2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

RIDE: Self-Supervised Learning of Rotation-Equivariant Keypoint Detection and Invariant Description for Endoscopy

Mert Asim Karaoglu[1,2], Viktoria Markova[2], Nassir Navab[1,3], Benjamin Busam[1], and Alexander Ladikos[2]

***Abstract*— Unlike in natural images, in endoscopy there is no clear notion of an up-right camera orientation. Endoscopic videos therefore often contain large rotational motions, which require keypoint detection and description algorithms to be robust to these conditions. While most classical methods achieve rotation-equivariant detection and invariant description by design, many learning-based approaches learn to be robust only up to a certain degree. At the same time learning-based methods under moderate rotations often outperform classical approaches. In order to address this shortcoming, in this paper we propose RIDE, a learning-based method for rotation-equivariant detection and invariant description. Following recent advancements in group-equivariant learning, RIDE models rotation-equivariance implicitly within its architecture. Trained in a self-supervised manner on a large curation of endoscopic images, RIDE requires no manual labeling of training data. We test RIDE in the context of surgical tissue tracking on the SuPeR dataset as well as in the context of relative pose estimation on a repurposed version of the SCARED dataset. In addition we perform explicit studies showing its robustness to large rotations. Our comparison against recent learning-based and classical approaches shows that RIDE sets a new state-of-the-art performance on matching and relative pose estimation tasks and scores competitively on surgical tissue tracking.**

I. INTRODUCTION

Minimally invasive endoscopic surgery has emerged as a modern alternative to traditional open surgery, reducing patient trauma and recovery times. During such surgeries, an endoscope is used to provide visual guidance and surveying to the operator. However, these devices usually have certain physical drawbacks affecting maneuverability, and their limited view can make navigation difficult. Modern computer vision techniques can be used to provide real-time solutions for simultaneous localization and mapping (SLAM) [1], [2], [3], [4], and tissue tracking [5], [6], [7], thereby assisting surgeons in performing surgeries with more precision [8]. Furthermore, 3D reconstruction applications [9], [10], [11] can be employed for diagnosis and longitudinal assessment procedures. Detecting and describing keypoints is a crucial step in such geometric computer vision tasks. However, due to illumination-inconsistencies and large rotational viewpoint changes, for example for matching between distant keyframes, this task is exceptionally difficult for endoscopic scenes. Classical methods like SIFT [13] compute rotation-invariant descriptors based on estimated orientations of the keypoints. Although they perform effectively on natural images, both their detection and matching performance substantially reduces when used for endoscopic images. This is mainly because of the strongly non-uniform illumination, non-Lambertian surface properties, limited visual textures and frequent occlusions that exist in the environment. More recently developed methods using convolutional neural networks (CNN) show significant improvements in detection and distinctiveness on visually demanding scenes, by learning to be robust against such inconveniences as well as viewpoint changes. However, as we show in our experiments, their success significantly decreases, and goes below that of classical methods when challenged by large rotational motions highlighting the limitations of learning-based robustness to rotations.

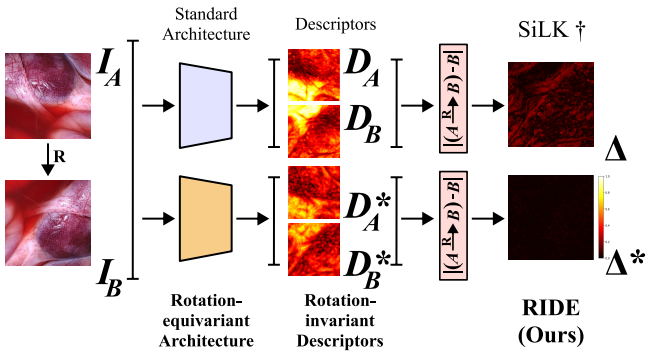

Fig. 1. In comparison to the state-of-the-art in keypoint detection and description, SiLK [12], († means that it is trained on endoscopic images); RIDE uses a rotation-equivariant architecture and creates rotation-invariant descriptors. Let I_B be the image I_A rotated by R and D_A, D_B the respective dense descriptor maps with their dimensionalities reduced by PCA and Δ is their difference after re-alignment. We use "*" to denote our results.

In this paper, we argue for the necessity of a rotation-equivariant design for reliable keypoint detection and description on endoscopic scenes, see Fig. 1. As a solution, we propose a novel method that we name RIDE. We design RIDE substantially influenced by some of the most recent state-of-the-art works [12], [14], [15] in the field, guided by the challenges of the targeted domain. RIDE employs a rotation-equivariant [16] steerable CNN to predict rotation-equivariant keypoints and invariant descriptors. Its lightweight architecture makes it a great option for possible real-time applications for navigation and 3D reconstruction. We use a simple yet effective self-supervised training scheme on homogpraphically augmented images from a large collec-

[1]Mert Asim Karaoglu, Nassir Navab, Benjamin Busam are with Technical University of Munich, Munich, Germany mert.karaoglu@tum.de
[2]Mert Asim Karaoglu, Viktoria Markova, Alexander Ladikos are with ImFusion GmbH, Munich, Germany
[3]Nassir Navab is with Johns Hopkins University, Baltimore, MD, USA

tion of endoscopic datasets requiring no manual labeling. We extensively evaluate our method for relative pose estimation and surgical tissue tracking. In addition we test its reliability under large rotation changes through a matching task.

In summary, we contribute:

- A novel, self-supervised rotation-equivariant keypoint detection and invariant description method with real-time capability, designed for endoscopic scenes.
- State-of-the-art results for endoscopic matching and relative pose estimation on the repurposed SCARED dataset [17].
- Competitive results for surgical tissue tracking on the SuPeR dataset [18].

II. RELATED WORK

Keypoint detection and description is one of the oldest tasks in computer vision. Traditional methods such as SIFT [13], ORB [19], and AKAZE [20], [21] often remain highly competitive due to their elegant architectures, which account for various symmetries including rotation-invariance by design. They employ various algorithms like histogram of gradient orientations [13] to compute the dominant orientations of the keypoints and use them to invariantize the descriptors. Even though these methods can achieve remarkable results in certain uses cases, in visually challenging endoscopic images they fall behind more recent learning based approaches.

Recently, learning-based methods [22], [23], [24], [25], [26], [12] have proven their advantages on demanding benchmarks on natural scenes. This success is also reflected in the surgical domain. Liu et al. [27] propose a dense descriptor method for sinus endoscopy. Similarly, ReTRo [28] introduces a dense descriptor model for endoscopic images and also learns to predict dense orientation maps. Unlike our approach, both of these methods rely on external keypoint extractors for sparse matching. Proposing an alternative to classical methods which typically detect considerably lower number of keypoints on retinal images, GLAMPoints [29] introduces a learning-based detector and uses classical descriptors for matching. Barbed et al. [30] study the performance of SuperPoint [30], which was designed for natural scenes, on colonoscopy images. Their study highlights the domain-gap that negatively effects a direct transfer of such models and proposes an adaption scheme.

Similar to equivariant designs of the classical methods, recent advancements in group-equivariant learning [31], [32], [16] enable constructing CNNs that are rotation-equivariant by design. REKD [14] utilizes a rotation-equivariant steerable CNN [16] to learn oriented keypoints using a histogram based orientation estimation inspired by SIFT [13]. RELF [15] employs a similar idea for rotation-invariant dense description. Both approaches show the advantages of rotation-equivariant learning for either detection or description under large rotational motion. However, as opposed to our method, they only focus on one part of the problem, keypoint detection or description, and do not propose a joint solution.

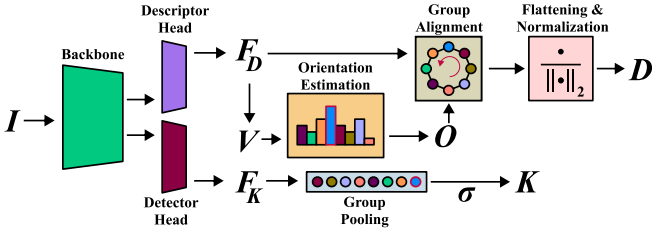

Fig. 2. **RIDE's architecture**. We utilize a rotation-equivariant architecture to extract dense maps F_D and F_K. F_D is used to estimate the orientation and generate invariant descriptors via group alignment. F_K is used to produce rotation-equivariant keypoint detections via group pooling.

III. METHOD

This work proposes a real-time capable keypoint detection and description learning pipeline that can handle strong illumination changes and abrupt camera motions in endoscopic scenes. For this, RIDE utilizes a rotation-equivariant steerable CNN and jointly learns to detect rotation-equivariant keypoints and invariant descriptors in a real-time capable architecture, see Fig. 2. Our model is trained in a self-supervised manner, which allows it to learn on various endoscopic datasets without any need for manual labeling. In the following sections, we'll explain in more detail the rotation-equivariant architecture, the keypoint detection, and the feature description components of RIDE.

A. RIDE's Architecture

Following SuperPoint [22] and SiLK [12] we use a single model that does detection and description in two separate heads with a shared backbone. Unlike them, our model is constructed with rotation-equivariant steerable convolutions [16], inside a VGG-style architecture, that act on cyclic group $C_{|G|}$, $|G|$ stands for the order of the discrete group. Although steerable convolutions come with a small overhead in the training time, at inference time, they act like standard convolutions not affecting the leanness of our approach. The detector and descriptor heads have compact architectures and output dense maps in the regular representation of the action group G; respectively, $F_K \in \mathbb{R}^{|G| \times H' \times W'}$ and $F_D \in \mathbb{R}^{C_D \times |G| \times H' \times W'}$, where H' and W' are the height and width of the maps and C_D is the size of the channel dimension.

B. Orientation Estimation and Rotation-invariant Description

RIDE's descriptor head serves for both dense orientation and description estimation. Feature map $F_D \in \mathbb{R}^{C_D \times |G| \times H' \times W'}$ contains C_D number of features for each element of the action group G.

We follow the prior works [33], [14], [15] and consider its first element along the channel dimension as the dense orientation histogram map $V \in \mathbb{R}^{|G| \times H' \times W'}$. The indices of the histogram directly relate to the elements of the cyclic group $C_{|G|}$ that samples the continuous rotational space $SO(2)$ with $\frac{360}{|G|}$ increments. During training we extract the relative in-plane rotation from the known homography that transforms an image, I_A, to its warped version I_B. Then,

we discretize them to their nearest neighbors on the cyclic group and assign them to be the ground-truth orientations $\tilde{O_B^A}$. We define the orientation loss term based on the histogram alignment loss introduced by Lee et al. [33]. In our implementation, we first apply a softmax on V along the group dimension. Then, we use a cyclic shift operator that works along the group dimension, defined as $\tau_G(X, O)$, to shift the elements of the histogram V_B and align to V_A using the ground-truth orientations $\tilde{O_B^A}$ as:

$$\begin{aligned} V'_A &= \text{softmax}(V_A)_G, \\ V'_B &= \tau_G(\text{softmax}(V_B)_G, -\tilde{O_B^A}). \end{aligned} \tag{1}$$

Our orientation loss is defined as follows:

$$\mathcal{L}_O = -\frac{1}{|\tilde{M}||G|} \sum_{(i_A,i_B)\in\tilde{M}} \sum_{k=1}^{|G|} V'_A(i_A)_k \log(V'_B(i_B)_k), \tag{2}$$

where $\tilde{M}$ is the list of ground-truth corresponding-pixel locations as i_A on I_A and i_B on I_B.

To generate rotation-invariant descriptors, we employ group aligment [15] and apply the cyclic shift operator $\tau_G(X, O)$ on the dense feature map F_D. Similar to histogram alignment, applied on F_D, τ_G shifts the feature elements along the group dimension by the index of their orientations O in the cyclic group $C_{|G|}$. Then, we concatenate them along the channel dimension to create our rotation-invariant descriptors $D \in \mathbb{R}^{(C_D|G|)\times H'\times W'}$. Finally, we normalize them to fit onto a unit-hypersphere.

For training, we use the dense descriptors, D_A and D_B of the image pairs and construct a dense score matrix S, where $S(i, j) = D_A(i) \cdot D_B(j)$. Following [34], [35], [15], we apply dual-softmax on S and apply a temperature value to acquire the soft mutual matching probabilities P. We apply negative log-likelihood on ground-truth matching pairs $\tilde{M}$ as our description loss:

$$\mathcal{L}_D = -\frac{1}{|\tilde{M}|} \sum_{(i_A,i_B)\in\tilde{M}} \log(P(i_A, i_B)). \tag{3}$$

Please refer to our implementation and training details section for more information how $\tilde{M}$ is generated.

During inference, we use the orientation estimated from the histogram while for training we use the ground-truth orientations. Furthmore, we leave the user the option to choose the matching algorithm, i.e. mutual nearest neighbor (MNN) or dual-softmax.

C. Rotation-equivariant Keypoints

In RIDE, we use a keypoint detection scheme constructed with simplicity in mind. We omit using a cell-based detection which imposes spatial constraints on keypoint locations. Instead, we follow [29], [12] and do a pixel-wise classification.

We generate the rotation-equivariant keypoint score map, $K \in \mathbb{R}^{H'\times W'}$, by collapsing the group dimension of the detector head output F_K using group pooling and applying a sigmoid on it.

Following [29], [36], [12], we take the cyclic matching success of the descriptors to generate the ground-truth labels. To generate ground-truth keypoint labels, we apply MNN between D_A and D_B. If ground-truth correspondences (i_A, i_B) are correctly matched, we label them as keypoints on the ground-truth keypoint label maps $\tilde{K}_A$, $\tilde{K}_B$. We use binary cross entropy (BCE) for the computing the keypoint loss as:

$$\begin{aligned} \mathcal{L}_K = -\sum_{i\in\{A,B\}} \frac{1}{|K_i|} \sum_{j=1}^{|K_i|} &(\tilde{K}_i(j)\log(K_i(j)) \\ &+ (1-\tilde{K}_i(j))\log(1-K_i(j))) \end{aligned} \tag{4}$$

In our experiments, we do not use non-maximum suppression (NMS), but we leave it to the user's choice.

D. Training Objective

Our training loss is the combination of the orientation, description and keypoint losses:

$$\mathcal{L} = \lambda_O\mathcal{L}_O + \mathcal{L}_D + \mathcal{L}_K, \tag{5}$$

where λ_O is defined as the weighting factor of the orientation loss.

IV. Evaluation

A. Implementation and Training Details

RIDE is implemented in PyTorch [37] using e2cnn [16] for group-equivariant operations.

For the experiments we train two different variations: RIDE and RIDE-L. Even though they both follow the structure of the VGGnp-4 backbone [12] and the corresponding detection and description heads, RIDE-L has twice as many parameters in the channel dimension and outputs 256 dimensional descriptors. In comparison RIDE's descriptors are 128 dimensional. We modify this architecture by removing the bias term in the convolutions, and swapping the positions of the BatchNorm and ReLU layers so that BatchNorm is applied first. Keeping the size of the channel dimensions the same while increasing the order of the cyclic group ($|G|$) results in high computational cost. Based on the findings of [38] we decide to set $|G|$ to 8 as an optimal point between runtime efficiency and performance. Because 3×3 kernels defined on C_8 are not well represented when steered by 45 degrees, we employ 5×5 kernels on all convolutions [16]. We change the size of the channel dimensions in RIDE so that at each layer the combined size of the group and channel dimensions is equal to its corresponding layer's channel size in SiLK. Like [22], [12] our model also operates on grayscale images. Since we don't use pooling functions and padding in convolutions, the size of the output maps are equal to the input size center-cropped during convolutions. Specifically, for an input image I of size $H \times W$, the output size $H' \times W'$ equal to $H - 36 \times W - 36$.

We train our models on a curation of various endopcopic datasets bundled and shared by Batic et al. [39] in addition to MITI [40]. More specifically, our training set includes 179,132 images of laparoscopic operations on various anatomies, from MITI [40], DSAD [41], ESAD [42], GLENDA [43], LapGyn4 [44], and PSI-AVA [45] datasets.

We train RIDE on image pairs generated by applying known homographies similar to [22], [12]. Taking SiLK [12] as the baseline, we apply the same data augmentation strategy. This means that the angle of in-plane rotations for homography generation is limited to the range $[-22.34, 22.34]$ in degrees. We rely on the image augmentations to gain robustness to illumination-inconsistencies. We compute the pixel correspondences, $\tilde{M}$, using the known homogprahic transformations. Similar to SiLK [12] we keep only the bijective correspondences and define their positions at the pixel centers.

We train RIDE for 100,000 iterations on the curated datasets. To balance the training, at every sample, we randomly pick one of the datasets and a frame from it. We crop out all GUI elements visible in the images and then resize them to 480 pixels on the longest dimension. To achieve the equal output size as SiLK, we train our models on cropped images of size 182×182. We use the ADAM optimizer [46] with learning rate 1e-4 and (0.9, 0.999) as betas. Empirically found, we set the weight for the orientation loss λ_O to 10. During training we apply a temperature of 20^{-1} on the score matrix S. We train with a batch size of 2 on a single Nvidia RTX 3090 GPU with mixed precision and use the block-size computation of the score matrix S [12] to decrease the vRAM cost. Our training of RIDE takes approximately 7 hours.

To present a fair baseline, we train a SiLK model with the VGGnp-4 backbone from scratch on the same dataset of RIDE and denote it as SiLK†. For all the trained models (including SiLK†), we use the weights of the checkpoint that performed the best on the validation set. In all the experiments we use the exact same models and extract the top 10,000 detections as keypoints.

In the experiments we use both dual-softmax and mutual nearest neighbor (MNN) matching. For dual-softmax we always use a temperature value of 0.1 for the score matrix and matching threshold of 0.9. These parameters are empirically chosen without detailed testing, therefore we believe that there is room for improvement for task-specific-tuning.

B. Relative Pose Estimation

1) Dataset: In accordance with the assessment methods employed in previous research [47], [35] on natural scenes [48], we leverage the components from a porcine endoscopic stereo depth estimation dataset, SCARED [17], to create an evaluation protocol specifically for this purpose. Pair sampling is done within all sequences from frames that have pose labels and at least 50% point cloud overlap with the initial frame of their sequence. First, with a stride of 40, we consider each frame as the source and its temporally consecutive ones as the targets in pairs. Then, to keep only the feasible and interesting pairs, we filter out the ones that have a point cloud overlap outside of the range between $40\% - 95\%$. This way, we generate in total 1,223 image pairs with challenging illumination and pose variations, see Fig. 3. We believe our sampling heuristics are more suitable for the evaluation of the task at hand in comparison to sampling based on temporal distance [30] or among left-right stereo pairs [28]. In this experiment, we resize the images to 640×512.

2) Metrics: For evaluation we compute the AUC of the pose error under thresholds at $(5^\circ, 10^\circ, 20^\circ)$ as introduced by Yi et al. [49]. The pose error is defined as the maximum of the angular error computed from the estimated rotation matrix and the unit-scale translation vector. The relative poses are extracted from the essential matrix computed using the OpenCV [50] implementation with MAGSAC++ [51].

3) Baselines: In addition to SiLK† we also compare our model against the publicly shared weights of SiLK [12] with the same architecture trained by the authors on the COCO dataset [52]; this model is simply denoted as SiLK [12]. Also we benchmark against the OpenCV [50] implementations of some of the very successful traditional methods: ORB [19], AKAZE [20], [21], and SIFT [13] with and without the ratio-test.

4) Results: As shown in Table I, both RIDE-L and RIDE achieve first and second best performances on the thresholds when coupled with the dual-softmax matcher. We believe that the improvement shown by RIDE-L is connected to the more capacity it has due to its larger channel size. Trailing RIDE on most metrics, SiLK†, trained on endoscopic images, significantly improves over SiLK [12] trained on COCO [52]. This suggests that the domain-gap between the real-world and the endoscopic scenes is too large for learning-based methods to generalize to and therefore training, or fine-tuning on the endoscopic scenes is necessary. The results also highlight that, unlike the other classical descriptors, SIFT can still perform respectively well under challenging conditions surpassing the SoTA learning-based algorithm, SiLK [12], trained on out-of-domain data. However, the significantly higher number of keypoint detections, and matches (see Fig. 3), make learning-based approaches great options for SLAM and 3D reconstruction pipelines.

TABLE I

RELATIVE POSE ESTIMATION EVALUATION ON THE REPURPOSED SCARED DATASET[17]. MNN STANDS FOR MUTUAL NEAREST NEIGHBOR. THE MOST RIGHT TWO COLUMNS SHOW THE NUMBER OF DETECTED AND MATCHED KEYPOINTS. **BOLD** VALUES ARE THE HIGHEST IN THEIR CATEGORY AND UNDERLINED ONES ARE THE RUNNER-UPS.

Method	Matching	Pose estimation AUC at 5°	10°	20°	# Detected	# Matched
ORB [19]	MNN	0.70	2.46	6.92	500	157
AKAZE [20], [21]	MNN	3.22	11.23	24.48	443	172
SIFT [13]	MNN	6.14	20.12	39.13	1,312	506
	Ratio test	9.79	27.51	48.94	1,312	202
SiLK [12]	MNN	5.17	16.05	31.21	10,000	3,183
	Dual-softmax	8.28	22.47	40.48	10,000	1,395
SiLK† [12]	MNN	8.14	23.16	41.29	10,000	3,274
	Dual-softmax	10.56	28.68	47.86	10,000	1,238
RIDE	MNN	7.15	21.71	41.58	10,000	3,135
	Dual-softmax	<u>11.03</u>	<u>31.19</u>	<u>54.43</u>	10,000	1,136
RIDE-L	MNN	8.87	26.26	48.00	10,000	3,046
	Dual-softmax	**12.05**	**33.98**	**57.84**	10,000	1,273

C. Matching Under Large In-plane Rotations

1) Dataset: Unlike most real-life applications like autonomous driving or indoor navigation, in endoscopy it is

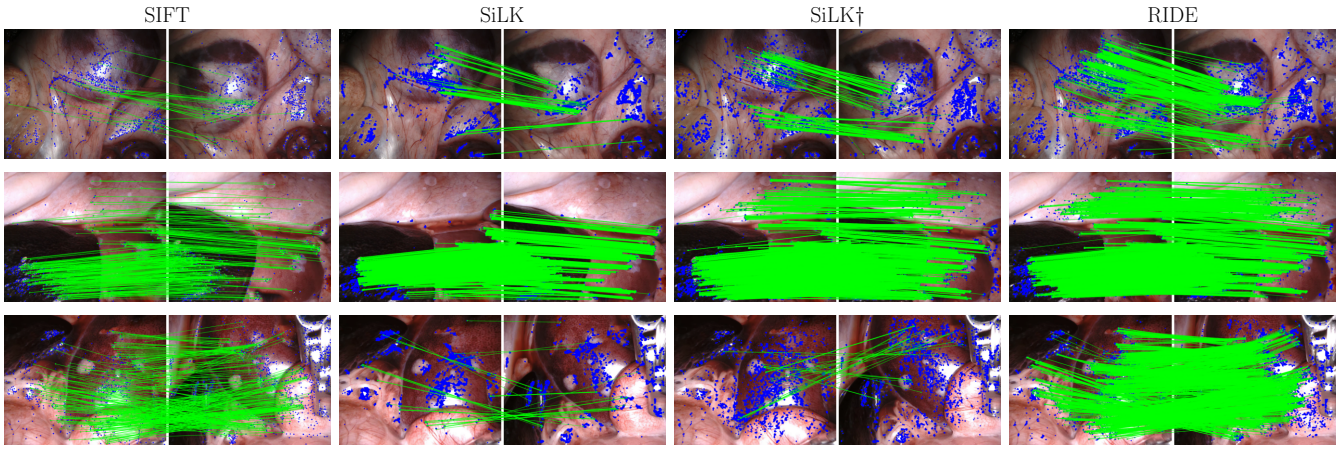

Fig. 3. Qualitative matching results on pairs from SCARED dataset [17]. Blue points show all the detected keypoints and green lines refer to inlier matches extracting using MNN.

common to have large-in plane rotations as part of abrupt viewpoint changes, for example in keyframe-based SLAM systems. Since it is very challenging to get pixel-wise correspondences on endoscopic images, we create a test setup by applying known rotations.

We extract 10 temporally equally distanced images from all images in the SCARED dataset [17] to increase the visual variation. Similar to the rotation invariance experiment conducted in RELF [15], pairs of images are generated by taking the original images as the source images and their in-plane rotated version as the targets. For each source image, the target images are its rotated version with 10 degrees increments from 0 to 350 degrees. In total, we end-up with 360 image pairs with known pixel-wise correspondences. We ensure that the image transformation preserves the image content scale, which generates empty parts in the resulting image. These are replaced with a smooth grayscale background. Accordingly, while the source images are kept at 640×512, the target image sizes depend on the rotation angle.

2) Metrics: For evaluation we compute the mean matching accuracy of the matching error under thresholds at $(3, 5, 10)$px following [15]. For a fair comparison, all the methods employ MNN matching.

3) Baselines: We use the same baselines as the previous experiment.

4) Results: By combining strong distinctiveness and description capabilities of the learning-based methods with a rotation-invariant design that is often found in classical approaches, RIDE outperforms baselines of both classes. As shown in Table II and Fig. 5, both RIDE and RIDE-L achieve the top performance on all thresholds. Employing a standard CNN architecture, that is only translation-equvariant, SiLK and SiLK† perform better than the classical methods only within a limited range of rotations. Outside of this range, they both fail and drastically fall behind the classical methods that are engineered to be rotation invariant. This finding strongly supports the design choices of our approach that combines the strengths of both sides.

TABLE II

ROTATION ROBUSTNESS EVALUATION ON IMAGES FROM SCARED DATASET [17]. ϵ REPRESENTS THE THRESHOLD VALUES IN PIXELS USED FOR MMA. **BOLD** VALUES ARE THE HIGHEST IN THEIR CATEGORY AND UNDERLINED ONES ARE THE RUNNER-UPS.

	Mean Matching Accuracy		
Method	$\epsilon = 3$px	$\epsilon = 5$px	$\epsilon = 10$px
ORB [19]	0.62	0.66	0.68
AKAZE [20], [21]	0.84	<u>0.85</u>	<u>0.86</u>
SIFT [13]	0.71	0.71	0.72
SiLK [12]	0.26	0.27	0.28
SiLK† [12]	0.20	0.21	0.22
RIDE	**0.87**	**0.88**	**0.89**
RIDE-L	<u>0.85</u>	**0.88**	**0.89**

D. Surgical Tissue Tracking

1) Dataset: In this experiment we follow ReTRo's [28] assessment for tracking deforming points on the SuPeR dataset [18]. SuPeR [18] is a perception dataset recorded on a Da Vinci Surgical System using tissue-like material to imitate a robotic surgical scene for various tasks. Its tissue tracking subset is a sequence of 522 frames with every 10^{th} frame manually labeled with the tracking information of 20 pre-defined points. Each frame is of size 640×480.

2) Metrics: We report the average distance error proportion to image height for each tracking point computed as follows. For each labeled frame, keypoints are extracted and matched against the initial labeled frame. From matching keypoints in the initial frame, the closest 4 keypoints are assigned to a tracking point. Their average motion is computed and recorded as the motion of the tracking point. And finally, for each tracking point, the average distance error in pixels divided by the image height is reported.

3) Baselines: In this experiment we compare against ReTRO [28], CAPS [53], ORB [19] and SIFT [13]. Their results are taken from the report of Schmidt et al. [28]. For RIDE and RIDE-L we compute the matches with dual-softmax and following [28], we employ the OpenCV [50] implementation of the GMS [54].

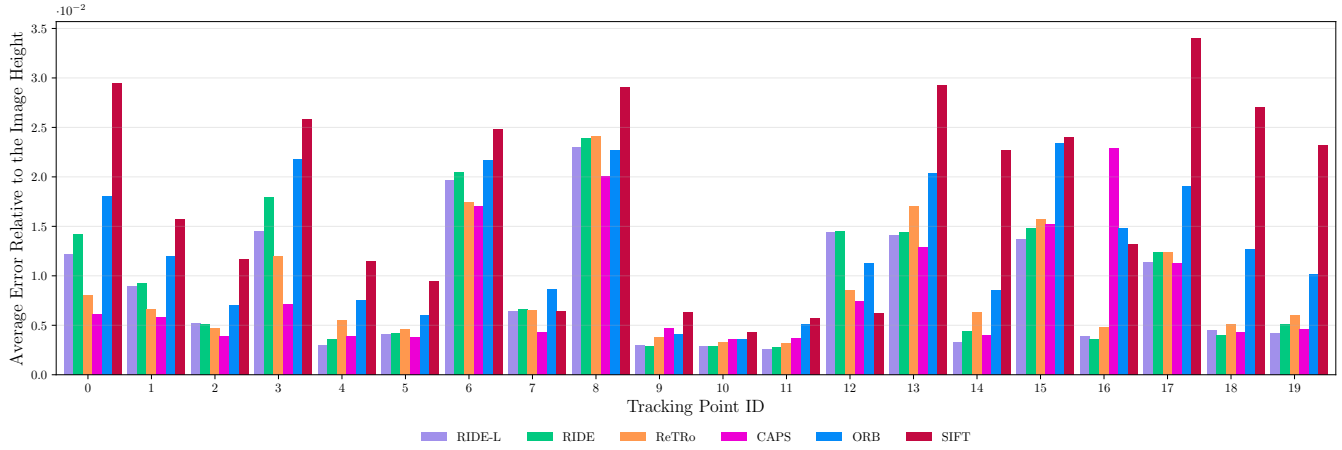

Fig. 4. Average point tracking errors relative to the image height on the SuPeR dataset [18]. The horizontal axis represents the id of the each tracking point and clustered on them are each method's errors represented with bars. Results of the other methods are taken from the report of Schmidt et al. [28].

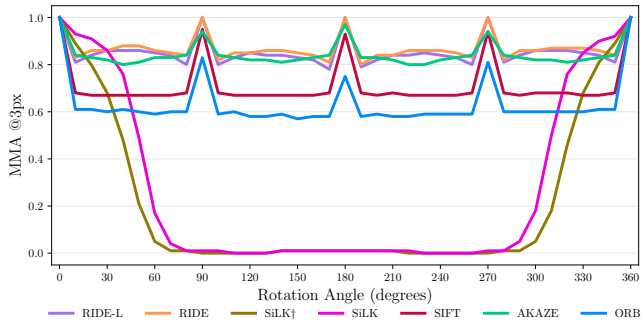

Fig. 5. Mean matching accuracy (MMA) computed at the threshold of 3px. The pairs are generated using images from the SCARED dataset [17] and applying known rotations with 10 degree increments.

4) Results: Based on the results depicted in the Fig. 4, our model demonstrates competitive performance compared to the learning-based methods [53], [28] exceeding well above the classical approaches [2], [13]. This shows the robustness of our method even on deforming tissues, which is a transformation it does not encounter during training.

E. Runtime Performance

Compiled with TensorRT on half precision, for an input of size 640×512 RIDE runs at 65.41 FPS while the larger model RIDE-L achieves 17.91 FPS on average on an NVIDIA GeForce RTX 3090.

V. DISCUSSION

In this work, we mainly focus on solving the task of reliable keypoint detection and description for endoscopic images that contain large rotational viewpoint changes. On the task of relative pose estimation, see Table I and Fig. 3, our method excels over the classical approaches and the state-of-the-art CNN-based method [12] proving the effectiveness of its rotation-equivariant architecture on providing reliable predictions regardless of the large viewpoint changes. This experiment also highlights that the domain-gap between natural and surgical images is too large for learning-based methods trained on the prior to generalize successfully. Moreover, our method's robustness is further challenged in the matching task showing consistent success across the whole spectrum of rotation angles making it robust to rotations beyond those seen during training. Finally, we test our method for deforming tissue tracking and perform competitively to the state-of-the-art learning-based method for endoscopy [28].

However, endoscopic videos consist of many more challenges other than large rotational viewpoint changes. Our method relies on image augmentations for learning to be robust against illumination-inconsistencies. Yet, as it is displayed in Fig. 3, RIDE also detects keypoints on the edges of reflections which can result in adversities. In future work, regularizing terms [30] can be studied and imposed on the detection objective. Furthermore, tissue deformation is a prominent issue in surgeries. A more sophisticated architecture directly targeting deformation-awareness [55] can be further explored in the context of endoscopy.

VI. CONCLUSION

We present RIDE, a self-supervised rotation-equivariant detection and invariant description method for endoscopic scenes. As a learning-based method, RIDE can predict distinctive descriptors and high numbers of salient keypoints. Furthermore, its rotation-equivariant design enables it to perform reliably under large rotational motion. RIDE achieves state-of-the-art performance on matching and relative pose estimation tasks and scores competitively on surgical tissue tracking, outperforming recent learning-based and classical approaches.

ACKNOWLEDGMENT

We thank Adam Schmidt for providing us with great support in reproducing the surgical tissue tracking experiment.

REFERENCES

[1] X. Liu, Z. Li, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, "Sage: slam with appearance and geometry prior for endoscopy," in *2022 International conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 5587–5593.

[2] L. Oliva Maza, F. Steidle, J. Klodmann, K. Strobl, and R. Triebel, "An orb-slam3-based approach for surgical navigation in ureteroscopy," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 11, no. 4, pp. 1005–1011, 2023.

[3] J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Tracking monocular camera pose and deformation for slam inside the human body," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 5278–5285.

[4] J. J. G. Rodriguez, J. Montiel, and J. D. Tardos, "Nr-slam: Non-rigid monocular slam," *arXiv preprint arXiv:2308.04036*, 2023.

[5] A. Schmidt, O. Mohareri, S. DiMaio, and S. E. Salcudean, "Fast graph refinement and implicit neural representation for tissue tracking," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1281–1288.

[6] A. Schmidt, O. Mohareri, S. P. DiMaio, and S. E. Salcudean, "Recurrent implicit neural graph for deformable tracking in endoscopic videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252369261

[7] A. Schmidt, O. Mohareri, S. DiMaio, and S. E. Salcudean, "Sendd: Sparse efficient neural depth and deformation for tissue tracking," *arXiv preprint arXiv:2305.06477*, 2023.

[8] Z. Fu, Z. Jin, C. Zhang, Z. He, Z. Zha, C. Hu, T. Gan, Q. Yan, P. Wang, and X. Ye, "The future of endoscopic navigation: A review of advanced endoscopic vision technology," *IEEE Access*, vol. 9, pp. 41 144–41 167, 2021.

[9] B. Busam, P. Ruhkamp, S. Virga, B. Lentes, J. Rackerseder, N. Navab, and C. Hennersperger, "Markerless inside-out tracking for 3d ultrasound compounding," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation. POCUS. MICCAI*. Springer, 2018, pp. 56–64.

[10] X. Liu, M. Stiber, J. Huang, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, "Reconstructing sinus anatomy from endoscopic video–towards a radiation-free approach for quantitative longitudinal assessment," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 3–13.

[11] M. A. Karaoglu, N. Brasch, M. Stollenga, W. Wein, N. Navab, F. Tombari, and A. Ladikos, "Adversarial domain feature adaptation for bronchoscopic depth estimation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. Springer, 2021, pp. 300–310.

[12] P. Gleize, W. Wang, and M. Feiszli, "Silk–simple learned keypoints," *arXiv preprint arXiv:2304.06194*, 2023.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[14] J. Lee, B. Kim, and M. Cho, "Self-supervised equivariant learning for oriented keypoint detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4847–4857.

[15] J. Lee, B. Kim, S. Kim, and M. Cho, "Learning rotation-equivariant features for visual correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 887–21 897.

[16] M. Weiler and G. Cesa, "General e (2)-equivariant steerable cnns," *Advances in neural information processing systems*, vol. 32, 2019.

[17] M. Allan, J. Mcleod, C. Wang, J. C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K. X. Fu, T. Zeffiro, W. Xia *et al.*, "Stereo correspondence and reconstruction of endoscopic data challenge," *arXiv preprint arXiv:2101.01133*, 2021.

[18] Y. Li, F. Richter, J. Lu, E. K. Funk, R. K. Orosco, J. Zhu, and M. C. Yip, "Super: A surgical perception framework for endoscopic tissue manipulation with surgical robotics," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2294–2301, 2020.

[19] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[20] P. F. Alcantarilla and T. Solutions, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 34, no. 7, pp. 1281–1298, 2011.

[21] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "Kaze features," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*. Springer, 2012, pp. 214–227.

[22] D. DeTone, T. Malisiewicz, and A. Rabinovich, "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.

[23] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. 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*, 2019, pp. 8092–8101.

[24] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger, "R2d2: repeatable and reliable detector and descriptor," *arXiv preprint arXiv:1906.06195*, 2019.

[25] A. Barroso-Laguna, Y. Verdie, B. Busam, and K. Mikolajczyk, "Hddnet: Hybrid detector descriptor with mutual interactive learning," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[26] X. Zhao, X. Wu, W. Chen, P. C. Chen, Q. Xu, and Z. Li, "Aliked: A lighter keypoint and descriptor extraction network via deformable transformation," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[27] X. Liu, Y. Zheng, B. Killeen, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, "Extremely dense point correspondences using a learned feature descriptor," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4847–4856.

[28] A. Schmidt and S. E. Salcudean, "Real-time rotated convolutional descriptor for surgical environments," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. Springer, 2021, pp. 279–289.

[29] P. Truong, S. Apostolopoulos, A. Mosinska, S. Stucky, C. Ciller, and S. D. Zanet, "Glampoints: Greedily learned accurate match points," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 732–10 741.

[30] O. L. Barbed, F. Chadebecq, J. Morlana, J. M. Montiel, and A. C. Murillo, "Superpoint features in endoscopy," in *MICCAI Workshop on Imaging Systems for GI Endoscopy*. Springer, 2022, pp. 45–55.

[31] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International conference on machine learning*. PMLR, 2016, pp. 2990–2999.

[32] T. S. Cohen and M. Welling, "Steerable cnns," *arXiv preprint arXiv:1612.08498*, 2016.

[33] J. Lee, Y. Jeong, and M. Cho, "Self-supervised learning of image scale and orientation," in *31st British Machine Vision Conference 2021, BMVC 2021, Virtual Event, UK*. BMVA Press, 2021. [Online]. Available: https://www.bmvc2021-virtualconference.com/programme/accepted-papers/

[34] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," *Advances in neural information processing systems*, vol. 31, 2018.

[35] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[36] M. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 254–14 265, 2020.

[37] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[38] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2786–2795.

[39] D. Batić, F. Holm, E. Özsoy, T. Czempiel, and N. Navab, "Whether and when does endoscopy domain pretraining make sense?" *arXiv preprint arXiv:2303.17636*, 2023.

[40] R. Hartwig, D. Ostler, J.-C. Rosenthal, H. Feußner, D. Wilhelm, and D. Wollherr, "Miti: Slam benchmark for laparoscopic surgery," *arXiv preprint arXiv:2202.11496*, 2022.

[41] M. Carstens, F. M. Rinner, S. Bodenstedt, A. C. Jenke, J. Weitz, M. Distler, S. Speidel, and F. R. Kolbinger, "The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science," *Scientific Data*, vol. 10, no. 1, p. 3, 2023.

[42] V. S. Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, A. Leporini, C. Landolfo, A. Stabile, F. Setti, R. Muradore, E. Oleari *et al.*, "Esad: Endoscopic surgeon action detection dataset," *arXiv preprint arXiv:2006.07164*, 2020.

[43] A. Leibetseder, S. Kletz, K. Schoeffmann, S. Keckstein, and J. Keckstein, "Glenda: gynecologic laparoscopy endometriosis dataset," in *International Conference on Multimedia Modeling*. Springer, 2019, pp. 439–450.

[44] A. Leibetseder, S. Petscharnig, M. J. Primus, S. Kletz, B. Münzer, K. Schoeffmann, and J. Keckstein, "Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology," in *Proceedings of the 9th ACM multimedia systems conference*, 2018, pp. 357–362.

[45] N. Valderrama, P. Ruiz Puentes, I. Hernández, N. Ayobi, M. Verlyck, J. Santander, J. Caicedo, N. Fernández, and P. Arbeláez, "Towards holistic surgical scene understanding," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 442–452.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[47] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "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.

[48] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[49] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 2016, pp. 467–483.

[50] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[51] D. Barath, J. Noskova, M. Ivashechkin, and J. Matas, "Magsac++, a fast, reliable and accurate robust estimator," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1304–1312.

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[53] Q. Wang, X. Zhou, B. Hariharan, and N. Snavely, "Learning feature descriptors using camera pose supervision," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 757–774.

[54] J. Bian, W.-Y. Lin, Y. Liu, L. Zhang, S.-K. Yeung, M.-M. Cheng, and I. Reid, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," *International Journal of Computer Vision (IJCV)*, 2020.

[55] G. Potje, , F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento, "Enhancing deformable local features by jointly learning to detect and describe keypoints," in *2023 IEEE / CVF Computer Vision and Pattern Recognition (CVPR)*, 2023.