# NICE-Transeg: Non-iterative Coarse-to-fine Transformer Network For Joint Affine and Deformable 2D Image Registration and Segmentation

JiaXin Xie*
HKUST
Clear Water Bay, Hong Kong
jxiebn@connect.ust.hk

Yau-Meng Wong*
Columbia University
New York, New York
yw3809@columbia.edu

## Abstract

*Extensive research has focused on developing end-to-end deep registration methods that can yield accurate results at low computation cost. Recently, Meng et al. proposed the Non-Iterative, Coarse-to-finE Transformer (NICE-Trans) network. This novel architecture increases runtime efficiency while improving registration accuracy, achieving cost-effective performance that matches or surpasses state-of-the-art methods. Along another direction, segmentation-assisted registration systems have also shown to enhance model outcomes because it provides indirect supervision for learning registration tasks. Notably, ProGressively Coupling Network (PGCNet) for joint registration and segmentation is recently proposed, which adds feature interaction between segmentation and registration branches to enable a highly coupled joint network.*

*This paper sets out to investigate whether integrating a transformer-based segmentation branch would further benefit registration performance. Specifically, we propose the NICE-Transeg network for joint 2D image registration and segmentation. The model sets out to progressively integrate the segmentation task into the NICE-Trans registration process to further refine model performance in a few-shot scenario. We test on two public MRI datasets, OASIS and IXI. While NICE-Transeg doesn't outperform NICE-Trans in 2D image registration, it demonstrated strong performance in segmentation outcome. Code is available on Github.*

## 1. Introduction

Image registration is the procedure of spatially aligning a moving image to a fixed image. This procedure has significant implications in medical image analysis. By resolving misalignment between images taken from different subjects, times, or mediums, image registration facilitates comparison between complex medical images, thereby helping monitor disease progression, detect abnormalities, provide image guided interventions in surgical settings, etc.

Traditional methods of image registration involves carefully manipulating pre-designed feature detection and transformation algorithms in a case by case basis. These type of models are not generalizable across populations and are computationally inefficient. [1] [11]. With the rise of deep learning, deep registration methods have become widely studied to deliver more cost-efficient registration results with minimal supervision [2] [5]. Recent developments introduced multi-step cascading networks, in which deformation fields are generated in a coarse-to-fine manner, yielding results that match or even surpass traditional methods at a significantly lower cost [12] [15]. However, such networks remains sub-optimal due to their reliance on repetitive feature extractions steps. One recent study proposes the NICE-Trans model, a novel non-iterative coarse-to-fine transformer-based model that performs joint affine and deformable registration [10]. While prior deep registration methods must repeatedly perform feature extraction to iteratively improve the deformation field, the NICE-Trans model adapts an auto-encoder architecture to avoid such repeated calculations, thereby reducing computation loads while achieving the same outcome. The model is also the first to join affine and deformable registration in a single network, providing a much simpler architecture compared to previous deep registration models, where affine registration was often performed separately through traditional registration methods or extra networks. In addition, the model incorporates transformer blocks in the decoder, which are shown to yield substantially better registration results due to their large receptive field capturing more precise spatial correspondence [3]. These features allowed NICE-Trans to outperform state-of-the-art methods while further improving computational efficiency.

However, like most other registration approaches, NICE-Trans is trained in an unsupervised fashion. Because the ground truth for registration tasks are hard to obtain, there is an interest in exploring ways of introducing some su-

*These authors contributed equally to this work pervision during training to help the model more quickly identify regions of interest, preserve anatomical structures, and obtain more realistic result. In recent years, studies have suggested that segmentation labels can provide valuable supervision for registration. Labels identify specific areas of interests for the registration network, effectively bolstering registration accuracy while preventing unrealistic image distortion [13]. Several studies have demonstrated the benefits of coupling supervised segmentation with registration [6] [14] [17].

Under this context, a natural question arises on how the NICE-Trans model may fit into the current work in segmentation-assisted registration networks. In this paper, we set out to answer whether incorporating a segmentation component into the NICE-Trans network would further improve its registration performance, and how it would effect the computational cost of inference. Specifically, we add a segmentation network component to the current NICE-Trans model by referencing the recently proposed ProGressively Coupling Network (PGCNet) architecture [16]. Previous segmentation-assisted registration networks utilized only joint supervision without any additional information sharing between the registration and segmentation branches. PGCNet is set apart from previous networks because it infuses the warped image into the segmentation branch at each registrations step, thereby increasing the influence of segmentation in regularizing the correct registration projection.

Due to limited access to GPUs and low memory availability, NICE-Transeg is a 2D segmentation-assisted registration network rather than 3D like NICE-Trans and PGCNet. We investigated the performance of NICE-Transeg using two public brain MRI dataset, OASIS and IXI. The contribution of our study is three fold. First, we proposed NICE-Transeg, which is a joint 2D image segmentation and registration network that combined ideas from both NICE-Trans and PGCNet. Secondly, our result suggests that while NICE-Trans did not benefit significantly from the addition of the segmentation branch, the segmenetation outcome of such a joint network is good. Thirdly, we made available a preprocessed, 2D version of the IXI dataset for future researchers with limited computational resources.

## 2. Method

As a segmentation-assisted registration network, NICE-Transeg learn segmentation and registration concurrently. We parameterize the network as function $F_\theta(I_f, I_m) = \phi, S_f, S_m$. $\theta$ denotes the set of learnable parameters. The function takes in one fixed image $I_f$ and one moving image $I_m$, both of which are single-channel, gray-scale volumes in the 2D spatial domain. The output includes three elements. $\phi$ denotes the predicted spatial transformation field. $S_f$ denotes the segmentation map of the fixed image, and $S_m$ denotes the segmentation map of the moving image.

### 2.1. NICE-Transeg Model Architecture

NICE-Transeg directly adopts the autoencoder backbone proposed in NICE-Trans. The weight-sharing dual path encoder will separately extracts the features of one moving image $I_m$ and one fixed image $I_f$. The encoder consists of 5 convolutional block. In each block, the encoder applies one 2x2 max pooling and one 2D convolutional module, which consists of two 3x3 convolutional layers and a LeakyReLU activation function with a parameter of 0.2. The result of the encoder consists of five feature maps for moving $F_m \in \{F_{m,1}, F_{m,2}, F_{m,3}, F_{m,4}, F_{m,5}\}$ and five feature maps for fixed image $F_f \in \{F_{f,1}, F_{f,2}, F_{f,3}, F_{f,4}, F_{f,5}\}$.

We keep the same registration decoder used in NICE-Trans, which includes four Swin Transformer modules and one 3D convolutional module. Each registration decoder module is shown in Figure 2. At each step i, the feature maps $F_{m,i}$ is transformed using the previous predicted registration $\phi_{i-1}$, then concatenated with $F_{f,i}$. In particular, patch expanding layer will double feature resolution and halves the feature dimension between steps. The result then passes through the Swin transformer block that is made up of layer normalization, shifted window based multi-head self-attention, multi-layer perceptron, and residual connection. The output of each step consists of the the current displacement field $\phi_i$ for a five-step coarse-to-fine registration process. Because we are implementing the progressively coupling idea from PGCNet, we not only map $F_{m,i}$ to $F_{f,i}$ to obtain the regular registration result $phi$ , but we will also map $F_{f,i}$ to $F_{m,i}$ to obtain the inverse registration result $\phi^{-1}$ that will be used in the segmentation branches.

On top of NICE Trans, we add two segmentation decoder branches where the segmentation maps of moving and fixed images are predicted with shared weights. In the moving image segmentation branch, we propagate the registration result at each step to the corresponding segmentation step for the moving image. We concatenate the transformed extracted fixed feature $F_{f,i} \circ \phi_i$, the moving feature $F_{m,i}$, with the patch-expanded segmentation map $S_{m,i-1}$ from the previous step. The result will then pass through a Swin transformer block to obtain the next predicted segmentation map. The same procedure is used in the fixed image segmentation branch. Here, however, we utilize the inverse registration result for the fixed image and concatenate the transformed moving image feature $F_{m,i} \circ \phi_i^{-1}$, the fixed image feature $F_{f,i}$, with the patch-expanded segmentation map $S_{f,i-1}$.

The design of NICE-Transeg is intended to incorporate the best of both world in NICE-Trans and PGCNet. On the one hand, it reaps the benefit of model efficiency and long-range spatial relevance learning capability of the NICE-Trans backbone. On the other hand, we leverage the progressively coupling design proposed in the PGCNet

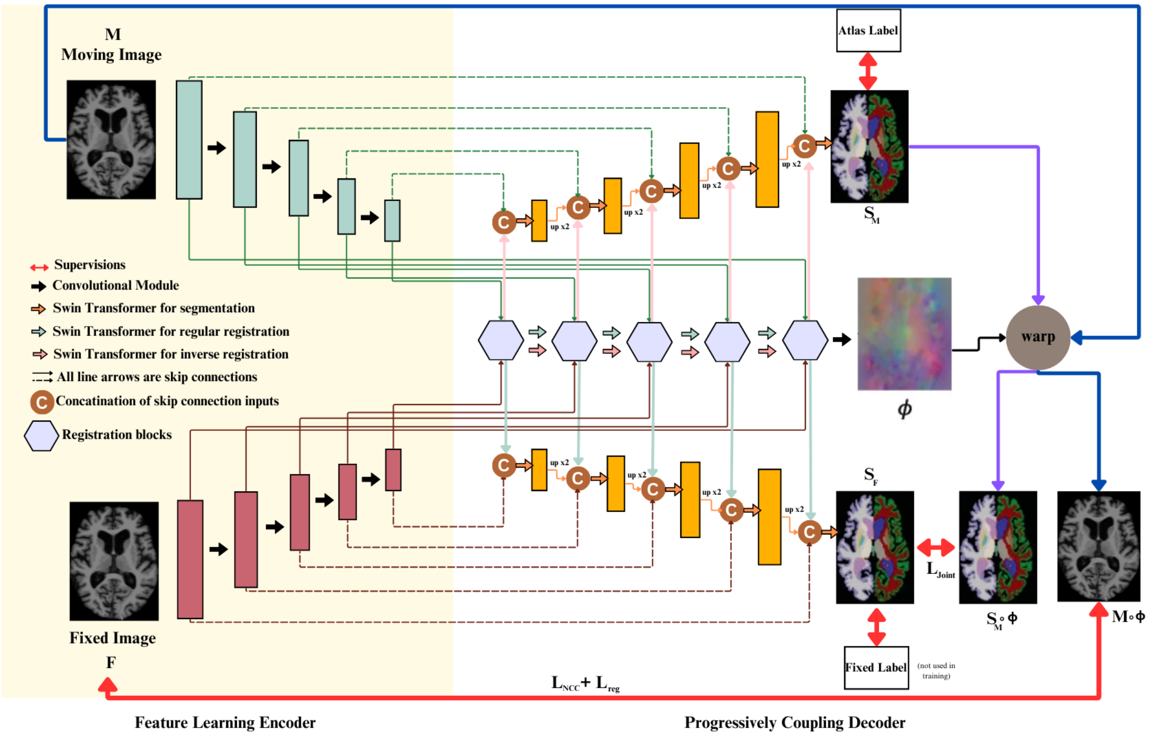
M
Moving Image
Atlas Label
S_M
Supervisions
Convolutional Module
Swin Transformer for segmentation
Swin Transformer for regular registration
Swin Transformer for inverse registration
All line arrows are skip connections
C Concatination of skip connection inputs
Registration blocks
up x2
warp
ϕ
S_F
L_Joint
S_M∘ϕ
M∘ϕ
Fixed Image
F
Fixed Label
(not used in training)
L_NCC + L_reg
Feature Learning Encoder
Progressively Coupling Decoder

Figure 1. The encoder-decoder pipeline for NICE-Transeg.

to amplify the indirect supervision provided by co-training segmentation and registration. Ultimately, the goal is for NICE-Transeg to efficiently yield more accurate, realistic registration result.

## 2.2. Loss Calculation

Concurring with NICE-Trans, the registration branch in NICE-Transeg is trained through unsupervised learning. At each pixel $p$ in the image $\Omega$, loss in the registration task is calculated by $L_r = L_{sim} + \sigma L_{reg}$. In particular, $L_{sim}$ is the negative local normalized cross-correlation(NCC) for penalizing any differences between the fixed image and the warped moving image. $L_{reg}$ is a diffusion regularize used on the displacement map to encourage the smoothness of the predicted transformation.

$$L_{reg} = \sum_{p \in \Omega} || \bigtriangledown \phi(p)||^2$$

In the segmentation task, we calculate the loss $L_s = L_{wce} + L_{dice}$, where $L_{wce}$ is weighted cross entropy and $L_{dice}$ is Dice loss as used in PGCNet.

$$L_{wce}(S, S^*) = -\frac{1}{NP} \sum_{n=0}^{N} \sum_{p=0}^{P} w_n S^*_{n,p} log S_{n,p}$$

$$L_{dice}(S, S^*) = -\frac{1}{N} \sum_{n=0}^{N} \frac{\sum_{p=0}^{P} S^*_{n,p} \cdot S_{n,p}}{\sum_{p=0}^{P} S^*_{n,p} + \sum_{p=0}^{P} S_{n,p}}$$

$N$ is the number of semantic classes, while n represents the channel that corresponds to each semantic class. $P$ indicates the total number of pixels in each channel. Instead of using the true segmentation labels of each data point, we use five atlas segmentation labels to simulate the semi-supervised setting when training the segmentation branch. $S$ represents the segmentation result while $S^*$ denotes the ground truth atlas segmentation map. Finally, a joint loss for registration and segmentation $L_{joint}$ is calculated by $L_{dice}(S_f, S_m \circ \phi)$.

# 3. Experimental Setup

## 3.1. Dataset and Preprocessing

Two public preprocessed datasets, OASIS [9] and IXI are used. Both are 2D versions, as we didn't have enough compute to train on the 3D datasets. First, we used the public preprocessed OASIS dataset from the HyperMorph paper [7]. A total number of 410 T1–weighted 2D brain MRI slices were split into 285, 42, and 83 (7:1:2) images for training, validation, and test sets. FreeSurfer was used to pre-process the MRI images. All images were cropped to size of 160 × 192. Label maps of 25 anatomical structures were used to evaluate segmentation and registration perfor-

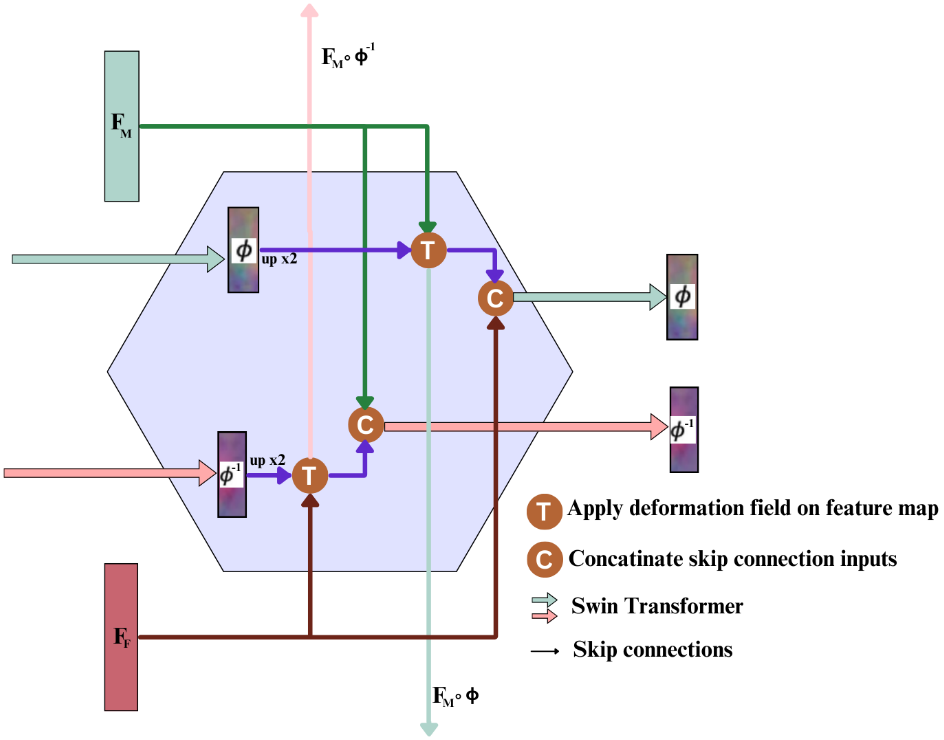

Figure 2. Registration module for NICE-Transeg

mances. For IXI we took the preprocessed version used in TransMorph [4] and converted it to 2D. This version of IXI and the code we used to convert it is available on Github. The original Transmorph IXI dataset has 256 semantic segmentation labels, but in the 2D version only 31 are used. The images are cropped to 160 × 224 and split into training, validation, and test sets in the same fashion as OASIS.

In training both OASIS and IXI, we randomly select 5 scans from the training set to act as atlas images. These are the moving images for registration, and their labels are the only labels seen in training. This few-shot training method for segmentation mimics the training procedure of PGCNet [16].

### 3.2. Implementation Details

We trained our model for 200 epochs on two Nvidia GeForce RTX-3080 GPUs. The model was trained with Adam, a learning rate of 1E-4, and a batch size of 1. For each training iteration, an image from the atlas set acts as the moving image and an image from the training set acts as the fixed image. During an epoch each image from the training set (minus the atlas) is seen exactly once. For validation, pairs of images are chosen from the validation set. Similarly for testing.

### 3.3. Comparison Methods

The underlying registration model is the same as NICE-Trans. The segmentation decoder is also similar to the registration decoder of NICE-Trans in its sequential Swin-transformers and convolutions. The coupling of the segmentation and registration branches is nearly identical to that in PGCNet, though here the decoders contain transformers in addition to the convolutional layers. We followed the same one encoder/two decoders structure as PGCNet, as opposed to the two encoders/two decoders structure of previous models, as PGCNet's positive results showed that the single encoder can be beneficial in more closely joining the registration and segmentation learning.

## 4. Results

As displayed in Table 1, NICE-Transeg did not outperform 2D NICE-Trans in 2D image registration on either 2D IXI or 2D OASIS. Specifically, NICE-Transeg achieves 0.757 DSC on 2D OASIS and 0.55 DSC on 2D IXI with an inference runtime of 54.393 ms and 54.6793 ms respectively. In contrast, NICE-Trans obtained 0.758 DSC on 2D OASIS and 0.551 DSC on 2D IXI but at a significantly lower inference runtime of around 16 ms. This suggests that the addition of the segmentation branch and the progressively coupling of the decoder networks did not contribute

Table 1. Segmentation and Inference Runtime Results

| Method | OASIS 2D | | IXI 2D | | Inference Runtime (ms) | |
|---|---|---|---|---|---|---|
| | DSC | Segmentation Accuracy | DSC | Segmentation Accuracy | OASIS 2D | IXI 2D |
| Before Registration | 0.533 | - | 0.434 | - | - | - |
| NICE-Trans-2D (Affine Only) | 0.558 | - | 0.448 | - | - | - |
| NICE-Trans-2D | 0.758 | - | 0.551 | - | 16.7977 | 16.7725 |
| FCN-Resnet | - | 0.699 | - | 0.59 | 20.3483 | 16.7725 |
| NICE-Transeg (Affine only) | 0.555 | - | 0.447 | - | - | - |
| NICE-Transeg (Ours) | 0.757 | 0.817 | 0.55 | 0.654 | 54.393 | 54.6793 |

any improvement to registration performance.

On the other hand, NICE-Transeg produced strong segmentation result. Compared with a common semantic segmentation benchmark FCN-Resnet [8], which was trained on the same 5 atlas labels to simulate a few-shot scenario, NICE-Transeg outperforms it by 11.8% and 6.4% on OASIS and IXI respectively. This suggests that the registration branch may be more beneficial for the segmentation branch than the other way around.

## 5. Discussion

It appears that the segmentation branch benefits more from the progressively coupling design than the registration branch. Since the registration branch learns to map the moving image to the fixed image, this may help the segmentation branch better capture the defining features of anatomical structures. On the other hand, it remains unclear why the segmentation branch did not have a strong impact on the registration outcome. There is also the question of whether this difference in performance generalizes to 3D MRIs as well. This could be a future research direction. It is possible that the segmentation branch did not contribute significantly because 2D image registration is less challenging to learn compared to 3D image registration. In 3D image registration, the supervision provided by the segmentation branch may offer more important guidance to the model to better learn the 3D structure. Another future direction could be to include positional embeddings into the registration branch, as this was shown to improve performance in PGCNet.

Although NICE-Transeg incurred additional runtime compared to FCN-Resnet, its segmentation accuracy was impressive enough to perhaps compare NICE-Transeg with more few-shot semantic segmentation benchmarks. Doing so would offer a more holistic insight into where NICE-Transeg stacks up to current SOTA methods, and whether the additional computation expense of the registration branch is worth it.

## 6. Conclusion

In this paper, we outlined the Non-Iterative Coarse-to-finE Transformer network for joint Segmentation and registration (NICE-Transeg). This network integrates the joint task design of PGCNet into the NICE-Trans backbone. NICE-Transeg did not show improved registration performance when compared with NICE-Trans. However, NICE-Transeg did achieve a much better few-shot semantic segmentation performance when compared to common benchmark FCN-Resnet. While our study sets out to answer the research question of whether the progressive coupling technique introduced in PGCNet would improve the performance of NICE-Trans, the result of our study opened up many new, larger questions to be answered for future studies. The most important and intriguing of all is perhaps what makes the joint network to benefit registration tasks and segmentation tasks at different degrees.

## References

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

[2] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019. 1

[3] Junyu Chen, Eric C. Frey, Yufan He, William P. Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, Nov. 2022. 1

[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 Analysis*, 2022. 4

[5] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019. 1

[6] Théo Estienne, Marvin Lerousseau, Maria Vakalopoulou, Emilie Alvarez Andres, Enzo Battistella, Alexandre Carré, Siddhartha Chandra, Stergios Christodoulidis, Mihir Sahasrabudhe, Roger Sun, et al. Deep learning-based concurrent brain registration and tumor segmentation. *Frontiers in computational neuroscience*, 14:17, 2020. 2

[7] Andrew Hoopes, Malte Hoffmann, Douglas N. Greve, Bruce Fischl, John Guttag, and Adrian V. Dalca. Learning the effect of registration hyperparameters with hypermorph. *Machine Learning for Biomedical Imaging*, 1(IPMI 2021):1–30, Apr. 2022. 3

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2014. 5

[9] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, Sept. 2007. 3

[10] Mingyuan Meng, Lei Bi, Michael Fulham, Dagan Feng, and Jinman Kim. *Non-iterative Coarse-to-Fine Transformer Networks for Joint Affine and Deformable Image Registration*, page 750–760. Springer Nature Switzerland, 2023. 1

[11] 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. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010. 1

[12] Tony CW Mok and Albert CS Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 211–221. Springer, 2020. 1

[13] Sarah Parisot, William Wells III, Stéphane Chemouny, Hugues Duffau, and Nikos Paragios. Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs. *Medical image analysis*, 18(4):647–659, 2014. 2

[14] Liang Qiu and Hongliang Ren. Rsegnet: a joint learning framework for deformable registration and segmentation. *IEEE Transactions on Automation Science and Engineering*, 19(3):2499–2513, 2021. 2

[15] Yucheng Shu, Hao Wang, Bin Xiao, Xiuli Bi, and Weisheng Li. Medical image registration based on uncoupled learning and accumulative enhancement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–13. Springer, 2021. 1

[16] Zuopeng Tan, Hengyu Zhang, Feng Tian, Lihe Zhang, Weibing Sun, and Huchuan Lu. Progressively coupling network for brain mri registration in few-shot situation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–633. Springer, 2023. 2, 4

[17] Zhenlin Xu and Marc Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 420–429. Springer, 2019. 2