



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Master's Thesis

Deformable Image Registration Using
Convolutional Neural Networks for
Connectomics

In-Wan Yoo

Department of Computer Science and Engineering

Graduate School of UNIST

2018

Deformable Image Registration Using Convolutional Neural Networks for Connectomics

In-Wan Yoo

Department of Computer Science and Engineering

Graduate School of UNIST

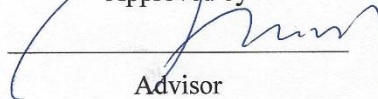
Deformable Image Registration Using Convolutional Neural Networks for Connectomics

A thesis
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Master of Science

In-Wan Yoo

01.03.2018

Approved by



Advisor

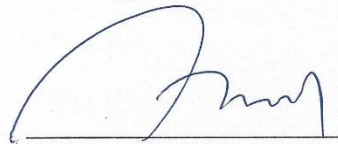
Won-Ki Jeong

Deformable Image Registration Using
Convolutional Neural Networks for
Connectomics

In-Wan Yoo

This certifies that the thesis of In-Wan Yoo is approved.

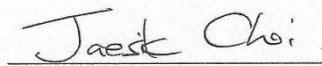
01.03.2018



Advisor: Won-Ki Jeong



Se-Young Chun: Thesis Committee Member #1



Jaesik Choi: Thesis Committee Member #2

Abstract

In this thesis, a new novel method to align two images with recent deep learning scheme called *ssEMnet* is presented. The reconstruction of serial-section electron microscopy (ssEM) images gives critical insight to neuroscientist understanding real brains. However, alignment of each ssEM plane is not straightforward because of its densely twisted circuit structures. In addition, dynamic deformations are applied to images in the process of acquiring ssEM dataset from specimens. Even worse, non-matched artifacts like dusts and folds occur in the EM images.

In recent deep learning researches, especially related with convolutional neural networks (CNNs) have shown to be able to handle various problems in computer vision area. However, there is no clear success on ssEM image registration problem using CNNs. *ssEMnet* is constructed with two parts. The first part is a spatial transformer module which supports differentiable transformation of images in deep neural network. A convolutional autoencoder (CAE) which encodes dense features follows. The CAE is trained by unsupervised fashion and its features give wide receptive field information to align the source and target images. This method is compared with two other major ssEM image registration methods and increases accuracy and robustness, although it has less number of user parameters.

Contents

1. Introduction.....	1
1.1. Problem definition	1
1.2. Background.....	1
1.3. Motivation	4
1.4. Contribution.....	4
1.5. Goal	5
2. Related Work.....	6
2.1. Convolutional Neural Networks	6
2.2. Image Registration.....	8
2.3. Spatial Transformer Networks.....	9
3. Method	11
3.1. Training a Convolutional Autoencoder	11
3.2. Deformable Image Registration using a Spatial Transformer Network	13
4. Results	16
4.1. Implementation & Experiment Details.....	16
4.2. 3D Volume Reconstruction	17
4.3. Recovery of Artificially Deformed Dataset.....	18
4.4. Robust Alignment of Data with Fold Artifacts.....	20
5. Conclusion and Future Work	22

List of Figures

1	An example of EM image registration. (a) A fixed image, (b) a moving image, and (c) a warped image of (b) with grid visualization.....	1
2	The concept overview of image registration on connectome dataset. After imaging, each EM slice is not matched to its neighbors, so an image registration procedure is required... 3	3
3	Comparison between the pixel intensity based and the CAE feature based registration with backpropagation. (a) the fixed image, (b) the moving image, (c) the heat map of NCC of the pixel intensity-based result (NCC: 0.167), and (d) the heat map of NCC of the CAE-based result (NCC: 0.28) in red box region.	12
4	The model overview of the convolutional autoencoder consisting encoder 6 layers and decoder 6 layers.	15
5	The overview of the proposed method. The right network is the encoder of pretrained convolution autoencoder (CAE). The alignment is processed by backpropagation with loss of CAE features.	15
6	Drosophila melanogaster TEM dataset. Left: Pre-aligned result. Right: After registration using our method.	17
7	Vertical view of the alignment result of the randomly deformed CREMI dataset. (a) bUnwarpJ, (b) elastic stack alignment, and (c) our method. Each neuron is assigned a unique color.	19
8	An example slice of mouse lateral geniculate nucleus dataset with fold artifact.	20
9	Visual comparison of mouse ssEM image registration results. (a) before alignment, (b) bUnwarpJ, (c) elastic stack alignment, and (d) our method. The red box is the region near the folds (shown as black spots).....	21

List of Tables

- 1 The graph of the weighted Dice coefficient of each slice and the averaged values through slices for each method. Y-axis represents weighted Dice coefficient values. X-axis notices the slice number from the 1 to 31 (Total 31 slices are aligned). 19
- 2 Normalized cross correlation values (NCC) of the inner region in each aligned result. .. 21

1 Introduction

1.1 Problem Definition

Image registration has long been studied and the aim of this problem is to find the optimal transformation between two source and target images. [16] Image registration problems can be formulated as finding a transformation T minimizing following objective function:

$$\min_T M(T(I_s), I_t) + R(T) \tag{1}$$

I_s and I_t are the source and target images, respectively. M is a metric for comparing the matching results. R is a regularization term that restricts transformation T to unnatural or trivial solutions.

(1) is a problem of finding T for defined M and R , but the image registration problem is not simply optimization, because the optimization of solving the problem greatly changes according to M and R . If M or R is too simple, the optimization of (1) is also too easy and the resulting T is likely to be trivial. Therefore, depending on the given source and target images, a different M and a different R should be given. Various previous image registration methods based on (1) have been proposed. However, the implementation and modeling of these methods is not generalized. This means that different modeling is needed when different images and data are given. Therefore, tools that are used practically are trying to perform image registration in arbitrary data by using various user parameters according to dataset. It is up to the user to find the parameters one by one. In this paper, I propose a framework for image registration in a more general way to solve these problems. I have verified the validity of this framework through implementation and experiment of unsupervised method.

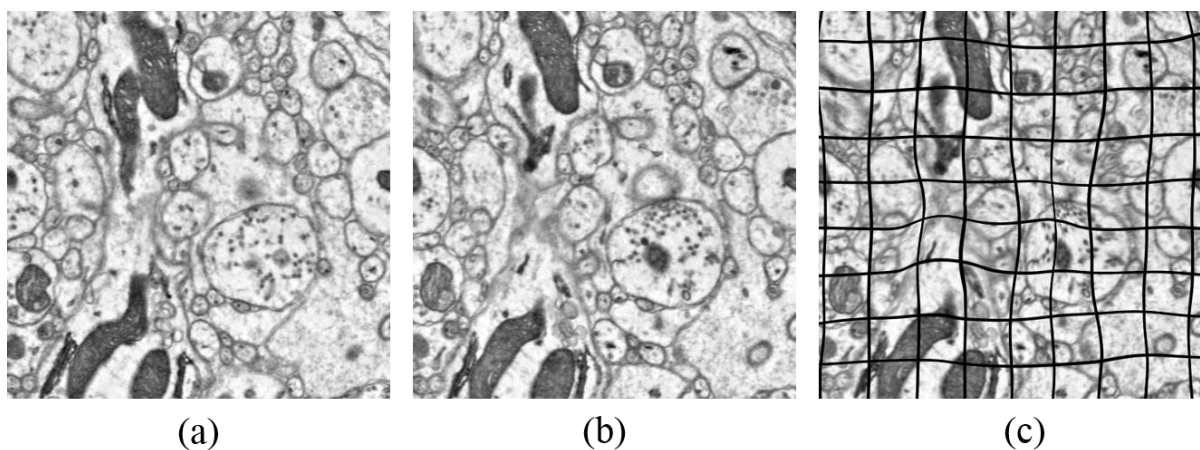


Fig. 1. An example of EM image registration. (a) A fixed image, (b) a moving image, and (c) a warped image of (b) with grid visualization.

1.2 Background

Connectomics research focuses to construct brain connectivity that represents how neurons are related with each other. Reconstructing 3D volume of brain tissue is challenging because the high packing density of neuronal-circuit. [7] The physical approaches of finding connection of the neural structure rely on slicing brain tissue and scanning these sections by electron microscopy (EM). For example, ATUM [6], the automated tape-collecting ultramicrotome generates serial-section EM (ssEM) images with 30nm thick and 5nm in-place resolution. However, the acquired dataset has slice-wise transformation including scale, rotation, translation and non-linear deformation because of the heat and pressure applied on the tissue. Therefore, these distorted ssEM sections should be stitched and aligned first to reconstruct the 3D volume as close as possible to the original specimen and to process next step of analysis on it.

Although Image registration problem has been studied with both real and medical images in decades, ssEM image registration has different challenging issues with traditional registration problem handling the real world images. Large deformation and non-predictable artifacts such as dusts and folds easily occurred on the acquiring process. An open-source toolset for ssEM data processing called TrakEM2 [4] supports several ssEM image registration methods including *bUnwarpJ* [2] and *Elastic stack alignment* [3] which are commonly used for ssEM image alignment.

bUnwarpJ is a B-spline based elastic registration with consistency constraint. This method registers two source and target images by solving minimization problems with three terms of similarity term, soft landmark term and regularization term. However, divergence and curl parameters in the regularization term are not intuitive for users to control regularization. In addition, the regularization more depends on quality setting (e.g. resolution of control vector grid), because its main similarity energy term is based on pixel-wise mean-square distance which is vulnerable to outlying features caused by artifacts from imaging process or anisotropic characteristic of ssEM datasets.

Elastic stack alignment is a more practical tool to align ssEM dataset. It has two alignment steps: A global alignment based on SIFT [26] and spring mesh alignment with block matching features. By matching SIFT descriptor which is rotate and scale invariant, given pairs of two images are approximately aligned. Based on these matching, each points of triangular grids in source image are matched to neighbor target images by block matching. Virtual springs are placed in these inter-image mappings and intra-image triangular grids, which control the distance through relaxation and deform images.

Both tools require users to find parameters on hand. The process of finding the parameters that match the characteristics of certain data depends entirely on the users' intuition and on their several tedious attempts. Elastic stack alignment finds the matching points of the two images by brute force search of matching block. The users should check whether each block size is suitable for the alignment or not. Also, Gaussian blur should be applied according to the texture characteristics. A testing tool for block matching is provided, but this test tool just helps users to check block size repeatedly rather than finding it. Even worse, no tool is not supported to find the parameters for spring mesh relaxation. Therefore, users should try all registration step to ensure the spring mesh. Because of these limitations of existing tools, data-driven image registration algorithm with less number of parameters is demanded for large scale ssEM datasets.

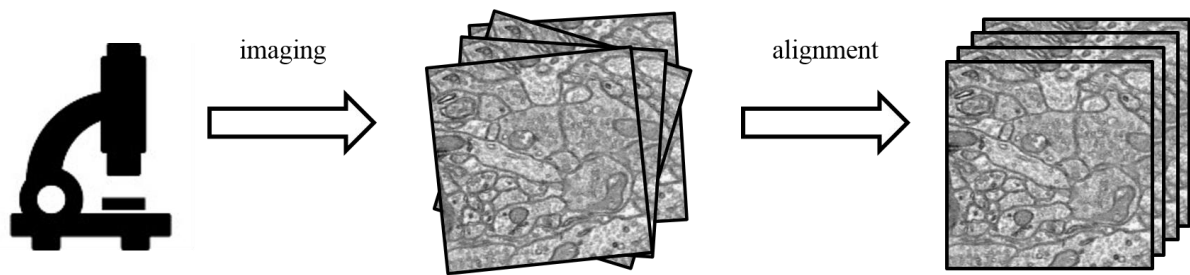


Fig.2. The concept overview of image registration on connectome dataset. After imaging, each EM slice is not matched to its neighbors, so an image registration procedure is required.

1.3 Motivation

Convolutional Neural Networks (CNNs) addresses many of the problems in the computer vision area with superior performance. CNN looks for features and functions that are difficult to define directly through the dataset. Therefore, researchers do not have to model hand-crafted features. (e.g. [14]) Image registration is very important for modeling metrics and regularization term according to data. As mentioned earlier, for practical tools, many user parameters are needed to solve this problem. In order to deal with complex data such as ssEM, users have to search directly through many attempts. It is necessary to find a feature encoder that knows the characteristics of each data through an automatic method rather than a user for a large dataset. In recent research, abstract objects such as artistic style have also been successfully transferred through pre-trained CNN. [5] Here, VGG-19 [12], a pre-trained CNN, is used as a feature descriptor for the loss function. In other words, the deep neural network itself becomes a component for a metric. This approach can also be applied to image registration. If there is a learned network the target dataset, there will be no need for the user parameter for that dataset, and it can be applied to complicate data like ssEM.

CNN is capable of feature encoding, but it does not represent the transformation of the image. Therefore, CNN alone cannot be an image registration model. In addition, the problem with the CNN feature is that the convolution operator is not rotational-invariant and scale-invariant in nature. This problem has been addressed in [8]. [8] suggests localization through scale, translation, affine transformation and thin plate spline (TPS) deformation using a differentiable transformation module called spatial transformer. In this way, robust features are learned from the given image to solve the fine-grained classification. A spatial transformer can also be applied to image registration. With this and the feature network, M and T in (1) can be constructed based on deep learning.

1.4 Contribution

There are three major contributions to this study. First, in this study, a general deep learning based image registration framework is presented. In this framework, new networks, transformations and objective functions can be used to construct new image registration models. Here, deep neural networks become feature descriptors, and thus metrics for matching. As mentioned in [8], various types of transformations can be implemented with ST module. Since these are differentiable, backpropagation-based optimizers used in existing deep learning can be used.

Second, I confirmed that convolutional Autoencoder (CAE) can act as a feature network for image registration. CAE is one of the simplest CNNs that can learn from a dataset without supervision [22]. [5] used a pre-trained network through a large dataset, but there is no labeled dataset for ssEM

datasets in my knowledge. Therefore, we want to verify that proposed image registration can work even if there is no powerful pre-trained network for arbitrary data. In this study, unsupervised CAE was shown to perform as a sufficient metric for matching.

Finally, a new loss-drop technique has been proposed that can robustly match ssEM's various artifacts. The unpredictable features appearing in the ssEM dataset preclude correct matching. To avoid falling into local optimum due to these features, loss-drop drops large feature errors. This enables matching even in severely damaged data which could not be matched by the conventional method, and it can be applied in neural network architecture to improve image registration performance.

1.5 Goal

The goal of this study is to propose a novel method so-called ssEMnet, which is a general image registration framework with pre-trained deep neural networks. I explore various studies related with CNNs, including the neural style transfer algorithm and STNs. In addition, several existing image registrations are introduced to compare the proposed method. In order to apply it directly to ssEM dataset without supervision, I proposed a CAE that can be directly used as a feature encoder without supervision, and proposed a loss-drop technique for robust image registration in non-matching artifacts. Experiments have shown that CAE with loss-drop provides a proper metric to align a damaged ssEM dataset.

2 Related Work

2.1 Convolutional Neural Networks

Convolutional Neural networks (CNNs) are deep learning architectures based on the characteristics of visual reception. This is suitable for learning image datasets because it uses fewer parameters than the previously used deep fully connected network [27] to handle a number of pixels of input images. More parameters require more time and data to learn, which can be a performance limitation. The basic CNN layer follows Eq. (2) [29].

$$z_{i,j,k}^l = w_k^{lT} x_{i,j}^l + b_k^l \quad (2)$$

Where l denotes a specific layer, and i and j denote spatial coordinates. $x_{i,j}$ is the input patch centered on i, j , and w_k and b_k are the bias of the k th filter respectively. As shown in (2), w is independent of input size, so CNNs can be trained with fewer parameters.

In [28], CNN has been used to analyze hand-written digits, and various networks have been proposed to solve image classification problem. AlexNet [30], which achieved the best records on ILSVRC-2012 large scale image classification challenge, applies ReLU non-linearity activation and layer response normalization for the first time. GoogleLeNet [31] proposed the Inception module to improve performance by applying various types of convolutional filters to one layer and achieved state-of-the-art in ILSVRC-2014. VGG-19 [12] improved performance by learning a deeper network of 19 layers, and ResNet [32] won the ILSVRC-2015 challenge by reducing the gradient vanishing problem of deep networks by introducing a residual module.

Besides image classification, CNNs have outperformed many areas like semantic segmentation [17] and super-resolution [18]. Unlike hand-crafted features, features from CNNs are directly derived from the data to solve the problem. For image registration, various hand-crafted features such as pixel intensity-based features (e.g. block matching) or SIFT were used, recent studies have shown that these existing descriptors can be replaced with learning-based descriptors. ([14], [15]).

[23] is an interesting example for understanding the learning of Deep learning. Normally, we put input into CNN, but in this study the authors find information inside CNN through backpropagation. An example of solving a specific problem through inverting a neural network is to solve the optical flow through inversion of the frame interpolation network [24]. Video itself is made up of successive frames, so you can use it to train a frame interpolation network. After the frame interpolation network

is trained, an interpolated frame is obtained from two frames to compute the optical flow. Gradients from each pixel in the interpolated frame to the given two frames are computed as a chain rule. Since the pixel giving the largest absolute gradient value has the greatest influence on interpolation, we can find the matching of two frames.

Gatys et al. [5] performed style transfer using pre-trained VGG-19. VGG-19 network learned from ImageNet [13], a large scale dataset, serves as a good descriptor for extracting various features of an image [14]. The authors made two assumptions in this paper: 1) if the features of the two images coming through VGG-19 are similar, they should be morphologically similar, and 2) image style is the correlation of filter responses in one layer of CNN. Based on their assumptions, they defined the content loss (3) and style loss (4) as follows, and performed style transfer by minimizing the total loss function (5).

$$L_{content}(I_c, I_g, l) = \frac{1}{2} \sum_{i,j} \left(F_{I_c}^l(i, j) - F_{I_g}^l(i, j) \right)^2 \quad (3)$$

$$L_{style}(I_s, I_g) = \sum_{l=0}^L \frac{w_l}{4N_l^2 M_l^2} \sum_{i,j} \left(\sum_k F_{I_s}^l(i, k) F_{I_s}^l(j, k) - \sum_k F_{I_g}^l(i, k) F_{I_g}^l(j, k) \right)^2 \quad (4)$$

$$L_{total}(I_c, I_s, I_g) = \alpha L_{content}(I_c, I_g) + \beta L_{style}(I_s, I_g) \quad (5)$$

I_c , I_s , I_g are content image, style image, generated image respectively. $F_l^l(i, j)$ is l th layer feature of input image I where the pixel positions are denoted as i and j . In (4), the correlation of filter responses is computed by Gram matrix. In this study, the style loss is removed and only the content loss is applied to align a pair of images.

Along with the recent development of CNN, CAE has appeared in image registration field. Wu et al. [10] acquires multi-dimensional features through a convolutional stacked autoencoder (SAE) to reduce the large 3-D image patch. They proposed multi-dimensional features through max-pooling and stacked autoencoder for MRI volume and matching them through multichannel version of Demons algorithm [20] and feature-based HAMMER algorithm [21]. However, their convolutional stacked autoencoder doesn't have any convolution operators, but spatially reduced max-pooling of patch-wise features trained with sparse SAE. This allows for a wider receptive field than a simple patch-wise SAE feature, but max-pooling is a hand-crafted reduction. The convolutional SAE feature is translation invariant, but there is no guarantee that it is also rotation and scale invariant.

2.2 Image Registration

Image registration is the task of aligning two matching images into the same coordinate space. As described in (1), the main elements in image registration are metric, transformation and regularization. Image registration methods are largely categorized as the difference between these three factors. Depending on the metrics, it can be broadly divided into feature-based and intensity-based methods.

Feature-based methods extract feature descriptors from source and target images and then transform the source image to reduce the distance of the sparse feature descriptors. An example of such a feature is Scale Invariant Feature Transform (SIFT) [26]. The SIFT algorithm finds scale, noise, rotation and illumination-invariant feature points. This means that it can find feature points robustly even if the four conditions are changed. However, the matching through the feature points such as SIFT has a limit when the deformation is strong. Globally, images can be matched, but matching in fine-resolution is difficult. [3] And, for image registration, the user needs to filter to the appropriate threshold according to matching of matching SIFT features. If there is an erroneous feature matching, the performance of the matching can be greatly reduced. The proposed method registers a pair of images through encoded features rather than pixel-intensities, but is close to intensity-based methods. In this respect, it is suitable to match finer local features rather than global alignment. This can be complemented by using SIFT for global alignment as in previous [5] and matching the detail with the proposed method.

On the other hand, intensity-based image registration uses all the pixels of a given image. For example, UnwarpJ [33] using pixel intensity to match two source and target images with vector-spline regularization. In [33], the authors find solutions by minimizing the energy terms including data term, landmark term and two regularization terms. The landmark acts as a kind of human-labeled feature point. The reason for the landmark term is to prevent potential mismatches. This is because pixel intensity-based image registration is limited to local minima. A multi-channel intensity-based approach has emerged to compensate for the limitations of pixel intensity-based registration ([20], [21]). ssEMnet presented in this study matches through multi-channel vectors. However, the multi-channel vectors are encoded based on deep learning. It differs from the single pixel location presented in [20], because it has a wider receptive field. In addition, it encodes the significant pixels through a convolutional filter compared to [21] which just uses multi-resolution pixels.

2.3 Spatial Transformer Networks

CNN is a powerful tool for solving problems through data-driven, but the disadvantage is that the convolution operation on which it is based is not invariant to rotation or scale. There have been many attempts to solve these problems. (Reference addition)

Spatial transformer (ST) is a CNN module designed to solve this problem. It consists of localization network, grid generator, and sampler. Localization network finds suitable transformation parameters from a given input feature map. The number of parameters coming out of the output depends on the transformation. If the given transformation is an affine transformation, then the output parameter must be 6-dimensional.

Through the parameters obtained from the localization network, the grid generator creates a sampling grid. The sampling grid indicates the mapping of the input and output feature maps of ST. For affine transformation, the pointwise sampling is expressed as:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (6)$$

Where (x_i^t, y_i^t) is the target coordinate, (x_i^s, y_i^s) is the source coordinate, and both coordinates are represented by relative coordinates with values from -1 to 1. More practically, ST also can perform attention by following transformation in (7) with 3 parameters.

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = \begin{bmatrix} \theta_1 & 0 & \theta_2 \\ 0 & \theta_1 & \theta_3 \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (7)$$

The sampler then uses the sampling grid to sample the input feature map. The actual reference values are contained in the grid feature map, so they are referenced using bilinear sampling. For the input feature map U , the output feature map V is expressed as (8) where c notates channel.

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \quad (8)$$

The key point of ST is that all parts are made up of differentiable elements to enable backpropagation. For the output feature map V , we can obtain the following partial derivatives for the input feature map U and source coordinate (x_i^s, y_i^s) .

$$\frac{\partial V_i^c}{\partial U_{mn}^c} = \sum_n^H \sum_m^W \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \quad (9)$$

$$\frac{\partial V_i^c}{\partial x_i^s} = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |y_i^s - n|) \begin{cases} 0 & \text{if } |x_i^s - m| \geq 1 \\ 1 & \text{if } m \geq x_i^s \\ -1 & \text{if } m < x_i^s \end{cases} \quad (10)$$

The partial derivative of (9), (10) can be used to learn the localization network, which can complement the lack of the rotational and scale invariant properties of the convolution operator. In this study, no localisation network is used in the proposed method. ssEMnet finds an optimal deformation by backpropagation rather than training on certain dataset, because basically, there is no ground-truth for ssEM image registration.

3 Method

3.1 Training a Convolutional Autoencoder

VGG-19 [12] is trained based on the real world image dataset, ImageNet. Style transfer with CNN also used VGG16. There is no such large scale of classification dataset on EM image, so a convolutional autoencoder (CAE) was used as a CNN to give a feature metric. The reason for using CAE as a feature network is that it can be learned in an unsupervised fashion without labels and has already been used as a feature in previous studies ([10], [22]) to produce meaningful results. The CAE used in this study consists of two components. The first is the convolutional encoder composed with convolutional layers followed by ReLU activation. The convolutional encoder reduces spatial resolution through strides without pooling. This is because the average pooling and max pooling are inadequate for feature encoding because of the loss of information. Instead, the spatial resolution was reduced by applying a convolution with a stride of 2 in height and width to some layers in the encoder network. By reducing spatial resolution, a network with a wider receptive field was constructed. The second part is a transposed convolutional decoder consisting of transposed convolutional layers with ReLU activations. Transposed convolutional layers are configured such that the kernel, strides, and channel dimension settings are mirrored with the encoder layers. Since the network consists entirely of convolution and transposed convolution, it can be applied to any data set regardless of spatial resolution. It can be designed formally as following equations.

$$h = f_{\theta}(I) \quad (11)$$

$$\tilde{I} = g_{\phi}(h) \quad (12)$$

$$L_{\theta, \phi} = \sum_{i=1}^N \|I_i - \tilde{I}_i\|_2^2 + \lambda \left(\sum_k \|\theta_k\|_2^2 + \sum_k \|\phi_k\|_2^2 \right) \quad (13)$$

Encoders and decoders are denoted by f_{θ} and g_{ϕ} , respectively, θ and ϕ are parameters of each network. h is applied as feature vector in the next step. The objection function (13) consists of reconstruction term and L2 weight regularization term used to reduce overfitting. The reconstruction term serves to contain as much information of input image as possible in the bottleneck feature vector h . The weight regularization term limits the absolute value of the encoded feature h to not be too large. Fig. 3 shows the comparison between pixel intensity-based registration and CAE feature-based registration. The normalized cross correlation (NCC) value measured on (c) is larger than (d). The solution obtained from pixel-wise loss seems to easily fall to a local optimum.

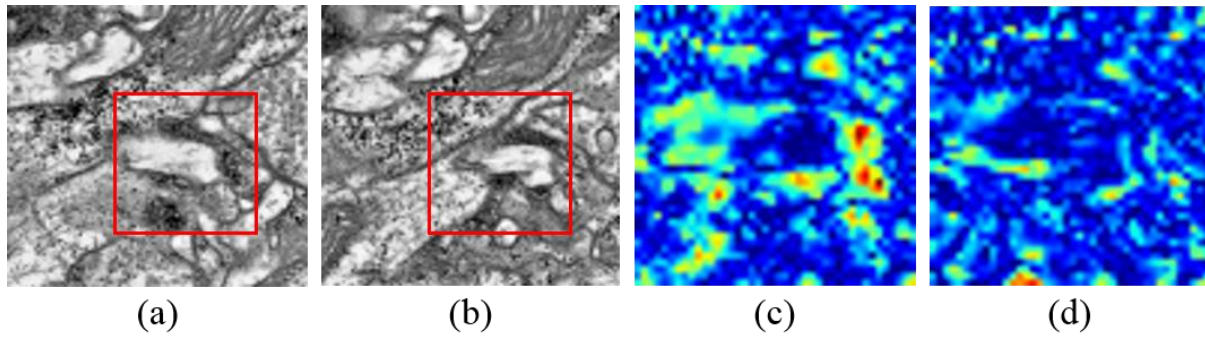


Fig. 3. Comparison between the pixel intensity based and the CAE feature based registration with backpropagation. (a) the fixed image, (b) the moving image, (c) the heat map of NCC of the pixel intensity-based result (NCC: 0.167), and (d) the heat map of NCC of the CAE-based result (NCC: 0.28) in red box region.

2.2 Deformable Image Registration using a Spatial Transformer

After learning the CAE, the encoder of the CAE is taken out and attached to a ST. This applies CNN to the image registration specific problem shown in (1). ST plays the role of transformation T , and the pre-trained encoder becomes the feature descriptor for metric M . This network architecture is designed to find proper deformation for image registration using ST as a differentiable transformer. The objective function for registration error is expressed by (14). Here, I_0 is the source image to be transformed, and I_1 is the target image. The deformation warp is generated by spanning the coarse vector map v to the control points. Here, ST provides a gradient for updating v to backpropagation. In [8], the authors applied a thin plate spline (TPS) [11] for smooth deformation with only a few parameters for their experiments. TPS, however, tended to be more difficult to match EM images if folds occur in the imaging process. Therefore, ST using bilinear interpolation is applied in the experiment.

$$L_v(I_0, I_1) = \|f_\theta(I_1) - f_\theta(T_v(I_0))\|_2^2 + \alpha \|v\|_2^2 + \beta \|\nabla v_x\|_2^2 + \gamma \|\nabla v_y\|_2^2 \quad (14)$$

Eq. (14) consists of a total of four terms. The first term is a contextual different measurement of two images through the learned CAE features. If both images are well matched, the encoded features of both images should appear similar. This term is designed as a loss that reduces the difference between the encoded features of the warped image and the target image with an L2-norm. The remaining terms are the regularization term for the vector map v . The second term is to prevent images from drifting when dealing with multiple images. The third and fourth terms are smoothness loss terms to produce the deformation as smooth as possible. α , β , γ are the weights of each regularization term. Since all ssEMnet components including ST are differentiable, v is optimized by backpropagation via chain rule. At this time, the encoder of CAE is fixed and only v is updated. The ADAM optimizer [9] is used to perform stable gradient descent in this study. Fig. 4 shows the image registration using the backpropagation with CAE features

Optimizing the objective function (14) only aligns two adjacent images. During the imaging, various damage and artifacts can occur on EM sections. Matching a pair of images can cause accumulation of errors through multiple sections of alignment. For more robust registration, (14) is extended to make several sections of the neighbor as target images.

$$L_v(I_0, \dots, I_n) = \sum_{i=1}^N w_i \left\| M(T_v) \left(f_\theta(I_i) - f_\theta(T_v(I_0)) \right) \right\|_2^2 + \alpha \|v\|_2^2 + \beta \|\nabla v_x\|_2^2 + \gamma \|\nabla v_y\|_2^2 \quad (15)$$

As (14), the moving image is I_0 and the remaining target images are I_1 to I_n . Since each target image is a neighbor slice that is a different distance from the source image, we can assign a different weight to each target image. These weights are each represented as w_i . Eq. (15), a modified version of (14) measures the registration error from multiple images and is robust from finding false matches due to outlying artifacts or large deformation.

An image deformed by the ST module can refer to a pixel outside the image boundary by a regularization term. In the implementation of ssEMnet, the outer part of all images was kept at pixels of intensity 0. Instead, an empty space mask representing the empty area outside the image was created after image deformation. This mask is applied with bilinear interpolation to have same spatial resolution with CAE features. (e.g. $M(T_v)$ in (15)). This is to ensure that CAE features with unnecessary information do not affect matching.

A new technique called *loss-drop* is introduced for more robust matching from artifacts such as dusts and folds. Since the artifacts that occur during the imaging process create large feature errors, reducing the feature error of these parts from the beginning of registration makes it easy to fall into local minima. Loss-drop technique drops the top k% feature error to zero. In the experiments, k was initially set to 50 and this k was reduced by half for every iteration. This prevented v from falling into the wrong solution early in the registration process and generated a smoother registration result. Since multiple EM sections should be aligned at the same time, sliding-window method is applied to minimize the objective function iteratively through the entire section sequences with out-of-core fashion.

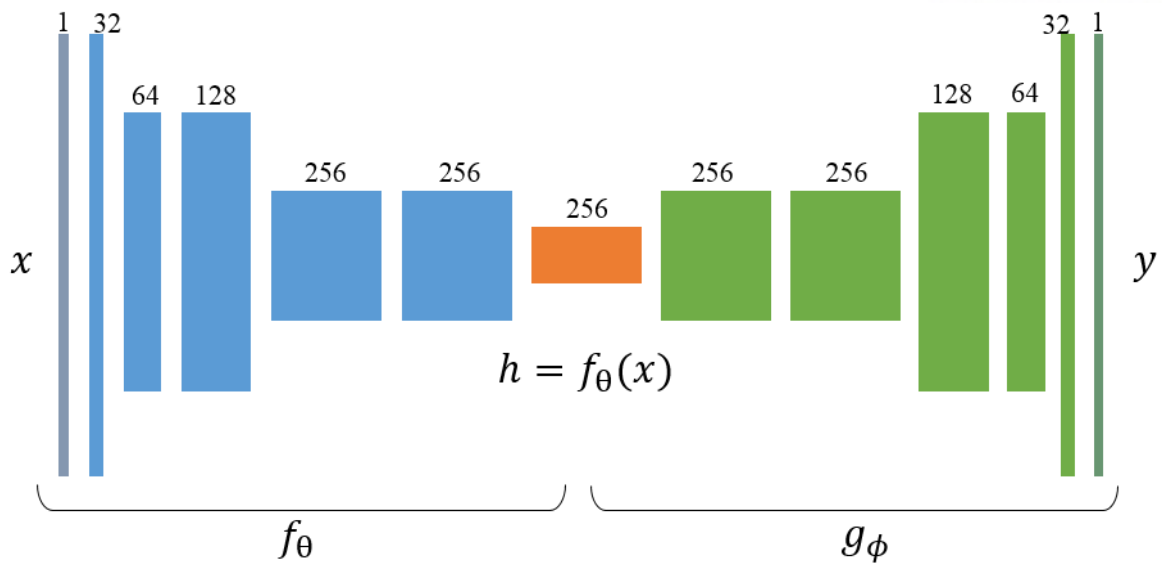


Fig. 4. The model overview of the convolutional autoencoder consisting encoder 6 layers and decoder 6 layers.

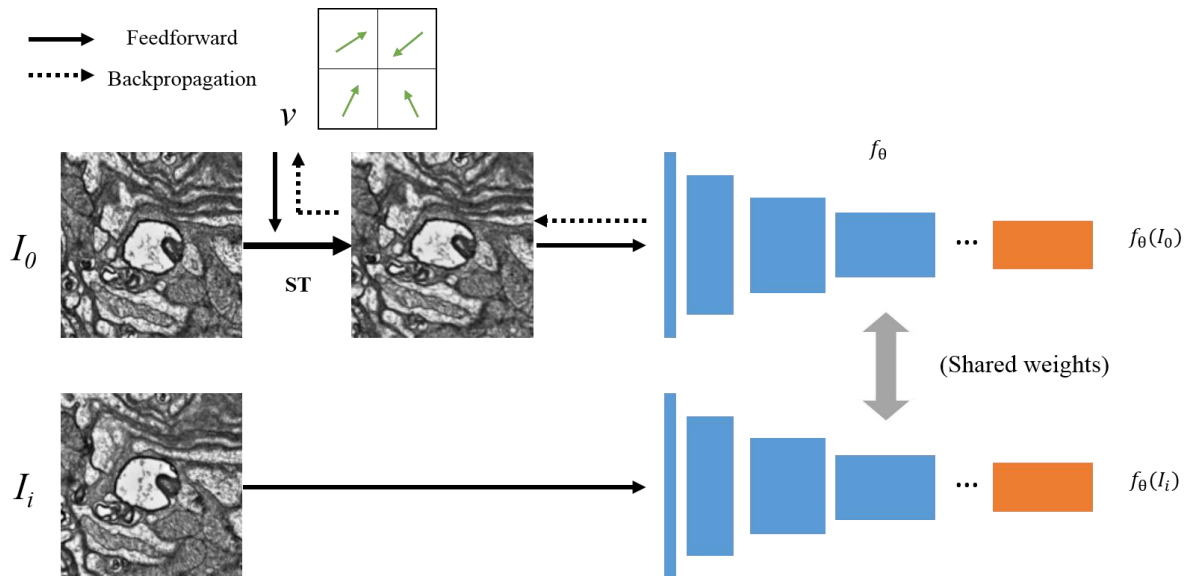


Fig. 5. The overview of the proposed method. The right network is the encoder of pretrained convolution autoencoder (CAE). The alignment is processed by backpropagation with loss of CAE features.

4 Results

4.1 Implementation & Experiment Details

The proposed method is developed using TensorFlow [1]. All experiments were done using a GPU workstation equipped with an NVIDIA Titan X GPU. Three experiments on three datasets are performed to test the method. Each dataset is transmission EM (TEM) images of Drosophila brain, human-labeled TEM images of another Drosophila brain provided by CREMI challenge (<https://cremi.org/>). Mouse brain scanning EM (SEM) images are corrupted by fold artifacts that creates a strong deformation and loss of image parts. Drosophila images are acquired separately on the different imaging techniques.

The characteristics of the CAE features, the receptive field and the computational burden of backpropagation vary greatly depending on the structure of the CAE. Therefore, two different autoencoders were implemented according to the experiment. One is a deeper network using a 3×3 convolutional filters (Fig. 4). This was used to match Drosophila TEM datasets. The other is a shallower network consisting of encoder 3 layers and decoder 3 layers. The network was implemented with 7×7 filters instead of 3×3 convolutional filters. This shallow network was used for experiments on the mouse SEM dataset. bUnwarpJ and elastic stack alignment were tested and applied the best parameters to give the best results.

4.2 3D Volume Reconstruction

The original dataset of an adult female *Drosophila* brain cut has 4 million plane images and occupies a capacity close to 50TB. The dataset was firstly aligned with AlignTK (<http://mmbios.org/aligntk-home>). AlignTK is a batch-oriented alignment toolkit for 2D or 3D dataset. It finds corresponding positional matching, so called ‘map’ between adjacent slices. The grid points of each source image is matched to neighbor target images based the pixel intensity of images. After finding mapping, it applies relaxation method with virtual springs as elastic stack alignment does. This tool supports parallel execution with Message Passing Interface (MPI) that is able to process large amount of images with parallel fashion efficiently. Though AlignTK is very powerful tool to align such large scale dataset of EM images, it still requires persistent efforts of users and large amount of CPU resources like a huge computing cluster. The alignment via AlignTK allows researchers to complete manual neural annotation. However, additional fine-grained re-alignment is required for automated segmentation of membranes. Therefore, a small volume of size 512 x 512 x 47 is cropped for re-alignment. Fig. 6 shows the result of ssEMnet. The left and right of Fig. 6 is cross-sectional view before and after the re-alignment. The red circles are annotated to compare the correction by the proposed method. The membranes of the sub-volume after re-alignment are more smoothly connected through sections than before.

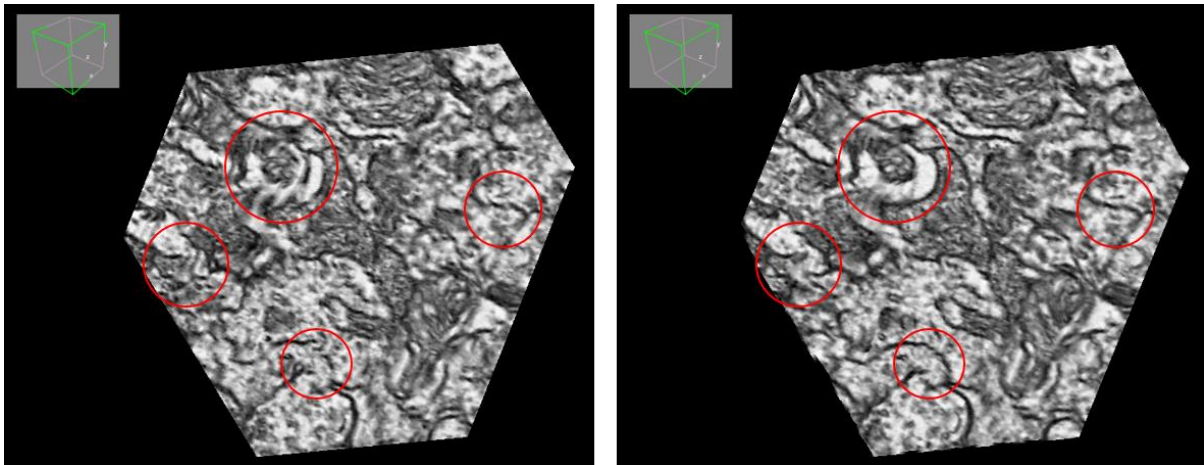


Fig. 6. *Drosophila melanogaster* TEM dataset. Left: Pre-aligned result. Right: After registration using our method.

4.3 Recovery of Artificially Deformed Dataset

Pixel intensity-based quantitative measurements are limited to making a good comparison in EM images where the two images have large variations and many features are not matched. Also, simple slice-by-slice measurements do not detect drifting of the stack. Thus, in this study, a new measurement based on the labeled dataset, the CREMI challenge dataset, was presented. First, raw and labeled images sampled at a size of $512 \times 512 \times 31$ from CREMI dataset were deformed by TPS parameterized with a random vector on random positions. The random positions were uniformly sampled within the image size and the random vector are sampled from the normal distribution of zero mean value. By doing this, the image was subjected to various smooth random deformations. The labeled image was deformed as well as the raw EM image. For qualitative comparison, the weighted Dice coefficients of the label images after registration and ones before the random deformation were obtained. Through these weighted Dice coefficients, it is qualitatively measurable how much the deformation in the imaging process is restored to the existing data by matching. The weighted Dice coefficients were obtained by assigning weights according to the size of cell cross section for each slice of the largest cell 50 labels in the sampled area, and averaged over all slices. Tab. 1 shows a graph for each slices weighted Dice coefficient and the mean of the weighted Dice coefficient of all slices.

This experiment is performed using bUnwarpJ, elastic stack alignment and the proposed method. Fig. 7 show orthogonal views of the results registered with each method. Fig. 7 (a) shows the result of applying the bUnwarpJ method to the data. Although slices in Fig. 7 (a) are continuously matched, the slices are drifting due to excessive deformation. On the other hand, elastic stack alignment results in less deformation but less matching. The vertical section through the proposed image registration method is smoother and more continuous than which of the other methods. The proposed method is also more robust to random deformation and less drifting.

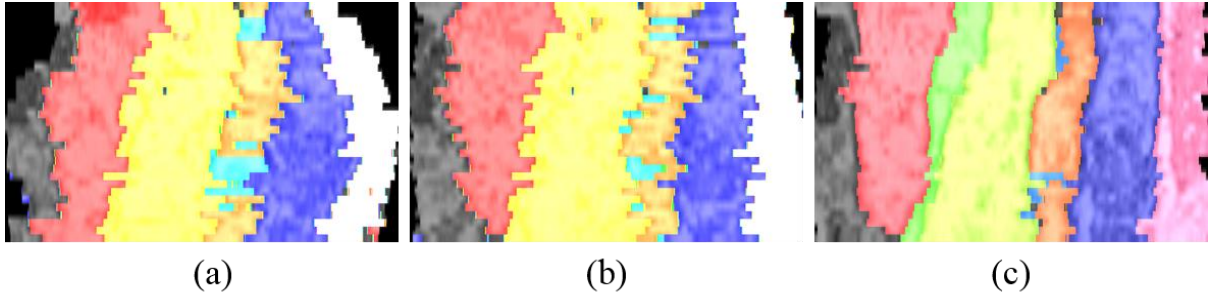
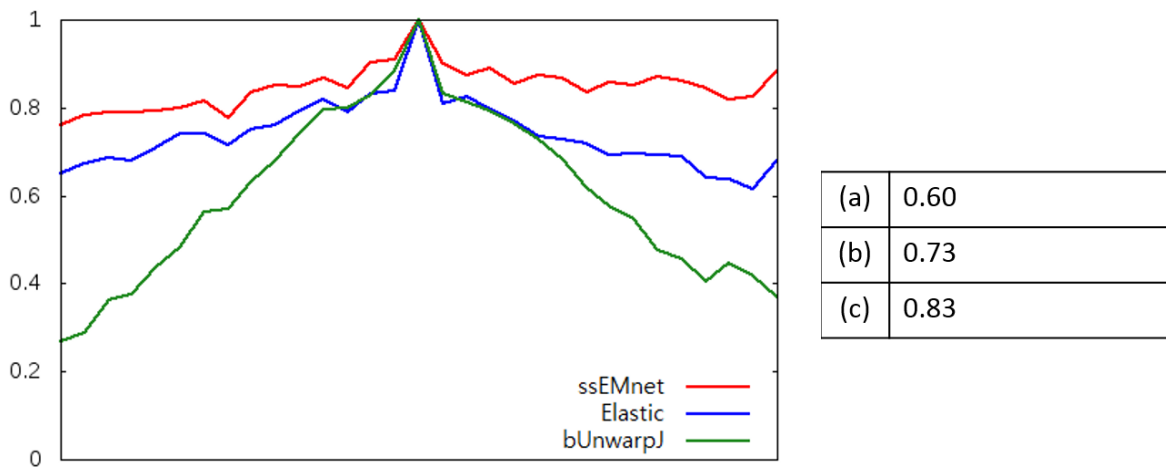


Fig. 7. Vertical view of the alignment result of the randomly deformed CREMI dataset. (a) bUnwarpJ, (b) elastic stack alignment, and (c) our method. Each neuron is assigned a unique color.



Tab. 1. The graph of the weighted Dice coefficient of each slice and the averaged values through slices for each method. Y-axis represents weighted Dice coefficient values. X-axis notices the slice number from the 1 to 31 (Total 31 slices are aligned).

4.4 Robust Alignment of SEM Data with Fold Artifacts

During imaging, a variety of damages can be imposed on the acquired EM image. One of them is fold artifact. Fold artifacts literally appear as wrinkled folds of paper in thin EM sections. These folds do not fit anywhere in adjacent sections. There is also a very strong deformation around this fold. Fig. 8 is an example of the EM section where this fold is applied. Pixel intensity-based deformable image registration methods have a dilemma for matching these folded EM images. To fit unfolded features, the folded area must be pulled out to widen, resulting in a greater matching error. Experiments were performed to verify that the CAE feature loss using the loss-drop technique robustly matches these folded data with a mouse lateral geniculate nucleus dataset [19]. A sub-volume cropped to 1520 x 2500 x 100 size was aligned. Fig. 9 is the vertical section view of the registration results. A fold artifact is a part of the vertical section view that looks dark. The red colored box was marked to compare whether the matching is correct near the folds. In other methods, there are many misaligned sections in the red box, but the proposed ssEMnet employing the loss-drop technique has smoother and more continuous results. Tab shows the average normalized cross correlation (NCC) values inside the resulted volumes. To exclude the out-of-boundary pixels, the NCC values were calculated using only the pixels in the inner region.

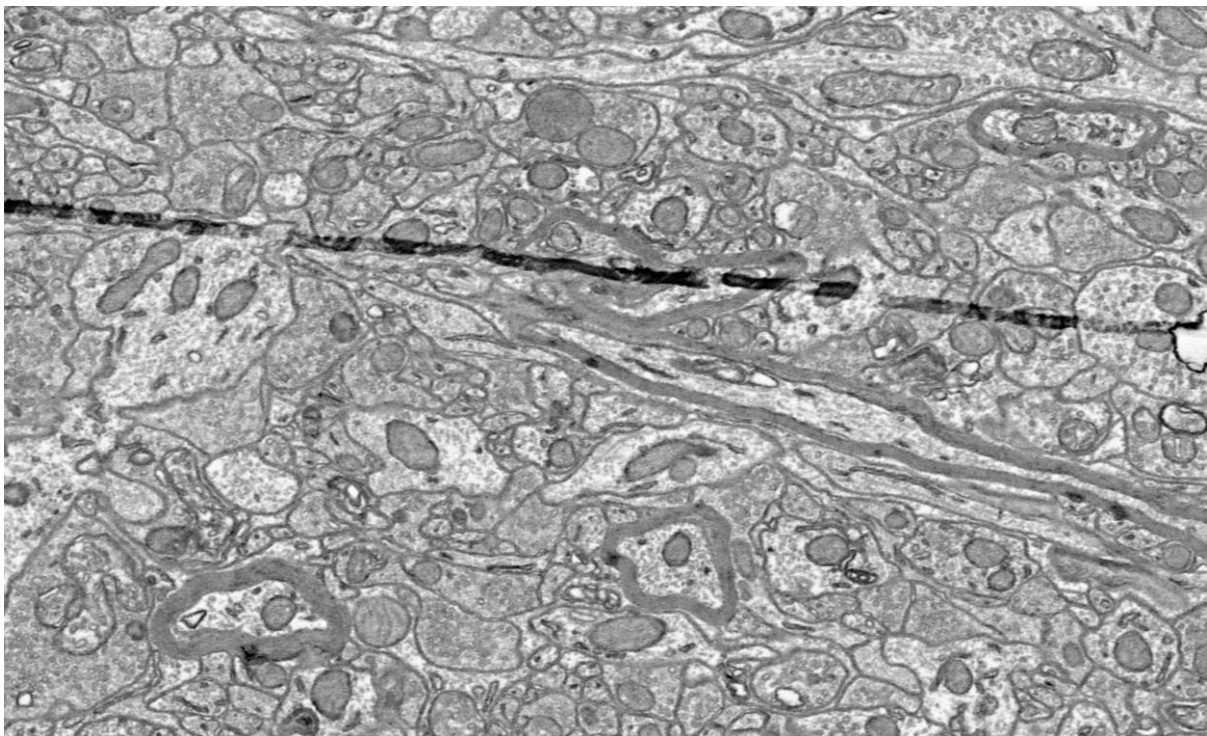


Fig. 8. An example slice of mouse lateral geniculate nucleus dataset with fold artifact.

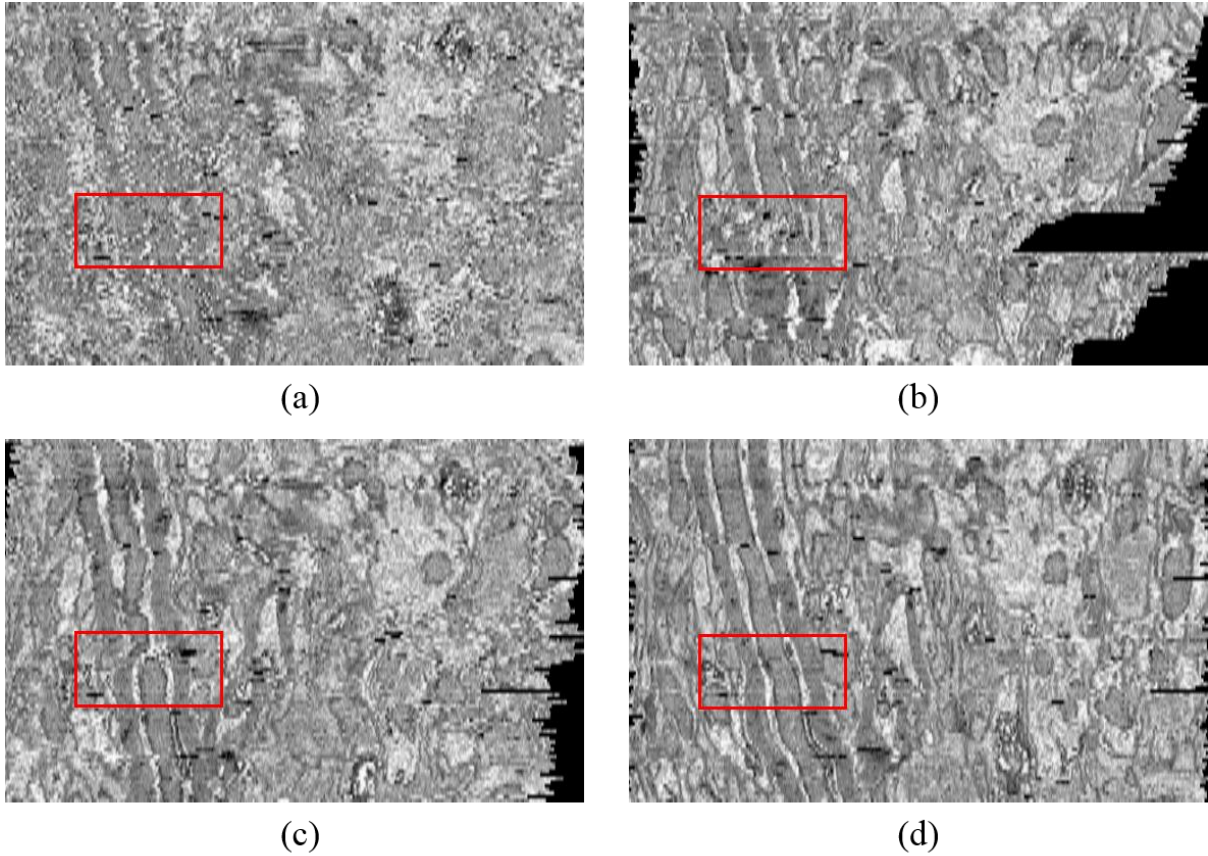


Fig. 9. Visual comparison of mouse ssEM image registration results. (a) before alignment, (b) bUnwarpJ, (c) elastic stack alignment, and (d) our method. The red box is the region near the folds (shown as black spots).

	(a)	(b)	(c)	(d)
NCC	0.1995	0.3562	0.2931	0.4305

Tab. 2. Normalized cross correlation values (NCC) of the inner region in each aligned result.

5 Conclusion and Future Work

CNN is powerful in many computer vision areas, but CNN's fundamental component, the convolution operator, is neither scale-invariant nor rotational invariant. Spatial transformer networks (STN) were proposed to solve this problem. In this work, a deformable image registration method using a deep learning feature using ST and the commonly used backpropagation of deep learning has been proposed. The proposed method has a remarkably small number of parameters for matching because of the characteristic of learning based approach. In addition, the registration results are more robust and better than other methods. In this work, z-axis alignment is performed through the sliding window method and the sparsity term for the vector map, but applying the same relaxation scheme as the spring mesh will align the stack of EM images faster.

The localization network of the ST module is missing in this work. This is because training the complex deformation of the EM image is difficult without supervision. [25] suggests how to learn deformable image registration through unsupervised end-to-end training, but does not handle regularization and appropriate metrics. Therefore, it can be a future work to train the feed forward network with the regularization on vector map and an appropriate loss function rather than the backpropagation based direct optimization of the vector map.

References

- [1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
- [2] Arganda-Carreras, I., Sorzano, C. O., Marabini, R., Carazo, J. M., Ortiz-de-Solorzano, C., & Kybic, J. (2006, May). Consistent and elastic registration of histological sections using vector-spline regularization. In International Workshop on Computer Vision Approaches to Medical Image Analysis (pp. 85-95). Springer, Berlin, Heidelberg.
- [3] Saalfeld, S., Fetter, R., Cardona, A., & Tomancak, P. (2012). Elastic volume reconstruction from series of ultra-thin microscopy sections. *Nature methods*, 9(7), 717-720.
- [4] Cardona, A., Saalfeld, S., Schindelin, J., Arganda-Carreras, I., Preibisch, S., Longair, M., ... & Douglas, R. J. (2012). TrakEM2 software for neural circuit reconstruction. *PloS one*, 7(6), e38011.
- [5] Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.
- [6] Hayworth, K. J., Morgan, J. L., Schalek, R., Berger, D. R., Hildebrand, D. G., & Lichtman, J. W. (2014). Imaging ATUM ultrathin section libraries with WaferMapper: a multi-scale approach to EM reconstruction of neural circuits. *Frontiers in neural circuits*, 8.
- [7] Helmstaedter, M. (2013). Cellular-resolution connectomics: challenges of dense neural circuit reconstruction. *Nature methods*, 10(6), 501-507.
- [8] Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. In *Advances in Neural Information Processing Systems* (pp. 2017-2025).
- [9] Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- [10] Wu, G., Kim, M., Wang, Q., Munsell, B. C., & Shen, D. (2016). Scalable high-performance image registration framework by unsupervised deep feature representations learning. *IEEE Transactions on Biomedical Engineering*, 63(7), 1505-1516.

- [11] Bookstein, F. L. (1991, July). Thin-plate splines and the atlas problem for biomedical images. In Biennial International Conference on Information Processing in Medical Imaging (pp. 326-342). Springer, Berlin, Heidelberg.
- [12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- [13] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). IEEE.
- [14] Moo Yi, K., Verdie, Y., Fua, P., & Lepetit, V. (2016). Learning to assign orientations to feature points. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 107-116).
- [15] Yi, K. M., Trulls, E., Lepetit, V., & Fua, P. (2016, October). Lift: Learned invariant feature transform. In European Conference on Computer Vision (pp. 467-483). Springer International Publishing.
- [16] Sotiras, A., Davatzikos, C., & Paragios, N. (2013). Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7), 1153-1190.
- [17] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431-3440).
- [18] Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2), 295-307.
- [19] Morgan, J. L., Berger, D. R., Wetzel, A. W., & Lichtman, J. W. (2016). The fuzzy logic of network connectivity in mouse visual thalamus. *Cell*, 165(1), 192-206.
- [20] Forsberg, D., Rathi, Y., Bouix, S., Wassermann, D., Knutsson, H., & Westin, C. F. (2011). Improving registration using multi-channel diffeomorphic demons combined with certainty maps. *Multimodal Brain Image Analysis*, 19-26.
- [21] Shen, D. (2007). Image registration by local histogram matching. *Pattern Recognition*, 40(4), 1161-1172.

- [22] Masci, J., Meier, U., Cireřan, D., & Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. *Artificial Neural Networks and Machine Learning–ICANN 2011*, 52-59.
- [23] Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5188-5196).
- [24] Long, G., Kneip, L., Alvarez, J. M., Li, H., Zhang, X., & Yu, Q. (2016, October). Learning image matching by simply watching video. In *European Conference on Computer Vision* (pp. 434-450). Springer International Publishing.
- [25] de Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M., & Išgum, I. (2017). End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network. *arXiv preprint arXiv:1704.06065*.
- [26] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91-110.
- [27] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems* (pp. 396-404).
- [28] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- [29] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Wang, G. (2015). Recent advances in convolutional neural networks. *arXiv preprint arXiv:1512.07108*.
- [30] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
- [31] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
- [32] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[33] Sorzano, C. O. S., Thévenaz, P., & Unser, M. (2005). Elastic registration of biological images using vector-spline regularization. *IEEE Transactions on Biomedical Engineering*, 52(4), 652-663.