Bi-level Learning of Task-Specific Decoders for Joint Registration and One-Shot Medical Image Segmentation

Xin Fan[*] Xiaolin Wang Jiaxin Gao Jia Wang Zhongxuan Luo Risheng Liu
School of Software Technology, Dalian University of Technology, Dalian, China
{xin.fan,zxluo,rsliu}@dlut.edu.cn, {wxl1009,wangjia}@mail.dlut.edu.cn,
jiaxinn.gao@outlook.com

Abstract

One-shot medical image segmentation (MIS) aims to cope with the expensive, time-consuming, and inherent human bias annotations. One prevalent method to address one-shot MIS is joint registration and segmentation (JRS) with a shared encoder, which mainly explores the voxel-wise correspondence between the labeled data and unlabeled data for better segmentation. However, this method omits underlying connections between task-specific decoders for segmentation and registration, leading to unstable training. In this paper, we propose a novel Bi-level Learning of Task-Specific Decoders for one-shot MIS, employing a pretrained fixed shared encoder that is proved to be more quickly adapted to brand-new datasets than existing JRS without fixed shared encoder paradigm. To be more specific, we introduce a bi-level optimization training strategy considering registration as a major objective and segmentation as a learnable constraint by leveraging inter-task coupling dependencies. Furthermore, we design an appearance conformity constraint strategy that learns the backward transformations generating the fake labeled data used to perform data augmentation instead of the labeled image, to avoid performance degradation caused by inconsistent styles between unlabeled data and labeled data in previous methods. Extensive experiments on the brain MRI task across ABIDE, ADNI, and PPMI datasets demonstrate that the proposed Bi-JROS outperforms state-of-the-art one-shot MIS methods for both segmentation and registration tasks. The code will be available at https://github.com/Coradlut/Bi-JROS.

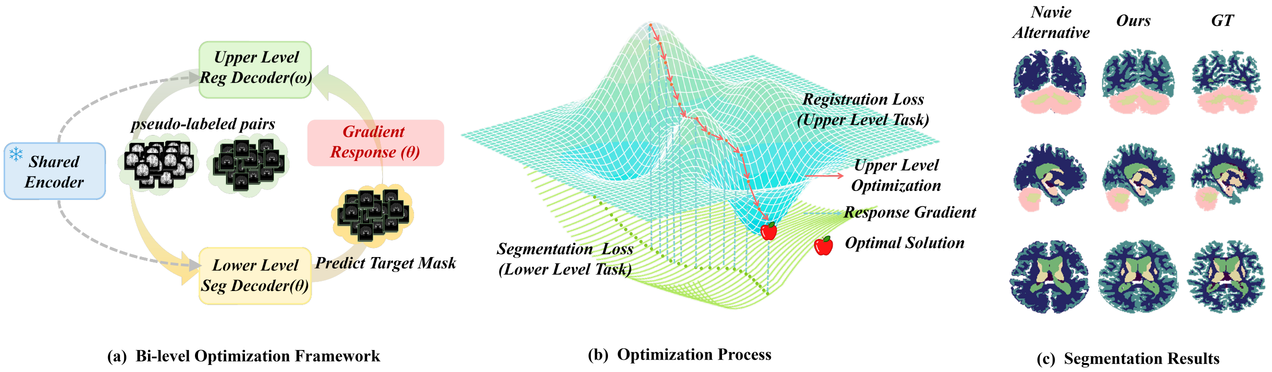
1. Introduction

Medical image segmentation (MIS), playing a key role in medical image analysis, is widely used in clinical scenarios, such as tumor detection, atlas construction, and organ quantification analysis [10, 12, 33, 34, 37]. One-shot MIS has demonstrated considerable potential for alleviating the need for extensive manual labeling. Zhao *et al.* [42] employ registration techniques to learn the spatial transformations offline between a single annotated template image and all unannotated images. This approach enables the model to propagate annotations from one image to multiple unannotated images, thereby reducing the dependency on extensive manual annotation. Conversely, many registration methods [3, 13, 25] incorporate segmentation maps as auxiliary information to aid the registration model in aligning anatomical features more accurately, particularly in areas where these features are less discernible in original images. These methods consider registration and segmentation as two independent tasks, and how to effectively combine these two tasks remains a challenge.

Recent methods [18, 23, 36, 41] have introduced a novel Joint Registration and Segmentation (JRS) paradigm, which combines medical image registration with segmentation, fostering a reciprocal enhancement of both tasks. Specifically, the output of registration serves as an input to the segmentation model, and conversely, the output from segmentation guides and constrains the learning trajectory of the registration model. This symbiosis enhances the segmentation accuracy through improved spatial consistency afforded by registration, while precise segmentation yields additional structured information to direct the registration model toward a more accurate alignment of anatomical features. However, most JRS methods [18, 39, 41, 44] tend to build two independent encoder-decoder structures to perform the registration and segmentation tasks, leading to a large increase in the overall parameters of the model and structural redundancy. Several methods [1, 11, 43] have proposed the use of a shared encoder to concurrently learn both tasks, enabling rapid and accurate registration and segmentation in a single inference.

However, the approach of employing a shared encoder still encounters two challenges: i) Inadequate use of deformed image features: existing shared encoder meth-

[*] Corresponding author



(a) Bi-level Optimization Framework

(b) Optimization Process

(c) Segmentation Results

Figure 1. (a) The bi-level optimization framework which establishes the coupling dependencies between registration- and segmentation-specific decoders. (b) Illustrating the bi-level optimization process with the feedback of segmentation optimization to the registration learning process. (c) Visualization of segmentation results.

ods [11, 43] all utilize the deformed segmentation map to guide the learning of the segmentation decoder, which limits the potential of the registration task to increase data diversity. ii) Training stability: existing methods [1, 43] mainly use a naive jointly or an alternative optimization training strategy and fall short in precisely characterizing or acknowledging the intricate coupling relationships between tasks, e.g., alternate training optimizes network parameters of a task while fixing another. It hinders the model's capacity to adequately capture and exploit the dynamic changes and interrelationships between tasks and leads to a continual adaptation to inaccurately misaligned areas, thereby initiating an unstable learning process.

In this paper, we propose a bi-level optimization learning framework to model the coupled dependencies between task-specific decoders for registration and segmentation, thereby guiding a collaborative optimization process to stably converge to an optimum. Our framework, comprising a fixed shared encoder and two task-specific decoders, leverages a fixed shared encoder in multi-task learning to markedly boost computational efficiency and speed up training upon transitioning to new datasets. In Fig. 1(a), with the registration decoder positioned as the upper-level task and the segmentation decoder as the lower-level task. Simultaneously, we introduce an appearance conformity constraint that indirectly utilizes the template image to increase data diversity and integrate the constraint into the segmentation task. Our bi-level optimization framework is capable of deriving a cooperative training algorithm that encapsulates step-wise coupled gradient responses, as opposed to the traditional simple alternating iterative method without inter-task interaction. Fig. 1(b) illustrates the optimization process with the gradient response from the lower level. Fig. 1(c) presents the visual results of our method compared with those obtained using a simple alternating approach. Our contributions can be summarized as follows:

- We propose a **B**i-level optimization-based framework for **J**oint **r**egistration and **O**ne-shot **S**egmentation, termed as Bi-JROS, which precisely characterizes the coupling constraints between decoders specific to registration and segmentation tasks.
- We design an iterative **G**radient **R**esponse (GR) algorithm to tackle the nested bi-level optimization challenge. It leverages the gradient response of the segmentation decoder to the registration decoder during each step of the optimization process, ensuring more effective and stable training compared to simple alternating learning strategy.
- We propose an **A**ppearance **C**onformity **C**onstraint (ACC) to avoid the texture gap between target and atlas images and increase the diversity of the data. This is integrated into the segmentation task to strengthen the interconnection between registration and segmentation.

2. Related Works

2.1. One-shot medical image segmentation

The primary aim of one-shot Medical Image Segmentation (MIS) [15, 16, 18, 38, 41, 42] is to leverage the unsupervised registration to help the supervised segmentation task. Such methods learn the voxel-wise correspondence from the labeled to the unlabeled data, thereby indirectly constructing the fake label for unlabeled data to perform segmentation, which is a promising MIS paradigm and many studies have emerged. Zhao et al. [42] introduced DataAug, which incorporates an appearance transformation within this paradigm to generate diverse data. Xu *et al.* [41] employed the segmentation model to constrain the registration generating higher-quality data. Wang *et al.* [38] proposed LT-Net, which improved spatial transformations for

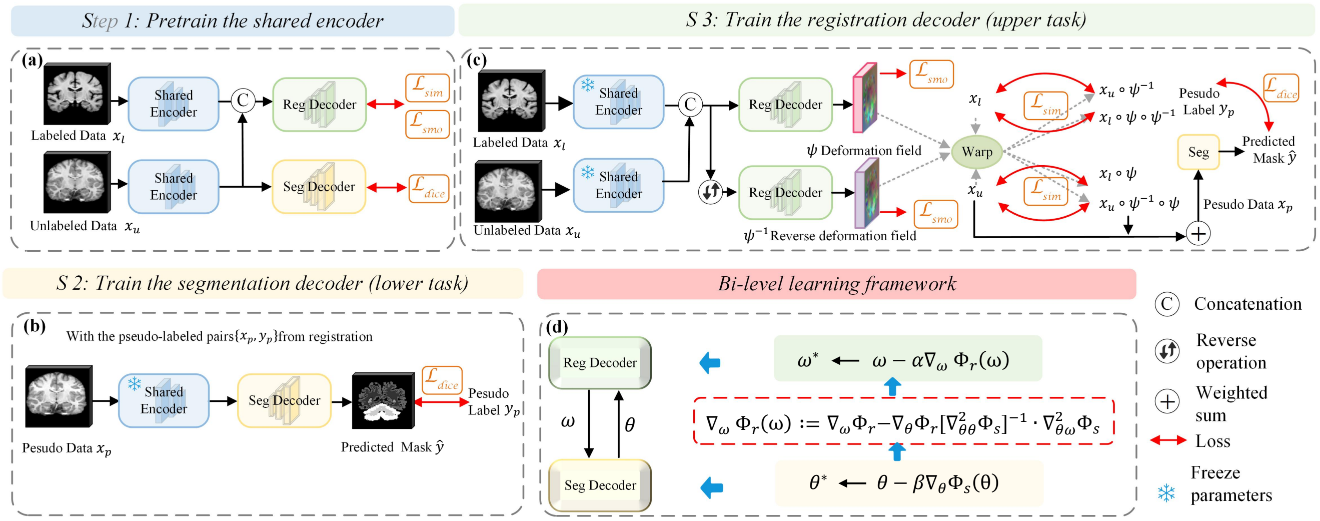


Figure 2. Overall framework of the proposed Bi-JROS. (a) demonstrates the pretraining process of the shared encoder, (b) and (c) together constitute the bi-level optimization learning phase and (d) illustrates the mechanism of gradient updating.

data augmentation and adds forward-backward consistency to boost registration. Ding *et al.* [9] modeled the probability distribution of unlabeled medical images using VAEs for one-shot image segmentation. He *et al.* [17] enhanced model pre-training in an unlabeled setting by exploring geometric visual similarity in medical images, significantly impacting subsequent tasks like segmentation. More recently, He *et al.* [18] proposed BRBS from the perspective of authenticity, diversity, and robustness of the one-shot MIS, achieving a remarkable improvement.

Despite promising performance achieved, the inherent issue with these existing methods is that different encoder is employed in their framework to extract similar feature representations resulting in time-consuming. To alleviate this problem, Zhao *et al.* [43] first proposed to learn shared features for both registration and segmentation tasks by one encoder module achieving favorable performance. Recently, Andresen *et al.* [1] proposed the joint non-correspondence segmentation and image registration network to handle the problem of missing correspondences caused by inter-patient variations. Different from the previous methods, we adopt a fixed encoder trained by datasets with various styles to learn shared features for both tasks, which can find the optimal solution faster when using other datasets.

2.2. Bi-level optimization

Bi-level Optimization is the hierarchical mathematical program where the feasible region of the upper-level task is restricted by the solution set mapping of the lower-level task and the two tasks are mutually reinforced [24, 27]. Subsequently, the bi-level optimization framework has been investigated in view of many important applications in the fields of machine learning and computer vision e.g., hyper-parameter optimization [21, 28], multi-task and meta-learning [26, 28]. Motivated by the above observations, we construct a bi-level optimization that can help address mutual learning by explicitly considering the impact of the follower segmentation task on the leader registration task during the optimization process.

3. Method

3.1. Overview

In one-shot medical image segmentation (MIS) scenario, the training data comprises a single labeled image pair, denoted by $(\mathbf{x}_l,\ \mathbf{y}_l)$, and a substantial volume of unlabeled data, denoted by $D_{se} = (\mathbf{x}_u)_{i=1}^N$ for the shared encoder and $D_{de} = (\mathbf{x}_u)^M$ for the task-specific decoders, where N and M denote the medical volumes with $N \gg M$ and $\mathbf{y}_l$ being the ground truth. Our goal is to improve registration and segmentation accuracy and facilitate rapid adaptation of the model to new datasets by adopting joint registration and segmentation (JRS) paradigm with a shared encoder. To this end, we proposed a **Bi**-level optimization-based framework for **J**oint **R**egistration and **O**ne-shot medical image **S**egmentation, termed as Bi-JROS. The training process of Bi-JROS is shown in Fig. 2 (a) to (c), consisting of two phases: we first train the shared encoder by using datasets D_{se} with different styles in a joint training way as shown in Fig. 2 (a). Then, we employ the pretrained shared encoder with fixed parameters Ω to train two task-specific decoders for segmentation and registration as shown in Fig. 2 (b) and (c). Moreover, we proposed a bi-level modeling to enable stable training for these two task-specific decoders, which has been proven to be effective through the broad range of experiments mentioned in Sec. 4.3.

Appearance conformity constraint strategy. In JRS paradigm, registration serves as an auxiliary learning process to produce a deformation filed ψ^i from $\mathbf{x}_l$ to an unlabeled image $\mathbf{x}_u^i$ and generates a pseudo-labeled pair $\left\{\mathbf{x}_p^i, \mathbf{y}_p^i\right\}$ by applying ψ^i to warp the labeled image pair. However, the appearance gap between labeled image and unlabeled data will result in a mismatch between deformed labeled and unlabeled data. Such mismatch, if not handled appropriately, will be further magnified by the subsequent segmentation. Inspired by [18, 22, 38], we propose an appearance conformity constraint strategy (ACC) to generate robust pseudo-labeled pairs, aiming to simplify the operational process while ensuring algorithm performance, which predicts bidirectional deformation *avoiding directly using labeled image to perform data augmentation*. Specifically, We first predict the deformation filed ψ^i from the labeled image to unlabeled data. Different from the previous methods that directly perform data augmentation on labeled image, we re-perform registration to obtain the reverse deformation field $(\psi^i)^{-1}$ to obtain the *pseudo-labeled image*, then perform spatial transformation on the fake labeled image instead of directly using labeled image. Regarding the inevitable noise in the re-warped fake labeled image, we employ a weighted fusion operation that integrates information with the unlabeled data $\mathbf{x}_u$. The above process can be summarized as

$$\psi^i = f_{rd}(F_l^i; F_u^i), (\psi^i)^{-1} = f_{rd}(F_u^i; F_l^i), \tag{1}$$

$$\mathbf{x}_p^i = \gamma \cdot \mathbf{x}_u^i \circ (\psi^i)^{-1} \circ \psi^i + (1-\gamma) \cdot \mathbf{x}_u^i, \mathbf{y}_p^i = \mathbf{y}_l \circ \psi^i, \tag{2}$$

where F_l^i and F_u^i are features of the labeled image and unlabeled data generated from the fixed share encoder f_{se} and γ is a random number. Finally, the generated pseudo-labeled pair $\left\{\mathbf{x}_p^i, \mathbf{y}_p^i\right\}$ is used for constraint segmentation. Such a simple design not only can alleviate the appearance gap between labeled data and unlabeled data, but also generate diversity distribution data for subsequent segmentation.

3.2. Bi-level learning

Existing JRS-based methods treat registration and segmentation as two independent optimization tasks, alternatively updating parameters for one task with those for the other frozen. However, we find that registration and segmentation are two tightly coupled learning tasks by scrupulously reviewing a large number of experimental results (see Fig. 3 and the appendix). We further find *Stackelberg game theory* [29], which is a strategic model in economics where participants make sequential decisions, with one player acting as a leader who anticipates and influences the subsequent actions of the follower.

Motivated by the above exploration, we provide a bi-level formulation to explicitly characterize the coupling dependency between two task-specific decoders for registration and segmentation, which can be formulated as

$$\begin{aligned} &\min_{w} \Phi_r \left[(w, \Omega, \theta^*); \{\mathbf{x}_l, \mathbf{x}_u^i\}\right], s.t., \theta^* \in \mathcal{C}_s(\omega), \\ &\mathcal{C}_s(w) := \arg\min_{\theta} \Phi_s \left[(\theta, \Omega, w); \left\{\mathbf{x}_p^i, \mathbf{x}_u^i, \mathbf{y}_p^i\right\}\right], i \in M. \end{aligned} \tag{3}$$

where w and θ denote the parameters of f_{rd} and f_{sd}, $\circ$ denotes the warp operation, Φ_r and Φ_s represent the energies of the leader (upper) and follower (lower) levels of registration decoder and segmentation decoder, respectively. The leader task aims to optimize the parameters of f_{rd}, with respect to w, where θ^* is the best response constraint drawn from the constraint set $\mathcal{C}_s$ representing the solution set of the follower-level segmentation decoder problem. The follower-level sub-problem becomes optimizing Φ_s with respect to θ given pseudo-labeled pairs $\left\{\mathbf{x}_p^i, \mathbf{y}_p^i\right\}$.

Equation 3 explicitly formulates such a coupling relationship between two task-specific decoders that the optimization of θ in the follower-level constraints that of w in the leader level through the set $\mathcal{C}_s$ shown as vertical dashes.

3.3. Gradient Response Algorithm

Training the two task-specific decoders, f_{rd} and f_{sd}, turns out to resolve the bi-level optimization tasks *w.r.t.* w and θ in Eq. 3. We develop a numerical solution to the complicated optimization problem and start from the leader objective, computing its gradient *w.r.t.* w

$$\nabla_w \Phi_r\big(w, \theta^*(w)\big) = \nabla_w \Phi_r\big(w, \cdot\big) + \nabla_\theta \Phi_r\big(\cdot, \theta^*(w)\big). \tag{4}$$

The first term is a direct gradient in terms of w and the second term depicts the latent coupled connection with the follow-up segmentation decoder which is a challenging problem. Motivated by the Gaussian-Newton approximation that provides a first-order computation to address continuous learning [18], we propose a gradient response (GR) Algorithm, since the best response θ^* couples θ with w, which can be written as

$$\nabla_\theta \Phi_r = \nabla_\theta \Phi_r\big(w, \theta^*(w)\big) \nabla_w \theta^*(w). \tag{5}$$

The cross gradient term $\nabla_w \theta^*(w)$ inevitably evokes optimizing Φ_s in the follower level of Eq. 3. Previous studies in the context of bi-level optimization employ an explicit scheme to approximate the gradient by recurrent products of second-order Hessian matrices that are computationally expensive [14, 26].

We resort to the implicit function theorem targeting the optimal solution to the follower-level task. Letting $\partial \Phi_s / \partial w = 0$, we obtain

$$\nabla_w \theta^*(w) = -\left[\nabla_{\theta\theta}^2 \Phi_s\left(w, \theta^*(w)\right)\right]^{-1} \cdot \mathcal{W}_{Hes}, \tag{6}$$

where ∇^2 denotes the second-order partial derivatives and the Hessian $\mathcal{W}_{Hes} = \nabla_{\theta,w}^2 \Phi_s\big(w, \theta^*(w)\big)$. Hence, this gradient computation demands an inversion of second-order

Algorithm 1 Bi-level optimization learning

```
Require: The features of labeled and unlabeled data: $F_l$, $F_u$. Two parameterized registration and segmentation decoders for $w$ and $\theta$. Initialize $w$, $\theta$, and necessary hyper-parameters($\alpha$ and $\beta$ : learning rate)
repeat
    % Perform transformation
    $\psi^i = f_{rd}(w; F_l^i, F_u^i)$, $(\psi^i)^{-1} = f_{rd}(w; F_u^i, F_l^i)$
    % Generate a random number $\gamma$
    $\mathbf{x}_p^i = \gamma \cdot \mathbf{x}_u^i \circ (\psi^i)^{-1} \circ \psi^i + (1-\gamma) \cdot \mathbf{x}_u^i$
    $\mathbf{y}_p^i = \mathbf{y}_l \circ \psi^i$
    % Supervised learning segmentation
    Update $\theta$ to obtain approximation $\hat{\theta}$.
    $\hat{\theta}(\theta) := \theta - \beta \nabla_\theta \Phi_s(\theta, w)$
    Calculate $\nabla_w \Phi_s$, $\nabla_{\hat{\theta}} \Phi_s(\theta, w)$, and $\nabla_{\hat{\theta}} \Phi_r$
    Calculate $\nabla_w \Phi_r(w)$ by Eq. 4 with $\hat{\theta}$ and $w$
    $w := w - \alpha \nabla_w \Phi_r(w)$
until training convergence
return $(w^*, \theta^*)$ (Optimal solution)
```

derivatives and a Hessian matrix. Leveraging the outer product approximation in the Gauss-Newton method, we can further simplify calculating GR as products of first-order derivatives. Based on the above formula, we have derived an optimization algorithm in Alg. 1.

3.4. Loss function

In this part, we will elaborate on the concrete loss function to define Φ_r and Φ_s.

Leader-level loss for registration. We first adopt a smoothness loss function $\mathcal{L}_{smo}$ to constrain the deformation field, ensuring its smoothness

$$\mathcal{L}_{smo}(\psi) = \sum_{p \in \phi} \|\nabla \psi(p)\|^2, \tag{7}$$

where $\nabla_\psi(p)$ represents the gradient of the deformation field ψ at point p. Subsequently, we employ normalized cross-correlations as the similarity loss function $\mathcal{L}_{sim}$ to constrain the similarity of post-image registration, ensuring the precision and reliability of the registration outcomes. Thus, our complete registration loss $\mathcal{L}_{reg}$ is

$$\begin{aligned}\mathcal{L}_{reg} &= \lambda_1 \big(\mathcal{L}_{sim}(\mathbf{x}_l, \mathbf{x}_u \circ \psi^{-1}) + \mathcal{L}_{sim}(\mathbf{x}_u, \mathbf{x}_l \circ \psi)\big) \\ &+ \lambda_2 \big(\mathcal{L}_{sim}(\mathbf{x}_l, \mathbf{x}_l \circ \psi \circ \psi^{-1}) + \mathcal{L}_{sim}(\mathbf{x}_u, \mathbf{x}_u \circ \psi^{-1} \circ \psi)\big) \\ &+ \lambda_3 \big(\mathcal{L}_{smo}(\psi) + \mathcal{L}_{smo}(\psi^{-1})\big),\end{aligned} \tag{8}$$

where λ_1, λ_2 and λ_3 are the weights of the losses in $\mathcal{L}_{reg}$.

Considering the challenge in fully revealing the details of anatomical structures with registration methods based on image intensity, especially in blurred or structurally similar areas, we further utilize segmentation maps generated by the segmentation decoder f_{sd} to explicitly delineate structural boundaries and regional information and then Φ_r is

$$\Phi_r = \mathcal{L}_{reg} + \lambda_4 \mathcal{L}_{dice}(\hat{\mathbf{y}}^i, \mathbf{y}_p^i), \tag{9}$$

where $\hat{\mathbf{y}}^i$ represents the predicted segmentation result and λ_4 signifies the weight of the corresponding loss term.

Follower-level loss for segmentation. With the proposed ACC strategy, we can not only perform data augmentation but also maintain its appearance conformity by indirectly using labeled data. The segmentation decoder is trained with the generating pseudo-labeled pairs $\{\mathbf{x}_p^i, \mathbf{y}_p^i\}$ in Eq. 2 , and generates the predicted mask $\hat{\mathbf{y}}^i$. Φ_s is

$$\Phi_s = \mathcal{L}_{dice}(\hat{\mathbf{y}}^i, \mathbf{y}_p^i). \tag{10}$$

4. Experiments

Our proposed method is validated on brain MR image registration and segmentation tasks. A series of comparative experiments are detailed in Sec. 4.2 to reveal the exceptional performance of our method. The ablation experiments in Sec. 4.3 demonstrate the efficacy of our ACC strategy, bi-level modeling, and gradient response and validate the stability of our proposed Bi-JROS.

4.1. Experiments configurations

Data preparation: The proposed method and comparison methods are evaluated on mixed brain MRI datasets (ABIDE [8], ANDI [32] and PPMI [28]) and OASIS [30]. The publicly available atlas from [3] is the single labeled template image in training.

Mixed dataset: We divide our data into 295, 114 volumes for training and testing. All test volumes are anatomically segmented with FreeSurfer, resulting in 13 anatomical structures.

OASIS: To assess our method's capability for rapid adaptation, we conducted experiments on an entirely new dataset, and we split the data into 86, 25 for training and testing. The dataset has the same segmentation labels as the mixed dataset mentioned above.

Standard preprocessing steps including motion correction, NU intensity correction, normalization and affine normalization are done with FreeSurfer and FSL [40]. All scans are cropped and resized to 128 × 128 × 128 with 1 mm isotropic resolution.

Implementation details and metrics: The training process is divided into two phases: a) the pre-training phase, which draws on the optimization method of [11] as baseline, i.e., pre-training the network by combining multiple loss functions, and b) the bi-level optimization phase, we take the encoder weights derived from pre-training on a mixed dataset with different styles as initial values and freeze them to optimize the decoder for a specific task from scratch.

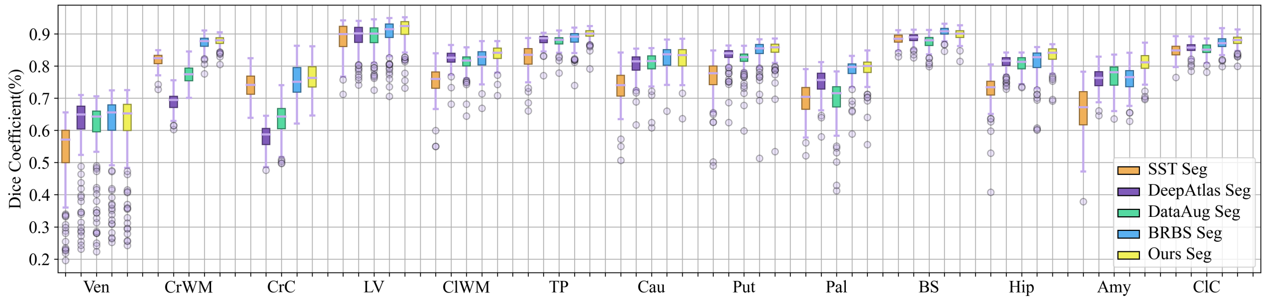


Figure 3. Boxplots of performance towards various segmentation methods of Dice scores with 13 categories of brain anatomical structures.

Our framework is implemented using PyTorch on a NVIDIA A40 with 48 GB of RAM. During training, we set the trade-off factor $\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 1$ and $\lambda_4 = 1$. We evaluate the performance using two widely used metrics, including Dice and NCC. Higher values of the Dice coefficient implies an increase in region overlap, which is indicative of superior alignment or segmentation. Correspondingly, a higher Normalized Correlation Count (NCC) indicates a higher degree of similarity between the deformation image and the target image, indicative of superior alignment performance.

4.2. Comparison Experiments

Comparison settings: To evaluate the excellence of the proposed Bi-JROS, we conducted comparative analyses with 14 prevalent medical image registration and segmentation algorithms under a one-shot setting, including a) traditional registration methods: SyN [2], NiftyReg [31], and deedsBCV [19]; b) deep learning-based registration approaches such as VoxelMorph [3], LKU-Net [20], and TransMorph [4]; c) deep learning-driven segmentation algorithms like U-Net [7], MASSL [5], and CPS [6]; and d) state-of-the-art (SOTA) JRS methodologies, namely DeepAtlas [41], SST [35], DataAug [42], UReSNet [11], and BRBS [18].

Comparison results: Tab. 1 reports the performance of our Bi-JROS with other methods in medical image registration and segmentation tasks. We can conclude that *firstly*, most JRS methods outperform standalone registration or segmentation approaches, underscoring the significance of implementing combined registration and segmentation. *Secondly*, the employment of a bi-level optimization framework facilitates the effective establishment of task-specific decoder interdependencies. Consequently, our framework achieved the highest segmentation Dice coefficient at 82.8%, the highest registration Dice coefficient at 80.8%, and the highest registration NCC value at 0.387, surpassing other JRS methods.

Fig. 3 illustrates the segmentation performance of the JRS method across various anatomical structures. For the segmentation task, our approach exhibited the best segmentation performance in 10 out of 13 structures, matched the performance of BRBS in 2 structures, and secured the second rank in one structure. The registration results are largely consistent with the segmentation results and are provided in the supplementary material.

Methods	Segmentation	Registration	
	Dice(%) ↑	Dice(%) ↑	Ncc ↑
Initial	-	63.8 ± 5.4	0.133 ± 0.009
SyN[2]	-	65.1 ± 6.0	0.232 ± 0.007
NiftyReg[31]	-	73.6 ± 3.0	0.238 ± 0.008
deedsBCV[19]	-	75.0 ± 2.7	0.238 ± 0.008
VoxelMorph[3]	-	76.5 ± 2.2	0.328 ± 0.006
LKU-Net[20]	-	76.6 ± 2.5	0.295 ± 0.008
TransMorph[4]	-	77.4 ± 2.1	0.341 ± 0.005
UNet[7]	45.0 ± 12.2	-	-
MASSL[5]	63.4 ± 6.8	-	-
CPS[6]	75.1 ± 4.2	-	-
SST[35]	75.8 ± 2.8	75.4 ± 5.4	0.364 ± 0.005
DeepAtlas[41]	78.1 ± 1.9	77.5 ± 2.8	0.293 ± 0.007
DataAug[42]	78.4 ± 2.5	77.7 ± 2.5	0.358 ± 0.008
UReSNet[11]	81.2 ± 2.1	79.1 ± 2.0	0.353 ± 0.008
BRBS[18]	81.8 ± 2.4	80.0 ± 2.4	0.324 ± 0.008
Ours	**82.8** ± **2.2**	**80.8** ± **2.0**	**0.387** ± **0.007**

Table 1. Quantitative comparison among various methods for registration and segmentation tasks on the mixed dataset. The top-ranked method is highlighted in **bolded** form.

We further show the segmentation results and registration results of six representative methods in Figs. 4 and 5 from three perspectives. In the segmentation task, our method achieves the segmentation effect with the highest overlap with GT on both the large structure Cerebral White Matter and the small tissue Lateral Ventricle. Due to the lack of supervisor information, CPS shows significant errors in the segmentation task. Most JRS methods enhance image diversity by employing style transformation or appearance transformation, and the ensuing image intensity transformation may lead to mismatches between deformed images and deformed labels, which in turn suppresses segmentation performance. Our proposed loss indirectly uses

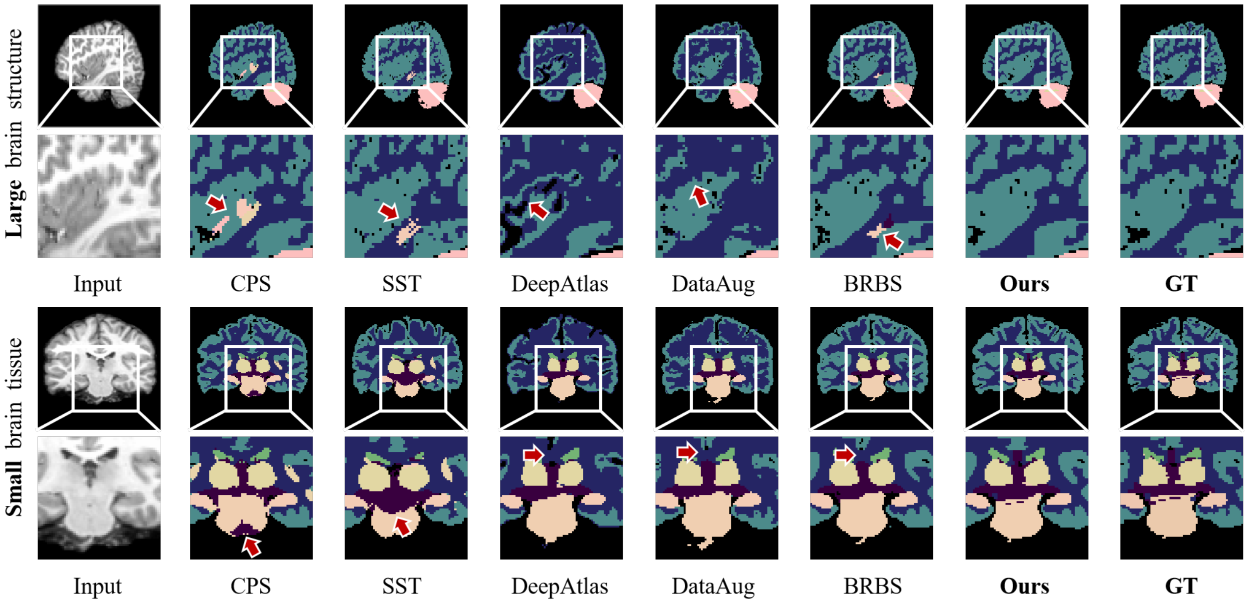


Figure 4. Segmentation visualization results of different methods on large brain structure Cerebral White Matte (CrWM) and small brain tissue 3rd/4th Ventricle (Ven). The red arrows point to the segmentation errors.

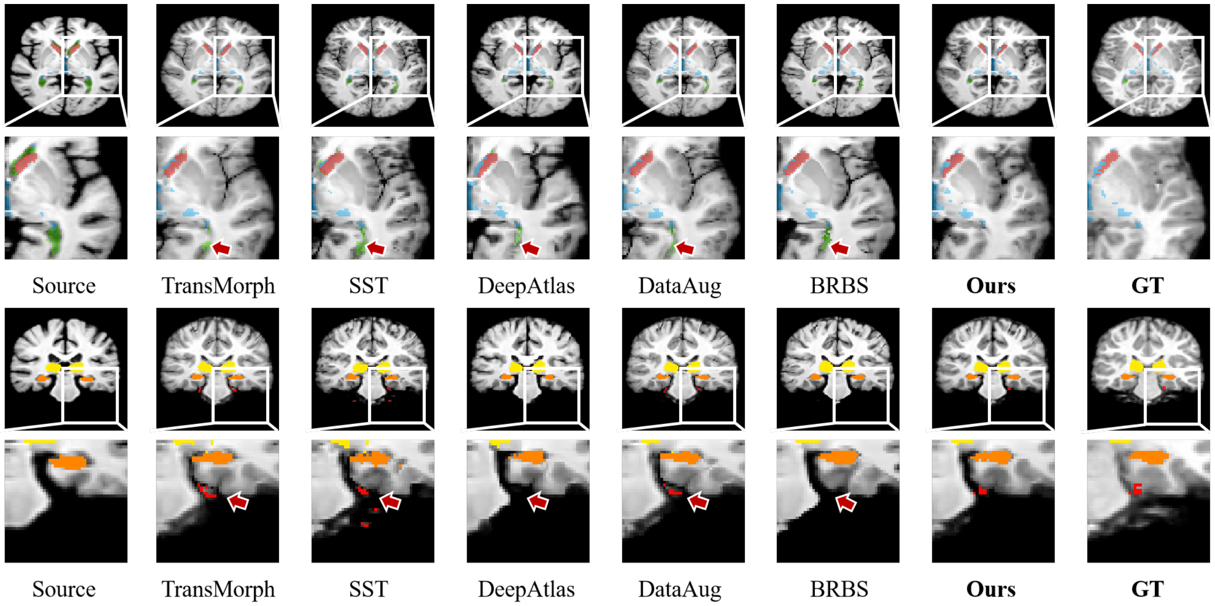
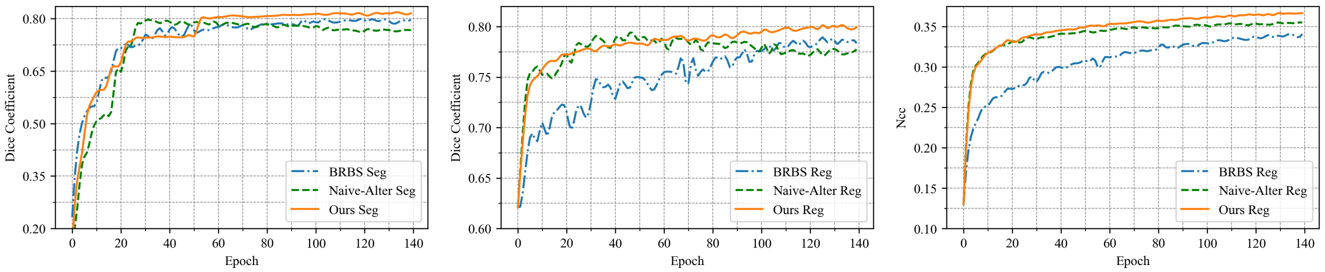


Figure 5. Visual results of performance comparison towards various registration methods of brain anatomical structures. The red arrows point to the segmentation errors.

atlas, which skillfully avoids the problem of mismatch between deformation images and deformation labels by applying two deformation transformations to the target image. In the registration task, we achieve the best alignment on the 3rd/4th Ventricle, Lateral Ventricle and Thalamus Proper compared to other methods.TransMorph, SST and DataAug do not introduce segmentation map information during the training alignment process, resulting in obvious misalignment. DeepAtlas and BRBS significantly improved the alignment error thanks to the introduction of segmentation maps as auxiliary information for alignment learning, which we likewise took into account. In addition, we consider the dynamic response of segmentation to alignment in each iteration, further improving the accuracy of alignment.



(a) Segmentation results in terms of dice scores.

(b) Registration results in terms of dice scores.

(c) Registration results in terms of ncc scores.

Figure 6. Illustrating the change in performance with the iterative process.

Baseline	FE Alter	ACC Alter	GR	S-Dice(%)	R-Dice(%)
✓				81.2 ± 2.1	79.1 ± 2.0
✓	✓			81.5 ± 1.9	79.8 ± 2.0
✓	✓	✓		82.1 ± 2.1	80.2 ± 2.1
✓	✓	✓	✓	82.8 ± 2.2	80.8 ± 2.0

Table 2. The ablation study demonstrates the contribution of our innovations.

4.3. Ablation Experiments

Effectiveness of our innovations. Tab. 2 precisely displays the results of the ablation experiments, effectively demonstrating the efficacy of each component in our study. In the first row of the table, we present the baseline results obtained by training the encoder and two decoders using a joint training approach. By comparing different experimental setups, we can observe three significant findings: *i)* adopting a strategy of freezing the encoder and starting from scratch to alternately train the two decoders, we achieved a 0.7% improvement in registration performance; *ii)* when introducing the ACC strategy for segmentation tasks under the same experimental settings, we observed a 0.6% increase in segmentation performance; and *iii)* by implementing a bi-level optimization strategy to model the interaction between decoders and applying Gradient Response techniques, we further enhanced the registration performance by 0.6% and the segmentation performance by 0.7%. Compared to the baseline, our method ultimately achieved performance improvements of 1.7% and 1.6% in two tasks.

Stability and rapid adaptability of our Bi-JROS. To validate the stability and rapid adaptability of our bi-level optimization, we conducted a series of experiments on a new dataset. In these experiments, we kept the encoder learned from the mixed dataset in a frozen state and fine-tuned the decoder. The experimental setup consisted of three different comparison methods: *1)* the BRBS method retrained on OASIS, *2)* naive alternative training (with the same frozen encoder parameters as ours, but without the bi-level optimization and GR), and *3)*: our Bi-JROS model.

Methods	Seg	Reg		Time
	Dice(%)	Dice(%)	Ncc	(hours)
Initial	-	62.1 ± 2.7	0.130 ± 0.007	-
Naive Alter	80.1 ± 1.0	79.4 ± 0.6	0.356 ± 0.004	**6**
BRBS	80.3 ± 1.1	79.5 ± 1.0	0.342 ± 0.006	9.6
Ours	**81.4 ± 0.6**	**80.0 ± 0.6**	**0.362 ± 0.004**	7.8

Table 3. Results among BRBS, naive alternative training and our Bi-JROS for registration (Reg) and segmentation (Seg) tasks.

As can be seen in Tab. 3, our Bi-JROS achieves the highest Dice scores and Ncc values on both tasks and our adaptation to new datasets is markedly swifter than BRBS. Fig. 6 further depicts the performance variation of Bi-JROS, BRBS, and naive alternating training on both tasks during the iteration process. From these graphs, we can intuitively observe that our model exhibits higher stability compared to simple alternative training, which shows a trend of performance degradation in the later part of the iteration.

5. Conclusion

In this paper, we present a bi-level optimization formulation for registration-specific and segmentation-specific decoders. By integrating the ACC strategy into the segmentation task, we effectively mitigate the risk of overfitting to homogeneous data styles, thereby improving the model's rapid adaptability. We introduce the Response Gradient (RG) to replace the naive alternating learning approach, ensuring efficient and stable training. Our extensive experiments, which yield state-of-the-art results on various datasets, demonstrate our excellent performance in both registration and segmentation tasks.

Acknowledgements This work was supported in part by the National Key R&D Program of China (2020YFB1313503), the National Natural Science Foundation of China (Nos. 61922019), LiaoNing Revitalization Talents Program (XLYC1807088), and the Fundamental Research Funds for the Central Universities.

References

[1] Julia Andresen, Timo Kepp, Jan Ehrhardt, Claus von der Burchard, Johann Roider, and Heinz Handels. Deep learning-based simultaneous registration and unsupervised non-correspondence segmentation of medical images with pathologies. *Int. J. Comput. Assist. Radiol. Surg.*, 17(4):699–710, 2022. 1, 2, 3

[2] Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 2011. 6

[3] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE Trans. Medical Imaging*, 38(8):1788–1800, 2019. 1, 5, 6

[4] Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Anal.*, 82: 102615, 2022. 6

[5] Shuai Chen, Gerda Bortsova, Antonio García-Uceda Juárez, Gijs van Tulder, and Marleen de Bruijne. Multi-task attention-based semi-supervised learning for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 457–465, 2019. 6

[6] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 6

[7] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 424–432, 2016. 6

[8] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014. 5

[9] Yuhang Ding, Xin Yu, and Yi Yang. Modeling the probabilistic distribution of unlabeled data for one-shot medical image segmentation'. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1246–1254, 2021. 3

[10] Mohamed S Elmahdy, Laurens Beljaards, Sahar Yousefi, Hessam Sokooti, Fons Verbeek, Uulke A Van Der Heide, and Marius Staring. Joint registration and segmentation via multi-task learning for adaptive radiotherapy of prostate cancer. *IEEE Access*, 9:95551–95568, 2021. 1

[11] Théo Estienne, Maria Vakalopoulou, Stergios Christodoulidis, Enzo Battistela, Marvin Lerousseau, Alexandre Carre, Guillaume Klausner, Roger Sun, Charlotte Robert, Stavroula Mougiakakou, et al. U-resnet: Ultimate coupling of registration and segmentation with deep nets. In *Medical Image Computing and Computer Assisted Intervention*, pages 310–319, 2019. 1, 2, 5, 6

[12] Théo Estienne, Marvin Lerousseau, Maria Vakalopoulou, Emilie Alvarez Andres, Enzo Battistella, Alexandre Carré, Siddhartha Chandra, Stergios Christodoulidis, Mihir Sahasrabudhe, Roger Sun, Charlotte Robert, Hugues Talbot, Nikos Paragios, and Eric Deutsch. Deep learning-based concurrent brain registration and tumor segmentation. *Frontiers Comput. Neurosci.*, 14:17, 2020. 1

[13] Xin Fan, Zi Li, Ziyang Li, Xiaolin Wang, Risheng Liu, Zhongxuan Luo, and Hao Huang. Automated learning for deformable medical image registration by jointly optimizing network architectures and objective functions. *IEEE Trans. Image Process.*, 2023. 1

[14] Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1165–1173, 2017. 4

[15] Yuting He, Tiantian Li, Guanyu Yang, Youyong Kong, Yang Chen, Huazhong Shu, Jean-Louis Coatrieux, Jean-Louis Dillenseger, and Shuo Li. Deep complementary joint model for complex scene registration and few-shot segmentation on medical images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 770–786, 2020. 2

[16] Yuting He, Tiantian Li, Rongjun Ge, Jian Yang, Youyong Kong, Jian Zhu, Huazhong Shu, Guanyu Yang, and Shuo Li. Few-shot learning for deformable medical image registration with perception-correspondence decoupling and reverse teaching. *IEEE J. Biomed. Health Informatics*, 26(3): 1177–1187, 2021. 2

[17] Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9538–9547, 2023. 3

[18] Yuting He, Rongjun Ge, Xiaoming Qi, Yang Chen, Jiasong Wu, Jean-Louis Coatrieux, Guanyu Yang, and Shuo Li. Learning better registration to learn better few-shot medical image segmentation: Authenticity, diversity, and robustness. *IEEE Trans. Neural Networks Learn. Syst.*, 35(2): 2588–2601, 2024. 1, 2, 3, 4, 6

[19] Mattias P Heinrich, Mark Jenkinson, Michael Brady, and Julia A Schnabel. Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE Trans. Medical Imaging*, 32(7):1239–1248, 2013. 6

[20] Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. U-net vs transformer: Is u-net outdated in medical image registration? In *Machine Learning in Medical Imaging*, pages 151–160. Springer, 2022. 6

[21] Dian Jin, Long Ma, Risheng Liu, and Xin Fan. Bridging the gap between low-light scenes: Bilevel learning for fast adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2401–2409, 2021. 3

[22] Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. Cyclemorph: Cycle consistent unsupervised deformable image registration. *Medical Image Anal.*, 71:102036, 2021. 4

[23] Ziyang Li, Zi Li, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coupling deep deformable registration with contextual re-

finement for semi-supervised medical image segmentation. In *19th IEEE International Symposium on Biomedical Imaging, ISBI*, pages 1–5. IEEE, 2022. 1

[24] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12): 10045–10067, 2021. 3

[25] Risheng Liu, Zi Li, Xin Fan, Chenying Zhao, Hao Huang, and Zhongxuan Luo. Learning deformable image registration from optimization: perspective, modules, bilevel training and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44 (11):7688–7704, 2021. 1

[26] Risheng Liu, Jiaxin Gao, Xuan Liu, and Xin Fan. Revisiting gans by best-response constraint: Perspective, methodology, and application. *arXiv preprint arXiv:2205.10146*, 2022. 3, 4

[27] Risheng Liu, Jiaxin Gao, Xuan Liu, and Xin Fan. Learning with constraint learning: New perspective, solution strategy and various applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2024. 3

[28] Matthew MacKay, Paul Vicol, Jon Lorraine, David Duvenaud, and Roger Grosse. Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions. In *Proceedings of the 7th International Conference on Learning Representations*, 2019. 3, 5

[29] Sabita Maharjan, Quanyan Zhu, Yan Zhang, Stein Gjessing, and Tamer Basar. Dependable demand response management in the smart grid: A stackelberg game approach. *IEEE Trans. Smart Grid*, 4(1):120–132, 2013. 4

[30] Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *J. Cogn. Neurosci.*, 22(12):2677–2684, 2010. 5

[31] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Comput. Methods Programs Biomed.*, 98(3):278–284, 2010. 6

[32] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005. 5

[33] Annemie Ribbens, Jeroen Hermans, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Sparc: unified framework for automatic segmentation, probabilistic atlas construction, registration and clustering of brain mr images. In *Proceedings of the 2010 International Symposium on Biomedical Imaging*, pages 856–859, 2010. 1

[34] Matthew Sinclair, Andreas Schuh, Karl Hahn, Kersten Petersen, Ying Bai, James Batten, Michiel Schaap, and Ben Glocker. Atlas-istn: joint segmentation, registration and atlas construction with image-and-spatial transformer networks. *Medical Image Anal.*, 78:102383, 2022. 1

[35] Devavrat Tomar, Behzad Bozorgtabar, Manana Lortkipanidze, Guillaume Vray, Mohammad Saeed Rad, and Jean-Philippe Thiran. Self-supervised generative style transfer for one-shot medical image segmentation. In *Winter Conference on Applications of Computer Vision*, pages 1737–1747, 2022. 6

[36] Devavrat Tomar, Behzad Bozorgtabar, Manana Lortkipanidze, Guillaume Vray, Mohammad Saeed Rad, and Jean-Philippe Thiran. Self-supervised generative style transfer for one-shot medical image segmentation. In *Winter Conference on Applications of Computer Vision*, pages 1998–2008, 2022. 1

[37] Liesbeth Vandewinckele, Siri Willems, David Robben, Julie Van Der Veen, Wouter Crijns, Sandra Nuyts, and Frederik Maes. Segmentation of head-and-neck organs-at-risk in longitudinal CT scans combining deformable registrations and convolutional neural networks. *Comput. methods Biomech. Biomed. Eng. Imaging Vis.*, 8(5):519–528, 2020. 1

[38] Shuxin Wang, Shilei Cao, Dong Wei, Renzhen Wang, Kai Ma, Liansheng Wang, Deyu Meng, and Yefeng Zheng. Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 9162–9171, 2020. 2, 4

[39] Yiqian Wang, Junkang Zhang, Cheolhong An, Melina Cavichini, Mahima Jhingan, Manuel J Amador-Patarroyo, Christopher P Long, Dirk-Uwe G Bartsch, William R Freeman, and Truong Q Nguyen. A segmentation based robust deep learning framework for multimodal retinal image registration. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 1369–1373, 2020. 1

[40] Mark W Woolrich, Saad Jbabdi, Brian Patenaude, Michael Chappell, Salima Makni, Timothy Behrens, Christian Beckmann, Mark Jenkinson, and Stephen M Smith. Bayesian analysis of neuroimaging data in fsl. *NeuroImage*, 45(1): S173–S186, 2009. 5

[41] Zhenlin Xu and Marc Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 420–429, 2019. 1, 2, 6

[42] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 8543–8553, 2019. 1, 2, 6

[43] Fenqiang Zhao, Zhengwang Wu, Li Wang, Weili Lin, Shunren Xia, Gang Li, and UNC/UMN Baby Connectome Project Consortium. A deep network for joint registration and parcellation of cortical surfaces. In *Medical Image Computing and Computer Assisted Intervention*, pages 171–181, 2021. 1, 2, 3

[44] Bo Zhou, Zachary Augenfeld, Julius Chapiro, S Kevin Zhou, Chi Liu, and James S Duncan. Anatomy-guided multimodal registration by learning segmentation without ground truth: Application to intraprocedural cbct/mr liver segmentation and registration. *Medical Image Anal.*, 71:102041, 2021. 1