Optron: Better Medical Image Registration via Training in the Loop

Yicheng Chen *
Tongji University
Shanghai, China
2053186@tongji.edu.cn

Shengxiang Ji *
Huazhong University of Science and Technology
Wuhan, China
u202015362@hust.edu.cn

Yuelin Xin *
University of Leeds
Leeds, UK
sc20yx2@leeds.ac.uk

Kun Han
University of California, Irvine
Irvine, CA, USA
khan7@uci.edu

Xiaohui Xie
University of California, Irvine
Irvine, CA, USA
xhx@ics.uci.edu

Abstract

Previously, in the field of medical image registration, there are primarily two paradigms, the traditional optimization-based methods, and the deep-learning-based methods. Each of these paradigms has its advantages, and in this work, we aim to take the best of both worlds. Instead of developing a new deep learning model, we designed a robust training architecture that is simple and generalizable. We present ***Optron****, a general training architecture incorporating the idea of training-in-the-loop. By iteratively optimizing the prediction result of a deep learning model through a plug-and-play optimizer module in the training loop, Optron introduces pseudo ground truth to an unsupervised training process. And by bringing the training process closer to that of supervised training, Optron can consistently improve the models' performance and convergence speed. We evaluated our method on various combinations of models and datasets, and we have achieved state-of-the-art performance on the IXI dataset, improving the previous state-of-the-art method TransMorph by a significant margin of +1.6% DSC. Moreover, Optron also consistently achieved positive results with other models and datasets. It increases the validation DSC for VoxelMorph and ViT-V-Net by +2.3% and +2.2% respectively on IXI, demonstrating our method's generalizability. Our implementation is publicly available at https://github.com/miraclefactory/optron*

1. Introduction

Medical image registration is a crucial task in medical image analysis. It is a problem in which we want to find transformations (i.e., a deformation field) to fit two medical images (such as volumetric CT scans) together in such a way that they look as similar as possible.

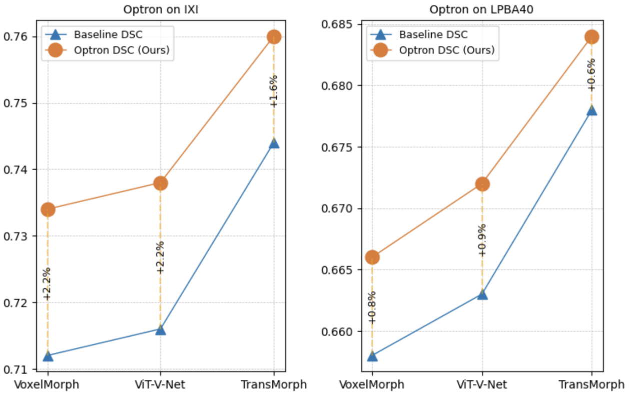

Figure 1. We benchmarked Optron with various deep learning models (VoxelMorph [7], ViT-V-Net [12], TransMorph [11]) and datasets (showing IXI [25] and LPBA40 [39] here). We observed significant improvement over the purely deep-learning-based methods on Dice score. We have achieved +1.6% higher DSC on IXI compared with the previous state-of-the-art method TransMorph [11].

Traditionally, optimization-based methods like [4, 6, 32], are used to iteratively improve the deformation field, doing so will enable the pair of images (usually referred to as fixed and moving image) to fit better together. More recently, with the emergence and dominance of deep learning methods, the field of image registration has also adopted these types of methods which predict deformation fields directly from the fixed and moving images. These two types of methods have been producing impressive results, however, since the emergence of models using Transformer-

[*] Equal contribution

based [15] backbone networks [11, 12], the deep learning approach has gradually surpassed traditional methods as the new bar of image registration tasks. Nonetheless, traditional optimization-based methods still have their merits. While these optimization-based methods tend to be more computationally intensive, they are more stable and produce smoother deformation fields. If we find a way to combine the advantages of both paradigms, we can potentially improve upon the current state-of-the-art methods for medical image registration.

In this paper, we will present and explore a novel training architecture, called ***Optron***, which utilizes optimized deformation fields as pseudo ground truth to provide pseudo supervision for deep networks during the training process. In our method, the training architecture will be divided into two stages, the prediction stage and the optimization stage. Firstly, a deep network is used to predict a deformation field during the forward pass, then, this predicted deformation field will be used as the initialization parameters of the optimizer module. The optimizer module will then iteratively refine its parameters (the deformation field) and produce an optimized deformation field. The optimized field will be used as pseudo ground truth to supervise the deep network. This pseudo supervision is derived by calculating the mean square error loss between the optimized deformation field and the one predicted by the deep learning network. This heuristic will be used to guide the unsupervised training process and make it pseudo-supervised. By bringing the training process closer to that of supervised training, we can improve the model's performance and increase its convergence speed.

We benchmarked our method using various deep learning models and datasets. Optron showed a consistent advantage over the original deep-learning-based methods. In order to demonstrate the architecture's robustness and generalizability, we didn't introduce any model-specific adjustment or hyperparameter settings. On the previous state-of-the-art model TransMorph [11], we were able to improve its performance by +1.6% on DSC on the IXI dataset [25] by training it with the Optron architecture. We have also observed a significant decrease in Jacobian determinant which suggests the model now predicts smoother deformation fields with less spatial folding and topological anomalies.

The main contributions of our work are summarized as follows:

- We present the Optron training architecture, an effective, generalizable, plug-and-play method to improve the training outcome of deep learning registration models.

- We incorporated the idea of training-in-the-loop to build a close collaboration between deep learning models and the optimizer module within the training loop. Using pseudo supervision provided by the optimizer module, we can train the models to a higher capacity compared with unsupervised training.

- We benchmarked our method with various models and datasets, and we have achieved state-of-the-art performance with TransMorph [11] on the IXI dataset [25], along with consistent performance increase with other models tested.

2. Related Work

Deformable image registration computes a dense correspondence between two images, typically involving two steps: an initial affine transformation for global alignment, and a subsequent deformable transformation with a higher degree of freedom. In this work, our primary focus is on the latter step.

2.1. Traditional Registration Methods

Traditional registration methods solve an optimization problem in the space of deformation fields with constraints. They usually minimize a custom energy function iteratively [1]:

$$L(I_m, I_f, \phi) = L_{sim}(I_m \circ \phi, I_f) + L_{reg}(\phi) \qquad (1)$$

where I_m and I_f represent the moving image and fixed image respectively, ϕ denotes the deformation field, and $\circ$ represents the transformation function which warps X to Y using the deformation field ϕ. The energy function typically consists of two terms. The first term L_{sim} evaluates the differences between the warped image and reference image, and the second term L_{reg} quantifies the smoothness of the deformation field.

These traditional methods include elastic-type methods [6, 32], free-form deformations with b-splines [24], statistical parametric mapping [2], discrete methods [3, 17], and Demons [38]. Additionally, some diffeomorphic methods have demonstrated comparable or even superior results, such as Large Diffeomorphic Distance Metric Mapping (LDDMM) [8, 16], symmetric normalization (SyN) [4] and stationary velocity fields (SVF) [30].

These approaches typically formulate the registration problem as an iterative optimization task, making them computationally intensive and relatively impractical for use in clinical settings.

2.2. Deep-Learning-based Registration Methods

The learning-based registration algorithms [7, 13, 19, 27–29, 31, 33–36, 40] optimize the energy function in Eq. (1) for a whole training dataset, and therefore acquire a global representation of the images being registered. As a result,

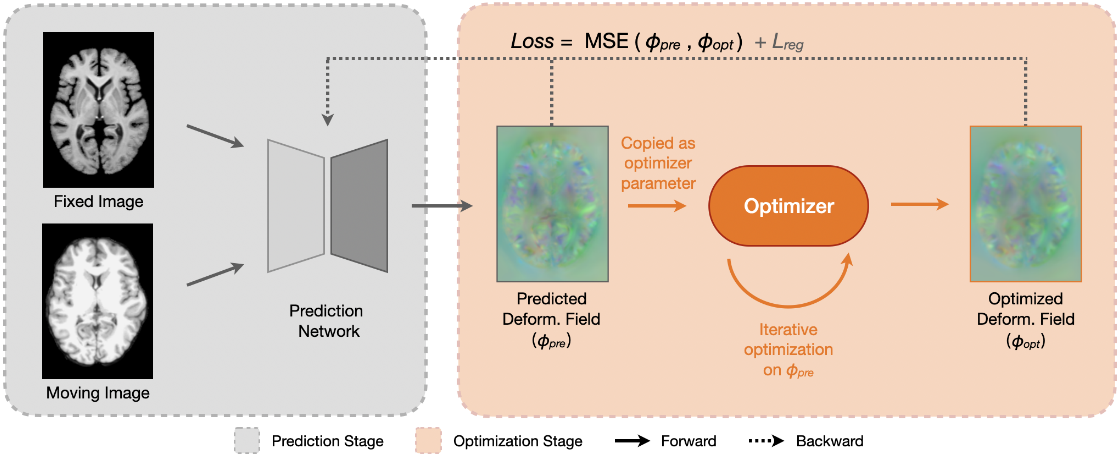
$Loss$ = MSE (ϕ_{pre} , ϕ_{opt}) + L_{reg}
Fixed Image
Moving Image
Prediction Network
Predicted Deform. Field (ϕ_{pre})
Copied as optimizer parameter
Optimizer
Iterative optimization on ϕ_{pre}
Optimized Deform. Field (ϕ_{opt})
Prediction Stage
Optimization Stage
Forward
Backward

Figure 2. The overall structure of the proposed Optron architecture. The whole architecture is divided into two parts, the prediction stage (any network that can produce a deformation field given a pair of images), and the optimization stage. Optron uses the idea of training-in-the-loop to integrate the optimizer into the training process of the prediction network. The optimizer will iteratively refine the deformation field produced by the prediction network, the derived optimized deformation field will then be used as pseudo ground truth to train the prediction network. As this graph illustrates, a loop is formed between the prediction network and the optimizer module. Detailed internal structure of the optimizer module is in Figure 3.

they can easily predict deformation vector fields (DVFs) for unseen data. Previous learning-based methods are categorized into two main types: ConvNet-based methods and Transformer-based methods.

ConvNet-based methods. Some ConvNet-based approaches require the supervision of ground truth deformation vector fields. These ground truths are typically generated by traditional registration methods or manual annotation. Quicksilver [40] is a supervised deep encoder-decoder network that generates the momentum-parameterization of LDDMM [8] from image patches. RegNet [36] estimates the DVF from a pair of input images and is trained by artificially generated DVFs. However, a limitation of these supervised learning methods is that the ground-truth deformation fields are usually computationally expensive to generate, and these methods often require high-quality ground truths to achieve accurate results.

Therefore, unsupervised ConvNet-based methods have been developed to address the limitations of supervised learning methods. One key innovation is the spatial transformer network (STN) [21]. VoxelMorph [7] uses a CNN to predict the deformation field, then, STN is used to reconstruct one image from another. Bob D. de Vos et al. [14] proposed a framework for affine and deformable image registration, offering a method for coarse-to-fine image registration. However, these ConvNet-based methods above are usually limited in learning global and long-range dependent information.

Transformer-based methods. In recent developments, Transformer-based architectures have been applied in many computer vision tasks, such as image recognition [15], object detection [9], and segmentation [37]. In medical image registration, transformer-based methods solve the limitations of ConvNets and achieve state-of-the-art performance. DTN [41] uses a transformer block over the CNN backbone to capture semantic contextual relevance. ViT-V-Net [12] is the first to apply Vision Transformer [15] in volumetric image registration, which embeds a ViT block at the bottleneck of a U-Net architecture [20] to learn long-distance relationships between high-level features of the moving and fixed images. TransMorph [11] is a new hybrid Transformer-ConvNet model, which uses Swin Transformer [23] as the encoder and ConvNet as the decoder.

3. Method

In this work, we present a novel, robust, two-stage training architecture called Optron, utilizing the idea of training-in-the-loop. It demonstrated consistent improvement across various deep learning models and datasets for medical image registration tasks. In the following section, we will discuss the overall architecture of Optron, the design of the optimizer module, alongside the implementation details of the architecture.

3.1. Optron Architecture

Figure 2 presents the overall structure of the proposed Optron architecture. There are two stages in our method. The first stage consists of a prediction model. The prediction model takes in the fixed image I_f and the moving image I_m, and constructs a nonlinear transformation function

F through its deep network which generates a dense deformation field ϕ for each image pair I_f and I_m, i.e.,

$$F_\theta(I_f, I_m) = \phi \tag{2}$$

where θ denotes the parameters of the deep neural network. Since the first stage can be any deformable registration model that predicts a deformation field, we used several popular models, such as VoxelMorph [7], ViT-V-Net [12] and TransMorph [11], as our prediction model in the experiment later. It is the nature of our method that the Optron architecture could in theory work with any of such models, and it is a plug-and-play method that can be easily applied. This generalizability will be further examined in later sections.

The second stage uses an optimizer module that iteratively optimizes the initial deformation field ϕ_{pre} predicted by the first stage. Subsequently, the optimized deformation field ϕ_{opt} is utilized as the pseudo ground truth to provide supervision for the prediction model during training time, forming a loop between the prediction model and the optimizer module. More details about the internal implementation of the optimizer module can be found in Sec. 3.2.

Our method, Optron, establishes a strong collaboration between learning-based methods (first stage) and iterative optimization-based methods (second stage). Within a training loop, the deformation field generated by the prediction model is used to initialize the optimizer module, which then iteratively refines the deformation field. Subsequently, the optimized deformation field in turn is used to supervise the training of the prediction model. During inference time, we exclude the optimizer module and only use the prediction model itself to generate deformation fields enabling time-efficient registration. Using this architecture will provide several crucial benefits, most notably the two below.

Computational Efficiency. In contrast to traditional registration methods, in which optimization is typically very slow without a good initialization from the prediction model, our optimization-based approach is much faster. With a reasonable initial estimate, the iterative optimization process can be accelerated dramatically. Intuitively, this initialization, often approaching the actual optimal solution, enables the optimizer to bypass extensive search over the parameter space to a large extent. Moreover, a limited number of iterations is set to make this approach more practical within the training loop.

Self-improving. Our proposed architecture is self-improving by nature. A well-estimated initial deformation field from the prediction model can lead to a better optimized deformation field. Likewise, a better-deformation field from iterative optimization offers better pseudo supervision to the prediction model. This underscores the significance of training-in-the-loop since it enables a tight collaboration between these two components.

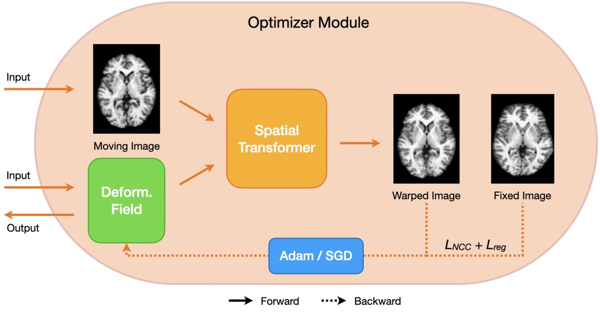

Figure 3. A diagram of the internal structure of the proposed optimizer in Optron. The optimizer takes as inputs the moving image and an initial deformation field predicted by a deep learning model, and it outputs the optimized deformation field.

3.2. Optimizer

An efficient optimizer module is the key for training-in-the-loop. We explored three different optimization strategies: cascaded and downsample approaches and the proposed optimizer as shown in Figure 3. Below we provide the details on these methods.

Cascaded Optimizer. One possible approach is to use a cascaded network that sequentially warps the moving image [42]. This approach refines the warped image at each cascade. Intuitively, this optimization process is similar to the idea of residuals, as it predicts small displacements towards the optimal field in each cascade. However, this approach is computationally intensive, which is typically impractical within the training loop.

Downsample Optimizer. Optimizing the entire deformation field at once might be challenging [18]. By downsampling the deformation field, we can potentially decrease the number of parameters being updated for every optimization iteration. Additionally, downsampled deformation field can provide coarse-resolution supervision [27], which is crucial to long-range displacement and the smoothness of the deformation field. This approach might sound reasonable, but our experiments showed that this method was ineffective and could not optimize the deformation field adequately due to the loss of information during the downsampling procedure.

Optron Optimizer. The basic structure of the proposed Optron optimizer module is illustrated in Figure 3. The optimizer takes in the deformation field generated by the prediction model as learnable parameters. Internally, it contains a single Spatial Transformer Network (STN) [21] layer, which introduces no additional parameter. Consequently, the only learnable parameters in the optimizer is the deformation field. To optimize a deformation field, the optimizer evaluates a loss using the current deformation field, then backpropagate the gradient directly back to the deformation field and update it using methods like Adam [22] or Stochastic Gradient Descent (SGD). Compared with the approaches we mentioned above, it is a simple yet effective design. The purpose of using this design over the other ones is that:

- In order to prove the usefulness of the Optron architecture, we do not need a complex optimizer. In fact, a simple one is more desirable for its structural simplicity ensures minimal effect on the optimization outcome.
- Optimizing a unique deformation field for different pairs of images instead of using a shared set of parameters can provide more image-specific heuristics and thus benefit more when used as pseudo ground truth.
- Optimizing the deformation field directly provides higher degrees of freedom with parameter adjustment, facilitating the production of better pseudo ground truth.

In the ablation study later on, we empirically demonstrated that our optimizer design achieves the highest registration accuracy compared to other approaches.

3.3. Implementation Detail

Overall Loss. Let ϕ_{pre}, ϕ_{opt} denote the predicted deformation field and optimized deformation field respectively. Given these two values, we can train our prediction model with loss supplied by pseudo supervision:

$$L_{opt} = ||\phi_{pre} - \phi_{opt}|| \tag{3}$$

In our experiments, we used MSE for L_{opt}.

The overall loss function L_{all} for the prediction model training consists of two parts: L_{opt} that penalizes the difference between ϕ_{pre} and ϕ_{opt}, and L_{reg} that regularizes the predicted deformation field ϕ_{pre}:

$$L_{all} = L_{opt} + \lambda L_{reg} \tag{4}$$

where λ is the weight of the regularization loss. λ is preferably set to 0.02. L_{reg} acts as a diffusion regularizer on the spatial gradients of deformation field ϕ:

$$L_{reg}(\phi) = \sum_{p \in \Omega} ||\nabla \phi(p)||^2 \tag{5}$$

Optimizer Module Loss. Like Eq. (1), the objective function to be minimized in the optimizer module consists of two components: a similarity loss term that computes the difference between I_m and I_f and a regularization loss term that imposes smoothness in ϕ.

The regularization term we used is the same as Eq. (5) and the similarity metric we used is the local normalized cross-correlation (NCC) between warped image $I_m \circ \phi$ and fixed image I_f:

$$LNCC(I_f, I_m \circ \phi) = \sum_{p \in \Omega} \frac{(\sum_{p_i}(f(p_i) - \hat{f}(p))([I_m \circ \phi](p_i) - [\hat{I}_m \circ \phi](p)))^2}{(\sum_{p_i}(f(p_i) - \hat{f}(p))^2)(\sum_{p_i}([I_m \circ \phi](p_i) - [\hat{f}_m \circ \phi](p))^2)} \tag{6}$$

where $\hat{I}_f(p)$ and $\hat{I}_m(p)$ represent the mean voxel value within a local window of size n^3 centered at voxel p.

4. Experiments

4.1. Datasets and Preprocessing

Here, we provide a concise overview of datasets we used for both training and evaluation. Our reported results are mainly based on three publicly available datasets of 3D brain MRI: the Information eXtraction from Images (IXI) dataset [25], the OASIS dataset [26] and LONI Probabilistic Brain Atlas 40 (LPBA40) dataset [39]. FreeSurfer [5] was used to perform standard preprocessing procedures on structural brain MRI data for all datasets. These procedures include skull stripping, resampling, and affine transformation. Subsequently, the preprocessed volumes were cropped to a uniformed size. Detailed descriptions of each dataset can be found below.

IXI dataset. In the context of atlas-to-patient brain MRI registration, we used the pre-processed IXI dataset provided by [11]. This dataset contains a total of 576 T1-weighted brain MRI images and each image has a resolution of $160 \times 192 \times 224$. We randomly selected 200 volumes for training and 20 volumes for validation. 30 anatomical structures in segmentation maps were used to evaluate registration performance.

OASIS dataset. For inter-patient registration task, we used the preprocessed OASIS dataset provided by [11]. This dataset consists a total of 451 brain T1 MRI images with size $160 \times 192 \times 224$. We randomly selected 200, 19 volumes for training and validation, respectively. 35 anatomical structures in segmentation maps were used to evaluate registration performance.

LPBA40 dataset. We additionally used the LPBA40 dataset [39] to validate our method. It includes 40 patients with brain MRI images and ground truth segmentation masks. Each image has a resolution of $160 \times 192 \times 160$. We used 30 volumes for training, 9 volumes for validation, and 1 volume as atlas. 54 anatomical structures in segmentation maps were used to evaluate registration performance.

4.2. Evaluation Metrics

We used Dice Similarity Coefficient (DSC), Jacobian Determinant and Registration Speed to evaluate the performances of our deformable registration methods. DSC measures the volume overlap between anatomical segmen-

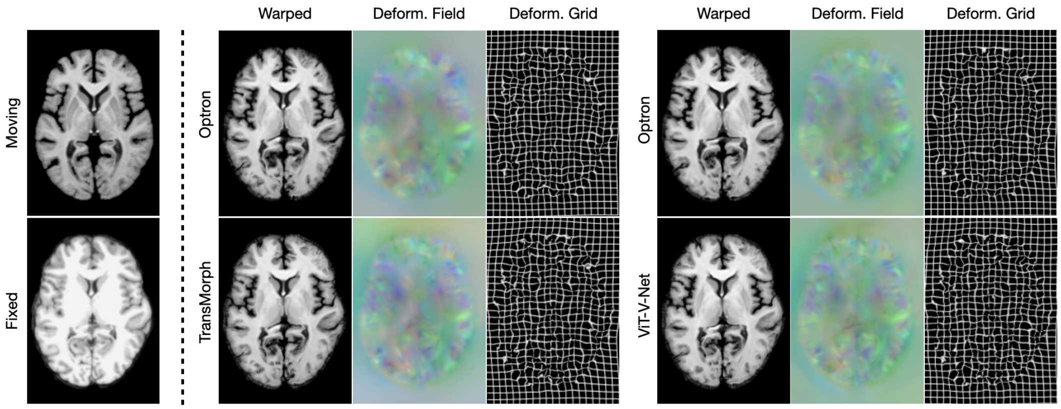

Figure 4. Visualization of registration results. This is an arbitrary demo extracted from the comparison results between baseline TransMorph, ViT-V-Net (row 2) and their respective model trained with Optron (row 1), demo from the IXI dataset [25]

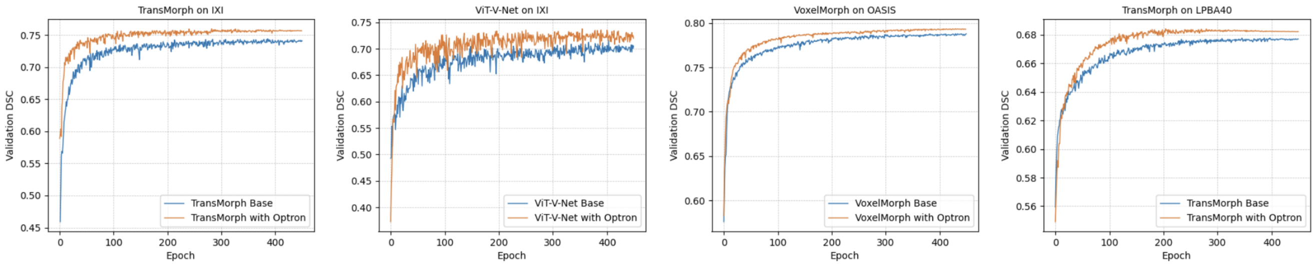

Figure 5. Demo visualization of training process vs. validation DSC for some combinations of models and datasets.

tations of the fixed image and warped moving image. Negative Jacobian Determinant captures local distortion in the neighborhood. We quantify the regularity of the deformation fields using the percentage of non-positive values in the determinant of the Jacobian matrix on the deformation fields. A lower percentage of negative Jocobian determinant indicates a smoother deformation field. We additionally computed the average time of registration of each image pair to evaluate the registration speed.

4.3. Experiment Settings

All models were trained on a NVIDIA RTX 4090 GPU for 500 epochs using the Adam optimization algorithm, with an initial learning rate of $1e-4$ and a batch size of 1. The dataset was augmented with flipping in random directions during training. For the optimization model in our architecture, we used an initial learning rate of 0.1 and an optimization iteration number of 10 during training.

4.4. Baseline Methods

We validated our proposed architecture based on various registration methods that have previously demonstrated state-of-the-art performance in registration tasks. We compared our method to two traditional methods. Detailed hyperparameter settings for each method are as follows:

- SyN [4]: For all datasets, the mean squared error(MSE) was used as the objective function, along with three scales with 160, 80, 40 iterations, respectively.
- NiftyReg [10]: The sum of squared difference (SSD) was used as the objective function. We used three scales with 300 iterations each by default.

We also compared our method with several state-of-the-art deep-learning-based methods. For a fair comparison and more reliable results, we used LNCC (Eq. (6)) and a regularizer (Eq. (5)) as the loss function (Eq. (4)) for all models. The regularization hyperparameter λ was set to 1 suggested in [7] as optimal values. Note that we didn't use Dice loss. Detailed hyperparameter settings for each method are as follows:

- VoxelMorph [7]: This registration network was based on U-net [20]. We used the default parameters of VoxelMorph-1 proposed in [7].
- ViT-V-Net [12]: This registration network was developed based on ViT [15]. We applied the default network hyperparameter settings suggested in [12].

Datasets	Methods	Base. DSC	**Optron DSC**	Base. $\lvert J_\phi \rvert < 0$ (%)	**Optron** $\lvert J_\phi \rvert < 0$ (%)	Inference Time (s)
IXI [25]	SyN [4]	0.647	N/A	1.96e-6	N/A	277(CPU)
	NiftyReg [10]	0.585	N/A	0.029	N/A	22.4(CPU)
	VoxelMorph [7]	0.714	**0.737**	1.398	**0.516**	0.061(GPU)
	ViT-V-Net [12]	0.716	**0.738**	1.543	**0.545**	0.615(GPU)
	TransMorph [11]	0.744	**0.760**	1.433	**0.794**	0.114(GPU)
OASIS [26]	SyN [4]	0.769	N/A	1.58e-4	N/A	258(CPU)
	NiftyReg [10]	0.762	N/A	0.011	N/A	25.0(CPU)
	VoxelMorph [7]	0.788	**0.794**	0.911	**0.490**	0.061(GPU)
	ViT-V-Net [12]	0.794	**0.809**	0.887	**0.487**	0.646(GPU)
	TransMorph [11]	0.818	**0.818**	0.765	**0.517**	0.160(GPU)
LPBA40 [39]	SyN [4]	0.703	N/A	1.18e-4	N/A	172(CPU)
	NiftyReg [10]	0.691	N/A	1.13e-3	N/A	22.8(CPU)
	VoxelMorph [7]	0.658	**0.666**	0.288	**0.023**	0.046(GPU)
	ViT-V-Net [12]	0.663	**0.672**	0.390	**0.112**	0.446(GPU)
	TransMorph [11]	0.678	**0.684**	0.438	**0.150**	0.327(GPU)

Table 1. Evaluation results for different methods on various datasets. The Optron architecture provides significant improvement on the purely deep learning methods. All inference time measured on CPU is tested on a 8-core Intel Xeon CPU (Skylake), all inference time measured on GPU is tested on a NVIDIA RTX 4090 GPU.

- TransMorph [11]: This registration network was developed based on Swin Transformer [23]. We applied the default hyperparameter settings of TransMorph in [11].

4.5. Results

4.5.1 Optron on IXI

We evaluated our method on the IXI dataset which is also used in TransMorph [11], so that we can make quick comparison on the enhancement of using the Optron architecture on the state-of-the-art method. Our evaluation result on the IXI dataset is shown in Tab. 1. The first image in Figure 5 visualizes the training process of TransMorph on IXI.

Notably, Optron considerably increased DSC and convergence speed over the original deep learning models. In particular, we were able to improve upon the previous state-of-the-art method TransMorph [11] on the IXI dataset by +1.6% on DSC and decreasing its percentage of $|J_\phi| < 0$ by 44.6%. This suggests that not only can Optron improve the model's performance, but it can also prevent the model from producing an over-sharpened deformation field, which can be easily seen by visualizing actual registration results as shown in Figure 4.

We have also improved other deep learning methods considerably, with a +2.3% increase on DSC for VoxelMorph [7] and a +2.2% increase on DSC for ViT-V-Net [12], further validating the effectiveness and generalizability of our method.

4.5.2 Optron on OASIS

OASIS [26] is another commonly used dataset in medical image registration. Methods commonly achieved higher DSC on this dataset. The results are available in Tab. 1.

We found that Optron introduced a limited improvement over TransMorph on OASIS. This can be explained as the OASIS dataset is less challenging for TransMorph that has strong fitting capability. This claim can be easily validated by comparing the convergence loss and baseline DSC of all the models across different datasets as shown in Tab. 2 and Tab. 1 respectively. Across mutiple combinations of 3 models and 3 datasets, TransMorph on OASIS has the lowest convergence loss and highest DSC, which verifies our thoughts.

Additionally, the training process of ViT-V-Net and VoxelMorph on OASIS are visualized in the second and third images in Figure 4, which further proves that our method can significantly improve performances and convergence speed.

4.5.3 Optron on LPBA40

LPBA40 [39] is another dataset we used to validate our method. Evaluation results are presented in Tab. 1. With only 40 volumes, deep learning models can easily overfit on this dataset. This can be partially observed in the 4th image in Figure 4.

On a small dataset, the extra supervision provided by our method can be crucial due to the lack of training data. A su-

Converg. Loss	IXI [25]	OASIS [26]	LPBA40 [39]
VoxelMorph [7]	-0.222	-0.259	-0.183
ViT-V-Net [12]	-0.242	-0.266	-0.215
TransMorph [11]	-0.262	**-0.283**	-0.232

Table 2. Comparison between the 3 models' convergence losses on 3 datasets. Note that the loss function settings were the same. For TransMorph on OASIS [26], the convergence loss is clearly the lowest.

α:β	1:1	1:10	1:100	1:1000	0
DSC	0.427	0.407	0.453	0.489	**0.728**

Table 3. Ablation study on the impact of the ratio of α (weight of L_{sim}) to β (weight of L_{opt}). The weight of L_{reg}, λ, is always 0.02. Specifically, for 1:10, we simply set α and β to 1 and 10 respectively, etc. In the case that the ratio of α to β equals 0 (i.e. Eq. (4)), α and β equal to 0 and 1 respectively. This study was conducted using TransMorph [11] with Optron on the IXI dataset [25] and we compared DSC at the 30th epoch. Note that the raw DSC after affine preprocessing is 0.407. Except the overall loss function, other experiment settings remains the same.

pervision that can motivate the training process by giving the model more challenge, can benefit the model and yielding a better result.

4.6. Ablation Study

4.6.1 Loss Composition

Since we developed our method based on widely-used unsupervised methods that utilize image similarity metrics to train networks, our initial idea was to add L_{opt} to the unsupervised loss function in [7] as an extra and auxiliary pseudo supervision, i.e.

$$\begin{aligned} L_{all} &= L_{us} + \beta L_{opt} \\ &= \alpha L_{sim} + \lambda L_{reg} + \beta L_{opt} \end{aligned} \tag{7}$$

We experimented with several different ratios of α to β. However, all these ratios led to abnormal validation results, in which the dice score did not increase during the training process and even decreased and the model failed to converge.

Tab. 3 shows the experimental results of different ratios of the L_{sim} weight to the L_{opt} weight. The reason is likely that L_{sim} and L_{opt} had very different effects on training the network and led the model towards two different directions. When they were put together, their effects offset each other, and even led the model to produce worse results. Hence, we removed the term L_{sim} in Eq. (7), and set β, λ to $1, 0.02$ respectively (i.e. Eq. (4)) aiming to maximize the advantage of pseudo ground truth. Experimental results demonstrate that our overall loss function(i.e. Eq. (4)) is reasonable and can obtain better registration accuracy.

Optimizer Designs	DSC	Time (s)	VRAM Usage
Cascaded-1	0.547	6.109	27.10 GB
Cascaded-2	0.599	7.372	19.27 GB
DownSample	0.610	0.094	12.44 GB
Ours	**0.654**	3.228	13.07 GB

Table 4. Ablation study with different optimizer schemes on LPBA40 dataset. We evaluated their performance using the validation dataset during the first 30 epochs.

4.6.2 Optimizer Design

Tab. 4 shows the performance of different optimizer schemes, namely: (1) Cascaded Optimizer, (2) Downsample Optimizer, and (3) the proposed Optron optimizer. The hyperparameter configurations for each optimizer module are as follows: 2 cascades and 5 optimization iterations for Cascaded-1 Optimizer, 1 cascade and 10 optimization iterations for Cascaded-2 Optimizer, and 10 optimization iterations for (2) and (3). The initial learning rate for network-based optimizer is $1e-4$, for (2) and (3) is 0.1.

Among all the compared strategies, the proposed Optron optimizer module achieves the highest dice score while maintaining practical optimization time within the training loop. Surprisingly, despite requiring more time for the optimization of a pair of images, the network-based optimizer modules (Cascaded optimizer) fails to yield better performance. It appears that the Cascade network works more effectively as an integrated network rather than as an optimizer module, though this integration also result in longer inference time.

5. Conclusion

This work introduces Optron, a general architecture that combines learning-based methods and an optimization-based method to train a deep learning network for medical image registration tasks. Our approach uses the learning-based models to predict an initial deformation field for the optimizer module, which then refines the deformation field iteratively and provides pseudo ground truth for the training of the learning-based models.

Our proposed architecture outperforms previous approaches by large margins, and we were able to achieve state-of-the-art performance on multiple datasets, proving its effectiveness and generalizability. Our method demonstrates promising performance, and it provides a solid base for future works to build upon. The training process of Optron is still more time-consuming compared with pure deep learning methods due to the extra optimization steps for each pair of images. Future work could consider improving the design and efficiency of the optimizer module which can benefit this architecture to a greater extent.

References

[1] Sotiras A, Davatzikos C, and Paragios N. Deformable medical image registration: a survey. *IEEE Trans Med Imaging*, 2013. 2

[2] John Ashburner and Karl J. Friston. Voxel-based morphometry—the methods. *NeuroImage*, 2000. https://doi.org/10.1006/nimg.2000.0582. 2

[3] Dalca AV, Bobu A, Rost NS, and Golland P. Patch-based discrete registration of clinical brain images. *Patch Based Tech Med Imaging (2016)*, 2016. 2

[4] 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. *Medical Image Analysis*, 12(1):26–41, 2008. Special Issue on The Third International Workshop on Biomedical Image Registration – WBIR 2006. 1, 2, 6, 7

[5] Fischl B. Freesurfer. *Neuroimage*, 2012. https://surfer.nmr.mgh.harvard.edu/. 5

[6] Ruzena Bajcsy and Stane Kovačič. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, 46(1):1–21, 1989. 1, 2

[7] 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, aug 2019. 1, 2, 3, 4, 6, 7, 8

[8] Ceritoglu C, Wang L, Selemon LD, Csernansky JG, Miller MI, and Ratnanather JT. Large deformation diffeomorphic metric mapping registration of reconstructed 3d histological section images and in vivo mr images. *Front Hum Neurosci*, 2010. 2, 3

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. *arXiv e-prints*, page arXiv:2005.12872, May 2020. 3

[10] UK Centre for Medical Image Computing, University College London. Niftyreg, 2023. http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg. 6, 7

[11] 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, 2, 3, 4, 5, 7, 8

[12] Junyu Chen, Yufan He, Eric C. Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration, 2021. 1, 2, 3, 4, 6, 7, 8

[13] Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration. *arXiv e-prints*, page arXiv:1805.04605, May 2018. 2

[14] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, feb 2019. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2, 3, 6

[16] Beg M. Faisal, Miller Michael I, Trouvé Alain, and Younes Laurent. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 2005. 2

[17] Ben Glocker, Nikos Komodakis, Georgios Tziritas, Nassir Navab, and Nikos Paragios. Dense image registration through mrfs and efficient linear programming. *Medical Image Analysis*, 12(6):731–741, 2008. Special issue on information processing in medical imaging 2007. 2

[18] Kun Han, Shanlin sun, Xiangyi Yan, Chenyu You, Hao Tang, Junayed Naushad, Haoyu Ma, Deying Kong, and Xiaohui Xie. Diffeomorphic Image Registration with Neural Velocity Field. *arXiv e-prints*, page arXiv:2202.12498, Feb. 2022. 4

[19] Jing Hu, Ziwei Luo, Xin Wang, Shanhui Sun, Youbing Yin, Kunlin Cao, Qi Song, Siwei Lyu, and Xi Wu. End-to-end multimodal image registration via reinforcement learning. *Medical Image Analysis*, 68:101878, 2021. 2

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018. 3, 6

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks, 2016. 3, 4

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 3, 7

[24] Dirk Loeckx, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Nonrigid image registration using free-form deformations with a local rigidity constraint. In Christian Barillot, David R. Haynor, and Pierre Hellier, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004*, pages 639–646, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. 2

[25] Imperial Collage London. Information extraction from images, 2023. https://brain-development.org/ixi-dataset/. 1, 2, 5, 6, 7, 8

[26] 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, 09 2007. 5, 7, 8

[27] Tony C. W. Mok and Albert C. S. Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks, 2020. 2, 4

[28] Tony C. W. Mok and Albert C. S. Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks, 2021. 2

[29] Tony C. W. Mok and Albert C. S. Chung. Affine medical image registration with coarse-to-fine vision transformer, 2022. 2

[30] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: Learning deformable image registration using shape matching. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, pages 266–274, Cham, 2017. Springer International Publishing. 2

[31] Ameneh Sheikhjafari, Michelle L. Noga, K. Punithakumar, and Nilanjan Ray. Unsupervised deformable image registration with fully connected generative neural network, 2018. 2

[32] Dinggang Shen and C. Davatzikos. Hammer: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging*, 21(11):1421–1439, 2002. 1, 2

[33] Zhengyang Shen, Xu Han, Zhenlin Xu, and Marc Niethammer. Networks for Joint Affine and Non-parametric Image Registration. *arXiv e-prints*, page arXiv:1903.08811, Mar. 2019. 2

[34] Zhengyang Shen, François-Xavier Vialard, and Marc Niethammer. Region-specific Diffeomorphic Metric Mapping. *arXiv e-prints*, page arXiv:1906.00139, May 2019. 2

[35] Jiacheng Shi, Yuting He, Youyong Kong, Jean-Louis Coatrieux, Huazhong Shu, Guanyu Yang, and Shuo Li. XMorpher: Full Transformer for Deformable Medical Image Registration via Cross Attention. *arXiv e-prints*, page arXiv:2206.07349, June 2022. 2

[36] Hessam Sokooti, Bob de Vos, Floris Berendsen, Boudewijn P. F. Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, pages 232–239, Cham, 2017. Springer International Publishing. 2, 3

[37] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for Semantic Segmentation. *arXiv e-prints*, page arXiv:2105.05633, May 2021. 3

[38] J.-P. Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998. 2

[39] Laboratory of Neuro Imaging University of Southern California. Loni probabilistic brain atlas (lpba40), 2023. https://loni.usc.edu/research/atlases. 1, 5, 7, 8

[40] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration - a deep learning approach, 2017. 2, 3

[41] Yungeng Zhang, Yuru Pei, and Hongbin Zha. Learning dual transformer network for diffeomorphic registration. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 129–138, Cham, 2021. Springer International Publishing. 3

[42] Shengyu Zhao, Yue Dong, Eric Chang, and Yan Xu. Recursive cascaded networks for unsupervised medical image registration. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 4