Learning Rotation-Equivariant Features for Visual Correspondence

Anonymous authors
Paper under double-blind review

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 show 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 (Sattler et al., 2018; Lynen et al., 2020), simultaneous localization and mapping (DeTone et al., 2017; 2018; Mur-Artal et al., 2015), and 3D reconstruction (Agarwal et al., 2011; Heinly et al., 2015; Schonberger & Frahm, 2016). 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 (Yi et al., 2016; Revaud et al., 2019; 2022). Learning *rotation*-invariant local descriptors, however, remains challenging; the classical techiniques (Lowe, 2004; Rublee et al., 2011) 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 (Weiler & Cesa, 2019) to jointly extract rotation-equivariant local features and their orientations from an image. To extract reliable orientations, we propose an orientation alignment loss, 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. 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 to obtain invariant features. Conventional methods to yield invariant features from group-equivariant features collapse the group dimension by group-pooling, *e.g.,* max-pooling or bilinear-pooling (Liu et al., 2019), 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 for the first time.
- We propose group aligning, a method to shift a group-equivariant descriptor 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 dominant 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 (Balntas et al., 2017) and the MVS dataset (Strecha et al., 2008).

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 (Lowe, 2004; Rublee et al., 2011). However, these methods cannot be applied to standard neural networks as the process of histogram construction is not differentiable. Therefore, we propose a fully end-to-end pipeline to obtain orientation-normalized local descriptors, with differentiable components for equivariant feature extraction and dominant orientation prediction.

Learning-based invariant local descriptors. A branch of learning-based methods learns to obtain invariant local descriptors in an explicit manner. GIFT (Liu et al., 2019) 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 (Pautrat et al., 2020) 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 (Ebel et al., 2019; Tian et al., 2020; 2019) or image-based (DeTone et al., 2018; Mishkin et al., 2018; Revaud et al., 2019; Tyszkiewicz et al., 2020; Lee et al., 2021b). 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 steerable networks (Weiler & Cesa, 2019), 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 Memisevic (2012); Memisevic & Hinton (2010); Sohn & Lee (2012); Marcos et al. (2017); Zhou et al. (2017); Weiler & Cesa (2019). For example, G-CNNs (Cohen & Welling, 2016a) use group equivariant convolutions that reduce sample complexity by exploiting symmetries on discrete isometric groups; SFCNNs (Weiler et al., 2018) and H-Nets (Worrall et al., 2017) 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 (Sosnovik et al., 2021; Lee et al., 2021a; Barroso-Laguna et al., 2022). Han et al. (2021); Pielawski et al. (2020); Lee et al. (2022) leverage equivariant

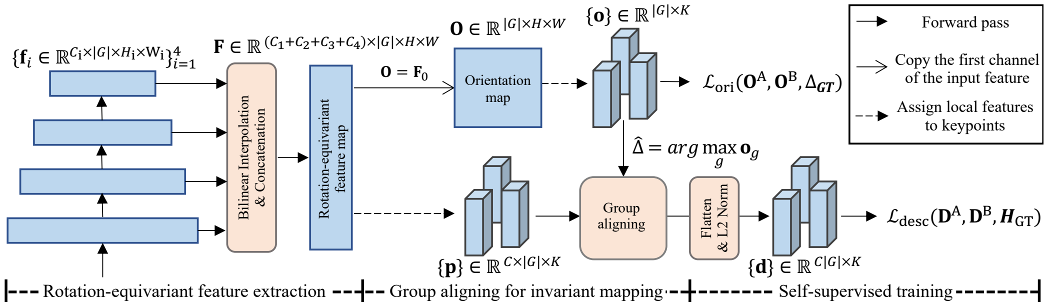

$\{\mathbf{f}_i \in \mathbb{R}^{C_i \times |G| \times H_i \times W_i}\}_{i=1}^{4}$
$\mathbf{F} \in \mathbb{R}^{(C_1+C_2+C_3+C_4)\times|G|\times H\times W}$
Bilinear Interpolation & Concatenation
Rotation-equivariant feature map
$\mathbf{O} = \mathbf{F}_0$
$\mathbf{O} \in \mathbb{R}^{|G|\times H\times W}$
Orientation map
$\{\mathbf{o}\} \in \mathbb{R}^{|G|\times K}$
$\mathcal{L}_{\text{ori}}(\mathbf{O}^{\text{A}}, \mathbf{O}^{\text{B}}, \Delta_{GT})$
Forward pass
Copy the first channel of the input feature
Assign local features to keypoints
$\hat{\Delta} = arg \max_g \mathbf{o}_g$
$\{\mathbf{p}\} \in \mathbb{R}^{C\times|G|\times K}$
Group aligning
Flatten & L2 Norm
$\{\mathbf{d}\} \in \mathbb{R}^{C|G|\times K}$
$\mathcal{L}_{\text{desc}}(\mathbf{D}^{\text{A}}, \mathbf{D}^{\text{B}}, \boldsymbol{H}_{\text{GT}})$
Rotation-equivariant feature extraction
Group aligning for invariant mapping
Self-supervised training

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 $\mathcal{H}_{\text{GT}}$.

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 Sec. A.1 of the appendix for a brief introduction to group equivariance.

3 Rotation-equivariant Features, Rotation-invariant descriptors

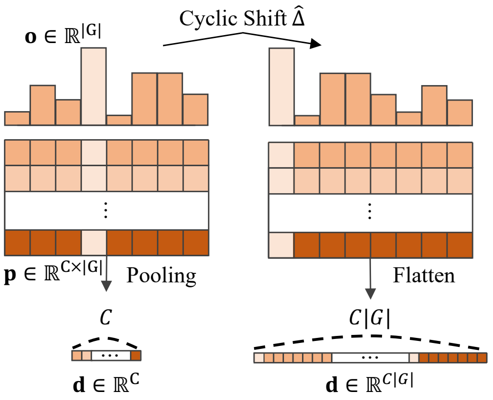
In this section, we first draw the line between the terms *feature* and *descriptor* which we will be used throughout this paper. Therefore, the goal of our work is to learn to extract rotation-equivariant local *features* from our rotation-equivariant backbone network, and then to align them by their 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 align 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). Fig. 1 shows the overall architecture of our method.

3.1 Rotation-equivariant feature extraction

As the feature extractor backbone, we use ReResNet18 (Han et al., 2021), which has the same structure as ResNet18 (He et al., 2016) but is constructed using rotation-equivariant convolutional layers Weiler & Cesa (2019). The 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 transformed to the vector field of the group representation (Weiler & Cesa, 2019). 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 (Hariharan et al., 2015; Min et al., 2019; Kim et al., 2022). 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.


$\mathbf{o} \in \mathbb{R}^{|G|}$
Cyclic Shift $\hat{\Delta}$
$\mathbf{p} \in \mathbb{R}^{C \times |G|}$
Pooling
Flatten
C
$C|G|$
$\mathbf{d} \in \mathbb{R}^{C}$
$\mathbf{d} \in \mathbb{R}^{C|G|}$

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

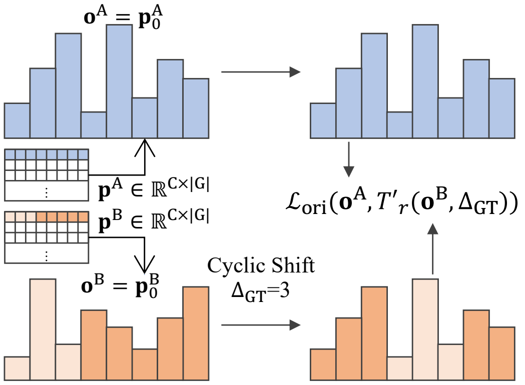

$\mathbf{o}^{\mathrm{A}} = \mathbf{p}_0^{\mathrm{A}}$
$\mathbf{p}^{\mathrm{A}} \in \mathbb{R}^{C \times |G|}$
$\mathbf{p}^{\mathrm{B}} \in \mathbb{R}^{C \times |G|}$
$\mathcal{L}_{\mathrm{ori}}(\mathbf{o}^{\mathrm{A}}, T'_r(\mathbf{o}^{\mathrm{B}}, \Delta_{\mathrm{GT}}))$
$\mathbf{o}^{\mathrm{B}} = \mathbf{p}_0^{\mathrm{B}}$
Cyclic Shift
Δ_{GT}=3

Figure 3: **Illustration of orientation alignment loss.** Given two rotation-equivariant tensors $p^{\mathrm{A}}, p^{\mathrm{B}} \in \mathbb{R}^{C \times |G|}$ 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 resulting orientation vectors of the two descriptors to be the same.

3.2 Assigning local features to keypoints

During training, we extract K keypoints from the source image using Harris corner detection (Harris et al., 1988). We then use the ground-truth homography $\mathcal{H}_{\mathrm{GT}}$ 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 use SuperPoint (DeTone et al., 2018) as the keypoint detector during inference.

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 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. Fig. 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. 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 $\mathbf{O}$. 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 $\hat{\Delta} = \frac{|G|}{360}\hat{\theta}$. 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 $\mathbf{p}$ 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:

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

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 orientation histogram in the classical descriptor SIFT (Lowe, 2004), but we apply this concept on the equivariant neural features. The proposed group aligning preserves the group information, so our invariant descriptors have more representative power than the existing group pooling methods which collapse the group dimension for invariance.

3.4 Orientation alignment loss

To learn to obtain the dominant orientations from the orientation vectors, we propose an orientation alignment loss to supervise the orientation histograms in $\mathbf{O}$ to be rotation equivariant under the photometric/geometric transformations. Fig. 3 shows the illustration of orientation alignment loss. The cyclic shift of an orientation histogram map at train 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 loss as follows:

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

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 Eq. 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.

3.5 Contrastive descriptor loss

We propose to use a descriptor similarity loss motivated by contrastive learning (Chen et al., 2020) 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:

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

where sim is cosine similarity and τ is the softmax temperature. Our overall self-supervised loss is formulated as $\mathcal{L} = \alpha\mathcal{L}^{\mathrm{ori}} + \beta\mathcal{L}^{\mathrm{desc}}$, where α and β are balancing terms.

3.6 Scale invariance

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 (Revaud et al., 2019). 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 (Lowe, 2004), for improved robustness to scale changes.

4 Experiment

Implementation details. We use rotation-equivariant ResNet-18 (ReResNet-18) (Han et al., 2021) implemented using the rotation-equivariant layers of $E(2)$-CNN framework (Weiler & Cesa, 2019)

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

	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 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.

	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

as our backbone network. We remove the first maxpool layer to preserve the spatial size, so that the spatial resolution of the rotation-equivariant feature **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 0.0001, and a weight decay of 0.1. We train our model for 12 epochs with 1000 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. Loss balancing factors α and β are 10 and 1, respectively. The final output descriptor size is 1,024, with $C = 64$, $|G| = 16$. We use SuperPoint (DeTone et al., 2018) as the keypoint detector to evaluate our method. 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 our proposed Roto-360 dataset and show the transferability on real image benchmarks, *i.e.,* HPatches (Balntas et al., 2017) and MVS (Strecha et al., 2008) datasets.

Training dataset. We generate a synthetic dataset for self-supervised training from the MS-COCO dataset (Lin et al., 2014). 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 as follows: $\theta_{\text{GT}} = \arctan(\frac{\mathcal{H}_{21}}{\mathcal{H}_{11}})$, where $\mathcal{H}$ is a 3×3 homography matrix. We sample $K = 512$ keypoints using Harris corner detector (Harris et al., 1988), 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 (Liu et al., 2019) 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 (Balntas et al., 2017). Roto-360 is more suitable to evaluate the rotation invariance of our descriptors, as the extreme rotation (ER) dataset (Liu et al., 2019) 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 and the number of predicted matches following D2-Net (Dusmanu et al., 2019).

HPatches (Balntas et al., 2017) 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 (Strecha et al., 2008) 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 given ground-truth keypoint pairs without training, *i.e.,* no keypoint deviation, on the Roto-360 dataset. As ground-truth keypoint pairs are used, we use MMA@1px as the evaluation metric. We demonstrate the upper-bound

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. See text for details.

Method	MMA			pred.	total.
	@10px	@5px	@3px		
SIFT (Lowe, 2004)	71.67	71.42	71.29	194.1	382.8
ORB (Rublee et al., 2011)	78.73	85.29	86.78	607.6	1005.2
SuperPoint (DeTone et al., 2018)	22.85	22.10	21.83	462.6	1161.0
LF-Net (Ono et al., 2018)	75.05	74.30	72.61	386.7	1024.0
RF-Net (Shen et al., 2019)	15.64	15.18	14.58	1602.5	5000.0
D2-Net (Dusmanu et al., 2019)	15.56	9.30	5.21	386.9	1474.5
R2D2 (Revaud et al., 2019)	15.80	14.97	13.50	197.9	1500.0
GIFT (Liu et al., 2019)	42.35	42.05	41.59	589.2	1161.0
LISRD (Pautrat et al., 2020)	16.96	16.04	15.64	323.6	1781.1
PosFeat (Li et al., 2022)	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

performance of group aligning using Δ_{GT} to shift the equivariant features, where it shows to find nearly perfect keypoint correspondences with 97.54% matching accuracy. Using group pooling *i.e.,* max pooling or average pooling, largely loses 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 existing bilinear pooling (Liu et al., 2019) 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 (Liu et al., 2019). 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 (DeTone et al., 2018; Dusmanu et al., 2019; Lowe, 2004; Mur-Artal et al., 2015; Ono et al., 2018; Revaud et al., 2019; Shen et al., 2019), or combined with off-the-shelf detectors (Liu et al., 2019; Pautrat et al., 2020; Li et al., 2022). While the classical methods (Lowe, 2004; Rublee et al., 2011) achieve better matching accuracy than the existing learning-based methods, our method achieves the best results overall. While GIFT (Liu et al., 2019) and ours use the SuperPoint detector (DeTone et al., 2018), our method finds more matches than SuperPoint and GIFT albeit exhibiting the same number of total extracted keypoints, which are also more accurate as can be seen from the higher MMA results. This shows that our descriptors obtained using group aligning show the highest matching accuracy under rotation changes compared to existing methods. The significant outperformance 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. While we apply group-aligning to the group-equivariant features using the dominant orientation value, we can also extract multiple differently aligned descriptors by using multiple orientation candidates. 'ours*' denotes a setting where we use multiple orientation candidates, whose scores are at least 60% of the maximum score in the orientation histogram, to align the group-equivariant features. We consider a single keypoint position with k differently aligned descriptors as k detected keypoints. This multiple descriptor extraction compensates for the case of incorrect orientation prediction, thereby further improving

Table 4: **Evaluation with predicted keypoint pairs on real image benchmarks.** The first group of methods includes existing local descriptor 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 dominant 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. Results in bold indicate the best result, and underlined results indicate the second best results.

Method	HP-all		HP-illu		HP-view		Pose		
	@5px	@3px	@5px	@3px	@5px	@3px	20°	10°	5°
SIFT (Lowe, 2004)	46.85	44.20	45.97	43.30	47.70	45.07	0.02	0.00	0.00
ORB (Rublee et al., 2011)	52.22	47.40	50.85	46.29	53.55	48.47	0.06	0.00	0.00
SuperPoint (DeTone et al., 2018)	69.71	61.75	74.63	67.53	64.96	56.17	0.20	0.07	0.01
LF-Net (Ono et al., 2018)	56.45	52.22	62.21	57.63	50.88	47.00	0.06	0.03	0.01
RF-Net (Shen et al., 2019)	59.08	54.42	61.63	57.46	56.62	51.49	0.10	0.04	0.01
D2-Net (Dusmanu et al., 2019)	50.18	32.54	63.80	44.09	37.02	21.38	0.11	0.05	0.01
GIFT (Liu et al., 2019)	76.03	67.31	**79.71**	**71.89**	72.48	62.88	**0.60**	0.28	0.09
LISRD (Pautrat et al., 2020)	62.16	56.12	70.09	63.64	54.50	48.85	0.05	0.02	0.00
PosFeat (Li et al., 2022)	53.52	45.97	62.73	55.51	44.62	36.75	0.19	0.06	0.01
$\text{ours}_{\text{avgpool}}$	64.10	57.94	62.28	56.27	65.85	59.55	0.27	0.10	0.05
$\text{ours}_{\text{maxpool}}$	61.57	55.81	59.66	53.91	63.42	57.64	0.27	0.11	0.03
$\text{ours}_{\text{bilinearpool}}$ (Liu et al., 2019)	45.59	41.90	45.13	41.57	46.03	42.22	0.35	0.17	0.09
$\text{ours}_{\text{bilinearpool}}$† (Liu et al., 2019)	58.72	53.77	57.32	52.67	60.06	54.83	0.24	0.11	0.03
$\text{ours}_{\text{groupalign}}$	70.69	63.42	70.39	62.88	70.97	63.95	0.58	0.26	0.12
$\text{ours}_{\text{groupalign}}$*	73.92	66.37	73.13	65.33	74.69	67.38	0.56	0.30	0.12
$\text{ours}_{\text{groupalign}}$†	**78.00**	**69.70**	77.94	69.35	**78.06**	**70.03**	0.56	**0.33**	**0.14**

the matching accuracy. We guide the readers to Sec. A.2 of the appendix for more details on multiple descriptor extraction.

Consistency of matching accuracy with respect to rotation changes. Fig. 4 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 and ORB are high at the upright rotations, they tend to fluctuate significantly with varying rotations. The existing learning-based group-invariant descriptor, GIFT (Liu et al., 2019), fails to find correspondences beyond 60°.

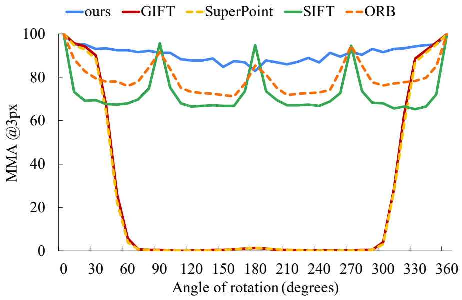


Figure 4: **Matching accuracies according to varying degree of rotations on Roto-360.**

4.4 Transferability of local descriptors to real image benchmarks

Table 4 demonstrates the matching performance of local descriptors on HPatches illumination/viewpoint (Balntas et al., 2017) and pose estimation (Strecha et al., 2008). Our model shows the highest performance overall on the HPatches dataset. While GIFT shows a higher performance under illumination changes that only contain identity mappings, ours†, which uses a larger backbone network (ReWRN), improves 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. It should be noted that the core difference between $\text{ours}_{\text{bilinearpool}}$ and GIFT is the usage of explicit rotation-equivariant CNNs (Weiler & Cesa, 2019), 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 ($\text{ours}_{\text{bilinearpool}}$†) shows significantly lower results compared to $\text{ours}_{\text{groupalign}}$†.

In the MVS dataset (Strecha et al., 2008) to evaluate relative camera pose estimation, our method shows a higher performance than GIFT at finer rotation 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 proposed local descriptors using rotation-equivariant representations exhibit strong transferability to real-world matching datasets.

4.5 Ablation study and design choice

Table 5 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.

Table 5: **Ablation test on HPatches and Roto-360.** 'params.' denotes the number of model parameters.

	HP-all		Roto-360		params.
	@5px	@3px	@5px	@3px	(millions)
ours (proposed $\lvert G\rvert = 16$)	**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
$\lvert G\rvert = 64$	63.96	57.35	85.12	83.32	**0.16M**
$\lvert G\rvert = 36$	68.17	60.95	87.78	85.89	0.26M
$\lvert G\rvert = 32$	69.44	62.08	89.10	87.31	0.31M
$\lvert G\rvert = 24$	69.72	62.21	90.27	88.34	0.39M
$\lvert G\rvert = 8$	65.74	58.92	87.16	85.57	1.24M

We also carry out experiments to show the effect of the order of cyclic group G on the performance of our method in the second group of Tab. 5. 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 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 performance 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

Fig. 5 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 dominant 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 (Ono et al., 2018) and RF-Net (Shen et al., 2019), our method predicts more consistent dominant orientations of keypoints.

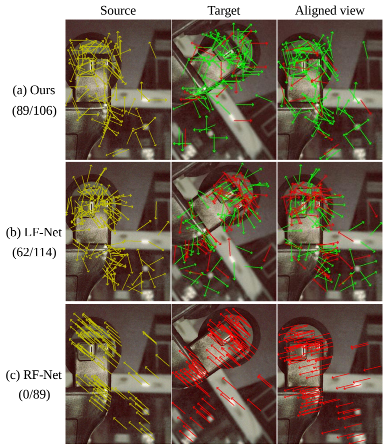


Figure 5: **Visualization of consistency of dominant orientation estimation.** Best viewed in electronics and colour.

5 Conclusion

In this paper, we propose to apply rotation-equivariant networks to the task of visual correspondence to improve the discriminability of local descriptors. Specifically, we propose group-aligning, a novel method to shift the group-equivariant descriptors along the group dimension to yield rotation-invariant descriptors without having to collapse the group information, preserving the feature discriminability. Our proposed pipeline is trained in a self-supervised manner, leveraging orientation alignment loss for dominant orientation prediction and contrastive descriptor loss for robust descriptor extraction. We demonstrate state-of-the-art performances in obtaining rotation-invariant descriptors and strong transferability to the tasks of keypoint matching and camera pose estimation. We anticipate that this work proposes the potential of expanding the rotation group to more general geometric transformation groups, and will motivate the use of group-equivariant representation for more practical applications of computer vision.

REFERENCES

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.

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*, pp. 5173–5182, 2017.

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*, pp. 12808–12818, 2022.

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

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016a.

Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016b.

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*, pp. 9145–9156, 2019.

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

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Deep Learning for Visual SLAM Workshop*, 2018. URL `http://arxiv.org/abs/1712.07629`.

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*, pp. 8092–8101, 2019.

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*, pp. 253–262, 2019.

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.

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*, pp. 2786–2795, 2021.

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*, pp. 447–456, 2015.

Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pp. 10–5244. Citeseer, 1988.

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*, pp. 770–778, 2016.

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)*, pp. 3287–3295, 2015. doi: 10.1109/CVPR.2015.7298949.

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*, pp. 8697–8707, 2022.

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.

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, 2021a. URL `https://www.bmvc2021-virtualconference.com/programme/accepted-papers/`.

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*, pp. 898–908, 2021b.

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*, pp. 4847–4857, 2022.

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*, pp. 15838–15848, 2022.

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*, pp. 740–755. Springer, 2014.

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.

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

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.

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

Roland Memisevic. On multi-view feature learning. In *ICML*, 2012.

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

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. 2019.

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)*, pp. 284–300, 2018.

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.

Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: learning local features from images. In *Advances in neural information processing systems*, pp. 6234–6244, 2018.

Rémi Pautrat, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *European Conference on Computer Vision*, pp. 707–724. Springer, 2020.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18433–18444. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/d6428eecbe0f7dff83fc607c5044b2b9-Paper.pdf`.

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.

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*, pp. 3926–3936, 2022.

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

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.

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*, pp. 8601–8610, 2018.

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

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*, pp. 8132–8140, 2019.

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

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*, pp. 1092–1097, 2021.

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*, pp. 1–8. Ieee, 2008.

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*, pp. 11016–11025, 2019.

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.

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

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

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*, pp. 849–858, 2018.

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*, pp. 5028–5037, 2017.

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

Hao Zhou, Torsten Sattler, and David W Jacobs. Evaluating local features for day-night matching. In *European Conference on Computer Vision*, pp. 724–736. Springer, 2016.

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

A Appendix

Contents

In this appendix, 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 A.1 explain a formal definition of equivariance and group equivariant networks. Section A.2 shows an example of multiple descriptor extraction using dominant orientation candidates, and different strategies of multiple descriptor extraction. Section A.3 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 A.4 shows the matching results with increasing the number of samples of the Roto-360 dataset. Section A.5 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 (Balntas et al., 2017; Liu et al., 2019).

A.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 $\Phi : X \to Y$ as

$$\Phi[T_g(x)] = T'_g[\Phi(x)], \tag{6}$$

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 $\mathbb{Z}^2 \to \mathbb{R}^K$ given convolution filter weights $\psi \in \mathbb{R}^{2\times K}$, the translation equivariance of a convolutional operation can be expressed as follows:

$$[T_t f] * \psi(x) = [T_t[f * \psi]](x), \tag{7}$$

where $*$ indicates the convolution operation.

Recent studies (Cohen & Welling, 2016a; Cohen et al., 2019; Cohen & Welling, 2016b; Weiler & Cesa, 2019; Weiler et al., 2018) 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:

$$[T_g f] * \psi(g) = [T_g[f * \psi]](g), \tag{8}$$

by replacing $t \in (\mathbb{R}^2, +)$ with $g \in G$ in Eq. 7. 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

$$L_i[T_g(g)] = T_g[L_i(g)], \tag{9}$$

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.

$$[\Pi_{i=1}^{M} L_i](T_g I) = T_g[\Pi_{i=1}^{M} L_i](I). \tag{10}$$

A.2 Elaboration of multiple descriptor extraction

In this section, we present an example of the multiple descriptor extraction scheme which was mentioned in Sec. 4.3, Tabs. 3 and 4, and we show the results of different configurations of the multiple descriptor extraction scheme.

A.2.1 An example of multiple extractions using orientation candidates

Fig. 6 shows an example of multiple descriptor extraction using 0.6 as the score ratio threshold. The distribution denotes an orientation histogram $\mathbf{o} \in \mathbb{R}^{16}$, and the scores are confidence values of each bin obtained from the group-equivariant features. The indices pointed by arrows denote the orientation candidates to be used for multiple descriptor extraction. The example shows that 3 orientations are selected to obtain 3 candidate descriptors for the feature point, which is possible as we predict a score for each orientation.

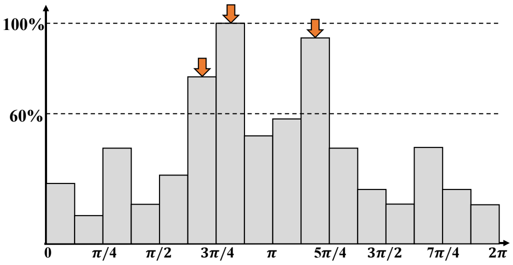


Figure 6: **An example of multiple descriptor extraction using 0.6 as the score ratio for orientation candidates.**

A.2.2 Different strategies for multiple descriptor extraction

Tab. 6 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 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 (Lowe, 2004). 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.

Table 6: **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.

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

A.3 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 (Pautrat et al., 2020), which is originally from the DNIM dataset (Zhou et al., 2016).

A.3.1 Data generation

The source dataset DNIM (Zhou et al., 2016) 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 (Pautrat et al., 2020) dataset is generated by warping the target image of each pair with homographies as in SuperPoint (DeTone et al., 2018) 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 (Pautrat et al., 2020). Fig. 7 shows some examples of ERDNIM image pairs.

A.3.2 EXAMPLES OF ERDNIM IMAGE PAIRS

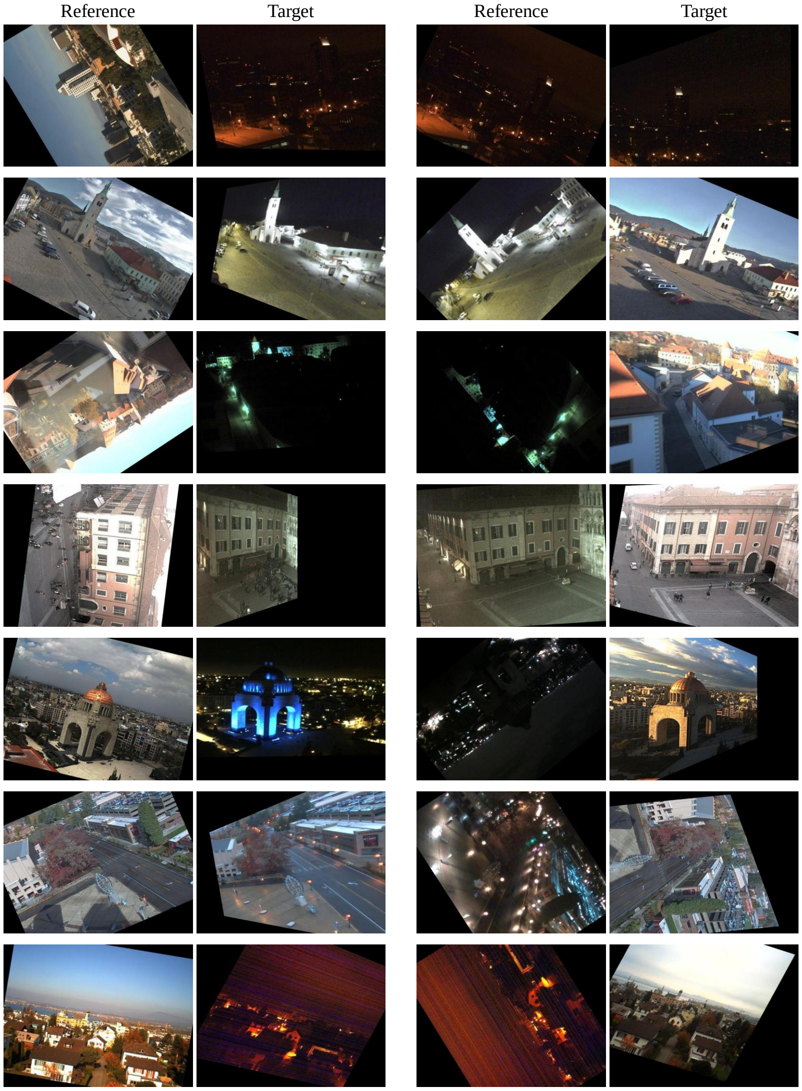


Figure 7: **Example of ERDNIM image pairs augmented from (Pautrat et al., 2020; Zhou et al., 2016).** The left two columns show the day reference benchmark with day-day and day-night image pairs. The right two columns show the night reference benchmark with night-day and night-night image pairs. The 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.

Table 7: **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 underlined results indicate the second best scores.

		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	0.232	**0.272**
	MMA	0.049	0.082	0.024	0.054	0.068	0.123	0.270	0.245	**0.277**
Night	HEstimation	0.108	0.092	0.002	0.062	0.097	0.151	0.291	0.316	**0.364**
	MMA	0.082	0.111	0.033	0.076	0.093	0.177	0.358	0.362	**0.404**

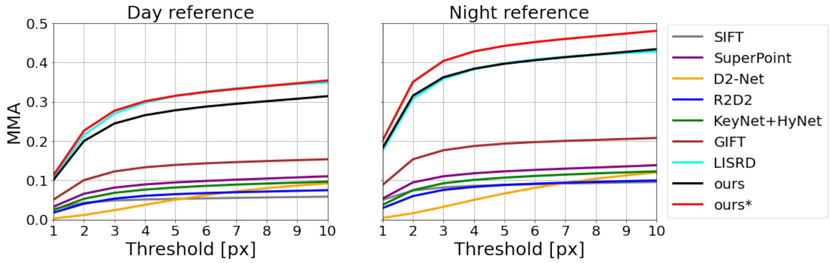


Figure 8: **Results of MMA with different pixel thresholds on the ERDNIM dataset.** 'ours*' uses k differently group-aligned features based on top-k selection. We use $k = 4$ in this experiment.

A.3.3 Evaluation metrics

We use two evaluation metrics, HEstimation and mean matching accuracy (MMA), following LISRD (Pautrat et al., 2020). We measure the homography estimation score (DeTone et al., 2018) using RANSAC (Fischler & Bolles, 1981) 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 (Dusmanu et al., 2019; Revaud et al., 2019) 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.

A.3.4 Results

Table 7 shows the evaluation results on the ERDNIM dataset. We compare the descriptor baselines SIFT (Lowe, 2004), SuperPoint (DeTone et al., 2018), D2-Net (Dusmanu et al., 2019), R2D2 (Revaud et al., 2019), KeyNet+HyNet (Laguna & Mikolajczyk, 2022; Tian et al., 2020), GIFT (Liu et al., 2019), and LISRD (Pautrat et al., 2020). Our proposed model with the rotation-equivariant network (ReResNet-18) achieves state-of-the-art performance in terms of homography estimation. GIFT (Liu et al., 2019), 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 (Lin et al., 2014) to GIFT (Liu et al., 2019). LISRD (Pautrat et al., 2020), 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.

Fig. 8 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.

A.4 The number of sampled images for Roto-360

Fig. 8 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.

Table 8: **Results on Roto-360 constructed using a different number of source images.**

# 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

A.5 Additional qualitative results

A.5.1 Visualization of the consistency of orientation estimation

We provide more examples for Figs. 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 Fig. 9, we create a sequence of 480×640 images augmented by random in-plane rotation with Gaussian noise sourced by ILSVRC2012 (Russakovsky et al., 2015). Fig. 9 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 (Lee et al., 2021a) consistently estimates more correct keypoint orientations than LF-Net (Ono et al., 2018) and RF-Net (Shen et al., 2019).

A.5.2 Visualization of the similarity maps of a keypoint under varying rotations

Fig. 10 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 (Liu et al., 2019) and LF-Net (Ono et al., 2018). Specifically, although GIFT (Liu et al., 2019) 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 (Weiler & Cesa, 2019) yield better rotation-invariant features than constructing the group-equivariance features with image augmentation (Liu et al., 2019).

A.5.3 Visualization of the predicted matches on the extreme rotation

Figs. 11 visualize the predicted matches on the ER dataset (Liu et al., 2019). 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 (Liu et al., 2019) and LF-Net (Ono et al., 2018).

A.5.4 Visualization of the predicted matches on the HPatches viewpoint

Fig. 12 visualize the predicted matches on the HPatches (Balntas et al., 2017) 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 (Liu et al., 2019) and LF-Net (Ono et al., 2018).

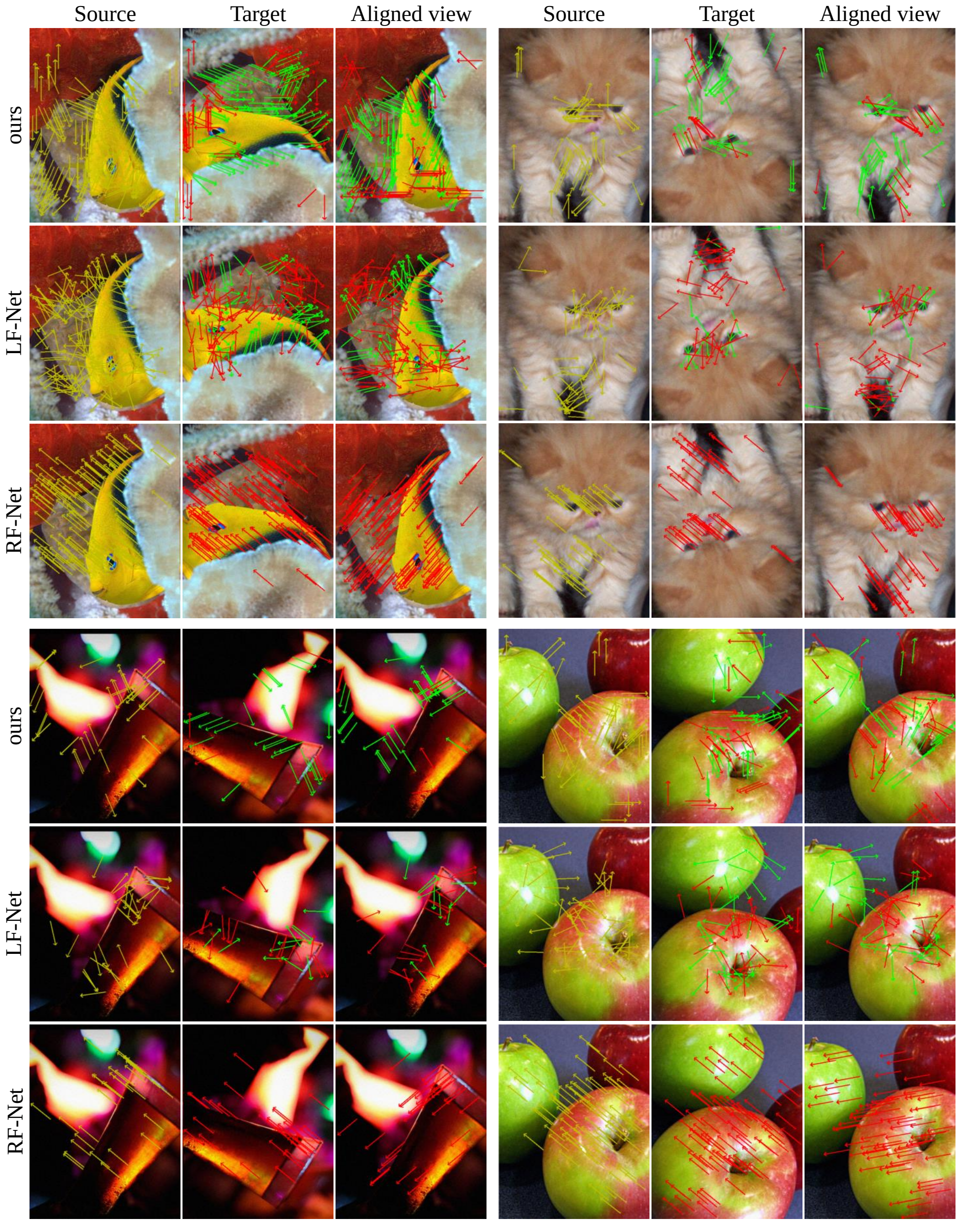


Figure 9: **Visualization of consistency of dominant orientation estimation.** We extract the source keypoints using SuperPoint (DeTone et al., 2018) and obtain the target keypoints using GT homography. We evaluate the consistency of orientation estimation by comparing the relative angle difference and 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 (Ono et al., 2018) and RF-Net (Shen et al., 2019).

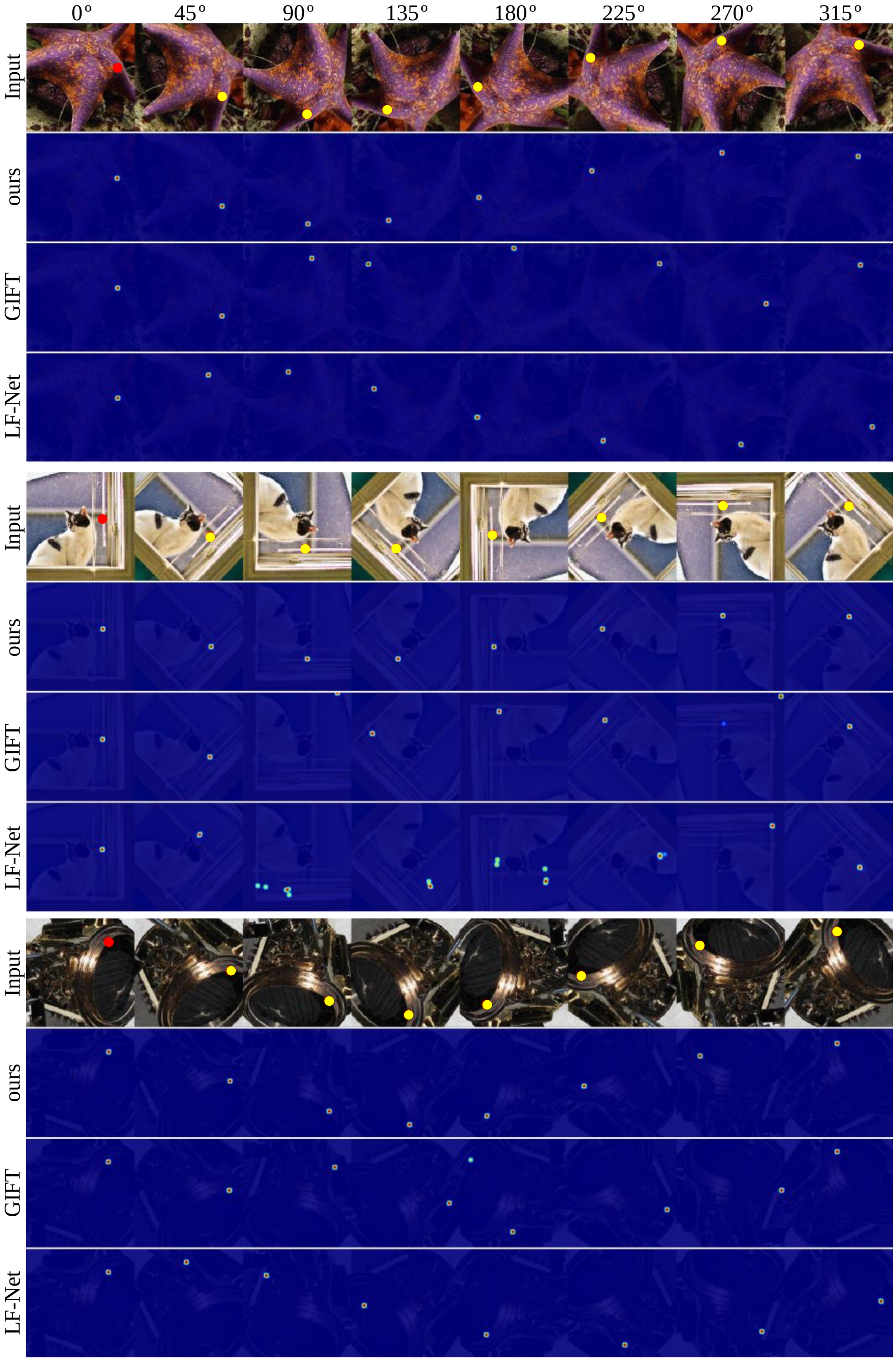


Figure 10: **Similarity maps with respect to a keypoint under rotation.** We compare one descriptor about the red keypoint from the source image at 0° to all other descriptors extracted across the rotated images, with yellow circles representing corresponding keypoints. For better visibility, we visualize the top 5 pixels with the highest similarity to the keypoints.

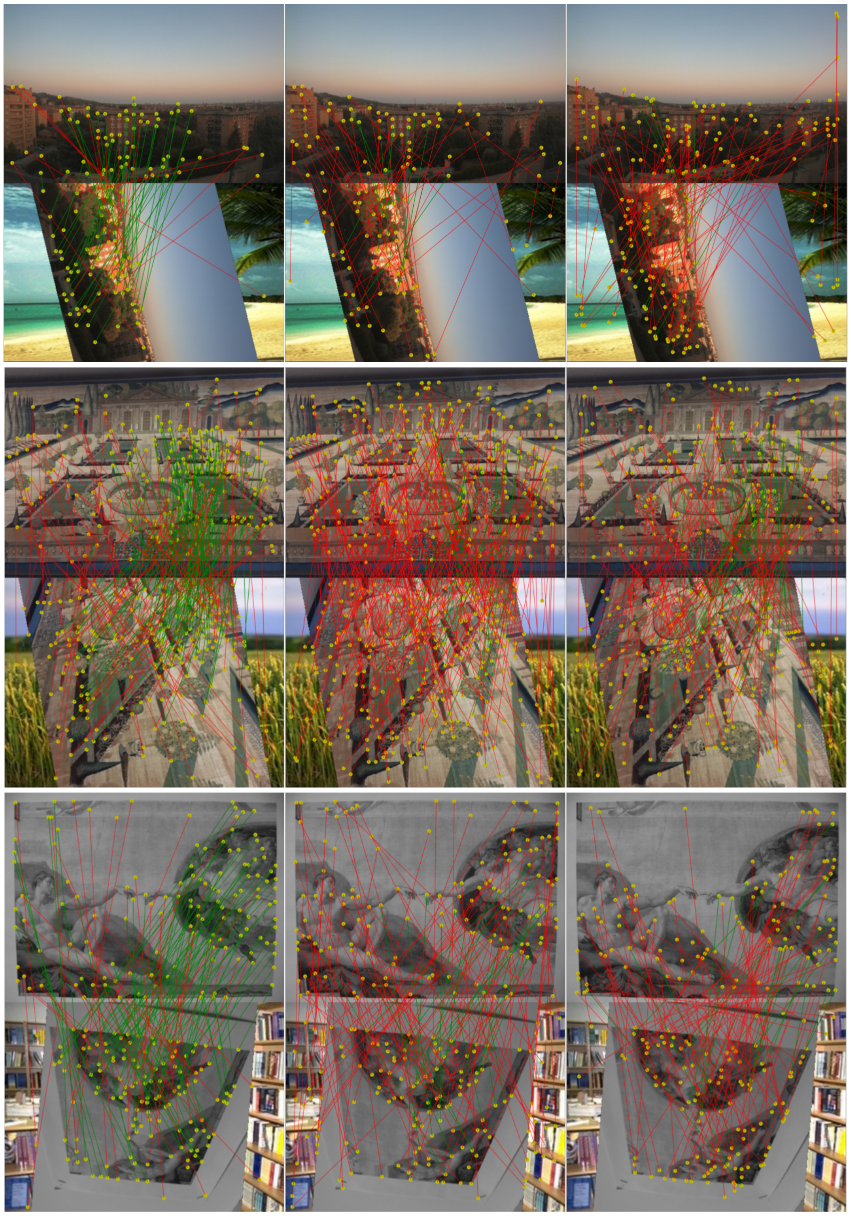
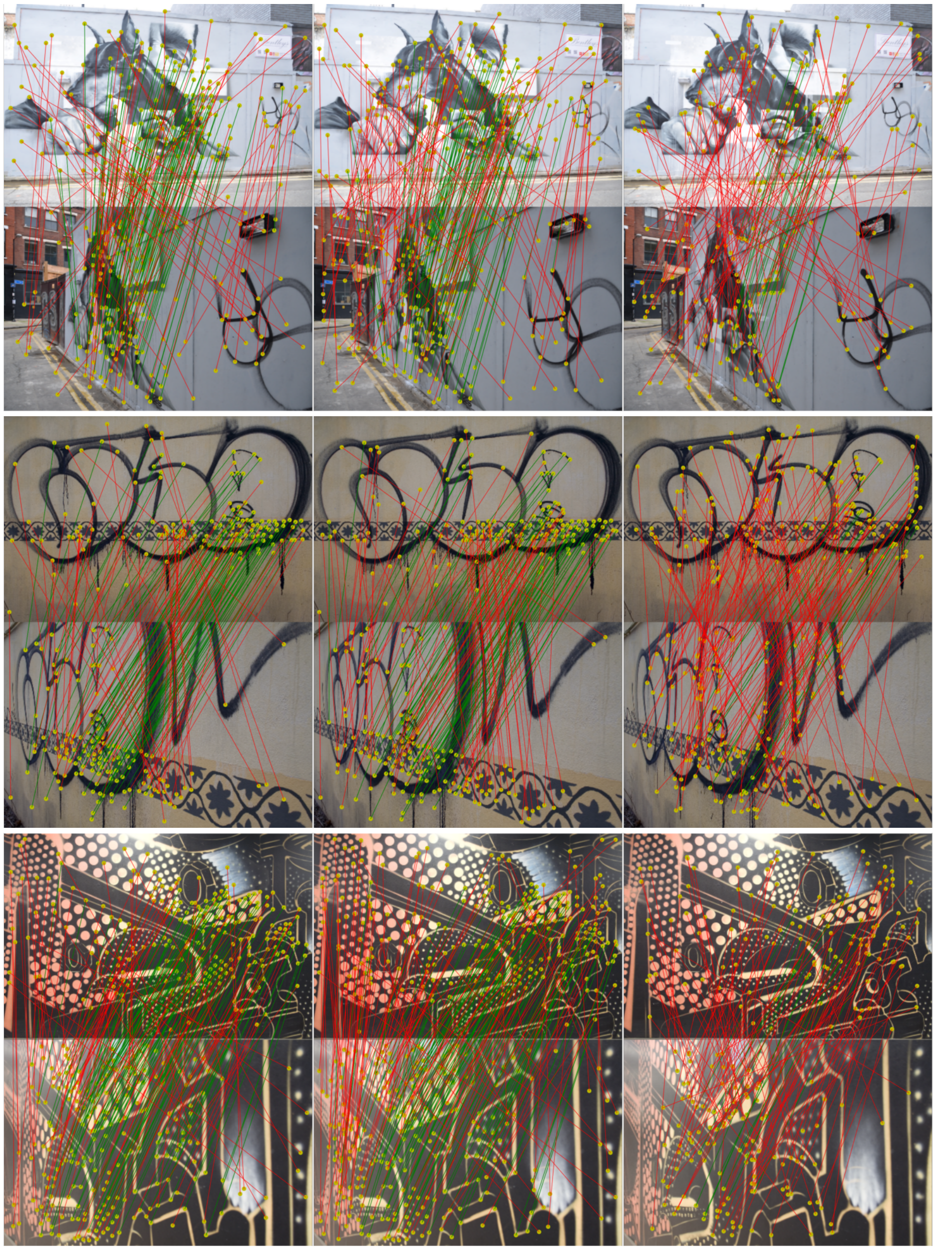

(a) ours (b) GIFT (c) LF-Net

Figure 11: **Visualization of predicted matches in the ER dataset (Liu et al., 2019).** 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 12: **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.