WEAKLY SUPERVISED LEARNING IN DEFORMABLE EM IMAGE REGISTRATION USING SLICE INTERPOLATION

Thanh Nguyen-Duc[1,‡], *Inwan Yoo*[1,‡], *Logan Thomas*[2], *Aaron Kuan*[2], Wei-chung Lee[2], and Won-Ki Jeong[1,*]

ABSTRACT

Alignment of large-scale serial-section electron microscopy (ssEM) images is crucial for successful analysis in nano-scale connectomics. Despite various image registration algorithms proposed in the past, large-scale ssEM alignment remains challenging due to the size and complex nature of the data. Recently, the application of unsupervised machine learning in medical image registration has shown promise in efforts to replace an expensive numerical computation process with a once-deployed feed-forward neural network. However, the anisotropy in most ssEM data makes it difficult to directly adopt such learning-based methods for the registration of these images. Here, we propose a novel deformable image registration approach based on weakly supervised learning that can be applied to registering ssEM images at scale. The proposed method leverages slice interpolation to improve registration between images with sudden and large structural changes. In addition, the proposed method only requires roughly aligned data for training the interpolation network while the deformation network can be trained in an unsupervised fashion. We demonstrate the efficacy of the method on real ssEM datasets.

Index Terms— EM Image Registration, Electron Microscope, Deep Learning, GPU

1. INTRODUCTION

In cellular-level connectomics research, high-resolution electron microscopy (EM) is used to reveal nano-scale neuronal structures critical for orchestrating the activity of cellular networks hundreds of micrometers to millimeters in size. Brain tissue is sliced into extremely thin sections (about 30 to 50 nanometers in thickness) for two-dimensional (2D) EM imaging with a lateral resolution of 3 to 5 nanometers. Because each tissue section is imaged independently, accurate alignment of 2D EM images to form three-dimensional (3D) EM volumes is crucial for subsequent analyses. The main challenge in EM image registration lies in the large and complex nature of the data: multi-terabyte data size, non-linear tissue distortions, imaging artifacts, and large and sudden structural deformations across sections. Several open-source microscopy image registration tools (AlignTK[1], bUnwarpJ[2], and Elastic [1]) partially address these issues by adopting parallel computing and sparse feature-based image matching, but are often problematic, requiring significant manual parameter tuning and intervention to optimize inter-slice alignment accuracy.

Recent advances in deep learning have yielded encouraging results for many computer vision problems, including optical flow and image registration. Dosovitskiy *et al.* [2] proposed convolutional neural networks trained in an end-to-end fashion using pairs of data (i.e., adjacent frames and the corresponding flow field). Sokooti *et al.* [3] proposed 3D convolutional neural networks trained with the data deformed by randomly generated synthetic vector fields for use in 3D medical volume registration. These supervised learning-based methods achieve fast and accurate image-to-image matching but encounter problems in attempts to generate realistic training data. Earlier work on unsupervised learning in image registration has focused mostly on finding image features automatically via a learning process. Wu *et al.* [4] used patch-based 3D convolutional autoencoders to generate features in 3D MRI volumes. Yoo *et al.* [5] proposed an end-to-end trained 2D convolutional autoencoder to generate feature maps for measuring similarity between EM sections. More recently, unsupervised learning has been used to replace time-consuming iterative minimization processes in image registration [6].

In this paper, we propose a novel image registration method based on unsupervised learning augmented with slice interpolation from light-weight supervised learning (i.e., weakly supervised learning). The main motivation behind our work is that EM image registration is more complicated to solve by using either unsupervised or supervised learning alone. The thickness of conventional EM sections is almost 10 times larger than the lateral pixel resolution; therefore, there can be large structural changes (including the appearance or disappearance of structures) across adjacent sections, which makes registering images more challenging. Since

‡Both authors contributed equally to this work
*Corresponding author. E-mail: wkjeong@unist.ac.kr
[1]Ulsan National Institute of Science and Technology
[2]Harvard Medical School

[1]http://mmbios.org/aligntk-home
[2]https://imagej.net/BUnwarpJ

there is no groundtruth data for EM image alignment, supervised learning-based registration approaches usually generate synthetic training data by deforming the given input image based on the assumption that the source and target images are similar, which is not always the case in real data. Unsupervised learning-based registration approaches learn the deformation field by minimizing the difference between each pair of input images, but large structural changes across EM sections can cause problems. The proposed method is built mainly on an unsupervised learning model that assesses the similarity between input images, but it additionally leverages slice interpolation to predict structural changes between adjacent sections, which contributes to improving robustness and accuracy. Specifically, we employ two interpolation approaches, one is calculation of a simple average (e.g., linear interpolation of two images) and the other is predicted interpolation using a pre-trained deep network that requires only roughly aligned training data. Our method can quickly generate a dense deformation field without manual parameter optimization via the one-way deployment of a feed-forward network that readily scales to align large EM images. We demonstrate the performance of our method by aligning real CREMI dataset and mouse cerebellum dataset (about $250{,}000 \times 150{,}000 \times 372$ voxels; see Fig. 3).

2. METHOD

The proposed EM image registration method consists of two end-to-end deep neural networks: a pre-trained interpolation network, and a deformation network to be trained in an unsupervised fashion (see Fig. 1).

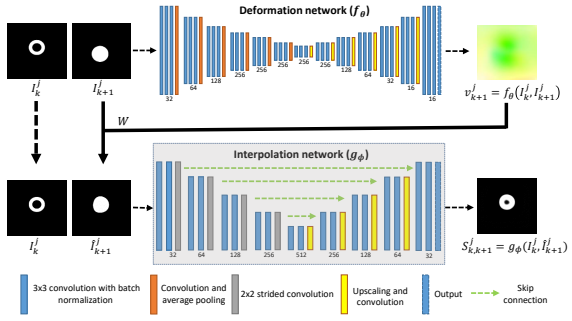

Fig. 1: Overview of the proposed image registration method and network architectures. Two input images (top left) have different structures, so conventional pixel-based methods may fail to correctly align them. Our method uses a pre-trained interpolation network (gray dotted box) to predict the semantically correct intermediate image (bottom right), which allows correct alignment of the images (bottom left).

2.1. Robust slice interpolation network from pre-aligned dataset

A simple, conventional deformation network can be constructed via unsupervised training by minimizing pixel-wise intensity differences between random image pairs [6]. Such an approach is similar to iterative minimization using a pixel difference objective function. However, such a simple intensity-based learning can lead to incorrect image alignment when structures change rapidly across images, which is often found in real-world EM data. To address this issue, we propose using intermediate slice prediction between images (i.e., slice interpolation) during unsupervised training of the deformation network. To get a good intermediate slice prediction, we use an interpolation network trained using roughly aligned EM stacks (which can be done by way of global alignment in a coarse resolution). The proposed interpolation network accepts two consecutive images and predicts the intermediate image between them (therefore, three consecutive EM sections are used to train the network). Let I_k^j be the jth image patch of the kth slice in the training EM image stack. Then, an interpolation network g , parameterized by , that predicts the intermediate slice from two adjacent slices can be trained by minimizing the following loss function:

$$L = \sum_k \sum_j \| g(I_k^j; I_{k+2}^j) - I_{k+1}^j \|_2^2 \quad (1)$$

We employed a simplified U-net architecture [7] to implement the interpolation network. In our implementation, we reduced the number of features by half compared to the original U-net, and used exponential linear units (ELU) [8] for the activation function. For the fast and stable convergence, we employed batch normalization to all except the last layer. We applied zero-padding in decoder layers to create an interpolated image of the original scale.

Because the pre-aligned stacks used for training are not perfectly aligned, the interpolation network is easily overfitted and may not perform well. To remedy this problem, we develop a special training scheme, *random noise inpainting*. The proposed training technique randomly discards subpatches from the two input images and fills in uniform random values, which makes the interpolation network more robust to image artifacts and structural changes across EM sections (see Fig. 2).

2.2. Deformation network with pre-trained interpolation network

The proposed model shown in Fig. 1 is designed using a deformation network, a spatial transformer, and the pre-trained interpolation network to learn weakly supervised deformable image registration. Let f be a deformation network parameterized by . For the input fixed and moving input patches I_k^j and I_{k+1}^j, the pixel-wise deformation v_{k+1}^j is computed by f :

$$v_{k+1}^j = f(I_k^j; I_{k+1}^j) \quad (2)$$

With a differentiable bilinear sampler W, we can warp the moving image I_{k+1}^j to $\hat{I}_{k+1}^j$:

$$\hat{I}_{k+1}^j = W(I_{k+1}^j; v_{k+1}^j) \quad (3)$$

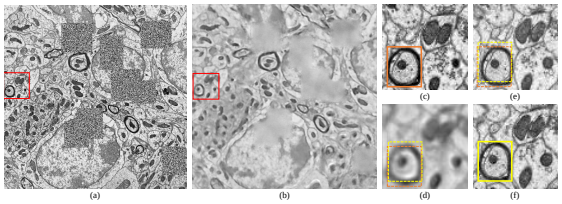

Fig. 2: Random noise inpainting technique for robust training of an interpolation network. (a) and (b): an example of an input section with random noise inpainting and its predicted result; (c), (e), and (f): zoomed-in region of first, second, and third EM sections in which the location of a myelin axon is marked with red and yellow boxes; (d): the predicted result using the interpolation network.

$$S^j_{k;k+1} = g\ (I^j_k; \tilde{I}^j_{k+1}) \tag{4}$$

The interpolation network then takes these fixed patch I^j_k and warped image $\tilde{I}^j_{k+1}$ and predicts an intermediate slice $S^j_{k;k+1}$ to measure the similarity to train the deformation network f . We found that multi-scale loss functions can help in stably training the deformation network. We use average pooling to get a down-sampled loss function. We defined the set of kernel sizes for average pooling as P, and p denotes the pooling kernel size of the down-sampled patches. The objective loss function for training the deformation network is defined as follows:

$$L_d = \sum_k \sum_p \| I^{j;p}_k - \tilde{I}^{j;p}_{k+1} \|^2_2 \tag{5}$$

$$L_i = \sum_k \sum_p \| I^{j;p}_k - S^{j;p}_{k;k+1} \|^2_2 + \| S^{j;p}_{k;k+1} - \tilde{I}^{j;p}_{k+1} \|^2_2 \tag{6}$$

$$L_s = \sum_k (\| r_x v^j_{k+1} \|^2_2 + \| r_y v^j_{k+1} \|^2_2) \tag{7}$$

$$L = \frac{L_d + L_i}{2} + L_s \tag{8}$$

where Eq. 5 and 6 represent data loss terms that measure image matching accuracy using simple averaging and slice interpolation, respectively. A smoothness term for the vector map is given in Eq. 7 in which its contribution to the final loss can be controlled by in Eq. 8. Note that the parameter in the pre-trained interpolation network is fixed during the training of the deformation network.

Our interpolation network is also an end-to-end convolutional neural network with encoding and decoding layers but without skip connections to promote smooth deformation (architecture details can be found in Fig. 1). Its input size and output size are both 512 512. To align large EM sections, we decompose the input image into overlapping patches (stride 256), and process each of them using Eq. 2 in parallel. Generated per-patch vector maps are then blended together using Gaussian weights to construct the entire vector field.

3. RESULTS

A mouse cerebellum (CB) dataset which was acquired using TEMCA imaging [9] and automated GridTape sample handling (in preparation) is used for our evaluation results. The raw image data were approximately 26 TB, with the final stitched and registered EM dataset encompassing approximately 11 TB, spanning, 800 600 20 m and encompassing all three layers of a cerebellar lobule cut along the parasagittal axis (Fig. 3). To assess the alignment quality, we aligned a small cropped test volume (512 512 64 voxels) using several loss functions in our method and other well-known image registration methods, and measured inter-slice similarity using normalized cross correlation (NCC) and structural similarity index (SSIM) metrics (see Tab. 1). The baseline result is generated using AlignTK by aligning the original mouse cerebellum image stacks before cropping out the test volume. We get the best result when both slice average and slice interpolation losses are combined together (Eq. 8), which also outperforms bUnwarpJ and Demons deformable registration results. We applied our method to align the entire mouse cerebellum EM stacks (Fig. 3). We generated pairwise dense vector maps using our method on the images down-sampled by a factor of 16 to match the spring mesh resolution (maximum level 4; $2^4 = 16$), and used AlignTK to run spring mesh relaxation and rendering to generate the aligned full-resolution 3D volume (250; 000 150; 000 372 voxels). As shown in the cross-sectional view in Fig. 3, our method significantly improves the structural continuity across sections.

		NCC	SSIM	DICE
CB	Baseline	0.4153	0.2194	_
	(5)+(7)	0.4984	0.2872	_
	(6)+(7)	0.4813	0.2674	_
	(8)	0.5277	0.3210	_
	Demons	0.4893	0.2870	_
	bUnwarpJ	0.5143	0.3147	_
CREMI	(5)+(7)	0.716	0.517	0.798
	(6)+(7)	0.704	0.505	0.792
	(8)	0.780	0.563	0.848
	Demons	0.611	0.445	0.721
	bUnwarpJ	0.623	0.481	0.736

Table 1: Alignment quality comparison on cropped mouse CB and CREMI datasets

CREMI dataset[3] is also used to quantitatively assess more about the registration quality of our method. We trained the slice interpolation network with *random noise inpainting* and then used it for the deformation network. Finally, the registration was conducted in a volume (1250x1250x124) of raw EM images. The quantitative results are shown that proposed approach achieves better dice coefficient, NCC and

[3] https://cremi.org