DRKF: Distilled Rotated Kernel Fusion for Efficiently Boosting Rotation Invariance in Image Matching

Chao Li\textsuperscript{1}, Jiancheng Cai\textsuperscript{2}, Ranran Huang\textsuperscript{2*}, Xinmin Liu\textsuperscript{2}

Abstract—Most existing learning-based image matching pipelines are designed for better feature detectors and descriptors which are robust to repeated textures, viewpoint changes, etc., while little attention has been paid to rotation invariance. As a consequence, these approaches usually demonstrate inferior performance compared to the handcrafted algorithms in circumstances where a significant level of rotation exists in data, due to the lack of keypoint orientation prediction. To address the issue efficiently, an approach based on knowledge distillation is proposed for improving rotation robustness without extra computational costs. Specifically, based on the base model, we propose Multi-Oriented Feature Aggregation (MOFA), which is subsequently adopted as the teacher in the distillation pipeline. Moreover, Rotated Kernel Fusion (RKF) is applied to each convolution kernel of the student model to facilitate learning rotation-invariant features. Eventually, experiments show that our proposals can generalize successfully under various rotations without additional costs in the inference stage.

I. INTRODUCTION

Image matching which aims to build correspondences across images with overlapping areas is essential for multiple computer vision tasks, including Structure from Motion (SfM) \cite{1}, Simultaneous Localization and Mapping (SLAM)\cite{2}, etc.

In the early stage, image matching tasks were mostly based on handcrafted features, including SIFT \cite{3}, SURF \cite{4}, ORB \cite{5}, etc. These methods explicitly assign one or more orientations to each keypoint based on which the future operations are performed, thereby providing a degree of invariance to orientation as shown in Fig. 1(a). However, they lack the ability to cope with the complex cases caused by repeated textures, viewpoint changes, etc.

In recent years, numerous studies that use learning-based pipelines have proved their potential in dealing with the complicated cases. In terms of feature detection and description pipelines, existing learning-based approaches can be classified into two categories. The first branch, including LIFT \cite{6}, LF-Net \cite{7}, and RF-Net \cite{8}, follows the idea of SIFT \cite{3} and attempts to predict the shape parameters (including orientations, scales, etc.) for the detected feature points. Subsequently, image patches around these key points are cropped and warped based on corresponding shape properties for further description. The other branch, including methods such as SuperPoint \cite{9}, D2-Net \cite{10}, R2D2 \cite{11}, ASLFeat \cite{12}, as well as DISK \cite{13}, directly locates the feature points and computes their descriptors using convolution neural networks (CNN) \cite{14} without considering orientations.

For the first branch of learning-based pipelines, predicting feature shapes such as orientation and scale is difficult due to the lack of explicit labels, and the rotation invariance is barely satisfactory accordingly, as shown in Fig. 1(b). Worse still, extracting descriptors from those cropped patches brings huge computational costs, especially when sampling patches around dense points generates considerable overlapping areas. For the other branch, features are directly extracted without taking the orientation into account, therefore the rotation invariance is even worse, as shown in Fig. 1(c).

Rotation invariance is essential in the image matching task, especially under circumstances where cameras move in 3D space and severe rotations are commonplace, such as in

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Visualized matching results of SIFT, LF-Net and ASLFeat under rotation of $\pi$. The correspondences colored green are inliers while those colored red are outliers.}
\end{figure}
drone-view localization tasks. To boost rotation invariance efficiently, we propose a framework based on knowledge distillation (KD) to transfer knowledge from the teacher which is robust to orientation changes, to the student in a computationally efficient way. Specifically, our contributions can be concluded in the following aspects:

- Compared to some classical methods that predict orientations explicitly, we propose Multi-Oriented Feature Aggregation (MOFA) shown in Fig. 2(a), which obtains diverse orientation information from the multi-branch input and is further used as a teacher model in the knowledge distillation pipeline. To be specific, we introduce multiple rotation transformations to the input images of the base model, and implement an aggregation of the corresponding extracted features.
- As for the student model, we propose Rotated Kernel Fusion (RKF) shown in Fig. 2(c) and Fig. 4 to impose kernel rotations and feature fusion on each convolution layer of the base model. By means of RKF, we enhance the learning capability of the student model and improve knowledge transfer efficiency. Eventually, reparameterization is adopted, by which computational costs remain the same in the inference stage.
- Instead of directly transferring knowledge from MOFA to the base model, we propose to adopt the RKF model as the student to learn from MOFA, and the distilled model is referred to as Distilled Rotated Kernel Fusion (DRKF) which is shown in Fig. 2. According to our experimental results, our DRKF improves rotation invariance considerably without introducing extra computational costs in the inference phase. To make quantitative comparisons, we augment HPatches [15] and YFCC100M [16] with diverse rotations, and the proposed DRKF outperforms other state-of-the-art approaches remarkably.

II. RELATED WORKS

A. Feature Rotation Invariance

a) Handcrafted approaches: To improve rotation robustness, conventional methods compute the orientations of detected features, followed by the descriptor construction process. There are different ways to obtain the orientations. For example, SIFT [3] and SURF [4] first select the feature points using local extrema detection in the pyramid of Difference of Gaussian (DoG) and compute orientations with the histogram of oriented gradients. When constructing descriptors, the coordinate axis for each feature point is rotated according to its orientation. Some other methods extract the orientation based on intensity. For instance, ORB [5] firstly detects feature points with FAST [17] in different scales and then takes the direction from the corner’s center to the intensity centroid in a local window as the orientation.

b) Learning-based approaches: There are two branches of learning-based approaches. Methods including LIFT [6], LF-Net [7], RF-Net [8], etc., attempt to imitate the conventional algorithms and predict the orientation and scale parameters. These usually consist of the following steps: feature detection (including positions, orientations as well as scales),
patch sampling with spatial transform network (STN) [18], and feature description. Due to the lack of explicit labels, it is difficult to regress feature scales and orientations. In addition, patch sampling and warping for numerous detected feature points are computationally expensive, and overlapping areas among the patches also cause redundant computation. On the contrary, methods from the other branch, such as SuperPoint [9], D2-Net [10] and ASLFeat [12], do not take shape parameters (including orientations, scales, etc.) into consideration, therefore fail to demonstrate steady performance under various rotation transformations.

B. Knowledge Distillation (KD)

Knowledge Distillation (KD) is firstly proposed by Hinton et al. [19], which supervises a light student model with a cumbersome teacher model and is commonly adopted for model compression. The major concern of KD is how to transfer the knowledge from the large model to the small model. Based on the forms of knowledge, current methods can be summarized into three categories: response-based, feature-based, and relation-based [20]. KD has proved to be an effective method in multiple tasks, including classification [21], [22], object detection [23], [24] and semantic segmentation [25], [26], etc.

In this paper, we propose to employ the distillation strategy in the image matching field and adopt MOFA as the teacher model, and our proposed RKF model as the student model. As the aggregation is conducted on the output layer, we transfer response-based knowledge from MOFA to the student model.

C. Reparameterization

Reparameterization is a significant technique that has been frequently used for boosting performance while reducing computational costs in the inference stage, including methods such as RepLkNet [27], RepVGG [28], ACNet[29], etc. Generally, models with multi-branch aggregation demonstrate stronger learning ability than the ones that only contain a single one [30], [28]. Nevertheless, these multi-branch models are usually computationally expensive. To improve inference efficiency, reparameterization is adopted to integrate the multiple branches equivalently.

For our proposed RKF model, a multi-branch structure is applied within each convolution layer for better representation ability, and with the reparameterization approach, the model can be equally slimmed to the original single-branch structure, thus reducing the computation cost considerably.

III. METHODS

In this paper, we consider incorporating diverse orientation information into the base model to improve rotation invariance. Specifically, we first propose the Multi-Oriented Feature Aggregation (MOFA) to integrate features of the base model based on multi-oriented inputs. To be more computationally efficient, we make the most of MOFA by applying it as a teacher to supervise the base model in a manner of knowledge distillation. Furthermore, due to the limited representation ability of the base model, Rotated Kernel Fusion (RKF) is developed which imposes rotations on each convolution kernel of the base model. Contrary to MOFA, it won’t introduce any extra computation cost in the inference phase with subsequent reparameterization process.

A. Multi-oriented feature aggregation (MOFA)

Inspired by ensembling methods [31], [32], we propose MOFA which conducts feature aggregation on multi-oriented features to boost rotation invariance of feature detectors and descriptors. Firstly, the base model shown in Fig. 2(b) is trained with rotation augmentation, and its features based on multi-oriented inputs are subsequently integrated by MOFA. Given the input image denoted as $I$, MOFA can be theoretically formulated as:

$$F_{mofa}(I) = \frac{1}{2\pi} \int_0^{2\pi} g^{-\theta} (F_{base}(g^\theta(I))) \, d\theta, \quad (1)$$

where $g^\theta$ refers to rotation operation with radian $\theta$, $g^{-\theta}$ is the corresponding inverse operation, and $F_{base}(\cdot)$ represents the base model that projects the input to the descriptor or score map. According to Eq. 1 features extracted from all-oriented inputs are aggregated, therefore the rotation robustness can be improved.
However, integration in Eq. 1 is intractable, and an alternative pipeline is shown in Fig. 2(a). The images are rotated by $n$ different radians respectively, and then fed into the base model to get the corresponding outputs, which are further aggregated as:

$$ F_{mofa}(I) = \frac{1}{n} \sum_{i=0}^{n-1} g^{-\theta_i} (F_{base}(g^{\theta_i}(I))), $$ (2)

where $\theta_i$ is the rotation radian uniformly sampled from $[0, 2\pi)$.

We can understand MOFA intuitively from Fig. 2. Multiple rotation transformations convert the original large orientation gap to multiple smaller gaps, which makes it much easier for the model to match keypoints with severe orientation changes, thereby generating more rotation-invariant features.

B. Knowledge distillation (KD)

For MOFA, the increase in rotation invariance comes at a cost of multiple feedforwards, making it computationally expensive for embedded devices. On the other hand, for the base model in Fig. 2(b), although it can be trained under rotation augmentation to obtain better generalization ability for rotation, there is still large room for improvement according to our following experimental verifications. Therefore, we propose to utilize knowledge distillation and take advantage of MOFA to transfer knowledge to the base model, where MOFA serves as the teacher and the base model acts as the student. In the KD stage, training loss for the student consists of the general part and the distillation part, as explained in the following.

a) General loss for detection-and-description: We follow most existing detection-and-description approaches like D2-Net [10] and ASLFeat [12] and adopt the loss function as Eq. 3 to jointly optimize the detection and description objectives, which is

$$ L_{ori} = \frac{1}{\hat{C}} \sum_{c \in \hat{C}} \sum_{g \in C} \sum_{q \in C} s_c s'_c \mathcal{M}(f_c, f'_c), $$ (3)

where $\hat{C}$ is the ground-truth correspondence set, $s_c$ and $s'_c$ are corresponding scores, $f_c$ and $f'_c$ refer to corresponding descriptors, and $\mathcal{M}(\cdot, \cdot)$ is the ranking loss for representation learning, which is usually defined as triplet [33] or contrastive loss. [34].

b) Distillation loss in knowledge transfer: The distillation loss is imposed on the output of the teacher and student model. Suppose the normalized dense descriptor map and score map generated by the teacher model are $D^{(t)}$ and $S^{(t)}$ respectively, and those generated by the student model are denoted as $D^{(s)}$ and $S^{(s)}$ respectively, thus the size of $D$ is $\frac{H}{r} \times \frac{W}{r} \times C$, and the size of $S$ is $H \times W \times 1$, where $H$ and $W$ refer to height and width of the input image respectively, and $r$ is the downsampling rate. The distillation loss of descriptors is defined in the form of mean distance error, which is formulated as:

$$ L_{dis}^{(desc)} = \frac{r^2}{HW} \sum_{i,j} \left\| D^{(s)}_{ij} - D^{(t)}_{ij} \right\|_2, $$ (4)

And the distillation loss of the score map is in the form of local cross-entropy loss. We reshape the two score maps $S^{(s)}$ and $S^{(t)}$ to $\frac{H}{r} \times \frac{W}{r} \times r^2$ with inverse pixel shuffle [35] operation, and define the distillation loss of the score map as:

$$ L_{dis}^{(score)} = - \frac{r^2}{HW} \sum_{i,j} \sum_{k=1}^{r^2} P^{(t)}_{ijk} \log P^{(s)}_{ijk}, $$ (5)

where

$$ P_{ijk} = \exp(S_{ijk}) / \sum_{k} \exp(S_{ijk}). $$ (6)

The total distillation loss is a combination of Eq. 4 and Eq. 5 which is

$$ L_{dis} = L_{dis}^{(desc)} + \lambda_1 L_{dis}^{(score)}. $$ (7)

Therefore, the total training loss of our method is formulated as:

$$ L = L_{ori} + \lambda_2 L_{dis}. $$ (8)

Compared to approaches where only general loss in Eq. 3 is adopted in training, we innovatively employ the distillation strategy, which introduces extra supervision and transfers knowledge of rotation invariance from MOFA to the base model.

C. Distilled Rotated kernel fusion (DRKF)

Although the base model can learn rotation-invariant features from MOFA through knowledge distillation, it still lacks representation ability for rotation invariance due to the inherent nature of convolution kernels. To this end, we apply Rotated Kernel Fusion (RKF) which is shown in Fig. 4 to each convolution kernel to act similarly to MOFA. Contrary to MOFA which instills orientation information in the input level and aggregate features in a multi-feedforward manner, RKF is applied within each convolution stage and employs multi-oriented kernels to conduct convolution. Because reparameterization can be further used to fuse multiple kernels into one for each convolution layer, RKF won’t introduce extra computation in the inference phase. Overall, instead of using the base model as the student model, we adopt the enhanced student model, which is the RKF model, as the student to learn from MOFA, and we refer to the distilled model as Distilled Rotated Kernel Fusion (DRKF).

Concretely, the RKF is depicted in Fig. 4. We rotate the conventional kernel by $n$ different radians, and the corresponding feature maps are concatenated and convolved with a $1 \times 1$ kernel for dimension reduction and eventually merged with the original feature using a learnable parameter $z \in [0, 1]$. The formulation can be written as:

$$ y = zf^{\theta_0}(x) + (1-z)h([f^{\theta_1}(x), \cdots, f^{\theta_{n-1}}(x)]), $$ (9)

where $\theta_i$ is the rotation radian uniformly sampled from $[0, 2\pi)$, $f^{\theta_i}$ refers to convolution with the kernel rotated by $\theta_i = \frac{2\pi}{n} i$, and $h$ represents the $1 \times 1$ convolution.

As shown in Fig. 4, the RKF involves no non-linear function, therefore multiple kernels within one convolution layer can be merged into one single kernel equally in the
inference stage, making its computational costs the same as the original base model.

IV. EXPERIMENTS

A. Datasets and experimental settings

For the training dataset, we use GL3D [36], which contains 181,280 high-resolution images in 1073 different scenes, along with the image matches, the intrinsics, and extrinsics, as well as depth data. As for evaluation datasets, we adopt HPatches [15] and YFCC100M [16]. The former contains 116 scenes, of which 57 scenes are under illumination changes and the left 59 scenes are under viewpoint changes. Each scene contains 6 sequences, where the first one is the reference image, and the rest are offered with the corresponding homography. For YFCC100M, we follow SuperGlue[37] which samples 4000 test pairs from the entire dataset, and each pair has ground-truth poses and sparse 3D models obtained from an off-the-shelf SfM tool.

To analyze the effectiveness of rotation augmentation in the training phase, we implement the training process under the following two settings:

- Train w/o rot: Train without rotation augmentations.
- Train w rot: Train with rotation augmentations of radians in $[0, 2\pi]$.

In the training phase, the base model is trained for 400,000 iterations, and the proposed MOFA is constructed with the base model trained with rotation augmentations. We set $n = 4$ in Eq. 2 and Eq. 3, therefore the rotation radians in MOFA and RKF are $0, \pi/2, \pi, 3\pi/2$. During the KD stage, the base model and RKF model are trained with 100,000 iterations respectively. For loss hyperparameters, we set $\lambda_1 = \lambda_2 = 1$ in Eq. 7 and Eq. 8.

Furthermore, to evaluate rotation invariance, apart from the original public dataset mentioned above, we also augment them with various rotations in the testing phase. Therefore, our testing settings are as follows:

- Test w/o rot: Test on original datasets.
- Test w rot: Test on rotation-augmented datasets under rotation of radians in $[0, 2\pi]$.

B. Comparison with SOTA approaches

In this part, we first compare our proposals with SIFT [3] and existing state-of-the-art learning-based approaches including SuperPoint (SP) [9], LF-Net (LF) [7], as well as ASLFeat (ASL) [12]. Here, we mainly focus on feature extraction stage, therefore we do not compare our approach to SOTA feature matching approaches, including SuperGlue [37], LofTR [38], CoTR [39], ClusterGNN [40], etc. Besides, for SuperPoint, data augmentation is essential in training for pseudo label generation, therefore we do not evaluate the model trained without rotation augmentations.

a) HPatches: The results on HPatches are shown in TABLE I, in terms of repeatability (Rep), mean matching accuracy (MMA), and homography estimation accuracy (HE). Based on the results, we observe that when trained without rotation augmentation, those learning-base methods perform poorly on rotation-augmented datasets, while SIFT show steady performance under rotation due to its elaborate design for orientation prediction. When trained with rotation augmentation, SuperPoint, LF-Net and ASLFeat obtain better generalization when a significant level of rotation is introduced to the test data, while our DRKF still surpasses
them by a large margin, for example, 10.86% better than SuperPoint with regards to homography estimation. It is notable that the proposed DRKF even outperforms SIFT considerably when testing on rotation-augmented datasets, while retaining the performance on original datasets. This demonstrates that DRKF effectively enables the network to extract more rotation-invariant features.

b) YFCC100M: For relative pose estimation on the YFCC100M dataset, we compare different methods in terms of AUC@5°, AUC@10°, and precision (PR). The results are shown in TABLE II, based on which it can be seen that DRKF also achieves the best performance, for instance, outperforms the augmented ASLFeat by 10.98% on precision. Besides, the performance of LF-Net trained with rotation augmentations even becomes worse when testing on the original dataset. This implies that excessive data augmentation sometimes can be harmful to the model. On the contrary, DRKF performs more steadily on the original and rotation-augmented datasets.

c) Qualitative comparison: Qualitative comparison can provide an intuitive observation of the matching results of different methods. We compare our DRKF with SIFT and other learning-based methods trained with rotation augmentations, including SuperPoint, LF-Net as well as ASLFeat in Fig. 5. And we can see that compared to other methods, DRKF obtains the largest number of correct correspondences under rotation of $\pi$, which shows that it can generalize effectively under large orientation changes, and even outperforms SIFT which explicitly assigns orientation to each keypoint.

C. Ablation Studies

To analyze the effectiveness of each component of our DRKF, we conduct experiments on four variants: Base (the original base model), MOFA, DBase (the base model distilled from MOFA), DRKF (the RKF model distilled from MOFA). The experiments are carried out on HPatches, and the results are shown in TABLE III, from which it can be concluded that:

- For the base model, when trained without rotation augmentation, it shows inferior performance on the rotation-augmented datasets. And it is clear that rotation augmentation on the training data can improve the rotation robustness in the testing phase.
- MOFA explicitly integrates multi-oriented features of the base model trained with rotation augmentation, therefore outperforms the base model on rotation-augmented datasets remarkably.
- With the supervision of MOFA, DBase overtakes the base model significantly, while still lags behind MOFA because of limited learning capability.
- Based on DBase, DRKF utilizes rotated kernels to enhance the representation ability for rotation invariance, therefore surpasses any other variants, including its teacher, MOFA.

To summarize, three components of our proposed method, MOFA, KD, and RKF, all contribute to rotation invariance.

D. Efficiency Analysis

For efficiency analysis, we compare the inference time of each method on Nvidia TX2 at the input size of 640×360 in TABLE IV. It can be seen that compared to previous methods, MOFA consumes the longest inference time due to muti-feedforward feature aggregation. DBase considerably shortens the computation time, and by virtue of the reparameterization technique, DRKF can be accelerated, and become computationally comparable to DBase.

V. CONCLUSION

In this paper, we propose Distilled Rotated Kernel Fusion (DRKF) under the supervision of Multi-Oriented Feature Aggregation (MOFA), and experiments reveal that it can boost the rotation invariance of extracted features remarkably. Thanks to the reparameterization approaches, the rotated kernels can be further merged into one so that the computational cost remains the same. In fact, despite the significance of rotation invariance for image matching, it is still an open question. Therefore, we hope our research can inspire the community to pay closer attention to this topic, we will also explore more effective methods in the future.
