MCMR-SM: A Multilevel Cross-Modal Registration Framework Using Semantic Masks

Wenbo Liu
Translational Medicine Research Center,
WeiFang People's Hospital,
Shandong Second Medical University
Weifang, Shandong, China
boboliucn@126.com

Shuqing Sun
Intensive Care Unit ,
WeiFang People's Hospital,
Shandong Second Medical University
Weifang, Shandong, China
shuqingsun@163.com

Yang Chen
Shuiping Gou
Kehong Liu
Yingping Li*
xduchenyang@163.com
shpgou@mail.xidian.edu.cn
kehongliu@stu.xidian.edu.cn
liyingping@xidian.edu.cn
School of Artificial Intelligence,
Xidian University
Xi'an, Shaanxi, China

Ying Zhang*
Department of Radiology,
Xijing Hospital,
Air Force Medical University
Xi'an, Shaanxi, China
48600658@qq.com

ABSTRACT

In abdominal tumor radiotherapy, combining MRI with CBCT improves target delineation and dose accuracy, essential for effective treatment planning. However, traditional registration methods rely on internal features and face challenges in handling diverse grayscale distributions and anatomical variations across modalities. This paper introduces a multilevel cross-modal registration framework using semantic masks (MCMR-SM), specially tailored to address large deformations across modalities. This multilevel framework begins with training a segmentation network such as Mask-RCNN to obtain organ segmentation, which serves as weak supervision for image registration. Subsequently, an initial affine registration network aligns organ structures globally, followed by a deformation registration network refining organ shape and structure with precision. The registration process relies primarily on semantic masks, a novel contribution of our work, constructed from segmentation masks and images. These semantic masks reduce dependence on original images, thus reinforcing the network against anatomical and modality variations and promoting robust registration across different imaging modalities. Experimental results demonstrate the superiority of our method over traditional iterative techniques and other deep learning methods.

CCS CONCEPTS

• **Computing methodologies** → **Biometrics**.

KEYWORDS

Cross-modal image registration, Semantic masks, Multilevel registration, Large deformations, Multimodal image registration

ACM Reference Format:
Wenbo Liu, Shuqing Sun, Yang Chen, Shuiping Gou, Kehong Liu, Yingping Li, and Ying Zhang. 2024. MCMR-SM: A Multilevel Cross-Modal Registration Framework Using Semantic Masks. In *2024 9th International Conference on Biomedical Imaging, Signal Processing (ICBSP 2024), October 18–20, 2024, Hong Kong, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Medical image registration is the process of aligning multiple medical images of the same anatomical region, acquired from different patients, time points, perspectives, or imaging modalities [8, 10]. This process establishes correspondences between the images, facilitating clearer visualization and comprehensive analysis. Medical image registration plays a crucial role in various clinical applications such as diagnosis, treatment planning, image-guided interventions, and research. For example, Cone Beam Computed Tomography (CBCT) are often utilized in image-guided radiation therapy for abdominal tumors due to their rapid scanning, minimal radiation exposure, and low cost. However, their limited capacity to visualize soft tissues necessitates the use of Magnetic Resonance Imaging (MRI), which particularly excels at imaging soft tissues within the

*Corresponding authors.

ICBSP 2024, October 18–20, 2024, Hong Kong, China

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

body. In clinical practice, doctors typically combine CBCT with MRI observations to accurately locate organs and lesions on abdominal CT images. In this context, registering the CBCT images and MRI is crucial and can improve the effectiveness of radiation therapy.

Integrating CBCT with MRI is crucial in clinical practice, enhancing target delineation and dose calculation accuracy in radiotherapy planning for abdominal tumors. However, abdominal imaging often presents challenges due to large deformations, including organ overlapping, obscuration, adhesion, and expansion. These challenges complicate registration tasks and hinder model interpretability. To address these issues, we propose a multilevel cross-modal registration framework based on semantic masks, named MCMR-SM. Our framework consists of segmentation, affine registration, and deformation registration stages, effectively extracting common features across modalities and incorporating prior knowledge through the proposed semantic masks. By distributing registration tasks across multiple levels and utilizing various semantic masks, our approach achieves superior performance in handling large deformations, modal differences, and inconsistent grayscale levels. We evaluate our method on cervical cancer MRI and CBCT datasets, focusing on organs such as the bladder, uterus, and rectum. Our results demonstrate significant improvements in registration performance compared to existing approaches, as evidenced by metrics such as mutual information (MI) and DICE coefficient.

Our contributions are summarized as follows:

1) Several semantic masks are proposed to encode organ shapes and structures, serving as prior knowledge to alleviate the influence of large deformations and modal differences during registration.
2) Leveraging the proposed semantic masks, we introduce a multilevel framework for cross-modal image registration named MCMR-SM. It is tailored for challenging scenarios characterized by large deformations, modal differences, and inconsistent grayscale distributions. This framework consists of an initial affine transformation to globally align organ structures while effectively preserving internal textures, followed by a deformable registration network to refine organ shape and internal structure with higher precision. Throughout this process, specific semantic masks are constructed for each registration sub-tasks, facilitating the utilization of predefined features.
3) Experiments on CBCT-to-MRI registration task demonstrate the superiority of our proposed MCMR-SM method over traditional iterative techniques and other deep learning methods.

2 RELATED WORK

2.1 Weakly Supervised Image Registration

The image registration problem can be understood as an optimization problem that aims to find a registration field ϕ so that

$$\begin{aligned}\hat{\phi} &= \arg\min_{\phi} \mathcal{L}(f, m, \phi) \\ &= \arg\min_{\phi} \mathcal{L}_{sim}(f, m \circ \phi) + \lambda \mathcal{L}_{smooth}(\phi),\end{aligned} \tag{1}$$

where f and m correspond to the fixed and moving images, respectively. The transformation $m \circ \phi$ warps m by ϕ to approximate f. The function $\mathcal{L}_{sim}(\cdot,\cdot)$, $\mathcal{L}_{smooth}(\cdot)$ and parameter λ correspond to the image similarity loss, the smooth regularization loss and their trade-off parameter, respectively [4].

As an unsupervised loss, $\mathcal{L}_{sim}(f, m \circ \phi)$ measures the similarity between the fixed image f and the warped image $m \circ \phi$. For the cases where f and m have similar image intensity distributions and local contrast, Mean Square Error(MSE) can be an option for this similarity loss function $\mathcal{L}_{sim}$. However, in cases of cross-modal registration, where the fixed and moving images exhibit vastly different intensity distributions, as in our study with CBCT and MRI images, it becomes imperative to explore new methods for quantifying the similarity between these images. An alternative approach is to utilize segmentation labels of the Region of Interests (ROIs), such as organs or tumors. Compared to ground truth of deformation fields, obtaining segmentation labels for ROIs is much more feasible, whether through existing segmentation tasks or manual segmentation process. Subsequently, weakly supervised registration aims to optimize the similarity between the two segmented ROIs:

$$\begin{aligned}\hat{\phi} &= \arg\min_{\phi} \mathcal{L}(f, m, \phi) = \arg\min_{\phi} \mathcal{L}(R_f, R_m, \phi) \\ &= \arg\min_{\phi} \mathcal{L}_{sim}(R_f, R_m \circ \phi) + \lambda \mathcal{L}_{smooth}(\phi),\end{aligned} \tag{2}$$

where R_f and R_m represent the segmentation labels of the ROIs (such as organs or tumors) in the fixed and moving images, respectively. The transformation $R_m \circ \phi$ warps R_m by ϕ to approximate R_f.

Some weakly supervised image registration methods have been proposed. Hu et al. [14] propose a label-driven weakly supervised network for multimodal image registration, inferring voxel-level transformation from higher-level anatomical labels. The proposed approach utilizes convolutional neural networks to predict displacement fields for aligning multiple labeled corresponding structures, demonstrating versatility in training with diverse types of anatomical labels and real-time, fully automated inference without requiring any anatomical labels or initialization. Hering et al. [12] introduce a memory-efficient weakly supervised deep learning model to achieve successful large deformation registration across multi-modal images. Their proposed weakly supervised model combines the complementary strengths of global semantic information (weakly supervised learning with segmentation labels) and local distance metrics borrowed from conventional medical image registration that support the alignment of surrounding structures.

While these existing methods excel in addressing registration challenges involving small deformations and modal differences, they falter when confronted with tasks involving significant deformations and modal disparities. Considering the inherent characteristics of medical images, researchers have utilized image features like centroids and principal axes[21, 28] for registration purposes. Moreover, the binary images themselves possess numerous physical properties that are valuable for registration tasks. Additionally, superpixel methods[15, 23] are commonly employed as a precursor to segmentation tasks. Collectively, these approaches offer valuable insights for us to develop weakly supervised registration methods with enhanced constraints and robustness.

2.2 Unimodal and Multimodal Registration

Existing medical image registration methods primarily fall into two categories: unimodal registration and multimodal (also called cross-modal) image registration. Unimodal registration deals with images of the same modality, where the gray values of corresponding tissues typically show minimal variation. Methods such as optical flow [19, 25] and shape matching are commonly employed in unimodal registration [6]. On the other hand, cross-modal registration involves aligning images from different modalities, posing a challenge due to differing pixel value distributions. To address this, three kinds of methods can be investigated. The first is to investigate the similarity function robust to intensity variations across scans, like local cross-correlation. The second is to map the image into a latent feature space, and then calculate the similarity between the two latent spaces[7, 20]. The third is to adopt an image-to-image translation method like Generative Adversarial Network (GAN)[9] to convert the cross-modal image registration task into a unimodal one[18, 24].

2.3 Multilevel Registration

Rigid registration methods are commonly employed as preprocessing steps for non-rigid registration techniques[4]. In regions like the brain and lungs, where minor positional and angular discrepancies often arise due to factors such as patient positioning during image acquisition, preprocessing steps like translation and rotation prove highly effective. Many comprehensive registration tools, such as ANTs[3] and Elastix[17], incorporate both rigid and non-rigid registration stages. Aside from addressing deviations resulting from external image scanning conditions, multi-level registration is also employed to handle large deformations. In this situation, rather than serving solely as a preprocessing step, rigid registration is integrated into the non-rigid registration process, sometimes iteratively, within the method. The cascade optimization approach leverages multiple Volume Tweening Networks (VTNs) [30] to progressively refine the deformation field, yielding smoother transformations and facilitating registration tasks involving substantial deformations[29].

3 METHOD

3.1 Mask R-CNN for Segmentation

Mask-RCNN [11] is an advanced instance segmentation model capable of detecting and delineating multiple objects within an image simultaneously. Its unique ability lies in its capacity to perform precise segmentation while also accurately identifying object instances, making it an invaluable tool for various computer vision tasks. Unlike traditional methods, it excels in completing these tasks without the need for pixel-level accurate labels during training, making it particularly valuable for areas where obtaining precise annotations is challenging or requires specialized expertise. In this study, Mask R-CNN is trained to automatically obtain the organ segmentation masks.

3.2 Generation of Different Semantic Masks

In this section, as listed below, we propose four semantic masks constructed by the original image and the organ segmentation. Figure 1 shows an example of the generated semantic masks.

3.2.1 Fixed-value mask. The fixed-value mask assigns a constant value to the segmented organ region. By assigning distinct values to the boundary and interior points, it further emphasizes the contour details. This type of mask offers the advantage of simplicity in construction and reduced texture information, enabling the registration network to prioritize the organ's position and shape details. Consequently, registration networks trained with this mask can demonstrate superior performance in organ positioning and shaping.

3.2.2 Centroid mask. Centroid mask is constructed to force the registration network to focus on the internal structure and texture information of the organ. The center-of-mass coordinates $(\bar{x}, \bar{y})$ of the organ can be calculated by

$$\begin{aligned} \bar{x} &= \frac{\iint xv(x,y)dxdy}{\iint v(x,y)dxdy} \\ \bar{y} &= \frac{\iint yv(x,y)dxdy}{\iint v(x,y)dxdy} \end{aligned} \tag{3}$$

where x and y correspond to the coordinates of the pixels in the image, $v(x, y)$ corresponds to the pixel value at the coordinate (x, y). Let $M_{atten} = (m_{xy})_{m\times n}$ be the center attenuation matrix of the image, then each element m_{xy} of M_{atten} can be calculated as

$$m_{xy} = 255 - \lambda_1\sqrt{(x-\bar{x})^2 + (y-\bar{y})^2} \tag{4}$$

where λ_1 corresponds to the attenuation coefficient. Then, the centroid mask can be calculated by the Hadamard product (also known as element-wise product) [13] of the center attenuation matrix M_{atten} and the original organ segmentation mask M_{orig}, namely,

$$M_{centroid} = M_{atten} \odot M_{orig} \tag{5}$$

3.2.3 Centroid axis mask. The centroid mask has limited effectiveness in constraining the rotation of the anatomical structure, potentially leading to errors and inadequate information in polar coordinate space. To address this, we incorporate the principle axis as an extra constraint. Typically, the principal axis—representing the object's axis of least inertia, or least second moment—is employed to ascertain the object's distribution direction.

Let θ be the angle between this principle axis and the x-axis, then

$$\cos 2\theta = \frac{\pm b}{\sqrt{b^2 + (a-c)^2}} \tag{6}$$

where

$$\begin{cases} a = \iint x'^2 v(x', y')dx'dy' \\ b = 2\iint x'y'v(x', y')dx'dy' \\ c = \iint y'^2 v(x', y')dx'dy' \end{cases} \tag{7}$$

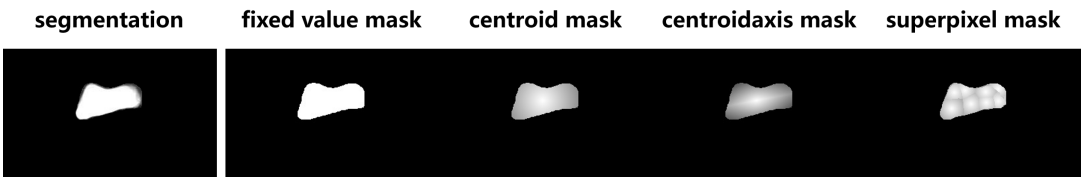

Figure 1: Four different semantic masks proposed in this study.

and $x^{'} = x - \bar{x}$, $y^{'} = y - \bar{y}$. According to the half angle formulas and the relation shape between $\sin\theta$, $\cos\theta$ and $\tan\theta$, we can get

$$\tan\theta = \frac{\sin\theta}{\cos\theta} = \pm\sqrt{\frac{1-\cos 2\theta}{1+\cos 2\theta}} \tag{8}$$

Thus, the line of principle axis can be denoted as

$$y = \bar{y} + (x - \bar{x}) \cdot \tan\theta, \tag{9}$$

then, for each point (x_0, y_0), the distance from (x_0, y_0) to the principle axis can be calculated as

$$d = \frac{|x_0 \tan\theta - y_0 + \bar{y} - \bar{x}\tan\theta|}{\sqrt{\tan^2\theta + 1}} \tag{10}$$

Then the element m_{xy} in the center attenuation matrix M_{atten}, expressed in Equation (4), can be updated by

$$m_{xy} = 255 - \lambda_1\sqrt{(x-\bar{x})^2 + (y-\bar{y})^2} - \lambda_2 d \tag{11}$$

where λ_2 is the attenuation coefficient of the principle axis.

Thus, we construct the centroid mask based on the principle axis, named as centroid axis mask in our study. The centroid axis mask can better constrain the transformation in the polar coordinate space.

3.2.4 Superpixel mask. The centroid mask captures internal organ structure to some extent, but fails to represent complex asymmetries. Thus, a new method is needed. Superpixel mask groups adjacent pixels with similar features like texture and color, providing a region-based representation of the organ structure. Superpixel masks serve as an effective method to encode organ structure.

3.3 Overall Multilevel Registration Framework

Multilevel registration techniques typically employ cascaded registration networks, leveraging deformation fields generated by multiple networks to address large deformations. This study adopts such a framework, beginning with the training of a segmentation network, such as Mask-RCNN, for segmenting multiple organs to be registered. This is followed by multilevel registration, comprising an initial affine registration network followed by a deformation registration network. The affine registration network uses geometric transformations like translation, scaling, rotation, and shearing to maintain the affine properties of shapes. It globally aligns organ structures while effectively preserving internal textures. Then, the deformation registration network refines organ shapes and internal structures with greater precision. Figure 2 shows the multilevel registration framework proposed in this study.

In our proposed registration framework, the segmentation network, affine registration network, and deformable registration network can be selected based on dataset requirements. Our primary emphasis lies in constructing and utilizing semantic masks. By constructing these semantic masks from both the original images and the organ segmentation masks, reliance on the original images is reduced. This improves adaptability across intra-patient organ variability and inter-modality variations, thereby promoting robust registration. Thus, this holistic approach strengthens the registration network against variations in patient anatomy and imaging modalities.

3.3.1 Organ segmentation network. Mask R-CNN is an advanced instance segmentation model capable of detecting and delineating multiple objects within an image simultaneously. Its notable advantage lies in its ability to train without the need for pixel-level annotations. So we use Mask-RCNN to perform the automatic organ segmentation in our study. Significantly, the segmentation network functions autonomously from the registration process, allowing users to choose their preferred segmentation network based on their specific dataset requirements.

3.3.2 Affine registration Network. Utilizing the segmentation outcomes from our organ segmentation network, Mask-RCNN, we derive organ segmentation masks for both the original fixed image and the moving image. These masks, in conjunction with the original images, are combined to create semantic masks. These semantic masks are then utilized as inputs for the first-level affine registration network, powered by Spatial Transformer Network (STN) [16]. During this phase, fixed-value masks are utilized as the selected semantic masks to emphasize the shape and spatial orientation of the organs, aiming to globally align the organ structures while maintaining the internal textures.

3.3.3 Deformable registration network. According to the learned affine transformation in the first-level registration, we calculate the transformed images and the transformed organ segmentation masks. Subsequently, new semantic masks are constructed using the original fixed images and organ segmentation masks, as well as the transformed images and their corresponding organ segmentation masks. To meticulously enhance the organ shape and internal structure during deformation registration, we employ centroid or superpixel masks, containing detailed internal structure information, to refine organs with intricate architectures. These constructed semantic masks are then fed into the second-level deformable registration network, which generates a deformable field. Adjusting the transformed image from the first registration level using this deformable field yields the final registration results.

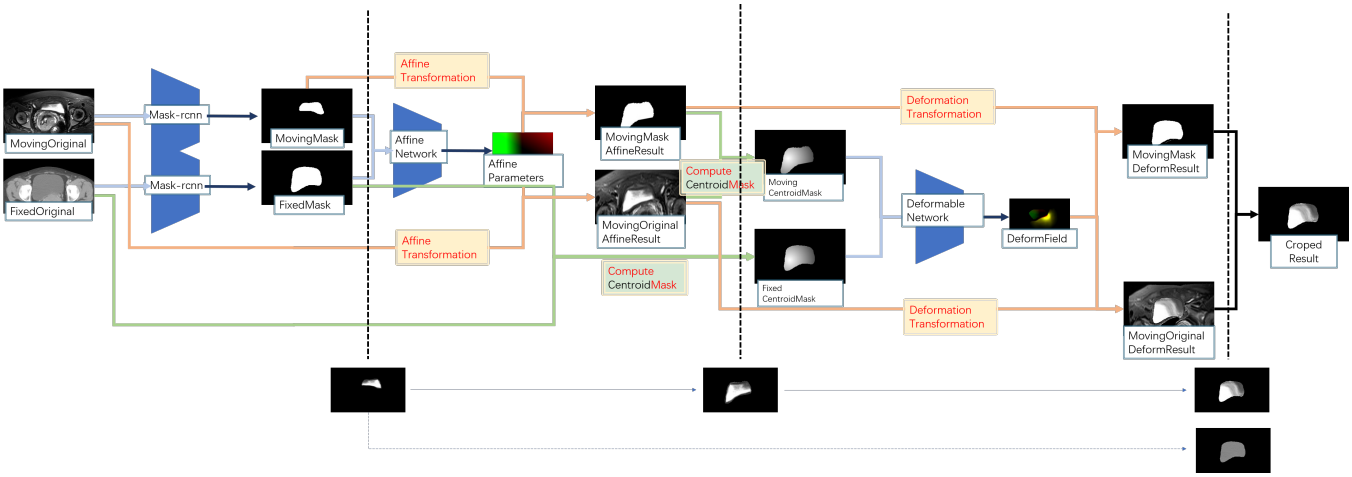

Figure 2: The proposed multilevel registration framework.

4 EXPERIMENTAL SETTINGS

4.1 Dataset and Image Preprocessing

This study collected data from 158 patients who underwent radiotherapy as part of their cancer treatment plans at the Department of Radiology, Xijing Hospital, Air Force Medical University. For each patient, the lower abdomen was scanned to obtain the T2 MRI and Cone beam computed tomography (CBCT) images.

As image preprocessing steps, we first adjust the window width of the medical images. Each MRI-CBCT image pair includes more than 60 MRI slices and more than 100 CBCT slices. After manual screening and alignment, each group has more than ten aligned slices. The segmentation and registration are done on the 2D image slices, thus no necessary to resample the layer thickness, thereby mitigating issues related to inconsistent multi-center spacing. Then we use the grayscale histogram equalization to adjust the grayscale distribution of each histogram. The dataset was randomly split into the training and test sets, with 1200 MRI-CBCT slice pairs as training and 150 pairs as a test.

The study was approved by the ethics committee of Xijing Hospital (Xi'an, China) and conforms to the ethical guidelines of the 1975 Declaration of Helsinki. The need for informed consent was waived.

4.2 Implementation Details

The multi-organ segmentation and the image registration were conducted on the 2D images, with the input size 192×320. For the cross-modal registration, the setups during the modal training were as follows: i) the batch size is set to 2; ii) Adam was used as the optimizer; iii) the initial learning rate was set to 0.0001; iv) the models were trained for 50 epochs; v) Data augmentation strategies, such as random rotation, scaling, and sample pairing, are employed.

All experiments were carried out on a Windows 10 system equipped with an Intel Core i7-8700K CPU, 32GB RAM, NVIDIA GeForce GTX 1080Ti GPU, and CUDA 10.0. Additionally, Keras 2.2.4 and TensorFlow 1.13.0 were utilized.

4.3 Evaluation Metrics

Mutual Information (MI) [1] quantifies the similarity between two images from the perspective of information entropy. Compared to other evaluation matrices, MI is non-linear invariant and is robust with respect to variations of illumination [27], thus being used as an evaluation matrix for cross-modal registration in this study. Dice score and Average Surface Distance (ASD) are commonly used evaluation metrics in segmentation tasks. Here we employ DICE and ASD to evaluate whether the target organs aligned accurately during registration.

4.4 Comparison Methods

The following classical registration methods were compared with the proposed cross-modal registration methods: Elastic methods [5, 22], Demons [19, 25], diffeomorphic demons [26], standard symmetric normalization (SyN) [2] and VoxelMorph [4].

5 RESULTS

In this section, we conduct a series of experiments, both quantitatively and visually, to demonstrate the superior performance of our proposed method.

5.1 Quantitative Comparison of the Classical and Proposed Registration Methods

Table 1 presents the registration performance of the commonly used image registration methods and our proposed MCMR-SM method. As we can see, Elastix and SyN yield unsatisfactory outcomes, possibly due to line noise interference across different modalities. Similarly, the Demons method based on optical flow also fails to provide satisfactory registration results. The diffeomorphic Demons considers the smoothness of the deformation field but at the expense of weakening its deformation capability, leading to incomplete registration for large deformations. VoxelMorph, a deep learning-based method, struggles to learn effective information due to high gray distribution inconsistency between moving and fixed images. Compared to these methods, our proposed MCMR-SM, which leverages

Table 1: Quantitative comparison of the classical and proposed registration methods on our MRI-CBCT image registration task.

Method	MI (↑)	Dice (↑)	ASD (↓)
Elastix	0.194039	0.315895	15.876984
Demons	0.093330	0.280000	16.075311
Symmetric Demons	0.012995	0.086978	nan
SyN	0.128072	0.412759	13.076204
VoxelMorph	0.116464	0.337425	16.062301
MCMR-SM (ours)	**0.322582**	**0.971510**	**0.561818**

Table 2: Quantitative comparison of different semantic masks applied on different registration methods.

	Method	MI (↑)	Dice (↑)	ASD (↓)
Original Image	Elastix	**0.194039**	0.315895	15.876984
	Demons	0.093330	0.280000	16.075311
	Symmetric Demons	0.012995	0.086978	nan
	SyN	0.128072	**0.412759**	**13.076204**
	VoxelMorph	0.116464	0.337425	16.062301
	MCMR-SM (ours)	0.137974	0.186163	32.645654
Fixed-value Mask	Elastix	0.189894	0.930892	0.987932
	Demons	0.280041	0.879704	3.587272
	Symmetric Demons	0.185174	0.626522	9.447205
	SyN	0.287542	0.884215	2.487896
	VoxelMorph	0.281402	0.890459	2.663311
	MCMR-SM (ours)	**0.317281**	**0.994057**	**0.363935**
Centroid Mask	Elastix	0.175194	0.874370	2.148309
	Demons	0.205719	0.715195	5.051722
	Symmetric Demons	0	0	nan
	SyN	0.273350	0.852520	3.010766
	VoxelMorph	0.214077	0.730218	6.326664
	MCMR-SM (ours)	**0.316696**	**0.984384**	**0.577105**

semantic mask information and employs multilevel strategies, excels in all metrics (MI, Dice, and ASD).

5.2 Effectiveness of Semantic Masks

TABLE 2 presents the performance of various algorithms employing distinct semantic masks. All methods exhibit enhanced effectiveness with the inclusion of semantic masks. The fixed-value mask primarily captures contour and shape details, resulting in generally higher Dice score. Conversely, the centroid mask, which contains more information, proves vital in preserving internal texture. However, in scenarios involving large deformations, the single-level grayscale algorithm does not take full advantage of the additional information provided by the centroid mask, resulting in no improvements in the MI index. That is, within a multilevel registration framework, leveraging distinct levels to align diverse masks proves effective in enhancing information extraction and utilization.

5.3 Comparison of Various Semantic Masks at Different Registration Stages

TABLE 3 presents the comparison of different semantic masks at affine and deformable registration stages. While the use of semantic masks undoubtedly enhances model deformability, the selection of specific masks at each stage depends on organ characteristics. As illustrated in TABLE 3, our strategies - using different semantic masks at two registration levels - exhibit different improvements for different organs. Overall, using tailored semantic masks at the affine and deformable registration stages archives better registration results.

The Dice metric measures the overlap between the organs, while ASD is more sensitive to the organ's boundary and shape. Typically, higher Dice score corresponds to lower ASD. However, for specific cases like rectal and bladder in our study, higher Dice scores do not always mean lower surface distances. This phenomenon arises when the registration result closely aligns with the fixed image, either for organs that are too large or too small, resulting in either complete coverage by the fixed organ or being entirely covered by it. Consequently, while DICE scores improve, ASD measurements worsen.

5.4 Qualitative Comparison of the Classical and Proposed Registration Methods

Achieving both deformability and smoothness in registration tasks often presents a challenge, as these factors tend to constrain each other. The goal of registration goes beyond mere alignment, it involves balancing the deformability and smoothness of the deformation fields. We use Figure 3 to illustrate this balance. Figure 3 presents a visual comparison of the deformation fields and moved bladder segmentation masks using different registration methods. As shown in Figure 3, methods such as Elastix, Symmetric Demons, and SyN produce notably smooth deformation fields but yield subpar registration outcomes. Conversely, Demons and VoxelMorph lack sufficient constraints on deformation field smoothness, leading to less satisfactory results. Our approach prioritizes both smooth deformation fields and fidelity of the bladder to the fixed image, ensuring deformations remain as smooth as possible while closely aligning with the reference image.

Figure 4 displays the registration results of different methods. Each row illustrates a specific case, showing the bladder contour from both the moving and registered images superimposed onto the fixed image. The bladder contours in the moving and registered images are outlined in pink and yellow, respectively. As shown in Figure 4, Elastix and VoxelMorph exhibit uneven contours, SyN yields minimal deformation, Demons' contours appear fragmented, and Symmetric Demons even results in some contour disappearance. Only our proposed MCMR-SM method successfully accomplishes the registration task. That is, the bladder in the registered images (enclosed by the yellow contour) overlaps best with the actual bladder in the fixed image.

6 DISCUSSION AND CONCLUSION

In this study, we propose a multi-level deep learning framework for cross-modal image registration, demonstrating its efficacy through

Table 3: Comparison of different semantic masks at affine and deformable registration stages: 1) using original images at both levels; 2) using original segmentation masks at both levels; 3) utilizing fixed-value masks at both levels; 4) utilizing centroid masks at both levels; and 5) using different masks(affine: fixed-value, deformable: centroid) at both levels. Underlined italics represents the best metrics at the affine registration stage, while bold text highlights the best metrics at the deformable registration stage, both across different semantic masks.

		MI (↑)			Dice (↑)			ASD (↓)		
		bladder	rectum	uterus	bladder	rectum	uterus	bladder	rectum	uterus
orig image	affine	0.052897	0.012061	0.053027	0.126520	0.046932	0.093335	36.894659	33.952816	35.459843
	deformable	0.137974	0.012563	0.045212	0.186163	0.031250	0.055413	32.645654	35.614420	32.097432
orig seg mask	affine	0.235191	0.069100	0.169691	0.625512	0.459140	0.60837	7.801295	12.662587	5.500784
	deformable	0.310691	0.083247	0.202042	0.987672	0.842773	0.982316	**0.354547**	4.966201	0.365346
fixed-value mask	affine	0.235463	0.064524	*0.177066*	*0.695675*	0.481711	0.616553	*5.934081*	14.108211	5.244624
	deformable	0.317281	0.084538	**0.215482**	**0.994057**	0.912018	0.995518	0.363935	4.865964	0.078851
centroid mask	affine	0.234324	0.069399	0.166608	0.678380	0.549829	0.613584	6.134080	*11.305571*	5.240941
	deformable	0.316696	0.084837	0.208906	0.984384	0.917690	0.986501	0.577105	**0.164944**	0.298001
different masks	affine	*0.245900*	*0.072458*	0.165149	0.681808	*0.568951*	*0.636536*	6.707808	11.748494	*4.899527*
	deformable	**0.319054**	**0.086309**	0.212010	0.975355	**0.928337**	**0.995595**	0.529227	0.873523	**0.054906**

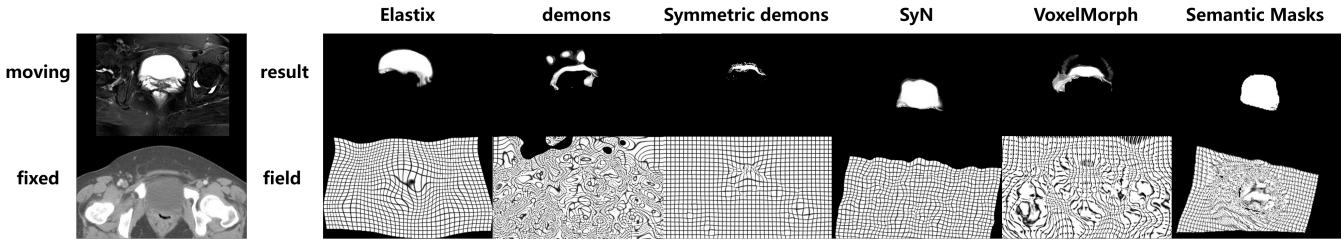

Figure 3: Visual comparison of the deformation fields and the moved organ segmentation masks using different registration methods. The moving image and fixed image are displayed on the left. On the right side, the bottom row shows deformation fields from various methods, while the top row displays the corresponding moved organ segmentation masks.

experiments on registering abdominal CBCT and MRI medical images, even amidst significant deformations. Our method includes three steps that combine deep learning and prior knowledge. Initially, we utilize Mask-RCNN for image segmentation in separate modalities, obtaining key organ labels for registration. Subsequently, semantic masks are constructed based on the segmented labels and medical images. Finally, multilevel registration is performed using tailored masks at each level, ensuring the inclusion of necessary semantic information.

Through the design of semantic masks, our method leverages prior knowledge to address issues of inconsistent grayscale distribution and learning from shape information across modalities. This approach aids in achieving specific registration goals by choosing different semantic masks, particularly in large deformation scenarios. Additionally, the mask constructed based on organ segmentation can be regarded as a type of data resampling, which enhances the robustness of the method to a certain extent.

Compared to fully supervised registration algorithms that require the ground truth of the deformation field, our approach only requires the segmentation labels of the key organs. Thanks to Mask-RCNN, the segmentation labels used for training is not pixel-level. In our registration process, information is primarily derived from modal characteristics and prior knowledge.

We conducted ablation experiments to assess the impact of semantic masks on multilevel registration across various organs. Results demonstrate that employing appropriate semantic masks at different registration steps facilitates the registration. Our strategies are effective across different organs, with the flexibility to use different masks for different registration levels. Comparative analysis with other deep learning and traditional methods reveals our method's superiority in handling large deformation across modalities.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant number 62372358 and 62302355), the Shandong Provincial Natural Science Foundation Project (grant number ZR2023LZY012), the Technology Project of Xianyang City (Key R&D) (grant number JBGS-013), the Key R&D Program of Shaanxi Province (grant number 2024GH-ZDXM-35), and the Fundamental Research Funds for the Central Universities (grant number XJSJ24071 and XJSJ24072).

REFERENCES

[1] Paluck Arora, Rajesh Mehta, and Rohit Ahuja. 2024. An adaptive medical image registration using hybridization of teaching learning-based optimization with

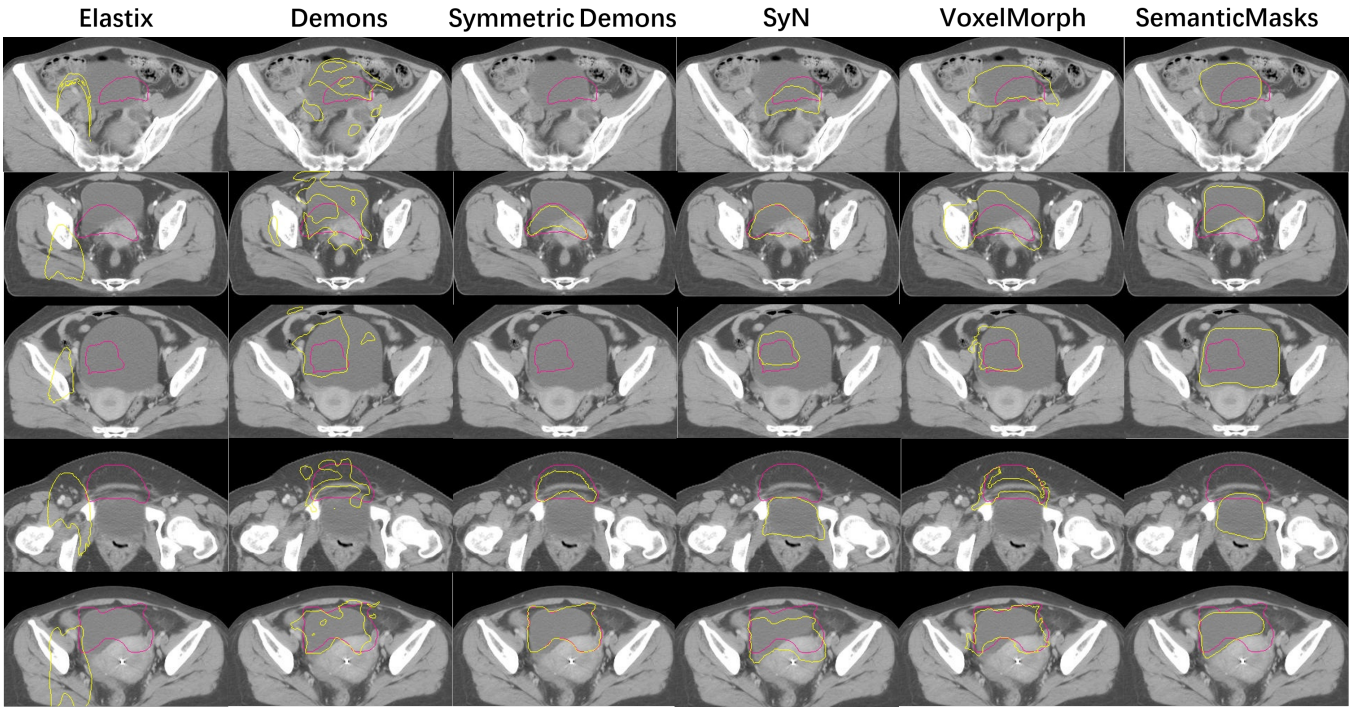

Figure 4: Visual comparison of the registration results using different registration methods. Each row illustrates a specific case, showing the bladder contour from both the moving and registered images superimposed onto the fixed image. The bladder contours in the moving and registered images are outlined in pink and yellow, respectively.

affine and speeded up robust features with projective transformation. *Cluster Computing* 27, 1 (2024), 607–627.

[2] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis* 12, 1 (2008), 26–41.

[3] Brian B Avants, Nick Tustison, Gang Song, et al. 2009. Advanced normalization tools (ANTS). *Insight j* 2, 365 (2009), 1–35.

[4] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. 2019. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging* 38, 8 (2019), 1788–1800.

[5] Christos Davatzikos. 1997. Spatial transformation and registration of brain images using elastically deformable models. *Computer Vision and Image Understanding* 66, 2 (1997), 207–222.

[6] Xin Deng, Enpeng Liu, Shengxi Li, Yiping Duan, and Mai Xu. 2023. Interpretable multi-modal image registration network based on disentangled convolutional sparse coding. *IEEE Transactions on Image Processing* 32 (2023), 1078–1091.

[7] Xin Deng, Enpeng Liu, Shengxi Li, Yiping Duan, and Mai Xu. 2023. Interpretable multi-modal image registration network based on disentangled convolutional sparse coding. *IEEE Transactions on Image Processing* 32 (2023), 1078–1091.

[8] Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. 2020. Deep learning in medical image registration: a review. *Physics in Medicine & Biology* 65, 20 (2020), 20TR01.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[10] Grant Haskins, Uwe Kruger, and Pingkun Yan. 2020. Deep learning in medical image registration: a survey. *Machine Vision and Applications* 31, 1 (2020), 8.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[12] Alessa Hering, Sven Kuckertz, Stefan Heldmann, and Mattias P Heinrich. 2019. Memory-efficient 2.5 D convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart CT and MRI scans. *International journal of computer assisted radiology and surgery* 14 (2019), 1901–1912.

[13] Roger A Horn. 1990. The hadamard product. In *Proc. Symp. Appl. Math*, Vol. 40. 87–169.

[14] Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester Bonmati, Guotai Wang, Steven Bandula, Caroline M Moore, Mark Emberton, et al. 2018. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis* 49 (2018), 1–13.

[15] Abdelhameed Ibrahim and El-Sayed M El-kenawy. 2020. Image segmentation methods based on superpixel techniques: A survey. *Journal of Computer Science and Information Systems* 15, 3 (2020), 1–11.

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. *Advances in neural information processing systems* 28 (2015).

[17] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. 2009. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging* 29, 1 (2009), 196–205.

[18] Dwarikanath Mahapatra, Bhavna Antony, Suman Sedai, and Rahil Garnavi. 2018. Deformable medical image registration using generative adversarial networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 1449–1453.

[19] Xavier Pennec, Pascal Cachier, and Nicholas Ayache. 1999. Understanding the "demon's algorithm": 3D non-rigid registration by gradient descent. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 597–605.

[20] Chen Qin, Bibo Shi, Rui Liao, Tommaso Mansi, Daniel Rueckert, and Ali Kamen. 2019. Unsupervised deformable registration for multi-modal images via disentangled representations. In *International Conference on Information Processing in Medical Imaging*. Springer, 249–261.

[21] Henry Rusinek, Wai-Hon Tsui, Alejandro V Levy, Marilyn E Noz, and Mony J de Leon. 1993. Principal axes and surface fitting methods for three-dimensional image registration. *Journal of nuclear medicine* 34, 11 (1993), 2019–2024.

[22] Dinggang Shen and Christos Davatzikos. 2002. HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE transactions on medical imaging* 21, 11 (2002), 1421–1439.

[23] Subhashree Subudhi, Ram Narayan Patro, Pradyut Kumar Biswal, and Fabio Dell'Acqua. 2021. A survey on superpixel segmentation as a preprocessing step in hyperspectral image analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), 5015–5035.

[24] Christine Tanner, Firat Ozdemir, Romy Profanter, Valeriy Vishnevsky, Ender Konukoglu, and Orcun Goksel. 2018. Generative adversarial networks for MR-CT deformable image registration. *arXiv preprint arXiv:1807.07349* (2018).
[25] J-P Thirion. 1998. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical image analysis* 2, 3 (1998), 243–260.
[26] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. 2009. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage* 45, 1 (2009), S61–S72.
[27] Paul Viola and William M Wells III. 1997. Alignment by maximization of mutual information. *International journal of computer vision* 24, 2 (1997), 137–154.
[28] Hengwang Zhao, Zhidong Liang, Chunxiang Wang, and Ming Yang. 2021. CentroidReg: A global-to-local framework for partial point cloud registration. *IEEE Robotics and Automation Letters* 6, 2 (2021), 2533–2540.
[29] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. 2019. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10600–10610.
[30] Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. 2019. Unsupervised 3D end-to-end medical image registration with volume tweening network. *IEEE journal of biomedical and health informatics* 24, 5 (2019), 1394–1404.

Received 15 August 2024; revised xx September 2024; accepted xx September 2024