Self-Supervised Endoscopic Image Key-Points Matching

Manel Farhat[a], Houda Chaabouni-Chouayakh[a,b], Achraf Ben-Hamadou[a,b,*]

[a]*Centre de Recherche en Numérique de Sfax, Technopôle de Sfax, 3021 Sfax, Tunisia*
[b]*Laboratory of Signals, systems, artificial Intelligence and networks, Technopôle de Sfax, 3021 Sfax, Tunisia*

Abstract

Feature matching and finding correspondences between endoscopic images is a key step in many clinical applications such as patient follow-up and generation of panoramic image from clinical sequences for fast anomalies localization. Nonetheless, due to the high texture variability present in endoscopic images, the development of robust and accurate feature matching becomes a challenging task. Recently, deep learning techniques which deliver learned features extracted via convolutional neural networks (CNNs) have gained traction in a wide range of computer vision tasks. However, they all follow a supervised learning scheme where a large amount of annotated data is required to reach good performances, which is generally not always available for medical data databases. To overcome this limitation related to labeled data scarcity, the self-supervised learning paradigm has recently shown great success in a number of applications. This paper proposes a novel self-supervised approach for endoscopic image matching based on deep learning techniques. When compared to standard hand-crafted local feature descriptors, our method outperformed them in terms of precision and recall. Furthermore, our self-supervised descriptor provides a competitive performance in comparison to a selection of state-of-the-art deep learning based supervised methods in terms of precision and matching score.

Keywords: self-supervised learning, feature matching, endoscopic images,

*Corresponding author
Email addresses: farhat.manel@hotmail.fr (Manel Farhat), houdachaabouni@gmail.com (Houda Chaabouni-Chouayakh), achraf.benhamadou@crns.rnrt.tn (Achraf Ben-Hamadou)

deep learning, image key-points matching.

1. Introduction

Bladder cancer is the sixth most common cancer and the ninth cancer causing mortality for men, and it is the fifteenth cancer causing mortality for both sexes worldwide (Sung et al., 2020). Endoscopy remains the gold standard visual examination for detecting bladder cancer in its early stages. It consists in inserting an endoscope (A lighted, tubular instrument) through the urethra into the bladder. The doctor explores the bladder's inner walls by navigating the endoscope to scan the organ for lesions. Despite this being easy to perform, the endoscopic examination has several limitations. The main one being the very limited field of view, in addition to the reduced maneuverability inside the bladder. Because bladder malignancy (*e.g.,* lesions) is multi-focal and typically spreads over larger areas than the endoscope field of view, lesion partitions are divided across multiple video frames. As a result, this makes it time consuming and difficult for the doctor to locate and determine the spatial distribution of the lesions. Furthermore, finding an image of interest for diagnosis within a few minutes of the sequence is also a tough task. An intuitive solution to these limitations is to provide the doctor with a panoramic image of the clinical sequence instead of the image sequence. In this case, the extraction of discriminative local features for accurate image matching is a fundamental step in the construction of a panoramic image.

Despite extensive research into panoramic image generation for endoscopy, the development of a robust and accurate feature matching between endoscopic images is particularly difficult due to the specificity of such images. In fact, the texture of the inner surfaces of the bladder varies greatly between patients. Endoscopic images are also typically reddish with a weak texture. This is in

addition to the scarcity of data due to ethical concerns and the personal data protection regulations.

There are three basic approaches to endoscopic image matching that are currently being used in the literature: motion-based methods, global image matching methods, and local image matching methods. For motion based methods, we can cite (Hernandez-Mier et al., 2010; Sharib et al., 2013; Sharib et al., 2016; Chu et al., 2020; Zenteno et al., 2022) who used optical flow based approaches to find key-point correspondences between bladder images. Despite their efficiency in registering consecutive image frames, motion-based matching methods are still sensitive to challenging situations such as lighting changes, weak textures, and large view-point changes. As a result of this, these methods are limited to matching consecutive image frames and cannot be used for images separated in time, such in case of patient follow-up or the closing loop in a same endoscopic video.

Global image matching strategies ensure the global coherence of the obtained panoramic images and penalize discontinuities by using the entire image content (*e.g.,* contours, color, graph, *etc.*) and specific smoothness constraints. For example, Miranda-Luna *et al.* (Miranda-Luna et al., 2008) proposed a mosaicing algorithm for endoscopic images based on the maximization of the mutual information using the entire image frames. Weibel *et al.* (Weibel et al., 2012) used instead graph-cuts to minimize a global energy function computed on the entire image pixels. Global matching methods are memory-intensive and time-consuming since the energy function is usually computed over the entire image pixels. Furthermore, they are better suited to consecutive image matching with minor changes as they are sensitive to the initial geometric transformation between the input images. It is also worth noting that the strong assumption on the planarity of organ surfaces is not usually verified.

Local image matching methods, on the other hand, find the correspondence between image key-points by extracting discriminative features from local image data. These methods are computationally efficient, making them suitable for real-time applications. They usually begin by detecting image key-points and then compute a feature vector for each detected key-point using descriptor techniques such as SIFT (Lowe, 2004) and SURF(Bay et al., 2008).

Deep learning techniques have gained traction in the field of local feature description and matching in the recent years. In the domain of endoscopy, to the best of our knowledge, no previous work has been proposed to build discriminative local feature descriptors for endoscopic images using deep learning techniques. Nevertheless, we can find deep learning studies focusing on other purposes like endoscopic images denoising (Zou et al., 2019), polyp classification (Kim et al., 2021), and bleeding zone semantic segmentation (Ghosh et al., 2018). They are all based on supervised learning schemes that rely heavily on the availability of annotated data.

Unsupervised learning, in contrast to supervised learning, does not require a labeled dataset, which increases its popularity, particularly in the medical imaging field (Li et al., 2020; Chen & Frey, 2020). Recently, the research community has become increasingly interested in self-supervised learning, a new subset of unsupervised methods. Basically, it consists in training neural networks with automatically generated labels known as pseudo labels (Jing & Tian, 2019), and without any manual annotation. Self-supervised learning paradigm contributes to overcome the barrier of labeled data scarcity by leveraging the availability of large amounts of unlabeled data.

In this article, we deal with local image matching methods to find correspondences between image key-points for the purpose of generating panoramic images in endoscopy. In particular, we propose the first self-supervised approach

for endoscopic image key-points matching based on deep learning techniques. We designed a convolutional neural network (CNN) as a local feature descriptor to transform patches extracted around key-points into a discriminative embedding space for an effective key-points matching. Our main contribution is the design of a training procedure for the proposed CNN model without need of any labeled data. Indeed the training requires only raw video frames. The source code related to this study is released on https://github.com/abenhamadou/Self-Supervised-Endoscopic-Image-Key-Points-Matching.git

The remainder of this paper is organized as follows: First, an overview on existing local handcrafted and learning-based feature descriptors as well as their applications in the medical field is provided in section 2. Then, section 3 presents the proposed self-supervised endoscopic image key-points matching approach. After that, experiments and results are presented and discussed in section 4. Finally, section 5 gives some conclusions and perspectives.

2. Related works

In this section, we provide an overview on state-of-the-art local feature descriptors, focusing on the trendy shift from handcrafted to deep learning based feature descriptors. Furthermore, we discuss the use of local feature descriptors in the medical field, emphasizing the labeled data scarcity issue.

2.1. Handcrafted Local Feature Descriptors

According to (Ma et al., 2021), handcrafted local feature descriptors can be classified into floating and binary descriptors based on the discriminative vector type. One of the well-known floating descriptors is SIFT (Lowe, 2004), originally calculated based on image gradients. Inspired by SIFT, (Bay et al., 2008) proposed later the SURF descriptor which is much faster than SIFT. SURF used integral images and Haar wavelet responses in a circular neighborhood around

the SURF key-points, which allows to reduce computational costs and speed up the descriptor extraction. (Alcantarilla et al., 2012) proposed KAZE algorithm whose detection pipeline is similar to SURF. However, it used the MU–SURF method (Agrawal et al., 2008) to create a non-linear scale space after applying a non-linear diffusion filter. Despite the efficiency of floating descriptors, they are still not suitable for real time applications because of their important computational time. Hence, the rise of binary descriptors. The main benefits of these descriptors are their ease of implementation, simplicity and efficiency. The key idea behind is to generate a binary feature vector by comparing and encoding the intensity of each pixel relatively to its neighbors (Pietikäinen et al., 2011). Among these methods, we cite BRIEF (Calonder et al., 2010), ORB (Rublee et al., 2011), an extended version of BRIEF with rotation invariance and AKAZE (Alcantarilla et al., 2013), an accelerated method of KAZE algorithm. Another example is BRISK (Leutenegger et al., 2011), which exploits a concentric circles pattern and presents an optimization of BRIEF and ORB.

When it comes to the medical context, handcrafted local feature descriptors are widely used for image matching (Saha et al., 2016; Du et al., 2011). In (Hernandez-Matas et al., 2017), the authors proposed a retinal image registration method which combined local feature descriptors with vascular bifurcation. They tested SIFT, SURF and Harris-PIIFD (Chen et al., 2010) as feature descriptors and showed that the combination of SIFT with vascular bifurcations outperforms other combinations. For retinal image mosaicing, Jalili *et al.* (Jalili et al., 2020) used SIFT descriptor to extract local features and then selected optimal features based on a Voronoi diagram. In fluorescence endoscopy, Behrens *et al.* (Behrens et al., 2011, 2009) proposed a real-time bladder mosaicing method based on SURF feature descriptor. Du *et al.* (Du et al., 2011) designed a SIFT-based zone matching approach for endoscopic images. This approach im-

proves matching results especially with regards to computing time. Also, the work of (Liu et al., 2022) proposed an improved feature point pair purification algorithm for endoscopic image matching based on the SIFT descriptor. In the same vein, Zhang *et al.* (Zhang et al., 2022) proposed to improve the standard ORB-oriented algorithm using the Gaussian Pyramid method for endoscopic image mosaicing purposes. It should be noted that the design of a suitable local feature descriptor plays a critical role in this family of methods.

2.2. Deep Learning based Local Feature Descriptors

Among the first successful learning-based image matching approaches, we can cite MatchNet (Han et al., 2015). It is presented as the combination of two networks: the first one is the feature extracting network, inspired by Siamese network, and the second one is the learned metric network composed of 3 fully connected layers. DeepDesc (Simo-Serra et al., 2015) proposed a Siamese network with L2 distance trained by selecting only hardest pairs samples to match in order to increase the descriptor performance. Similarly, in (Tian et al., 2017), the authors proposed a network named L2-Net with seven convolutional layers that outperformed traditional descriptors, including SIFT. GeoDesc (Luo et al., 2018b) proposed to learn local descriptors by including geometry constraints from multi-view reconstructions, achieving thus significant improvements in terms of loss computation, data sampling and data generation during the learning process. In designing SOSNet, Tian *et al.* (Tian et al., 2019) used second-order similarity into the learning of local descriptors, achieving state-of-the-art performance on several standard benchmarks for different tasks. Later, the same authors (Tian et al., 2020) proposed HyNet, a modified version of L2-Net where all batch normalization layers are replaced by the off-the-shelf Filter Response Normalization (FRN) layers, which outperforms previous architectures on standard benchmarks.

Recently, a dense feature descriptor namely DGD-net is designed in (Liu et al., 2021) where the training is guided by the reliability of the descriptor in matching. DGD-net proposes a backtracking method to enhance the localization accuracy. The authors of (Zhou et al., 2021) designed Patch2Pix in a detect-to-refine manner, which begins with establishing correspondences between patches and then regresses pixel correspondences according to matched patches using a local search. A novel local descriptors extraction method named CNDesc is introduced in (Wang et al., 2022) where cross normalization technology is used as an alternative to the common L2 normalization. An efficient feature reuse backbone is designed providing the network with a strong descriptive ability and an image-level distribution consistent loss is used for regularization to enhance the robustness and stability of local descriptors.

Recently, in (Wiles et al., 2021), authors proposed a new image matching approach using a co-attention module to condition learned descriptors on both images and a distinctiveness score computed to select the best matches at test time, leading thus to an improved correspondence between image pairs under challenging conditions. Sarlin *et al.* (Sarlin et al., 2020) designed SuperGlue, an attention-based graph neural networks for local feature matching based on transformer (Vaswani et al., 2017). Inspired by SuperGlue, LoFTR (Sun et al., 2021) used self and cross attention layers to extract local feature descriptors and match images. It used a linear transformer to reduce the computational complexity.

Motivated by this success of deep learning techniques in various computer vision tasks, the medical imaging community has investigated the transition from handcrafted based systems to learning based systems (Khan & Yong, 2016; Liu et al.). However, this transition has been gradual over the past few years. In (Khan & Yong, 2016), Khan *et al.* presented a comparison between handcrafted

and CNN features in medical image modality classification based on local image features. They showed that handcrafted features outperform CNN features. This finding was explained by the data intensive nature of the used architecture which make it not enough discriminant when using a limited amount of training data. Luo *et al.* (Luo et al., 2018a) have shown also that CNNs require a relatively large amount of training data to achieve high accuracy. However, the scarcity of labeled data especially in the medical field is a major bottleneck for training such models.

Transfer learning from natural images has been widely used in medical imaging as one of the potential options to mitigate this inherent data issue (Liu et al., 2020; Menegola et al., 2017; Shan et al., 2020). Tajbakhsh *et al.* (Tajbakhsh et al., 2016) proposed to fine-tune adequately a pre-trained network for colonoscopy frame classification. Promising results were obtained despite the difference between medical images and ImageNet dataset (Deng et al., 2009) on which the network has been trained. Several studies show also that transfer learning can improve performances in medical applications (Alzubaidi et al., 2020; Li et al., 2017; Morid et al., 2021).

Most recently, self-supervised learning has gained popularity as it is a solution to deal with the problem of scarce labeled data. This method has shown a great success in several applications (Misra & Maaten, 2020; Jing & Tian, 2020; Goyal et al., 2019), but less attention in medical image analysis (Spitzer et al., 2018; Bai et al., 2019; Zhuang et al., 2019; Jing & Tian, 2019). Azizi *et al.* (Azizi et al., 2021) showed that self-supervised strategy, with unlabeled medical images, significantly outperforms transfer learning strategy.

3. Proposed approach

In this section, we first outline the general principle of our proposed matching method for determining the correspondence between endoscopic image key-points. Then we go over our model architecture in detail, as well as the training steps.

3.1. Image matching approach

The general principal of our image matching approach is depicted in Figure 1. Let I_t and I_{t+1} be two consecutive endoscopic images to be matched. Image key-point detector is applied to the input two images to extract N_t and N_{t+1} key-points located in I_t and I_{t+1} respectively. We then crop around each key-point a squared patch of 128×128 pixels yielding in two sets of patches $\{P_i^t\}_{i=1:N_t}$ and $\{P_j^{t+1}\}_{j=1:N_{t+1}}$. The aim of the remaining processing is to find the correspondence between these two sets of patches. To do this, we use a CNN to transform each patch into a more discriminative representation space, and then solve the matching problem by minimizing the Euclidean distance between patches in that embedding space following equation 1. In 1, $f(.)$ represents the CNN transformation of a given patch, $\hat{j} \in [1 : N_{t+1}]$ is the optimal index matching the i-th patch from I_t, and $\|.\|_2$ stands for the Euclidean distance.

$$\hat{j} = \underset{j}{\operatorname{argmin}} \left\| f(P_j^{t+1}) - f(P_i^t) \right\|_2 \tag{1}$$

3.2. CNN model architecture

The architecture of our network is inspired by L2-Net (Tian et al., 2017). It consists of seven convolutional layers with 128-D feature vector output. As shown in Figure 2, the first six convolution layers have small kernel size (3×3) and are followed by batch normalization and ReLU. The last layer has a kernel

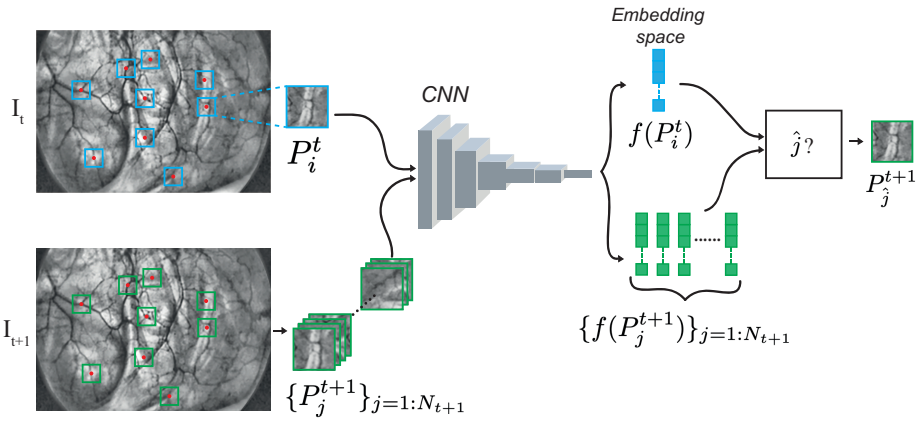

Figure 1: Overview of the matching process.

size of (8×8) and is followed only by batch normalization. Batch normalization is considered only for CNN training phase. The number of filters by convolution layer is respectively $\{16, 16, 32, 64, 128, 128, 128\}$. The padding is set to 1 for all layers (except in the last layer). We used convolution stride equals to 2 instead of pooling layers.

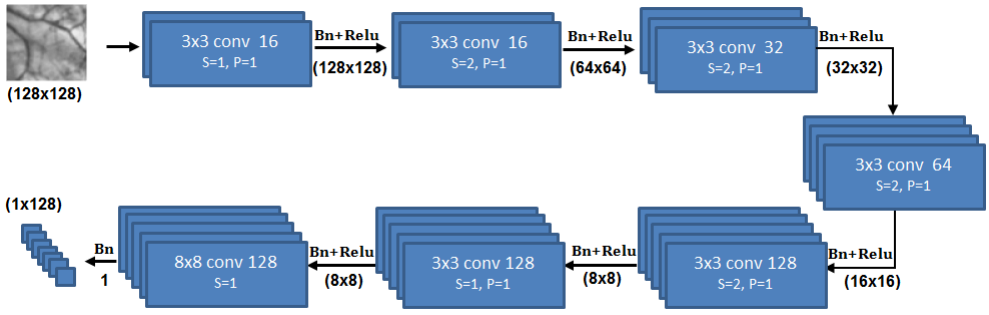

Figure 2: Network architecture.

3.3. CNN model training

In our approach, the CNN model training phase is critical because it must produce a model with high discriminative capability without using labeled data. A typical supervised-learning scheme would rely entirely on key-point matching

ground truth in addition to the input video frames for training. Instead, we designed a self-supervised training approach based on a triplet loss architecture that only requires raw endoscopic video frames. Indeed, triplet loss, introduced in (Schroff et al., 2015) as FaceNet model, has been successfully used in several tasks (Grati et al., 2020; Harvill et al., 2019; Kumar et al., 2021). As depicted

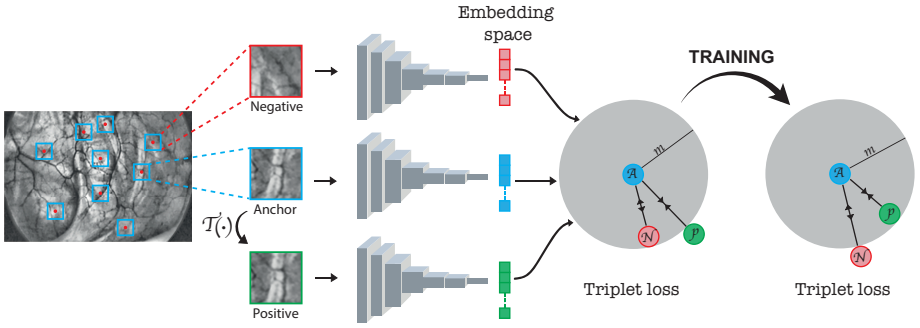

Figure 3: Overview of the triplet loss training approach. Anchor, positive and negative patches are fed to the CNN and trained so that the anchor-positive distance is minimized and the anchor-negative one is maximized in the feature space. As a result, a more discriminative embedding space is learned.

in Figure 3, the training requires a triplet input image patches. For sake of simplicity, we note them (P_i, P_i^+, P_i^-) where P_i is the anchor patch, P_i^+ is the positive patch (has a similar appearance as P_i) and P_i^- is the negative patch. An anchor patch is obtained by cropping a window around a given key-point. Its corresponding positive patch is generated by applying a homography transformation $T(.)$ to the anchor patch. $T(.)$ is a simulated transformation that often exists in endoscopic images. The negative patch could be any other patch selected from the same endoscopic image.

The method of selecting positive and negative patches to form triplets is critical, as random selection does not always produce good results. Various triplet selection approaches have been proposed in several studies (Hermans et al., 2017; Cui et al., 2016; Yu et al., 2018). In our case, we use the same strategy proposed in recent HardNet method (Mishchuk et al., 2017). It aims

at minimizing the distance between anchor and positive patches while maximizing the distance between the anchor and the nearest negative patch. The loss equation is defined as follows:

$$L = \frac{1}{N_t}\sum_{i=0}^{N_t} max\left(0, m + d(f(P_i) - f(P_i^+)) - min(d(f(P_i) - f(P_{j_{min}}^+)), d(f(P_{k_{min}}) - f(P_i^+))\right) \tag{2}$$

where $d(f(P_i), f(P_i^+)) = \sqrt{2 - 2f(P_i)f(P_i^+)}$, $P_{j_{min}}^+$ is the second nearest neighbor to P_i after P_i^+, $P_{k_{min}}$ is the nearest non-matching anchor to P_i^+, m is a margin scalar, j_{min} and k_{min} are defined respectively in equations 3 and 4.

$$j_{min} = \underset{j=1..N_t, j\neq i}{\operatorname{argmin}} d(f(P_i) - f(P_j^+)) \tag{3}$$

$$k_{min} = \underset{k=1..N_t, k\neq i}{\operatorname{argmin}} d(f(P_k) - f(P_i^+)) \tag{4}$$

4. Experiments and results

The evaluation of the proposed image key-points matching is performed in several stages to study different aspects of the approach. The evaluation experiments are conducted in three experimental sets:

- Intrinsic evaluation of our method: we assess the robustness of our method to typical geometric transformation variations between frames such as view point, scale, and blurring changes. In addition, we back up our HardNet loss choice by evaluating different triplet loss variants.
- Comparison to state-of-the-art local feature descriptor methods: we consider both handcrafted and deep learning methods.
- Use-case of endoscopic image mosaicing: we qualitatively demonstrate the efficiency of our image key-points matching in the use case of endoscopic

image mosaicing.

Before delving into the details of these experiments, we will first present our dataset, and go over our training settings.

4.1. Database description

To generate our training database, five human bladder endoscopic videos from different patients acquired with the same endoscope have been used. Samples of such images are shown in Figure 4. In all of the conducted experiments, we use 4 out of the 5 videos for training and keep the remaining video for validation in a cross-validation scheme.

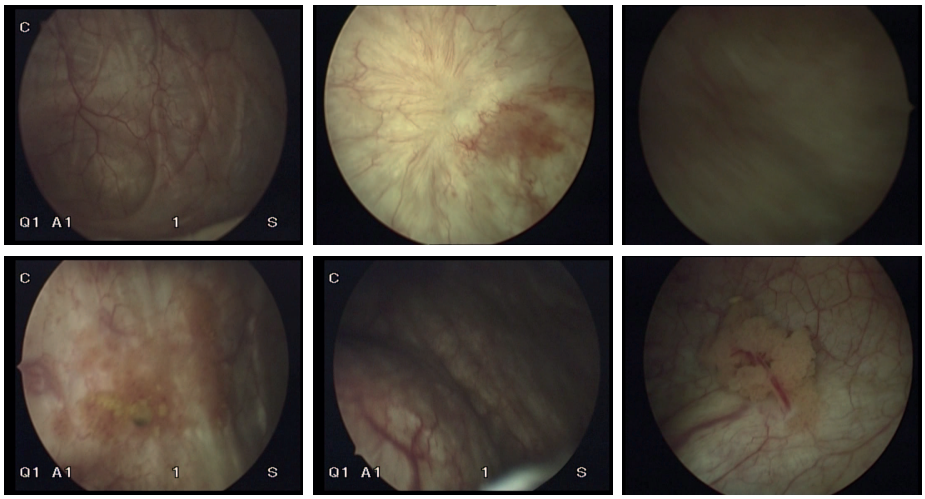

Figure 4: Samples of clinical endoscopic images taken from different patients.

We first convert raw data into gray-scale images and enhance contrast based on CLAHE method. Comes after key-points detection and anchors cropping. As explained in the previous section, positive samples are obtained after applying any appropriate transformation to endoscopic frames. In our case, we applied rotations with small angles ($\theta \in [5, 10, 15]$ degrees), scaling with small scale factor ($S_f = [0.9, 0.95, 1.05, 1.1, 1.15]$) and/or small translations in both axes

(*e.g.,* 8 pixels). In total, we created a training dataset composed of about 20k patch.

4.2. Training settings

The database was trained using Stochastic Gradient Descent with a batch size of 128 and an initial learning rate of 0.001 and a momentum set to 0.9. These are the final hyperparameters we considered after several runs, with the goal of improving convergence speed. The model converges in about 28 hours of training on a NVIDIA Titan V GPU with 12GB memory. The margin m of equation 2 is experimentally fixed to 1. The matching between 576×720 image pairs completes in less than 1 second. In our application context, the run-time is less important because panoramic image construction is typically done offline.

4.3. Intrinsic evaluation of the proposed method

4.3.1. Robustness to image transformation variations

To evaluate the robustness of the proposed local feature matching method against image transformation variations, we applied various combinations of geometric transformation to each image from the validation video in order to assess the robustness of our method against specific transformations (*i.e.,* viewpoint change, scale change, blurring). Performances have been evaluated in terms of *recall* with regards to *precision* (Mikolajczyk & Schmid, 2005). The number of correct and false matches are determined by the projection error PE instead of the overlap error used in (Mikolajczyk & Schmid, 2005). The projection error PE is defined as the Euclidean distance calculated between the matched key-points and the ground truth key-points (correct matches). In our experiments, we fixed experimentally the projection error to $PE = 5$. We also report the obtained results for well-known handcrafted descriptors (*i.e.,* AKAZE, KAZE, SURF, SIFT, ORB and BRISK) to put our results into perspective. It is worth

noting that since each original handcrafted descriptor is built on top of its own key-point detector, we decided to output the results for each of these detectors for a fair comparison. As an example, when applied to key-points detected with the code provided in the AKAZE implementation, we refer to our method as Proposed$_{AKAZE}$.

4.3.2. Robustness to viewpoint changes

To simulate typical viewpoint changes in endoscopic videos, we estimate geometric transformations between consecutive frames in clinical data yielding in a bench of homography 3×3 matrices. This estimation is performed in a standard registration scheme based on key-points detection, matching, and RANSAC algorithm (Fischler & Bolles, 1981). We applied a set of 10 randomly selected pre-computed transformations to each image from the validation endoscopic videos. Then, for each pair of images (input and transformed images), we detect key-points, extract local features and match descriptors according to the baseline to evaluate. The obtained recall and precision results are shown in Figure 5.

Figure 5 shows that the proposed descriptor outperforms the handcrafted ones. Indeed, the maximal recall value is reached for a precision value of about 99% for Proposed$_{AKAZE}$ setting. We notice also that AKAZE is outperforming the other handcrafted descriptors.

4.3.3. Robustness against scale changes

Considering that only small variations occur generally between two consecutive frames in endoscopic videos, we do not need to evaluate large scale changes. We evaluated the performance of the different descriptors only against small pre-defined scale factors $\{0.9, 0.95, 1, 1.05, 1.1, 1.15\}$. The result relative to the sensitivity of the proposed descriptor to scale changes is depicted in Figure

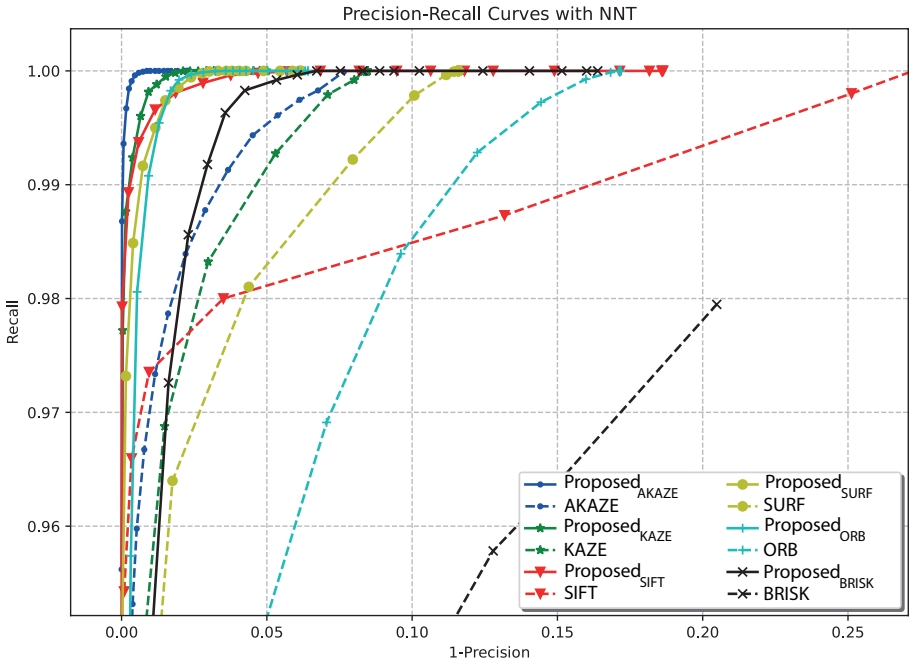

Figure 5: Recall/precision curves computed for all descriptors when varying viewpoints. The proposed descriptor outperforms the handcrafted ones, reaching the maximal recall value for a precision rate of 99% (when AKAZE is used as detector).

6

Figure 6 highlights that handcrafted descriptors are more sensitive to scaling factor changes than the proposed descriptor. Especially, for Proposed$_{ORB}$, Proposed$_{AKAZE}$, Proposed$_{KAZE}$, and Proposed$_{SURF}$ configurations, the precision of the proposed descriptor remains high (almost more than 90%) for scaling factors in the range of $[0.9, 1.1]$.

4.3.4. Robustness against blurring

Endoscopic videos are often captured with blur due to several factors such as small hand vibration, bad choice of lens focus, fast movement of endoscopic camera with a low frame rates, *etc.* Therefore, we have to quantitatively measure the descriptors efficiency regarding the blur artifact. Thus, we blurred the input

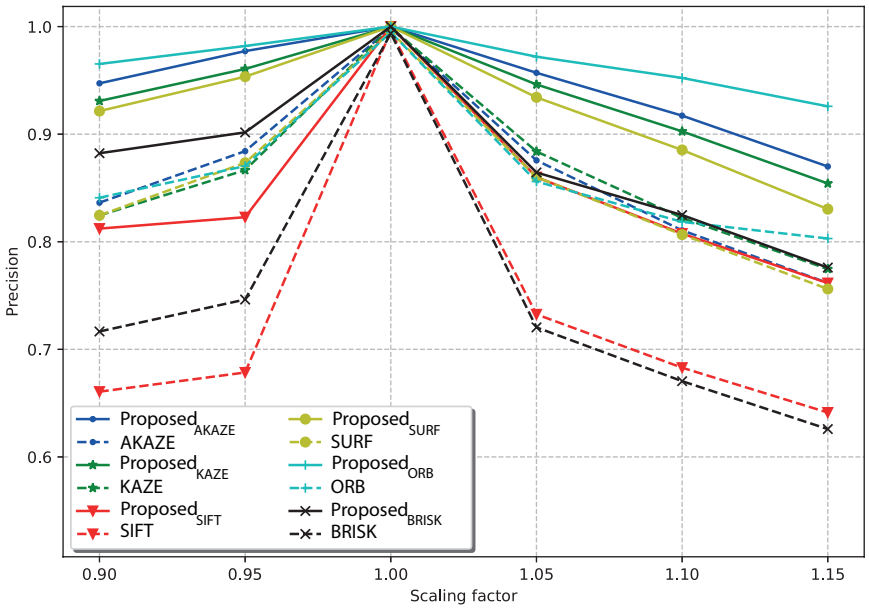

Figure 6: Precision rates computed for different scale factors. The proposed descriptor outperforms the handcrafted ones, reaching higher precision rates, showing thus less sensitivity to scaling changes.

video frames by applying various convolution kernels ($3\times3, 5\times5, 10\times10, 15\times15$).

Figure 7 illustrates quantitatively how blur affects matching precision. Compared to handcrafted descriptors, the proposed descriptor is the least sensitive to blur.

4.3.5. Comparison of triplet loss variants

In here, we consider different triplet Loss functions: the HardNet Loss (Mishchuk et al., 2017), the standard Triplet Loss with a fixed margin (Schroff et al., 2015), and the adaptive margin triplet Loss (Wang et al., 2018). The HardNet and the triplet loss functions are tested with a same margin $m = 1$. Similarly to (Mishchuk et al., 2017), we evaluate the different losses in terms of precision which is defined as the ratio of correct matches over the total number of matches and matching score which is the ratio of correct matches over the total number of detected key points. The obtained comparison results are illus-

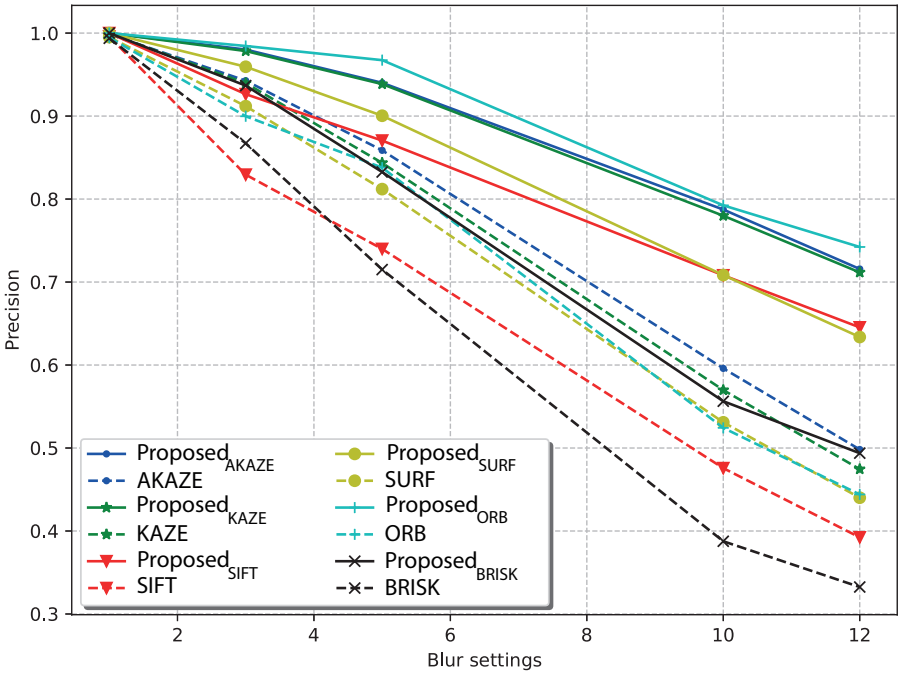

Figure 7: Precision rates computed with different blur created with different kernel sizes. The proposed descriptor is the least sensitive to blur.

trated in Table 1 showing the outperformance of HardNet loss which explains our decision to select it for our CNN model training.

4.4. Comparison to state-of-the-art local feature descriptor methods

There is no publicly available benchmark for evaluating endoscopic image matching at the moment. We decided in these experiments to annotate the frames of one validation video to cover more realistic transformations between

Table 1: Comparison of three different loss functions: HardNet Loss, Triplet Loss and the adaptive margin triplet Loss in terms of precision and score matching in %. HardNet loss shows better performance.

	Precision	*Matching Score*
HardNet Loss	**99.01**	**80.36**
Triplet Loss	95.80	79.13
Adaptative margin triplet Loss	92.19	80.34

consecutive frames. The annotation is carried out as follows. We first detect key-points in all the video frames, and then we manually annotate the correspondence between key-points with a customized tool yielding over 1000 annotated image pairs.

4.4.1. Comparison to state-of-the-art Handcrafted descriptors

These experiments are being carried out to consolidate the results previously obtained in the intrinsic validation of our method, demonstrating its superiority over all well-known handcrafted descriptors. In Figure 8, we depict the matching results between two endoscopic frames. Green and red lines refer respectively to correct and wrong matches. Compared to all the used descriptors, we can observe that a larger number of correct matches (green lines) is achieved with our approach.

For quantitative evaluation, we report the obtained matching performances in terms of recall and precision in Figure 9.

From Figure 9, we can observe that the maximal recall rate could be achieved for a precision rate of almost 98% when our descriptor is paired with SURF or KAZE. However, when handcrafted descriptors are used, the best recall value could be obtained for a maximal precision rate of 88% (in the case of SURF). A trade-off between precision and recall should be undertaken to choose the best descriptor.

4.4.2. Comparison to deep learning based supervised methods

As previously stated, no benchmarking for endoscopic image key-points matching exists in the literature. Nonetheless, to put our results into perspective, we compare the performance of our descriptor to the most recent state-of-the-art methods for key-point matching based on supervised learning: SosNet (Tian et al., 2019), SuperGlue (Sarlin et al., 2020), HyNet (Tian et al.,

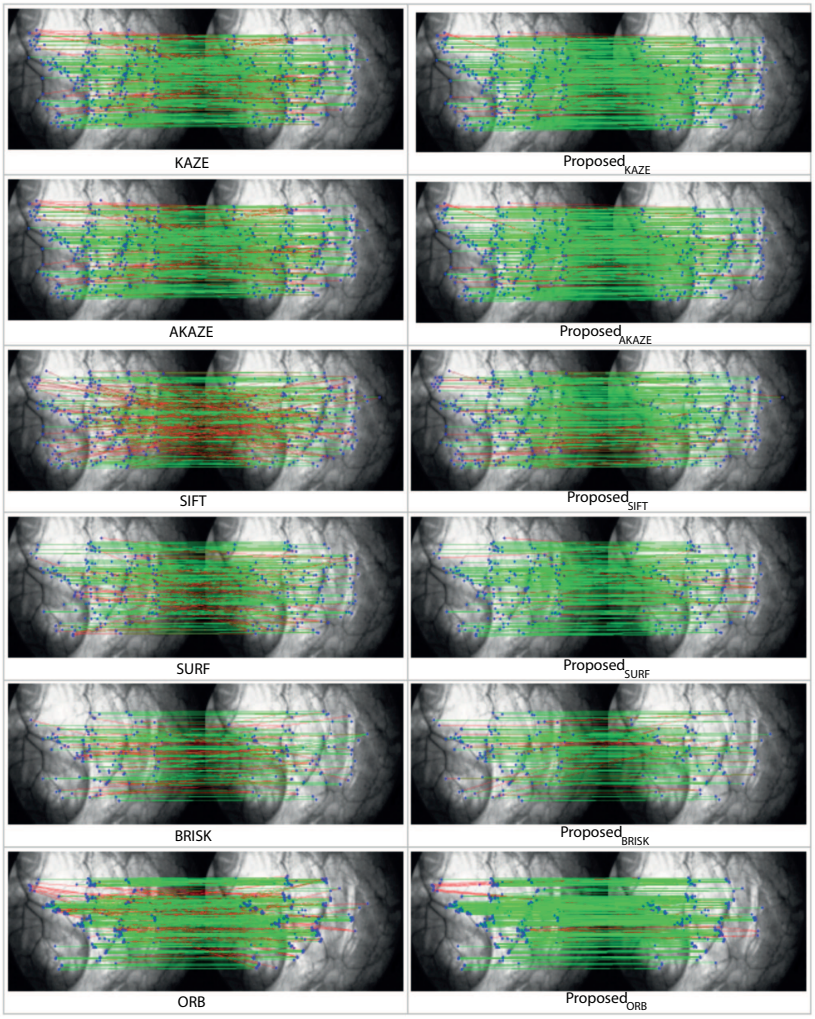

Figure 8: Qualitative evaluation of the proposed descriptor. Green and red lines refer respectively to correct and wrong matches. The best matching quality (number of green lines) is reached with the proposed descriptor.

2020), and CNDesc (Wang et al., 2022). We used 800 image pairs for training and left the remaining 200 images for testing.

Both SuperGlue and CNDesc originally require first to extract a set of feature key-points from the training and testing images. We followed the same procedure to train SuperGlue and CNDesc on our dataset. In our experiments,

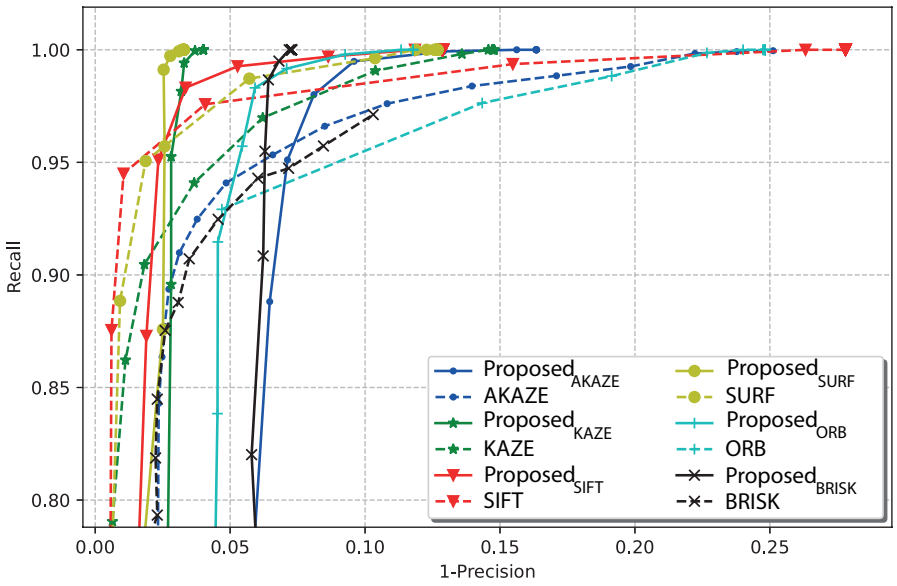

Figure 9: Quantitative evaluation of the proposed descriptor on real clinical dataset. The maximal recall rate could be achieved for a precision rate of almost 98% when our descriptor is used with SURF or KAZE detectors (*i.e.,* Proposed$_{SURF}$ and Proposed$_{KAZE}$, respectively.

feature key-points were detected using SIFT method as we did not observe any significant performance differences when choosing different key-point detectors. On the other hand, training the HyNet and SosNet models as well as our proposed model requires a set of extracted patches based on the localization of image key-points. To this end, the training patch dataset is extracted around the SIFT image key-points that have already been detected. The training for all five competing methods is therefore based on the same key-points.

Comparative results in terms of precision and matching score metrics are summarized in Table 2. Despite being the most recent work, CNDesc shows the lowest results and seems to be the least suitable to endoscopic videos. The other methods achieved competitively high performances. However, the fundamental difference between our method and the other competing ones is still the training mode and our ability to adapt automatically to the large texture variability in endoscopic videos. Indeed, we remember that in our case the training triplets are

automatically generated which means that a fine-tuning step could be carried out whenever needed on new endoscopic videos without any annotation.

Table 2: Comparison between our method and the most recent state-of-the-art methods for key-point matching based on supervised learning in terms of precision and matching score reported in (%). We notice that all the obtained results are very comparable.

	Precision	*Matching Score*
SosNet (Tian et al., 2019)	99.95	91.60
SuperGlue (Sarlin et al., 2020)	99.17	92.81
HyNet (Tian et al., 2020)	99.96	90.29
CNDesc (Wang et al., 2022)	98.37	90.51
Proposed	99.89	92.56

4.5. Use-case of endoscopic image mosaicing

To validate the effectiveness of the proposed descriptor in generating panoramic images, we feed the obtained matching results to a mosaicing system based on RANSAC algorithm (Fischler & Bolles, 1981). We construct a panoramic image with a set of 400 consecutive frames from bladder endoscopic video, as shown in Figure 10.

The obtained mosaic image is coherent and we can observe clearly the texture continuity which is a good indicator proving the precision of the images alignment. This experiment illustrates that the proposed descriptor provide reliable and robust matching features that can be utilized to construct panoramic images for endoscopic videos. However, a strong blur effect or large scale change can perturb the registration process and cause discontinuities in the resulting mosaic image. To overcome this problem, we can extend the mosaic image system with an artifact detection algorithm.

5. Conclusions and perspectives

Despite the success of deep learning approaches in a variety of computer vision tasks, a lack of labeled data remains a major barrier to the use of neural

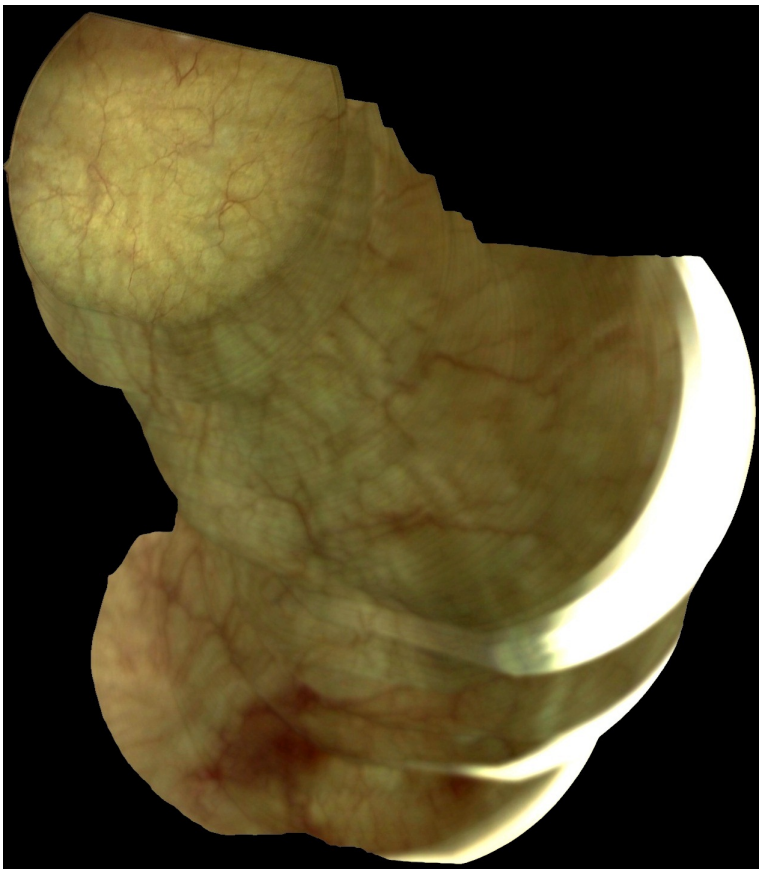

Figure 10: A 1258×1436 panoramic image constructed from a 400 consecutive frames from human bladder endoscopic video. The obtained mosaic image is coherent and we can observe clearly the texture continuity which is a good indicator proving the precision of the images alignment.

networks in medical applications. To address this issue, we proposed a self-supervised approach for endoscopic image matching in this paper, which is based on the automatic generation of a pseudo labeled data-set. As a result, our method allows us to train a local descriptor network using only endoscopic images, with no need for labeled data or manual annotation. The proposed self-supervised approach was evaluated and compared to different handcrafted image feature descriptors and also to recent deep learning based supervised methods: SosNet, SuperGlue, HyNet and CNDesc.

The experimental results proved the robustness of our descriptor against viewpoint, scaling factor, and blurring changes. Moreover, compared to the supervised state-of-the-art deep learning based methods, our approach achieves competitive performance in terms of precision and matching score while using unlabeled patches in a self-supervised training mode. For future works, we would like to investigate further our approach on endoscopic videos of other organs such as small intestine, large intestine and stomach. In addition, we can use different endoscopy system such as capsule endoscopy or blue laser endoscopy system.

References

Agrawal, M., Konolige, K., & Blas, M. (2008). Censure: Center surround extremas for realtime feature detection and matching. In *Proceedings of the European Conference on Computer Vision* (p. 102–115). Marseille, France.

Alcantarilla, P., Nuevo, J., & Bartoli, A. (2013). Fast explicit diffusion for accelerated features in nonlinear scale spaces. In T. Burghardt, D. Damen, W. Mayol-Cuevas, & M. Mirmehdi (Eds.), *Proceedings of the British Machine Vision Conference* (pp. 13.1–13.11). BMVA Press.

Alcantarilla, P. F., Bartoli, A., & Davison, A. (2012). Kaze features. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Proceedings of the British Machine Vision Conference* (pp. 214–227). Berlin, Heidelberg: Springer.

Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., J., S., Duan, Y., & Oleiwi, S. R. (2020). Towards a better understanding of transfer learning for medical imaging: A case study. *Applied Sciences*, *10*, 4523.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., & Norouzi, M. (2021). Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3478–3488).

Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S. E., Guo, Y., Matthews, P. M., & Rueckert, D. (2019). Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 541–549). Springer.

Bay, H., Ess, A., Tuytelaars, T., & V., G. L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, *110*, 346–359.

Behrens, A., Bommes, M., Stehle, T., Gross, S., Leonhardt, S., & Aach, T. (2011). Real-time image composition of bladder mosaics in fluorescence endoscopy. *Computer Science - Research and Development*, *26*, 51–64.

Behrens, A., Stehle, T., Gross, S., & Aach, T. (2009). Local and global panoramic imaging for fluorescence bladder endoscopy. *Annu Int Conf IEEE Eng Med Biol Soc*, *45*, 6990–3.

Calonder, M., Lepetit, V., Strecha, C., & Fua, P. (2010). Brief: Binary robust independent elementary features matching. In *European Conference on Computer Vision* (p. 778–792). Heraklion, Crete, Greece.

Chen, J., & Frey, E. C. (2020). Medical image segmentation via unsupervised convolutional neural network. *Medical Imaging with Deep Learning 2020*, .

Chen, J., Tian, J., Lee, N., Zheng, J., Smith, R. T., & Laine, A. F. (2010).

A partial intensity invariant feature descriptor for multimodal retinal image registration. *IEEE Transactions on Biomedical Engineering*, *57*, 1707–1718.

Chu, Y., Li, H., X., L., Ding, Y., Yang, X., Ai, D., Chen, X., Y., W., & J., Y. (2020). Endoscopic image feature matching via motion consensus and global bilateral regression. *Computer Methods and Programs in Biomedicine*, *190*, 105370.

Cui, Y., Zhou, F., Lin, Y., & Belongie, S. (2016). Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. (pp. 1153–1162).

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Du, P., Zhou, Y., Xing, Q., & H, X. (2011). Improved sift matching algorithm for 3d reconstruction from endoscopic images. In *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications* (p. 561–564).

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm formodel fitting with applications to image analysis and automated cartography. *Commu-nications of ACM*, .

Ghosh, T., Li, L., & Chakareski, J. (2018). Effective deep learning for semantic segmentation based bleeding zone detection in capsule endoscopy images. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 3034–3038).

Goyal, P., Mahajan, D., Gupta, A., & Misra, I. (2019). Scaling and benchmarking self-supervised visual representation learning. (pp. 6390–6399).

Grati, N., Ben-Hamadou, A., & Hammami, M. (2020). Learning local representations for scalable rgb-d face recognition. *Expert Systems with Applications*, *150*.

Han, X., Leung, T., Jia, Y., Sukthankar, R., & Berg, A. (2015). Matchnet: Unifying feature and metric learning for patch-based matching. *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, (pp. 3279–3286).

Harvill, J., Wahab, M. A., Lotfian, R., & Busso, C. (2019). Retrieving speech samples with similar emotional content using a triplet loss function. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7400–7404). IEEE.

Hermans, A., Beyer, L., & Leibe, B. (2017). In defense of the triplet loss for person re-identification, .

Hernandez-Matas, C., Zabulis, X., & Argyros, A. A. (2017). An experimental evaluation of the accuracy of keypoints-based retinal image registration. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 377–381).

Hernandez-Mier, Y., Blondel, W., Daul, C., Wolf, D., & Guillemin, F. (2010). Fast construction of panoramic images for cystoscopic exploration. *Comput Med Imaging Graph*, *34*, 579–92.

Jalili, J., Hejazi, S. M., Riazi-Esfahani, M., Eliasi, A., Ebrahimi, M., Seydi, M., M., A., Fard, M. A., & A., A. (2020). Retinal image mosaicking using scale-invariant feature transformation feature descriptors and voronoi diagram (erratum). *Journal of Medical Imaging*, *7*.

Jing, L., & Tian, Y. (2019). Self-supervised visual feature learning with deep

neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, .

Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, .

Khan, S., & Yong, S. (2016). A comparison of deep learning and hand crafted features in medical image modality classification. In *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)* (pp. 633–638).

Kim, Y., Bae, J., & Chung, J. e. a. (2021). New polyp image classification technique using transfer learning of network-in-network structure in endoscopic images. *Sci Rep*, *11*.

Kumar, P., Jain, S., Raman, B., Roy, P. P., & Iwamura, M. (2021). End-to-end triplet loss based emotion embedding system for speech emotion recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 8766–8773). IEEE.

Leutenegger, S., Chli, M., & Siegwart, R. Y. (2011). Brisk: Binary robust invariant scalable keypoints. In *2011 Int. Conf. on Computer Vision* (pp. 2548–2555).

Li, X., Zhang, H., Zhang, X., Liu, H., & Xie, G. (2017). Exploring transfer learning for gastrointestinal bleeding detection on small-size imbalanced endoscopy images. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1994–1997).

Li, Y., Li, W., Xiong, J., Xia, J., & Xie, Y. (2020). Comparison of supervised and unsupervised deep learning methods for medical image synthesis between

computed tomography and magnetic resonance images. *BioMed Research International*, *2020*.

Liu, D., Liu, Y., Li, S., Li, W., & Wang, L. (). Fusion of handcrafted and deep features for medical image classification. *Journal of Physics: Conference Series*, *1345*, 2019.

Liu, G., Hua, J., Wu, Z., Meng, T., Sun, M., Huang, P., He, X., Sun, W., Li, X., & Chen, Y. (2020). Automatic classification of esophageal lesions in endoscopic images using a convolutional neural network. *Annals of Translational Medicine*, *8*.

Liu, X., Meng, C., Tian, F.-P., & Feng, W. (2021). Dgd-net: Local descriptor guided keypoint detection network. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). doi:10.1109/ICME51207.2021.9428406.

Liu, Y., Tian, J., Hu, R., Yang, B., Liu, S., Yin, L., & Zheng, W. (2022). Improved feature point pair purification algorithm based on sift during endoscope image stitching. *Frontiers in Neurorobotics*, *16*. doi:10.3389/fnbot.2022.840594.

Lowe, D. G. (2004). Sift—the scale invariant feature transform. *Int. Journal of Computer Vision*, *60*, 91–110.

Luo, C., Li, X., Wang, L., He, J., Li, D., & Zhou, J. (2018a). How does the data set affect cnn-based image classification performance? In *2018 5th International Conference on Systems and Informatics (ICSAI)* (pp. 361–366).

Luo, Z., Shen, T., Zhou, L., Zhu, S., Zhang, R., Yao, Y., Fang, T., & Quan, L. (2018b). Geodesc: Learning local descriptors by integrating geometry constraints. In *ECCV*.

Ma, J., Jiang, X., Fan, A., Jiang, J., & Yan, J. (2021). Image matching from handcrafted to deep features: A survey. *Int. Journal of Computer Vision*, *129*.

Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F. V., Avila, S., & Valle, E. (2017). Knowledge transfer for melanoma screening with deep learning. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (pp. 297–300).

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *27*, 1615–1630.

Miranda-Luna, R., Daul, C., Blondel, W., Hernandez-Mier, Y., Wolf, D., & Guillemin, F. (2008). Ieee trans biomed eng. *Journal of Computer Science*, *55*, 541–553.

Mishchuk, A., Mishkin, D., Radenović, F., & Matas, J. (2017). Working hard to know your neighbor's margins: Local descriptor learning loss. In *NIPS*.

Misra, I., & Maaten, L. v. d. (2020). Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6707–6717).

Morid, M. A., Borjali, A., & Del Fiol, G. (2021). A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in biology and medicine*, *128*, 104115.

Pietikäinen, M., Hadid, A., Zhao, G., & Ahonen, T. (2011). *Computer Vision Using Binary Patterns*. (1st ed.). Verlag London: Springer.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). Orb: An efficient

alternative to sift or surf. In *2011 Int. Conf. on Computer Vision* (pp. 2564–2571). Barcelona, Spain.

Saha, S., Xiao, D., Frost, S., & Kanagasingam, Y. (2016). A two-step approach for longitudinal registration of retinal images. *Journal of Medical Systems*, *40*.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). SuperGlue: Learning feature matching with graph neural networks. In *CVPR*.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, .

Shan, H., Jia, X., Yan, P., Li3, Y., Paganetti, H., & Wang, G. (2020). Synergizing medical imaging and radiotherapy with deep learning. *Machine Learning: Science and Technology*, *1*.

Sharib, A., C., D., Galbrun, E., Guillemin, F., & Blondel, W. (2016). Anisotropic motion estimation on edge preserving riesz wavelets for robust video mosaicing. *Pattern Recognition*, *51*, 425–442.

Sharib, A., Daul, C., Weibel, T., & Blondel, W. (2013). Fast mosaicing of cystoscopic images from dense correspondence: Combined surf and tv-l1 optical flow method. In *2013 IEEE Int. Conf. on Image Processing* (pp. 1291–1295).

Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., & Moreno-Noguer, F. (2015). Discriminative learning of deep convolutional feature point descriptors. In *2015 IEEE Int. Conf. on Computer Vision (ICCV)* (pp. 118–126).

Spitzer, H., Kiwitz, K., Amunts, K., Harmeling, S., & Dickscheid, T. (2018). Improving cytoarchitectonic segmentation of human brain areas with self-

supervised siamese networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 663–671). Springer.

Sun, J., Shen, Z., Wang, Y., Bao, H., & Zhou, X. (2021). LoFTR: Detector-free local feature matching with transformers. *CVPR*, .

Sung, H., Ferlay, J., L., S. R., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2020). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *0*, 1–41.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, *35*, 1299–1312.

Tian, Y., Barroso Laguna, A., Ng, T., Balntas, V., & Mikolajczyk, K. (2020). Hynet: Learning local descriptor with hybrid similarity measure and triplet loss. In *NeurIPS*.

Tian, Y., Fan, B., & Wu, F. (2017). L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (pp. 6128–6136).

Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., & Balntas, V. (2019). Sosnet: Second order similarity regularization for local descriptor learning. In *CVPR*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (p. 6000–6010). Curran Associates, Inc. volume 30.

Wang, C., Xu, R., Xu, S., Meng, W., & Zhang, X. (2022). CNDesc: Cross Normalization for Local Descriptors Learning. *IEEE Transactions on Multimedia*, (pp. 1–1). doi:10.1109/TMM.2022.3169331.

Wang, J., Zhou, S., Wang, J., & Hou, Q. (2018). Deep ranking model by large adaptive margin learning for person re-identification. *Pattern Recognition*, *74*, 241–252.

Weibel, T., Daul, C., Wolf, D., Rosch, R., & Guillemin, F. (2012). Graph based construction of textured large field of view mosaics for bladder cancer diagnosis. *Pattern Recognition*, *45*, 4138–4150.

Wiles, O., Ehrhardt, S., & Zisserman, A. (2021). Co-attention for conditioned image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 15920–15929).

Yu, B., Liu, T., Gong, M., Ding, C., & Tao, D. (2018). Correcting the triplet selection bias for triplet loss. In *Computer Vision – ECCV 2018* (pp. 71–86). volume 11210.

Zenteno, O., Trinh, D.-H., Treuillet, S., Lucas, Y., Bazin, T., Lamarque, D., & Daul, C. (2022). Optical biopsy mapping on endoscopic image mosaics with a marker-free probe. *Computers in Biology and Medicine*, *143*, 105234.

Zhang, Z., Wang, L., andLirong Yin, W. Z., Hu, R., & Yang, B. (2022). Endoscope image mosaic based on pyramid orb. *Biomedical Signal Processing and Control*, *71*, 103261. doi:https://doi.org/10.1016/j.bspc.2021.103261.

Zhou, Q., Sattler, T., & Leal-Taixe, L. (2021). Patch2pix: Epipolar-guided pixel-level correspondences. In *CVPR*.

Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., & Zheng, Y. (2019). Self-supervised feature learning for 3d medical images by playing a rubik's cube.

In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 420–428). Springer.

Zou, S., Long, M., Wang, X., Xie, X., Li, G., & Wang, Z. (2019). A cnn-based blind denoising method for endoscopic images. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (pp. 1–4).