www.cad-journal.net
Computer-Aided Design AND APPLICATIONS

Multimodal Medical Image Registration Algorithm Based on DETR Model

Jianfeng Han, Jingxuan Zhao, Renjie Li and Yong Zhang

School of Information Engineering, Tianjin University of Commerce, Tianjin 300134, China,
hanjianfeng@toec-gdgs.com, 1667211650@qq.com, 18326653378@163.com,
zhangyong@tjcu.edu.cn

Corresponding author: Yong Zhang, zhangyong@tjcu.edu.cn

Abstract. Aimed at the problems of feature extraction with lots of outliers, unevenly distributed feature vectors, low accurate registration, and long processing time caused by the ultra-high image resolution, repeated image textures, and different tissue deformations in the medical image registration, a multimodal medical image registration algorithm based on the DETR model was proposed in this article. Firstly, with the demands of image registration, the two images feature extraction network has to be constructed based on DETR model to complete the image pairs information interaction. In this way, the feature information is estimated to have a wider receptive field and then used as the input vectors for the predictor. Secondly, the prediction network of feature points' coordinates is built, and these coordinates, as the basic information, should participate in the encoder-decoder of image features in the neural network so as to achieve the combination of these coordinates and the positional information of image blocks and to promote the estimation precision of key points' positions. Finally, the simulation experiments on the proposed model were conducted on the ANHIR medical image dataset and the FIR retinal image dataset. As for the experimental results, this proposed model, which has good properties, can provide a feasible and effective solution to medical image registration.

Keywords: Registration of medical images; Multi-Mode; DETR model.
DOI: https://doi.org/10.14733/cadaps.2024.S25.294-310

1 INTRODUCTION

At present, medical images can be formed by many methods, such as magnetic resonance imaging (MRI), X-rays, ultrasound, computed tomography (CT), etc., all of which provide extensive information related to the body structure, anatomy, and pathology from different perspectives [1,2]. During the medical diagnosis, it is often necessary to compare two images in different modes for monitoring diseases, analyzing lesions conditions, and evaluating the effects of operations. Due to the differences in instrument parameters and image processing algorithms in diverse imaging methods and the non-linear variations in the image collection of human tissues during acquisition, there may be deviations in the spatial position of images. Medical images of different modalities usually do not correspond directly to each other in spatial position, making it difficult to directly compare images. To solve this problem, image registration is usually required.

Image registration is a process based on similarities of images that seek spatial alternation by reflecting the reference image into the target image so that the two images can correspond in space [3,4]. With the correspondence, the multimodal images collected at diverse times, from different angles and by various sensors, can fuse into a new image. The new images have their own unique information from the original two images, which can provide great help for subsequent information processing and clinical diagnosis [5].

Image registration can be divided into three stages, namely, feature detection, feature description and feature matching [6-8]. In the feature detection stage, it is necessary to find the points with obvious features on each image as points of interest, called key points or feature points; in the feature description stage, a unique feature descriptor should be generated according to each feature point; then in the feature matching stage, the mapping relation of two groups of feature points with descriptors has to be produced through a matching algorithm. Therefore, it is apparent that the key to image registration is to identify the correspondences of feature points by finding the points similar to feature descriptors. By finding feature points with similar feature descriptors, their correspondences can be determined, and the correspondences can be used for registration. The accuracy and effectiveness of image registration largely depend on the quality and quantity of keypoint matching.

By now, image registration optimization methods can be divided into two categories: image registration based on traditional methods and image registration based on deep learning methods. The methods above both realize the three stages of registration and both need to find feature points as the matching criteria. The traditional image registration methods are generally based on local feature matching of detectors, such as Scale Invariant Feature Transform (SIFT) [9], Accelerated Robust Feature Transform (SURF) [10], and Radiation insensitive Feature Transform (RIFT) [11], which extract local features to achieve matching of image features such as points, lines, and surfaces. Usually, under the constraint of the objective function, feature matching is achieved by traversing image information, extracting feature points, and calculating the optimal spatial transformation. The advantage of this is that it has a small computational cost, which makes it easy and fast to detect image features, and there is no complex preprocessing process for the image.

But here are also some disadvantages of traditional image registration algorithms. Firstly, they can only extract shallow features [12], making it difficult to extract deeper and more global features of the image. The selection of feature points focuses on the position of the image's appearance shape transformation, and cannot freely select the position of feature points, nor can they generate specified feature symbols that describe the key points of the image. When facing high-resolution images, the number of similar feature points increases, which affects registration accuracy. Secondly, the detector-based local feature matching method utilizes techniques such as a sum of squared differences, normalized cross-correlation, and mutual information [13] to transform the matching problem into an optimization problem, which is greatly influenced by the objective function. Most similarity measurement methods have many local minima and may not necessarily obtain the global optimal solution [14]. Thirdly, for the multimodal medical images, there may be some poor textures, pattern duplication, deformations, stretching, etc., and big differences in the collected images. The collected images have significant differences, and simple transformation estimation alone cannot describe the matching situation, which greatly affects the matching results and makes it difficult to achieve complex feature medical image matching.

In response to some of the problems that have arisen in the traditional algorithms mentioned above, researchers have adopted deep learning methods to improve them, mainly through neural network training, learning image features, analyzing image data, and mining deep structures, in order to improve and optimize the image fusion effect and solve the complex feature extraction problems in traditional methods. The present deep learning method mainly using the dense matching algorithm or optical flow analysis, holds the basic thought of finding more key points and more excellent matching algorithms. Among them, in references of [15,16], convolutional neural network (CNN) is used to learn better feature detectors and descriptors from data; in reference[17], FlowNet, a rapid registration network, is capable to use the end-to-end fully convolutional network (FCN) to predicate the optical flows of two input images in a direct manner; in reference [18], it constructs feature vectors in the dense space and mentions learning the comparison of descriptors of the pixel features and optimizing the feature matching with the nearest neighbor searching algorithm; in reference [19], NCNet builds all the possible matching situations and uses 4D convolution which saves calculation cost; in reference [20], it conforms to the thoughts of dense matching and also puts forward a method that is from coarse to fine to make the dense matching more precise; in reference [21], SuperGlue network adopts graph neural network (GNN), a wide model and a simple generalized linear assignment method with an approximate solution to accomplish the image registration; The LoFTR [22] network uses self-attention layer and cross attention layer to obtain feature descriptors based on two images, effectively utilizing attention mechanism, which is a method that balances efficiency and accuracy.

The different methods above show good properties in the related image datasets, but medical images are different from natural images, optical images and scene images, and have their unique characteristics. Firstly, the resolution of medical images is usually higher than that of natural images, reaching 10k*10K or even higher, which means the computer needs more computing resources. Secondly, differing from optical images, medical images need to manifest the body tissues and structures through virtual staining. The relevant human tissue may exhibit mismatches in color tone, contrast, and brightness. Thirdly, medical images have more repetitive textures than scene images, which poses higher requirements for image registration. Therefore, based on the registration models above and combined with DETR model [23], this article proposes a new network model applied in the key points' positioning of multimodal image registration by improving the self-attention and cross-attention layers of LoFTR, which is used to cope with the problems of inaccurate key points' positioning and long-process matching of medical images.

2 IMPROVED DETR MODEL ON KEY POINTS' POSITIONING

2.1 Basic DETR Model

Detection Transformer (DETR), launched by Facebook, is an object detection model that demonstrates considerable accuracy and timeliness while also having the characteristic of being easy to generalize. As is illustrated in Figure 1, the whole framework of DETR is composed of three main components, namely, a CNN backbone network to extract the representation of image features, an encoder-decoder transformer structure, and a simple prediction network [24]. DETR uses a CNN backbone network to learn the feature information of the input images. The CNN divides images into uniform small blocks and embeds positional encoding information, which is then passed to the Transformer encoder. After encoding, the feature vector is passed to the Transformer decoder. The decoder takes a fixed number of object query vectors as input and extracts the encoder's output vector. Pass each output of the decoder to a shared feedforward network (FFN), which predicts the detection results.

The core of DETR is the encoder-decoder structure, which is based on the attention mechanism [25]. Figure 2 shows that in the attention mechanism, query, key, and value vectors are all involved in operations. The attention mechanism selects related information by measuring the similarities between query elements and other key elements. The output vector is the sum of value vectors weighted by similarity scores. If the similarities are higher, more relevant information is extracted from the value vectors.

Figure 2, Q , K , V represents the query vector, key vector and value vector respectively. The weight of attention is calculated by the dot products of Q and K , and then retrieve information from V according to the weight. During the training process, each feature vector will use its own query vector to take calculations with all other features to renew its weight values and train its own vector globally every time attention is calculated.

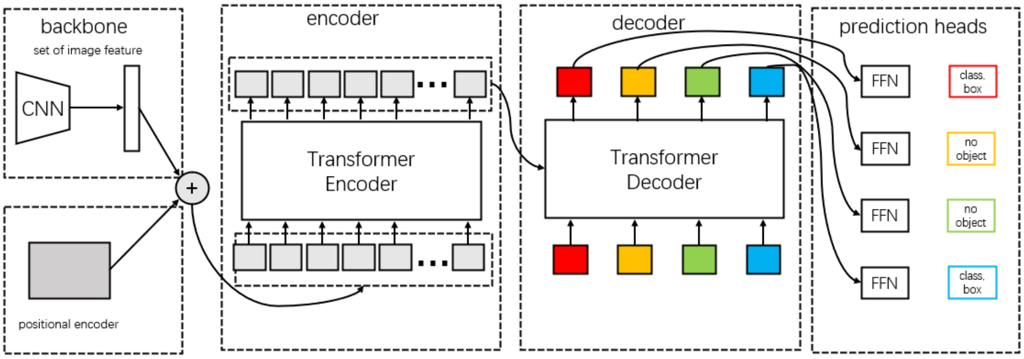

Figure 1: DETR network model structure.

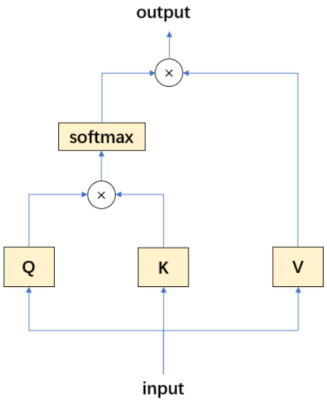

Figure 2: The computing process of attention mechanism.

2.2 Construction of the Improved DETR Model

After the comparison between the traditional registration method and the deep learning registration method, based on the encoder-decoder structure of the DETR model, combined with the optical flow analysis and self-attention and cross-attention methods, this paper proposes a new network structure to complete the specified key point localization task. In order to adapt to the key point-matching task, two improvements have been made to the DETR model. Firstly, a feature extraction network for two images has been added in the feature extraction section of the DETR architecture, and information exchange between image pairs is completed in the encoder section. Secondly, a prompt encoder module has been added to construct the desired location prediction network. Given a group of flexible key points as the query objects, the new model will prompt the location of network key points through the prompt information encoding module. Then after the new model infers the information of key points and the context of the whole image, it will directly output the final predicted positioning results of target key points in a parallel manner, achieving the end-to-end matching of key points.

2.2.1 Feature extraction and receptive field

The image registration task requires feature extraction of two images, usually referred to as the source image and target image. The improved DETR must employ the CNN twice to extract the features of these two images respectively. However, using only Convolutional Neural Networks cannot accurately extract information because the receptive field of convolutional neural networks is very small. The receptive field in the neural network refers to the receptive range of differently positioned neurons on the source image. The larger the receptive field, the larger the range of the original image that the neuron can contact. In convolutional neural networks, the deeper neurons have wider receptive ranges. As is shown in Figure 3, passing one layer of 3×3 convolution, one position in layer 1 can receive 9 positions in layer 0. And with a bigger convolution kernel and a deeper network structure, there would be a wider receptive field. But using big-size kernels like 5×5 or 7×7 ones, is often considered to lose some information about features. A deeper neural network structure leads to a heavier calculation complexity, and only stacking a certain number of network layers will result in poor performance.

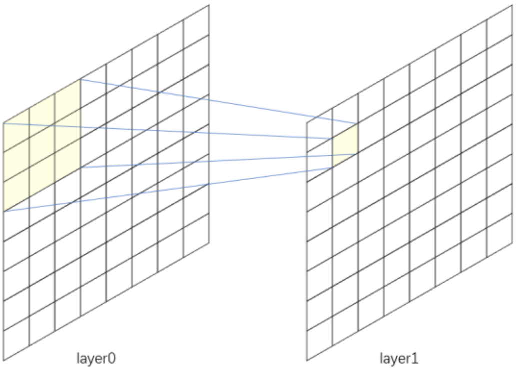

Figure 3: One element in layer 1 receives nine elements in layer 0 with 3×3 convolution.

After the improved DETR extract features, the information is transmitted into the encoder and takes further encoding to secure a wider receptive field. The encoder based on the attention mechanism calculates the dot products through query vectors and the image's own key vectors, which is called the self-attention mechanism; and then the process of calculating the dot products through query vectors and key vectors of another image is called the cross-attention mechanism. By alternately using the self-attention mechanism and cross-attention mechanism in the encoder, the vector of the current image block and the similarity relationship between different image blocks are calculated, thereby completing the feature extraction encoding of the image.

2.2.2 Position coding network.

The attention mechanism processes all information in a parallel way. If positional information is not provided to the model, the model cannot get the semantic grammar differences in the sequential order of image blocks' feature vectors and the related structures need to be added to complement the positional information. There are two mainstream ways to represent location information, absolute position encoding and relative position encoding. For the improved DETR, employs the relative position encoding to encode the images' positions with the coordinates of key points. The encoded positional information of key points is the initialization status of prediction heads of DETR to engage in the encoder-decoder model. Because the position of the key points comes from the source image, the feature information of the same source image in the detection head is added to the training in the encoder section. In the decoder section, the prediction heads decode the feature information of the target image and get the predicted positional features in order to complete the matching of key points.

3 IMPROVED MODEL’S NETWORK STRUCTURE AND LOSS FUNCTION

3.1 Network Structure of Improved Model

Given image pairs: source image A and target image B, as well as N key point positions on image A. The model task is to predict the specific positions of key points on target image B. Set x_i^A, y_i^B as the coordinate of the number i key point on image A, x_i^B, y_i^B as the coordinate of the corresponding coordinate on image B and x_i^P, y_i^P as the coordinate of the predicated coordinate. Obviously, the closer the distance between x_i^B, y_i^B and x_i^P, y_i^P , the better the predicted effect. Figure 4 overviews the basic framework of the improved model.

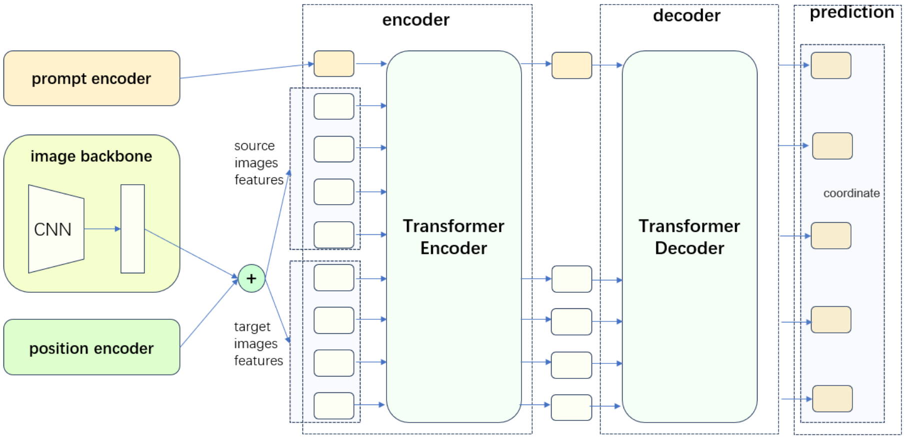

Figure 4: The structural schematic of the improved model in this article.

As is shown in Figure 4, this model is based on the DETR framework and consists of four main components: firstly, a CNN backbone network to extract the feature representation of images; secondly, a prompt encoder to fit the matching of key points; thirdly, an encoder-decoder transformer; fourthly, a simple prediction network to make the final predicated coordinates with the extracted representation of features.

3.1.1 Improved feature extraction network

Different from feature detection which extracts features of one image, image registration requires the feature extraction of both the source image A and target image B. In this article, the feature extraction is taken by CNN, which with the inductive bias of translation equivariance and locality is suitable to make the local extraction of image blocks, to let the characteristic length uniform and to manage the computing cost. Figure 5 illustrates that after the two grayscale images pass CNN’s four layers, layer0, layer1, layer2, layer3, the dimensions of feature vectors turn into 128, 128, 196 and 256 respectively, and the sizes of feature images change from 480*640 to 240*320, 240*320, 120*160 and 60*80 respectively. With the local feature extraction, the information of images is transformed into 4800 256-dimensional feature vectors, each of which represents the feature information in the 8*8 block of images. Then the sequential information will be processed as the basic elements by the following transformer framework, and the basic feature vectors are called tokens. Finally, image A and image B will both get 4800 256-dimentional tokens which are named F_{A} and F_{B} respectively.

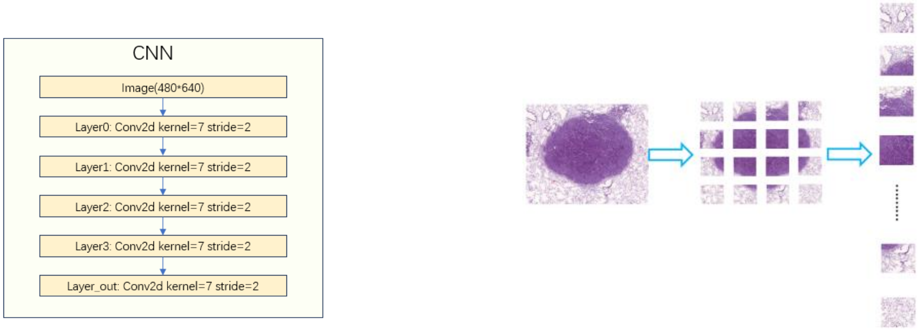
CNN
Image(480*640)
Layer0: Conv2d kernel=7 stride=2
Layer1: Conv2d kernel=7 stride=2
Layer2: Conv2d kernel=7 stride=2
Layer3: Conv2d kernel=7 stride=2
Layer_out: Conv2d kernel=7 stride=2

Figure 5: CNN used in feature extraction.

3.1.2 Positional encoder and prompt encoder

The prompt encoder turns the positional information of key points to the prompt information and inputs it to the neural network. The key points are expressed by position coding [25]. And N sparse position encodings and Gaussian matrices are combined to form a sparse position encoding tensor F_p, which is the token for the hint information. Then, generate dense hints based on sparse position encoding tensors as position information F_C. The position information F_C has a spatial correspondence with the image, and the channel dimension is mapped to 256 dimensions using a 1x1 convolution, which is added to the image elements F_{A}, F_{B} and embedded into the network. The summation of F_{A}, F_{B} and the corresponding part F_C can effectively combine the image information of F_{A} and F_{B} with the prompt information F_p and the position-coding F_C [26, 27]. It is convenient for the encoder and decoder to utilize the attention mechanism when the dynamic relative position coding is employed to insert the positional information of the token. After embedding the position information and features F_{A} and F_{B} of images A and B, a new feature vector of F_A' and F_{B}' is obtained, which, together with the prompt information F_p, enters the encoder layer for encoding operation.

3.1.3 Image encoder and decoder network

The features of F_{A}' and F_{B}' got by adding the information on image features and positions and taking the image information coding, which makes use of the framework of the transformer [28]. The encoder consisted of encoder layers connected in order, alternately employing the self-attention coding and cross-attention coding [22] to finish the image coding. Use a self-attention mechanism to encode the self-feature information of source image A and target image B, and use a cross-attention mechanism to interact with the feature information of source image A and target image B. Figure 6 reveals the inner framework of encoder layers.

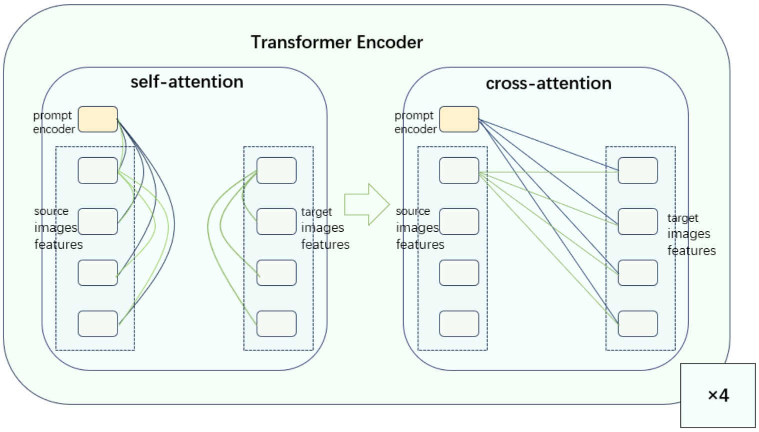

Figure 6: Transformer encoder network.

Formula (1) expresses the calculation method of attention mechanism, where Q, K and V are the representatives of query, key and value vectors respectively. The self-attention and cross-attention mechanisms can be achieved by dot product calculation with the query vectors and the image's own key vectors or the key vectors of another image.

$$Attention\ \ Q,K,V\ \ = softmax\ \ QK^T\ \ V \tag{1}$$

The key points' positional information from source image A in tensor F_p , takes part in the following calculation with $F_{\mathrm{A}}{}'$. The self-attention mechanisms of image A and image B take $F_{\mathrm{A}}{}'$ and $F_{\mathrm{B}}{}'$ as the input to take calculations. This kind of self-attention allows for the maximum receptive field of one's own image in the initial stage of training. The cross-attention mechanism between image A and image B is used $F_{\mathrm{A}}{}'$ as the query vector, $F_{\mathrm{B}}{}'$ the key vector and the value vector. Conversely, the cross-attention mechanism between image B and image A is used $F_{\mathrm{B}}{}'$ as the query vector and $F_{\mathrm{A}}{}'$ as the key vector and value vector. In this way, the cross-attention mechanism is capable of making information interact as much as possible between these two images.

The encoder is made up of the alternate self-attention layers and cross-attention layers, and the self-attention and cross-attention mechanisms are used alternately four times to complete the image encoding process. The above attention mechanism adopts an 8-head attention mechanism, which uses 8 identical attention structures and operates in parallel with 256-dimensional vectors under different initialization parameters. Each attention structure can learn different feature information. Finally, the resulting feature vectors of $F_p{}''$ and $F_{\mathrm{B}}{}''$ taken from the studied F_p and $F_{\mathrm{B}}{}'$ to participate in the following decoding process. With the mechanisms of self-attention and cross-attention, the information in every token possesses the structural information of two images, which is beneficial to accomplishing the positional prediction of key points. Moreover, the decoder, by the framework of the transformer, takes the multi-head self-attention mechanism and the decoder-attention mechanism to convert the feature vectors. The converted vectors are the positional features of the corresponding key points in image B. The structure of the decoder is illustrated in Figure 7.

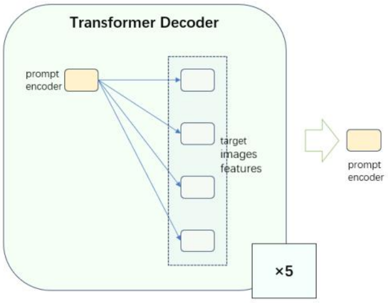

Figure 7: Transformer decoder.

Take N tokens F_p'' with a vector size of 256 as the query vector for the attention mechanism and use the sum of the corresponding elements of feature vector F_{B}'' and position encoding information F_C generated by target image B as the key vector and value vector. F_p'' are parallelly decoded in each decoder layer and the predicated vector F_i'' refers to the predicted result of number i token. For the information in F_p is got from the learning of different key points' positions and N various positional results will be generated when the information of N positions parallelly passes the same decoder, F_i'' which means the positional feature of the number i key point in target image B. F_i'' is reasoned as the feature vector by the model which fully uses the encoder-decoder framework of the attention mechanism, and then the vector employs the corresponding relations between two images to take reasoning at the global level while preserving its own unique positional feature.

3.1.4 Prediction network

The final prediction network is composed of a three-layer perception with a middle dimension of 256 to calculate the coordinates of corresponding key points on the target image. Then take a parallel calculation to predicate the positions of N key points, which x_i^p, y_i^p represents the coordinate of the number i of key points.

3.2 Loss Function of Model

One of the difficulties of training is to assess the predicted positions in the real situation. The loss function l is made up of positional loss l_c and structural loss l_f.

The error parameter to measure the registration of key points is TRE(Target Registration Error), namely, the Euclidean distance between the predicted position x_i^p, y_i^p and the real one x_i^B, y_i^B . The computing method is shown in formula (2).

$$TRE=\sqrt{x_i^p-x_i^B{}^2+\ y_i^p-y_i^B{}^2} \tag{2}$$

The model takes the normalized TRE (Registration Target Registration Error) as the positional loss l_c. The computing method is manifested in formula (3).

$$l_c=\frac{1}{\mathrm{N}}\sum_{i=1}^{N}\frac{TRE}{\sqrt{w^2+h^2}} \tag{3}$$

In this formula, w and h mean the width and height of the image, respectively. However, if only the positional loss l_c is used, the predicted positions will focus on the image center, falling into the local optimal solution in the training process. Hence, the structural loss l_f is also needed.

Regard key points' coordinates as a set, where the distance matrix formed by the distances of each point pair is considered as the set's structural information. Utilize x^p,y^p and x^B,y^B to calculate the distance matrixes named M^B and M^P of the inner set, respectively. For the parameter values of these two matrixes should be as equal as possible, the computing method l_f is shown in formula (4).

$$l_f=\frac{1}{N^2}\sum_{i=1}^{N^2}\sqrt{M_i^P-M_i^B{}^2} \tag{4}$$

In this formula, M_i^P and M_i^B mean the value of number i element in M^P and M^B respectively. M^B and M^P both belong to the $N\times N$ matrix.

The formula to calculate the loss function l is $l=l_c+\lambda l_f$, where the hyper-parameter λ will be smaller as the training goes on so that the model effect can be measured by l_c (rTRE） more conveniently.

4 SIMULATION EXPERIMENTS AND RESULTS ANALYSIS

4.1 Experimental Platform

The experimental platform is NVIDIA RTX4080 GPU adopting version 1.13.1 of pytorch and version 12.0 of CUDA. The training takes 60 rounds with an initial learning rate of 1×10^{-5} and batch size of 1×10^{-5}. After about 7 hours, the model converges. The whole model uses the weights initialized at random to carry out the end-to-end training with the 256-dimentional chained vectors between modules, each of which is designed with a softmax activation function and a normalized module. In the present situation, the average running time of modules every time is 113 milliseconds.

In this article, two datasets are selected. The first dataset is (1) ANHIR dataset [29], which is composed of images of pathological tissues, like the lung lobe, breast and kidney. Every group of images stained with various colors contains consecutive tissue slices, where the images can take registration freely. And the trained 230 image pairs in this article are at a medium solution with a pixel of about 8k×12K. The second dataset is (2) the FIRE dataset, which collects 39 patients' 129 fundus retinal images with pixels of about 3K×3K. Also, the data including the mask and the corresponding position marker of each image, is used to take the zero-shot verification. The ordinary images are illustrated in Figure 8 (a) and Figure 8 (b).

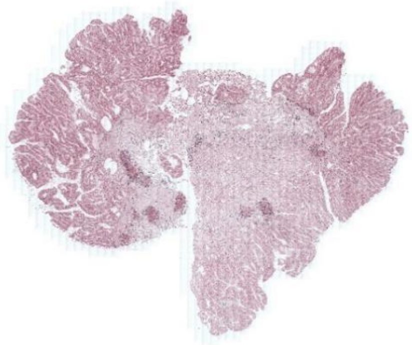

Figure 8 (a): ANHIR dataset.

Figure 8 (b): FIRE dataset.

4.2 Assessment Indicators and Experimental Analysis

In this article, rTRE and homography estimation are applied to assess the properties of the model.

4.2.1 Analysis of rTRE

According to section 3.2, rTRE (Registration Target Registration Error), l_c in experimental training, refers to the registration error between key points from the source and target images. The average registration error reflects the accuracy of registration; the median and maximum show its robustness; and the executive time reveals the efficiency of the algorithm. The model aims to pursue accuracy and timeliness in keypoint matching. Randomly pass image pairs into the model, calculate the relative target registration error (rTRE) and runtime of the model, and analyze the impact of different methods and images on the model.

Method	*Average-rTRE*	*Max-rTRE*	*Median-rTRE*	*Average-Time(S)*
UPENN	0.004057	0.023043	0.002791	1.451193
AGH	0.005636	0.030005	0.003804	6.863679
MEVIS	0.005191	0.026069	0.003852	0.145392
TUB	0.004731	0.014927	0.004099	0.000705
CKVST	0.006044	0.026128	0.004609	7.127142
TUNI	0.010363	0.038723	0.008724	10.320549
RVSS	0.047089	0.103180	0.045024	4.723187
UA	0.056887	0.119045	0.054878	1.470925
DROP	0.061602	0.122958	0.061336	3.406355
Elastix	0.069476	0.137054	0.068433	2.962337
ANTs	0.069322	0.134296	0.068621	43.092353
bUnwarpJ	0.079704	0.149613	0.079557	9.151172
NiftyReg	0.082488	0.151446	0.082781	0.151164
our	0.013054	0.242175	0.008343	0.134871

Table 1: Comparison results under different models.

The experimental results are shown in Table 1, The average registration error of the method proposed in this article is 0.013054, and the average matching time is nearly 0.134871 seconds.

In most cases, the matching of key points can be finished within a reasonable time and its real and predicated situations are shown in Figure 9 (a) and Figure 9 (b).

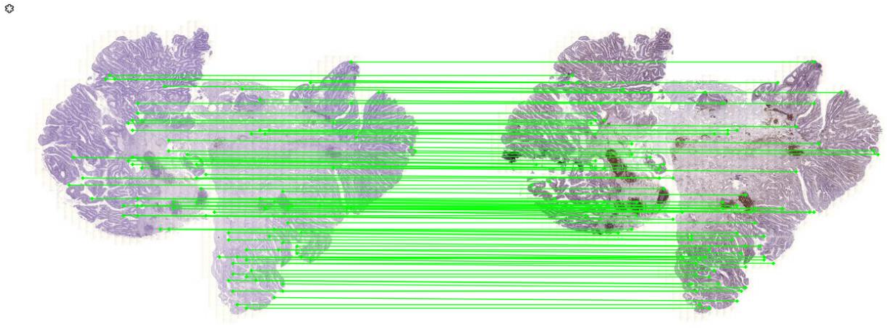

Figure 9 (a): Real situation.

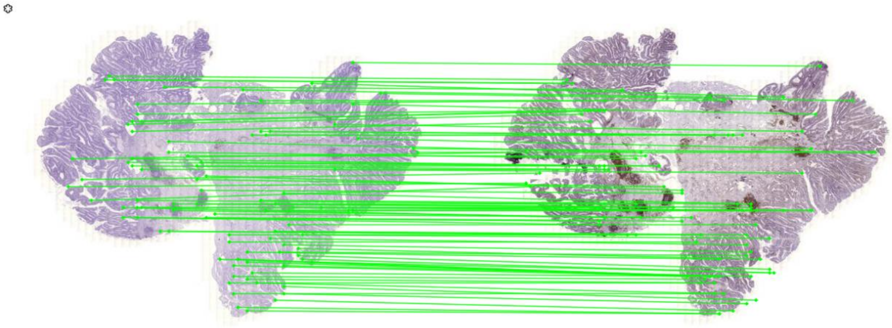

Figure 9 (b): Predicted situation.

Compared with other experimental results, these results show that although there are structural deformations, color changes and different detailed features in the experimental image pairs, the model mentioned in this article still accomplishes the matching of key points well, owns certain robustness for some situations, like deformation, and can make a balance between the precision and time efficiency of key points matching. Also, compared to the methods of UPENN, AGH, TUNI, CKVST, ANTs, RVSS and bUnwarpJ, in terms of time efficiency, the deep neural network model has advantages, which are caused by the model's parallel computing. Besides, about the deformation processing, the accuracy of the method proposed in this article is better than Elastix, ANTs, RVSS, bUnwarpJ and NiftyReg, for the expanded receptive field is capable of adding the structures of images into training in the early stage to boost the method's accuracy. However, compared to the methods of UPENN, AGH, MEVIS and TUB, the precision of the proposed method is lower, maybe since the image pyramid is not well integrated with this model, which means the detailed parts still need to be adjusted and strengthened.

4.2.2 Analysis of homography estimation

This study makes a comparison between the LoFTR model and the SuperGlue one, for the self-attention and cross-attention mechanisms used in the proposed model, are also employed by these two models. The method of RANSAC in OpenCV is adopted as the robustness estimator to calculate the angle error of image registration. The resolution of the experimental images from ANHIR dataset is 8K×12K, which is almost 20 times larger than 640×480 of images set by LoFTR.

Therefore, it is necessary to turn up the threshold of RANSAC to 60px, 100px and 200px and to employ the percentage form to express the conformity error.

Method	*homography estimation AUC*			*matches-points*
	60px	*100px*	*200px*	
SuperGlue	22.3	30.1	54.3	605
LoFTR	24.3	29.4	60.1	810
Our	30.2	42.8	80.4	134

Table 2: Comparison with other attention mechanism methods.

The experimental results are shown in Table 2, it is obvious to see that although the LoFTR algorithm can test a great number of feature points, its matching accuracy is adversely affected under the situation of a medical dataset. There may be two reasons for this. For one thing, this algorithm utilizes the scene datasets both indoors and outdoors to take training, while these datasets cannot fit the medical images well for there are more repeated textures and deformations than scene data in medical images; for the other thing, in this algorithm, the confidence threshold related to successful matching is set to be so low that it is only 0.2, while medical images need higher confidence interval to make calculations. Then, for the model of this article, since it only focuses on the interested key points, its accuracy is better than LoFTR and SuperGlue, while it also has to bear the cost of the amount of used key points being smaller.

4.3 rTRE Loss of Different Tissue Samples in Experimental Datasets

In the experimental datasets, the number of images has a great impact on the model. The datasets contain breast tissues, colon adenocarcinomas (COAD), gastric mucosa and gastric adenocarcinoma tissues, kidney tissues and mice-kidney tissues, lung lesions and lung lobes and mammary glands. Among them, there are 5 pairs of breast tissues, 84 pairs of COADs, 13 pairs of gastric mucosa and gastric adenocarcinoma tissues, 20 pairs of kidney tissues and mice-kidney tissues, 70 pairs of lung lesions and lung lobes and 38 pairs of mammary glands. The rTRE loss values of images are shown in Table 3:

Method	*Average-rTRE*	*Max-rTRE*	*Min-rTRE*	*Standard Deviation*
Breast tissue	0.102909	0.242175	0.022016	0.074842
COAD	0.010613	0.042750	0.002526	0.008008
Gastric mucosa	0.031411	0.084196	0.003162	0.033246
Kidney tissue	0.008982	0.019115	0.004092	0.004099
lung lesion	0.007942	0.014522	0.003553	0.002776
Mammary glands	0.011303	0.020974	0.004146	0.004616

Table 3: rTRE loss of different tissue samples in datasets.

The average rTRE loss value is a reflection of the fitting ability and accuracy of the model. Among all the tissues, the best tissues are kidney tissues and mice-kidney tissues lung lesions and lung lobes and the second one is COADs. Although their losses all outperform the average level, the effect of COADs with more data is a little bit worse than the first two, maybe COADs have a larger variability and more various shaped images. Then the loss values of breast tissues gastric mucosa and gastric adenocarcinoma tissues are the highest, and the loss of gastric mucosa is even better than that of breast tissue, which obviously shows that the loss has direct relations with fewer related training data in datasets. The standard deviation regarded as the reflection of the model's stability, is related to the amount of sample data. The standard deviations of COADs, lung lesions and lung lobes, and mammary glands which are all with large data amounts, are 0.008008, 0.002776 and 0.004616 respectively, which is a showcase of the matching stability. However, for the data with small amounts, like breast tissues gastric mucosa and gastric adenocarcinoma tissues, their standard deviations are 0.074842 and 0.033246, which are nearly 10 times larger than the standard deviations with excellent data amounts. Because there is a 10-times gap between these training data amounts, the model has a large dependence on the data amount.

The results above reveal that to optimize the model's accuracy and stability, the expansion of data is a vital direction, for data amount can make contributions to the accuracy, complexity and stability of the model. The larger the amount of data, the more accurate the statistical distribution of the model for real data. By providing more data samples, the model can better capture the true features and patterns of the data, thereby improving the accuracy of predictions. Also, when the data amount is larger, the complexity of the model will be advanced accordingly, for the larger datasets can support more complex model structures, which can better fit the non-linear relationship of data so as to improve the accuracy. In addition, with a bigger data amount, the model will be more stable in terms of the volatilities with different features, for the model may be too sensitive to some rare feature changes with a relatively smaller amount. In this way, larger datasets are capable of offering more samples to cover the changing range of different features so as to make the model better tackle diverse situations.

4.4 Model Generalization

This study takes a zero-shot verification on the model with FIRE dataset and observes that the rTRE of experimental results is 0.02414, showcasing that the model proposed in this article owns excellent generalization performance, an indicator to evaluate the adaptation and universality of models when they face the unknown data. Through this verification, the model has not only been proven to possess a good performance in data training but also to secure relatively accurate results in new fields and missions. The experimental results of FIRE dataset are illustrated in Figure 10.

The generalization performance of this model is not only shown in the experimental results but also in the model's design and algorithm. Based on the deep learning technology, this model with the strong ability, is able to learn from the limited samples and generalize the learning results. Besides, with certain robustness, the framework and training method of the model can process the changes of the input data and the noises to make the model's application in real cases more reliable and stable. According to the results and features, the model secures excellent properties and strong generalization ability so that it can be employed in various real cases and unknown fields. Therefore, important references for further research, engineering application, and technology development in the related fields are provided.

5 CONCLUSIONS

This article raises a Registration Algorithm based on the DETR model to solve the problem of medical image registration. Firstly, the features of the two images and the relations between the image pair are extracted through the working of CNN and the attention mechanism. Then, the model training and verification will be completed by designing the specific loss function. Finally, the experimental results demonstrate that the usage of CNN to pre-process the multi-sensor images to extract their features and the usage of feature matching method based on deep learning can effectively boost the accuracy and robustness of the matching algorithm, which is a showcase of the feasibility to process other downstream tasks through transformer model generalization.

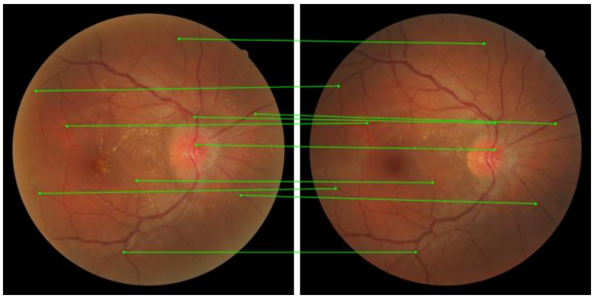

Figure 10 (a): The real results of FIRE dataset.

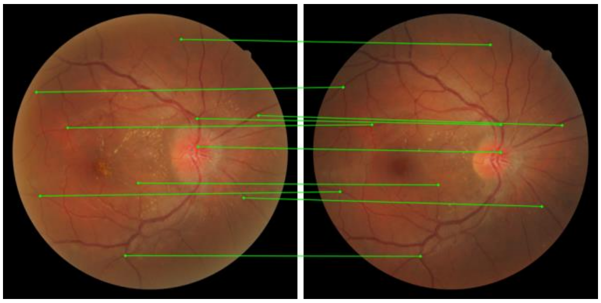

Figure 10 (b): The predicted results of the model in this article.

In addition, this article improves the network structure of the ordinary DETR to fit the image registration so that the quality of image generation is boosted and effective reference data are provided to deal with the registration of medical images and the processing of heterogeneous images

In the following research, the image features pyramid and the image deformation field will be introduced to the DETR model study for the registration speed acceleration and enhancing its accuracy, which would offer thoughts on image change detection or fusion in other fields.

Jianfeng Han, https://orcid.org/0009-0001-7702-4129
Jingxuan Zhao, https://orcid.org/0009-0007-1959-7002
Renjie Li, https://orcid.org/0009-0005-3488-1146
Yong Zhang, https://orcid.org/0000-0001-8854-1319

ACKNOWLEDGEMENT

This article was funded by The Natural Science Foundation of Tianjin (grant number 23JCQNJC00840).

REFERENCES

[1] Borovec, J; Munoz-Barrutia, A; Kybic, J: Benchmarking of image registration methods for differently stained histological slides, 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, 3368-3372. http://doi.org/10.1109/ICIP.2018.8451040.

[2] Han, S; Liu, X; Dong, J: Remote Sensing Multimodal Image Matching Based on Structure Feature and Learnable Matching Network, Applied Sciences, 13(13), 2023, 7701. https://doi.org/10.3390/app13137701

[3] Dong, L; Liu, F; Han, M; You, H: Mosaicing Technology for Airborne Wide Field-of-View Infrared Image, Applied Sciences, 13(15), 2023, 8977. https://doi.org/10.3390/app13158977

[4] Zhu, C; Zhang, Y; Xuhua, P; Qi, C; Qingyu, F: Improved Harris Hawks Optimization algorithm based on quantum correction and Nelder-Mead simplex method, Mathematical Biosciences and Engineering, 19(8), 2022, 7606-7648. https://doi.org/10.3934/mbe.2022358

[5] Fernandez‐Gonzalez, R; Jones, A; Garcia‐Rodriguez, E: System for combined three‐dimensional morphological and molecular analysis of thick tissue specimens, Microscopy Research and Technique, 59(6), 2002, 522-530. https://doi.org/10.1002/jemt.10233.

[6] Yong, Z; Xinyue, L; Li, W; Shurui, F; Lei, Z; Shuhao, J: An Autocorrelation Incremental Fuzzy Clustering Framework Based on Dynamic Conditional Scoring Model, Information Sciences, 648(11), 2023, 119567. https://doi.org/10.1016/j.ins.2023.119567

[7] Yong, Z; Renjie, L; Qi, C; Derui, Z; Xiaolin, W; Changzhou F; Jiahui S; Shuhao J: An improved bicubic interpolation SLAM algorithm based on multisensor fusion method for rescue robot, International Journal of Sensor Networks, 42(2), 2023, 125-136. https://doi.org/10.1504/ijsnet.2023.131656

[8] Zhang, Q; Xiang, W: Cross-Modal Image Registration via Rasterized Parameter Prediction for Object Tracking, Applied Sciences, 13(9), 2023, 5359. https://doi.org/10.3390/app13095359

[9] Lowe, D. G: Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60(2), 2004, 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[10] Bay, H; Tuytelaars, T; Van Gool, L: Surf: Speeded up robust features, Proceedings of the 9th European Conference on Computer Vision - Volume Part, 2006, 404–417. https://doi.org/10.1007/11744023_32

[11] Li, J; Hu, Q; Ai, M: RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform, IEEE Transactions on Image Processing, 29, 2019, 3296-3310. https://doi.org/10.1109/TIP.2019.2959244

[12] Rocco, I; Arandjelovic, R; Sivic, J: Convolutional neural network architecture for geometric matching, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 6148-6157. https://doi.org/10.1109/CVPR.2017.12

[13] Chen, H. M; Arora, M. K; Varshney, P. K: Mutual information-based image registration for remote sensing data, International Journal of Remote Sensing, 24(18), 2003, 3701-3706. https://doi.org/10.1080/0143116031000117047

[14] Boveiri, H R; Khayami, R; Javidan, R: Medical image registration using deep neural networks: a comprehensive review, Biomedical Signal Processing and Control, 73, 2022, 103444. https://doi.org/10.1016/j.bspc.2021.103444

[15] Gong, M; Zhao, S; Jiao, L: A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information, IEEE Transactions on Geoscience and Remote Sensing, 52(7), 2013, 4328-4338. https://doi.org/10.1109/TGRS.2013.2281391

[16] Zhao, S; Lau, T.; Luo, J.: Unsupervised 3D End-to-End Medical Image Registration With Volume Tweening Network, IEEE Journal of Biomedical and Health Informatics, 24(5), 2020, 1394-1404. https://doi.org/10.1109/JBHI.2019.2951024

[17] Dosovitskiy, A; Fischer, P; Ilg, E: Flownet: Learning optical flow with convolutional networks, Proceedings of the IEEE International Conference on Computer Vision, 2015, 2758-2766. https://doi.org/10.1109/ICCV.2015.316

[18] Schmidt, T; Newcombe, R; Fox, D: Self-supervised visual descriptor learning for dense correspondence, IEEE Robotics and Automation Letters, 2(2), 2016, 420-427. https://doi.org/10.1109/LRA.2016.2634089

[19] Rocco, I; Cimpoi, M; Arandjelović, R: NCNet: Neighbourhood Consensus Networks for Estimating Image Correspondences, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(2), 2022, 1020-1034. https://doi.org/10.1109/TPAMI.2020.3016711

[20] Li, X; Han, K; Li, S: DualRC: A Dual-Resolution Learning Framework With Neighbourhood Consensus for Visual Correspondences, IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(1), 2024, 236-249. https://doi.org/10.1109/TPAMI.2023.3316770

[21] Sarlin, P E; DeTone, D; Malisiewicz, T: Superglue: Learning feature matching with graph neural networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 4938-4947. https://doi.org/10.1109/CVPR42600.2020.00499

[22] Sun, J; Shen, Z; Wang, Y: LoFTR: Detector-free local feature matching with transformers, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 8922-8931. https://doi.org/10.1109/CVPR46437.2021.00881

[23] Carion, N; Massa, F; Synnaeve, G: End-to-end object detection with transformers, European Conference on Computer Vision, 2020, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[24] Kirillov A; Mintun E; Ravi N: Segment anything, 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, 3992-4003. https://doi.org/10.1109/ICCV51070.2023.00371

[25] Tancik, M; Srinivasan, P; Mildenhall, B: Fourier features let networks learn high-frequency fu nctions in low dimensional domains, CoRR 2020, 33, 2020, 7537-7547. https://doi.org/10.48550/arXiv.2006.10739

[26] Cheng, B; Schwing, A; Kirillov, A: Masked-attention Mask Transformer for Universal Image Segmentation, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 1280-1289. https://doi.org/10.1109/CVPR52688.2022.00135

[27] Yong, Z; Renjie, L; Fenghong, W; Weiting, Z; Qi, C; Derui, Z; Xinxin, C; Shuhao, J: An Autonomous Navigation Strategy Based on Improved Hector SLAM With Dynamic Weighted A* Algorithm, IEEE Access, 11, 2023, 79553-79571. https://doi: 10.1109/ACCESS.2023.3299293.

[28] Dosovitskiy, A; Beyer, L; Kolesnikov, A: An image is worth 16x16 words: Transformers for image recognition at scale, ICLR, 2021. https://doi.org/10.48550/arXiv.2010.11929

[29] Gupta, L; Klinkhammer, B M; Boor, P: Stain independent segmentation of whole slide images: A case study in renal histology, 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, 1360-1364. https://doi.org/10.1109/ISBI.2018.8363824