Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Rotation-equivariant correspondence matching based on a dual-activation mixer☆

Shuai Su [a], Ronghao Dang [a], Rui Fan [a], Chengju Liu [a,b], Qijun Chen [a,*]

[a] School of Electronics and Information Engineering, Tongji University, Shanghai, 201804, China
[b] Tongji Research Institute of Artificial Intelligence (Suzhou), Jiangsu, 215131, China

ARTICLE INFO

Communicated by J. Ma

Keywords:
Correspondence matching
Convolutional neural networks
Rotation-equivariant visual feature
Rotation-invariant visual feature
Inlier ratio

ABSTRACT

Learning-based correspondence matching methods have become the mainstream techniques in many computer vision and robotics applications due to their robustness to large illumination and viewpoint changes. However, it is difficult for conventional convolutional neural networks (CNNs) to extract rotation-equivariant local features. Recent work has shown that CNNs combined with group-equivariant architectures are surprisingly effective at matching correspondences even when the images are rotated to a dramatic extent. However, the inherent shape (square) of convolution kernels causes the performance bottleneck of such rotation-equivariant CNNs. To address this issue, we propose an adaptive dual rotation-equivariant correspondence matching algorithm, which performs stably at all angles. We mathematically analyze the effectiveness of our proposed rotation-equivariant correspondence matching approach and its performance with respect to different convolution kernels. Extensive experiments on the Rotated-HPatches, SIM2E, and MegaDepth datasets demonstrate the effectiveness of our proposed algorithm.

1. Introduction

Correspondence matching is a fundamental task within computer vision that involves identifying corresponding feature points, regions, or objects between two or more images [1,2], and it has various applications such as image stitching, object tracking, and 3D reconstruction [3–7]. Correspondence matching is often applied in various complex scenarios, thus it typically requires properties such as illumination invariance, viewpoint invariance, and rotation invariance.

Traditional methods [8–10] typically consist of detecting the significant feature points, describing feature points and matching them with nearest neighbor algorithm.

However, due to the handcrafted features being computed based on brightness variations in the neighborhood of keypoints, they are sensitive to changes in brightness. This sensitivity results in poor performance in scenarios with variations in lighting, seasonal changes, and other factors, leading to a significant degradation in tasks such as visual localization and object tracking in open environments [11–14].

Researchers have explored image keypoint detection and description based on deep learning techniques [15,16], learning features from extensive image data without relying on manually annotated data. This capability allows it to perform exceptionally well when abundant data is available. SuperPoint utilizes a self-supervised learning approach and introduces Homographic Adaptation to enhance its robustness to scale and homography transformations. It is trained on the MS-COCO general image dataset. On the other hand, R2D2 proposes joint learning of keypoint detection and description, along with a local descriptor discriminativeness predictor. Trained through self-supervision, it is capable of simultaneously producing sparse, repeatable, and reliable keypoints.

Nevertheless, traditional convolutional neural networks only possess translational invariance, and they exhibit poor rotational robustness. This leads to a significant decrease in performance, especially in applications with substantial rotation, often resulting in the failure of downstream tasks.

Researchers have introduced Group-equivariant CNNs (G-CNNs) [17] and explored their initial applications on the rotated-MNIST dataset. As the name suggests, equivariant neural networks, when subjected to a certain type of transformation operation on their input, produce an output that undergoes the same transformation. This is achieved by predefining the types of transformation operations and

☆ This paper is supported by the National Natural Science Foundation of China under Grants (62073245, 62233013). Supported by the Shanghai Municipal Science and Technology Major Project, China (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities, China. Suzhou Key Research and Development Project, China (SGC2021069); Special funds for Jiangsu Science and Technology Plan, China (BE2022119).

* Corresponding author.
E-mail addresses: sushuai@tongji.edu.cn (S. Su), qjchen@tongji.edu.cn (Q. Chen).

https://doi.org/10.1016/j.neucom.2023.127053
Received 27 June 2023; Received in revised form 31 October 2023; Accepted 14 November 2023
Available online 17 November 2023

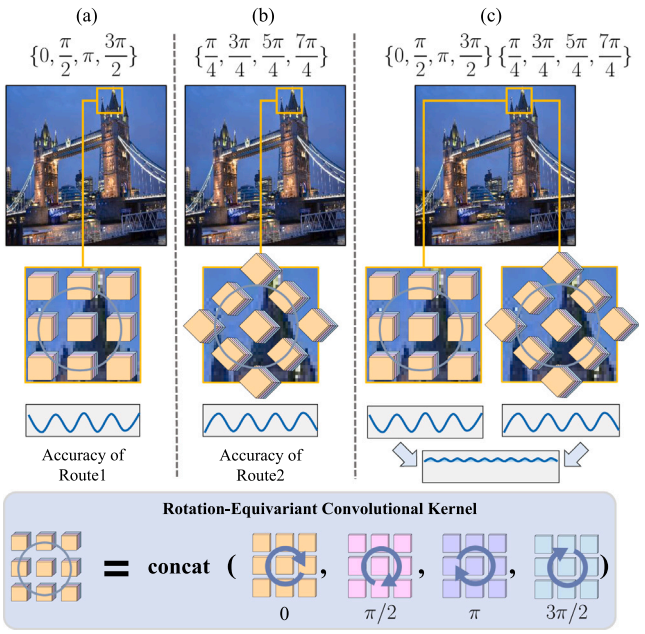


Fig. 1. An illustration of our approach, where a group-equivariant neural network is used for rotation-equivariant feature extraction. Such network archieves similar performance for each $\frac{\pi}{2}$. Our approach mixes two collections of visual features achieving significant performance at all angles.

then grouping convolutional filters to discretely sample the transformation space, aiming to obtain equivariant properties as much as possible. ReF [18] utilizes G-CNNs for correspondence matching. It employs a 5-layer fully convolutional equivariant neural network and then converts the rotation-equivariant feature maps into rotation-invariant ones using Group-Pooling operations. In terms of overall network architecture and training, it draws inspiration from the R2D2 approach, resulting in correspondence matching with robust resistance to large rotations.

However, we have observed that due to the inherent limitations of the storage format of digital images and the shape of convolutional kernels, ReF maintains consistently high performance only around 0 degrees, 90 degrees, 180 degrees, and 270 degrees, while its performance is almost entirely lost around 45 degrees, 135 degrees, 225 degrees, and 315 degrees. The performance curve exhibits an alternating pattern of peaks and valleys. This means that ReF meets the requirements of downstream tasks only in the vicinity of these peaks, and when the rotation angles are random, there is a significant risk when applying it to downstream tasks.

This paper explores the bottleneck of rotation-invariant correspondence matching and proposes a simple yet effective solution. We adopted a five-layer equivariant neural network. Initially, the input image is encoded into rotation-equivariant space. After feature extraction by the equivariant neural network, we use Group Pooling to transform rotation-equivariant features into rotation-invariant features. We employed the end-to-end training approach from R2D2, which generates reliability maps, repeatability maps, and descriptors separately from the rotation-invariant feature maps. For correspondence matching, our method takes one of the images and its rotated counterpart (rotated by 45 degrees) as input. Then, the output of the rotated image is projected back to the original image and overlaid with the original image's output. As shown in Fig. 1, on the left side of the image, we demonstrate ReF's performance curve as the rotation angle between two matching images varies from 0 to 360 degrees. In the middle, when one of the images and the convolutional kernel have a 45-degree rotation difference, the performance curve shifts along the x-axis by 45 degrees. The peaks of the performance curve in this case precisely align with the valleys in the left-side scenario. On the right side, when we combine the outputs of these two sets of matches, the performance curve exhibits a superimposed pattern of peaks and valleys. While fusing the outputs of the two paths, we also conducted ablation experiments, demonstrating that the correct filtering method brings further improvements. We conducted a comprehensive evaluation of our method for correspondence matching under significant rotations on the rotated-hpatches, SIM2E, and MegaDepth datasets. The results demonstrate that our approach exhibits superior performance, with a significant improvement compared to the baseline methods.

2. Related work

2.1. Traditional CNN-based methods

Traditional CNN-based methods for feature point matching have seen significant research progress. SuperPoint [15] employs data augmentation to enhance its robustness to rotation. However, even with training data augmented to cover a rotation range of 0 to 360 degrees, SuperPoint's performance still significantly deteriorates at 60 degrees. GIFT [19] uses image rotations or scaling to generate group-equivariant features and then collapses group dimensions through bilinear pooling to obtain invariant local descriptors. Its design, however, only considers the rotation range of ±90 degrees, making it unsuitable for downstream tasks with random rotations. RoRD [20] introduces a method for generating orthogonal views and employs it to improve the robustness of feature point matching under extreme viewing angles. In summary, traditional CNN-based methods attempt to approximate local features under different rotation angles, which makes the learning cost for achieving rotation robustness prohibitively high.

SuperGlue [21] leverages position information, self/cross-attention, and optimal transport to improve learning-based correspondence matching. SGMNet [22] is an improved version based on SuperGlue, which combines traditional methods like SIFT with deep learning techniques. It provides an open-source version that allows us to easily conduct comparative experiments.

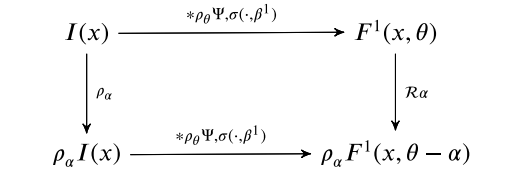
2.2. G-CNN-based methods

Equivariance is when a function, if a certain type of transformation is applied to its input, reflects that transformation in its output. Conversely, invariance is when a function's output remains unchanged when a specific transformation is applied to its input.

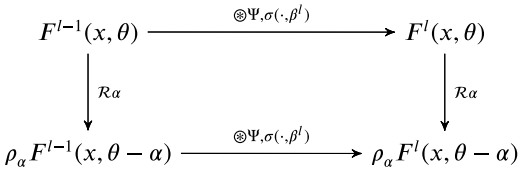
Group-equivariant convolutional neural networks (G-CNNs) are equivariant under a specific transformation (e.g., rotation, translation) which can also be represented by a special group. Researchers have designed G-CNNs using different mathematical approximations [17,23–28]. E2-CNN [26] is a general G-CNN framework that analyzes and models the orientation and symmetry of images.

The effectiveness of G-CNNs has shown promising results in areas such as point cloud processing, but its research in image feature point matching remains incomplete. ReF [18] replaces the backbone network of R2D2 with G-CNN and applies Group-Pooling to the output of the backbone network, transforming rotation-equivariant feature maps into rotation-invariant feature maps. However, the rotation-equivariant correspondence matching algorithm performance is very poor when rotating $\frac{\pi}{4}$, $\frac{3\pi}{4}$, $\frac{5\pi}{4}$ and $\frac{7\pi}{4}$.

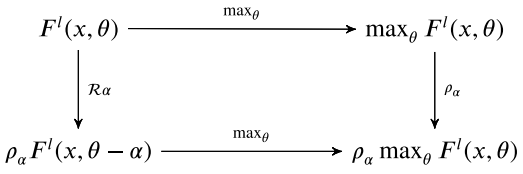
SE2LoFTR [29] employs G-CNN as its feature extraction backbone and, in comparison to LoFTR, demonstrates superior matching performance under significant rotation scenarios. It should be noted that our goal is to explore the key elements that affect the rotation-equivariance when using nearest neighbor (NN) based correspondence matching and improve the utilization strategy of G-CNNs, instead of proposing a new correspondence matching network. Therefore, this paper focuses entirely on improving the rotation-equivariant capability of correspondence matching based on NN.



(a) The lifting operation.



(b) The group convolution.



(c) The group (max) pooling operation.

Fig. 2. G-CNNs preliminaries.

3. Methodology

3.1. Preliminaries of G-CNNs for correspondence matching

This section presents the mathematical preliminaries of group-equivariant representations. The difference between group equivariance and invariance as well as their effectiveness on improving correspondence network performance is first provided. The strategy to apply G-CNNs to rotation-equivariant feature exaction in correspondence matching networks is then presented. Finally, we discuss the network architecture for group-equivariant correspondence matching.

3.1.1. Equivariance and invariance

Given a transformation group G, $\mathcal{T}$ is a linear representation of G, and $\mathcal{T}'$ is not required to be identical to $\mathcal{T}$. That is to say, $\mathcal{T}$ and $\mathcal{T}'$ may represent the same transform but acts on different space (*e.g.*, spatial coordinate space and group space) [17,30]. A group-equivariant function $\mathcal{F}_e : X \longrightarrow Y$ should observe that

$$\mathcal{F}_e(\mathcal{T}(x)) = \mathcal{T}'(\mathcal{F}_e(x)), \tag{1}$$

where $x \in X$, $\mathcal{T} \in G$. While the group-invariant operation $\mathcal{F}_i$ satisfies

$$\mathcal{F}_i(\mathcal{T}(x)) = \mathcal{F}_i(x). \tag{2}$$

3.1.2. Main steps of G-CNNs

Given an image I, a G-CNN backbone with L layers extracts the rotation-invariant feature maps. Let $F^l : G \to \mathbb{R}$ be the feature maps at layer l, $\Psi : G \to \mathbb{R}$ be a filter living on a group G, $*$ be the spatial convolution operator, $\circledast$ be the group convolution defined by $(F \circledast \Psi)(g) = \int_G F(h)\Psi(h^{-1}g)d\lambda(h)$ where $\forall h, g \in G$ and λ is the Haar measure, ρ_θ and ρ_α be the spatial rotation acts on an image or a feature map, and $\mathcal{R}\alpha$ be the group action defined by $\mathcal{R}_\alpha\Psi(\theta) := \rho_\alpha\Psi(\theta-\alpha)$ [30].

The first layer of G-CNNs is referred to as the lifting layer which lifts the input image from coordinate space to special group space. Then, group convolution is utilized for encoding the feature map in group space. Finally, the group max pooling operation is used for pooling the most relative orientation between input and filters. The equivariance of the lifting operation and group convolution procedure is shown in Figs. 2(a) and 2(b). The invariance of the group (max) pooling is shown in Fig. 2(c).

The equivariance of translation is common for CNNs and G-CNNs [17,30]. Therefore, the robustness of G-CNNs to SO(2) and SE(2) can be considered as the same problem. Therefore, we only consider SO(2) group in the remainder of this paper. Let $\Theta = \{0, \dots, 2\pi\frac{\Lambda-1}{\Lambda}\}$ be a total number of Λ equidistant orientations, where $\Lambda \in \mathbb{N}^+$. The implementation of the equivariant neural network is based on the assumption that signals are on a continuous domain $\mathbb{R}^2$. Then we can sample the orientations in Θ for the SO(2) group. The lifting operation is given by

$$F^1(x,\theta) = \sigma\left(\sum_{c=1}^{C}\left(I * \rho_\theta\Psi^1\right)(x) + \beta^1\right), \tag{3}$$

where x is the coordinate of a point in an image or a feature map, $\theta \in \Theta$ is one of the sampled orientations, β is the bias, and σ is the non-linear function.

After lifting operation, we obtain feature maps $F^l(\theta)$ on discrete group space. (3) can be rewritten as follows [30]:

$$\begin{aligned} F^l(x,\theta) &= \sigma\left(\sum_{c=1}^{C}\left(F^{l-1} \circledast \Psi^l\right)(x,\theta) + \beta^l\right) \\ &= \sigma\left(\sum_{c=1}^{C}\sum_{\phi\in\Theta}\left(F^{l-1}(\cdot,\phi) * \mathcal{R}_\phi\Psi^l(\cdot,\theta)\right)(x) + \beta^l\right). \end{aligned} \tag{4}$$

Then we get the lifting layer and group convolution layer. The feature maps generated from these layers are rotation-equivariant.

It is common to utilize multiple convolutional layers to extract a high-dimensional representation of local features. The feature maps generated from these layers should keep the rotation-equivariance. According to [18], the reason is that the group equivariance maintains the spatial structure of local features.

For visual correspondence matching, transforming rotation-equivariant feature maps to rotation-invariant is a key procedure. After the last group-convolutional layer, a max-pool operation is deployed. After group (max) pooling, the output feature map is transformed to the original space before lifting operation. Then the local features in this output feature map are rotation-invariant.

3.2. Dual rotation-equivariant activation mixer

This section analyzes the major factors limiting the performance upper bound of G-CNNs and details our proposed dual rotation-equivariant activation mixer that breaks the bottleneck of rotation-equivariant correspondence matching.

We believe the shape of the convolution kernel limits the rotation representation of networks at a fundamental level. Most computer vision tasks require a variety of visual features, and rarely require pure horizontal or pure vertical visual features. Therefore, the shape of the convolutional kernel in conventional CNNs is always square in computer vision. For example, the visual features in correspondence matching tasks are equally extracted from horizontal and vertical neighborhoods. The rotation operation takes into the loss of inliers (the part of pixel still inside the convolution kernel after rotation) and brings outliers (the part of pixel coming from outside the convolution kernel after rotation). Though the G-CNNs expand the representation of conventional CNNs, the performance of rotation-equivariance is still limited.

Therefore, we believe the overlapped area is the key to performance degradation, as shown in Fig. 4. When the local convolutional kernel rotates, the inlier ratio changes, highest at $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ and lowest at $\{\frac{\pi}{4}, \frac{3\pi}{4}, \frac{\pi}{4}, \frac{7\pi}{4}\}$. Our defined inlier ratio denotes the ratio of the overlapped area and kernel area, and the inlier ratio curve is assumed to be continuous in the spatial space of rotation. However, it is clear that the

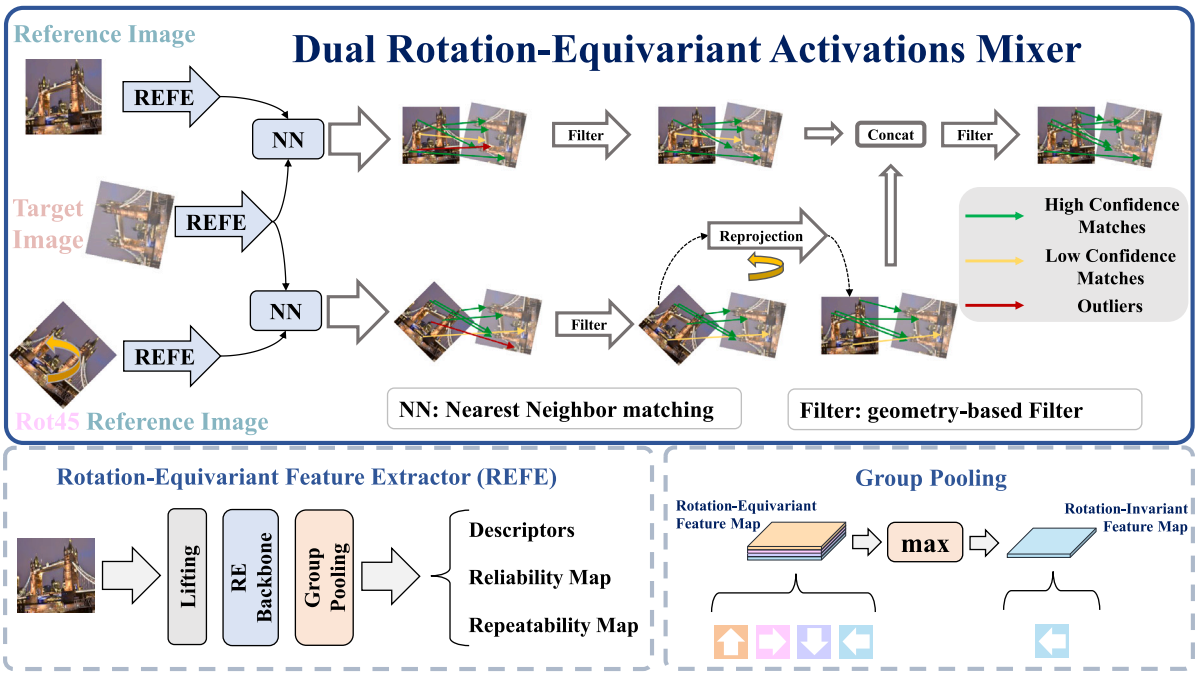


Fig. 3. The pipeline of the proposed approach. The original and the $\frac{\pi}{4}$-rotated images are simultaneously inserted in the rotation-equivariant local feature extractor. The results of two routes are generated by NN-based correspondence matching. Finally, the results of $\frac{\pi}{4}$-rotated route are reprojected and filtered with the original route. The group pooling operation is illustrated as the sequence changing inside the feature map. Then the pooled feature map is thought rotation-invariant.

relationship between the degradation of performance and the growth of loss area is not linear. The 15% loss of overlapped area is enough to degrade rotation-equivariant correspondence matching performance. Therefore, it is vital to get rid of the reduction of overlapped areas.

Our approach rotates the reference image $\frac{\pi}{4}$ only once, bringing high efficient way to boost the robustness of rotation-equivariance, as shown in Fig. 3. The image rotated by $\frac{\pi}{4}$ and the original image are two inputs of different routes. Then the outputs are obtained from the same $C4$-rotation-equivariant CNN. This procedure costs similar computation to the architectures with double group size but brings salient improvement rather than marginal changes. We collect correspondences and fuse them with three filter strategies based on random sample consensus (RANSAC).

The combination of two sets of $C4$-rotation-equivariant features can be likened to a superposition of periodic signals. The first route yields better matching results at angles of $0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$, while the other route performs better at angles of $\frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4}$. When the rotation angle between the reference image and the target image is close to $0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$, the first matching route generates significantly more accurate matches, whereas the second one does not. Conversely, if the rotation angle is close to $\frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4}$, the second matching route produces considerably more accurate matches, while the first one does not. Furthermore, when the rotation angle is near $\frac{\pi}{8}, \frac{3\pi}{8}, \frac{5\pi}{8}, \frac{7\pi}{8}, \frac{9\pi}{8}, \frac{11\pi}{8}, \frac{13\pi}{8}, \frac{15\pi}{8}$, both of these matching routes may result in numerous ambiguous matches in the spatial domain.

To address this issue, we employ different outlier filtering strategies to enhance the performance of rotation-equivariant correspondence matching. We tried three filtering strategies: the first one applies filtering separately to each of the two matching branches, the second applies filtering only after merging the two matching branches, and the third applies filtering both before and after merging the two matching branches.

3.3. Rotation-equivariant feature extractor

This section provides a detailed explanation of how G-CNNs are implemented for the task of correspondence matching, including the specific details of the implementation and the rationale behind the design choices.

We use E2-CNN [26] to establish a rotation-equivariant feature extraction network. We introduce the parameters information of G-CNNs to demonstrate that a large group is unsuitable for correspondence matching. The input of G-CNNs is an image $\boldsymbol{M_I} \in \mathbb{R}^{H\times W\times C}$, where W, H, and C denote the width, height, and the number of channels of the image. Then we define the group convolution filter at layer l as $\boldsymbol{M_{\Psi^l}} \in \mathbb{R}^{D^l\times S^l\times D^{l-1}\times S^{l-1}\times n\times n}$, where n is the size of kernel, D^{l-1} is the input dimension, D^l is the output dimension, S^{l-1} is the input group dimension, and S^l is the output group dimension. For SO(2) group, the standard rotation matrix $\boldsymbol{R}(\theta) \in \mathbb{R}^{2\times 2}$ is an example of a representation that acts on $\boldsymbol{I}$:

$$\boldsymbol{R}(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{5}$$

The feature map before group pooling $\boldsymbol{M_F} \in \mathbb{R}^{D^l\times S^l\times H\times W}$ and after group pooling $\boldsymbol{M_F} \in \mathbb{R}^{D^l\times H\times W}$. Thus, the group expansion brings high computation costs from the viewpoint of the feature map dimension. Furthermore, in the correspondence matching task, the transformations between image pairs are not SO(2) but Sim(2) and perspective transformation. The Sim(2) group operation consists of rotation, scaling, and translation. The perspective transformation can be represented by a 3×3 matrix with eight degrees of freedom. Therefore, these two groups are more complex for building group equivariant networks. However, achieving the perfect perspective transformation equivariance is unrealistic and inefficient. Therefore, it is very important to define the level of robustness we want to achieve. The SO(2) transformation is mainly concerned in this work. Because we believe learning the robustness of scaling and perspective transformation from data augmentation is easier. In computer vision, it is normal that use the cyclic group $C_N = \{0, \dots, 2\pi\frac{\Lambda-1}{\Lambda}\}$ to approximate continuous SO(2) group. For example, $C4 = \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$.

Additionally, referring to ReF [18] and C3PO [31], using more complex nonlinearity functions cannot bring improvements accordingly. Therefore, we focus on the inherent defects caused by the shape of images and convolutional kernels.

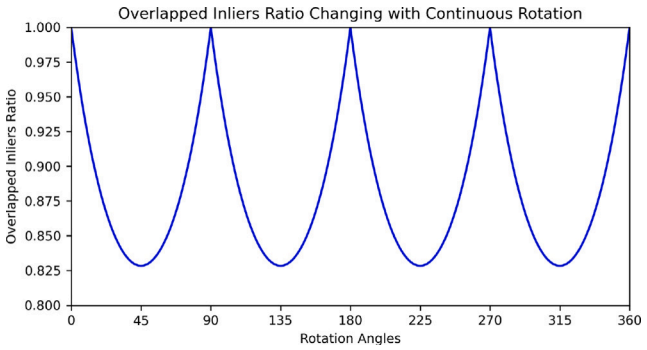


Fig. 4. Our defined inlier ratio of the convolutional kernel with continuous image rotation.

Similar to [18], we also use the network with five rotation-equivariant layers for rotation-equivariant visual features extraction. We follow the same pipeline of Revaud et al. [16], which extracts the positions and descriptions of interest points simultaneously. We mainly utilize $C4$ group equivariant convolution for sampling rotation angles from SO(2) group. The architecture is shown in Fig. 3.

We refer to the L2-Net [32] and adopt a 5-layers network structure. The dilation convolution [33] is used to obtain enough receptive fields without extra computation cost. However, if every layer utilizes dilation convolution, the feature extraction capability will be weak due to the gridding artifacts [34]. The first layer of our architecture is the lifting layer, which maps the image information from trivial representation to regular representation. Then, the next layer is an intermediate layer, which increases the depth of the network while keeping the rotation equivariance. At the third and fourth layers, we set the stride of dilation convolution to 2. The above four layers are all 3×3 rotation-equivariant convolution. At the final feature extraction layer, we use 3×3, 5×5, 7×7, and 9×9 rotation-equivariant convolutional kernels for ablation studies.

Additionally, we use the loss functions introduced in R2D2 [16]. These loss functions are enough to extract reliable and repeatable visual correspondences. Additionally, R2D2 achieves good performance while maintaining an easy network structure, which is very suitable for changing it to a G-CNN-based structure.

4. Experiments setup

We employed the official implementations of SuperPoint [15], R2D2 [16], GIFT [19], and RoRD [20] and their published weights in our experiments. ReF [18] and our proposed DREAM were implemented based on the G-CNNs introduced in [26]. We used an NVIDIA RTX 3090 GPU for model training and evaluation. Adam optimizer was used to minimize the loss function. The maximum epoch was set to 40. The batch size was set to 12. The learning rate was set to 0.0001. The weight decay was set to 0.0005. We also implemented an early-stopping mechanism to reduce over-fitting, where the model training will be terminated if the evaluation loss does not drop for five consecutive epochs.

4.1. Datasets

4.1.1. Training set

Following R2D2 [16], we used a synthetic dataset (basic images from the web) and a real dataset (the Aachen Day-Night dataset [35, 36]) to train the model.

4.1.2. Test sets

We use Rotated HPatches, SIM2E, and MegaDepth datasets for testing.

The HPatches [37] dataset is widely used to evaluate correspondence matching algorithms. It consists of 116 sequences of images, each of which has one reference image and five target images. Following C3PO [31] and ReF [18], we resize the images in the HPatches dataset to 300×300 pixels and rotate the images by every $\frac{\pi}{12}$. The Mean Matching Accuracy (MMA) is also commonly used for evaluating local feature matching tasks. A local feature is positive if projected to another image with a small location error between its matching feature. The pixel error threshold we used to show performance at all rotation angles is 3.

The SIM2E [38] dataset has three subsets named SIM2E-SO2S, SIM2E-SIM2S, and SIM2E-PersS. SO2S means rotation transformation only and synthetic background. SIM2S means rotation, scaling, translation transformations, and synthetic background. PersS means perspective transformation and synthetic background. These three subsets can help us comprehensively compare the rotation-equivariant capability of correspondence matching approaches. For SIM2E, we also use MMA as the evaluation metric. We show the performance in several thresholds: 1, 3, 5, and 10.

We also employ the MegaDepth [39] datasets to assess our suggested approach for relative camera pose estimation. MegaDepth comprises 1 million online images capturing 196 distinct outdoor scenes, and it offers the accurate real-world positioning information for every image within these scenes. Furthermore, MegaDepth supplies sparsely reconstructed data via COLMAP [40], along with depth maps derived from the multi-view stereo technique.

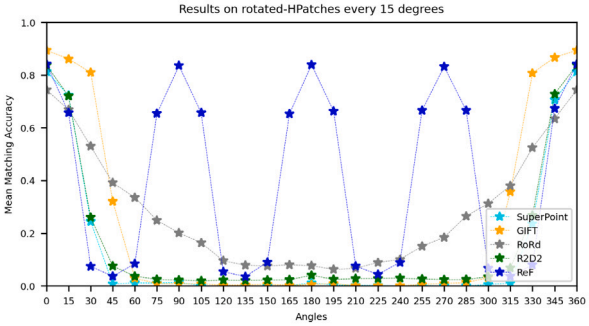
5. Results

5.1. Performance on the Rotated-HPatches dataset

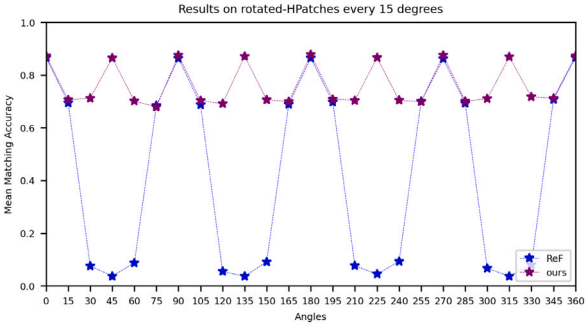
The results of learning-based correspondence matching approaches on Rotated-HPatches dataset are shown in Fig. 5(a). The SuperPoint [15] and R2D2 [16] only perform well in very narrow ranges, the performance range of GIFT [19] is slightly larger than SuperPoint, the performance of RoRD [20] drops slowly than SuperPoint and R2D2, and ReF has a good performance at several ranges around a multiple of $\frac{\pi}{2}$. It can be concluded that simply utilizing the G-CNNs rather than conventional CNN can greatly improve the rotation-equivariant capability of visual correspondences. In the range of $0 \sim \frac{\pi}{4}$ and $\frac{7\pi}{4} \sim 2\pi$, the performance of ReF is worse than original R2D2. The main reason is that the network learns limited knowledge from training data. Orientation is somehow a kind of local feature of an image.

Fig. 5(b) shows that our algorithm outperforms ReF. There is no obvious weakness in all rotation angles for our solution. The shape of the MMA curve in all angles also verified our previous view, fusing rotation-equivariant correspondences between the wave crest and wave trough.

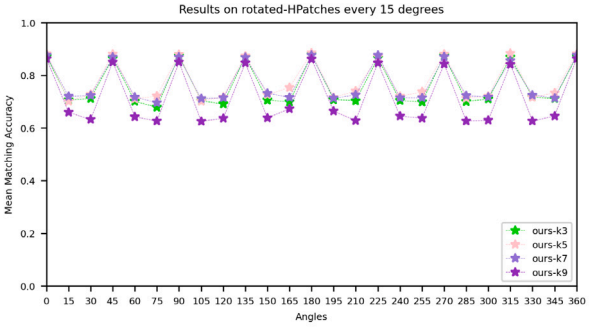
We display the feature maps with pseudo color mapping to prove the rationality of our method in Fig. 6. It can be observed that the activations of images with $\frac{\pi}{2}$ difference have similar visualizations. For example, image1 rotated by $-\frac{\pi}{4}$, image2 rotated by $\frac{\pi}{4}$ and image2 rotated by $\frac{3\pi}{4}$ have similar activations, and image1, image2 rotated by $\frac{\pi}{2}$ and image2 rotated by π have similar activations. These visualizations also demonstrate the rationality of rotating one of the input images with $\frac{\pi}{4}$. Besides, we show the activations of $\frac{\pi}{3}$ and $\frac{5\pi}{12}$. These two angles are existing between $\frac{\pi}{4}$ and $\frac{\pi}{2}$, where two angles has a $\frac{\pi}{4}$ difference. The appearance of pseudo maps changes gradually with the angle of $\frac{\pi}{4}$, $\frac{\pi}{3}$, $\frac{5\pi}{12}$ and $\frac{\pi}{2}$.



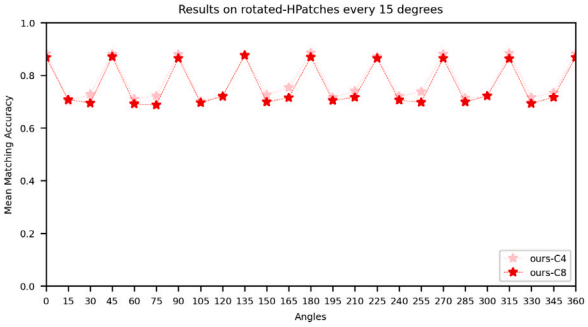
(a) The performance of SuperPoint, R2D2, GIFT, and ReF on the Rotated-HPatches dataset at different rotation angles.



(b) Comparison between our algorithm with the ReF on the Rotated-HPatches dataset.



(c) The effect of kernel size of the final layers of the feature extractor.



(d) The effect of the group size.

Fig. 5. MMA results over every $\frac{\pi}{12}$.

5.2. Performance on the SIM2E dataset

As shown in Fig. 7, our approach has a significant improvement in all three subsets of the SIM2E dataset. We compare our algorithm with R2D2 [16] to validate the effectiveness of the equivariant neural network. The RoRD [20] is considered for exploring the difference between data augmentation-based and G-CNNs-based approaches. We also visualized the matching results of our method and the compared methods in Fig. 8.

We provide results with different post-processing configurations for comprehensive evaluation and discussion. It can be observed that learning rotation-equivariance by data augmentation performs well at the SO2S dataset when the tolerance is large but fails on the SIM2S and PersS datasets. Our approach achieves significant overall performance while keeping a high accuracy on low thresholds of re-projection error tolerance.

The performance of SE2-LoFTR [29] and SGMNet [22] is also provided, which are not NN-based matching methods compared to others. SE2-LoFTR uses an end-to-end architecture based on a rotation-equivariant feature extractor, and SGMNet uses a classical rotation-equivariant feature extractor (SIFT [8]) and graph neural network. SGMNet only performs better than SE2-LoFTR at very small matching thresholds (1, 2). SE2-LoFTR has a strong capability, but our approach still has better performance under the thresholds (1, 2, 3, 4). That is to say, our method use only an NN matcher but has better performance for sub-pixel matching tasks. Furthermore, due to the detached module rather than end-to-end architectures, our approach has a more flexible way of deploying on many other applications.

5.3. Performance of camera pose estimation

Camera pose estimation is an important criterion for evaluating the quality of feature point matching. The relative pose between image pair consists of rotation and translation.

Following SuperGlue [21], we use MegaDepth-1500 dataset for relative pose estimation evaluation. To carefully evaluate the rotation-equivariance of the SoTA methods, we also provide two modified MegaDepth-1500 dataset, called MegaDepth-Rot90 and MegaDepth-RotRand. MegaDepth-Rot90 rotates the query image of image-pairs in $\frac{\pi}{2}$, π, $\frac{3\pi}{2}$, and 2π.

The estimated pose can be derived from matches by computing the essential matrix. Subsequently, we adhere to the same principle as presented in reference to determine the area under the curve (AUC) of pose error at specified thresholds (5 degrees, 10 degrees, 20 degrees). Here, the pose error is defined as the higher value between the angular error in rotation and translation. As the AUC of pose error employs RANSAC for pose estimation, it can arrive at an accurate pose estimate by disregarding numerous mismatches. However, it is important to note that the AUC of pose error might not comprehensively evaluate the matching method due to its reliance on RANSAC. To address this limitation, the precision of matches considers all matches, including mismatches, providing a more comprehensive assessment of the method's performance. Hence, we also adopt the approach outlined in SuperGlue to calculate the matching precision and utilize it as an additional metric for evaluating correspondence matching quality.

The experimental results are presented in Tables 2, 3, 4. It is easy to see that our method performs better than ReF. However, as our method is based on the Group-Pooling operation, we can only use a relatively simple network architecture. Otherwise, the dimension of the descriptor is too computationally expensive in the backbone network before pooling, so it is weaker than R2D2 on the MegaDepth-1500 dataset with less rotation. When the rotation angle is a multiple of 90 degrees, our method performs best. When the rotation is random, RoRD performs best and our method performs 2nd best. But RoRD performs not good enough when no rotation or the rotation angle is a multiple of 90 degrees. Especially on the MegaDepth-Rot-Rand dataset, our method exhibits nearly a twofold improvement in performance compared to ReF, validating that our Dual Activation Mixer fundamentally approaches a twofold performance increase.

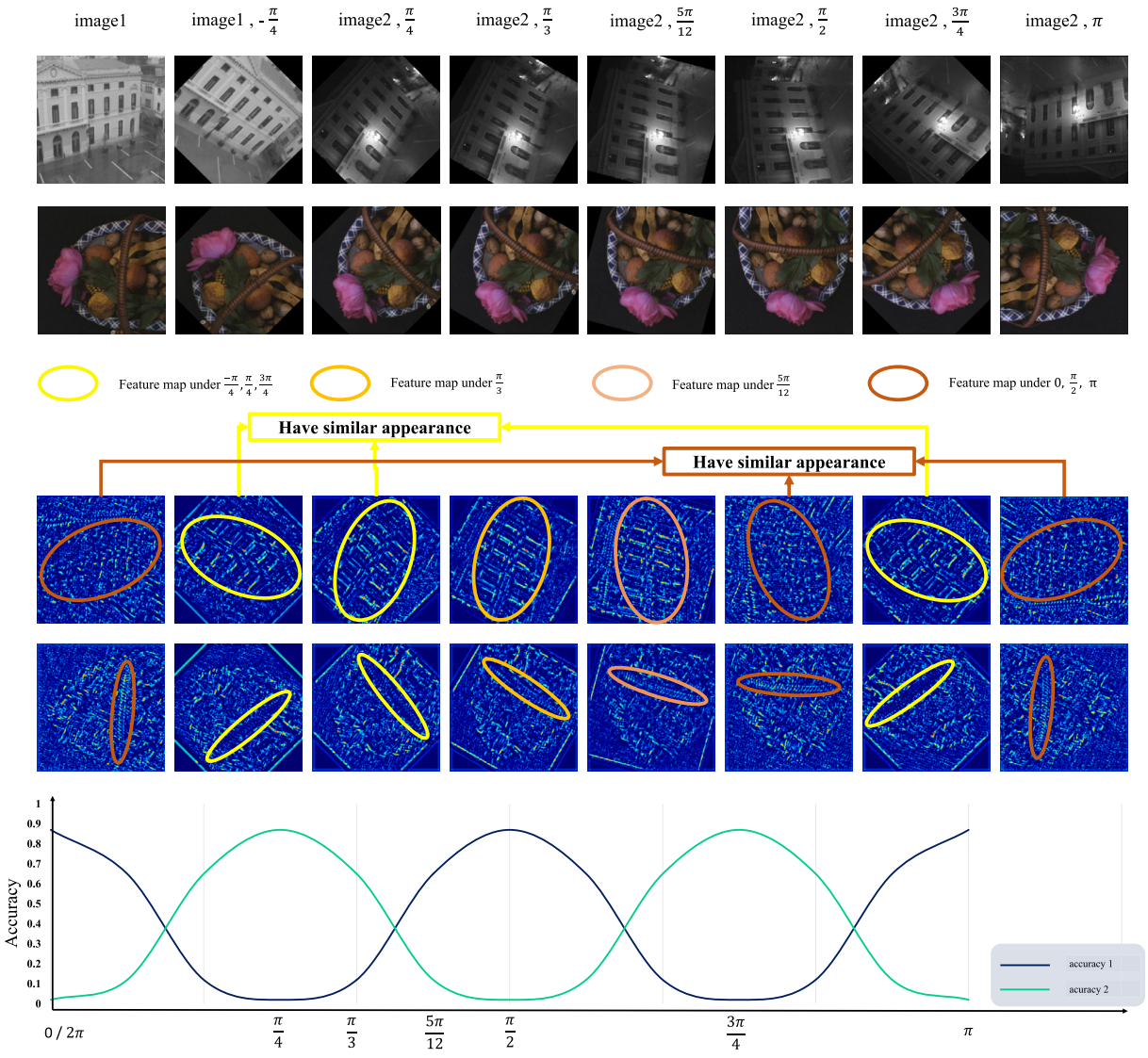


Fig. 6. We visualize the feature map of descriptors. Image1 and image2 have different illuminations. We observe that when the rotation angles between the two images are 45 degrees, 135 degrees, 225 degrees, and 315 degrees, the appearance of these feature maps is consistent. Similarly, when the rotation angles are 0 degrees, 90 degrees, 180 degrees, and 270 degrees, the appearance of these feature maps is also consistent.

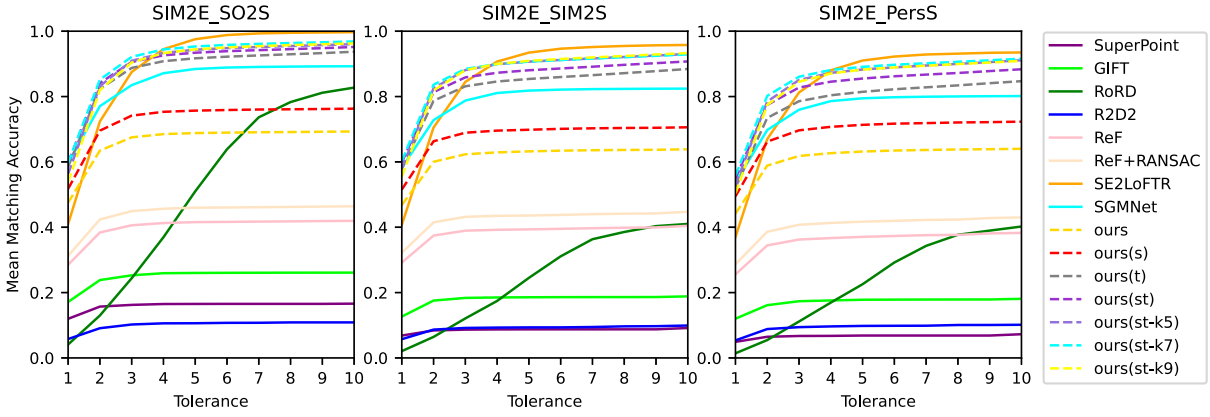


Fig. 7. Performance of SuperPoint, GIFT, RoRD, R2D2, ReF, SE2-LoFTR, SGMNet, and ours on three subsets of the SIM2E datasets.

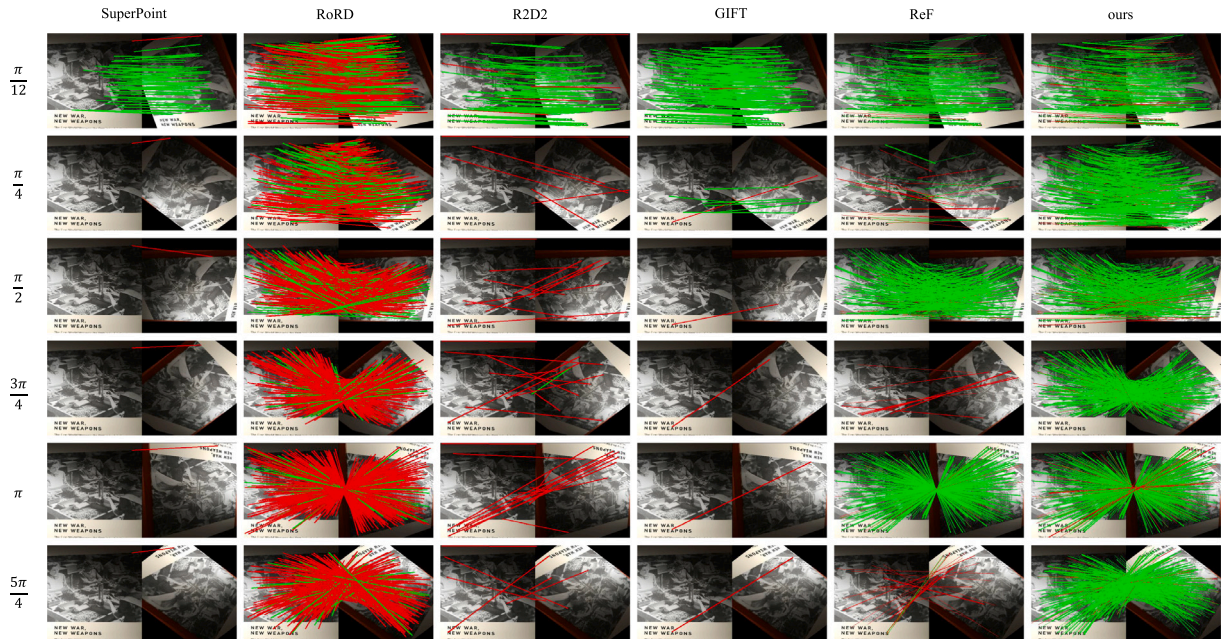


Fig. 8. Correspondence matching results under threshold 3 with respect to different rotation angles.

Table 1
MMA on three subsets of the SIM2E dataset. The params of s represents separately applying RANSAC on two collections of features from different routes, t represents applying RANSAC on all collected features, and st represents applying both s and t.

Dataset	Method	$\delta < 1$	$\delta < 3$	$\delta < 5$	$\delta < 10$
SO2S	SuperPoint [15]	0.120	0.162	0.165	0.166
	GIFT [19]	0.171	0.253	0.260	0.261
	RoRD [20]	0.041	0.244	0.510	**0.827**
	R2D2 [16]	0.058	0.102	0.106	0.109
	ReF [18]	0.285	0.406	0.416	0.420
	ours	**0.475**	**0.675**	**0.688**	0.693
	ReF+RANSAC	0.312	0.449	0.460	0.464
	ours(s)	0.517	0.742	0.757	0.763
	ours(t)	0.585	0.888	0.917	0.937
	ours(st)	**0.598**	**0.906**	**0.934**	**0.952**
	SE2LoFTR [29]	0.410	0.875	**0.975**	**0.997**
	SGMNet [22]	**0.596**	0.835	0.884	0.893
SIM2S	SuperPoint	0.068	0.087	0.087	0.092
	GIFT	0.126	0.184	0.186	0.188
	RoRD	0.021	0.121	0.245	0.410
	R2D2	0.057	0.092	0.093	0.099
	ReF	0.292	0.389	0.393	0.404
	ours	**0.468**	**0.623**	**0.632**	**0.638**
	ReF+RANSAC	0.322	0.432	0.436	0.447
	ours(s)	0.515	0.689	0.698	0.706
	ours(t)	0.578	0.831	0.854	0.884
	ours(st)	**0.599**	**0.859**	**0.880**	**0.907**
	SE2LoFTR	0.407	0.845	**0.934**	**0.958**
	SGMNet	0.582	0.788	0.818	0.824
PersS	SuperPoint	0.049	0.067	0.069	0.073
	GIFT	0.120	0.174	0.178	0.181
	RoRD	0.014	0.111	0.226	0.402
	R2D2	0.054	0.094	0.098	0.101
	ReF	0.256	0.362	0.371	0.383
	ours	**0.441**	**0.618**	**0.632**	**0.640**
	ReF+RANSAC	0.287	0.407	0.417	0.430
	ours(s)	0.495	0.697	0.713	0.723
	ours(t)	0.521	0.786	0.814	0.847
	ours(st)	**0.549**	**0.828**	**0.855**	**0.883**
	SE2LoFTR	0.370	0.812	**0.910**	**0.935**
	SGMNet	0.537	0.760	0.795	0.802

Table 2
The AUC and precision results on MegaDepth.

	MegaDepth			
Method	AUC@5	AUC@10	AUC@20	Prec
SuperPoint	29.43	45.93	61.03	71.04
GIFT	26.69	43.67	59.17	90.31
RoRD	14.42	28.57	44.8	90.39
R2D2	**35.07**	**52.83**	**68.03**	81.33
ReF(0.0)	16.81	29.35	42.89	39.53
ReF(0.7)	21.85	35.84	50.22	52.13
ours(0.0)	22.46	37.26	51.59	71.05
ours(0.7)	21.72	37.01	52.8	76.74

Table 3
The AUC and precision results on MegaDepth-Rot90.

	MegaDepth-Rot90			
Method	AUC@5	AUC@10	AUC@20	Prec
SuperPoint	8.3	12.79	17.23	29.8
GIFT	6.81	10.68	14.41	31.61
RoRD	11.6	25.01	41.67	89.21
R2D2	7.13	12.11	16.63	23.95
ReF(0.0)	16.3	28.75	42.77	39.57
ReF(0.7)	20.9	34.39	49.52	52.19
ours(0.0)	**21.78**	36.06	50.73	71.56
ours(0.7)	21.19	**36.12**	**51.95**	76.8

Table 4
The AUC and precision results on MegaDepth-Rot-Rand.

	MegaDepth-Rot-Rand			
Method	AUC@5	AUC@10	AUC@20	Prec
SuperPoint	5.27	8.36	11.38	24.29
GIFT	5.01	8.5	11.99	30.55
RoRD	11.08	**23.5**	**39.96**	89.17
R2D2	2.27	4.2	6.3	13.84
ReF(0.0)	4.68	9.09	15.08	20.99
ReF(0.7)	6.03	11.1	17.27	25.5
ours(0.0)	**11.49**	21.6	34.34	57.08
ours(0.7)	10.96	21.25	34.36	59.56

Table 5
Runtime of our method and baselines.

Method	Runtime(s)
R2D2	1.924
ReF	0.714
ours	1.187
ours(s)	1.195
ours(t)	1.194
ours(st)	1.251

5.4. Ablation studies

5.4.1. Kernel size

To validate that kernel size is not the main reason limiting the performance of rotation-equivariant local features matching, we compare our solution with kernel sizes 3 and 9. As shown in Fig. 5(c), the performance of kernel size 9 is worse than kernel size 3. The improved kernel size does not bring more robustness of rotation. On the contrary, the large kernel size makes the local feature more blurred.

To furthermore explore the effect of convolutional kernel size, we utilize models with different kernel sizes (3, 5, 7, 9) in the last layer of the feature extractor encoder and test them on the SIM2E dataset Table 6. It can be observed there are marginal changes between different kernel sizes. Kernel size 3 performs better at very tight matching thresholds, and kernel sizes (5, 7, 9) perform better at larger matching thresholds. A reason is that large kernel size leverages a larger receptive field, which is more robust to perspective transformation. Sim(2) and perspective transformations commonly exist in the camera localization task. Therefore, the larger kernel size is more robust to this application scenario, and the smaller kernel size is more suitable for SE(2) conditions.

5.4.2. Group size

To validate the marginal gains from improving the group size of G-CNNs, we compare our solution with group $C4$ and $C8$ on the Rotated-HPatches dataset, as shown in Fig. 5(d), it can be observed that the rotation-equivariant correspondence matching performances when using group sizes of $C4$ and $C8$ are similar. Therefore, it is futile to increase group size larger than $C4$ blindly.

5.4.3. Filtering strategies

As shown in Table 1, we compare several post-processing strategies that benefit our algorithm. The results show that ours(t) outperforms ours(s), which means the two collections of dual rotation-equivariant adaptive correspondence matching are not always able to cooperate well. Therefore, total RANSAC on two collections of correspondences is an effective way to bring about further improvement. The performance of total RANSAC and separate RANSAC with total RANSAC is very similar, which also supports our viewpoint that total RANSAC is vital for dual activation mixer.

5.5. Efficiency analysis

The efficiency of an algorithm is crucial for its deployability in practical applications. When matching a pair of images, R2D2 and ReF perform two forward passes through the network. In our algorithm, we perform three forward passes, which theoretically introduces an acceptable additional computational cost. We conducted efficiency tests on the MegaDepth dataset, where the longer side of the images was resized to 1200 pixels. The resulting runtimes are presented in the Table 5. Our method provides a significant performance improvement, and the increased computational cost compared to the baseline is acceptable (see Table 5).

Table 6
MMA of our method with different kernel sizes on three subsets of the SIM2E dataset.

Dataset	Kernel_Size	$\delta < 1$	$\delta < 3$	$\delta < 5$	$\delta < 10$
SO2S	3	**0.598**	0.906	0.934	0.952
	5	0.563	0.909	0.943	0.960
	7	**0.596**	**0.921**	**0.953**	**0.969**
	9	0.538	0.904	0.944	0.963
SIM2S	3	**0.599**	0.859	0.880	0.907
	5	0.579	**0.881**	**0.907**	**0.930**
	7	**0.603**	**0.885**	**0.907**	**0.929**
	9	0.554	**0.881**	**0.909**	**0.933**
PersS	3	0.549	0.828	0.855	0.883
	5	0.529	0.851	0.883	0.911
	7	**0.558**	**0.861**	**0.891**	**0.916**
	9	0.504	0.845	0.882	0.909

6. Discussions

In this work, the proposed dual rotation-equivariant activation mixer expands the rotation-equivariant capability from several fixed angles to all angles. The two different matching routes are mixed in the right way. However, there are some possible limitations, as below. First, the selected baseline [16] is convenient for implementation and validation but not fast enough. The main reason is there are no downsampling and upsampling operations in the backbone. Additionally, we follow R2D2 [16] to extract multi-scale features by resizing the input image multiple times. These results in much computation cost. In future work, we will focus on improving the computation efficiency of rotation-equivariant correspondence matching algorithm at all angles. For example, leveraging knowledge distillation to obtain more lightweight models.

Second, the features mix module relies on our post-processing procedure based on RANSAC. The rotation-equivariant dense feature map itself is still only rotation-equivariant on several fixed angles. In the future, we plan to explore techniques such as deformable convolutions to investigate convolution operations that preserve pixel-local information with minimal loss and exhibit greater robustness to rotations.

Third, our current feature point extractor relies solely on the receptive fields brought by layer-wise convolutions, lacking certain global information. In the future, we intend to explore ways to introduce global information while ensuring local feature rotation equivariance to further enhance feature point matching performance.

7. Conclusions

This paper proposes that the shape of kernels and the storage format of digital images constrain the rotation-equivariant capabilities of learning-based correspondence matching methods. We have validated our hypothesis through quantitative performance metrics in the preliminary experimental results and visualization of network feature maps. Therefore, we propose a simple yet highly effective solution: the dual activation mixer. This solution increases the inference computation by only 50%, yet it significantly enhances the performance of correspondence matching under large rotation conditions. Experimental results demonstrate the effectiveness of our algorithm. Our proposed dual activation mixer can also be applied to numerous other computer vision applications that require rotation-equivariant features. For example, tasks such as rotation object detection and scene re-identification from a bird's-eye view perspective.

CRediT authorship contribution statement

Shuai Su: Conceptualization, Methodology, Coding, WWriting – original draft. **Ronghao Dang:** Writing – review & editing. **Rui Fan:** Supervision, Writing – review & editing. **Chengju Liu:** Supervision, Writing – review & editing. **Qijun Chen:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Qijun Chen reports financial support was provided by Tongji University.

Data availability

Data will be made available on request.

References

[1] J. Ma, et al., Image matching from handcrafted to deep features: A survey, Int. J. Comput. Vis. 129 (2021) 23–79.
[2] X. Jiang, et al., A review of multimodal image matching: Methods and applications, Inf. Fusion 73 (2021) 22–71.
[3] J. Ma, et al., LMR: Learning a two-class classifier for mismatch removal, IEEE Trans. Image Process. 28 (8) (2019) 4045–4059.
[4] X. Jiang, et al., Robust feature matching using spatial clustering with heavy outliers, IEEE Trans. Image Process. 29 (2019) 736–746.
[5] X. Jiang, et al., Robust feature matching for remote sensing image registration via linear adaptive filtering, IEEE Trans. Geosci. Remote Sens. 59 (2) (2020) 1577–1591.
[6] X. Jiang, et al., Learning for mismatch removal via graph attention networks, ISPRS J. Photogramm. Remote Sens. 190 (2022) 181–195.
[7] A. Fan, et al., Efficient deterministic search with robust loss functions for geometric model fitting, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2021) 8212–8229.
[8] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, Vol. 2, IEEE, 1999, pp. 1150–1157.
[9] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[10] S. Leutenegger, et al., BRISK: Binary robust invariant scalable keypoints, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2548–2555.
[11] J. Guo, et al., Learning for feature matching via graph context attention, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–14.
[12] Z. Shi, et al., JRA-Net: Joint representation attention network for correspondence learning, Pattern Recognit. 135 (2023) 109180.
[13] S. Chen, et al., SSL-Net: Sparse semantic learning for identifying reliable correspondences, Pattern Recognit. (2023) 110039.
[14] X. Liu, et al., Pgfnet: Preference-guided filtering network for two-view correspondence learning, IEEE Trans. Image Process. 32 (2023) 1367–1378.
[15] D. DeTone, et al., Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 224–236.
[16] J. Revaud, et al., R2d2: Reliable and repeatable detector and descriptor, in: Advances in Neural Information Processing Systems, Vol. 32, 2019.
[17] T. Cohen, et al., Group equivariant convolutional networks, in: International Conference on Machine Learning, PMLR, 2016, pp. 2990–2999.
[18] A. Peri, et al., ReF–Rotation equivariant features for local feature matching, 2022, arXiv preprint arXiv:2203.05206.
[19] Y. Liu, et al., Gift: Learning transformation-invariant dense visual descriptors via group cnns, Adv. Neural Inf. Process. Syst. 32 (2019).
[20] U.S. Parihar, et al., RoRD: Rotation-robust descriptors and orthographic views for local feature matching, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2021, pp. 1593–1600.
[21] P.-E. Sarlin, et al., Superglue: Learning feature matching with graph neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4938–4947.
[22] J. Xu, et al., SGMNet: Learning rotation-invariant point cloud representations via sorted gram matrix, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10468–10477.
[23] C. Esteves, et al., Polar transformer networks, 2017, arXiv preprint arXiv:1709.01889.
[24] D. Marcos, et al., Rotation equivariant vector field networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5048–5057.
[25] T.S. Cohen, et al., Spherical CNNS, 2018, arXiv preprint arXiv:1801.10130.
[26] M. Weiler, et al., General e (2)-equivariant steerable CNNS, Adv. Neural Inf. Process. Syst. 32 (2019).
[27] M. Finzi, et al., Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data, in: International Conference on Machine Learning, PMLR, 2020, pp. 3165–3176.
[28] L. He, et al., Efficient equivariant network, Adv. Neural Inf. Process. Syst. 34 (2021) 5290–5302.
[29] G. Bökman, et al., A case for using rotation invariant features in state of the art feature matchers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5110–5119.
[30] M. Weiler, et al., Learning steerable filters for rotation equivariant CNNS, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 849–858.
[31] P. Bagad, et al., C-3PO: Towards rotation equivariant feature detection and description, in: 3rd Visual Inductive Priors for Data-Efficient Deep Learning Workshop, 2022, URL: https://openreview.net/forum?id=dXouQ9ubkPJ.
[32] Y. Tian, et al., L2-net: Deep learning of discriminative patch descriptor in Euclidean space, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 661–669.
[33] F. Yu, et al., Multi-scale context aggregation by dilated convolutions, 2015, arXiv preprint arXiv:1511.07122.
[34] Z. Wang, et al., Smoothed dilated convolutions for improved dense prediction, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2486–2495.
[35] T. Sattler, et al., Image retrieval for image-based localization revisited, in: BMVC, Vol. 1, no. 2, 2012, p. 4.
[36] T. Sattler, et al., Benchmarking 6dof outdoor visual localization in changing conditions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8601–8610.
[37] V. Balntas, et al., HPatches: A benchmark and evaluation of handcrafted and learned local descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5173–5182.
[38] S. Su, et al., SIM2E: Benchmarking the group equivariant capability of correspondence matching algorithms, 2022, arXiv preprint arXiv:2208.09896.
[39] Z. Li, et al., Megadepth: Learning single-view depth prediction from internet photos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2041–2050.
[40] J.L. Schonberger, et al., Structure-from-motion revisited, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4104–4113.

Shuai Su received the B.Sc. degree in Electronic and Information Engineering from Zhengzhou University in 2020. He is currently pursuing the Ph.D. degree (supervised by Prof. Qijun Chen) with the Artificial Intelligence Laboratory (RAIL) at Tongji University. His research interests include visual correspondence matching, equivariant representation learning, and online extrinsic calibration.

Ronghao Dang is currently pursuing the M.S. degree with the Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University. His research interests include action recognition and Vision-and-Language Navigation.

Rui Fan(Member, IEEE) received the B.Eng. degree in automation from the Harbin Institute of Technology in 2015 and the Ph.D. degree in electrical and electronic engineering from the University of Bristol in 2018. He worked as a Research Associate at the Hong Kong University of Science and Technology from 2018 to 2020 and a Postdoc Fellow at the University of California San Diego between 2020 and 2021. He is currently a (full) Research Professor at Tongji University and Shanghai Research Institute for Intelligent Autonomous Systems. His research interests include computer vision, machine learning, and robotics.

Chengju Liu received the Ph.D. in Control Theory and Control Engineering from Tongji University, Shanghai, China, in 2011. From October 2011 to July 2012, she worked in the BEACON Centre at Michigan State University, East Lansing, USA, as a research associate. She is currently a professor with the College of Electrical and Information Engineering, Tongji University, Shanghai, China. Her research interests include intelligent control, motion control of legged robots, and evolutionary computation.

Qijun Chen (Senior Member, IEEE) received the B.S. degree in automation from Huazhong University of Science and Technology, Wuhan, China, in 1987, the M.S. degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1990, and the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 1999. He is currently a Full Professor in the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include robotics control, environmental perception, and understanding of mobile robots and bioinspired control.