



S2P-Matching: Self-Supervised Patch-Based Matching Using Transformer for Capsule Endoscopic Images Stitching

Feng Lu , Dao Zhou, Haoyang Chen , Shuai Liu, Xianliang Ling, Lei Zhu , *Member, IEEE*, Tingting Gong, Bin Sheng , *Member, IEEE*, Xiaofei Liao , *Member, IEEE*, Hai Jin , *Fellow, IEEE*, Ping Li , *Member, IEEE*, and David Dagan Feng , *Life Fellow, IEEE*

Abstract—The Magnetically Controlled Capsule Endoscopy (MCCE) has a limited shooting range, resulting in capturing numerous fragmented images and an inability to precisely locate and examine the region of interest (ROI) as traditional endoscopy can. Addressing this issue, image stitching around the ROI can be employed to aid in the diagnosis of gastrointestinal (GI) tract conditions. However, MCCE images possess unique characteristics, such as weak texture, close-up shooting, and large angle rotation, presenting challenges to current image-matching methods. In this context, a method named S2P-Matching is proposed for self-supervised patch-based matching in MCCE image stitching. The method involves augmenting the raw data by simulating the capsule endoscopic camera's behavior around the GI tract's ROI. Subsequently, an improved contrast learning encoder is utilized to extract local features, represented as deep feature descriptors. This encoder comprises two branches that extract distinct scale

features, which are combined over the channel without manual labeling. The data-driven descriptors are then input into a Transformer model to obtain patch-level matches by learning the globally consented matching priors in the pseudo-ground-truth match pairs. Finally, the patch-level matching is refined and filtered to the pixel-level. The experimental results on real-world MCCE images demonstrate that S2P-Matching provides enhanced accuracy in addressing challenging issues in the GI tract environment with image parallax. The performance improvement can reach up to 203 and 55.8% in terms of NCM (Number of Correct Matches) and SR (Success Rate), respectively. This approach is expected to facilitate the wide adoption of MCCE-based gastrointestinal screening.

Index Terms—Capsule endoscopy, image stitching, multi-view simulation, patch-level matching, self-supervised contrastive learning, transformer.

Received 31 December 2023; revised 22 April 2024 and 22 August 2024; accepted 11 September 2024. Date of publication 20 September 2024; date of current version 22 January 2025. This work was supported in part by the Key Project of the National Natural Science Foundation of China under Grant 62232012 and Grant 62272298, in part by the Hubei Big Data Analysis Platform and Intelligent Service Project for Medical and Health, in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102, and in part by The Hong Kong Polytechnic University under Grant P0030419 and Grant P0035358. (Feng Lu and Dao Zhou contributed equally to this work.) (Corresponding author: Bin Sheng.)

Feng Lu, Haoyang Chen, Shuai Liu, Xianliang Ling, Xiaofei Liao, and Hai Jin are with the National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, China.

Dao Zhou is with the Key Laboratory of Cognitive Science of State Ethnic Affairs Commission, College of Biomedical Engineering, South-Central Minzu University, China.

Lei Zhu is with the ROAS Thrust, System Hub, The Hong Kong University of Science and Technology, China.

Tingting Gong is with the Department of Gastroenterology, Ruijin Hospital, School of Medicine, Shanghai Jiao Tong University, China.

Bin Sheng is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shengbin@sjtu.edu.cn).

Ping Li is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, and also with the School of Design, The Hong Kong Polytechnic University, Hong Kong.

David Dagan Feng is with the Biomedical and Multimedia Information Technology Research Group, School of Computer Science, The University of Sydney, Australia.

Digital Object Identifier 10.1109/TBME.2024.3462502

I. INTRODUCTION

THE non-invasive, painless, and non-cross infection capsule endoscopic equipment, such as Wireless Capsule Endoscopy (WCE) [1], [2] and Magnetically Controlled Capsule Endoscopy (MCCE) [3], has exhibited the potential to replace traditional endoscopy in clinical applications, including gastrointestinal (GI) ulcers and bleeding, inflammation, and cancers early detection. Unfortunately, the capsule endoscopy's movement is primarily dependent on GI peristalsis due to hardware limitations [4], [5], [6]. The limitations in controlling the lens during endoscopic procedures restrict the surgeon's ability to accurately investigate the region of interest (ROI) [7]. Instead, they typically capture continuously taken, fragmented, and area-overlapping endoscopic image frames [8], as depicted in Fig. 1. For instance, in MCCE, physicians commonly capture redundant and fragmented image frames while maneuvering around the ROI. Using artificial intelligence, these fragmented images can be seamlessly stitched together, enabling doctors to observe the candidate lesion area within a broader field of view and facilitating more accurate lesion detection.

Stitching capsule endoscopic images together is complex and presents several obstacles. One significant challenge is that each image captures only a limited area, typically ranging from 20 mm ~ 60 mm with the MCCE lens, resulting in fragmented

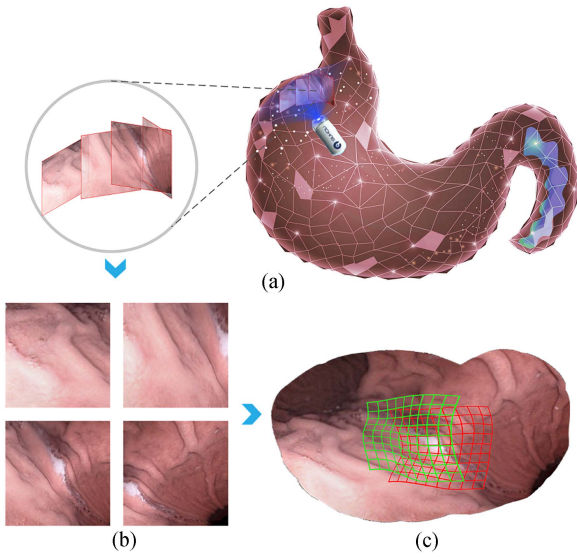


Fig. 1. Capsule endoscopic image stitching. (a) The capsule endoscopic moving in the 3D GI environment captures images. (b) The fragmented images captured by the floating camera are tiny, weakly textured, and rotation and transformation caused by close-up shooting. (c) An instance illustrates that our self-supervised patch-based matching is capable of producing a reliable match when subjected to a range of transformations, including multi-view camera position and rotation transformations.

images. Moreover, these images often contain repeated weak texture regions that are difficult to distinguish using existing feature descriptor-based methods [9], [10], [11], which focus on designing suitable handcrafted local feature descriptors like SIFT and ORB. As a result, these methods struggle to detect a sufficient number of reliable matching points in environments with sparse features. Additionally, they are not specifically optimized to address challenges encountered in capsule endoscopy, such as changes in camera perspective. Therefore, their effectiveness in handling the stitching of capsule endoscopic images with weak textures and complex scenes is limited.

An effective strategy to solve the problem of stitching images with insignificant discrimination is the patch-based matching method [12], [13]. It uses patch descriptors based on deep learning to reflect patch similarity and generalize well against indistinctive regions with low-textures, motion blur, or repetitive patterns. Similar to Sun et al. [13] used the Transformer model to find correspondences in the indistinctive regions based on the local neighborhood and a global context to identify the fuzzy patches. We also use the Transformer model to find the matching patches with minor discrimination in the capsule endoscopic images after expanding the receptive field of features. They are more likely to be distinguished by their location and relationships from the global perspective.

The primary prerequisite for deep learning is sufficient labeled training data. For example, the training datasets used by Sun et al. [13] include ScanNet [14] and MegaDepth [15], which contains 1613 monocular sequences with ground-truth poses and depth maps and One million internet images of 196 labeled outdoor scenes, respectively. However, obtaining a ground-truth dataset for capsule endoscopic image stitching takes much work. The images of the entire GI tract are massive and fragmented.

For example, the number of MCCE images per patient examination is about 57,600 [16]. Therefore, manually labeling these images is usually time-consuming, subjective, and error-prone. Especially when targeting fragmented images at a specific angle in the GI scene, the available images for matching training are very scarce. Fortunately, we can solve this obstacle with self-supervised contrastive learning with non-labeled training data. We build a pseudo-ground-truth dataset with matching pairs by simulating the shooting scene of capsule endoscopy in the GI tract. With this dataset, we can train models to find patch-based matching pairs without labeling.

This paper presents a self-supervised patch-based matching (S2P-Matching) that overcomes challenges in capsule endoscopic image stitching. Our methodology involves simulating the behavior of the capsule endoscopic camera in the GI tract to enhance the raw data. We then utilize modified contrastive learning to extract deep feature descriptors as local features. To achieve this, we employ a two-branch encoder to extract features from both the patch and the background patch, which are then connected over the channel. In this way, the feature of the original image's patch is packed into a $1D - d_e$ dimension vector. The compressed features are then fed into the Transformer model to obtain matches by learning the globally consented matching priors in the pseudo-ground-truth match pairs. As a result, we get the final accurate pixel-level matching by refining and filtering the patch-level matching to pixel-level one. After comparing the typical matching methods, we find that S2P-Matching can stitch capsule endoscopic images with more accurate results and offers the best tradeoff in simplicity and accuracy. Our contributions can be summarized as follows:

- We propose a S2P-Matching framework for matching capsule endoscopy images. By leveraging improved contrastive learning for patch-level matching, our framework effectively addresses the challenge of complex matching where images exhibit weak textures, close-up shooting, and significant rotations without labels.
- We have enhanced the contrastive learning feature extraction process with a dual-branch encoder that can hierarchically acquire image patch and background patch features. After consolidating these features on a multi-scale channel, the downstream transformer module can accurately discern feature correlations for patch-level matching.
- We evaluate our S2P-Matching on real-world datasets. The results show that our method could stitch more accurately than the state-of-the-art matching methods. The performance improvement can reach at most 203 and 55.8% in terms of NCM (Number of Correct Matches) and SR (Success Rate), respectively. Our method's stitching effect is natural, with no obvious texture misplacement or excessive scaling texture connection.

II. RELATED WORK

Current methods for registering capsule endoscopic images rely on handcrafted local feature descriptors designed to match endoscopic images. These descriptors are used to calculate the similarity distance between feature points and generate matches.

For instance, Xie et al. [9] used the ORB feature descriptor to reconstruct feature maps of the human colon, while Liu et al. [10] extracted features based on SIFT feature descriptors and combined matching algorithms. Zhang et al. [11] designed a Gaussian pyramid ORB endoscopic feature descriptor. While these descriptors can reduce the search space for matches and are often sufficient for a common task, they may struggle with extracting feature points between capsule endoscopic images due to challenges such as poor texture, viewpoint change, close-up shooting, and repetitive patterns [13].

Researchers recently used the patch-level matching method based on deep learning to match medical images. They remove the feature detector phase and use deep learning to produce dense descriptors representing the medical image's patches to refine the matching results. For instance, KdO-Net [18] has improved the efficiency of deep convolutional neural networks applied in the 3D pairwise point feature matching with patches. Zhou et al. [17] proposed a detect-to-refine method. They first predict and refine match proposals at the patch level. The features of the patches are extracted using the ResNet34 backbone and are fed into a correspondence network for the detection of match proposals. In image matching, homography estimation, and localization tasks, they found that this method significantly improves the performance of correspondence networks. However, it is still not accessible to capsule endoscopic image stitching. This is because regions of low texture or repetitive patterns occupy most of the field of view. It is only possible to find correct correspondences with specific repeatable interest patches.

Recent natural image-matching methods prefer Transformer networks to expand the receptive field and extract accurate matching regions with little discrimination. For example, SuperGlue [19] learns the matches of two interest point sets with a Transformer network, a graph neural network. LoFTR [13] extends this idea. With an FPN to extract image features, a local feature Transformer module generates stable features for matching. The Transformer model has a global receptive field, but its computation will significantly increase. Since the patch-level method can reduce the amount of input, the Transformer may work well for patch-level matching of medical images.

Handcrafted features may outperform deep learning-based features in medical image analysis. Lee et al. [20] conducted a study comparing the performance of handcrafted and CNN features in modality classification and found that handcrafted features were superior to deep learning-based ones. This may be due to the limited availability of training data, a significant bottleneck hindering the application of deep learning in medical images. To address this issue, researchers have been exploring self-supervised and unsupervised learning techniques [21], [22], [23], [24]. For instance, Liao et al. [22] introduced a dual-scale unsupervised deep-learning method for matching ROIs in consecutive WCE frames. Their approach outperformed non-deep-learning methods on the WCE dataset. Similarly, Farhat et al. [23] devised a self-supervised training technique for endoscopic image key-point matching using a triplet loss architecture to address the issue of limited labeled data availability.

In recent years, contrastive learning applied to self-supervised deep learning has led to state-of-the-art performance. The core

idea of contrastive learning is to pull an anchor and a 'positive' sample together in the embedding space and to push the anchor away from many 'negative' samples. For example, Chen et al. [25] proposed a contrastive learning approach that improves the quality of learned representation by using a large number of minibatch instances to obtain negative samples for each training instance. He et al. [26] construct a dynamic dictionary containing queues and moving average encoders from the perspective of contrast learning. It enables the construction of an extensive, consistent dictionary on the fly, thus facilitating contrast unsupervised learning. Researches has reported that data augmentation is the key for self-supervised training. Simon et al. [27] find that the matching algorithm can suppose a multiview bootstrapping method to augment massive labeled data and applied successfully on hand keypoint detection. SimCLR [25] shows that stronger data augmentations help to bring accuracy gains with contrastive learning and introduced spatial/geometric and appearance transformations. Experiments have shown that this approach can improve the performance of models on datasets such as ImageNet-100. NICE-Trans et al. [28] performs joint affine and deformable coarse-to-fine registration outperform the coarse-to-fine or transformer-based deep registration methods on registration accuracy. It inspired us to believe that the ground truth of matching points can be obtained by generating appropriate simulations on original data.

III. PROPOSED METHOD

As shown in Figs. 2, 3 and Algorithm 1, the S2P-Matching has five parts: data augmentation by capsule endoscopic camera behavior simulation, deep feature descriptor extraction, patch-level matching via the Transformer, refining patch-level matching to pixel-level matching, and correct correspondence filtering. We first use affine transformation to simulate the shooting behavior of the capsule endoscope camera in the GI tract, generating multiple simulated images at different positions and angles. Then, we combine the simulated images to acquire an extended data set of reference images and get the pseudo-ground-truth matching pair dataset. We use the deep feature descriptor extraction module based on the dataset to narrow the feature distance between the reference image and the simulated matching one and obtain the images' feature descriptors without labeling the matching points. The patch-level matching module next matches the patch features via the Transformer [13] to get the matching pool. After that, the matching granularity is refined by the PatchtoPixel [17] method to obtain pixel-level matching results. Finally, the correct correspondence filtering module filters the matching pairs to get accurate pixel-level matching results.

A. Data Augmentation by Capsule Endoscopic Camera Behavior Simulation

A capsule endoscopy camera always floats in the GI tract for close-up photography. As a result, we get massive fragmented GI tract images. They have many repetitive, weakly textured regions and are challenging to annotate. We use data augmentation to obtain pseudo-ground truth datasets with matching relationships, avoiding tedious manual annotation. Based on

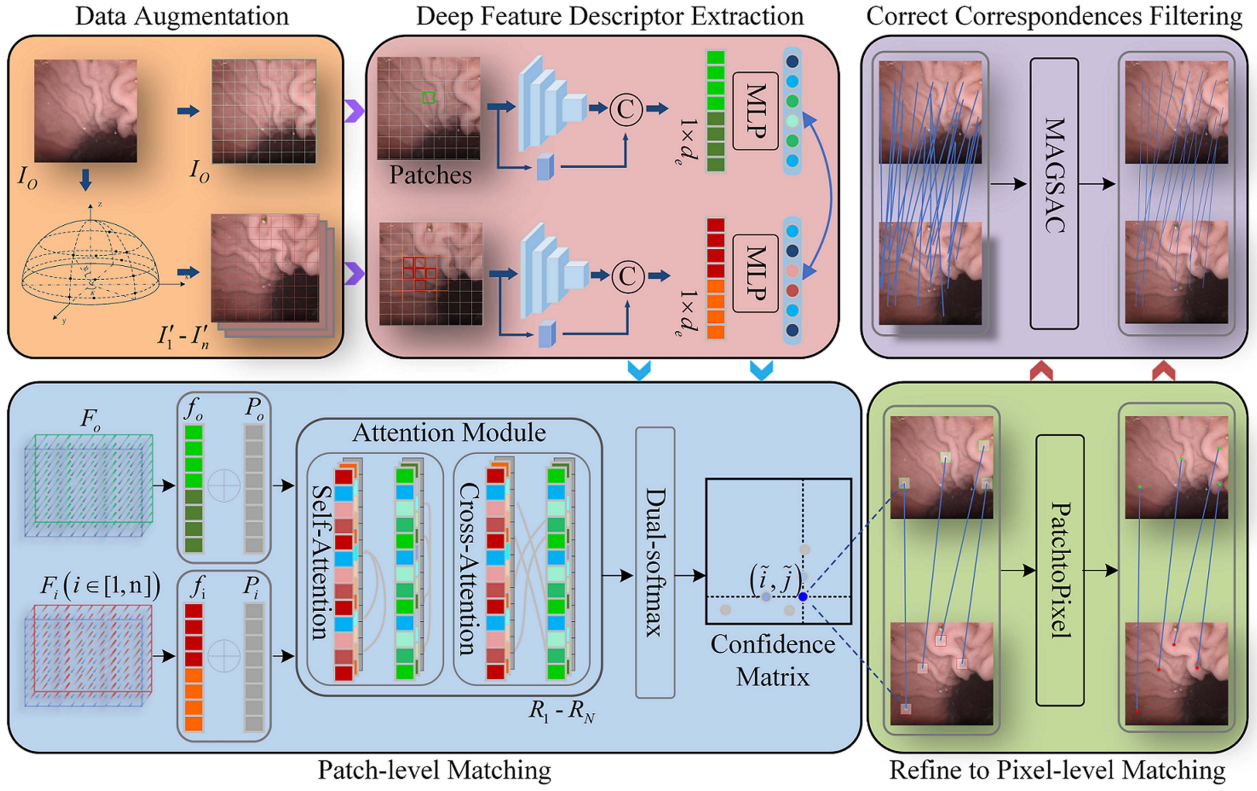


Fig. 2. The overview of the S2P-Matching. The data augmentation module generates additional simulated images for each reference image (I_0) and gets the pseudo-ground-truth match pair ($I'_{i, i \in [1, n]}$) dataset. Then, the match pairs are segmented into patches, and deep feature descriptors are extracted. The patch features are combined and fed into the Transformer [13] module to obtain the globally-consented matching priors by learning the matching pairs. Next, the PatchtoPixel [17] method was used to refine the matching granularity to obtain the pixel-level matching results. Finally, the correct correspondence filtering module filters the matches, obtaining more accurate pixel-level matching pairs.

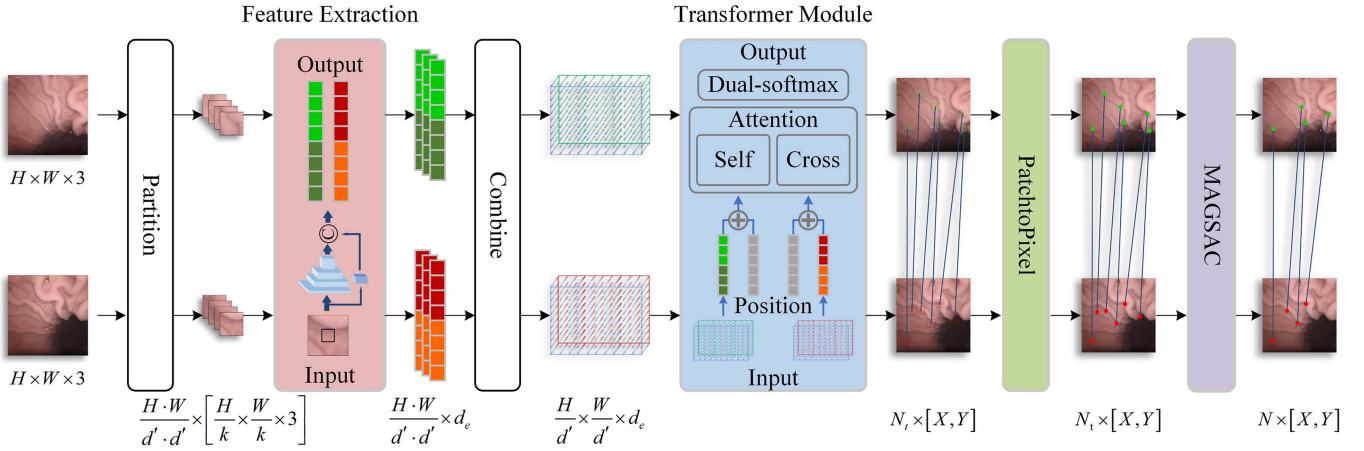


Fig. 3. The process illustration of S2P-Matching. Two images are fed into the network simultaneously and are segmented into $\frac{H}{k} \times \frac{W}{k}$ patches in d' -pixel steps. In this paper, k is finally set to 60. By contrastive learning, each patch is extracted into a $1 \times d_e$ feature vector and combined, involving the initial position. After this, each original image is transformed into a feature map of dimensions $\frac{H}{d'} \times \frac{W}{d'} \times d_e$. Next, the original images' feature encoding and position encoding are united to find the patch correspondence in the Transformer [13] module. The PatchtoPixel [17] module is responsible for carving the patch-level match into pixel-level matching. Finally, MAGSAC [29] filters the pixel matches to obtain pixel-level matching results.

the affine transformation and changing the spatial position of the camera optical axis, we introduce a multi-view camera position transformation method [30], [31] to simulate different viewpoints of the reference images, as shown in Fig. 1. We first fix the camera position and perspective by determining the

angle of view, distance, and rotation angle and estimate an affine matrix. Then, the affine matrix transforms the reference image and obtains a realistic simulation of the endoscopic image.

The key to generating a simulated image is to obtain the affine matrix A , which can be decomposed as $A = H_\lambda R_1(\omega) T_\theta R_2(\varphi)$,

Algorithm 1: S2P-Matching's main learning algorithm.

Input: $I_Patches [\mathcal{H} \times \mathcal{W} \times 3]$, and the initial model weights $(\mathcal{F}_e, \mathcal{T}, \mathcal{P})$.
Output: $\mathcal{F}_e(\cdot) \leftarrow$ Contrastive Learning Encoder;
 $\mathcal{T}(\cdot) \leftarrow$ Transformer_EncoderDecoder (Q, K, V) ;
 $\mathcal{P}(\cdot) \leftarrow$ PatchtoPixel_EncoderDecoder;
1: $\mathcal{I}'_{0,n} = Simulate_{affine}(\mathcal{I}_0)$, n is the number of generated simulated images;
2: $\mathcal{I}_{0,p} = Split_{rule}(\mathcal{I}_0)$, $\mathcal{I}_{i,p} = Split_{rule}(\mathcal{I}'_i)$, $p \in [0, N_d]$; rule: splitting images of size $\mathcal{H} \times \mathcal{W}$ into N_d patches of size $\frac{\mathcal{H}}{k} \times \frac{\mathcal{W}}{k}$ with a step size of d' ;
3: $f_{out_{i,p}} = \mathcal{F}_e(\mathcal{I}_{i,p})$, $i \in \{0, n\}$;
 $\mathcal{F}_e(\cdot) \leftarrow$ Training parameters;
4: **repeat**
5: $\mathcal{F}_c = Stitch_{position}(f_{out_c})$, $c \in \{o, j\}$;
 $\mathcal{M}_t = Compute_{Transformer}(F_o, F_i)$;
 $\mathcal{T}(\cdot) \leftarrow$ Training parameters;
6: Mid-level regressor:
 $\mathcal{M}_{lm} = \mathcal{F}_{mr}(\mathcal{M}_t)$, $\mathcal{M}'_t = (\mathcal{M}_t + \mathcal{M}_{lm})/2$;
Fine-level regressor:
 $\mathcal{M}_{lf} = \mathcal{F}_{fr}(\mathcal{M}'_t)$, $\mathcal{M}_p = (\mathcal{M}'_t + \mathcal{M}_{lf})/2$;
 $\mathcal{P}(\cdot) \leftarrow$ Training parameters;
7: **until** $i = n$
8: **return** Trained models $\mathcal{F}_e(\cdot)$, $\mathcal{T}(\cdot)$, and $\mathcal{P}(\cdot)$.

where T_θ is the transformation parameter. R_1, R_2 are rotations and $H_\lambda > 0$, $\omega \in [0, 2\pi)$, $\theta \in (-\pi/2, \pi/2)$, $\varphi \in [0, \pi)$. H_λ represents the scaling parameter. It can be understood as the distance between the camera position and the part being photographed. ω is the rotation angle around the camera axis, representing the camera's rotation when shooting. θ is the angle between the camera's optical axis and the routine image plane. φ is the angle between the camera's optical axis projection on the image plane and the fixed vertical plane.

To ensure the rotational invariance of the S2P-Matching, the sampling accuracy of the camera location's inclination $\sec \theta$ and angle φ needs to meet certain constraints. Here, multiple experiments on natural images obtain the sampling intervals [32]. When $\sec \theta > 2\sqrt{2}$, the tilt angle has reached 70° , and the simulation image has minimal help for splicing. Therefore, the range of $\sec \theta$ is fixed to $[1, 2\sqrt{2}]$ to reduce computation. Then we can calculate the simulation image I' as:

$$I'_i = \left\{ T_{\theta_i} \frac{\exp\left(-\frac{I_0^2}{2(\sec^2 \theta_i - 1)}\right)}{\sqrt{2\pi(\sec^2 \theta_i - 1)}} \mid \theta_i = \theta_0 + i\Delta\theta, i \in [1, n] \right\}. \quad (1)$$

where I_0 is the reference image and T_θ is the transformation parameter.

The simulated images are reliable by simulating the spatial position of the camera's optical axis. According to the affine matrix, the positions of the matching points in the simulated image corresponding to the reference image can be calculated

and thus get the matching labels. As a result, we get a reliable pseudo-ground-truth dataset for capsule endoscopic image stitching.

B. Deep Feature Descriptor Extraction

We use the patch-level matching method based on deep learning to match capsule endoscopic images with low-texture regions or repetitive patterns. The patch-level matching can also help overcome the annotation from pixel-level errors in the pseudo-ground-truth dataset. Before matching the patches in the capsule endoscopic images, we use contrastive learning to extract the patches' deep feature descriptors. In contrastive learning, we should distance patches' features among the different source images and narrow features of the same source image.

After obtaining the simulated images corresponding to the source capsule endoscopic images by an affine transformation, we split the image of size $H \times W$ into $N_d = \frac{H}{d'} \times \frac{W}{d'}$ patches of size $\frac{H}{k} \times \frac{W}{k}$ with a step size of d' . As shown in Fig. 4, we can use the affine transformation equation to determine the mapping relationship between the patches. Considering the capsule endoscopic image set of the same patient, a reference patch usually corresponds to multiple patches. The matching pair obtained by the affine transformation equation can be determined. We consider the original patches and the pseudo-patches corresponding to the affine transformation as positive sample pairs so that their feature distances are as small as possible. The original patch and the pseudo-patch of another patient are treated as opposite sample pairs to make the feature distance as large as possible. However, if there is no clear affine transformation correspondence with reference patches, we skip processing patches from images of the same patient.

As shown in Fig. 4, the patches of the sample pairs are input into the encoder network (backbone) for feature extraction. Specifically, we modified the encoder to enhance contrast learning. However, the small size of each patch posed a challenge to the conventional encoding approach, which uses a basic convolutional block. We needed to consider the surrounding information while matching. To realize this, we utilized the ViT [33] model as the foundational encoder network. This allowed us to evaluate the problem at multi-scale and obtain f_p by passing a patch through a ViT-based feature encoder. We also obtained f_s by feeding the surrounding patches into a ViT feature encoder. We then combined these features to create a new feature, $f_{out} = concatenate(f_p, f_s)$, which was designed to be used for downstream tasks. Then, the f_{out} are projected into a $1D - d_c$ dimensional potential space through a Multilayer Perceptron (MLP). The projected feature vector f_{cd} calculates the L_2 distance (\mathcal{D}) in (2). The L_2 distances between the deep feature descriptor vector of the reference patch and the other patches are calculated by:

$$\mathcal{D}(i, j) = \frac{e^{\|\Omega(i, j)\|_2}}{\sum_{p \in P(i)} e^{\|\Omega(i, p)\|_2} + \sum_{n \in N(i)} e^{\|\Omega(i, n)\|_2}} \quad (2)$$

where $\Omega(i, j) = Vec(x_i) - Vec(x_j)$, $Vec(\cdot)$ means the deep feature vector obtained from the last layer in the contrastive

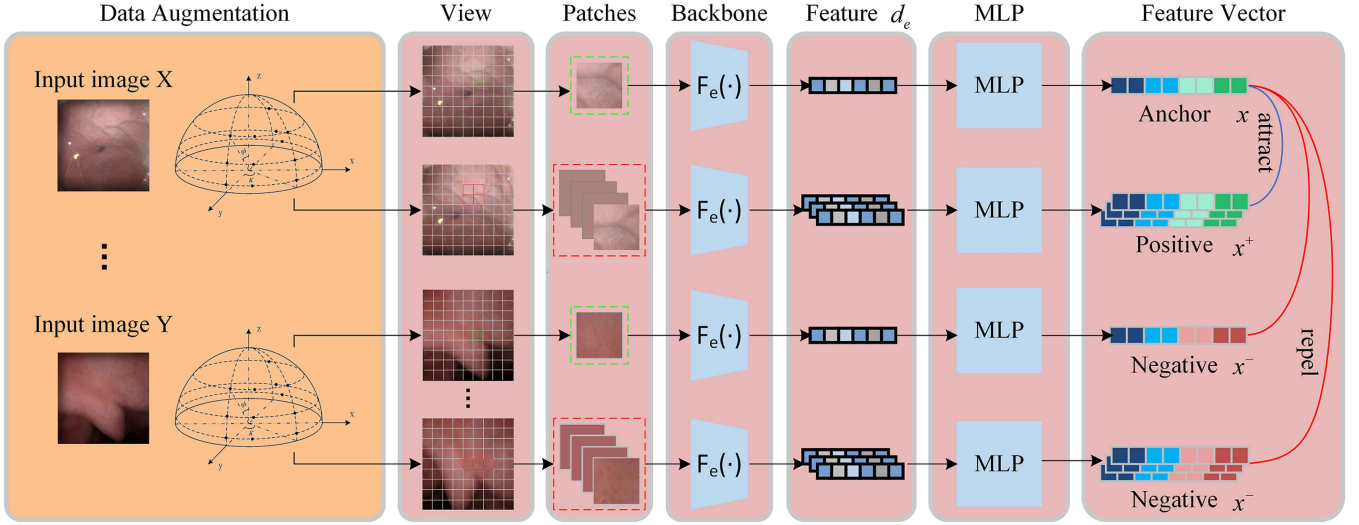


Fig. 4. Model structure of contrastive learning. We augment the images in contrastive learning using the multi-view simulation transformation. The other difference is that we use patches instead of images to extract the features. According to the corresponding relation obtained by the multi-view simulation, the features of patches with an intersection between image patches are drawn close by contrastive learning. The image patches' features determined to be irrelevant are drawn far away.

learning network, i.e., f_{cd} . x_i represents the reference patch at a specific location. x_j represents the patch in a different position. $P(i)$ and $N(i)$ are the sets of positive and negative samples of i respectively.

Our contrastive learning is trained on the pseudo-ground-truth dataset, where multiple positive and negative sample pairs exist for a reference patch, and is supervised [34]. Therefore, the loss function \mathcal{L}_f updates the network parameter in (3).

$$\mathcal{L}_f = - \sum_{i \in B} \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\mathcal{D}_{(i,p)}/\tau}{\sum_{n \in N(i)} \mathcal{D}_{(i,n)}/\tau} \quad (3)$$

where B is the batch data set, $|P(i)|$ is the cardinality of $P(i)$, and τ is the scalar temperature parameter.

The contrastive learning agent task is to close the feature distance between the reference patch and the corresponding pseudo-patch and pull apart the distance between the original and irrelevant patches. It is consistent with our behavior of seeking matches according to the patch features' similarity. Therefore, we can use contrastive learning to extract the most appropriate features for subsequent patch matching.

C. Patch-Level Matching Via Transformer

After obtaining deep feature descriptors of the patches in the capsule endoscopic images, we proceed with patch-level matching by finding the correspondence of the patches. Although medical images always use the traditional matching method to find nearest neighbor matching pairs, it does not match weakly textured regions in the capsule endoscopic images [35]. We use the Transformer [36], built on multi-head self-attention (MHSA), to expand the receptive field and find accurate correspondences in obscure regions. As given in (4) and (5), the MHSA [37] consists of multiple self-attention layers and plays an increasingly important role in all areas of deep learning due

to its large receptive field.

$$\begin{aligned} SA^{(i)}(I) &= Softmax \left(\tau IW_Q^{(i)} \left(IW_K^{(i)} \right)^T \right) IW_V^{(i)} \\ &= Softmax \left(\tau Q^{(i)} K^{(i)T} \right) V^{(i)} \end{aligned} \quad (4)$$

$$MHSA(I) = Concat_{i \in [n_i]} [SA^{(i)}(I)] W_p + \beta_p \quad (5)$$

where I is the input elements, (i) is the head index, W is the weight matrix, and τ is a scaling parameter. The MHSA layer with n_i heads aggregates the self-attention outputs with $W_p \in R^{n_i \mathcal{D}_v \times \mathcal{D}_{out}}$ and $\beta_p \in R^{\mathcal{D}_{out}}$.

We construct a Transformer model similar to the LoFTR [13]. The differences exist in the ground-truth dataset that guides the Transformer matching. The LoFTR adopts the labeled ground-truth dataset from the real world, while we use the pseudo-ground-truth dataset. Moreover, the LoFTR uses coarse-grained feature maps of the original image, simply extracted by a local feature CNN. However, our contrastive learning mentioned above can naturally extract the most appropriate features for our patch-level matching. The upper half of every feature is derived from a specific patch, while the lower half is obtained from the corresponding region where the patch is situated. This combination of features has been designed to facilitate feature matching for upcoming transformers. Then, the deep feature descriptors of the patches are stitched according to the image position to obtain a feature map for matching. Using the feature map, our Transformer model can support the discovery of the similarity between the patch features of two capsule endoscopic images.

Specifically, after cropping the original $H \times W$ image in d' -pixel steps to obtain small patches of size $\frac{H}{k} \times \frac{W}{k}$, we convert each patch into a d_e -dimensional feature. For each $H \times W$ capsule endoscopic image, it becomes a combined feature map

of $\frac{H}{d'} \times \frac{W}{d'} \times d_e$, also known as F_o and $F_{i,i \in (1,n)}$ in Fig. 2. F_o is the combined feature map of our reference capsule endoscopic images, and F_i is the combined feature map of one of the $n - 1$ feature images of I_o after an affine transformation. Using the affine transformation matrix, we can know the position relationship between homologous image patches. This patch-level position correspondence constitutes our pseudo-ground-truth sample dataset for training.

1) Positional Encoding: In our Transformer model for patch-level matching, the standard position encoding must also be input to the attention module simultaneously. The patch alone does not contain position information. However, its position in the combined feature map is known. We can add the position information to the feature by positional encoding and make the feature position-dependent. Taking a cue from LoFTR [13], We add the standard positional encoding to the backbone output using a 2D extension of the standard positional encoding in Transformer. Intuitively, the position encoding gives unique positional information to each element in the sine wave format. Adding position encoding to f_o and f_i makes the transformed features f'_o and f'_i position-dependent, which is critical to S2P-Matching's ability to produce matches in indistinctive regions.

2) Attention: Self and Cross: The self-attentive layer (shown in Fig. 2) focuses on the relationships within each input feature (either f'_o or f'_i). The cross-attention layer focuses on the relationships between different features (f'_o and f'_i). Following [19], we interleave the self-attention and cross-attention layers in the S2P-Matching module by R_N times. f'_o and f'_i are transformed into F'_o and F'_i after the attention module.

3) Establishing Patch-Level Matches: In S2P-Matching, we use the dual-softmax operator as the matching generation layer. After obtaining the transformed features, the score matrix S_m between the features is calculated as $S_m(o, i) = \frac{1}{\tau} \cdot \langle F'_o, F'_i \rangle$. We apply softmax to the two dimensions of S_m to obtain the probability of soft mutual nearest neighbor matching. The matching probability matrix P_m can be calculated as follows:

$$P_m(o, i) = \text{Softmax}(S_m(o, \cdot))_i \cdot \text{Softmax}(S_m(\cdot, i))_o. \quad (6)$$

4) Match Selection: Based on the matching probability matrix P_m , also called the confidence matrix, we can easily select matches with confidence higher than the threshold t_c . Here, we use the mutual nearest neighbor (M_{NN}) [42] criterion to filter possible outlier coarse matches. The patch-level match predictions are denoted as:

$$M_t = \{(o, i) | \forall (o, i) \in M_{NN}(P_m), P_m(o, i) \geq t_c\}, \quad (7)$$

D. Refining Patch-Level Matching to Pixel-Level Matching

In patch-level matching, the Transformer finds the most matched patches with greater confidence in the whole image according to the global information. However, since the patches are obtained by segmenting in d' -pixel point steps, patch-level

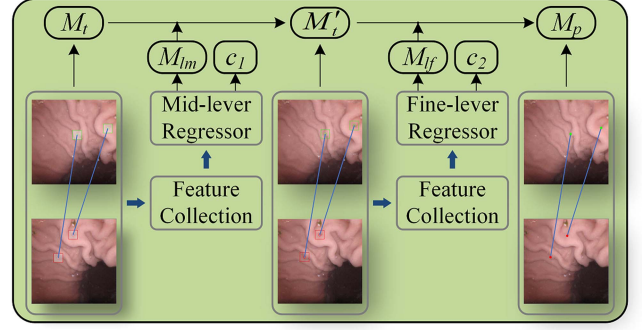


Fig. 5. The structure of PatchtoPixel. Depth feature descriptors extracted by contrastive learning are used as local features for pixel-level matching. Two regressors perform match generation with the same structure, i.e., a mid-level regressor and a fine-level regressor. The generated matches update the match set according to the confidence score and finally obtain pixel-level matches.

matching can only guarantee the similarity of a specific range of images. It cannot ensure a high degree of matching between the centers of patches. Therefore, it is necessary to obtain pixel-level matching using a refinement of patch-level matching. We use Patch2Pix to search for pixel-level matches in the patch-level match spaces. That is, refine the patch-level matches $M_t = (p_i^A, p_i^B)_{i=1}^{N_t} = (x_i^A, y_i^A, x_i^B, y_i^B)_{i=1}^{N_t} \in R^4$ to pixel-level matches $M_p = (p_i^A, p_i^B)_{i=1}^{N_t}$.

For patch-level matching $M_t(i) = (x_i^A, y_i^A, x_i^B, y_i^B)$, the patch block size is $\frac{H}{k} \times \frac{W}{k}$. To improve the matching accuracy, the local search spaces are set to $(x_i^A \pm \Delta d, y_i^A \pm \Delta d)$ and $(x_i^B \pm \Delta d, y_i^B \pm \Delta d)$. Pixel-level matching expands the range by d pixels with the patch as the center. The expanded search space covers the original space, which can be fault-tolerant for patch matching to some extent.

We use the deep feature descriptors extracted by contrastive learning as local features for pixel-level matching to reduce the computational overhead while achieving a better matching result. As shown in Fig. 5, we have used two regressors with the same structure for match generation, i.e., a mid-level regressor and a fine-level regressor. In the deep feature descriptor extraction module, we have obtained the features F_o and F_c of patches p_i^A and p_i^B . After collecting the features, the channels are connected into a vector F_{ci} . The vector F_{ci} is entered into the mid-level regressor. The regressor is a double-headed network. The two-headed network generates the matching pairs while the matches are evaluated for classification. In the regressor, after processing the features through a fully connected network, in the match generation head, the local match $M_{lm}(i) = (x_i^A, y_i^A, x_i^B, y_i^B)$ is output. In the classification header, a confidence score s_i is obtained for the production using sigmoid, indicating the degree of detection validity. We use the pseudo-correspondence obtained during affine transformation as labels to guide network training. The geometric error between the predicted and actual matches is measured by calculating the

Sampson distance, as shown in (8).

$$\Psi_i^{Sam} = \frac{\left(\left(p_i'^B \right)^T F p_i'^A \right)^2}{\sum_{v \in \mathbb{V}} \left[\left(F p_i'^A v \right)^2 + \left(F^T p_i'^B v \right)^2 \right]} \mathbb{V} = \{\varsigma_1, \varsigma_2\}, \quad (8)$$

where F is the image relative pose matrix, $p' = (x, y, 1)^T$ and ς_i represents the one-hot vector with the i -th value is 1.

The Sampson distance calculates the error degree of the predicted match. So we can obtain the classification expectation label of match accordingly, i.e., the error is less than the threshold value t_m to consider the classification as correct. Thus the classification labels l_i can be obtained. If $\Psi_i^{Sam} < t_m$, $l_i = 1$, otherwise $l_i = 0$. Using the binary cross-entropy loss function, the network parameters are corrected as shown in (9).

$$\mathcal{L}_r(l, s) = -\frac{1}{N_t} \sum_{i=1}^{N_t} (\beta l_i \log s_i + (1 - l_i) \log (1 - s_i)), \quad (9)$$

where β is the weight that balances the quantities of the two classes, and s_i is predicted confidence score.

The generated mid-level matches M_{lm} update the match set M_t according to the confidence score, and the exact refinement matching is performed once more to obtain the final match result M_p .

E. Correct Correspondence Filtering

After refining the matches at the patch level, we obtain the initial pixel-level matching pairs. However, due to the fragmentation and discrimination of capsule images, these matching pairs will inevitably be mixed with impurity matches, mismatches, or poorly matched pairs. They will lead to bad or even wrong stitching of the final image. Therefore, the matching results must be filtered to select the correct matches.

In pixel-level matching of capsule endoscopic images, two types of matching pairs significantly impact the matching results. One is the impurity matching pairs. Impurity pixel points are usually found as matching pairs because of their particular presentation. However, when the impurities stick to the lens, the impurity pixels will always appear in the image's fixed position. Stitching according to this will result in significant deviations. The other one is the hyper-boundary matching pairs. As shown in Fig. 6, S2P-Matching always matches the results by aggregating images taken with simulated multi-perspective cameras. In the back projection of the reference image, some matching points will fall back outside the image boundary. These matching pairs need to be filtered out.

After removing impurity and hyper-boundary matching pairs, we use the MAGSAC algorithm [29] to screen for poorly matched pairs. When MAGSAC fits matching pixels, it makes continuous attempts on different target space parameters. Increasing the number of iterations to determine the model parameters most matches meet improves the accuracy rate. At this point, matching pixels that satisfy the model are inliers, while those

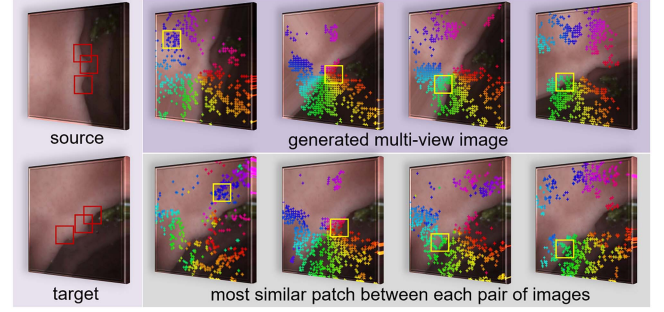


Fig. 6. The comprehensive correlation maps between matched image pairs. The first row is the source image and the generated four multi-view camera simulation images. The second row is the target image and its generated dense corresponding matching results. We mark the matching points in the image pair with the same color.

that do not satisfy are outliers. The threshold σ of the interior point is regarded as a random variable, and by marginalizing σ , the probability that a point is an internal point can be calculated. We treat the possibility that each point is an inlier as a weight for each point and use a weighted least squares fit to optimize the model according to the weight. The number of iterations is calculated as follows:

$$k_{ni}(\epsilon, \alpha) = \frac{1}{\sigma_{\max}} \sum_{i=1}^n \frac{(\sigma_i - \sigma_{i-1}) \ln(1 - \eta)}{\ln(1 - (|I(\alpha, \sigma_i, \epsilon)| / |\epsilon|)^t)} \quad (10)$$

where α is the homography matrix parameter, ϵ is the set of matching points, σ_{\max} is the maximum inlier threshold, n is the number of points whose distance to the model is less than σ_{\max} , σ_i is the interior threshold of the i th point, η is a manually set confidence in the results, t the size of the minimal sample needed for the estimation, and $|I(\alpha, \sigma_i, \epsilon)|$ is the inlier number of the model.

IV. EXPERIMENTS AND RESULTS

Our research primarily focuses on analyzing a series of consecutive frames captured by healthcare professionals in the clinical setting around the ROI. To evaluate the efficacy of S2P-Matching, we have opted to utilize capsule endoscopic images acquired continuously during periods of relative stability as the test dataset. This deliberate selection ensures that the compiled images incorporate overlapping areas, enabling us to assess the accuracy of our S2P-Matching algorithm. In this case, the shooting interval is 0.5 seconds, and the spatial resolution of each image is 480×480 pixels. Accordingly, we conduct statistical experiments to compare our method against state-of-the-art methods, CAPS [38], ASIFT [39], DeepMatching [35], R2D2 [40], SuperPoint [41], SuperGlue [19], LoFTR [13] and TransforMatcher [37]. The experimental hardware environments include Intel(R) Xeon(R) E5-2680 v4 CPU@ 2.40GHz, NVIDIA TESLA-P100 GPU, 16GB memory, and 3TB hard disk. Software environments include Ubuntu14.04 and Python 3.6. Here, Python is the development language to contain the PyTorch and OpenCV as the runtime library.

TABLE I

MATCHING ACCURACY OF DIFFERENT DATA SETS. OUR APPROACH PERFORMS BEST WHEN DEALING WITH WEAK TEXTURE, CLOSE-UP TRANSFORMATION, AND LARGE ANGLE ROTATION PROBLEMS. ACCORDING TO THE AVERAGE CALCULATION, OUR METHOD'S IMPROVEMENT OF NCM AND SR SCORES CAN REACH UP TO 203 AND 55.8%

method	weak texture		close-up transformation		large angle rotation		Average	
	NCM	SR%	NCM	SR%	NCM	SR%	NCM	SR%
CAPS [38]	104	32.5	107	24.3	172	20.9	128	25.9
ASIFT [39]	40	58.2	149	69.8	134	65.1	108	64.4
DeepMatching [35]	133	66.5	284	74.3	221	70.0	256	70.3
R2D2 [40]	250	34.8	169	65.1	278	71.2	232	57.0
SuperPoint [41]	192	45.0	233	65.7	302	43.9	242	51.5
SuperGlue [19]	275	72.8	213	66.4	217	74.3	235	71.2
LoFTR [13]	258	79.3	314	73.2	306	82.9	293	78.5
TransforMatcher [37]	271	81.3	335	74.9	317	82.9	308	79.7
S2P-Matching (Ours)	262	83.6	348	76.3	323	85.1	311	81.7

A. Datasets and Evaluation Metrics

Our datasets comprise records from capsule endoscopy examinations conducted at a domestic hospital between 2016 and 2019. For the purpose of facilitating randomization and achieving optimal stitching outcomes for comparative analysis, a sample of 213 patients was selected. Subsequently, $n \times 10$ consecutive image frames were extracted for each patient, with n ranging from 5 to 15, resulting in a total of 21,526 images. After filtering, we obtained 20,862 images. We then separated images from 20 patients for testing and used the remaining for training. We selected 528 images as the test set with matching points annotated by two collaborating physicians. The images were then classified into three categories: weak texture, close-up transformation, and large angle rotation, with 138, 204, and 186 images, respectively. Additionally, we chose two consecutive 10-frame sequences for testing purposes.

To achieve a more precise quantitative evaluation, obtaining a ground-truth geometric transformation for each pair of images is necessary. Unfortunately, real datasets are often subject to various factors that interfere with acquiring an actual ground-truth geometric transformation. Therefore, we follow RIFT [43] and utilize an approximate ground-truth geometric transformation for evaluation purposes. The collaborating doctors manually select five uniformly distributed correspondences for each image pair, allowing for the estimation of an accurate affine transformation that closely approximates the ground-truth geometric transformation.

After obtaining the correct matching points, we employ a traditional image fusion technique that utilizes perspective transformation [44] to produce image mosaics. Our approach maintains the experimental parameters throughout the mosaicking phase to guarantee conformity with established formats and to facilitate an impartial evaluation.

To assess the performance of various image-matching algorithms, we employ two standard metrics: the number of correct matches (NCM) and the success rate (SR). A superior matching method is characterized by higher NCM and SR scores. Additionally, to further appraise the quality of the matching points acquired by different algorithms for the final image stitching process, we employ SSIM (Structural Similarity) [45], FSIM (Feature Similarity) [46], and PSNR (Peak Signal to Noise Ratio) [47] as evaluation metrics for image stitching.

B. Comparison of Feature Point Matching

Table I summarizes the matching accuracy of different methods on three data types (i.e., weak texture, close-up transformation, and large angle rotation). The S2P-Matching method has the highest average NCM and SR scores of 311 and 81.7%, respectively.

To comprehensively evaluate our S2P-Matching approach concerning capsule endoscopy images with impurities and low-texture regions and its ability to handle transformations and rotations common in close-up photography, we have selected three sets of images from distinct datasets. Featured in Fig. 7 is a visual comparison of the matching results yielded by our method against those obtained through state-of-the-art techniques. Each pair of input images comprises two capsule endoscopic images taken at 0.5-second intervals. Notably, all three image pairs were taken in close proximity, with rotational variance. We have employed a white line to indicate corresponding pairs, clearly depicting the matching results.

From Fig. 7, it can be seen that from the first row to the third row, the texture becomes weaker, the degree of region repetition becomes higher, and the matching pairs matched by each method are reduced to different degrees. For example, CAPS [38] and ASIFT [39] can only extract a small number of matching pairs, and there are incorrect matching pairs that result in the final image splicing error. In this case, the DeepMatching [35] algorithm can also extract only a limited number of matching pairs. R2D2 [40] and SuperPoint [41], the number of matches is large, but more inaccurate matches are counterproductive to mosaic. In regions of the graph with transformations caused by close-up shooting, such as the third row, SuperGlue [19], LoFTR [13], and TransforMatcher [37] are difficult to function, with fewer correct matches. Compared with other methods, our S2P-Matching matching results achieve the best feature-matching performance, see the last column of Fig. 7. S2P-Matching can extract a sufficient number of significant matching pairs without interference from impurities and obvious transformation, which guarantees the final stitching.

C. Comparison of Consecutive Frames Stitching

In clinical practice, capsule endoscopy is constrained by its limited capture area per image. This presents a challenge for

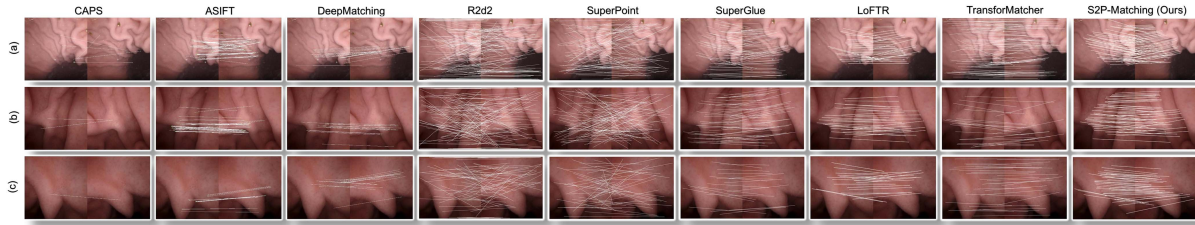


Fig. 7. Visual comparisons of image matching results. From left to right: CAPS [38], ASIFT [39], DeepMatching [35], R2d2 [40], SuperPoint [41], SuperGlue [19], LoFTR [13] and TransforMatcher [37] and S2P-Matching (our method). From top to bottom: (a) two continuous images with solid texture areas, impurities, rotation and transformation caused by close-up shooting. (b) continuous images with solid and weak texture areas, impurities, rotation and transformation due to close-up shooting. (c) two continuous images without noticeable changes in the overall texture and with rotation or transformation by close-up shooting. We use the white line to mark the matching pairs.

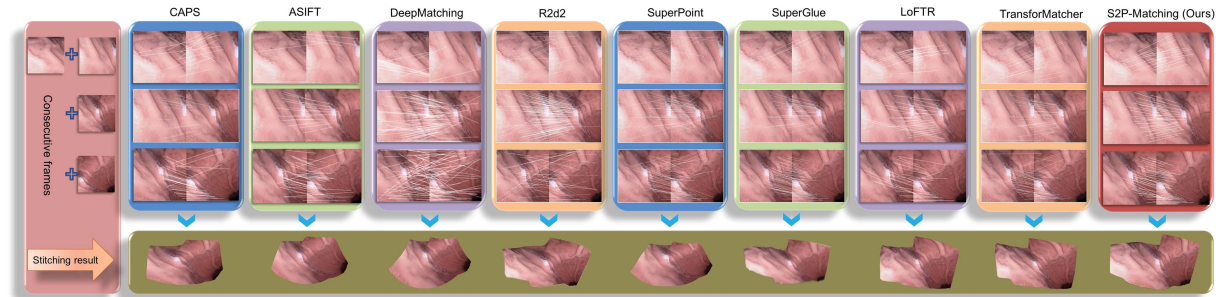


Fig. 8. Visual comparisons of image stitching results of continuous frames between our S2P-Matching and other methods. In the S2P-Matching-based stitching, the connection between two original images is natural because of the massive matching numbers and high accuracy. Furthermore, there is no obvious texture misplacement or excessive scaling texture connection. In the results of other methods, some methods fail to find enough accurate matches in the matching of the first two weakly textured images to cause splicing difficulties.

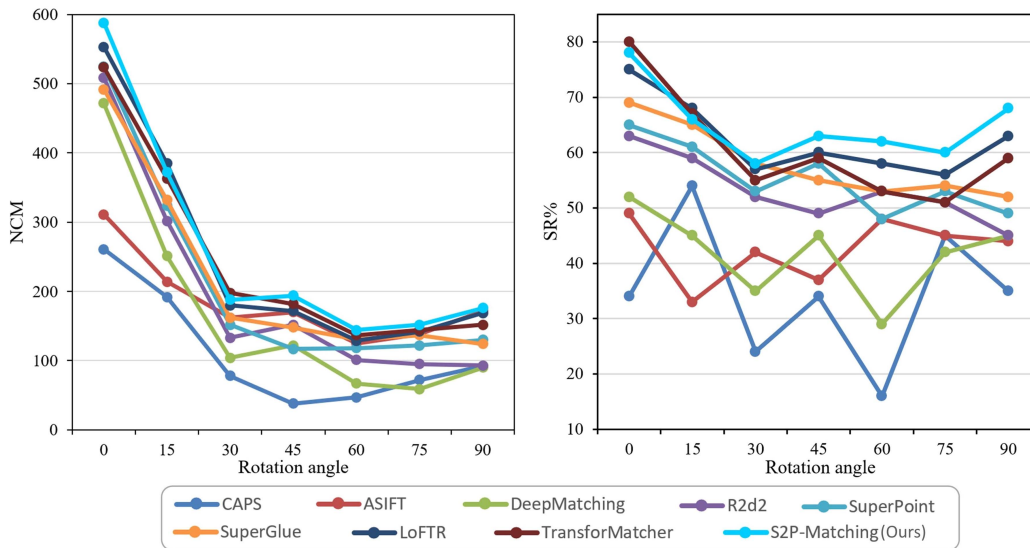


Fig. 9. Comparing NCM and SR scores of our S2P-Matching with CAPS [38], ASIFT [39], DeepMatching [35], R2d2 [40], SuperPoint [41], SuperGlue [19], LoFTR [13] and TransforMatcher [37] under different rotation angles.

physicians in observing ROI within a relatively wide field of view, thereby impacting diagnosis. Typically, an ROI region involves multiple sequential images with overlapping areas. As such, the continuous stitching of capsule endoscopy frames is crucial. Continuous image frames usually contain complex situations with multiple matching challenges, including weak

texture, transformation caused by close-up shooting, and large angle rotation. From the experimental results, it can be seen that our method has obvious advantages over other methods for the stitching of endoscopic continuous frame images. As shown in Fig. 8, the stitching effect of our method is more accurate and natural.

TABLE II

MATCHING RESULTS OF 10 CONSECUTIVE FRAMES REGARDING NCM AND SR FOR IMAGE MATCHING AND SSIM, FSIM, AND PSNR FOR IMAGE STITCHING

Method	dataset1					dataset2				
	image matching		image stitching			image matching		image stitching		
	NCM	SR%	SSIM	FSIM	PSNR	NCM	SR%	SSIM	FSIM	PSNR
CAPS [38]	275	34.6	0.529	0.292	10.080	105	34.2	0.508	0.452	10.026
ASIFT [39]	463	52.7	0.677	0.500	16.900	56	26.9	0.423	0.314	10.502
DeepMatching [35]	368	64.5	0.698	0.505	15.648	419	65.3	0.620	0.436	11.462
R2D2 [40]	376	59.1	0.609	0.391	13.580	341	67.3	0.575	0.345	11.251
SuperPoint [41]	368	69.9	0.561	0.434	14.280	345	62.4	0.569	0.292	10.080
SuperGlue [19]	425	73.1	0.628	0.436	14.967	434	70.8	0.550	0.261	10.160
LoFTR [13]	493	73.9	0.683	0.492	16.462	452	66.5	0.677	0.653	12.934
TransforMatcher [37]	511	74.2	0.691	0.505	16.836	449	65.7	0.653	0.632	12.131
S2P-Matching (Ours)	531	75.4	0.741	0.533	17.103	463	66.9	0.709	0.687	13.462

TABLE III

COMPARING MATCHING THE PERFORMANCE OF THE MODEL IN AN ABLATION STUDY UNDER TWO CONSECUTIVE CAPSULE ENDOSCOPIC IMAGE SEQUENCES. THE BASELINE IS THE DENSE CORRESPONDENCE MATCHING FRAMEWORK FOR A SINGLE SET OF IMAGE PAIRS USING GRADIENT FEATURE DESCRIPTORS. "IDA" DENOTES IMAGE DATA AUGMENTATION, WHILE "DFD" REPRESENTS DEEP FEATURE DESCRIPTORS

Methods	IDA	DFD	dataset1		dataset2	
			NCM	SR%	NCM	SR%
basic (dense correspondence matching framework using gradient feature descriptors)	×	×	469	61.3	442	63.1
basic + image data augmentation	✓	×	496	70.9	477	69.4
basic + deep feature descriptors	×	✓	482	70.2	459	65.8
basic + image data augmentation + deep feature descriptors (Our S2P-Matching)	✓	✓	528	73.9	493	71.1

In addition, two sets of capsule endoscopic images were selected for assessment, each containing ten image frames captured in a relatively stable shooting process. Table II presents the averaged NCM and SR scores and the SSIM, FSIM, and PSNR scores to assess image stitching accuracy. The results indicated that our S2P-Matching method resulted in the highest number of matching pairs. CAPS and ASIFT yielded the lowest number of matching pairs, particularly in dataset 2, posing challenges for subsequent image stitching. Furthermore, compared to DeepMatching, R2D2, SuperPoint, and SuperGlue, the overall performance of LoFTR and TransforMatcher was superior, approaching the level of S2P-Matching, as evidenced in Fig. 8. SuperGlue also demonstrated commendable performance. Our S2P-Matching method consistently delivered superior results in complex capsule endoscope stitching tasks.

D. Ablation Study Experiments

We further conduct ablation study experiments to evaluate the effectiveness of significant modules of our S2P-Matching by implementing different image-matching frameworks based on our S2P-Matching. To do so, we conduct image-matching experiments on two datasets of capsule endoscopic image data taken consecutively over a period of time. Table III summarizes the image-matching results of our S2P-Matching and constructed baseline methods by showing the average NCM and SR performance on two long capsule endoscopic image sequences.

From the quantitative results, we can find that "basic+IDA" has more significant NCM and SR scores on two datasets than "basic", which indicates that the image-matching effect of the

method is slightly improved after using the image derivation module simulated by the data augmentation module. This stems from the model's capacity to accommodate the effects of rotation and transformation resulting from close-up shooting after combining images generated from different angles with multiple virtual cameras. Then, "basic + DFD" performs better than "basic" regarding NSM and SR scores on dataset1 and dataset2. It shows that the image-matching results in the weak texture area are also improved when we replace the CNN descriptor with the deep feature descriptor. More importantly, our method's superior NCM and SR scores over "basic + IDA" and "basic + DFD" demonstrate that combining the image derivation and deep feature descriptors in our S2P-Matching framework can further improve the image matching accuracy. This is because our complete S2P-Matching framework fully considers various difficulties that may be encountered in capsule endoscopic image matching.

E. Rotation Invariance Analysis

We test the effect of rotation angle on the matching results to analyze the method's adaptability to different rotation angles of close-up shooting. This experiment selects seven sets of capsule endoscopic image pairs with different rotation angles ranging from small to large and close-up transformation, and various methods are used to match them.

The results in Fig. 9 showcase the variability of NCM and SR scores among the different methods based on the rotation angle. The comparison of results indicates that the S2P-Matching method demonstrates superior accuracy when applied to capsule endoscopy images captured at various rotation angles.

V. CONCLUSION

We have outline a self-supervised patch-based matching (S2P-Matching) for matching capsule endoscopy images. Unlike the existing image matching methods, our S2P-Matching performs joint affine and patch-level matching via the Transformer. Experimental findings with real-world MCCE images have demonstrated the efficacy of S2P-Matching in enhancing accuracy, particularly in addressing challenges in GI scenarios around ROI. Our next research phase aims to investigate more complex real-life scenarios, including illumination variations, presence of bubbles, defocus, uninformative and motion blur, occlusion, and reflected images. The algorithm is being adapted to encompass capsule endoscopy images throughout the entire gastrointestinal tract.

REFERENCES

- [1] B. Sushma and P. Aparna, "Recent developments in wireless capsule endoscopy imaging: Compression and summarization techniques," *Comput. Biol. Med.*, vol. 149, 2022, Art. no. 106087.
- [2] S. Kadian et al., "Smart capsule for targeted detection of inflammation levels inside the GI tract," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 5, pp. 1565–1576, May 2024.
- [3] J.-H. Kim and S.-J. Nam, "Capsule endoscopy for gastric evaluation," *Diagnostics*, vol. 11, no. 10, 2021, Art. no. 1792.
- [4] M. Szalai et al., "First prospective European study for the feasibility and safety of magnetically controlled capsule endoscopy in gastric mucosal abnormalities," *World J. Gastroenterol.*, vol. 28, no. 20, 2022, Art. no. 2227.
- [5] X. Wang et al., "A systematic review on diagnosis and treatment of gastrointestinal diseases by magnetically controlled capsule endoscopy and artificial intelligence," *Ther. Adv. Gastroenterol.*, vol. 16, 2023, Art. no. 17562848231206991.
- [6] B. Akpunonu et al., "Capsule endoscopy in gastrointestinal disease: Evaluation, diagnosis, and treatment," *Clev. Clin. J. Med.*, vol. 89, no. 4, pp. 200–211, 2022.
- [7] L. Lavenir et al., "Miniaturized endoscopic 2D US transducer for volumetric ultrasound imaging of the auditory system," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 9, pp. 2624–2635, Sep. 2023.
- [8] T. Rahim et al., "A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging," *Comput. Med. Imag. Graph.*, vol. 85, 2020, Art. no. 101767.
- [9] C. Xie et al., "Endoscope localization and gastrointestinal feature map construction based on monocular slam technology," *J. Infect. Public Heal.*, vol. 13, no. 9, pp. 1314–1321, 2020.
- [10] Y. Liu et al., "Improved feature point pair purification algorithm based on SIFT during endoscope image stitching," *Front. Neurobot.*, vol. 16, 2022, Art. no. 840594.
- [11] Z. Zhang et al., "Endoscope image mosaic based on pyramid ORB," *Biomed. Signal Process.*, vol. 71, 2022, Art. no. 103261.
- [12] J. Fan et al., "Deep feature descriptor based hierarchical dense matching for X-ray angiographic images," *Comput. Methods Programs Biomed.*, vol. 175, pp. 233–242, 2019.
- [13] J. Sun et al., "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8922–8931.
- [14] A. Dai et al., "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5828–5839.
- [15] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from internet photos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2041–2050.
- [16] L. Barducci et al., "Fundamentals of the gut for capsule engineers," *Prog. Biomed. Eng.*, vol. 2, 2020, Art. no. 042002.
- [17] X. Chen et al., "Tabletop transparent scene reconstruction via epipolar-guided optical flow with monocular depth completion prior," in *2023 IEEE-RAS 22nd Int. Conf. Humanoid Robots (Humanoids)*, 2023, pp. 1–8.
- [18] R. Zhang et al., "KdO-Net: Towards improving the efficiency of deep convolutional neural networks applied in the 3 D pairwise point feature matching," *Remote Sens.*, vol. 14, no. 12, 2022, Art. no. 2883.
- [19] P.-E. Sarlin et al., "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4938–4947.
- [20] S. L. Lee et al., "Late fusion of deep learning and handcrafted visual features for biomedical image modality classification," *IET Image Process.*, vol. 13, no. 2, pp. 382–391, 2019.
- [21] G. Wu et al., "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1505–1516, 2016.
- [22] C. Liao et al., "Deep learning for registration of region of interest in consecutive wireless capsule endoscopy frames," *Comput. Methods Programs Biomed.*, vol. 208, 2021, Art. no. 106189.
- [23] M. Farhat et al., "Self-supervised endoscopic image key-points matching," *Expert Syst. Appl.*, vol. 213, no. part, 2023, Art. no. 118696.
- [24] P. Yan et al., "Repeatable adaptive keypoint detection via self-supervised learning," *Sci. China Inf. Sci.*, vol. 65, no. 11, pp. 1–25, 2022.
- [25] T. Chen et al., "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
- [26] K. He et al., "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
- [27] T. Simon et al., "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4645–4653.
- [28] M. Meng et al., "Non-iterative coarse-to-fine transformer networks for joint affine and deformable image registration," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2023, pp. 750–760.
- [29] U. Aulia et al., "Visual place recognition for autonomous mobile robot navigation using LoFTR and MAGSAC++," *J. Polimesin*, vol. 22, no. 2, pp. 272–277, 2024.
- [30] X. Wang et al., "An ASIFT-based local registration method for satellite imagery," *Remote Sens.*, vol. 7, no. 6, pp. 7044–7061, 2015.
- [31] H. Zhou and J. Jayender, "Real-time nonrigid mosaicking of laparoscopy images," *IEEE Trans. Med. Imag.*, vol. 40, no. 6, pp. 1726–1736, Jun. 2021.
- [32] X. Ma et al., "A fast affine-invariant features for image stitching under large viewpoint changes," *Neurocomputing*, vol. 151, pp. 1430–1438, 2015.
- [33] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–8.
- [34] P. Khosla et al., "Supervised contrastive learning," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 18661–18673, 2020.
- [35] J. Revaud et al., "Deepmatching: Hierarchical deformable dense matching," *Int. J. Comput. Vis.*, vol. 120, no. 3, pp. 300–323, 2016.
- [36] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [37] S. Kim et al., "TransforMatcher: Match-to-match attention for semantic correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8697–8707.
- [38] Z. Yang et al., "Learning feature descriptors for pre- and intra-operative point cloud matching for laparoscopic liver registration," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 18, no. 6, pp. 1025–1032, 2022.
- [39] J.-M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," *SIAM J. Imag. Sci.*, vol. 2, pp. 438–469, 2009.
- [40] J. Revaud et al., "R2D2: Reliable and repeatable detector and descriptor," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 12405–12415.
- [41] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 224–236.
- [42] L. Haghverdi et al., "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors," *Nat. Biotechnol.*, vol. 36, no. 5, pp. 421–427, 2018.
- [43] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020.
- [44] V. Kwatra et al., "Graphcut textures: Image and video synthesis using graph cuts," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 277–286, 2003.
- [45] Z. Wang et al., "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [46] L. Zhang et al., "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
- [47] U. Sara et al., "Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study," *J. Comput. Commun.*, vol. 7, no. 3, pp. 8–18, 2019.