReResNet18 backbone to construct output features as follows:

$$\mathbf{F} = \bigoplus_{i \in l} \eta(\mathbf{f}_i), \quad \mathbf{f}_i = [\Pi_{j=1}^{i} L_j](I), \tag{1}$$

where $\mathbf{f}_i \in \mathbb{R}^{C_i \times |G| \times H_i \times W_i}$ is an intermediate feature from L_i, L_i is the i-th layer of the equivariant network, η denotes bilinear interpolation to $H \times W$, and $\bigoplus$ denotes concatenation along the C dimension. We utilize the multi-layer feature maps to exploit the low-level geometry information and high-level semantics in the local descriptors [13,19,35]. The output features $\mathbf{F} \in \mathbb{R}^{C \times |G| \times H \times W}$ contains rotation-equivariant features with multiple layers containing different semantics and receptive fields. We set $H = H_1$ and $W = W_1$, which are $\frac{1}{2}$ of the input image size.

3.2. Assigning local features to keypoints

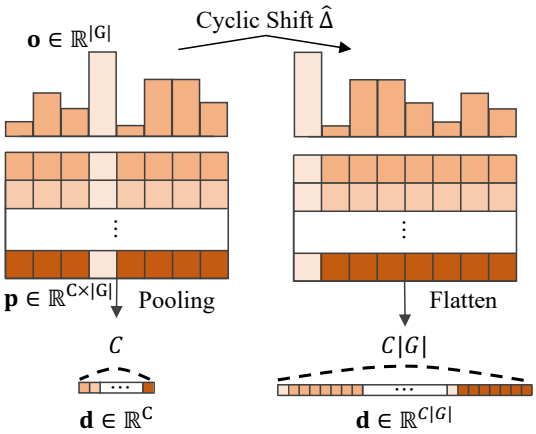
During training, we extract K keypoints from the source image using Harris corner detection [14]. We then use the ground-truth homography $\mathcal{H}_{\text{GT}}$ to obtain ground-truth keypoint correspondences. Also, we allocate a local feature $\mathbf{p} \in \mathbb{R}^{C \times |G| \times K}$ to each keypoint, using the interpolated location of the equivariant feature map $\mathbf{F}$. We experiment our descriptor with SIFT [27], LF-Net [40], SuperPoint [8], and KeyNet [20] as the keypoint detector during inference time.

3.3. Group aligning for invariant mapping

To transform the rotation-equivariant feature to a rotation-invariant descriptor, we propose group aligning, an operation to shift the group-equivariant feature in the G-dimension using the dominant orientation $\hat{\theta}$. Unlike existing methods that use group pooling, *e.g.,* average pooling or max pooling, which collapses the group dimension, group aligning preserves the rich group information. Figure 2 illustrates the difference between group pooling and group aligning on an equivariant representation.

Estimating the dominant orientation and the shifting value. We obtain the orientation histogram map $\mathbf{O} \in \mathbb{R}^{|G| \times H \times W} = \mathbf{F}_0$ by selecting the first channel of the rotation-equivariant tensor $\mathbf{F}$ as an orientation histogram map. Note that the first channels of each group action are simultaneously used as the channels of the descriptors and to construct the orientation histogram. The histogram-based representation of $\mathbf{O}$ provides richer information than directly regressing the dominant orientation, as the orientation histogram enables predicting multiple (*i.e.,* top-k) candidates as the dominant orientation. We first select an orientation vector $\mathbf{o} \in \mathbb{R}^{|G|}$ of a keypoint from the orientation histogram map $\mathbf{O}$ using the coordinates of the keypoint. Next, we estimate the dominant orientation value $\hat{\theta}$ from the orientation vector $\mathbf{o}$ by selecting the index of the maximum score, $\hat{\theta} = \frac{360}{|G|} \arg\max_g \mathbf{o}$. Using the dominant orientation value $\hat{\theta}$, we obtain the shifting value $\hat{\Delta} = \frac{|G|}{360}\hat{\theta}$ in G-dim. At training time, we use the ground-truth rotation θ_{GT} instead of the predicted dominant orientation value $\hat{\theta}$ to generate the shifting value Δ_{GT}.

Group aligning. Given a keypoint-allocated feature tensor $\mathbf{p} \in \mathbb{R}^{C \times |G|}$ from the equivariant representation $\mathbf{F}$, we obtain the rotation-invariant local descriptor $\mathbf{d} \in \mathbb{R}^{C|G|}$ by group aligning using Δ. After computing the dominant orientation $\hat{\theta}$ and the shifting value $\hat{\Delta}$ from $\mathbf{o}$, we obtain the orientation-normalized descriptor $\mathbf{d}' \in \mathbb{R}^{C|G|}$ by shifting $\mathbf{p}$



(a) Group pooling (b) Group aligning

Figure 2. **Difference between group pooling and group aligning.** In group pooling, the group dimension is collapsed to yield an invariant descriptor ($\mathbb{R}^{C\times|G|} \rightarrow \mathbb{R}^{C}$). In group aligning, the entire feature is cyclically shifted in the group dimension to obtain an invariant descriptor ($\mathbb{R}^{C\times|G|} \rightarrow \mathbb{R}^{C|G|}$) while preserving the group information and discriminability.

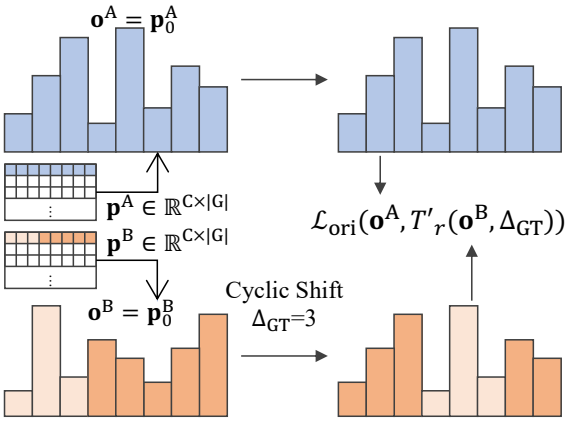


Figure 3. **Illustration of orientation alignment loss.** Given two rotation-equivariant tensors $\mathbf{p}^{\mathrm{A}}, \mathbf{p}^{\mathrm{B}} \in \mathbb{R}^{C\times|G|}$ obtained from two different rotated versions of the same image, we apply cyclic shift on one of the descriptors in the group dimension using the GT difference in rotation. The orientation alignment loss supervises the output orientation vectors of the two descriptors to be the same.

in the G-dimension by $-\hat{\Delta}$ and flattening the descriptor to a vector. We use cyclic shifting in consideration of the cyclic property of rotation. We finally obtain the L2-normalized descriptor $\mathbf{d}$ from the orientation-normalized descriptor $\mathbf{d}'$, such that $||\mathbf{d}||^2 = 1$. Formally, this process can be defined as:

$$\begin{aligned} \mathbf{p}'_{:,i} = T'_r(\mathbf{p}_{:,i}, \hat{\Delta}) = \mathbf{p}_{:,(i+\hat{\Delta}) \bmod |G|}, \\ \mathbf{d}'_{|G|i:|G|(i+1)} = \mathbf{p}'_i, \\ \mathbf{d} = \frac{\mathbf{d}'}{||\mathbf{d}'||_2}, \end{aligned} \tag{2}$$

where T'_r is shifting operator in vector space, and $\mathbf{p}'$ is a group-aligned descriptor before flattening. This shifting by $\hat{\Delta}$ aligns all the descriptors in the direction of their dominant orientations, creating orientation-normalized descriptors. This process is conceptually similar to subtracting the dominant orientation value of the orientation histogram in the classical descriptor SIFT [27], but we apply this concept to the equivariant neural features. The proposed group aligning preserves the group information, so our invariant descriptors have more representative power than the existing group-pooled descriptors which collapse the group dimension for invariance.

3.4. Orientation alignment loss

To learn to obtain the dominant orientations from the orientation vectors, we use an orientation alignment loss [21, 23, 63] to supervise the orientation histograms in $\mathbf{O}$ to be rotation equivariant under the photometric/geometric transformations. Figure 3 shows the illustration of orientation alignment loss. The cyclic shift of an orientation histogram map at the training time is formulated as follows:

$$T'_r(\mathbf{O}_i, \Delta_{\mathrm{GT}}) = \mathbf{O}_{(i+\Delta_{\mathrm{GT}}) \bmod |G|}, \tag{3}$$

where $\Delta_{\mathrm{GT}} = \frac{|G|}{360}\theta_{\mathrm{GT}}$ is the shifting value calculated from the ground-truth rotation θ_{GT}. We formulate the orientation alignment loss in the form of a cross-entropy as follows:

$$\begin{aligned} \mathcal{L}^{\mathrm{ori}}(\mathbf{O}^{\mathrm{A}}, \mathbf{O}^{\mathrm{B}}, \Delta_{\mathrm{GT}}) = \\ -\sum_{k\in K}\sum_{g\in G} \sigma(\mathbf{O}^{\mathrm{A}}_{g,k}) \log(\sigma(T'_r(\mathbf{O}^{\mathrm{B}}_{g,k}, \Delta_{\mathrm{GT}}))), \end{aligned} \tag{4}$$

where $\mathbf{O}^{\mathrm{A}}$ is the source orientation histogram map and $\mathbf{O}^{\mathrm{B}}$ is the target orientation histogram map obtained from a synthetically warped source image, σ is a softmax function applied to the G-dimension of the orientation histogram map to represent the orientation vector as a probability distribution for the cross-entropy loss to be applicable. Using Equation 4, the network learns to predict the characteristic orientations robustly against different imaging variations, such as photometric transformations and geometric transformations beyond rotation, as these transformations cannot be handled by equivariance to discrete rotations alone. Note that it is not straightforward to define the characteristic orientation of a keypoint to provide strong supervision. However, we facilitate the learning of characteristic orientations by formulating it as a self-supervised learning framework, leveraging the known relative orientation between two keypoint orientation histogram maps obtained from differently rotated versions of the same image.

	MMA	pred.
	@1px	
Align	**97.54**	**84.90**
Avg	33.72	33.72
Max	57.92	57.92
None	23.97	23.97
Bilinear	43.60	26.42

Table 1. **Evaluation with GT keypoint pairs on Roto-360 without training.** 'Align' uses GT rotation difference to apply group-aligning to demonstrate the upper-bound. 'None' does not use pooling nor aligning, demonstrating the lower-bound. We use an average of 111 keypoint pairs extracted using SuperPoint [8].

	MMA			pred.
	@10px	@5px	@3px	
Align	**93.08**	**91.35**	**90.18**	688.3
Avg	85.84	82.12	81.05	**705.9**
Max	82.61	78.00	77.79	686.0
None	19.68	18.81	18.57	349.1
Bilinear	42.69	41.03	40.51	332.5

Table 2. **Evaluation with predicted keypoint pairs on Roto-360 with training.** 'Max' and 'Avg' collapses the group dimension of the features through max pooling or average pooling. 'pred.' denotes the average number of predicted matches. We use an average of 1161 keypoint pairs extracted using SuperPoint [8].

3.5. Contrastive descriptor loss

We propose to use a descriptor similarity loss motivated by contrastive learning [5] to further empower the descriptors to be robust against variations apart from rotation, *e.g.,* illumination or viewpoint. The descriptor loss is formulated in a contrastive manner as follows:

$$\mathcal{L}^{\text{desc}}(\mathbf{D}^{\text{A}}, \mathbf{D}^{\text{B}}) = \sum_{(\mathbf{d}_i^{\text{A}}, \mathbf{d}_i^{\text{B}}) \in (\mathbf{D}^{\text{A}}, \mathbf{D}^{\text{B}})} -\log \frac{\exp(\text{sim}(\mathbf{d}_i^{\text{A}}, \mathbf{d}_i^{\text{B}})/\tau)}{\sum_{k \in K \setminus i} \exp(\text{sim}(\mathbf{d}_i^{\text{A}}, \mathbf{d}_k^{\text{B}}))/\tau)}, \tag{5}$$

where sim is cosine similarity and τ is the softmax temperature. Unlike the triplet loss with one hard negative sample, the contrastive loss can optimize the distance for all negative pairs. This contrastive loss with InfoNCE [41] maximizes the mutual information between the encoded features and effectively reduces the low-level noise. Our overall self-supervised loss is formulated as $\mathcal{L} = \alpha\mathcal{L}^{\text{ori}} + \mathcal{L}^{\text{desc}}$, where α is a balancing term.

3.6. Scale robustness

While we employ a rotation-equivariant network, it does not ensure that the descriptors are robust to scale changes. Thus, at inference time, we construct an image pyramid using a scale factor of $2^{1/4}$ from a maximum of 1,024 pixels to a minimum of 256 pixels as in R2D2 [44]. After constructing the scale-wise descriptors $\in \mathbb{R}^{S \times C|G| \times K}$ with S varying scales, we finally generate the scale-invariant local descriptors $\in \mathbb{R}^{C|G| \times K}$ by max-pooling in the scale dimension inspired by scale-space maxima as in SIFT [27,33], for improved robustness to scale changes.

4. Experiment

Implementation details. We use rotation-equivariant ResNet-18 (ReResNet-18) [12] implemented using the rotation-equivariant layers of $E(2)$-CNN [60] as our backbone. We remove the first maxpool layer to preserve the spatial size, so that the spatial resolution of the rotation-equivariant feature $\mathbf{F}$ is $H = \frac{H'}{2}$ and $W = \frac{W'}{2}$, where H' and W' are the height and width of an input image. We use 16 for the order of cyclic group G. We use a batch size of 8, a learning rate of 10^{-4}, and a weight decay of 0.1. We train our model for 12 epochs with 1,000 iterations using a machine with an Intel i7-8700 CPU and an NVIDIA GeForce RTX 3090 GPU. We use the temperature τ of $\mathcal{L}^{\text{desc}}$ as 0.07. The loss balancing factor α is 10. The final output descriptor size is 1,024, with $C = 64$, $|G| = 16$. We use SuperPoint [8] as the keypoint detector to evaluate our method except Table 4. For all descriptors, we use the mutual nearest neighbour matcher to predict the correspondences.

4.1. Datasets and metrics

We use a synthetic training dataset to train our model in a self-supervised manner. We evaluate our model on the Roto-360 dataset and show the transferability on real image benchmarks, *i.e.,* HPatches [2] and MVS [53] datasets.

Training dataset. We generate a synthetic dataset for self-supervised training from the MS-COCO dataset [25]. We warp images with random homographies for geometric robustness and transform the colors by jitter, noise, and blur for photometric robustness. As we need the ground-truth rotation θ_{GT} for our orientation alignment loss, we decompose the synthetic homography $\mathcal{H}$ as follows: $\theta_{\text{GT}} = \arctan(\frac{\mathcal{H}_{21}}{\mathcal{H}_{11}})$, where we assume that a 3×3 homography matrix $\mathcal{H}$ with no significant tilt can be approximated to an affine matrix. We sample $K = 512$ keypoints for an image using Harris corner detector [14], obtaining 512 corresponding keypoint pairs for each image pair using homography and rotation. Note that this dataset generation protocol is the same as that of GIFT [26] for a fair comparison.

Roto-360 is an evaluation dataset that consists of 360 image pairs with in-plane rotation ranging from $0°$ to $350°$ at $10°$ intervals, created using ten randomly sampled images from HPatches [2]. Roto-360 is more suitable to evaluate the rotation invariance of our descriptors, as the extreme rotation (ER) dataset [26] only covers $180°$, and includes photometric variations. We use mean matching accuracy (MMA) as the evaluation metric with pixel thresholds of 3/5/10 pixels

Method	MMA			pred.	total.
	@10px	@5px	@3px		
SIFT [27]	78.86	78.59	78.23	774.1	1500.0
ORB [46]	86.78	85.29	78.73	607.6	1005.2
SuperPoint [8]	22.85	22.10	21.83	462.6	1161.0
LF-Net [40]	75.05	74.30	72.61	386.7	1024.0
RF-Net [50]	15.64	15.18	14.58	1602.5	5000.0
D2-Net [9]	15.56	9.30	5.21	386.9	1474.5
R2D2 [44]	15.80	14.97	13.50	197.9	1500.0
GIFT [26]	42.35	42.05	41.59	589.2	1161.0
LISRD [42]	16.96	16.04	15.64	323.6	1781.1
ASLFeat [28]	19.34	16.38	13.13	1366.9	6764.2
DISK [57]	13.22	12.43	12.04	359.1	2048.0
PosFeat [24]	13.76	11.79	9.82	717.2	7623.5
ours	93.08	91.35	90.18	688.3	1161.0
ours*	**94.35**	**92.82**	**91.69**	1333.0	2340.4

Table 3. **Comparison to existing local descriptors on Roto-360.** We use mutual nearest matching for all methods to establish matches between images. 'total.' and 'pred.' denotes the average number of detected keypoints and predicted matches, respectively. 'ours*' denotes selecting multiple candidate descriptors based on the ratio of max value in the orientation histogram. We use SuperPoint keypoint detector [8] same to the GIFT descriptor [26].

Det.	Desc.	MMA			pred.	total.
		@10px	@5px	@3px		
SIFT [27]	SIFT [27]	78.86	78.59	78.23	774.1	1500
	GIFT [26]	37.97	36.82	36.09	531.2	1500
	ours	84.67	79.85	77.96	558.3	1500
	ours*	**84.91**	**80.09**	**78.18**	759.8	2219
LF-Net [40]	LF-Net [40]	75.05	**74.30**	**72.61**	386.7	1024
	GIFT [26]	35.56	33.82	32.29	426.3	1024
	ours	79.90	71.63	67.39	431.8	1024
	ours*	**80.32**	71.99	67.62	591.4	1503
SuperPoint [8]	SuperPoint [8]	22.85	22.10	21.83	462.6	1161
	GIFT [26]	42.35	42.05	41.59	589.2	1161
	ours	93.08	91.35	90.18	688.3	1161
	ours*	**94.35**	**92.82**	**91.69**	1333	234
KeyNet [20]	HyNet [55]	24.43	22.82	20.64	288.7	995
	GIFT [26]	34.08	32.31	29.17	275.7	995
	ours	72.95	61.36	41.33	257.2	995
	ours*	**72.48**	**60.69**	**40.95**	356.6	1484

Table 4. **Comparison to existing local descriptors when using the same keypoint detector on Roto-360.** Results in bold indicate the best result, and underlined results indicate the second best.

and the number of predicted matches following [9, 34].

HPatches [2] has 57 scenes with illumination variations and 59 scenes with viewpoint variations. Each scene contains five image pairs with ground-truth planar homography. We use the same evaluation metrics to Roto-360 to show the transferability of our local descriptors.

MVS dataset [53] has six image sequences of outdoor scenes with GT camera poses. We evaluate the relative pose estimation accuracy at $5°/10°/20°$ angular difference thresholds.

4.2. Comparison to other invariant mappings

Table 1 compares group aligning to various group pooling methods on the Roto-360 dataset using ground-truth keypoint pairs, *i.e.,* no keypoint deviation, without training. The purpose is to compare the invariant mapping operations only while keeping the backbone network and the number of keypoints fixed. We use Δ_{GT} to shift the equivariant features, and group aligning shows almost perfect keypoint correspondences with 97.54% matching accuracy. Group pooling, such as max pooling or average pooling, significantly reduces discriminative power compared to group aligning. The results show that group aligning shows the best results, proving that leveraging the full group-equivariant features instead of collapsing the groups shows higher discriminability. Note that the bilinear pooling [26] does not guarantee the rotation-invariant matching.

Table 2 compares the proposed group aligning to the existing group pooling methods on the Roto-360 dataset, this time with predicted keypoint pairs and with training. Note that while other methods are trained only with $\mathcal{L}^{\mathrm{desc}}$, our method is trained also with $\mathcal{L}^{\mathrm{ori}}$ to facilitate group aligning. While the number of predicted matches is the highest for average pooling, the MMA results are significantly higher for group aligning, which shows group-aligned descriptors have a higher precision. Overall, incorporating group aligning demonstrates the best results in terms of MMA compared to average pooling, max pooling or bilinear pooling [26]. Note that pooling or aligning the group-equivariant features to obtain invariant descriptors shows consistent improvements over not pooling nor aligning the group-equivariant features.

4.3. Comparison to existing local descriptors

Table 3 shows the matching accuracy compared to existing local descriptors on the Roto-360 dataset. We evaluate the descriptors using their own keypoint detectors [8, 9, 27, 39, 40, 44, 50], or combined with off-the-shelf detectors [24, 26, 42]. While the classical methods [27, 46] achieve better matching accuracy than the existing learning-based methods, our method achieves the best results overall. This is because the learning-based methods learn only a limited degree of invariance in a data-driven manner without guaranteeing full invariance to rotation by design.

Table 4 shows the performance of our method in comparison to existing local descriptors when using the same keypoint detector, where our method shows consistent performance improvement. In particular, our rotation-invariant descriptor shows consistently higher matching accuracy than GIFT [26], which is a representative learning-based group-invariant descriptor. While our model shows a lower MMA than the LF-Net [40] descriptor when using the LF-Net detector at 5px and 3px thresholds, we conjecture that this is due to the better integrity of the detector and descriptor of LF-Net due to their joint training scheme.

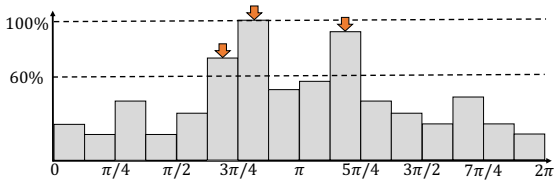


Figure 4. **An example of multiple descriptor extraction.** The distribution is an orientation histogram $\mathbf{o} \in \mathbb{R}^{16}$, and the scores are confidence values for each bin from group-equivariant features. Arrows indicate the orientation candidates for multiple descriptor extraction. The example shows selecting three orientations to obtain three candidate descriptors for a feature point, which is possible as we predict a score for each orientation.

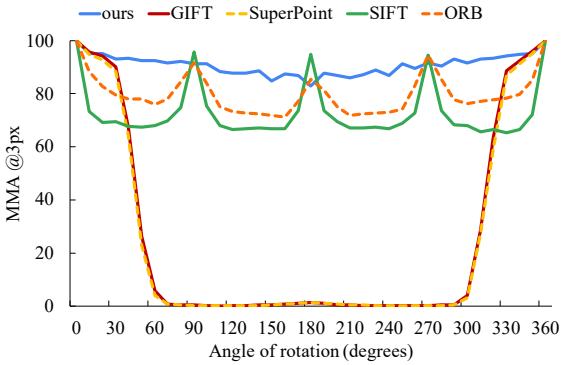


Figure 5. **Matching accuracies according to varying degree of rotations on Roto-360.**

These results show that our descriptors obtained using the proposed group aligning show the highest matching accuracy under rotation changes compared to existing methods. The improvement of our method is also attributed to the usage of rotation-equivariant networks, which have a higher sampling efficiency, *i.e.,* do not require intensive rotation augmentations to learn rotation invariance.

Multiple descriptor extraction using orientation candidates. Group aligning can extract multiple descriptors with different alignments by using multiple orientation candidates, denoted by 'ours*', whose scores are at least 60% of the maximum score in the orientation histogram. When there is a single keypoint position with k descriptors that are differently aligned, we treat it as if there are k detected keypoints. Multiple descriptor extraction compensates for incorrect orientation predictions and further enhances matching accuracy. Figure 4 illustrates an example of multiple descriptor extraction with a score ratio threshold of 0.6.

Consistency of matching accuracy with respect to rotation changes. Figure 5 illustrates how the matching accuracy changes with respect to varying degrees of rotation. Our method shows the highest consistency, proving the enhanced invariance of descriptors obtained using group aligning against different rotations. While MMA of SIFT [27] and ORB [46] are high at the upright rotations, they tend to fluctuate significantly with varying rotations. The existing learning-based group-invariant descriptor, GIFT [26], fails to find correspondences beyond 60°.

4.4. Transferability to real image benchmarks

Table 5 shows the matching performance of local descriptors on HPatches illumination/viewpoint [2] and pose estimation [53]. Our model shows the highest performance overall on the HPatches dataset. The performance gain of ours becomes smaller compared to the Roto-360 dataset due to the absence of extreme rotations in HPatches. While GIFT shows a higher performance under illumination changes that only contain identity mappings, ours†, which uses a larger backbone network (ReWRN), improves

Method	HP-all		HP-illu		HP-view		Pose		
	@5px	@3px	@5px	@3px	@5px	@3px	20°	10°	5°
SIFT [27]	51.36	46.32	49.08	44.62	53.57	47.96	0.02	0.00	0.00
ORB [46]	52.22	47.40	50.85	46.29	53.55	48.47	0.06	0.00	0.00
SuperPoint [8]	69.71	61.75	74.63	67.53	64.96	56.17	0.20	0.07	0.01
LF-Net [40]	56.45	52.22	62.21	57.63	50.88	47.00	0.06	0.03	0.01
RF-Net [50]	59.08	54.42	61.63	57.46	56.62	51.49	0.10	0.04	0.01
D2-Net [9]	50.18	32.54	63.80	44.09	37.02	21.38	0.11	0.05	0.01
GIFT [26]	76.03	67.31	**79.71**	**71.89**	72.48	62.88	**0.60**	0.28	0.09
LISRD [42]	62.16	56.12	70.09	63.64	54.50	48.85	0.05	0.02	0.00
ours$_{\text{avgpool}}$	64.10	57.94	62.28	56.27	65.85	59.55	0.27	0.10	0.05
ours$_{\text{maxpool}}$	61.57	55.81	59.66	53.91	63.42	57.64	0.27	0.11	0.03
ours$_{\text{bilinearpool}}$ [26]	45.59	41.90	45.13	41.57	46.03	42.22	0.35	0.17	0.09
ours$_{\text{bilinearpool}}$† [26]	58.72	53.77	57.32	52.67	60.06	54.83	0.24	0.11	0.03
ours$_{\text{groupalign}}$	70.69	63.42	70.39	62.88	70.97	63.95	0.58	0.26	0.12
ours$_{\text{groupalign}}$*	73.92	66.37	73.13	65.33	74.69	67.38	0.56	0.30	0.12
ours$_{\text{groupalign}}$†	**78.00**	**69.70**	77.94	69.35	**78.06**	**70.03**	0.56	**0.33**	**0.14**

Table 5. **Evaluation with predicted keypoint pairs on real image benchmarks.** The first group of methods includes existing local feature extraction methods. The second group of methods includes comparisons to other group pooling methods by replacing our group aligning with them. 'ours*' denotes the extraction of multiple descriptors using the orientation candidates, whose scores are at least 60% of the maximum score in the orientation histogram. 'ours†' denotes our method using the rotation-equivariant WideResNet16-8 (ReWRN) backbone for feature extraction. We use SuperPoint [8] keypoint detector to evaluate ours.

matching accuracy by 7.15%p at 3px and 5.58%p at 5px, and ours* improves by 4.5%p at 3px, 2.21%p at 5px under viewpoint changes compared to GIFT [26]. It should be noted that the core difference between ours$_{\text{bilinearpool}}$ and GIFT is the usage of explicit rotation-equivariant CNNs [60], which clearly shows that bilinear pooling is not well-compatible with the equivariant CNNs in comparison to group aligning. Using the same network with bilinear pooling (ours$_{\text{bilinearpool}}$†) proposed in [26] shows significantly lower results compared to ours$_{\text{groupalign}}$†.

In the MVS dataset [53] to evaluate relative camera pose estimation, our model shows a higher performance than GIFT at finer error thresholds of 10° and 5°. This shows that our model can find more precise correspondences under 3D viewpoint changes. Overall, these results show that our descriptors using rotation-equivariant representation ex-

	HP-all		Roto-360		params.
	@5px	@3px	@5px	@3px	(millions)
ours (proposed $\|G\| = 16$)	**70.69**	**63.42**	91.35	90.18	0.62M
w/o orientation loss	66.41	58.61	85.29	83.26	0.62M
w/o descriptor loss	27.49	24.83	25.64	24.98	0.62M
w/o image scale pyramid	68.77	62.25	**91.47**	**90.43**	0.62M
w/o equivariant backbone	47.25	42.52	8.65	8.51	11.18M
$\|G\| = 64$	63.96	57.35	85.12	83.32	**0.16M**
$\|G\| = 36$	68.17	60.95	87.78	85.89	0.26M
$\|G\| = 32$	69.44	62.08	89.10	87.31	0.31M
$\|G\| = 24$	69.72	62.21	90.27	88.34	0.39M
$\|G\| = 8$	65.74	58.92	87.16	85.57	1.24M

Table 6. **Ablation test on HPatches and Roto-360.** 'params.' denotes the number of model parameters.

hibit strong transferability to the real-world examples.

4.5. Ablation study and design choice

Table 6 shows the results of ablation studies on the HPatches and Roto-360 datasets. The matching accuracy drops when either the orientation alignment loss or the contrastive descriptor loss is not used. Specifically, even when using the ground truth rotation difference for group alignment, not using the descriptor loss results in lower performance, highlighting the importance of robustness to other sorts of variations, *e.g.*, illumination or viewpoint. Not using the image pyramid at inference time results in a slight drop in HPatches, but the performance on Roto-360 remains nearly unchanged. When training without equivariant layers, ResNet-18 with conventional convolutional layers was used - this results in a drastic drop in performance especially on Roto-360, with a rapid increase in the number of model parameters. This demonstrates the significance of high sample efficiency of group-equivariant layers.

We also demonstrate the effect of the order of cyclic group G on the performance of our method in the second group of Table 6. We fix the computational cost $C \times |G| = 1,024$, and vary the order of group to show the parameter efficiency of the group equivariant networks. Our design choice $|G| = 16$ yields the best results, and the performance drops gracefully as G increases. This is because with a higher order of groups, the precision of dominant orientation estimation is likely to decrease, leading to lower results. Reducing the order of group to $|G| = 8$ reduces the MMA in both benchmarks as well, which we suspect is because the range of rotation covered by one group action becomes too wide, leading to increased approximation errors.

4.6. Qualitative results

Figure 6 visualizes the consistency of dominant orientation estimation. From the source (left) and target (middle) images, we estimate the dominant orientation for the same set of predicted keypoints. We use the ground truth rotation to align the estimated orientation and the target image for better visibility (right). The green and red arrows (middle, right) represent the consistent and inconsistent orientation predictions with respect to the initial estimations (left) at a 30° threshold. The numbers on the left represent the number of consistent estimations/number of detected keypoints. Compared to LF-Net [40] and RF-Net [50], our method predicts more consistent dominant orientations of keypoints.

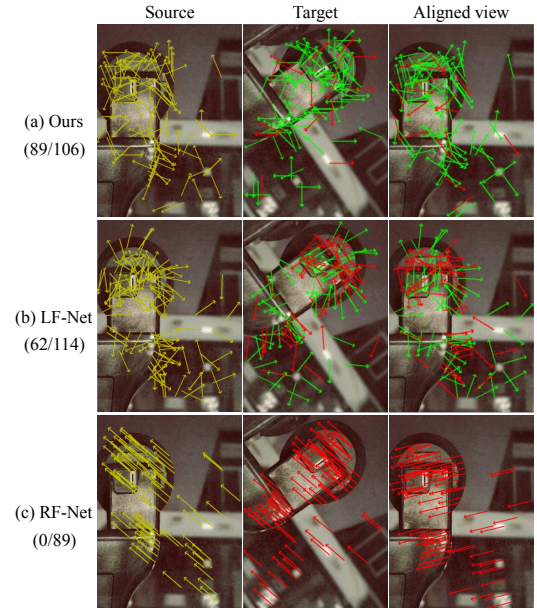


Figure 6. **Visualization of consistency of dominant orientation estimation.** Best viewed in electronics and colour.

5. Conclusion

We have proposed a self-supervised rotation-equivariant network for visual correspondence to improve the discriminability of local descriptors. Our invariant mapping called group-aligning shifts the rotation-equivariant features along the group dimension based on the orientation value to produce rotation-invariant descriptors while preserving the feature discriminability, without collapsing the group dimension. Our method achieves state-of-the-art performance in obtaining rotation-invariant descriptors, which are transferable to tasks such as keypoint matching and camera pose estimation. We believe that our approach can be further extended to other geometric transformation groups, and will motivate group-equivariant learning for practical applications of computer vision.

Acknowledgement. This work was supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-TF2103-02 and also by the NRF grant (NRF-2021R1A2C3012728) funded by the Korea government (MSIT).

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 1

[2] 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, 5, 6, 7

[3] Axel Barroso-Laguna, Yurun Tian, and Krystian Mikolajczyk. Scalenet: A shallow architecture for scale estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12808–12818, 2022. 2

[4] Fabio Bellavia and Carlo Colombo. Rethinking the sgloh descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):931–944, 2017. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

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

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Toward geometric deep slam. *arXiv preprint arXiv:1707.07410*, 2017. 1

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Deep Learning for Visual SLAM Workshop*, 2018. 1, 2, 3, 5, 6, 7

[9] 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, 6, 7

[10] Patrick Ebel, Anastasiia Mishchuk, Kwang Moo Yi, Pascal Fua, and Eduard Trulls. Beyond cartesian representations for local descriptors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 253–262, 2019. 2

[11] Bin Fan, Fuchao Wu, and Zhanyi Hu. Rotationally invariant descriptors using intensity order pooling. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):2031–2045, 2011. 1, 2

[12] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021. 2, 5

[13] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015. 3

[14] Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988. 3, 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[16] Jared Heinly, Johannes L. Schönberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world* in six days. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3287–3295, 2015. 1

[17] 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

[18] Seungwook Kim, Yoonwoo Jeong, Chunghyun Park, Jaesik Park, and Minsu Cho. SeLCA: Self-supervised learning of canonical axis. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022. 2

[19] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher: Match-to-match attention for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8697–8707, 2022. 3

[20] 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. 3, 6

[21] 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. 1, 2, 4

[22] Jongmin Lee, Yoonwoo Jeong, Seungwook Kim, Juhong Min, and Minsu Cho. Learning to distill convolutional features into compact local descriptors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 898–908, 2021. 2

[23] 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. 1, 2, 4

[24] 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. 2, 6

[25] 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. 5

[26] 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, 5, 6, 7

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1, 2, 3, 4, 5, 6, 7

[28] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Aslfeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6589–6598, 2020. 2, 6

[29] Simon Lynen, Bernhard Zeisl, Dror Aiger, Michael Bosse, Joel Hesch, Marc Pollefeys, Roland Siegwart, and Torsten Sattler. Large-scale, real-time visual–inertial localization revisited. *The International Journal of Robotics Research*, 39(9):1061–1084, 2020. 1

[30] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5048–5057, 2017. 2

[31] Roland Memisevic. On multi-view feature learning. In *ICML*, 2012. 2

[32] Roland Memisevic and Geoffrey E Hinton. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural computation*, 22(6):1473–1492, 2010. 2

[33] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004. 5

[34] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630, 2005. 6

[35] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, 2019. 3

[36] 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

[37] 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. 2

[38] Daniel Moyer, Esra Abaci Turk, P Ellen Grant, William M Wells, and Polina Golland. Equivariant filters for efficient tracking in 3d imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 193–202. Springer, 2021. 2

[39] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 1, 6

[40] 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, 6, 7, 8

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[42] 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. 2, 6, 7

[43] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Nataša Sladoje. CoMIR: Contrastive multimodal image representation for registration. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18433–18444. Curran Associates, Inc., 2020. 2

[44] 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. 1, 2, 5, 6

[45] Jérome Revaud, Vincent Leroy, Philippe Weinzaepfel, and Boris Chidlovskii. Pump: Pyramidal and uniqueness matching priors for unsupervised learning of local descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3926–3936, 2022. 1

[46] 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, 6, 7

[47] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *European conference on computer vision*, pages 752–765. Springer, 2012. 1

[48] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018. 1

[49] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[50] 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. 2, 6, 7, 8

[51] Kihyuk Sohn and Honglak Lee. Learning invariant representations with local transformations. In *ICML*, 2012. 2

[52] Ivan Sosnovik, Artem Moskalev, and Arnold Smeulders. How to transform kernels for scale-convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1092–1097, 2021. 2

[53] Christoph Strecha, Wolfgang Von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008. 2, 5, 6, 7

[54] Yurun Tian, Axel Barroso Laguna, Tony Ng, Vassileios Balntas, and Krystian Mikolajczyk. Hynet: Learning local descriptor with hybrid similarity measure and triplet loss. *Ad-*

vances in Neural Information Processing Systems, 33:7401–7412, 2020. 2

[55] 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. 6

[56] 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. 2

[57] Michal Jan Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 6

[58] Rui Wang, Robin Walters, and Rose Yu. Data augmentation vs. equivariant networks: A theory of generalization on dynamics forecasting. *arXiv preprint arXiv:2206.09450*, 2022. 1

[59] Zhenhua Wang, Bin Fan, Gang Wang, and Fuchao Wu. Exploring local and overall ordinal information for robust feature description. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2198–2211, 2015. 2

[60] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 32:14334–14345, 2019. 1, 2, 3, 5, 7

[61] 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. 2

[62] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017. 2

[63] Pei Yan, Yihua Tan, Shengzhou Xiong, Yuan Tai, and Yansheng Li. Learning soft estimator of keypoint scale and orientation with probabilistic covariant loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19406–19415, 2022. 1, 4

[64] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016. 1

[65] Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Oriented response networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–528, 2017. 2

[66] Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan. Very large-scale global sfm by distributed motion averaging. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4568–4577, 2018. 1