Unsupervised Histological Image Registration Using Structural Feature Guided Convolutional Neural Network

Lin Ge, Xingyue Wei, Yayu Hao, Jianwen Luo, *Senior Member, IEEE*, and Yan Xu

***Abstract*— Registration of multiple stained images is a fundamental task in histological image analysis. In supervised methods, obtaining ground-truth data with known correspondences is laborious and time-consuming. Thus, unsupervised methods are expected. Unsupervised methods ease the burden of manual annotation but often at the cost of inferior results. In addition, registration of histological images suffers from appearance variance due to multiple staining, repetitive texture, and section missing during making tissue sections. To deal with these challenges, we propose an unsupervised structural feature guided convolutional neural network (SFG). Structural features are robust to multiple staining. The combination of low-resolution rough structural features and high-resolution fine structural features can overcome repetitive texture and section missing, respectively. SFG consists of two components of structural consistency constraints according to the formations of structural features, i.e., dense structural component and sparse structural component. The dense structural component uses structural feature maps of the whole image as structural consistency constraints, which represent local contextual information. The sparse structural component utilizes the distance of automatically obtained matched key points as structural consistency constraints because the matched key points in an image pair emphasize the matching of significant structures, which imply global information. In addition, a multi-scale strategy is used in both dense and sparse structural components to make full use of the structural information at low resolution and high resolution to overcome repetitive texture and section missing. The proposed method was evaluated on a public histological dataset (ANHIR) and ranked first as of Jan 18th, 2022.**

***Index Terms*— Convolutional neural network, histological image, registration, structural feature, unsupervised learning.**

I. Introduction

REGISTRATION of multiple stained histological images is an important prerequisite step for many computer-aided histological image analysis tasks, such as the combination of gene expression maps from multiple specimens [1], [2] for grading or classification of diseases [3], [4].

There exists non-rigid deformation between histological images. Therefore, non-rigid registration is needed. There are many traditional methods for non-rigid histological image registration, such as B-spline deformation based methods [5]–[10], finite element models based elastic deformation model [11], large deformation diffeomorphic metric (LDDMM) image matching algorithm [12], greedy diffeomorphic algorithm [13], and multimodal algorithm [14]. Recently, deep learning (DL) has made big leaps in many fields of image processing [15]. In particular, there are some studies on histological image registration based on DL [16]–[18]. One method proposed by Zhao *et al.* as mentioned in [16] is based on a supervised convolutional neural network (CNN) with manually labeled key points. In such supervised learning-based methods, the ground-truth data with known correspondences across the set of training images is required [19]. Obtaining this type of data can be a very laborious and subjective process because human experts are typically needed [19], [20]. Thus, unsupervised DL methods are desirable for histological image registration.

Typically, unsupervised learning methods have inferior results compared with supervised learning methods [21]. Therefore, improvements in unsupervised learning methods are expected. Besides, the registration of histological images is challenging for the following reasons (Fig. 1): 1). Histological images have appearance variance due to multiple staining; 2). There are repetitive textures in histological images because

Manuscript received 20 January 2022; revised 21 March 2022; accepted 24 March 2022. Date of publication 1 April 2022; date of current version 31 August 2022. This work was supported in part by the National Key R&D Program of China under Grant 2017YFC0110903; in part by the National Natural Science Foundation in China under Grant 61871251, Grant 62022010, Grant 81771910, and Grant 62027901; in part by the Fundamental Research Funds for the Central Universities of China from the State Key Laboratory of Software Development Environment, Beihang University, China, under Grant SKLSDE; in part by the 111 Project in China under Grant B13003; in part by the Beijing Hope Run Special Fund of Cancer Foundation of China under Grant LC2018L02; and in part by the High Performance Computing (HPC) Resources at Beihang University. *(Lin Ge and Xingyue Wei contributed equally to this work.) (Corresponding authors: Jianwen Luo; Yan Xu.)*

Lin Ge, Xingyue Wei, Yayu Hao, and Jianwen Luo are with the Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 100084, China (e-mail: luo_jianwen@tsinghua.edu.cn).

Yan Xu is with the State Key Laboratory of Software Development Environment and the Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Research Institute of Beihang University, Beihang University, Beijing 100191, China (e-mail: xuyan04@gmail.com).

Digital Object Identifier 10.1109/TMI.2022.3164088

of homogeneous tissues; 3). There may exist section missing when processing tissue sections [16]. With the existence of these challenges, common strategies for performance optimization in registration do not perform well in histological images [16].

A common approach to appearance variance challenges in registration is to integrate metrics that are robust to intensity variance [22]. The structure information is robust to intensity variance [22] and structural patterns can be used for measurements of the correlation between multiple stained slices [23]. Besides, the structural information at low resolution as rough scale that describes the structural features in a large region can overcome repetitive texture [24]–[26]. Meanwhile, the structural information at high resolution as fine scale that describes the structural features in a small region can overcome the section missing [24]. Structural features at rough and fine scales not only overcome the challenges in histological image registration, but also are important in diagnosis [27]–[30], which is the purpose of histological image registration. To improve the results of unsupervised DL methods, overcome the challenges of histological images, and better serve the following diagnosis, we propose a structural feature guided unsupervised CNN (SFG) for non-rigid histological image registration.

The proposed method has two components of structural consistency constraints according to the formations of structural features: dense structural component and sparse structural component. In the dense structural component, we compute the structural consistency by computing the staining type-independent structural feature maps from the histological images. A structure-consistency loss is defined as the correlation between the structural feature maps of a pair of images. In the sparse structural component, we compute the structural consistency using the distance between matched key points per pair, which can describe the significant structures that imply anatomical information in the histological images [31]. It is worth mentioning that the matched key points are automatically obtained. We also utilize a multi-scale strategy to make full use of the structural information at low resolution and high resolution to overcome repetitive texture and section missing. The registration performance of the proposed algorithm ranked first in the challenge of Automatic Non-rigid Histological Image Registration (ANHIR)[1] (as of Jan 18th, 2022), the only open comparison of image registration algorithms on microscopic images. The proposed algorithm achieved the lowest Median rTRE (0.00067) among all the submissions to the ANHIR challenge website (as of Jan 18th, 2022). We have released our source code and pre-trained models of the proposed method at https://github.com/wendy127green/SFG/tree/master/SFG.

Our contributions are summarized as follows:

An unsupervised structural feature guided CNN (SFG) for non-rigid histological image registration is proposed, which deals with the image registration task well without any other manually annotated data. It contains dense and sparse structural components.

[1]https://anhir.grand-challenge.org/evaluation/challenge/leaderboard/

The dense structural component contains the comprehensive structural features of the whole image, and the sparse structural component emphasizes the regions with significant structures. The combination of the two components makes full use of the global and local information of the histological image, which improves the performance of the network.

A multi-scale strategy is utilized in both dense and sparse structural components to overcome repetitive texture and section missing in histological images.

II. Related Work

In this section we introduce some work related to this paper.

A. Traditional Algorithms for Image Registration

There are many traditional image registration algorithms proposed for multiple stained histological images. We select typical ones and introduce them here.

The greedy diffeomorphic algorithm [13] firstly performs affine registration based on the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm and then obtains the finer results based on the greedy algorithm. The greedy algorithm is computed iteratively and utilizes normalized cross-correlation (NCC) as the similarity metric. It also introduces the smoothness of the deformation field into the iterative equation based on the insight segmentation and registration toolkit (ITK) recursive Gaussian smoothing classes. The authors participated in the ANHIR challenge as the UPENN team.

Some researchers complete the histological image registration task in multiple steps [14]. After pre-processing, they calculate the feature-based affine registration or exhaustive initial alignment. Specifically, they select the best result among the results estimated by the random sample consensus (RANSAC) algorithm using scale-invariant feature transform (SIFT), speeded up robust features (SURF), and oriented fast and rotated brief (ORB) features as the initial transformation. If all three results are bad, they use the exhaustive search method to obtain the initial transformation. Then an iterative operation is performed to refine the results. Finally, the deformation field is calculated based on the Demons algorithm using a modality independent neighborhood (MIND) descriptor. The authors also participated in the ANHIR challenge as the AGH team.

The method proposed in [32] focuses on the registration of the whole slide histological images. They firstly register the whole image at low resolution. After registering the image patches at high resolution, they combine the patches' deformation fields and form a global deformation field for the whole image.

B. Image Registration Algorithms Based on DL

In addition to these traditional methods, many algorithms based on DL have been proposed for histological image registration. The volume tweening network (VTN) [33] performs the registration task using CNN which is trained in an unsupervised manner. This network uses several cascaded

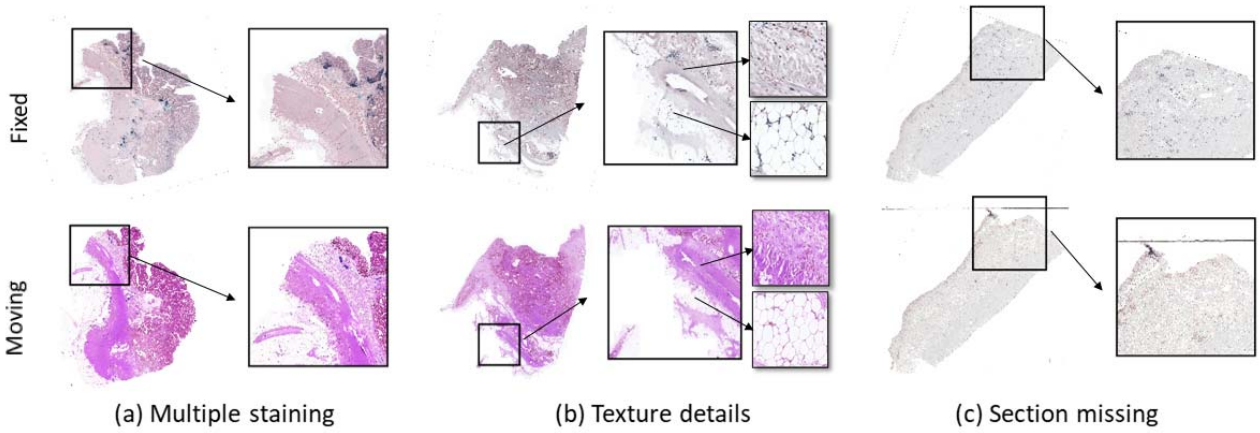

Fig. 1. Challenges of histological image registration. Examples of challenges of histological image registration compared with other medical image registration are presented: the appearance variance due to multiple staining (a), repetitive texture (b), and section missing (c). Corresponding regions of interest in the fixed images and moving images are enlarged for visualization. Repetitive textures are further enlarged in the black box in (b).

sub-networks to refine the deformation fields. The correlation coefficient is calculated as the similarity loss. In addition, the recursive cascaded networks proposed by the same team [34] utilize the similarity between the moving image and the warped image. The work also introduces an orthogonality loss, a total variation loss, and a determinant loss into the total loss function. This team also participated in the ANHIR challenge as the TUB team. They used a network similar to VTN to perform the registration. The network was trained in an unsupervised manner but fine-tuned using provided landmark positions. The registration results of TUB are in Table III in our manuscript.

There are also many weakly-supervised algorithms dealing with the image registration task [35]–[39]. Compared with supervised algorithms, weakly-supervised algorithms use labels that are easily accessible but are inexact or inaccurate to train the network. The structures and locations of the solid organs, ducts, vessels, and other ad hoc structures can be considered as the weakly-supervised labels. To be exact, the supervised DL algorithms mentioned in this paper should be classified into weakly-supervised algorithms, because the number of manually annotated key points used in the training process is limited. For simplicity, we uniformly call these DL algorithms using manually annotated data as supervised algorithms.

C. Image Registration Algorithms Based on Dense Structural Features

For image registration algorithms, an essential step is feature extraction. Commonly used features in an image include sparse structural features (feature maps of points) and dense structural features (feature maps of images or patches).

There are also many algorithms based on dense structural features. The study in [40] extracts feature maps at different scales to train the network which performs well in the autism brain imaging data exchange (ABIDE) data set. The algorithm proposed in [41] utilizes the SIFT features in elastic image registration and achieves good results. Studies in [42] achieve automatic image registration based on the SIFT features. In addition, many other algorithms [43]–[45] deal with image registration based on SIFT features. These studies demonstrate that dense structural features play an important role in image registration.

D. Image Registration Algorithms Based on Sparse Structural Features

Many algorithms based on sparse structural features are proposed to match two point-sets, which consist of many key points of the moving and fixed images respectively, by optimizing multifarious distance functions. A deep neural network architecture called free point transformer (FPT) proposed in [46] utilizes unordered and unstructured point pairs and has been validated in magnetic resonance image registration and ultrasound image registration. A point descriptor calculated by the triangle-area representation (TAR) of the K nearest neighbors (KNN-TAR) is used in [47] to perform image registration. Other studies [48]–[50] also utilize sparse structural features in image registration. There is also a review about sparse structural features in medical image registration [51], highlighting the value and advantage of sparse structural features in image registration.

This paper combines both dense and sparse structural features to improve registration efficiency.

III. Methods

The pipeline of the proposed unsupervised SFG network for non-rigid histological image registration is shown in Fig. 2. Image preprocessing and affine registration are performed before non-rigid registration. The details are described in Section IV-C.1. The moving and fixed images after preprocessing and affine registration are the inputs to our SFG network and the deformation field is the output. The deformation field is optimized by increasing the structural consistency between the fixed image and the warped image. In SFG, we propose two components of structural consistency constraints according to the formations of structural features: dense structural component and sparse structural component. The former utilizes pixel-level feature maps and the latter utilizes key points which are automatically matched by SIFT

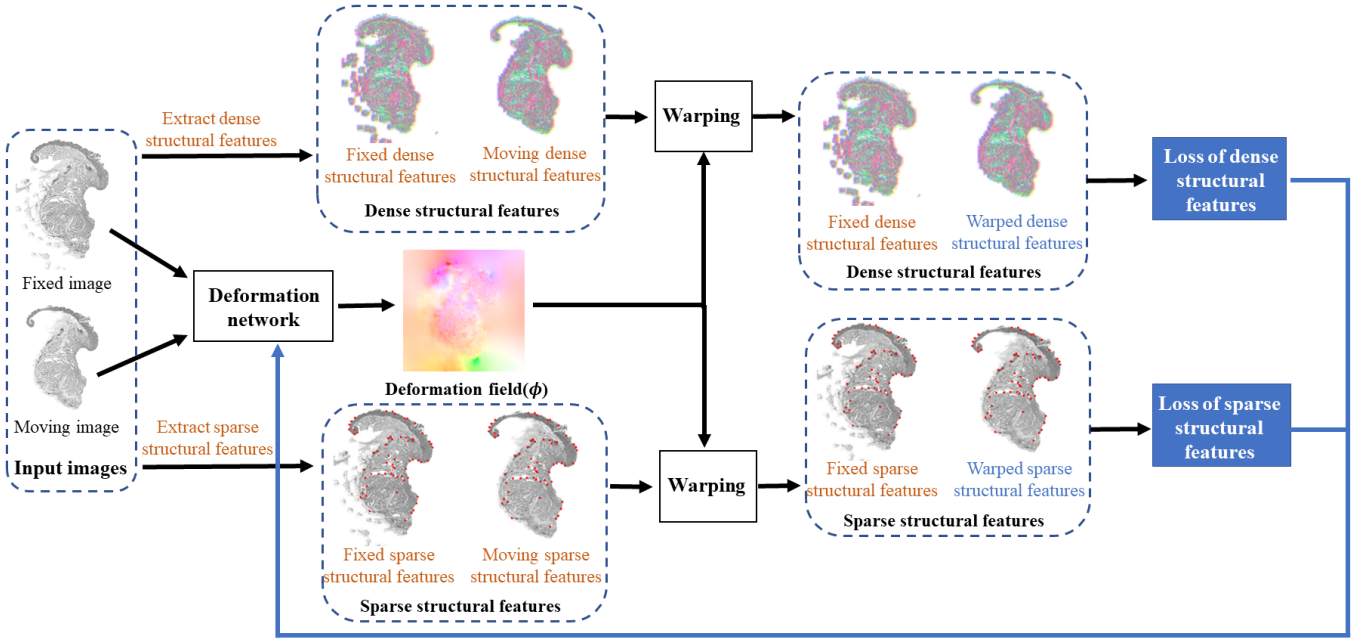

Fig. 2. Pipeline of the proposed unsupervised SFG network for non-rigid image registration. The fixed and moving images are the input to the deformation network and the deformation field is the output. The deformation field is optimized by increasing the structural consistency between the fixed image and the warped image. The structural consistency constraints consist of dense structural component and sparse structural component. The former utilizes pixel-level feature maps and the latter utilizes key points which are automatically matched by SIFT descriptors. Therefore, the dense structural features and sparse structural features of the moving image and the fixed image need to be extracted. Then the moving dense structural features and the moving sparse structural features are warped using the deformation field to obtain the warped dense structural features and the warped sparse structural features. The structural consistency between the warped dense and sparse structural features and the fixed dense and sparse structural features respectively is calculated as a part of the loss function to update the network.

descriptors. Therefore, the dense structural features and sparse structural features of the moving image and the fixed image need to be extracted. Then the moving dense structural features and the moving sparse structural features are warped using the deformation field to obtain the warped dense structural features and the warped sparse structural features. Finally, the structural consistency between the warped dense and sparse structural features and the fixed dense and sparse structural features respectively is calculated as a part of the loss function to update the network.

A. Baseline

1) Architecture: In this paper, we use a CNN similar to PWC-Net [52] and MaskFlownet-S [53] as our backbone. The architecture is shown in Fig. 3. Firstly, the feature maps of the moving and fixed images are extracted by a learnable feature pyramid with 6 levels. From the first to the sixth levels, the numbers of feature channels are respectively 16, 32, 64, 96, 128, and 196. The higher the level, the larger the feature channels. In each level of the pyramid, the feature map of the moving image is deformed using the $\times 2$ upsampled flow from the adjacent higher level to obtain the deformed feature map. Then the correlation between the feature map of the fixed image and the deformed feature map is calculated and sent to an optical flow estimator to estimate the corresponding deformation field. Finally, the moving image is warped according to the output of the network, i.e., the $\times 4$ upsampled flow from the second level, to obtain the warped image. Then the similarity between the warped and fixed images is calculated as the main part of the loss function.

TABLE I
HYPERPARAMETERS OF THE NETWORK

Convolution Layer		Input Channels	Output Channels	Kernel	Stride	Padding	Dilation
Conv1	Conv1a	3	16	3×3	2	1	1
	Conv1b	16	16	3×3	1	1	1
	Conv1c	16	16	3×3	1	1	1
Conv2	Conv2a	16	32	3×3	2	1	1
	Conv2b	32	32	3×3	1	1	1
	Conv2c	32	32	3×3	1	1	1
Conv3	Conv3a	32	64	3×3	2	1	1
	Conv3b	64	64	3×3	1	1	1
	Conv3c	64	64	3×3	1	1	1
Conv4	Conv4a	64	96	3×3	2	1	1
	Conv4b	96	96	3×3	1	1	1
	Conv4c	96	96	3×3	1	1	1
Conv5	Conv5a	96	128	3×3	2	1	1
	Conv5b	128	128	3×3	1	1	1
	Conv5c	128	128	3×3	1	1	1
Conv6	Conv6a	128	196	3×3	2	1	1
	Conv6b	196	196	3×3	1	1	1
	Conv6c	196	196	3×3	1	1	1

The hyperparameters of the network's architecture can be seen in Table I.

It is worth mentioning that the warping operation is based on STN [54], which includes localization network, grid generator, and sampler. The localization network produces the transformation parameters and the grid generator generates the sampling grid, which represents the fixed feature map's coordinates in the moving feature map. The sampler is differentiable and able to back-propagate gradients with the help of bilinear interpolation, which makes it possible to update the network.

2) Loss Function: In the CNN for unsupervised registration, the loss function typically contains two parts as in (2) [17], [18], [55], [56]. The first part is the similarity such as the

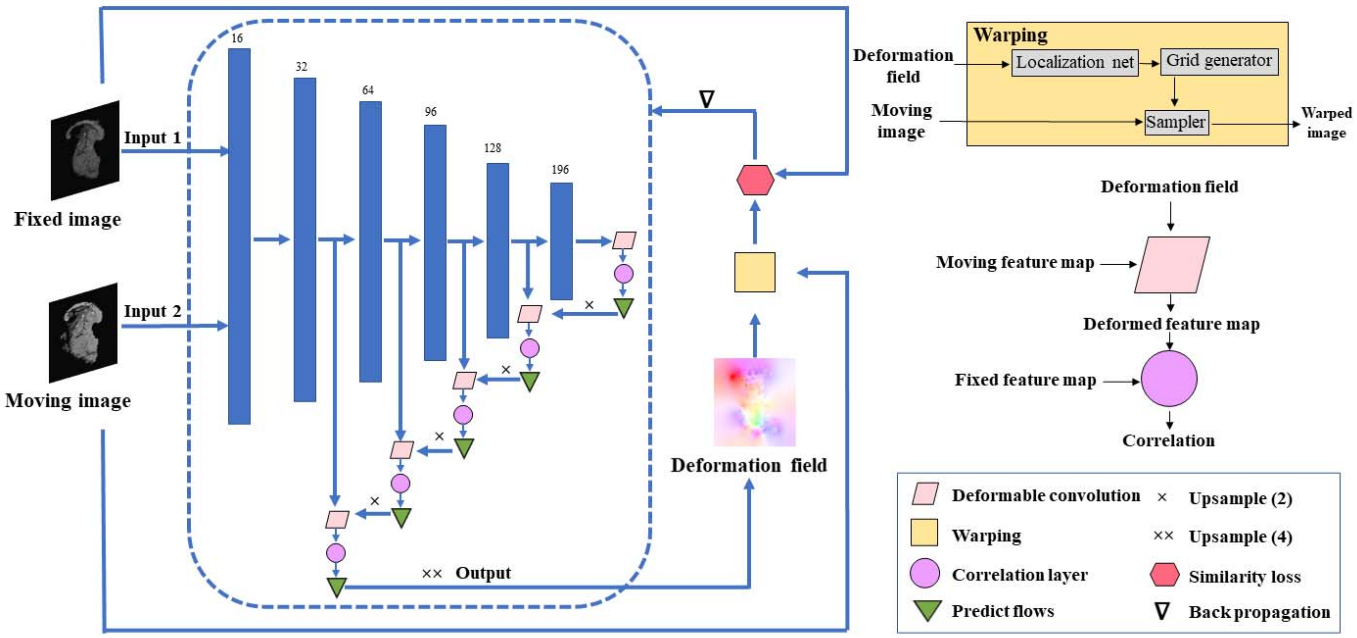

Fig. 3. The architecture of our network. The backbone of the deep network is similar to PWC-Net and MaskFlownet-S. A pyramidal structure (outlined by the blue boxes) is used to extract feature maps. From the first to the sixth levels, the numbers of feature channels are respectively 16, 32, 64, 96, 128, and 196. The higher the level, the larger the number of feature channels. In each level, the feature map of the moving image is deformed (outlined by the pink parallelograms) using the $\times 2$ upsampled (denoted by $\times$) flow from the adjacent higher level to obtain the deformed feature map. The correlation between the deformed feature map and the feature map of the fixed image is calculated (outlined by the purple circles) and sent to an optical flow estimator (outlined by the green triangles) to estimate the corresponding deformation field. Finally, the moving image is warped (outlined by the yellow rectangles) according to the output of the network, i.e., the $\times 4$ upsampled (denoted by $\times\times$) flow from the second level, to obtain the warped image. Then the similarity between the warped and fixed images is calculated (outlined by the red diamond) as the main part of the loss function. The warping operation is based on STN.

correlation between the warped image and the fixed image (3). The second part is the smoothness of the deformation field (4).

$$I'_m = \phi \circ I_m \tag{1}$$

$$L_I = L_c + \alpha_1 L_{reg} \tag{2}$$

$$L_c = 1 - \frac{mean((I_f - \overline{I_f}) \cdot (I'_m - \overline{I'_m}))}{\sqrt{mean((I_f - \overline{I_f})^2)} \cdot \sqrt{mean((I'_m - \overline{I'_m})^2)}} \tag{3}$$

$$L_{reg} = mean(flow_x^2) + mean(flow_y^2) \tag{4}$$

In the above equations, $\circ$ denotes the warping process, and ϕ denotes the flow field which warps the moving image I_m to obtain the warped image I'_m. L_c denotes the correlation between the input fixed image I_f and I'_m. L_I is the typical loss function for registration, with the intensity similarity as the constraint. The operator of $mean()$ is to obtain the average value. $\overline{I_f}$ and $\overline{I'_m}$ denote the averages of I_f and I'_m, respectively. L_{reg} denotes the regularization of the flow field with regularization parameter α_1. $flow_x$ and $flow_y$ stand for the gradient of the flow field along the x and y axes, respectively.

In this paper, we use the loss function in (2) in our baseline method. The components and corresponding regularization parameters of the loss function are also shown in Table II.

B. Dense Structural Component

Dense structural component utilizes the similarity between the pixel-level structural feature maps of a pair of images. Dense structural feature maps can present the local alignment information. Local consistency constraints of dense structural feature maps are better than intensities or gradient maps in histological image registration. One reason is that structural feature maps are robust to staining images. The other reason is that the anatomical structures are important for diagnosis and should be retained in histological images [27]–[30].

1) Structural Feature Maps: Structure-based descriptors such as SIFT descriptors [57], shape context descriptors [58], [59], histogram of oriented gradients (HOG) descriptors [60] are proved to be effective in many structure-based applications. In this paper, SIFT descriptors [57] are used as the feature descriptors. The neighborhood of each pixel is divided into a 4×4 cell array, and the orientation is quantized into 8 bins in each cell. A 128-dimension vector is then obtained as the SIFT representation for each pixel [61]. In this way, we obtain the structural feature maps based on SIFT descriptors, which characterize local image structures and encode contextual information. The top three principal components of a structural feature map are computed and visualized in the RGB space (Fig. 4(b)). It is illustrated that corresponding structures in the moving image and the fixed image are similar on structural feature maps.

2) Multi-Scale Structural Feature Maps: A structural feature map at a single scale, which is called as single-scale feature map in this paper, can only describe structure information at a certain resolution. Due to the repetitive textures and section

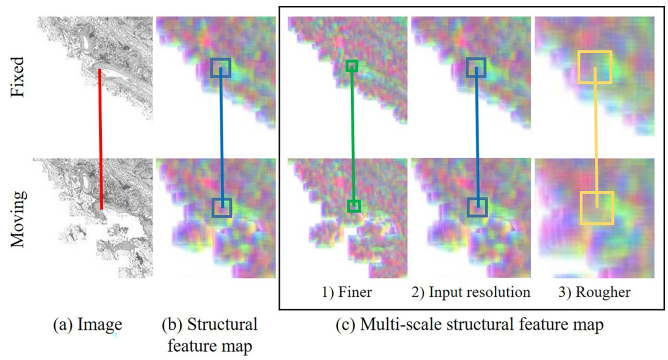

Fig. 4. Illustration of dense structural feature maps. (a) Gradient images. (b) The top three principal components of the structural feature map. The pixels with similar colors imply that they share similar local image structures. (c) The top three principal components of the multi-scale structural feature map. Structural feature descriptors for one pixel at different scales describe the neighborhoods with different sizes.

missing in histological images, structure information at rough levels and fine levels is both necessary. The structural features in one pixel at different scales describe the neighborhoods of different sizes as shown in Fig. 4(c). Structural features at finer scale describe more detailed textural information, which distinguishes adjacent pixels. Structural features at rougher scale describe a relatively larger neighborhood, which distinguishes repetitive texture at different regions. Therefore, we introduce the multi-scale feature maps in our network to integrate the local contextual information at different scales [61]. We downsample histological images to different scales and calculate the structural feature map at every single scale. Then these structural feature maps at different scales are weighted and concatenated together to obtain the multi-scale feature map. The multi-scale feature map can be expressed as:

$$F^{multi} = [w^1 F^1, \ldots, w^s F^s] \tag{5}$$

where F^i stands for the structural feature map obtained at the i-th scale. w^i is the weight of the feature map of the i-th scale, and s is the number of scales.

3) Constraints of Dense Structural Features: The single-scale or multi-scale dense structural feature maps of the moving image F_m are warped according to the deformation field obtained by the neural network as in (6), then we obtain the warped dense structural feature map F'_m. The correlation between the warped dense structural feature map and the dense structural feature map of fixed image F_f is used in the similarity component of the loss function in (7). The similarity component of the loss function is abbreviated as L_{fc}^{single} if a single-scale dense structural feature map is used and is abbreviated as L_{fc}^{multi} if a multi-scale dense structural feature map is used.

$$F'_m = \phi \circ F_m \tag{6}$$

$$L_{fc} = 1 - \frac{mean((F_f - \overline{F_f}) \cdot (F'_m - \overline{F'_m}))}{\sqrt{mean((F_f - \overline{F_f})^2)} \cdot \sqrt{mean((F'_m - \overline{F'_m})^2)}} \tag{7}$$

where F_m and F_f denote the structural feature maps of the moving and fixed images, respectively.

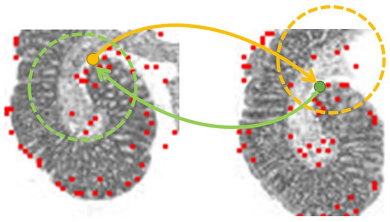

Fig. 5. Cycle match and search region restriction. Matched key points are selected if they are the best match points for each other (yellow point and green point) in the search region (green circle and yellow circle). Search region takes the coordinate of the key point as the center. Red points are automatically extracted key points.

C. Sparse Structural Component

We observe that histological images have many significant large structures related to anatomical structures [31]. Automatically obtained key points can annotate these significant structures [62]. Key point alignment can push the network to emphasize corresponding regions in a pair of images. Besides, there is section missing during making tissue sections. Therefore, automatically matched key points are utilized in the sparse structural component. It is worth mentioning that the deformation field is applied to the coordinates of key points on the moving image to get close to the coordinates of corresponding key points on the fixed images [63]. The distances of the warped key points on the fixed image and the key points on the moving image are used as the sparse structural constraints.

1) Automatically Obtaining Matching Key Points: It takes two steps to obtain matching key points: 1. Key points extraction on each image; 2. Key points matching from a pair of images. ORB [64] was used to extract key points on histological images. To match key points in a pair of histological images, we utilize several algorithms such as search region restriction, cycle match [65], and non-maximum suppression (NMS) [66]. Details are illustrated as follows:

For each key point on the fixed image (p_1, yellow points in Fig. 5), the search region is a circular region on the moving image with the coordinate of p_1 as the center (yellow circle in Fig. 5). We calculate the SIFT descriptor's similarity between p_1 and each key point in the moving image within the search region (p_r, red points in the yellow circle). Then the structural feature similarities between p_1 and all the p_r are ranked. If the maximum similarity is lower than the threshold of similarity or the second maximum similarity is nearly the same as the maximum similarity, p_1 is abandoned. To boost matching accuracy, we also utilize cycle match. A pair of key points are selected only if they are the best matching key point for each other, as the yellow point and the green point in the Fig. 5. In the last step of obtaining matched key points, we utilize NMS to avoid key points clustering, which leads to an overemphasis on some structures while ignoring the others.

2) Multi-Scale Strategy: We also use multi-scale structural features to combine rough and fine structural descriptors in key point matching. Rough structural information describes a larger neighborhood, which distinguishes points on repetitive textures at different regions. Details to distinguish adjacent pixels are obtained from finer structural information. Obtaining

TABLE II
COMPONENTS AND CORRESPONDING REGULARIZATION PARAMETERS (I.E., THE WEIGHT VALUES) OF LOSS FUNCTIONS

Method	Loss Function	L_c	L_{fc}	L_{kp}	L_{kp_manual}	L_{reg}
Baseline	L_I	1				5
Supervised baseline	L_S	1			200	5
D-SFG	L_F		1			5
S-SFG	L_K	1		500		5
SFG	L_{FK}		1	500		5
Supervised SFG	L_{SS}		1	500	200	5

key points' structural similarities on only one scale is called a single-scale matching algorithm. To improve the single-scale matching algorithm, we utilize the multi-scale structural similarity matching algorithm (abbreviated as multi-scale matching algorithm). In the multi-scale matching algorithm, the structural similarity on each scale is weighted with corresponding weight.

3) Constraints of Sparse Structural Features: With matching key points, we compute the new sparse point-distance loss function component L_{kp}, as follows:

$$L_{kp} = \frac{mean(\| \hat{p}_l^m - p_l^m \|_2)}{d_m} \tag{8}$$

where $\hat{p}_l^m$ is the coordinate of the l-th key point on the fixed images after warping, p_l^m is the coordinate of the corresponding key point on the moving images. The average of the Euclidean distance between $\hat{p}_l^m$ and p_l^m is normalized by d_m, which is the length of the image diagonal, and m is the index of the moving image.

D. Framework of SFG

To improve the performance of unsupervised registration, we replace the intensity similarity loss in the baseline with structural constraints. The loss function L_{fc} (7) in dense structural component utilizes comprehensive structural features all over the image. The loss function L_{kp} (8) in sparse structural component emphasizes matching significant structures. As shown in Table II, the proposed method uses both components and the loss function is abbreviated as L_{FK}, which can be expressed as:

$$L_{FK} = L_{fc} + \alpha_2 L_{reg} + \beta_2 L_{kp} \tag{9}$$

where α_2 and β_2 denote the regularization parameters.

To demonstrate the performance of dense and sparse structural components respectively, we also performed experiments with only dense structural component or sparse structural component, named respectively as D-SFG and S-SFG. Their loss functions are respectively abbreviated as L_F and L_K, as follows:

$$L_F = L_{fc} + \alpha_3 L_{reg} \tag{10}$$

$$L_K = L_c + \alpha_4 L_{reg} + \beta_4 L_{kp} \tag{11}$$

where α_3, α_4, and β_4 denote the regularization parameters. These two loss functions are also shown in Table II. In particular, the loss functions of L_F, L_K and L_{FK} are abbreviated as L_F^{single}, L_K^{single} and L_{FK}^{single} if single-scale structural features are used. They are abbreviated as L_F^{multi}, L_K^{multi} and L_{FK}^{multi} if multi-scale structural features are used.

IV. EXPERIMENTS

A. Dataset

We evaluated our algorithm on the dataset of the ANHIR challenge [16]. The ANHIR dataset contains 8 sub-datasets: lung lesion, lung lobe, mammary gland, mouse kidney, colon adenocarcinoma (COAD), gastric adenocarcinoma, human breast, and human kidney. This dataset includes 49 sets in total, with 3 to 9 images per set. There are 355 images with 18 different stains. The total number of image pairs is 481, 230 pairs for training and 251 pairs for evaluation, respectively. In this dataset, the key points representing significant structures are annotated in the images. These key points, also named landmarks, took about 250 hours for 9 annotators to annotate manually. The landmarks on the training image pairs are available, while the landmarks on the evaluation image pairs are not. Because the number of submissions to the ANHIR website is limited, we invited an expert to manually annotate 6 pairs of landmarks on each evaluation image pair for analyzing our methods. Except for the results in Table III, all other results are evaluated using 6 manually annotated landmarks per image unless specially noted.

B. Evaluation

To quantitatively evaluate the performance of different methods, the relative target registration error of landmarks (rTRE) for each pair of images (i, j) is used as the evaluation metric [16].

$$rTRE_l^{ij} = \frac{\left\| \hat{x}_l^j - x_l^j \right\|_2}{d_j}. \tag{12}$$

where $\hat{x}_l^j$ and x_l^j denote the l-th matching landmarks on the warped image and the fixed image, respectively. The Euclidean distance between $\hat{x}_l^j$ and x_l^j is normalized by d_j, which is the length of the image diagonal.

Furthermore, at the landmark level, the median, average, and maximum of all rTRE values in an image pair are calculated as median rTRE (MrTRE), average rTRE (ArTRE), and maximum rTRE (MxrTRE), respectively. At the case level, there is also aggregation by the median or the average. In total, the metrics are median-median rTRE (MMrTRE), average-median rTRE (AMrTRE), median-average rTRE (MArTRE), average-average rTRE (AArTRE), median-maximum rTRE (MMxrTRE) and average-maximum rTRE (AMxrTRE). All these evaluation metrics are the same as those in the ANHIR challenge [16]. Besides, robustness is evaluated using the relative number of successfully registered landmarks [16]. The average time, average rank of median rTRE (ARMrTRE), and average rank maximum rTRE (ARMxrTRE) [16] are also reported in Table III. The average time in Table III represents the time for each registration, including the time for data loading and writing the output files [16]. In detail, our methods take 0.02 min to achieve preprocessing and registration. Preprocessing is described in Section IV-C.1, including converting RGB images to grayscale, obtaining the gradient images, downsampling the gradient images, and affine registration. Registration refers

TABLE III

QUANTITATIVE COMPARISON OF THE PROPOSED SFG AND OTHER METHODS ON THE EVALUATION DATA USING THE ANHIR EVALUATION SYSTEM. THE EVALUATION METRICS ARE THE SAME AS THAT IN [16]. THE SIZE OF INPUT IMAGES IS 1024 × 1024 PIXELS FOR OUR METHODS. FOR DETAILED INFORMATION ABOUT ROW 6, PLEASE REFER TO [17]. FOR DETAILED INFORMATION ABOUT ROWS 7 TO 20, PLEASE REFER TO [16], BECAUSE THIS TABLE IS AN EXTENDED VERSION OF THE TABLE PRESENTED IN THE ANHIR SUMMARY ARTICLE [16]. 'TIME' REPRESENTS THE TIME FOR EACH REGISTRATION, INCLUDING THE TIME FOR DATA LOADING AND PREPROCESSING DESCRIBED IN SECTION IV-C.1. * = METHODS SUBMITTED BY OURSELVES. ** = THE CORRESPONDING METHOD IS PERFORMED ON GPU

Method	Average rTRE		Median rTRE		Max rTRE		Robustness		Median rTRE Average Rank	Max rTRE Average Rank	Average time[min]
	Average AArTRE	Median MArTRE	Average AMrTRE	Median MMrTRE	Average AMaxrTRE	Median MMaxrTRE	Average	Median	ARMrTRE	ARMxrTRE	
SFG (ours*)	0.0081	**0.0024**	0.0069	**0.0016**	0.0284	**0.0172**	0.9874	**1.0000**			**0.02****
D-SFG (ours*)	0.0083	0.0026	0.0071	**0.0016**	0.0291	0.0181	0.9864	**1.0000**			**0.02****
S-SFG (ours*)	0.0082	0.0025	0.0070	0.0017	0.0283	**0.0172**	0.9868	**1.0000**			**0.02****
Supervised baseline (ours*)	0.0082	0.0026	0.0070	0.0017	0.0287	0.0184	**0.9898**	**1.0000**			**0.02****
Baseline (ours*)	0.0089	0.0028	0.0075	0.0018	0.0312	0.0201	0.9844	**1.0000**			**0.02****
DeepHistReg	0.0061	0.0033	0.0047	0.0019	0.0276	0.0224	0.9799	**1.0000**			0.03**
Initial	0.1340	0.0684	0.1354	0.0665	0.2338	0.1157					
MEVIS	0.0044	0.0027	0.0029	0.0018	0.0251	0.0188	0.9880	**1.0000**	2.84	5.04	0.17
AGH	0.0053	0.0032	0.0036	0.0019	0.0283	0.0225	0.9821	**1.0000**	3.42	6.00	6.55
UPENN	**0.0042**	0.0029	0.0029	0.0019	**0.0239**	0.0190	**0.9898**	**1.0000**	3.47	4.29	1.60
CKVST	0.0043	0.0032	**0.0027**	0.0023	**0.0239**	0.0189	0.9883	**1.0000**	4.41	5.27	7.80
TUB	0.0089	0.0029	0.0078	0.0021	0.0280	0.0178	0.9845	**1.0000**	4.53	3.81	**0.02****
TUNI	0.0064	0.0031	0.0048	0.0021	0.0287	0.0204	0.9823	**1.0000**	5.32	5.80	9.73
DROP	0.0861	0.0042	0.0867	0.0028	0.1644	0.0273	0.8825	0.9892	7.06	7.43	3.99
ANTs	0.0991	0.0072	0.0992	0.0058	0.1861	0.0351	0.7889	0.9714	9.23	7.79	48.24
RVSS	0.0472	0.0063	0.0448	0.0046	0.1048	0.0275	0.8155	0.9928	9.65	8.42	5.25
bUnwarpJ	0.1097	0.0290	0.1105	0.0260	0.1995	0.0727	0.7899	0.9310	9.67	9.37	10.57
Elastix	0.0964	0.0074	0.0956	0.0054	0.1857	0.0353	0.8477	0.9722	10.04	8.88	3.50
UA	0.0536	0.0100	0.0506	0.0082	0.1124	0.0353	0.8209	0.9853	10.28	8.83	1.70
NiftyReg	0.1120	0.0372	0.1136	0.0355	0.2010	0.0714	0.7427	0.8519	11.08	10.08	0.14

to the model inference. Both preprocessing and registration include the time for data loading and transferring to GPU. All the average times in Table III are normalized by a JSON file provided by the ANHIR organizers with benchmark results on computer performance. However, the times in other tables are not normalized, because they are used for our ablation studies.

As mentioned in Section IV-A, our expert annotated 6 pairs of landmarks per evaluation image pair for evaluations in ablation studies. As there are only 6 manually annotated landmarks per image, we choose ArTRE as the evaluation metric instead of MrTRE for comparison. At the case level, the median (MArTRE) is good for describing the quality of registration, and the average (AArTRE) can be used to verify potential outliers [16], [17]. Therefore, both are used as evaluation metrics.

C. Implementation Details

Our work focuses on non-rigid registration. After preprocessing the histological images and affine registration, we performed non-rigid histological image registration.

1) Preprocessing: In preprocessing, RGB images were converted to grayscale. To keep the detailed structural information such as cells' locations, a Sobel descriptor with a kernel size of 7 was used to obtain the gradient images. After being smoothed with a median filter with a kernel size of 5, the gradient images were padded and downsampled as the input images of our networks. Affine registration was performed using the affine registration network introduced in the recursive cascaded networks [55]. In our study, the size of input images was 512 × 512 pixels unless specifically mentioned.

2) Pre-Training: In non-rigid registration, because the dataset of histological images is relatively small, we pre-trained our network with a large synthetic FlyingChair dataset [67]. There are 22872 image pairs in the FlyingChair dataset. We split them into 22232 training and 640 test samples. The size of the image is 512 × 384 pixels. The details for the FlyingChair dataset can be seen in [67]. The loss function used in the pre-training process was the same as that used in PWC-Net [52].

3) Comparison Network: As mentioned in Section III-D, we also performed experiments with only dense structural component or sparse structural component, named respectively as D-SFG and S-SFG to verify the impact of each component. The methods of the baseline, SFG, D-SFG, and S-SFG are all trained in an unsupervised manner. The components of their loss functions can be seen in Table II.

Finally, we designed a supervised baseline and a supervised SFG for comparison which share the same network architecture as the proposed unsupervised methods. Landmarks of the training image pairs provided by ANHIR are added in the two supervised methods in the training process with the same role as key points in S-SFG. The loss function of the supervised method can be respectively expressed as:

$$L_S = L_c + \alpha_5 L_{reg} + \beta_5 L_{kp\_manual} \tag{13}$$

$$L_{SS} = L_{fc} + \alpha_6 L_{reg} + \beta_6 L_{kp} + \gamma_6 L_{kp\_manual} \tag{14}$$

where α_5, α_6, β_5, β_6, and γ_6 denote the regularization parameters. These two loss functions can also be seen in Table II. Actually, these two supervised algorithms can also be classified as weakly-supervised algorithms in the field of image registration to some degree. In this paper, we uniformly call them supervised algorithms as mentioned in Section II-A.

4) Setting: The training and evaluation were implemented using Mxnet and Python 3.6. All computer operations were performed on a server with 4 cards of 12G NVIDIA TITAN V GPU and 2 cards of Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz CPU. The training stage ran 100 epochs and the

batch size was 2. The learning rate is set to 1.0×10^{-4} in the first 1.2×10^4 steps and after more than 1.2×10^4 steps, it is halved every 2000 steps. The regularization parameters in the proposed methods (as shown in Tables II and VII) were optimized carefully. We observed the loss values of different components of the loss function and found a large difference among them. Therefore, we optimized the regularization parameters with the ratio between the values of different components of the loss function in mind. The hyperparameters of the network (as shown in Table I) are same as that in PWC-Net [52] and MaskFlownet-S [53].

5) Submission: We submitted our results to the ANHIR website with the team name of 'SFG'. We obtained the evaluation results of the evaluation dataset from the evaluation system, which can be used to compare our results with other methods in a fair way [16]. It is worth mentioning that the results of SFG, S-SFG, D-SFG, and the baseline are all submitted to the website. Only the best method (i.e., SFG) is public on the leaderboard, which ranked 1st (as of Jan 18th, 2022). Another member of our team also submitted the result of the supervised baseline method as a comparison, which ranked the 4th in the leaderboard (as of Jan 18th, 2022). Other results can be seen in Table III, which shows the quantitative results on the evaluation data. We also provide quantitative results on both training and evaluation data in the Appendix.

D. Comparison With Other Methods

1) Comparison With Published Methods: Quantitative results of the algorithms that participated in the ANHIR challenge are shown in Table III. This table is an extended version of the table presented in the ANHIR summary article [16]. Therefore, more detailed information about Rows 7 to 20 in the table can be found in [16]. In addition, DeepHistReg [17] is a recently proposed DL method based on the ANHIR dataset and Row 6 in Table III are from [17]. It is worth mentioning that Rows 7 to 20 in Table III are sorted according to ARMrTRE while Rows 1 to 6 are sorted according to MMrTRE. It is mainly because MMrTRE is used to identify the registration quality of algorithms while ARMrTRE is obtained with a special ranking rule of the ANHIR challenge and is not possible to compute without the help of the ANHIR organizers.

The proposed SFG outperforms all the other methods presented in Table III in terms of MMrTRE and MArTRE on the evaluation dataset. As mentioned in Section IV-B, MMrTRE describes the quality of registration, and MArTRE verifies potential outliers [16]. In Table III, DROP, ANTs, RVSS, bUnwarpJ, Elastix, UA, and NiftyReg are the published traditional methods which were run by the organizers of the ANHIR challenge [16]. Their AArTRE and AMrTRE are tens of times larger than those of our method. It is evident that these traditional registration methods are not suitable for histological images. DeepHistReg, MEVIS, AGH, UPENN, CKVST, TUB, and TUNI are the methods that participated in the ANHIR challenge [16], [17]. Most of them used very similar and rather classical techniques and consisted of carefully designed multi-step pipelines. Only two methods (DeepHistReg and TUB) used the deep learning algorithm. In general, these methods obtain a smaller rTRE and higher robustness than the traditional methods. The proposed SFG method ranked first in the ANHIR challenge according to their special ranking rule depending on ARMrTRE and ARMxrTRE [16] by the time of our paper submission (as of Jan 18th, 2022). It achieves the lowest Median rTRE among all the submissions to the ANHIR challenge website. SFG also has good performance on the robustness metrics and achieves the lowest normalized processing time. As shown in Table III, our methods take 0.02 min to deal with an image pair, which is the same as that of TUB and shorter than that of DeepHistReg. As for other traditional algorithms, they are run on CPU. Therefore, the time cannot be directly compared.

TABLE IV
ABLATION STUDY OF SFG. 'AFFINE' IS FOR RESULTS AFTER AFFINE REGISTRATION AND BEFORE NON-RIGID REGISTRATION. 'SUPERVISED BASELINE' IS FOR RESULTS OF SUPERVISED BASELINE METHOD WITH LANDMARKS PROVIDED BY ANHIR. 'SUPERVISED SFG' IS FOR RESULTS OF SUPERVISED VERSION OF SFG. 'TIME' STANDS FOR THE TIME FOR REGISTERING AN IMAGE PAIR (I.E., MODEL INFERENCE) ON OUR DEVICE. THE SIZE OF INPUT IMAGES IS 512×512 PIXELS. THE RESULTS ARE OBTAINED USING 6 MANUALLY ANNOTATED LANDMARKS PER IMAGE INSTEAD OF THE EVALUATION DATA PROVIDED BY THE ANHIR CHALLENGE. STANDARD DEVIATIONS ACROSS THE CASES ARE IN PARENTHESES. THERE IS NO STANDARD DEVIATION FOR MARTRE BECAUSE MARTRE REPRESENTS THE MEDIAN VALUE

Method	MArTRE	AArTRE	Time (s)
Affine	0.00834	0.01060 (0.0112)	**0.008** (**0.012**)
Baseline	0.00251	0.00340 (0.00310)	0.016 (0.014)
Supervised baseline	0.00241	0.00312 (0.00241)	0.016 (0.014)
D-SFG	0.00231	0.00305 (0.00244)	0.016 (0.014)
S-SFG	0.00242	0.00321 (0.00282)	0.016 (0.014)
SFG	**0.00223**	0.00303 (0.00246)	0.016 (0.014)
Supervised SFG	0.00226	**0.00296** (**0.00239**)	0.016 (0.014)

2) Comparison on Different Tissues: We also compare the MMrTRE of different tissue types between different methods. Because all the data of lung lobes, lung lesions, and mammary glands are used for network training, we provide two figures here: Fig. 7 shows the results of all the datasets, i.e., the evaluation and the training dataset, while Fig. 8 shows the results of the evaluation datasets only. The corresponding values are shown in the Appendix. It is expected that the MMrTRE for DL algorithms in Fig. 7 is lower than the corresponding results in Fig. 8, because the network is trained on the training dataset and performs better on the training dataset than the evaluation dataset. For the evaluation dataset, our method achieves the lowest MMrTRE when dealing with the samples of colon adenocarcinoma (0.00198). The MMrTRE (0.00171) of our method for all the tissue types is close to the lowest one (0.00170). For the whole dataset, our method achieves the lowest MMrTRE when dealing with the samples of colon adenocarcinomas (0.00139), breasts (0.00128), and human kidneys (0.00163), and the second lowest MMrTRE

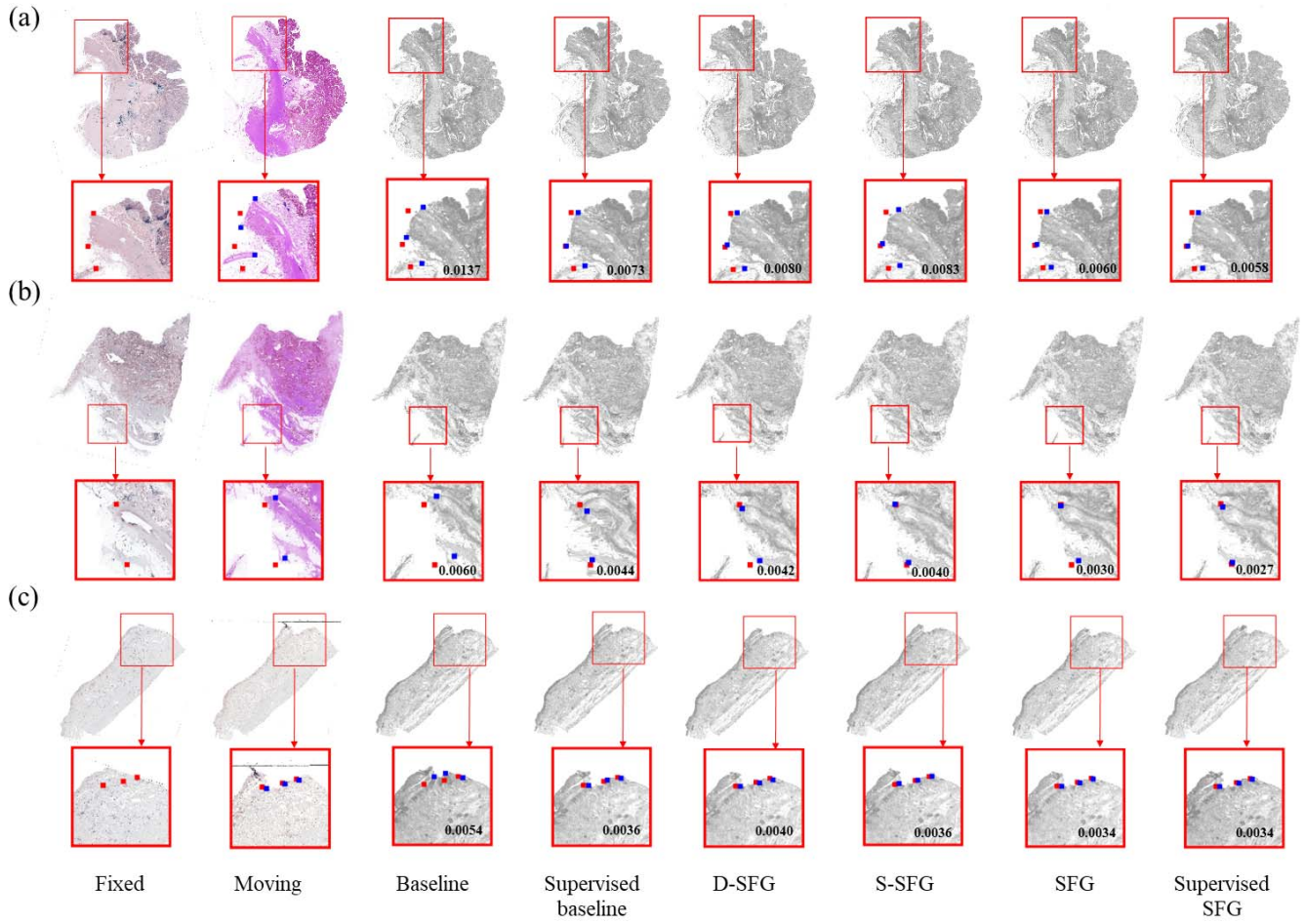

Fig. 6. Performance of different methods on three typical samples: (a) different staining, (b) repetitive texture, and (c) section missing. Supervised baseline, supervised SFG, SFG, D-SFG, and S-SFG all have improvements compared with the baseline method. SFG obtains the best performance except for supervised SFG, but the results of SFG are close to that of supervised SFG. The red points refer to specific structures in the fixed image. The blue points indicate the corresponding structures in the moving image (Column 2) or the warped image (Columns 3 to 8). The closer the red points and corresponding blue points, the better the registration. The numbers at the right bottom are the average rTREs on the corresponding whole images.

for mice kidneys' samples (0.00110). The MMrTRE for all types of tissues is the lowest (0.00098).

3) Comparison With Baseline and Supervised Methods: In addition to the proposed SFG, D-SFG, and S-SFG, we also submitted the results of the baseline and the supervised baseline method to the ANHIR evaluation system as a comparison (Table III). SFG performs significantly better than the baseline in terms of rTRE. D-SFG and S-SFG also have an improvement in rTRE compared with the baseline. The supervised baseline and S-SFG both utilize the sparse structural consistency constraints. The former uses the manually annotated landmarks while the latter uses the automatically extracted matched key points as sparse structural features. Their results are similar, which indicates that our automatic matched key point extraction algorithm performs well. The supervised baseline has better robustness than SFG. It is probably because the landmarks on the evaluation dataset are similar to those on the training dataset and thus less wrong deformation exists in the supervised baseline. Compared with the supervised baseline, SFG obtains better registration performances with additional dense structural consistency constraints. In addition, the supervised baseline method performs better than S-SFG, mainly because the supervised baseline method uses manually annotated landmarks while S-SFG uses automatically obtained landmarks, which may not be as accurate as the manually annotated landmarks. However, the results of S-SFG are only slightly worse than those of the supervised baseline method. It is mainly because we considered carefully the accuracy of the automatically obtained landmarks and adopted multiple strategies to assure the accuracy. The size of input images for the results submitted to the ANHIR evaluation system is 1024×1024 pixels. The results are evaluated in the ANHIR evaluation system which contains tens of landmarks on every image.

In ablation studies (Tables IV-X, and Figs. 6-12), we use the 6 pairs of landmarks annotated by our expert for evaluation and the size of input images is 512×512 pixels. The quantitative results are presented in Table IV. The improvement of SFG, D-SFG, and S-SFG in rTRE is significant compared with the baseline. The rTRE of SFG is smaller than that of the supervised baseline. The AArTRE of SFG is larger than that of the supervised SFG but the MArTRE of SFG is

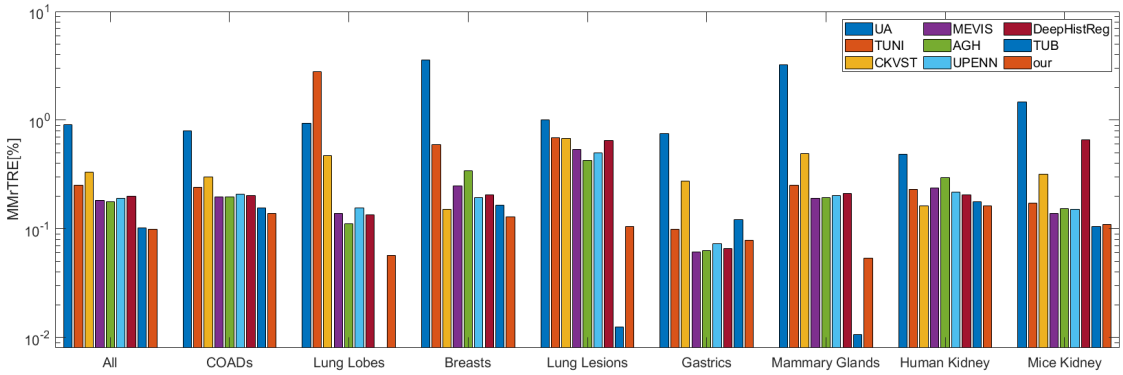

Fig. 7. The MMrTRE of different methods on both training dataset and evaluation dataset for each tissue type separately.

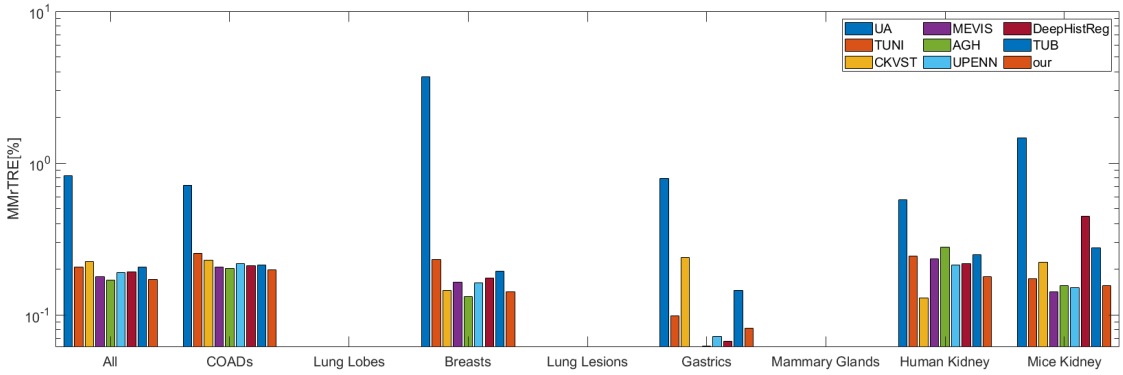

Fig. 8. The MMrTRE of different methods on evaluation dataset for each tissue type separately. All the images for lungs and mammary glands are used as training samples, so the results for these three datasets are missing.

smaller than that of the supervised SFG, demonstrating that the automatically extracted key points are accurate enough. In addition, the time for D-SFG and SFG is longer because of the high dimension of structural feature maps.

Performances of baseline, supervised baseline, D-SFG, S-SFG, SFG, and supervised SFG on three typical samples (same as that in Fig. 1) with different staining, repetitive texture, and section missing respectively are visualized in Fig. 6. The regions of interest are magnified and the key points are annotated via red and blue colors for comparison. The closer the corresponding red points and the blue ones, the better the alignment. The numbers at the bottom right of the images are the values of ArTREs for the whole images. The points and numbers have the same meaning in the following figures. SFG, D-SFG, and S-SFG all achieve a smaller ArTRE than the baseline, while SFG achieves a smaller ArTRE than the supervised baseline. In addition, SFG achieves a similar ArTRE to that of supervised SFG.

In Fig. 6 (a), the contrast between the bottom left and top right of the fixed image is higher than that of the moving image. The baseline method which uses the intensity similarity as the constraint, cannot obtain a good alignment. In SFG, we match the structural features instead of intensities and overcome the appearance variance. From Fig. 6 (b), similar gridding textures from the adipose tissue in different regions interrupt the alignment. SFG distinguishes different regions with rough structural features and achieves better performance than the baseline. From Fig. 6 (c), there is a section missing in the moving image due to the tearing in the tissue section preparation. There is also a slight section missing on the top right of the fixed image. SFG solves the problem of section missing by combining fine structural features at the dense level and rough structural features at the sparse level. In the baseline method, the moving image is directly deformed according to the contour of the fixed image and the deformation is unrealistic (Fig. 6 (c)). SFG achieves a more realistic deformation than the baseline due to structural consistency.

TABLE V
STANDARD DEVIATION OF JACOBIAN DETERMINANTS AND FRACTION OF FOLDING FOR DIFFERENT METHODS ON THE EVALUATION DATASET. AREAS WITH NEGATIVE JACOBIAN DETERMINANT ARE CONSIDERED AS FOLDING. STANDARD DEVIATIONS ACROSS THE CASES ARE IN PARENTHESES

Method	Std. Jacobian	Folding(%)
Baseline	0.099 (0.038)	0.020 (0.080)
Supervised baseline	0.125 (0.064)	0.087 (0.142)
D-SFG	0.104 (0.038)	0.028 (0.145)
S-SFG	0.107 (**0.037**)	0.011 (**0.030**)
SFG	**0.090** (0.039)	**0.007** (0.035)
Supervised SFG	0.123 (0.049)	0.058 (0.123)

Furthermore, we study the smoothness metrics [68] of the proposed methods. The quantitative results are shown in Table V. SFG achieves the best deformation smoothness with the lowest standard deviation of Jacobi determinants (0.090) and the lowest fraction of folding (0.007%). The visualization of Jacobi determinant can be seen in Fig. 9.

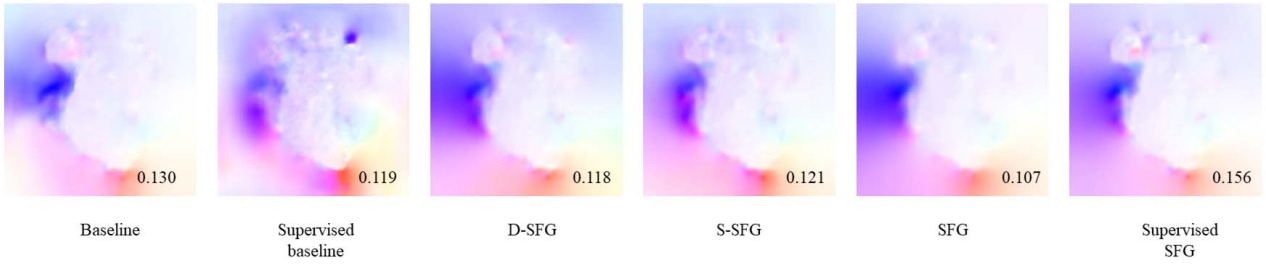

Fig. 9. Visualization of Jacobi determinant for different methods. The numbers at the right bottom are the standard deviations of Jacobian determinants of the corresponding flow images. The lower the standard deviations of Jacobian determinants, the better the deformation smoothness. SFG achieves the lowest standard deviation of Jacobi determinants.

E. Influence Factors of SFG

1) Impact of Dense Structural Component: Dense structural component utilizes comprehensive structural features at pixel-level. It restrains the unrealistic deformation on the whole image. We investigate its impact via the comparison of SFG and S-SFG in Fig. 6. With the incorporation of dense structural component, SFG performs better than S-SFG, especially in some regions without enough strong and conspicuous structures.

In addition, S-SFG is similar to those algorithms based on point-sets, as mentioned in Section II-D. Therefore, this experiment also demonstrates that SFG performs better than point-sets algorithms.

2) Impact of Sparse Structural Component: Sparse structural component emphasizes matching significant structures. It plays an important role, especially in regions where structures are significant. It is worth mentioning that the sparse point-distance loss used in sparse structural component is similar to the registration evaluation metric rTRE. We investigate the impact of sparse structural component via the comparison of SFG and D-SFG in Fig. 6. The results show that SFG performs better than D-SFG with the help of sparse structural component.

Besides, if time or memory needs to be saved in experiments, S-SFG is recommended. S-SFG is the fastest in Table IV. The GPU memory usages for multi-scale D-SFG, single-scale D-SFG, S-SFG, and the baseline are 5967, 3551, 2281, and 2277 MiB, respectively.

In conclusion, dense structural component and sparse structural component utilize local and global information of the histological image, respectively. Therefore, the combination of these two components can achieve the best performance in most cases.

3) Comparison of Multi-Scale and Single-Scale Structural Features: As described in Section III, both dense and sparse structural components use a multi-scale strategy to obtain more robust and accurate results. We perform the comparison of multi-scale and single-scale structural features in Table VI. Multi-scale strategy performs better than the single-scale strategy in SFG, S-SFG, and D-SFG. Comparison of multi-scale and single-scale strategies is also visualized on typical challenging samples (same as that in Fig. 1) in Fig. 10. Multi-scale SFG obtains a smaller ArTRE than single-scale SFG in typical samples with multiple staining (Fig. 10 (a)), repetitive textures (Fig. 10 (b)), and missing sections (Fig. 10 (c)).

TABLE VI
COMPARISON OF MULTI-SCALE AND SINGLE-SCALE STRATEGIES ON SFG. 'TIME' STANDS FOR REGISTERING AN IMAGE PAIR (I.E., MODEL INFERENCE) ON OUR DEVICE. THE SIZE OF INPUT IMAGES IS 512 × 512 PIXELS. THE RESULTS ARE OBTAINED USING 6 MANUALLY ANNOTATED LANDMARKS PER IMAGE INSTEAD OF THE EVALUATION DATA PROVIDED BY THE ANHIR CHALLENGE. STANDARD DEVIATIONS ACROSS THE CASES ARE IN PARENTHESES. THERE IS NO STANDARD DEVIATION FOR MARTRE BECAUSE MARTRE REPRESENTS THE MEDIAN VALUE

Method	MArTRE	AArTRE	Time (s)
Single-scale D-SFG	0.00231	0.00311 (**0.00224**)	0.016 (0.014)
Multi-scale D-SFG	**0.00230**	**0.00305** (0.00244)	0.016 (0.014)
Single-scale S-SFG	0.00245	0.00326 (0.00285)	0.016 (0.014)
Multi-scale S-SFG	**0.00242**	**0.00321** (**0.00282**)	0.016 (0.014)
Single-scale SFG	0.00227	0.00309 (**0.00198**)	0.016 (0.014)
Multi-scale SFG	**0.00223**	**0.00303** (0.00246)	0.016 (0.014)

4) Effect of Combining Similarity of Structural Feature Maps and Intensity Similarity: Compared with the loss function of the baseline, the intensity similarity L_c was replaced by the similarity of structural feature maps L_{fc} in the previous experiments of D-SFG and SFG. We explore the effect of combining L_c and L_{fc} in D-SFG and SFG in Table VII with the same weight. The MArTREs do not change, while the AArTREs slightly decrease. There is unrealistic deformation in the warped images to obtain aligned contours like that of the baseline method (Fig. 6) in some regions. To avoid unrealistic deformation, we replaced L_c with L_{fc} instead of adding L_{fc} in the loss functions of D-SFG and SFG.

5) Effect of Size of Input Images: We explore the influence of the size of input images. As shown in Table VIII, the registration errors of all the proposed methods are smaller when the size of input images is larger. But the time of training and the memory consumption also increase with the size of input images. Taking the baseline as an example, the time of training on our device with image sizes of 256 × 256, 512 × 512, and 1024 × 1024 pixels is 3.2, 24, and 96 hours, respectively. The time of test for all the evaluation data in the ANHIR dataset is 12.73, 15.60, and 30.41 seconds for the image sizes of 256 × 256, 512 × 512, and 1024 × 1024 pixels respectively. Considering the time cost, the image size of our ablation study is 512 × 512 pixels. The image size of the results submitted to the ANHIR challenge is 1024 × 1024 pixels to obtain the best performance.

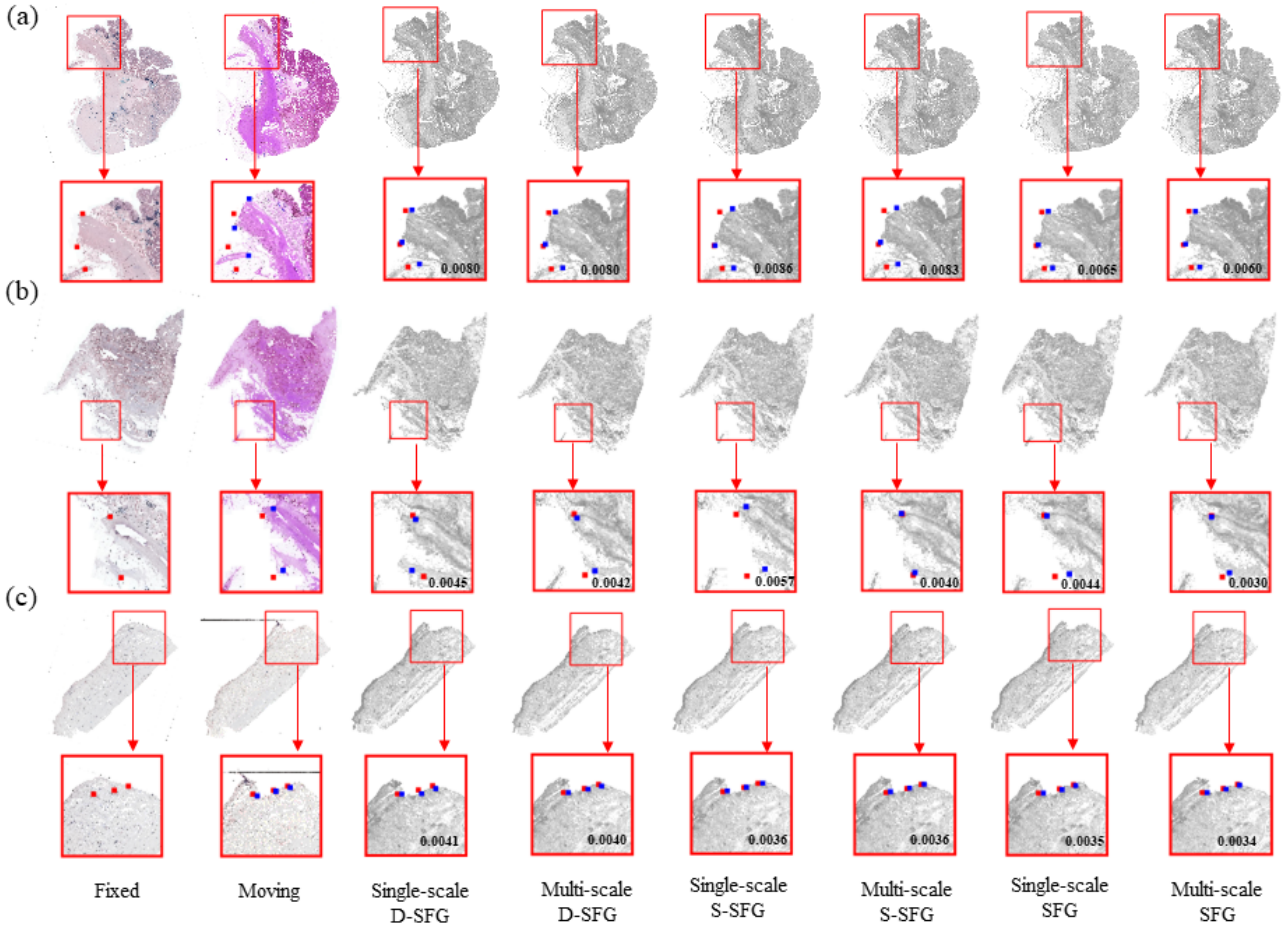

Fig. 10. Comparison of multi-scale and single-scale strategies on typical samples of different staining (a), repetitive texture (b), and missing section (c). The red points refer to specific structures in the fixed image. The blue points indicate the corresponding structures in the moving image (Column 2) or the warped image (Columns 3-8). The closer the red points and corresponding blue points, the better the registration. The numbers at the right bottom are the average rTREs of the corresponding samples. Multi-scale SFG performs better than single-scale SFG on all three samples.

TABLE VII

EFFECT OF COMBINING THE SIMILARITY OF STRUCTURAL FEATURE MAPS AND INTENSITY SIMILARITY IN MULTI-SCALE AND SINGLE-SCALE D-SFG AND SFG. COMPONENTS AND CORRESPONDING REGULARIZATION PARAMETERS (I.E., THE WEIGHT VALUES) OF LOSS FUNCTIONS ARE ALSO DISPLAYED. THE SIZE OF INPUT IMAGES IS 512 × 512 PIXELS. THE RESULTS ARE OBTAINED USING 6 MANUALLY ANNOTATED LANDMARKS PER IMAGE INSTEAD OF THE EVALUATION DATA PROVIDED BY THE ANHIR CHALLENGE. STANDARD DEVIATIONS ACROSS THE CASES ARE IN PARENTHESES. THERE IS NO STANDARD DEVIATION FOR MArTRE BECAUSE MArTRE REPRESENTS THE MEDIAN VALUE

Method	L_c	L_{fc}^{single}	L_{fc}^{multi}	L_{kp}^{single}	L_{kp}^{multi}	L_{reg}	MArTRE	AArTRE
Single-scale D-SFG		1				5	0.00231	0.00311 (**0.00224**)
Single-scale D-SFG+Intensity similarity	1	1				5	0.00231	**0.00306** (0.00237)
Multi-scale D-SFG			1			5	0.00230	0.00305 (0.00244)
Multi-scale D-SFG+Intensity similarity	1		1			5	0.00230	**0.00304** (**0.00231**)
Single-scale SFG		1		500		5	0.00227	0.00309 (**0.00198**)
Single-scale SFG+Intensity similarity	1	1		500		5	0.00227	**0.00307** (0.00216)
Multi-scale SFG			1		500	5	0.00223	**0.00303** (0.00246)
Multi-scale SFG+Intensity similarity	1		1		500	5	0.00223	**0.00303** (**0.00242**)

6) Effect of Receptive Field: We also explore whether adjusting the receptive field of the network can obtain better performance. We adjust the rate of downsampling for the first pyramid level, which consists of two feature extraction layers with a stride of 2. For the methods with the image size of 1024 × 1024, the original receptive field is small, causing the huge time-consuming. Therefore, we increase the rate of downsampling from 4 to 8, which means that one of the first two feature extraction layers needs to have a stride of 4 and the other has a stride of 2. We select the first and second feature extraction layers to have a stride of 4 and 2, respectively. We also increase the rate of downsampling from 4 to 16, i.e., the strides of the first two feature extraction layers are increased from 2 to 4 simultaneously. For the methods with an image size of 256 × 256, the original receptive field is so large that local details are suppressed greatly. Therefore,

TABLE VIII

INFLUENCE OF THE SIZE OF INPUT IMAGES. AVERAGE RTRE RESULTS WITH DIFFERENT SIZES OF INPUT IMAGES AND DIFFERENT METHODS ARE COMPARED. 'TIME' STANDS FOR THE AVERAGE TIME FOR REGISTERING AN IMAGE PAIR (I.E., MODEL INFERENCE) ON OUR DEVICE. THE RESULTS ARE OBTAINED USING 6 MANUALLY ANNOTATED LANDMARKS PER IMAGE INSTEAD OF THE EVALUATION DATA PROVIDED BY THE ANHIR CHALLENGE. STANDARD DEVIATIONS ACROSS THE CASES ARE IN PARENTHESES

Image Size	Baseline L_I	Single-scale D-SFG L_F^{single}	Multi-scale D-SFG L_F^{multi}	Single-scale S-SFG L_K^{single}	Multi-scale S-SFG L_K^{multi}	Single-scale SFG L_{FK}^{single}	Multi-scale SFG L_{FK}^{multi}	Time (s)
256	0.00410 (0.00300)	0.00378 (0.00261)	0.00370 (0.00270)	0.00391 (0.00274)	0.00385 (0.00249)	0.00372 (0.00277)	0.00368 (0.00284)	**0.013** (**0.012**)
512	0.00340 (0.00310)	0.00311 (**0.00224**)	0.00305 (**0.00244**)	0.00326 (0.00285)	0.00321 (0.00282)	0.00309 (**0.00198**)	0.00303 (0.00246)	0.016 (0.014)
1024	**0.00312** (**0.00287**)	**0.00299** (0.00258)	**0.00291** (0.00251)	**0.00287** (**0.00216**)	**0.00280** (**0.00203**)	**0.00279** (0.00202)	**0.00272** (**0.00213**)	0.030 (0.018)

TABLE IX

INFLUENCE OF THE RECEPTIVE FIELD. AVERAGE RTRE RESULTS WITH DIFFERENT SIZES OF INPUT IMAGES, DIFFERENT STRIDE FOR THE FIRST TWO FEATURE EXTRACTION LAYERS, AND DIFFERENT METHODS ARE COMPARED. 'STRIDE LAYER 1' AND 'STRIDE LAYER 2' STAND FOR THE STRIDES OF THE FIRST TWO FEATURE EXTRACTION LAYERS, RESPECTIVELY. 'TIME' STANDS FOR THE AVERAGE TIME FOR REGISTERING AN IMAGE PAIR (I.E., MODEL INFERENCE) ON OUR DEVICE. THE RESULTS ARE OBTAINED USING 6 MANUALLY ANNOTATED LANDMARKS PER IMAGE INSTEAD OF THE EVALUATION DATA PROVIDED BY THE ANHIR CHALLENGE. STANDARD DEVIATIONS ACROSS THE CASES ARE IN PARENTHESES.THE RESULTS WITH A IMAGE SIZE OF 512 × 512 ARE GIVEN HERE AS COMPARISON

Image Size	Stride Layer 1	Stride Layer 2	Baseline L_I	Single-scale D-SFG L_F^{single}	Multi-scale D-SFG L_F^{multi}	Single-scale S-SFG L_K^{single}	Multi-scale S-SFG L_K^{multi}	Single-scale SFG L_{FK}^{single}	Multi-scale SFG L_{FK}^{multi}	Time (s)
512	2	2	0.00340 (0.00310)	0.00311 (0.00224)	0.00305 (0.00244)	0.00326 (0.00285)	0.00321 (0.00282)	0.00309 (0.00198)	0.00303 (0.00246)	0.016 (0.014)
	1	1	**0.00359** (0.00301)	**0.00332** (**0.00243**)	**0.00325** (0.00256)	**0.00356** (0.00278)	**0.00354** (0.00276)	**0.00320** (**0.00236**)	**0.00316** (**0.00232**)	0.015 (0.016)
256	2	1	0.00365 (**0.00276**)	0.00337 (0.00285)	0.00335 (**0.00251**)	0.00361 (0.00283)	0.00357 (0.00284)	0.00332 (0.00255)	0.00329 (0.00265)	0.014 (0.013)
	2	2	0.00410 (0.00300)	0.00378 (0.00261)	0.00370 (0.00270)	0.00391 (**0.00274**)	0.00385 (**0.00249**)	0.00372 (0.00277)	0.00368 (0.00284)	**0.013** (**0.012**)
	2	2	**0.00312** (0.00287)	**0.00299** (**0.00258**)	**0.00291** (**0.00251**)	**0.00287** (**0.00216**)	**0.00280** (**0.00203**)	**0.00279** (**0.00202**)	**0.00272** (**0.00213**)	0.030 (0.018)
1024	4	2	0.00337 (**0.00273**)	0.00331 (0.00291)	0.00326 (0.00273)	0.00323 (0.00290)	0.00320 (0.00283)	0.00316 (0.00271)	0.00311 (0.00286)	0.027 (0.018)
	4	4	0.00392 (0.00313)	0.00387 (0.00305)	0.00381 (0.00313)	0.00378 (0.00321)	0.00375 (0.00310)	0.00372 (0.00290)	0.00370 (0.00290)	**0.023** (**0.016**)

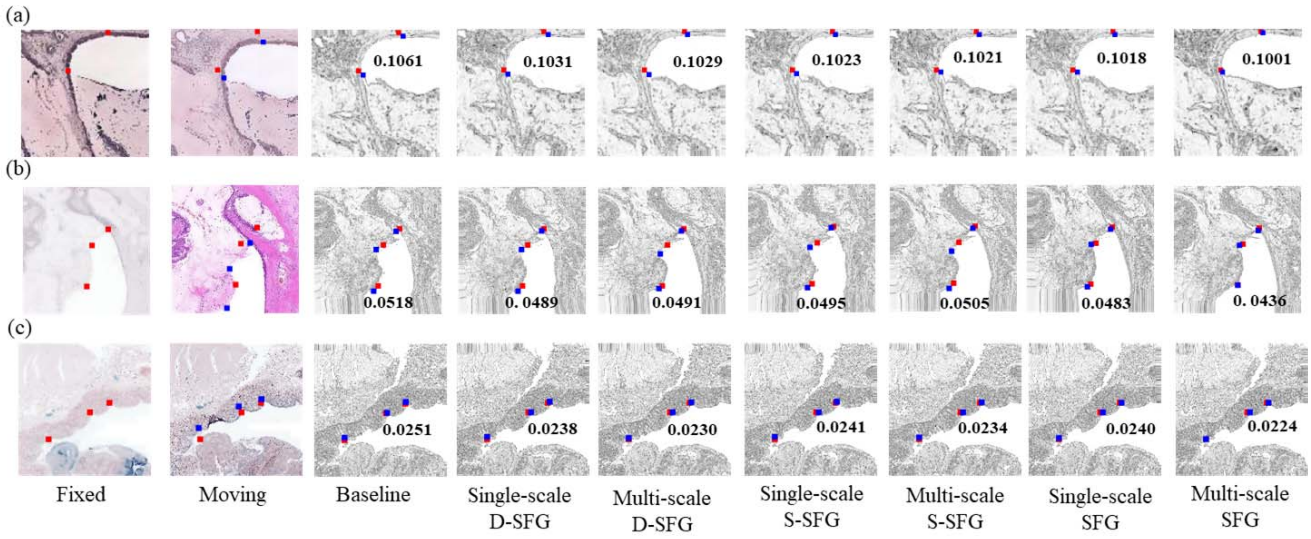

Fig. 11. Performance of all the proposed methods and the baseline on three patch-based datasets with patch size of 256 × 256 (a), 512 × 512 (b), and 1024 × 1024 pixels (c). The red points refer to specific structures in the fixed image. The blue points indicate the corresponding structures in the moving image (Column 2) or the warped image (Columns 3-8). The closer the red points and corresponding blue points, the better the registration. The numbers are the average rTREs on the corresponding whole images.

we decrease the rate of downsampling from 4 to 2, which means that one of the first two feature extraction layers needs to have a stride of 2 and the other has a stride of 1. We select the first and second feature extraction layers to have a stride of 2 and 1, respectively. We also decrease the rate of downsampling from 4 to 1, i.e., the strides of the first two feature extraction layers are decreased from 2 to 1 simultaneously. We repeated the proposed methods (i.e., baseline, single-scale S-SFG, multi-scale S-SFG, single-scale D-SFG, multi-scale D-SFG, single-scale SFG, and multi-scale SFG) and evaluated the average rTRE using 6 manually landmarks per image.

The quantitative results are shown in Table IX. For an image size of 256 × 256, decreasing the receptive field obtains smaller average rTRE at the cost of slightly longer time-consuming. When the strides for the first two layers are both 1, the average rTRE is close to that for an image size of 512 × 512. This result shows that decreasing the receptive field improves the network performance greatly. For an image size of 1024 × 1024, increasing the receptive field degrades the performance of the networks greatly. When the strides for the first two layers are both 4, the average rTRE is much higher than that for an image size of 512 × 512. When the strides for

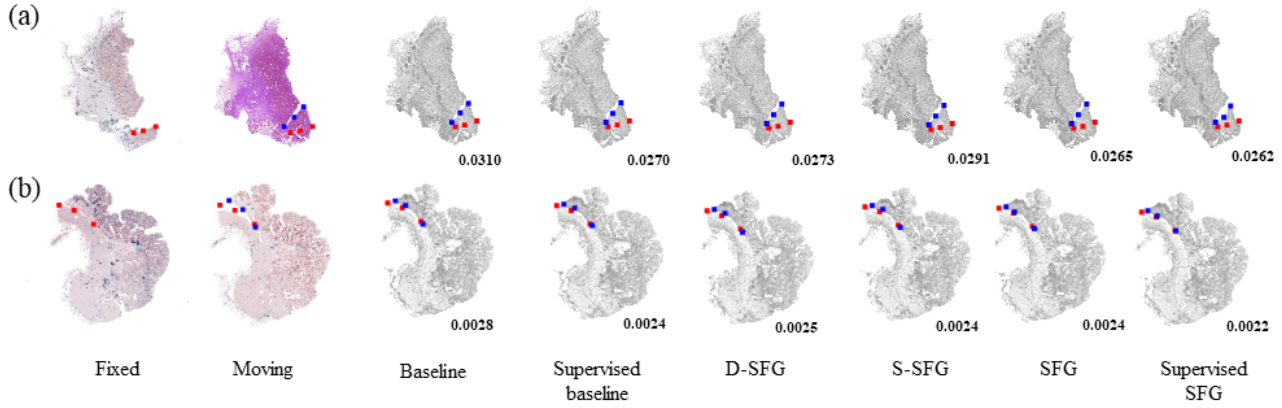

Fig. 12. Performance of SFG on image pairs with tearing regions. (a) SFG fails in registration of extremely tearing regions. (b) SFG performs well when the tearing region is relatively small. The red points refer to specific structures in the fixed image. The blue points indicate the corresponding structures in the moving image (Column 2) or the warped image (Columns 3-8). The closer the red points and corresponding blue points, the better the registration. The numbers at the right bottom are the average rTREs of the corresponding samples.

the first two layers are 4 and 2 respectively, the average rTREs of single-scale D-SFG, multi-scale D-SFG, single-scale SFG, and multi-scale SFG are higher than those for an image size of 512 × 512. Meanwhile, the time used for model inference reduces only slightly.

Therefore, when the image size is small, simply decreasing the receptive field can improve the performance of the networks. It is in line with expectations because the image size of 256 × 256 is so small that local details are limited. With the large receptive field of the original pyramid structure, local details are further suppressed. Decreasing the receptive field makes the network catch more helpful local information and obtain better performance. Meanwhile, the extra time-consuming caused by this operation is negligible because half of the model inference time is related to data loading. For large image sizes, simply increasing the receptive field reduces the performance of the network. It is reasonable because increasing the receptive field causes a reduction in local information. Meanwhile, the time used for model inference reduces slightly because half of the model inference time is related to data loading.

7) Effectiveness of Various Patch-Based Deformable Image Registration: We also explore the performance of the proposed methods on patch-based registration on the same dataset. We first resized the original images provided by ANHIR to 8000 × 8000 pixels and divided them into many patches with sizes of 1024 × 1024, 512 × 512, and 256 × 256 pixels, respectively. Then we repeated all the proposed methods (single-scale D-SFG, multi-scale D-SFG, single-scale S-SFG, multi-scale S-SFG, single-scale SFG, and multi-scale SFG) as well as the baseline on these three datasets.

The quantitative results can be seen in Table X. For different patch sizes, the proposed methods achieve lower MArTRE and AArTRE than the baseline. For patch size of 256 × 256 pixels, S-SFG performs better than D-SFG with lower MArTRE and AArTRE, while D-SFG performs better than S-SFG for patch sizes of 512 × 512 and 1024 × 1024 pixels. It is mainly because sparse structural component emphasizes global information while dense structural component emphasizes local information. The receptive field of 256 × 256 patches is so small that it is more difficult for D-SFG than for S-SFG to register the patched images according to the local information. The results also show that multi-scale strategy improves the performance of the network, because such strategy uses more information at different levels. This phenomenon is consistent with the conclusion of the image-based experiments. In addition, the time shown in Table X stands for the average time for registering a patched image pair on our device. The time used for registering the corresponding 8000 × 8000 image should be about 64 ×, 256×, and 1024× longer time for a patch size of 1024, 512, and 256, respectively. The performance of all the

TABLE X

PATCH-BASED STUDY OF THE PROPOSED METHODS. 'TIME' STANDS FOR THE AVERAGE TIME FOR REGISTERING A PATCHED IMAGE PAIR (I.E., MODEL INFERENCE) ON OUR DEVICE. THE RESULTS ARE OBTAINED USING 6 MANUALLY ANNOTATED LANDMARKS PER IMAGE INSTEAD OF THE EVALUATION DATA PROVIDED BY THE ANHIR CHALLENGE. STANDARD DEVIATIONS ACROSS THE CASES ARE IN PARENTHESES. THERE IS NO STANDARD DEVIATION FOR MARTRE BECAUSE MARTRE REPRESENTS THE MEDIAN VALUE

Patch Size	Method	MArTRE	AArTRE	Time (s)
256	Baseline	0.07295	0.10708 (0.09971)	0.013 (0.012)
	Single-scale D-SFG	0.07195	0.10405 (0.09722)	0.013 (0.011)
	Multi-scale D-SFG	0.06957	0.10338 (0.10047)	0.013 (**0.010**)
	Single-scale S-SFG	0.06862	0.10179 (0.09726)	0.013 (0.012)
	Multi-scale S-SFG	0.06562	0.09958 (**0.09377**)	0.013 (0.013)
	Single-scale SFG	0.06862	0.10179 (0.09647)	0.013 (0.012)
	Multi-scale SFG	**0.06477**	**0.09858** (0.09798)	0.013 (0.014)
512	Baseline	0.03087	0.05294 (0.06291)	0.016 (0.014)
	Single-scale D-SFG	0.02887	0.04998 (0.05738)	0.016 (0.012)
	multi-scale D-SFG	0.02692	0.04830 (0.05475)	0.016 (0.013)
	Single-scale S-SFG	0.03030	0.05166 (0.06259)	0.016 (**0.011**)
	multi-scale S-SFG	0.02823	0.05037 (0.05560)	0.016 (0.016)
	Single-scale SFG	0.02536	0.04578 (0.05359)	0.016 (0.014)
	multi-scale SFG	**0.02500**	**0.04168** (**0.05073**)	0.016 (0.014)
1024	Baseline	0.01407	0.02455 (**0.03467**)	0.030 (0.021)
	Single-scale D-SFG	0.01326	0.02436 (0.03733)	0.030 (**0.017**)
	multi-scale D-SFG	0.01280	0.02411 (0.03670)	0.030 (0.021)
	Single-scale S-SFG	0.01358	0.02454 (0.03761)	0.030 (0.019)
	multi-scale S-SFG	0.01336	0.02437 (0.03589)	0.030 (0.018)
	Single-scale SFG	0.01324	0.02418 (0.03744)	0.030 (0.018)
	multi-scale SFG	**0.01276**	**0.02378** (0.03707)	0.030 (0.019)

TABLE XI

QUANTITATIVE COMPARISON OF THE PROPOSED SFG AND OTHER METHODS ON BOTH THE TRAINING AND EVALUATION DATA USING THE ANHIR EVALUATION SYSTEM. THE EVALUATION METRICS ARE THE SAME AS THAT IN [16]. THE SIZE OF INPUT IMAGES IS 1024 × 1024 PIXELS FOR OUR METHODS. DATA IN ROW 6 IS OBTAINED FROM THE ANHIR WEBSITE (AS OF JAN 18TH, 2022). FOR DETAILED INFORMATION ABOUT ROWS 7 TO 20, PLEASE REFER TO [16], BECAUSE THIS TABLE IS AN EXTENDED VERSION OF THE TABLE PRESENTED IN THE ANHIR SUMMARY ARTICLE [16]. 'TIME' REPRESENTS THE TIME FOR EACH REGISTRATION, INCLUDING THE TIME FOR DATA LOADING AND PREPROCESSING DESCRIBED IN SECTION IV-C.1. * = METHODS SUBMITTED BY OURSELVES. ** = THE CORRESPONDING METHOD IS PERFORMED ON GPU

Method	Average rTRE		Median rTRE		Max rTRE		Robustness		Median rTRE	Max rTRE	Average
	Average AArTRE	Median MArTRE	Average AMrTRE	Median MMrTRE	Average AMaxrTRE	Median MMaxrTRE	Average	Median	Average Rank ARMrTRE	Rank ARMxrTRE	time[min]
D-SFG (ours*)	0.0046	**0.0011**	0.0038	**0.00067**	0.0186	0.0074	0.9928	**1.0000**			**0.02****
SFG (ours*)	0.0046	0.0012	0.0039	0.0008	0.0173	0.0076	**0.9931**	**1.0000**			**0.02****
S-SFG (ours*)	0.0048	0.0014	0.0040	0.0010	0.0192	0.0096	0.9923	**1.0000**			**0.02****
Supervised baseline (ours*)	0.0082	0.0026	0.0070	0.0017	0.0287	0.0184	0.9898	**1.0000**			**0.02****
Baseline (ours*)	0.0065	0.0028	0.0051	0.0018	0.0274	0.0181	0.9861	**1.0000**			**0.02****
DeepHistReg	0.0060	0.0033	0.0046	0.0020	0.0269	0.0203	0.9775	**1.0000**			0.03**
Initial	0.1340	0.0684	0.1354	0.0665	0.2338	0.1157					
TUB	0.0047	0.0012	0.0041	0.0010	**0.0149**	**0.0046**	0.9919	**1.0000**	2.84	2.47	**0.02****
MEVIS	0.0052	0.0029	0.0039	0.0018	0.0261	0.0186	0.9845	**1.0000**	2.98	4.83	0.15
UPENN	**0.0041**	0.0030	**0.0028**	0.0019	0.0230	0.0175	0.9888	**1.0000**	3.40	4.23	1.45
AGH	0.0056	0.0034	0.0038	0.0020	0.0300	0.0231	0.9770	**1.0000**	3.57	6.23	6.86
TUNI	0.0104	0.0037	0.0087	0.0025	0.0387	0.0234	0.8899	**1.0000**	5.99	6.37	10.32
CKVST	0.0060	0.0047	0.0046	0.0033	0.0261	0.0208	0.9730	**1.0000**	6.28	5.83	7.13
DROP	0.0616	0.0043	0.0613	0.0028	0.1230	0.0265	0.8861	0.9907	6.87	7.29	3.41
ANTs	0.0693	0.0087	0.0686	0.0067	0.1343	0.0359	0.8137	0.9718	9.04	7.84	43.09
RVSS	0.0471	0.0071	0.0450	0.0055	0.1032	0.0294	0.7958	0.9875	9.64	8.52	4.72
bUnwarpJ	0.0797	0.0256	0.0796	0.0246	0.1496	0.0652	0.7940	0.9310	9.65	9.57	9.15
Elastix	0.0695	0.0080	0.0684	0.0054	0.1371	0.0390	0.7668	0.9706	9.83	8.80	2.96
UA	0.0569	0.0110	0.0549	0.0090	0.1190	0.0360	0.8076	0.9737	10.14	8.91	1.47
NiftyReg	0.0825	0.0346	0.0828	0.0327	0.1514	0.0679	0.7495	0.8519	10.77	10.12	0.15

proposed methods on three patch-based datasets with different patch sizes is visualized in Fig. 11. It is worth mentioning that the edges of the warped images are blurred, which is reasonable because of the patch-slicing. A simple edge-cutting can deal with this problem.

8) Clinical Relevance: The patch-based experiment demonstrates that our methods can deal with images with high resolution and small field of view, which is valuable and necessary in the clinic. Actually, clinicians often enlarge the region of interest of the whole scale histological image to diagnose, because the size of computer screen is usually limited to display so much information on the whole scale image. Therefore, together with image-based experiments, our method can satisfy different demands of histological image registration, no matter for the whole scale or for the enlarged region of interest. We believe our methods can eventually provide assistance for clinicians in the future.

9) Reasons of Relatively High AArTRE and AMrTRE for SFG: As shown in Table III, the proposed SFG achieves the best MMrTRE and MArTRE but not the best AArTRE and AMrTRE. MMrTRE and MArTRE describe the quality of registration, and AArTRE and AMrTRE verify potential outliers at the case level. The AArTRE and AMrTRE of our method are not the best because some tissue sections are extremely torn during section staining and significantly destroy histological structure (Fig. 12). Landmarks on the extremely tearing regions cannot be matched well in our method because we keep the structure constraints (Fig. 12 (a)). But our method performs well on regions that are not torn so much (Fig. 12 (b)).

V. CONCLUSION

In this paper, we proposed an unsupervised structural feature guided CNN (SFG) for non-rigid registration of multiple stained histological images. Compared with supervised algorithms, it deals with image registration task well without any other manually annotated data. SFG contains two components to constrain the structural consistency between the fixed and the warped images in an unsupervised manner, i.e., dense structural component and sparse structural component. Dense structural component utilizes dense pairwise structural feature matching at pixel-level. Sparse structural component focuses on significant structural region matching. Multi-scale strategy is used in both dense and sparse structural components to learn rough and fine structural features simultaneously.

Experiments are conducted using the image-based and patch-based registrations on a public histological dataset. Our image-based experiments demonstrate that the proposed method SFG can deal with the whole scale image registration task. It overcomes the three challenges in histological image registration, i.e., multiple staining, repetitive texture, and section missing. The patch-based experiments demonstrate the good performance of our method on images with high resolution and small field of view. Therefore, our method can satisfy different demands of histological image registration, no matter the whole scale or the enlarged region of interest.

However, SFG cannot perform registration of images with extremely tearing regions because the structural feature is seriously damaged.

APPENDIX

An extended version of Table III that shows the quantitative results on both the training and evaluation data of different algorithms is provided as Table XI. The metrics in this table are the same as those in Table III. Different with Table III, data in Row 6 in Table XI are from the ANHIR website[2] (as of

[2]https://anhir.grand-challenge.org/evaluation/challenge/leaderboard/

TABLE XII
THE MMRTRE OF DIFFERENT METHODS ON BOTH TRAINING DATASET AND EVALUATION DATASET FOR EACH TISSUE TYPE SEPARATELY

Method	All	COADs	Lung Lobes	Breasts	Lung Lesions	Gastrics	Mammary-Glands	Human Kidneys	Mice Kidneys
UA	0.00904	0.00798	0.00930	0.03561	0.01011	0.00756	0.03236	0.00485	0.01468
TUNI	0.00250	0.00243	0.02772	0.00595	0.00693	0.00099	0.00253	0.00232	0.00171
CKVST	0.00331	0.00299	0.00470	0.00150	0.00683	0.00274	0.00492	**0.00163**	0.00319
MEVIS	0.00182	0.00196	0.00139	0.00249	0.00535	**0.00061**	0.00192	0.00236	0.00138
AGH	0.00179	0.00196	0.00111	0.00342	0.00423	0.00063	0.00194	0.00294	0.00153
UPENN	0.00192	0.00208	0.00156	0.00194	0.00498	0.00072	0.00203	0.00217	0.00151
DeepHistReg	0.00199	0.00202	0.00135	0.00206	0.00651	0.00066	0.00212	0.00207	0.00664
TUB	0.00102	0.00156	**0.00008**	0.00165	**0.00013**	0.00122	**0.00011**	0.00178	**0.00105**
our	**0.00098**	**0.00139**	0.00057	**0.00128**	0.00105	0.00078	0.00054	**0.00163**	0.00110

TABLE XIII
THE MMRTRE OF DIFFERENT METHODS ON EVALUATION DATASET FOR EACH TISSUE TYPE SEPARATELY

Method	All	COADs	Lung Lobes	Breasts	Lung Lesions	Gastrics	Mammary-Glands	Human Kidneys	Mice Kidneys
UA	0.00825	0.00714		0.03694		0.00793		0.00575	0.01459
TUNI	0.00206	0.00254		0.00233		0.00099		0.00243	0.00174
CKVST	0.00226	0.00230		0.00145		0.00238		**0.00130**	0.00223
MEVIS	0.00179	0.00207		0.00165		**0.00061**		0.00235	**0.00142**
AGH	**0.00170**	0.00203		**0.00132**		0.00062		0.00280	0.00156
UPENN	0.00190	0.00218		0.00162		0.00072		0.00214	0.00151
DeepHistReg	0.00192	0.00210		0.00174		0.00067		0.00217	0.00446
TUB	0.00207	0.00214		0.00195		0.00144		0.00249	0.00277
our	0.00171	**0.00198**		0.00142		0.00082		0.00179	0.00155

Jan 18th, 2022) because the authors did not public the results on both the training and evaluation data in [17].

Tables XII and XIII show the corresponding values of Figs. 7 and 8.

REFERENCES

[1] G. J. Metzger, S. C. Dankbar, J. Henriksen, A. E. Rizzardi, N. K. Rosener, and S. C. Schmechel, "Development of multigene expression signature maps at the protein level from digitized immunohistochemistry slides," *PLoS ONE*, vol. 7, no. 3, Mar. 2012, Art. no. e33520.

[2] O. Lobachev, C. Ulrich, B. S. Steiniger, V. Wilhelmi, V. Stachniss, and M. Guthe, "Feature-based multi-resolution registration of immunostained serial sections," *Med. Image Anal.*, vol. 35, pp. 288–302, Jan. 2017.

[3] A. J. Pierrot, E. E. Pujol, Y. Sharma, and M. D. Wang, "Autonomous point-based registration of prostate gland tissue images," in *Proc. 4th Int. Conf. Biomed. Eng. Informat. (BMEI)*, vol. 1, 2011, pp. 165–169.

[4] L. Solorzano, G. M. Almeida, B. Mesquita, D. Martins, C. Oliveira, and C. Wählby, "Whole slide image registration for the study of tumor heterogeneity," in *Computational Pathology and Ophthalmic Medical Image Analysis*. Cham, Switzerland: Springer, 2018, pp. 95–102.

[5] Y. Song, D. Treanor, A. Bulpitt, and D. Magee, "3D reconstruction of multiple stained histology images," *J. Pathol. Informat.*, vol. 4, no. 2, p. 7, 2013.

[6] I. Arganda-Carreras, C. O. Sorzano, R. Marabini, J. M. Carazo, C. Ortiz-de Solorzano, and J. Kybic, "Consistent and elastic registration of histological sections using vector-spline regularization," in *Proc. Int. Workshop Comput. Vis. Approaches Med. Image Anal.* Berlin, Germany: Springer, 2006, pp. 85–95.

[7] C.-W. Wang, E. Budiman Gosno, and Y.-S. Li, "Fully automatic and robust 3D registration of serial-section microscopic images," *Sci. Rep.*, vol. 5, no. 1, Dec. 2015, Art. no. 15051.

[8] D. F. G. Obando, A. Frafjord, I. Øynebråten, A. Corthay, J.-C. Olivo-Marin, and V. Meas-Yedid, "Multi-staining registration of large histology images," in *Proc. 14th Int. Symp. Biomed. Imag. (ISBI)*, 2017, pp. 345–348.

[9] J. Borovec, J. Kybic, M. Bušta, C. Ortiz-de Solórzano, and A. Munoz-Barrutia, "Registration of multiple stained histological sections," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, Dec. 2013, pp. 1034–1037.

[10] L. Gupta, B. M. Klinkhammer, P. Boor, D. Merhof, and M. Gadermayr, "Stain independent segmentation of whole slide images: A case study in renal histology," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1360–1364.

[11] A. du Bois d'Aische, M. D. Craene, X. Geets, V. Gregoire, B. Macq, and S. K. Warfield, "Efficient multi-modal dense field non-rigid registration: Alignment of histological and section images," *Med. Image Anal.*, vol. 9, no. 6, pp. 538–546, Dec. 2005.

[12] C. Ceritoglu, "Large deformation diffeomorphic metric mapping registration of reconstructed 3D histological section images and *in vivo* MR images," *Frontiers Hum. Neurosci.*, vol. 4, p. 43, Jan. 2010.

[13] L. Venet, S. Pati, M. D. Feldman, M. P. Nasrallah, P. Yushkevich, and S. Bakas, "Accurate and robust alignment of differently stained histologic images based on greedy diffeomorphic registration," *Appl. Sci.*, vol. 11, no. 4, p. 1892, Feb. 2021.

[14] M. Wodzinski and A. Skalski, "Multistep, automatic and nonrigid image registration method for histology samples acquired using multiple stains," *Phys. Med. Biol.*, vol. 45, Nov. 2020, Art. no. 025006.

[15] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[16] J. Borovec *et al.*, "ANHIR: Automatic non-rigid histological image registration challenge," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3042–3052, Oct. 2020.

[17] M. Wodzinski and H. Müller, "DeepHistReg: Unsupervised deep learning registration framework for differently stained histology samples," *Comput. Methods Programs Biomed.*, vol. 198, Jan. 2021, Art. no. 105799.

[18] M. Wodzinski and H. Müller, "Learning-based affine registration of histological images," in *Proc. Int. Workshop Biomed. Image Registration*. Cham, Switzerland: Springer, 2020, pp. 12–22.

[19] G. Wu, M. Kim, Q. Wang, B. C. Munsell, and D. Shen, "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1505–1516, Jul. 2016.

[20] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, pp. 977–1000, Oct. 2003.

[21] Y. Xu, J.-Y. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 964–971.

[22] C. Wachinger and N. Navab, "Structural image representation for image registration," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 95–100.

[23] K. Kayser, "Application of structural pattern recognition in histopathology," in *Syntactic Structural Pattern Recognition*. Berlin, Germany: Springer, 1988, pp. 115–135.

[24] M. Amintoosi, M. Fathy, and N. Mozayani, "Precise image registration with structural similarity error measurement applied to superresolution," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, pp. 1–7, Dec. 2009.

[25] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 1, pp. 121–130, Jan. 2021.

[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[27] G. H. B. Miranda, J. Barrera, E. G. Soares, and J. C. Felipe, "Structural analysis of histological images to aid diagnosis of cervical cancer," in *Proc. 25th SIBGRAPI Conf. Graph., Patterns Images*, 2012, pp. 316–323.

[28] R. Müller *et al.*, "Morphometric analysis of human bone biopsies: A quantitative structural comparison of histological sections and micro-computed tomography," *Bone*, vol. 23, no. 1, pp. 59–66, 1998.

[29] G. H. B. Miranda, E. G. Soares, J. Barrera, and J. C. Felipe, "Method to support diagnosis of cervical intraepithelial neoplasia (CIN) based on structural analysis of histological images," in *Proc. 25th Int. Symp. Computer-Based Med. Syst. (CBMS)*, 2012, pp. 1–6.

[30] Y. Song, D. Treanor, A. J. Bulpitt, N. Wijayathunga, N. Roberts, R. Wilcox, and R. D Magee, "Unsupervised content classification based nonrigid registration of differently stained histology images," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 1, pp. 96–108, Jan. 2014.

[31] S. Manivannan, W. Li, J. Zhang, E. Trucco, and S. J. McKenna, "Structure prediction for gland segmentation with hand-crafted and deep convolutional features," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 210–221, Jan. 2018.

[32] J. Lotz *et al.*, "Patch-based nonlinear image registration for gigapixel whole slide images," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 9, pp. 1812–1819, Sep. 2016.

[33] S. Zhao, T. Lau, J. Luo, E. I.-C. Chang, and Y. Xu, "Unsupervised 3D end-to-end medical image registration with volume tweening network," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 5, pp. 1394–1404, May 2020.

[34] S. Zhao, Y. Dong, E. Chang, and Y. Xu, "Recursive cascaded networks for unsupervised medical image registration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019.

[35] Y. Hu *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Med. Image Anal.*, vol. 49, pp. 1–13, Oct. 2018.

[36] Y. Hu *et al.*, "Label-driven weakly-supervised learning for multimodal deformable image registration," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, Dec. 2018, pp. 1070–1074.

[37] M. Blendowski, L. Hansen, and M. P. Heinrich, "Weakly-supervised learning of multi-modal features for regularised iterative descent in 3D image registration," *Med. Image Anal.*, vol. 67, 2021, Art. no. 101822.

[38] E. Ferrante, P. K. Dokania, R. M. Silva, and N. Paragios, "Weakly supervised learning of metric aggregations for deformable image registration," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1374–1384, Jul. 2019.

[39] T. Nguyen-Duc, I. Yoo, L. Thomas, A. Kuan, W.-C. Lee, and W.-K. Jeong, "Weakly supervised learning in deformable EM image registration using slice interpolation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 670–673.

[40] J. Zhang, F. Liu, X. Yu, Y. Ma, and X. Zhao, "A 3D medical image registration method based on multi-scale feature fusion," *J. Phys., Conf. Ser.*, vol. 1948, no. 1, Jun. 2021, Art. no. 012057.

[41] A. Franz, I. C. Carlsen, and S. Renisch, "An adaptive irregular grid approach using SIFT features for elastic medical image registration," in *Bildverarbeitung Für Medizin*, H. Handels, J. Ehrhardt, A. Horsch, H.-P. Meinzer, and T. Tolxdorff, Eds. Berlin, Germany: Springer, 2006, pp. 201–205.

[42] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang, "A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4328–4338, Jul. 2014.

[43] W. Ma *et al.*, "Remote sensing image registration with modified sift and enhanced feature matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 3–7, Jan. 2017.

[44] P. Schwind, S. Suri, P. Reinartz, and A. Siebert, "Applicability of the SIFT operator to geometric SAR image registration," *Int. J. Remote Sens.*, vol. 31, no. 8, pp. 1959–1980, Mar. 2010.

[45] Y. Chen and L. Shang, "Improved SIFT image registration algorithm on characteristic statistical distributions and consistency constraint," *Optik*, vol. 127, no. 2, pp. 900–911, Jan. 2016.

[46] Z. M. C. Baum, Y. Hu, and D. C. Barratt, "Real-time multimodal image registration with partial intraoperative point-set data," *Med. Image Anal.*, vol. 74, Dec. 2021, Art. no. 102231.

[47] K. Zhang, X. Li, and J. Zhang, "A robust point-matching algorithm for remote sensing image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 2, pp. 469–473, Feb. 2014.

[48] A. Myronenko, X. B. Song, and M. Carreira-Perpiñán, "Non-rigid point set registration: Coherent point drift," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2006, pp. 1–8.

[49] G. Wang, Z. Wang, Y. Chen, and W. Zhao, "Robust point matching method for multimodal retinal image registration," *Biomed. Signal Process. Control*, vol. 19, pp. 68–76, May 2015.

[50] J. Ma, J. Zhao, J. Tian, Z. Tu, and A. L. Yuille, "Robust estimation of nonrigid transformation for point set registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 2147–2154.

[51] S.-Y. Guan, T.-M. Wang, C. Meng, and J.-C. Wang, "A review of point feature based medical image registration," *Chin. J. Mech. Eng.*, vol. 31, no. 1, p. 76, 2018.

[52] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.

[53] S. Zhao *et al.*, "MaskFlownet: Asymmetric feature matching with learnable occlusion mask," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6278–6287.

[54] M. Jaderberg, K. Simonyan, A. Zisserman, and K. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing*, vol. 28, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015.

[55] S. Zhao *et al.*, "Recursive cascaded networks for unsupervised medical image registration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Dec. 2019, pp. 10600–10610.

[56] H. Qi *et al.*, "Non-rigid respiratory motion estimation of whole-heart coronary MR images using unsupervised deep learning," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 444–454, Jan. 2021.

[57] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[58] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[59] A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *Proc. CVPR*, vol. 1, 2005, pp. 26–33.

[60] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[61] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, Aug. 2011.

[62] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Jun. 2004.

[63] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 196–205, Jan. 2010.

[64] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[65] G. Kang, Y. Wei, Y. Yang, Y. Zhuang, and A. G. Hauptmann, "Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation," 2020, *arXiv:2011.00147*.

[66] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. Int. Conf. Pattern Recognit.*, vol. 3, Aug. 2006, pp. 850–855.

[67] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[68] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Med. Image. Anal.*, vol. 52, pp. 128–143, Feb. 2019.