Learning Better Registration to Learn Better Few-Shot Medical Image Segmentation: Authenticity, Diversity, and Robustness

Yuting He, Rongjun Ge, Xiaoming Qi, Yang Chen, Jiasong Wu, Jean-Louis Coatrieux, Guanyu Yang, Shuo Li

Learning Better Registration to Learn Better Few-Shot Medical Image Segmentation: Authenticity, Diversity, and Robustness

Yuting He, Rongjun Ge, Xiaoming Qi, Yang Chen, *Senior Member, IEEE*, Jiasong Wu, *Member, IEEE*, Jean-Louis Coatrieux, *Life Fellow, IEEE*, Guanyu Yang, *Senior Member, IEEE*, and Shuo Li, *Senior Member, IEEE*

***Abstract*— In this work, we address the task of few-shot medical image segmentation (MIS) with a novel proposed framework based on the learning registration to learn segmentation (LRLS) paradigm. To cope with the limitations of lack of authenticity, diversity, and robustness in the existing LRLS frameworks, we propose the better registration better segmentation (BRBS) framework with three main contributions that are experimentally shown to have substantial practical merit. First, we improve the authenticity in the registration-based generation program and propose the knowledge consistency constraint strategy that constrains the registration network to learn according to the domain knowledge. It brings the semantic-aligned and topology-preserved registration, thus allowing the generation program to output new data with great space and style authenticity. Second, we deeply studied the diversity of the generation process and propose the space-style sampling program, which introduces the modeling of the transformation path of style and space change between few atlases and numerous unlabeled images into the generation program. Therefore, the sampling on the transformation paths provides much more diverse space and style features to the generated data effectively improving the diversity. Third, we first highlight the robustness in the learning of segmentation in the LRLS paradigm and propose the mix misalignment regularization, which simulates the misalignment distortion and constrains the network to reduce the fitting degree of misaligned regions. Therefore, it builds regularization for these regions improving the robustness of segmentation learning. Without any bells and whistles, our approach achieves a new state-of-the-art performance in few-shot MIS on two challenging tasks that outperform the existing LRLS-based few-shot methods. We believe that this novel and effective framework will provide a powerful few-shot benchmark for the field of medical image and efficiently reduce the costs of medical image research. All of our code will be made publicly available online.**

***Index Terms*— Atlas, deep learning (DL), few-shot learning, generation, medical image registration (MIR), medical image segmentation (MIS).**

Manuscript received 21 December 2021; revised 19 June 2022; accepted 8 July 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0113202, in part by the National Natural Science Foundation under Grant 61828101 and Grant 62101249, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20210291, in part by the Excellence Project Funds of Southeast University, in part by the CAAI-Huawei MindSpore Open Fund, and in part by the Scientific Research Foundation of Graduate School of Southeast University under Grant YBPY2139. *(Corresponding author: Guanyu Yang.)*

Yuting He and Xiaoming Qi are with the Laboratory of Image Science and Technology and the Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, Nanjing 210096, China.

Rongjun Ge is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.

Yang Chen, Jiasong Wu, and Guanyu Yang are with the Laboratory of Image Science and Technology, the Key Laboratory of Computer Network and Information Integration, Ministry of Education, the Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, and the Centre de Recherche en Information Biomédicale Sino-Français (CRIBs), Southeast University, Nanjing 210096, China (e-mail: yang.list@seu.edu.cn).

Jean-Louis Coatrieux is with the Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing and the Centre de Recherche en Information Biomédicale Sino-Français (CRIBs), Southeast University, Nanjing 210096, China, and also with the Inserm, LTSI-UMR1099, Université de Rennes, 35000 Rennes, France.

Shuo Li is with the Department of Biomedical Engineering and the Department of Computer and Data Science, Case Western Reserve University, Cleveland, OH 44106 USA (e-mail: shuo.li11@case.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3190452.

Digital Object Identifier 10.1109/TNNLS.2022.3190452

I. Introduction

LEARNING registration to learn segmentation (LRLS, Fig.1) [1]–[6], as an effective paradigm of few-shot medical image segmentation (MIS), is creating its great popularity recently. It has pushed the label efficiency and accuracy of MIS models to soaring heights on a broad array of high-challenging few-shot medical segmentation scenes [2]–[4], [6] where the LRLS only needs nearly one label for delivered strikingly powerful performance. It has three steps (Fig. 1): *Step 1.* Learning a registration network (unsupervised) via a large unlabeled dataset to provide a basic unit in the generation program. *Step 2.* Generating a large (pseudo-) labeled dataset with the trained registration network via aligning a small atlas dataset and a large unlabeled dataset. *Step 3.* Learning the segmentation network (supervised) via the generated dataset for a powerful MIS ability. These three steps construct an unsupervised labeled data generation and a supervised learning process, driving the few-shot MIS.

However, three inherent limitations [Fig. 2(a)–(c)] of the LRLS paradigm bring large bottlenecks to the LRLS-based few-shot MIS models.

1) Lack of authenticity in the generated pseudo-labeled data [1], [2], [5] caused by registration error provides inaccurate information. Two kinds of distortion in the

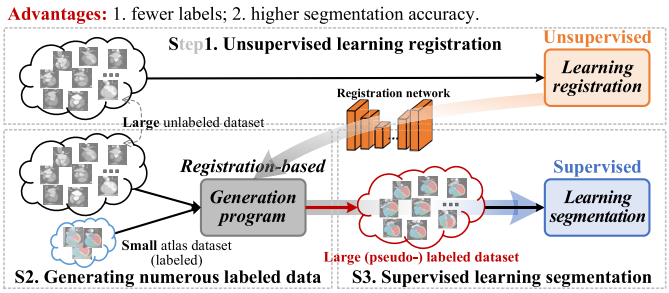

Fig. 1. Paradigm of LRLS for few-shot MIS.

generated data will transmit the error to the segmentation learning. Alignment distortion [3] makes the misaligned regions between the pseudo labels and images, bringing inaccurate supervision in generated data. Deformation distortion [7] makes inaccurate deformation such as the fold of regions in the registration, destroying the real topological structures in deformed images [1], [5]. Therefore, the large registration will make the segmentation network learn inaccurate representation.

2) Lack of diversity in the generation process [2], [4] limits the generalization ability of the segmentation network. Due to the limited amount of unlabeled images, the generation process, which directly couples and aligns the atlases and unlabeled images [1] or transforms space and style of atlases to unlabeled images [2], is only able to enlarge the training dataset to the size of the unlabeled dataset, generating images from a discrete distribution. If the distribution of the unlabeled dataset is very small and sparse, the generated data will be still in a small and sparse distribution, limiting the generalization of the segmentation network for more images.

3) Lack of robustness in segmentation learning makes the fitting of inaccurate information. The existing LRLS works [1]–[4], [6] are unable to resist the interference from the registration error. Their direct learning of pseudo-labeled datasets will make the segmentation model fit inaccurate feature representation and lose its robustness against registration errors such as the inaccurate labels in misaligned regions.

Motivation: To break through the bottleneck of the LRLS paradigm and achieve accurate MIS models from only extremely few labels, a novel LRLS framework, better registration better segmentation (BRBS), is proposed for powerful few-shot segmentation learning. It addresses the limitations [Fig. 2(a)–(c)] of authenticity, diversity, and robustness in the LRLS paradigm via our proposed three key innovations [Fig. 2(d)–(f)] that effectively generate high-authenticity image–label pairs requiring only few labels, stimulate the diversity of registration-based generation, and explicitly improve the robustness in the segmentation learning with inaccurate data.

1) Knowledge Consistency Constraint Strategy for Authenticity: Our knowledge consistency constraint strategy (KCC) [Fig. 2(d)] constrains the learning of the registration network according to two kinds of domain knowledge, reducing the registration error and improving the authenticity of generated data. *Knowledge 1* for alignment distortion. Aligned images have aligned semantic regions [4], [8], so we present semantic consistency (SeC) constraint for semantic-aligned registration. It utilizes the segmentation network in the training process to extract semantic regions and constrains the corresponding regions to align to improve the alignment accuracy between images. *Knowledge 2* for deformation distortion. A topology-preserved deformation is invertible and smooth [9], [10], so we fuse bidirectional consistency (BiC) constraint for topology-preserved registration. It constrains the network to learn both forward and reverse smooth deformation interimages for invertibility and smoothness [11], thus preserving the authenticity of structures in generated images.

2) Space-Style Sampling Program for Diversity: Our space-style sampling program (S3P) [Fig. 2(e)] densely samples data during the transformation process from atlases to unlabeled images, generating data in a dense distribution and improving the diversity. It builds continuous distributions between the atlases and unlabeled images via modeling their transformation process of style and space and densely samples data in these continuous distributions for the generated data with better diversity. For style, it builds a style transformation path between the atlas and an unlabeled image and samples the displacement degree on this path for new style features. For space, it builds a space transformation path between the atlas and another unlabeled image and samples the deformation degree on this path for new space features. The style and space features are fused into the atlas to generate a new pseudo-labeled image. With the dense sampling on the transformation paths (continuous distribution), diverse space and style features will be extracted to improve the diversity in our generation.

3) Mix Misalignment Regularization for Robustness: Our mix misalignment regularization (MMR) [Fig. 2(f)] simulates the misalignment and constrains the network to reduce the fitting degree of these misaligned regions, making a regularization for inaccurate information and improving the robustness of segmentation learning. It mixes two images to simulate the misalignment distortion and mixes their segmented maps to discover the misaligned regions. Then, it constrains the segmentation network to perceive these misaligned regions to improve the fit function's linearity of these regions, thus reducing the fitting degree of this inaccurate information. Therefore, the learned segmentation network will fit a simpler function and be more robust to misalignment in generated images [12], further improving the MIS performance.

Specifically, our contributions also include the following.

1) For the first time, we propose an efficient LRLS-based few-shot MIS framework, BRBS, which provides a powerful few-shot benchmark for medical image field and efficiently reduces the costs of medical image research.
2) We propose a domain-knowledge-driven registration learning strategy, KCC, which effectively reduces the registration error for better authenticity in the generation, being more in line with the real-world image generation.
3) We propose a novel data generation program, S3P, which densely samples data on continuous distributions

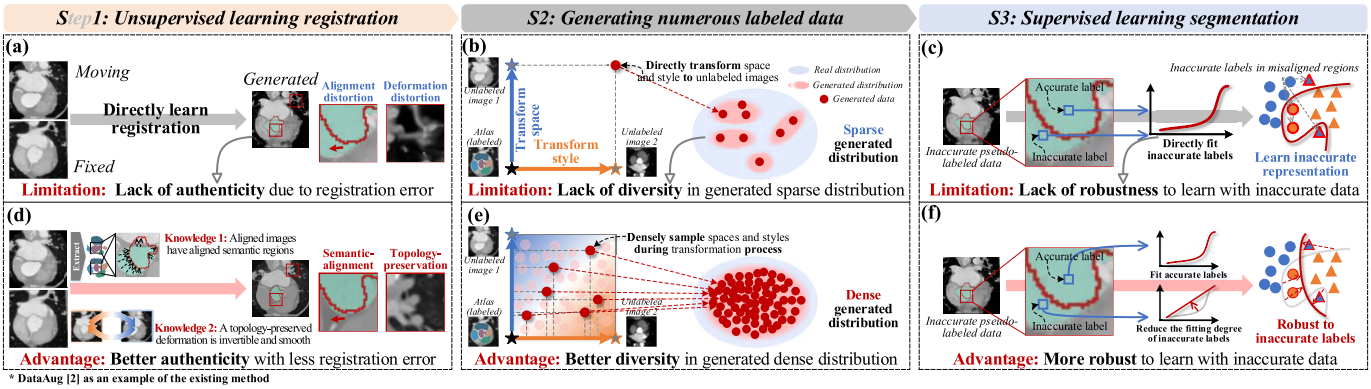

Fig. 2. Limitations of the existing LRLS frameworks and the advantages of our BRBS framework that break through the bottleneck of the LRLS paradigm and achieve accurate MIS models from only extremely few labels. From S1 to S3, they are the detailed steps of the LRLS paradigm, from (a) to (c), they are the limitations of the existing LRLS frameworks (DataAug [2] as an example), and from (d) to (f), they are the solutions and advantages of our BRBS, respectively. (a) Existing method directly learns registration. (b) Existing method generates data in the end of transformation. (c) Existing method fits inaccurate labels. (d) Our KCC constrains registration learning. (e) Our S3P generates data in a continuous transformation process. (f) Our MMR reduces the fitting degree of inaccurate labels.

between the atlases and unlabeled images for the generation with better diversity.

4) We propose a powerful data-driven regularization method, MMR, which constrains the segmentation network to reduce the fitting degree of misaligned regions in a data-driven way, so that the segmentation network will be more robust to learn with inaccurate labels.

Overall, our BRBS framework has three key advantages: **a) Fewer labels:** Compared with the full supervised methods [13]–[16], our framework is able to drive the whole training process with only nearly one label for a reliable MIS network, which effectively reduces the labeling costs in MIS tasks. **b) Higher accuracy:** Compared with other LRLS methods [1]–[4], [6], our advancements in the authenticity, diversity, and robustness of the LRLS paradigm bring a significant improvement in MIS performance, and our BRBS has achieved the state-of-the-art results on multiple challenging MIS tasks. **c) Great simplicity:** In the testing process, our framework only uses a single segmentation network for inference without any additional parts. We have implemented the proposed methods by the PyTorch [17] framework. We share our code and models at a companion website https://github.com/YutingHelist/BRBS.

II. Related Works

A. Few-Shot Learning

Few-shot learning targets learning methods from a limited number of examples with supervised information [18]–[21], effectively improving the label efficiency. Recently, it has achieved success in numerous tasks, such as the image classification [21], video classification [19], and neural architecture search [20]. Due to the high cost of annotation and the scarcity of medical data [22], the few-shot learning is urgent in medical image analysis and has achieved success in many medical image tasks [2]–[4], [6], [23]. However, as pointed out by Wang *et al.* [18], the core issue of few-shot learning is the unreliable optimization direction in learning (empirical risk minimizer). When the labeled training dataset is very small, empirical risk will be far from being a reliable approximation of the expected risk, bringing overfitted optimization direction from the few examples. Therefore, it is challenging for reliable learning in a few-shot situation. Although the above existing methods have introduced preliminary solutions to improve the reliability of optimization in a few-shot situation including the LRLS paradigm [1]–[6] in this article, it is still a bottleneck in few-shot learning.

1) Medical Image Registration and Segmentation: In the previous decade, medical image registration (MIR) [11] and MIS [24] are two key dense prediction tasks in medical image analysis. The MIR [4], [11], [25], [26] aligns the anatomical structures in medical images to the same spatial coordinate system, and the MIS [16], [24], [27] densely labels the semantics of the anatomical regions. With the development of deep learning (DL) [28], the DL-combined MIR and MIS models are being deeply studied and have achieved numerous remarkable results [4], [11], [24], [29]. However, large challenges are limiting these models. The unsupervised MIR models [11], [26] lack region perception ability, so they are always limited by the alignment accuracy on the regions of interest. The supervised MIS models [16], [24] have large label requirements resulting in large labeling costs. If only a few labels are available, the data-hungry nature of DL will bring serious overfitting risks.

2) Atlas-Based Segmentation: Based on the registration, a widely studied few-shot MIS paradigm is the atlas-based segmentation [5], [30]–[35]. It takes the key idea that uses the relationship between segmentation labels and images (i.e., atlas) in image–label pairs and registers the labeled images to unlabeled images, thus indirectly constructing the mapping from labels to unlabeled images for segmentation results [33]. In recent years, due to the powerful feature representation ability and high speed of the DL, some works [1]–[5], [30]–[32] have been devoted to combining the DL with atlas-based segmentation frameworks, achieving promising segmentation results with few labeling costs. An intuitive method [5], [31] is to use deep networks to learn

unsupervised registration models [11] for a faster and more accurate alignment between atlases and unlabeled images. However, it is limited by the similarity between atlases and unlabeled images, if the atlases are extremely different from the target images, it will bring poor results. Although some works [26], [36] use the constraint from segmentation labels for a better alignment, they also bring the limitation of label amount.

3) Learning Registration to Learn Segmentation: Recently, based on the research of the complementarity [3] between the DL registration and segmentation models, an effective few-shot MIS paradigm, LRLS [1]–[4], [6], is creating its great popularity. This paradigm takes the registration model to align the atlas and unlabeled images, thus generating numerous pseudo-labeled images to drive the learning of the DL segmentation model with few labels. The DataAug [2] used this paradigm and further introduced an appearance transformation that improves the diversity of generated data. However, the distortion in inaccurate registration brings interference in segmentation learning. Ding *et al.* [6] further advanced the diversity of DataAug and trained a variational autoencoder (VAE) to generate pseudo-labeled images from continuous latent space. However, it only works in the one-shot situation, which limits the scalability, and the distortion caused by the inaccurate registration is still limiting the final segmentation learning. The DeepRS [3], PC-Reg-RT [4], and DeepAtlas [1] took the complementarity of registration and segmentation tasks. They used the segmentation model to constrain the registration purposefully align target regions and bring more accurate alignment, reducing the distortion for higher quality generated data. However, these works did not consider the deformation distortion and the diversity of the generated data, so the segmentation model will still have the risk of interference in training and be challenging to generalize to more images. What is more, these works [1]–[6] lack further research on how to learn segmentation networks more robustly in the registration-based generated data.

4) Discussion for Related Works: The authenticity [5], [7], [37], diversity [6], [20], and robustness [38], [39] are the three key aspects of the LRLS paradigm on few-shot MIS tasks. Better authenticity of the generated image–label pairs will reduce the error information caused by registration and make the segmentation model learn a more reliable representation. Better diversity of the generated data will provide the segmentation network with more knowledge, achieving the efficient representation for better generalization ability. Better robustness of the segmentation learning will improve the ability to resist interference from inaccurate data, achieving accurate feature representation and reliable optimization. Therefore, the LRLS paradigm desiderates to advance its authenticity, diversity, and robustness, learning better registration to learn better few-shot MIS. Although some works have made preliminary studies on diversity [2], [3], [6] and authenticity [1], [3], [4], there is no deep and complete research on all of these three key aspects in the LRLS paradigm. In this article, our KCC, S3P, and MMR simultaneously improve the authenticity, diversity, and robustness, thus further advancing the LRLS paradigm to more powerful performance.

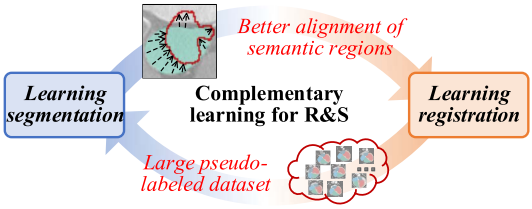

Fig. 3. Complementary learning process of segmentation (S) and registration (R). Throughout the whole training process, the registration and segmentation networks are trained cyclically, so their complementary constraints are transmitted in the cycling process, achieving mutual improvement.

III. Better Registration Better Segmentation

The proposed BRBS framework (Fig. 2) learns the semantic-aligned and topology-preserved registration for the generation with better authenticity (KCC, see Section III-A), densely samples data on the transformation paths for the generation with better diversity (S3P, see Section III-B), and learns misalignment-preserved segmentation for better robustness of inaccurate information in generated data (MMR, see Section III-C), thus achieving effective few-shot MIS. Our BRBS takes the complementary learning process [3] (Fig. 3), which trains registration and segmentation networks iteratively and alternately for complementary constraints throughout the training process.

A. Knowledge Consistency Constraint Strategy Improves Authenticity

Our KCC (Figs. 2(d) and 4) utilizes two kinds of domain knowledge to constrain the registration learning, generating pseudo-labeled data in line with the domain knowledge, greatly improving the authenticity. We denote the images sampled from our training dataset $\mathcal{D}$ as $\mathbf{M}$ and $\mathbf{F}$ to introduce this strategy. It has two key components.

1) Bidirectional Consistency: It constrains the registration network to learn a topology-preserved registration for anatomical structure, improving the structure's authenticity in generated images. It constrains invertibility and smoothness for topology preservation. For invertibility, it not only constrains forward deformation which deforms image $\mathbf{M}$ to image $\mathbf{F}$ via the forward deformation field ϕ like [11] but also constrains inverse deformation to deform image $\mathbf{F}$ to image $\mathbf{M}$ via the inverse deformation field ϕ^{-1}

$$\phi = \mathbb{R}(\mathbf{M}, \mathbf{F}), \phi^{-1} = \mathbb{R}(\mathbf{F}, \mathbf{M}). \tag{1}$$

The forward and inverse deformation fields are also constrained mutual inversion via an inverse loss $\mathcal{L}_I$

$$\begin{aligned} \mathcal{L}_I\left(\phi, \phi^{-1}\right) &= \sum\nolimits_{p\in\phi} \left\|\phi - \mathbb{I}\left(\phi^{-1}\right)\right\|^2 \\ \mathbb{I}\left(\phi^{-1}\right) &= -1 * \mathbb{T}\left(\phi^{-1}, \phi^{-1}\right) \end{aligned} \tag{2}$$

where p is the position of the voxels in ϕ, $\mathbb{T}$ is a differentiable spatial transformation operation [11], which deforms the input via the deformation field, and $\mathbb{I}$ is an inverse transformation process for the deformation field. This loss has two steps. First, $\mathbb{I}(\phi^{-1})$ aligns the inverse deformation field ϕ^{-1} to the space of the forward deformation field ϕ and negates it.

Therefore, the vectors in the inverse deformation field will have the same directions and space positions as their corresponding vectors in the forward deformable field. Then, $\mathcal{L}_I(\phi, \phi^{-1})$ is used to calculate the difference between the aligned inverse and forward deformation fields to optimize the network to perform invertible deformation, improving the topology-perversion ability. We simultaneously calculate the similarity $\mathcal{L}_{\text{sim}}$ between the deformed image $\mathbf{M}^{\phi}$ and target image $\mathbf{F}$ pair, and the similarity between the deformed image $\mathbf{F}^{\phi^{-1}}$ and target image $\mathbf{M}$ pair. Therefore, the deformation will obtain a better invertibility, bringing authenticity to topological structures. We take a popular normalized cross correlation loss [11] as $\mathcal{L}_{\text{sim}}$

$$\mathcal{L}_{\text{sim}}\left(\mathbf{F}, \mathbf{M}^{\phi}\right) = \sum_{p\in\Omega} \frac{\left(\sum_{p_i}\left(\mathbf{F}(p_i) - \hat{\mathbf{F}}(p)\right)\left(\mathbf{M}^{\phi}(p_i) - \hat{\mathbf{M}}^{\phi}(p)\right)\right)^2}{\left(\sum_{p_i}\left(\mathbf{F}(p_i) - \hat{\mathbf{F}}(p)\right)^2\right)\left(\sum_{p_i}\left(\mathbf{M}^{\phi}(p_i) - \hat{\mathbf{M}}^{\phi}(p)\right)^2\right)} \tag{3}$$

where Ω is the space of the images and $\hat{\mathbf{F}}(p)$ ($\hat{\mathbf{M}}^{\phi}(p)$) is the local mean intensity images: $\hat{\mathbf{F}}(p) = (1/n^3)\sum_{p_i}\mathbf{F}(p_i)$. p_i iterates over a n^3 volume around p. For smoothness, it further takes a gradient loss [11] $\mathcal{L}_{\text{smooth}}$ to constrain the gradient of the forward and inverse deformation fields to be small simultaneously, weakening local exaggerated distortion

$$\mathcal{L}_{\text{smooth}}(\phi) = \sum_{p\in\phi} \|\nabla\phi(p)\|^2 \tag{4}$$

where p is the position of the voxels in ϕ and $\nabla\phi(p)$ is the gradient of the pth position on the deformation field ϕ. Therefore, our BiC loss $\mathcal{L}_{\text{BiC}}$ is

$$\begin{aligned}\mathcal{L}_{\text{BiC}}&\left(\mathbf{F}, \mathbf{M}, \phi, \phi^{-1}\right) \\ &= \mathcal{L}_{\text{sim}}(\mathbf{F}, \mathbb{T}(\mathbf{M}, \phi)) + \mathcal{L}_{\text{sim}}\left(\mathbf{M}, \mathbb{T}\left(\mathbf{F}, \phi^{-1}\right)\right) \\ &\quad + \lambda_0\left(\mathcal{L}_{\text{smooth}}(\phi) + \mathcal{L}_{\text{smooth}}\left(\phi^{-1}\right)\right) + \lambda_1\mathcal{L}_I\left(\phi, \phi^{-1}\right)\end{aligned} \tag{5}$$

where λ_0 and λ_1 are the weights of the losses in $\mathcal{L}_{\text{BiC}}$. Following [11], we set $\lambda_0 = 1$, and following [40], we set $\lambda_1 = 0.1$ for better topology-preservation ability. Together with our SeC (below), our BiC achieves better topological preservation than [11] and a smaller risk of semantic misalignment caused by the excessive smoothing in [40].

2) Semantic Consistency: It constrains the registration network to learn the alignment ability of corresponding semantic regions between two images, improving the accuracy of style displacement for semantic authenticity in generated images. It segments the semantic regions on two original images $(\mathbf{M}, \mathbf{F})$ via our fixed segmentation network ($\mathbb{S}$) and constrains their corresponding semantic regions to align according to the spatial correspondence in the forward and inverse deformations (ϕ, ϕ^{-1}) via our SeC loss $\mathcal{L}_{\text{SeC}}$. We take a multiclass Dice coefficient [1] loss $\mathcal{L}_{\text{dice}}$ as the metric

$$\begin{aligned}\mathcal{L}_{\text{SeC}}&\left(\mathbf{F}, \mathbf{M}, \phi, \phi^{-1}\right) \\ &= \mathcal{L}_{\text{dice}}(\mathbb{S}(\mathbf{F}), \mathbb{T}(\mathbb{S}(\mathbf{M}), \phi)) + \mathcal{L}_{\text{dice}}\left(\mathbb{S}(\mathbf{M}), \mathbb{T}\left(\mathbb{S}(\mathbf{F}), \phi^{-1}\right)\right).\end{aligned} \tag{6}$$

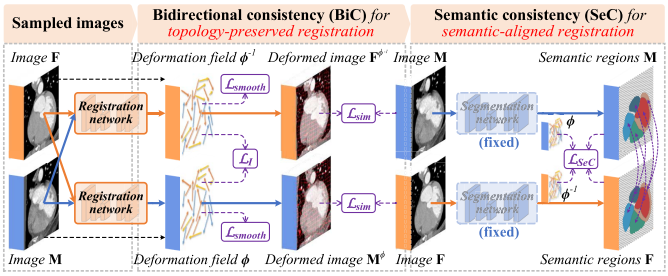

Fig. 4. Our KCC constrains the registration network to learn topology-preserved and semantic-aligned registration via our BiC and SeC, generating space and style transformation maps in line with the domain knowledge and improving the authenticity of generated data.

This loss will optimize the registration network to output the deformation with a great alignment of semantic regions. Therefore, the style displacement map calculated from finely aligned regions will improve the style authenticity.

In general, the learning process is formulated as minimizing an expected loss function $\mathcal{L}_{\mathbb{R}}(\cdot,\cdot,\cdot)$

$$\hat{\mathbb{R}} = \arg\min_{\mathbb{R}}\left[\mathbb{E}_{(\mathbf{F},\mathbf{M})\sim\mathcal{D}}[\mathcal{L}_{\mathbb{R}}(\mathbf{F}, \mathbf{M}, \mathbb{R}(\mathbf{F}, \mathbf{M}), \mathbb{R}(\mathbf{M}, \mathbf{F}))]\right] \tag{7}$$

where $\mathbf{F}$ and $\mathbf{M}$ are the images sampled from our training dataset $\mathcal{D}$. The complete loss is

$$\begin{aligned}\mathcal{L}_{\mathbb{R}}&(\mathbf{F}, \mathbf{M}, \mathbb{R}(\mathbf{F}, \mathbf{M}), \mathbb{R}(\mathbf{M}, \mathbf{F})) \\ &= \mathcal{L}_{\text{BiC}}\left(\mathbf{F}, \mathbf{M}, \phi, \phi^{-1}\right) + \lambda_2\mathcal{L}_{\text{SeC}}\left(\mathbf{F}, \mathbf{M}, \phi, \phi^{-1}\right)\end{aligned} \tag{8}$$

where λ_2 is the weights of the loss $\mathcal{L}_{\text{SeC}}$.

Discussion of the Innovation: Our KCC proposes a novel registration framework with high semantic alignment and topological preservation ability. It utilizes the complementary between the segmentation and registration for the semantic alignment and the cyclic consistency for the topology preservation. Therefore, it has effectively improved the registration accuracy to cope with the limitation of authenticity in the LRLS paradigm.

B. S3P Improves Diversity

Our S3P (Fig. 5) densely samples data on the transformation paths between atlases and unlabeled images in a registration-based generation program, generating image–label pairs with great diversity from few atlases. We consider the transformation of the space and style features in the medical images following [2]. We build an unlabeled set $\mathcal{U}$, which consists of numerous (M) unlabeled images for the varied space and style features, and an atlas set $\mathcal{A}$, which consists of extremely few $(N, N \ll M)$ labeled images for the basic anatomical structures. Therefore, it builds our training dataset $\mathcal{D}$ in the training

$$\mathcal{D} = \left\{\{(\mathbf{A}, \mathbf{y})\}^N, \{\mathbf{U}\}^M\right\}(N \ll M). \tag{9}$$

It first randomly samples (image-level) two images $\mathbf{U}_j$ and $\mathbf{U}_k$ from our unlabeled set $\mathcal{U}$ for their new spaces and styles and one image–label pair $(\mathbf{A}_i, \mathbf{y}_i)$ from our atlas set $\mathcal{A}$ for its basic space and style in $\mathcal{D}$: $\{(\mathbf{A}_i, \mathbf{y}_i), \mathbf{U}_j, \mathbf{U}_k\} \sim \mathcal{D}$.

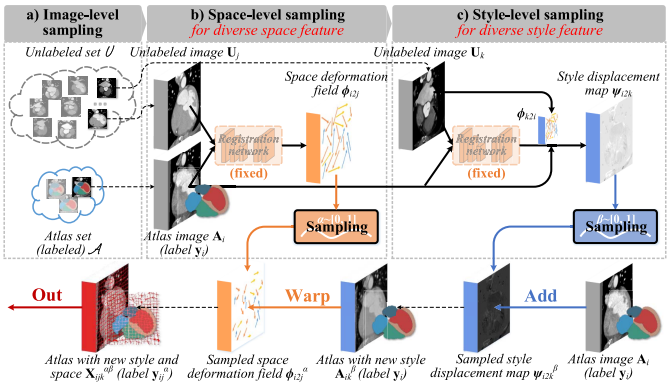

Fig. 5. Our S3P builds a registration-based probabilistic generation program with three levels that sample numerous space and style transformation maps throughout their transformation paths, generating more space and style features and improving the diversity.

For the sampling of space features (space level), one unlabeled image $\mathbf{U}_j$ and the atlas image $\mathbf{A}_i$ are put into the registration network $\mathbb{R}$ for a deformation field [see (10)]. The deformation field ϕ_{i2j} realizes the transformation of space features from the atlas image to the unlabeled image $\mathbf{U}_j$. Then, our S3P generates a random value $\alpha \in [0, 1]$ to sample the deformation degree for new space features in the transformation process

$$\phi_{i2j} = \mathbb{R}(\mathbf{A}_i, \mathbf{U}_j), \phi_{i2j}^{\alpha} = \alpha\phi_{i2j}. \tag{10}$$

For the sampling of style feature (style level), the other unlabeled image $\mathbf{U}_k$ is deformed to the atlas image $\mathbf{A}_i$ via the registration network for a deformed unlabeled image $\mathbf{U}_{k2i} = \mathbb{T}(\mathbf{U}_k, \mathbb{R}(\mathbf{U}_k, \mathbf{A}_i)))$. Then, the style displacement map ψ_{i2k} is calculated between the aligned images $(\mathbf{U}_{k2i}, \mathbf{A}_i)$ via pixelwise subtraction [see (11)], realizing the transformation of style features from the atlas image to the unlabeled image $\mathbf{U}_k$. Then, our S3P generates a random value β to sample the style displacement degree for new style features in the transformation process

$$\psi_{i2k} = \mathbf{U}_{k2i} - \mathbf{A}_i, \psi_{k2i}^{\beta} = \beta\psi_{k2i}. \tag{11}$$

Finally, the sampled style displacement map ψ_{k2i}^{β} is added to the atlas image for new style features, and the sampled space deformation map ϕ_{i2j}^{α} is used to warp the style displaced atlas image and the atlas label, generating image–label pairs $(\{\mathbf{X}_{ijk}^{\alpha\beta}, \mathbf{y}_{ij}^{\alpha}\})$ with new style and space features

$$\mathbf{X}_{ijk}^{\alpha\beta} = \mathbb{T}\left(\mathbf{A}_i + \psi_{i2k}^{\beta}, \phi_{i2j}^{\alpha}\right), \mathbf{y}_{ij}^{\alpha} = \mathbb{T}\left(\mathbf{y}_i, \phi_{i2j}^{\alpha}\right). \tag{12}$$

If β is 0, the generated image is the space deformed atlas image $\mathbf{A}_{ij}^{\alpha}$ without new style features, and if the α is 0, the generated image is the style displaced atlas image $\mathbf{A}_{ik}^{\beta}$ without new space features. Therefore, with the dense sampling of style and space features in our generation program, our S3C achieves denser generated data distribution via sampling data on the transformation paths.

In general, compared with the other typical registration-based generation methods [1]–[4] (Table I), our BRBS has three levels, including the image, space, and style levels,

TABLE I

PROBABILISTIC GENERATION PROGRAMS OF DIFFERENT TYPICAL LRLS METHODS SHOW THAT OUR BRBS HAS HIGHER RANDOMNESS, WHICH WILL GENERATE MORE DIVERSE IMAGES FOR FURTHER TRAINING

Method	Probabilistic program
DeepAtlas [1]	$p((\mathbf{A}_i, \mathbf{y}_i))p(\mathbf{U}_j)$
PC-Reg-RT [4]	$p(\phi_{i2j}\|\mathbb{R}, \mathbf{A}_i, \mathbf{U}_j)p((\mathbf{A}_i, \mathbf{y}_i))p(\mathbf{U}_j)$
DataAug [2]	$p((\mathbf{A}_i, \mathbf{y}_i))p(\mathbf{U}_j)p(\mathbf{U}_k)$
DeepRS [3]	$p(\alpha)p(\phi_{i2j}\|\mathbb{R}, \mathbf{A}_i, \mathbf{U}_j)p((\mathbf{A}_i, \mathbf{y}_i))p(\mathbf{U}_j)$
Our BRBS	$p(\beta)p(\psi_{i2k}\|\mathbb{R}, \mathbf{A}_i, \mathbf{U}_k)p(\alpha)p(\phi_{i2j}\|\mathbb{R}, \mathbf{A}_i, \mathbf{U}_j)p((\mathbf{A}_i, \mathbf{y}_i))p(\mathbf{U}_j)p(\mathbf{U}_k)$

constructing a three-level probabilistic generation program that provides a more diverse generation for further training. It uses three kinds of randomness including the randomness in the kinds of space and style (DataAug, $p((\mathbf{A}_i, \mathbf{y}_i))p(\mathbf{U}_j)p(\mathbf{U}_k)$), the randomness in the training process of the registration network (DeepRS, PC-Reg-RT, $p(\phi_{i2j}|\mathbb{R}, \mathbf{A}_i, \mathbf{U}_j)$), and the diversity in the style ($p(\beta)$) and space (DeepRS, $p(\alpha)$) transformation paths, further improving the diversity of generation.

Discussion of the Innovation: Our S3P proposes a novel registration-based data generation program with higher diversity and without additional training. It models the space and style transformations and densely samples their degree for more diverse space and style features. Therefore, it has effectively improved the diversity in the generated dataset to cope with the limitation of the diversity in the LRLS paradigm.

C. MMR Improves Robustness

Our segmentation network is trained by the generated image–label pair $(\mathbf{X}_{ijk}^{\alpha\beta}, \mathbf{y}_{ij}^{\alpha})$ and constrained by our MMR strategy, avoiding the interference caused by misalignment distortion and achieving robust learning of MIS (Fig. 7).

Our MMR constrains our segmentation network better linearity for the misaligned regions to reduce the fitting degree of misalignment distortion. Because this process does not require labels, we denote the images sampled from our training dataset $\mathcal{D}$ as $\mathbf{X}_A$ and $\mathbf{X}_B$ to introduce this strategy. A random value $\gamma \in [0, 1]$ is used to mix two images $\mathbf{X}_A, \mathbf{X}_B$ for a mixed image $\mathbf{X}_{AB}$ via weighted summation: $\mathbf{X}_{AB} = \gamma\mathbf{X}_A + (1-\gamma)\mathbf{X}_B$, simulating the misalignment distortion in the generated image. Then, our segmentation $\mathbb{S}(\cdot)$ infers the two original images $\mathbf{X}_A, \mathbf{X}_B$ and the mixed image $\mathbf{X}_{AB}$ for their segmented maps $\hat{\mathbf{y}}_A = \mathbb{S}(\mathbf{X}_A), \hat{\mathbf{y}}_B = \mathbb{S}(\mathbf{X}_B)$ and segmented mixed map $\hat{\mathbf{y}}_{AB} = \mathbb{S}(\mathbf{X}_{AB})$. The segmented maps $\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_B$ are also mixed with the random value γ via weighted summation for the mixed segmented map $\mathbf{y}_{AB}$, which has a high response for aligned regions and low response for misaligned regions. Finally, the segmented mixed map $\hat{\mathbf{y}}_{AB}$ is constrained to be similar to the mixed segmented map $\mathbf{y}_{AB}$ via the mix loss $\mathcal{L}_{\text{Mix}}$, thus guiding the segmentation network to perceive the aligned and misaligned regions for the robustness of the misalignment distortion caused by registration

$$\mathcal{L}_{Mix}(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_B, \hat{\mathbf{y}}_{AB}) = \| \gamma\hat{\mathbf{y}}_A + (1-\gamma)\hat{\mathbf{y}}_B - \hat{\mathbf{y}}_{AB} \|^2. \tag{13}$$

In the view of the optimization process [Fig. 6(a)], our MMR constrains our segmentation network to fit a function

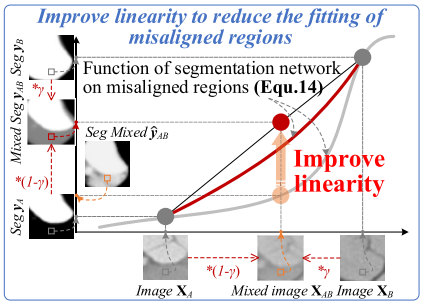

Fig. 6. Our MMR improves the robustness of segmentation training. In the view of optimization, it mixes two images to simulate the misalignment distortion and constrains the function of the DL model better linearity in the misaligned regions, thus reducing the fitting of this inaccurate information.

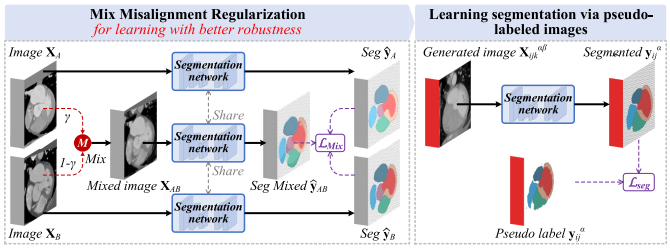

Fig. 7. Segmentation network learns the MIS via the constraints of our MMR $\mathcal{L}_{\text{Mix}}$ and the optimization of a segmentation loss $\mathcal{L}_{\text{seg}}$.

with better linearity on misaligned regions, thus reducing the risk of fitting for the inaccurate information. The misaligned regions of two images and their segmented maps $(\mathbf{X}_A, \hat{\mathbf{y}}_A)$, $(\mathbf{B}, \hat{\mathbf{y}}_B)$ are two points (gray) on the function of our segmentation network $\mathbb{S}(\cdot)$. Mixing these regions from two images and their segmented maps via weighted summation achieves sampling on the line connecting these two points for a mixed image and mixed segmented map [red point, $(\mathbf{X}_{AB}, \mathbf{y}_{AB})$]. Due to the different classes of the misaligned regions in the two images, the response of these regions will be lower than the aligned regions. Therefore, our MMR constrains the segmented mixed map $\hat{\mathbf{y}}_{AB}$ and the mixed segmented map $\mathbf{y}_{AB}$ to be consistent, resulting in the function's range between two points $(\mathbf{X}_A, \hat{\mathbf{y}}_A)$, $(\mathbf{X}_B, \hat{\mathbf{y}}_B)$ better linearity. This process is

$$\mathbb{S}(\gamma\,\mathbf{X}_A + (1-\gamma)\mathbf{X}_B) \Leftrightarrow \gamma\,\mathbb{S}(\mathbf{X}_A) + (1-\gamma)\mathbb{S}(\mathbf{X}_B). \quad (14)$$

In our experiment, the random value γ is sampled from a Beta distribution Beta(a, b) whose two parameters are all 0.3. If the mixed regions are aligned, their segmented maps will have the same value, and thus, the aligned regions in the mixed segmented map will be similar to their original value, which does not influence the fitting of such aligned regions. Therefore, the final fit function of the segmentation network will be simpler to reduce the fitting degree of inaccurate labels so that the trained segmentation network will have better performance.

The segmentation network is also optimized by a segmentation loss $\mathcal{L}_{\text{seg}}$ via the generated image–label pairs $(\mathbf{X}_{ijk}^{\alpha\beta}, \mathbf{y}_{ij}^{\alpha})$. As shown in Fig. 7, the segmentation loss $\mathcal{L}_{\text{seg}}$ evaluates the similarity between the segmented generated map $\hat{\mathbf{y}}_{ij}^{\alpha}$ and the generated label $\mathbf{y}_{ij}^{\alpha}$ to optimize the model for segmentation. In this article, we take the multiclass Dice coefficient [1] as the matric for the segmentation loss.

In general, the learning process is formulated as minimizing an expected loss function $\mathcal{L}_S(\cdot,\cdot,\cdot,\cdot,\cdot)$

$$\hat{\mathbb{S}} = \arg\min_{\mathbb{S}} \Big[\mathbb{E}_{((\mathbf{A}_i,\mathbf{y}_i),\mathbf{U}_j,\mathbf{U}_k,\mathbf{X}_A,\mathbf{X}_B)\sim\mathcal{D}} \times \big[\mathcal{L}_{\mathbb{S}}\big(\hat{\mathbf{y}}_{ij}^{\alpha}, \mathbf{y}_{ij}^{\alpha}, \hat{\mathbf{y}}_{AB}, \hat{\mathbf{y}}_A, \hat{\mathbf{y}}_B\big)\big]\Big]. \quad (15)$$

The complete loss for our segmentation network is

$$\mathcal{L}_{\mathbb{S}}\big(\hat{\mathbf{y}}_{ij}, \mathbf{y}_{ij}, \hat{\mathbf{y}}_{AB}, \hat{\mathbf{y}}_A, \hat{\mathbf{y}}_B\big) = \omega_0 \mathcal{L}_{\text{Mix}}\big(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_B, \hat{\mathbf{y}}_{AB}\big) + \mathcal{L}_{\text{seg}}\big(\hat{\mathbf{y}}_{ij}^{\alpha}, \mathbf{y}_{ij}^{\alpha}\big) \quad (16)$$

where ω_0 is the weight of the mix loss in complete loss.

Discussion of the Innovation: Our MMR proposes a novel data-driven regularization method without any labels. It constrains the segmentation network to reduce the fitting degree of misaligned regions in the generated data in a data-driven way. Therefore, it makes regularization for inaccurate information, coping with the limitation of robustness in the segmentation learning of the LRLS paradigm.

IV. Experiments Configurations

A. Datasets

We evaluate the excellent few-shot MIS performance of our BRBS framework on two important tasks with different imaging modalities and characteristics.

1) Few-shot cardiac structures segmentation [41] evaluates our framework on seven big cardiac structures on computed tomography (CT) images. Three public available datasets are introduced into our evaluation, including the MM-WHS [41], which has 20 images with cardiac structures' labels and 40 unlabeled images, ASOCA [42], which has 60 unlabeled images, and CAT08 [43], which has 32 images with cardiac structures' labels from.[1] Totally, 52 labeled and 100 unlabeled images are used in our evaluation. We crop the cardiac regions and resample them to 144 × 144 × 128. For our five-shot evaluation, we randomly select five labeled images as our atlas set and the remaining 47 labeled images as our testing set. For our one-shot evaluation, we randomly select one labeled image as our atlas and the remaining 51 labeled images as our testing set. We put the 100 unlabeled images as the unlabeled set.
2) Few-shot brain tissues segmentation [44] evaluates our framework on 28 small brain tissues on T1 magnetic resonance (MR) images. The CANDI dataset [45], which has 103 T1 brain MR images with brain tissues' labels, is used to evaluate our BRBS. Following [6], we crop a 160 × 160 × 128 volume from the center of the original images and split 20, 82, and one images as the test, unlabeled, and atlas sets in one-shot evaluation, respectively. For a five-shot evaluation, we resplit 20, 78, and five images as the test, unlabeled, and atlas sets, respectively. We perform the rotation in $[-20°, 20°]$ and scaling in [0.75, 1.25] times for all training processes as data augmentation.

[1] http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs/

1) Comparison Settings: To demonstrate the superiority of our proposed framework, we compare our BRBS framework with 16 widely used frameworks in the one- and five-shot situations.

1) Direct learning segmentation (LS) frameworks (3-D U-Net (2016) [13], SegNet (2017) [14], U-Net++ (2019) [15], and DBN (2020) [16]) are evaluated to give a basic performance of supervised MIS models in our few-shot situation.
2) Semisupervised LS (SLS) frameworks (UA-MT (2019) [46], MASSL (2019) [47], DPA-DBN (2020) [16], and CPS (2021) [48]) are compared to show the limitation of semisupervised methods due to the lack of domain knowledge in the few-shot situation.
3) Atlas-based medical image segmentation (ABS) frameworks with voting fusion (VoxelMorph (VM, 2018) [11], label-constrained VM (LC-VM, 2019) [26], and LT-Net (2020) [5]) are compared to show the limitation caused by the dissimilarity between the atlases and targeted images.
4) State-of-the-art LRLS-based methods (DeepAtlas (2019) [1], DataAug (2019) [2], DeepRS (2020) [3], PC-Reg-RT (2021) [4], and VAEAug (2021) [6]) are compared to demonstrate our excellent performance. Except for the LS, we use the same backbone, 3-D U-Net [13], for all methods to avoid the interference caused by network architecture.

2) Implementation Details: Our BRBS framework is implemented by PyTorch [17] on NVIDIA GeForce RTX 3090 GPUs with 24-GB memory. We set the values α, β, and γ to be sampled from a Beta distribution $Beta(a, b)$ whose two parameters are all 0.3 so that the probability distribution will be closer to 0 and 1 for a smaller distortion risk. We set $\lambda_0 = 1$ (following [26]), $\lambda_1 = 0.1$ (following [40]), and $\lambda_2 = 100$ (following Section V-D) for our registration network to achieve a semantic-aligned and topology-preserved registration. We further set $\omega_0 = 0.1$ (following Section V-D) to balance the training of our segmentation network. Following [4], we take the Adam whose learning rate is 1×10^{-4} to optimize our framework for fast convergence, and the batch size $= 1$ to save the memory for large 3-D images. Following [26], we perform an affine transformation on these images via AntsPy[2] to normalize the spatial position.

3) Evaluation Metrics: We evaluate our segmentation performance in two aspects [49]: **a)** The area-based metric: We take the Dice coefficient (DSC)[%] to evaluate the area-based overlap index. The higher DSC means the better region coincidence. **b)** The distance-based metric: We take the average Hausdorf distances (AVDs) to evaluate the coincidence of the surface. The lower AVD means the better surface coincidence. For further evaluation of our registration in our analysis, we calculate the percent of the Jacobian matrix <0 on each deformed voxel $J_\phi < 0$ to evaluate the distortion degree of deformation (the lower, the better) and calculate the DSC and AVD to evaluate the accuracy of deformation.

[2]https://github.com/ANTsX/ANTsPy

V. Results and Analysis

A. Quantitative Evaluation for Metric Superiority

The LRLS-based methods demonstrate their great superiority in few-shot MIS tasks (Table II). Three interesting observations can be found in Table II. **1)** SLS-based methods generally yield lower performance than the supervised LS-based methods. This is because the extremely few labels only provide little supervision, thus making extremely unreliable supervision information in their semisupervised setting and resulting in UA-MT's 17.9% DSC decrease compared with 3-D U-Net in the five-shot task (a). Although the CPS achieves the 3.1% DSC improvement in the five-shot task (a), it almost has no segmentation ability (25.3% and 37.1% DSC) on task (b) whose structures are really small. **2)** ABS methods have stable performance for small structures due to the consistency of the anatomical basic structures in medical images, so the VM achieves the third highest 83.1% DSC in the five-shot task (b). However, they are limited by the similarity between the atlas image and the test images, and the VM only has 81.0% DSC in the five-shot task (b) showing their bottleneck. **3)** The LRLS-based methods achieve general improvement over LS- and SLS-based methods due to their great representation ability of DL and the generated large amount of training data. Therefore, they have all achieved over 80% DSC and over 70% DSC in the five-shot tasks (a) and (b).

Compared with other LRLS-based models, our BRBS demonstrates its powerful MIS performance and large superiority in the few-shot situation in two aspects: **1)** Our BRBS has great MIS ability on both large [task (a)] and small structures [task (b)]. In the five-shot situation, it achieves the highest DSC (91.1%, 87.2%), and the lowest AVD (0.93 mm, 0.43 mm) on two tasks due to the great diversity and authenticity of generated images. The PC-Reg-RT achieves similar performance to our framework on large structures [task (a)], but it has very poor performance in small structures (task (b), more than 10% DSC lower than ours). This is because its serious mis-segmentation on small structures enlarges the alignment distortion in registration, making further degradation of segmentation in turn. The DataAug performs competitive performance on small structures (task (b), only 4.4% and 3.3% DSC lower than ours), but it is extremely limited on large structures [task (a)] due to the lack of semantic-alignment constraints and the alignment distortion caused by its poor registration making poor performance. **2)** Our BRBS has the best MIS ability in both one- and five-shot situations, illustrating our great robustness on label amounts. In task (a), the DeepRS has 87.0% DSC, which is only 4.1% lower than ours in the five-shot task (a), but this gap is enlarged to 15.8% in the one-shot task (a) due to the adversarial training, which enlarges the instability of data amount. The PC-Reg-RT has 88.5% DSC in the five-shot situation, which is only 2.6% DSC lower than ours, but in the one-shot situation, the gap is enlarged to 3.7% DSC. This is because the segmentation network in PC-Reg-RT extremely interferes with the registration network, so when the labeled images are extremely reduced, the degeneration of segmentation will bring bigger registration errors in generated data, further interfering with the segmentation learning.

TABLE II

QUANTITATIVE EVALUATION DEMONSTRATES THE ADVANTAGES OF OUR BRBS ON OUR TWO FEW-SHOT MIS TASKS IN THE ONE- AND FIVE-SHOT SITUATIONS. OUR BRBS ACHIEVES THE BEST PERFORMANCE BOTH ON LARGE (CARDIAC STRUCTURES) AND SMALL (BRAIN TISSUES) STRUCTURES COMPARED WITH 14 POPULAR METHODS. THE "UNABLE" MEANS THAT THE EXTREMELY POOR SEGMENTED RESULTS MAKE THE AVD UNABLE TO BE CALCULATED

Method	Type	(a)				(b)			
		1-shot$_{\pm std}$		5-shot$_{\pm std}$		1-shot$_{\pm std}$		5-shot$_{\pm std}$	
		DSC% ↑	AVDmm ↓	DSC% ↑	AVDmm ↓	DSC% ↑	AVDmm ↓	DSC% ↑	AVDmm ↓
3D U-Net [13]	LS	63.8$_{\pm 16.3}$	6.13$_{\pm 3.46}$	84.3$_{\pm 9.6}$	2.43$_{\pm 2.14}$	54.4$_{\pm 10.8}$	2.94$_{\pm 1.23}$	69.5$_{\pm 8.8}$	1.59$_{\pm 0.84}$
SegNet [14]	LS	57.5$_{\pm 17.4}$	7.01$_{\pm 4.53}$	78.8$_{\pm 10.5}$	2.68$_{\pm 1.72}$	52.3$_{\pm 4.9}$	3.18$_{\pm 0.37}$	62.7$_{\pm 7.0}$	1.98$_{\pm 0.72}$
U-Net++ [15]	LS	42.9$_{\pm 20.5}$	9.18$_{\pm 3.78}$	84.0$_{\pm 8.6}$	2.51$_{\pm 2.26}$	51.2$_{\pm 10.6}$	2.33$_{\pm 1.06}$	66.4$_{\pm 12.7}$	2.02$_{\pm 1.62}$
DBN [16]	LS	48.8$_{\pm 16.5}$	10.70$_{\pm 4.10}$	78.9$_{\pm 12.0}$	3.90$_{\pm 3.12}$	23.5$_{\pm 15.9}$	13.83$_{\pm 7.26}$	80.2$_{\pm 5.6}$	0.92$_{\pm 0.30}$
UA-MT [46]	SLS	54.8$_{\pm 17.0}$	9.44$_{\pm 4.77}$	66.4$_{\pm 16.2}$	4.69$_{\pm 2.27}$	36.7$_{\pm 8.4}$	8.69$_{\pm 2.29}$	75.5$_{\pm 3.4}$	1.31$_{\pm 0.95}$
CPS [48]	SLS	70.7$_{\pm 9.4}$	4.01$_{\pm 1.73}$	87.4$_{\pm 5.4}$	1.40$_{\pm 0.76}$	25.3$_{\pm 1.2}$	unable	37.1$_{\pm 1.8}$	unable
MASSL [47]	SLS	57.2$_{\pm 12.5}$	13.86$_{\pm 3.16}$	77.4$_{\pm 8.7}$	9.07$_{\pm 3.11}$	74.0$_{\pm 3.1}$	1.32$_{\pm 0.35}$	80.5$_{\pm 3.1}$	0.92$_{\pm 0.43}$
DPA-DBN [16]	SLS	49.0$_{\pm 14.4}$	10.47$_{\pm 3.81}$	68.0$_{\pm 14.5}$	5.75$_{\pm 3.89}$	28.1$_{\pm 7.6}$	7.75$_{\pm 1.78}$	68.7$_{\pm 8.2}$	3.90$_{\pm 2.39}$
VM [11]	ABS	77.6$_{\pm 6.0}$	2.49$_{\pm 0.73}$	81.0$_{\pm 6.1}$	2.13$_{\pm 0.78}$	78.7$_{\pm 1.8}$	0.73$_{\pm 0.07}$	83.1$_{\pm 1.8}$	0.56$_{\pm 0.08}$
LC-VM [26]	ABS	-	-	81.7$_{\pm 6.0}$	2.04$_{\pm 0.77}$	-	-	83.0$_{\pm 1.8}$	0.56$_{\pm 0.07}$
LT-Net [5]	ABS	67.2$_{\pm 6.5}$	3.55$_{\pm 0.90}$	77.8$_{\pm 7.8}$	2.25$_{\pm 0.95}$	76.9$_{\pm 1.5}$	0.75$_{\pm 0.51}$	82.6$_{\pm 1.2}$	0.57$_{\pm 0.05}$
DeepAtlas [1]	LRLS	85.4$_{\pm 4.5}$	1.59$_{\pm 0.56}$	87.9$_{\pm 4.3}$	1.30$_{\pm 0.57}$	73.0$_{\pm 2.4}$	1.02$_{\pm 0.10}$	79.3$_{\pm 2.6}$	0.74$_{\pm 0.12}$
DataAug [2]	LRLS	81.4$_{\pm 5.2}$	2.23$_{\pm 0.67}$	82.2$_{\pm 5.2}$	2.04$_{\pm 0.73}$	81.3$_{\pm 1.4}$	0.69$_{\pm 0.06}$	83.9$_{\pm 1.2}$	0.55$_{\pm 0.06}$
DeepRS [3]	LRLS	73.4$_{\pm 12.3}$	3.40$_{\pm 1.92}$	87.0$_{\pm 5.0}$	1.60$_{\pm 0.90}$	55.9$_{\pm 12.0}$	1.81$_{\pm 0.91}$	73.0$_{\pm 5.9}$	0.93$_{\pm 0.25}$
PC-Reg-RT [4]	LRLS	85.5$_{\pm 4.7}$	1.55$_{\pm 0.63}$	88.5$_{\pm 4.9}$	1.23$_{\pm 0.72}$	66.9$_{\pm 3.6}$	1.38$_{\pm 0.19}$	73.1$_{\pm 3.1}$	1.09$_{\pm 0.17}$
VAEAug [6]	LRLS	75.5$_{\pm 11.0}$	4.29$_{\pm 2.12}$	-	-	74.8$_{\pm 12.2}$	1.71$_{\pm 2.71}$	-	-
Our BRBS	LRLS	**89.2**$_{\pm 3.4}$	**1.24**$_{\pm 0.50}$	**91.1**$_{\pm 3.9}$	**0.93**$_{\pm 0.57}$	**85.7**$_{\pm 1.0}$	**0.49**$_{\pm 0.04}$	**87.2**$_{\pm 1.0}$	**0.43**$_{\pm 0.05}$

TABLE III

ABLATION STUDIES ON THE FIVE-SHOT CARDIAC STRUCTURES SEGMENTATION TASK DEMONSTRATE THE GREAT CONTRIBUTIONS OF OUR INNOVATIONS. WE EVALUATE THE SEGMENTATION AND REGISTRATION PERFORMANCE SIMULTANEOUSLY

KCC		S3P			MMR	Segmentation		Registration		
SeC	BiC	Image	Space	Style		DSC% ↑	AVDmm ↓	DSC% ↑	AVDmm ↓	$\lvert J_\phi\rvert \leq 0\%$ ↓
						84.3$_{\pm 9.6}$	2.43$_{\pm 2.14}$	-	-	-
		✓				80.7$_{\pm 9.6}$	2.52$_{\pm 1.52}$	72.6$_{\pm 13.8}$	2.89$_{\pm 1.18}$	3.3$_{\pm 0.7}$
		✓	✓			83.7$_{\pm 8.0}$	2.33$_{\pm 2.03}$	73.2$_{\pm 13.8}$	2.84$_{\pm 1.17}$	3.2$_{\pm 0.7}$
		✓	✓	✓		88.1$_{\pm 4.7}$	1.25$_{\pm 0.63}$	73.0$_{\pm 13.9}$	2.87$_{\pm 1.20}$	3.5$_{\pm 0.8}$
		✓	✓		✓	84.4$_{\pm 6.5}$	1.85$_{\pm 0.89}$	73.5$_{\pm 13.8}$	2.84$_{\pm 1.18}$	3.7$_{\pm 0.8}$
✓		✓	✓	✓		90.0$_{\pm 3.8}$	1.04$_{\pm 0.49}$	85.9$_{\pm 13.5}$	1.33$_{\pm 0.67}$	6.2$_{\pm 1.2}$
✓	✓	✓	✓	✓		90.4$_{\pm 3.4}$	1.00$_{\pm 0.44}$	86.0$_{\pm 13.5}$	1.31$_{\pm 0.64}$	2.5$_{\pm 1.1}$
✓	✓	✓	✓	✓	✓	**91.1**$_{\pm 3.9}$	**0.93**$_{\pm 0.57}$	**86.7**$_{\pm 13.6}$	**1.22**$_{\pm 0.62}$	**1.7**$_{\pm 0.8}$

B. Qualitative Evaluation for Visual Superiority

We show the results of five LRLS-based methods on five-shot cardiac structures and brain tissue segmentation tasks (Fig. 8). Our BRBS has high accuracy on the boundaries of large structures and fine segmentation quality on small structures. **1)** For large cardiac structures, our BRBS shows better segmentation accuracy on boundaries. As shown in the enlarged part of the first row, our BRBS and the DeepAtlas are able to finely segment the boundaries between two structures (yellow arrow), but the DataAug has a mis-segmentation in this region. This is because the DataAug has no semantic-alignment constraints in the registration, thus bringing too much alignment distortion on the boundaries when generating pseudo-labeled images via deformation. Although the DeepRS and the PC-Reg-RT have semantic-alignment constraints when training registration models, the interference of the adversarial process and the too much dependence on segmentation make their registration easy to distort on detail regions, resulting in poor segmentation accuracy on boundaries. **2)** For small brain tissues, our BRBS shows finer segmentation quality. The enlarged part in the last row has multiple adjacent small structures. Our BRBS has excellent integrity and consistency for these regions compared with other LRLS-based methods, due to our great authenticity of generated images. The PC-Reg-RT, DeepAtlas, and DeepRS seriously lose some small structures as pointed by the yellow arrow because their generation process did not consider authenticity, which is important for small structures. Once distorted, these tiny structures are easily disturbed by erroneous information, resulting in poor performance. DataAug performs better than the other three methods on small structures due to its three independent training stages that weaken the interference of mis-segmentation on registration, but it is still extremely limited by the misalignment.

C. Ablation Studies Show Improvements of Innovations

The ablation studies on the five-shot cardiac structures segmentation task demonstrate the great improvement of our innovations (Table III). The 3-D U-Net (first line, our baseline) only has 84.3% and 2.43 mm on DSC and AVD, respectively. When adding our S3P and only performing the image-level sampling, the large deformation distortion in the generated images reduces 3.6% DSC of 3-D U-Net. Our space-level

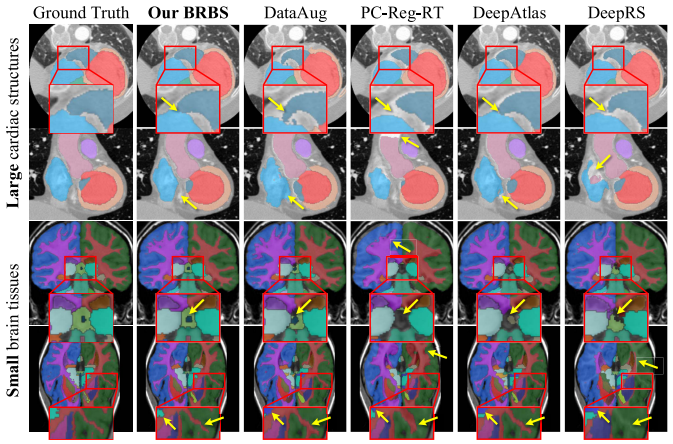

Fig. 8. Qualitative evaluation shows the visual superiority of our BRBS on our two few (five)-shot MIS tasks compared with four typical LRLS-based methods.

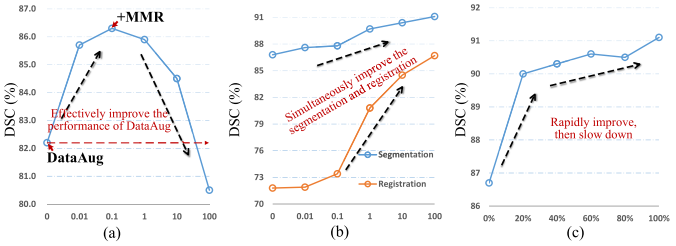

Fig. 9. Hyperparameters analysis on the five-shot cardiac structures segmentation task. Three hyperparameters are analyzed: (a) weight for our Mix loss ω_0 in our MMR, (b) weight for our SeC loss λ_2, and (c) size of the unlabeled dataset M.

sampling brings more diverse space features for 3.0% DSC improvement, but the distortion is still interfering with the segmentation learning, thus still only having 83.7% DSC. The style-level sampling makes 88.1% and 1.25 mm on DSC and AVD, respectively, which shows its large significance. This is because the variation of the style features caused by the different scanning protocols is a bottleneck for MIS and our style-level sampling provides more diverse style features for segmentation learning. When only adding our MMR (without style-level sampling), its regularization ability brings 0.7% DSC improvement. However, compared with the style-level sampling, it still lacks the directed supervision from the generated data with more style information, so it is still lower than the style-level sampling with 3.7% DSC. Our KCC brings semantic-aligned and topology-preserved registration for better authenticity in the generation. Our SeC constrains the registration network to produce better alignment on the same semantic regions, thus achieving 13.3% DSC improvement of registration and further having 1.9% DSC improvement of segmentation. However, $|J_\phi| \leq 0$ is enlarged to 6.2% due to the large deformation in the semantic field, distorting the topological structures. Our BiC further constrains the registration network to produce smoother and more invertible deformation so that it performs fewer folds and achieves 2.5% $|J_\phi| \leq 0$, which makes a 3.7% reduction. Due to the improvement of authenticity and diversity, our framework achieves 6.1% DSC improvement on segmentation than the 3-D U-Net. Our MMR provides the regularization for misaligned regions in generated pseudo-labeled images, further improving the robustness of the segmentation learning process. Therefore, our BRBS finally achieves 91.1% and 0.93 mm on DSC and AVD, respectively, with all our innovations, showing our powerful few-shot MIS performance.

D. Framework Analysis

1) Analysis of Hyperparameters: The analysis of three hyperparameters demonstrates the nature of our innovations (Fig. 9). **a)** Our MMR significantly improves the robustness of segmentation learning in DataAug [2] achieving higher performance, but when its weight is bigger than 0.1, it will reduce the performance. We evaluate our MMR on the DataAug model, which was extremely interfered with by the alignment distortion, to remove the interference of other innovations and evaluate our MMR clearly. When the weight ω_0 is 0.1, the DataAug model achieves the highest and significant improvement because our MMR performs a regularization, which reduces the fitting degree of inaccurate information. When enlarging the weight ω_0 to 100, our MMR extremely weakens the model's performance (even lower than the original DataAug). This is because the too powerful regularization makes the network difficult to fit the data, resulting in underfitting. **b)** Our SeC loss simultaneously improves the segmentation and the registration performance with the enlarging of its weight λ_2 (we set $\omega_0 = 0.1$). This is because the better alignment of semantic regions in the registration, the more authenticity in the generated data will be available. Therefore, the segmentation network trained by more authentic data will have better performance. **c)** With the enlarging of the unlabeled datasets' size, the MIS performance of our framework is rapidly improved with 20% of the unlabeled dataset and then slows down with more unlabeled data. This is because our S3P greatly improves the diversity of the generated data, and a small number of unlabeled images are able to generate the data with a large number of features. Therefore, it also illustrates that our framework is robust to the unlabeled data amount.

2) Analysis of Registration Performance: Compared with the registration models in other frameworks on the five-shot cardiac structures segmentation task, our BRBS has higher registration accuracy and low distortion owing to our KCC (Table IV). Two traditional (Trad) registration models (BSpline [50] and SyN [7]), two DL registration models (VM [11] and LC-VM [26]), and the registration part of three LRLS-based models (DeepAtlas [1], PC-Reg-RT [4], and DeepRS [3]) are compared in this section. The BSpline has 77.0%, 2.61 mm, and 8.4% on DSC, AVD, and $|J_\phi| \leq 0$, respectively, showing the basic registration performance on this task. The SyN is designed for diffeomorphic registration, thus achieving 0% $|J_\phi| \leq 0$ and its registration accuracy is improved to 79.2% DSC. The VM learns the registration via DL, but due to the lack of semantic perception ability, it only has 72.7% DSC, which is even lower than BSpline. The LC-VM trains VM only with additional five labels for

TABLE IV

COMPARED WITH THE REGISTRATION MODELS ON THE FIVE-SHOT CARDIAC STRUCTURES SEGMENTATION TASK, OUR BRBS HAS HIGHER REGISTRATION ACCURACY (DSC, AVD) AND LOWER DISTORTION ($|J_\phi| \leq 0$) DUE TO OUR KCC. THE "TRAD" IS THE TRADITIONAL METHOD AND "DL" IS THE DEEP LEARNING METHOD

Method	Type	DSC% ↑	AVDmm ↓	$\lvert J_\phi\rvert \leq 0$ % ↓
Initial		$62.1_{\pm 8.7}$	$4.37_{\pm 1.34}$	-
BSpline [50]	Trad	$77.0_{\pm 11.2}$	$2.61_{\pm 1.57}$	$8.4_{\pm 9.7}$
SyN [7]	Trad	$79.2_{\pm 8.9}$	$2.33_{\pm 0.92}$	**0**
VM [11]	DL	$72.7_{\pm 13.9}$	$2.89_{\pm 1.20}$	$3.5_{\pm 0.7}$
LC-VM [26]	DL	$73.0_{\pm 13.9}$	$2.85_{\pm 1.19}$	$3.1_{\pm 0.6}$
DeepAtlas [1]	DL	$79.7_{\pm 13.6}$	$2.10_{\pm 0.96}$	$2.5_{\pm 0.6}$
PC-Reg-RT [4]	DL	$80.8_{\pm 13.6}$	$1.94_{\pm 0.91}$	$0.5_{\pm 0.3}$
DeepRS [3]	DL	$76.9_{\pm 13.3}$	$2.39_{\pm 0.99}$	$5.3_{\pm 1.3}$
LT-Net [5]	DL	$68.2_{\pm 13.9}$	$3.26_{\pm 1.29}$	$2.5_{\pm 0.3}$
Our BRBS	DL	**86.7**$_{\pm 13.6}$	**1.22**$_{\pm 0.62}$	$1.7_{\pm 0.8}$

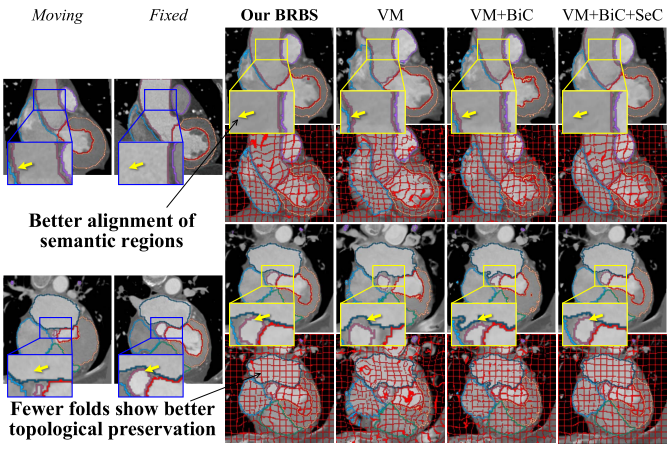

Fig. 10. Our KCC utilizes our BiC and SeC constraints, optimizing our registration network for great topological preservation (fewer folds) and semantic-alignment ability in registration.

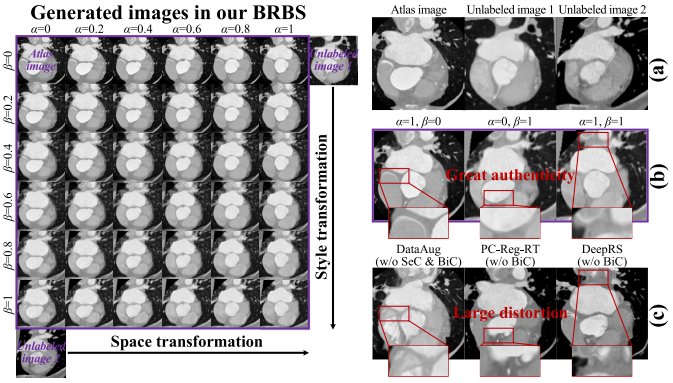

Fig. 11. Generated images of our BRBS have great diversity and authenticity. Left: our S3P densely samples images from the space and style transformation paths so that the features of the generated images will cover the transformation process for better diversity. Right: from (a) to (c), they are the images from the real world, the images from our S3P, and the images from other LRLS methods.

better semantic-alignment ability, achieving slight improvement (0.3% DSC) due to the too few labels. DeepAtlas takes a segmentation network for alignment of semantic regions, thus achieving a great 79.7% DSC on registration. PC-Reg-RT only performs deformation on the regions of interest, avoiding the potential distortion on large background, so it has 0.5% $|J_\phi| \leq 0$ showing great topology-preserving ability and having the second-highest performance. However, it has no deformation on the background regions, which lacks the alignment accuracy on some unlabeled but task-dependent regions. Our BRBS achieves the highest DSC (86.7%), lowest AVD (1.22 mm), and relatively low $|J_\phi| \leq 0$ (1.7%), demonstrating the powerful topology-preservation and semantic-alignment ability from our KCC. Such high alignment accuracy and the few deformation distortions in the registration will generate the data with great authenticity, training the segmentation with higher accuracy.

3) Analysis of Our KCC: Our KCC utilizes our BiC and SeC constraints and optimizes our registration network for topology-preserved and semantic-aligned registration (Fig. 10), reducing the registration error. Therefore, the generated images will have better authenticity, significantly improving the further learning of our MIS task. **a)** The VM has numerous folds and poor semantic alignment when deforming (Fig. 10), due to the lack of topology-preserved and semantic-aligned constraints. Therefore, as shown in Table IV, it has high $|J_\phi| \leq 0$ (3.5%) and low accuracy (72.7%). **b)** Our BiC utilizes the knowledge that a topology-preserved deformation performs an invertible and smooth mapping and bidirectional registration and smoothness constraints. Therefore, it achieves fewer folds in Fig. 10 and lower $|J_\phi| \leq 0$ (1.7%) in Table III. **c)** Our SeC utilizes the knowledge that two aligned images have aligned semantic regions and constrains the corresponding semantic regions in two images aligned via our segmentation network. Therefore, it achieves better alignment for semantics in Fig. 10 and higher DSC (86.7%) of registration in Table IV.

4) Analysis of the Authenticity of Generated Data: The generated images in our BRBS have great authenticity due to our BiC and SeC constraints [Fig. 11 (right)]. **a)** Compared with the images from the other LRLS-based methods [2]–[4], our BiC and SeC effectively improve the authenticity of our generated images. The DataAug has no constraints for semantic alignment and topological preservation, so the generated image has a large distortion and artifact on the boundaries. The PC-Reg-RT and DeepRS have targeted constraints for semantic alignment, but they make large distortions of topological structures due to the lack of constraints for topology preservation. Our KCC constrains the registration learning via two kinds of domain knowledge (BiC and SeC) for topology-preserved and semantic-aligned registration, thus generating the images with great authenticity. **b)** Compared with the images from the real world [Fig. 11(a)], the images generated from our BRBS have great authenticity with great semantic alignment and topological preservation so that the segmentation network trained with our generated data will learn a representation that matches real data. Therefore, the segmentation network will achieve great generalization for the testing data from the real world.

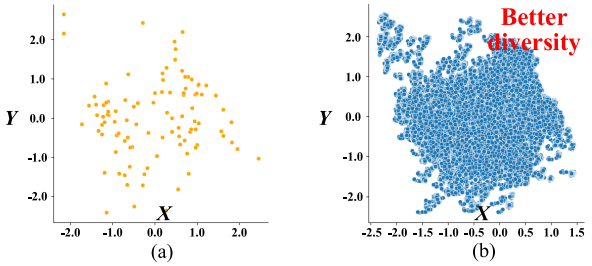

Fig. 12. Features from the bottleneck layer of the segmentation network (compressed to two values by PCA) show that our S3P provides much more diverse generated data for segmentation learning. (a) Without our S3P [1], only sparse generated data are available extremely limiting the diversity. (b) With our S3P, our BRBS has diverse and dense generated data for effective segmentation learning.

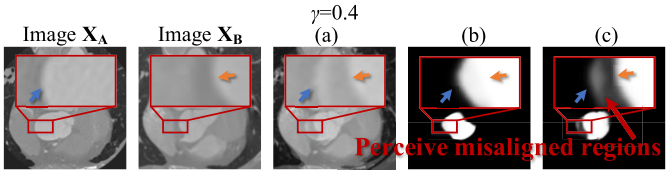

Fig. 13. Our MMR constrains the segmentation network to perceive the misaligned regions. (a) It mixes two images to simulate the misaligned regions in mixed images. (b) Without our MMR, the 3-D U-Net is unable to perceive the misaligned regions, fitting this inaccurate information. (c) With our MMR, our BRBS perceives the misaligned regions, thus reducing the fitting degree of such regions for better robustness.

5) Analysis of the Diversity in Generated Data: Our S3P provides much more diverse generated data for segmentation learning due to the dense sampling of the space and style transformation paths (Fig. 12). We demonstrate the distribution of the generated images via extracting their features from the bottleneck layer of a trained segmentation network and compressing these features to two values via the principal component analysis (PCA). **a)** Without our S3P [1], the generated images only have sparse feature distribution lacking the diversity of the image for segmentation training, and this is because the generation program directly samples data from sparse and discrete distribution from the unlabeled dataset so that the generated distribution is sparse and discrete limiting the diversity. **b)** Our S3P builds continuous distributions and densely samples the diverse space and style transformation maps in their transformation process, thus generating numerous images with diverse space and style features. Therefore, the feature distribution of the generated images from our S3P is dense and diverse, thus effectively improving the generalization of the segmentation network. **c)** Visually, the space and style features of the generated images cover the transformation process [Fig. 11 (left)]. With the enlarging of α and β, the space and style of the atlas image are gradually transformed into the two unlabeled images. Therefore, the features of the generated images will cover the transformation from the atlas images to the unlabeled images, for more diverse features in the generated dataset.

6) Analysis of the Regularization in Our MMR: Our MMR simulates the distortion of misalignment and constrains the network to perceive these regions to reduce the fitting degree of this inaccurate information (Fig. 13), thus achieving a regularization process. The images $\mathbf{X}_A$ and $\mathbf{X}_B$ are mixed via weighted summation so that the regions near the boundaries are misaligned (blue and yellow arrows) in mixed images that have distorted semantics, simulating the alignment distortion in our style-level sampling. Our MMR constrains the model to predict mixed segmentation maps and make a lower response on misaligned regions (c), thus improving the linearity of the function on these regions and achieving the robustness for such distortion in the generated images. The 3-D U-Net is unable to perceive the misaligned regions (b), so it has fit the inaccurate information and performs excessive confidence for the inaccurate segmentation of misaligned regions.

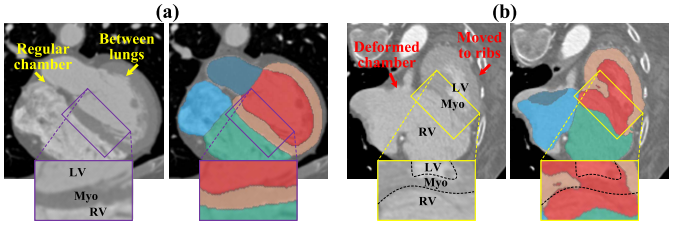

Fig. 14. Failure case study: the case whose structures and position are broken by the surgery made our BRBS that has failed segmentation between the LV and RV. (a) Regular case has regular chamber structures whose position is between lungs, having a very fine segmentation result. (b) However, the case after surgery has deformed chamber and is moved to ribs, resulting in distorted structures and context features in the image and making failed segmentation.

VI. Conclusion and Discussion

Our proposed BRBS framework provides a powerful few-shot benchmark for the field of MIS and efficiently reduces the costs of medical image research. It effectively advances the LRLS paradigm in diversity, authenticity, and robustness, strongly breaking through the large bottleneck of this paradigm. The proposed KCC further constrains the registration network to learn semantic-aligned and topology-preserved registration, thus allowing the generation program to output new data with great space and style authenticity. Our S3P introduces the modeling of the transformation paths of style and space changes between few atlases and numerous unlabeled images into the generation program, thus sampling for much more diverse space and style features in generated data. Our MMR finally simulates the misalignment distortion and constrains the network to reduce the fitting degree of the misaligned regions, thus improving the robustness of segmentation learning. Without any bells and whistles, the extensive experiments with state-of-the-art results on two challenging tasks reveal our powerful few-shot MIS performance and large application value. We also want to use BRBS on MindSpore,[3] which is a new DL computing framework. These problems are left for future work. All of our code will be made publicly available at https://github.com/YutingHe-list/BRBS.

A. Discussion of One Failure Cases and Our Potential Limitation

The failure case in our experiment demonstrates one of the potential limitations.

Failure Case Study: As shown in Fig. 14, the case after surgery made our BRBS has failed segmentation between the

[3]MindSpore: https://www.mindspore.cn/

LV and RA. Compared with the regular case [Fig. 14(a)] which has regular chamber structures and whose position is between lungs, the surgery makes the heart [Fig. 14(b)] have deformed chamber and be moved to ribs. Therefore, the features of this case are beyond the generalization range of the model trained for the images on regular cases.

Discussion of Limitation: Our framework is limited on the cases whose regular structures and position are broken manually, like the case in Fig. 14(b) whose structures and position are broken by the surgery. Our BRBS that is trained for the regular cases (never operated by surgery) will be unable to segment these broken cases whose structures and context features are out of the regular distribution. Fortunately, these broken cases are easy to be distinguished, and it is easy to overcome this limitation via only few simple doctor-assisted manual adjustments to the segmentation results. Our future work will also be devoted to promoting the solution of this limitation.

B. Discussion of Our Future Work

Our BRBS has demonstrated great superiority in medical images, and in the future, it also has great potential for further exploration. Therefore, our future works are mainly on two aspects.

1) Explore more registration-based efficient learning tasks in medical images. The registration utilizes the prior of anatomical consistency, which widely exists in numerous regular structured medical images (i.e., CT and MR), providing an effective way for the efficient learning of these images. Therefore, we will future study the paradigms that learning registration to learn other tasks such as quantification [51].
2) Explore the LRLS paradigm in domain adaptation [52]. The registration model is able to align the medical images with different styles (i.e., cross-modal registration [25]), having the ability to construct the relationship between different data domains. Once this cross-domain capability is integrated, the LRLS paradigm will have great potential for domain adaptation tasks. Therefore, we will further study the cross-domain capability in the LRLS paradigm, thus exploring it in domain adaptation tasks.

ACKNOWLEDGMENT

The authors thank the Big Data Computing Center, Southeast University, Nanjing, China, for providing the facility support on the numerical calculations.

REFERENCES

[1] Z. Xu and M. Niethammer, "Deepatlas: Joint semi-supervised learning of image registration and segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent*. Shenzhen, China: Springer, 2019, pp. 420–429.
[2] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2019, pp. 8543–8553.
[3] Y. He *et al.*, "Deep complementary joint model for complex scene registration and few-shot segmentation on medical images," in *Proc. 16th Eur. Conf. Comput. Vis.*, vol. 1, 2020, pp. 770–786.
[4] Y. He *et al.*, "Few-shot learning for deformable medical image registration with perception-correspondence decoupling and reverse teaching," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 1177–1187, Mar. 2022.
[5] S. Wang *et al.*, "LT-Net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9162–9171.
[6] Y. Ding, X. Yu, and Y. Yang, "Modeling the probabilistic distribution of unlabeled data for one-shot medical image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1246–1254.
[7] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, 2008.
[8] F. Liu *et al.*, "JSSR: A joint synthesis, segmentation, and registration system for 3D multi-modal image alignment of large-scale pathological CT scans," in *Computer Vision*. Glasgow, U.K.: Springer, Aug. 2020, pp. 257–274.
[9] M. Holden, "A review of geometric transformations for nonrigid body registration," *IEEE Trans. Med. Imag.*, vol. 27, no. 1, pp. 111–128, Jan. 2007.
[10] J. Ashburner, "A fast diffeomorphic image registration algorithm," *Neuroimage*, vol. 38, no. 1, pp. 95–113, 2007.
[11] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9252–9260.
[12] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 7, 2022, doi: 10.1109/TNNLS.2022.3152527.
[13] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent*. Athens, Greece: Springer, 2016, pp. 424–432.
[14] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[15] Z. Zhou *et al.*, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.
[16] Y. He *et al.*, "Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101722.
[17] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.
[18] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, p. 63, Jun. 2020.
[19] L. Zhu and Y. Yang, "Label independent memory for semi-supervised few-shot video classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 273–285, Jul. 2020.
[20] M. Zhang *et al.*, "One-shot neural architecture search: Maximising diversity to overcome catastrophic forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 2921–2935, Sep. 2021.
[21] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4077–4087.
[22] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.
[23] S. Puch, I. Sánchez, and M. Rowe, "Few-shot learning with deep triplet networks for brain imaging modality recognition," in *Domain Adaptation and Representation Transfer and Medical Image Learning With Less Labels and Imperfect Data*. Shenzhen, China: Springer, 2019, pp. 181–189.
[24] Y. He *et al.*, "Meta grayscale adaptive network for 3D integrated renal structures segmentation," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102055.
[25] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: A survey," *Mach. Vis. Appl.*, vol. 31, nos. 1–2, pp. 1–18, Feb. 2020.
[26] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.

[27] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, Feb. 2015.
[29] Y. He *et al.*, "DPA-DenseBiasNet: Semi-supervised 3D fine renal artery segmentation with dense biased network and deep priori anatomy," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* vol. 11769, 2019, pp. 139–147.
[30] M. S. Elmahdy, J. M. Wolterink, H. Sokooti, I. Isgum, and M. Staring, "Adversarial optimization for joint registration and segmentation in prostate ct radiotherapy," in *Medical Image Computing and Computer Assisted Intervention*. Shenzhen, China: Springer, 2019, pp. 366–374.
[31] N. K. Dinsdale, M. Jenkinson, and A. I. L. Namburete, "Spatial warping network for 3D segmentation of the hippocampus in mr images," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2019, pp. 284–291.
[32] H. Yang, J. Sun, H. Li, L. Wang, and Z. Xu, "Neural multi-atlas label fusion: Application to cardiac MR images," *Med. Image Anal.*, vol. 49, pp. 60–75, Oct. 2018.
[33] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.
[34] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. van Ginneken, "Multi-atlas-based segmentation with local decision fusion—Application to cardiac and aortic segmentation in CT scans," *IEEE Trans. Med. Imag.*, vol. 28, no. 7, pp. 1000–1010, Jul. 2009.
[35] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 611–623, Mar. 2013.
[36] W. Ding, L. Li, X. Zhuang, and L. Huang, "Cross-modality multi-atlas segmentation using deep neural networks," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2020, pp. 233–242.
[37] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, "Emotion classification with data augmentation using generative adversarial networks," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2018, pp. 349–360.
[38] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101759.
[39] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.
[40] J. Zhang, "Inverse-consistent deep networks for unsupervised deformable image registration," 2018, *arXiv:1809.03443*.
[41] X. Zhuang *et al.*, "Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101537.
[42] R. Gharleghi *et al.*, "Automated segmentation of normal and diseased coronary arteries—The ASOCA challenge," *Comput. Med. Imag. Graph.*, vol. 97, p. 102049, 2022.
[43] M. Schaap *et al.*, "Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms," *Med. Image Anal.*, vol. 13, no. 5, pp. 701–714, 2009.
[44] K. Payette *et al.*, "An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset," *Sci. Data*, vol. 8, no. 1, p. 167, Dec. 2021.
[45] D. N. Kennedy, C. Haselgrove, S. M. Hodge, P. S. Rane, N. Makris, and J. A. Frazier, "CANDIShare: A resource for pediatric neuroimaging data," *Neuroinformatics*, vol. 10, no. 3, pp. 319–322, 2012.
[46] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Medical Image Computing and Computer Assisted Intervention* (Lecture Notes in Computer Science). Shenzhen, China: Springer, 2019, pp. 605–613.
[47] S. Chen, G. Bortsova, A. G.-U. Juárez, G. van Tulder, and M. de Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2019, pp. 457–465.
[48] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 2613–2622.
[49] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, p. 29, Aug. 2015.
[50] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.
[51] R. Ge *et al.*, "K-Net: Integrate left ventricle segmentation and direct quantification of paired echo sequence," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1690–1702, May 2020.
[52] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 766–785, Mar. 2021.