Scribbles for Metric Learning in Histological Image Segmentation

Daisuke Harada[1], Ryoma Bise[1], Hiroki Tokunaga[1],
Wataru Ohyama[1], Sanae Oka[2], Toshihiko Fujimori[2], and Seiichi Uchida[1]

***Abstract*— Segmentation is a fundamental process in biomedical image analysis that enables various types of analysis. Segmenting organs in histological microscopy images is problematic because the boundaries between regions are ambiguous, the images have various appearances, and the amount of training data is limited. To address these difficulties, supervised learning methods (*e.g.*, convolutional neural networking (CNN)) are insufficient to predict regions accurately because they usually require a large amount of training data to learn the various appearances. In this paper, we propose a semi-automatic segmentation method that effectively uses scribble annotations for metric learning. Deep discriminative metric learning re-trains the representation of the feature space so that the distances between the samples with the same class labels are reduced, while those between ones with different class labels are enlarged. It makes pixel classification easy. Evaluation of the proposed method in a heart region segmentation task demonstrated that it performed better than three other methods.**

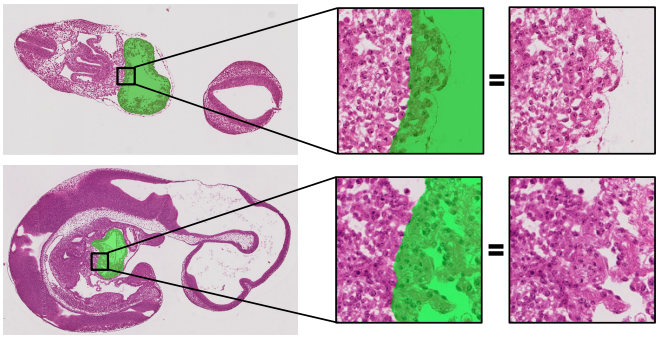

Fig. 1. Examples of sliced images. Left: Entire images for two examples, in which the slice angles differ. Appearance and color differ greatly. Middle: Enlarged images of boundary between heart region and other regions, where green shows heart regions. Right: Enlarged original images corresponding to middle images. Difference in texture along boundary is insignificant.

I. Introduction

Sliced histological imaging is a promising technology for investigating 3D structures during embryogenesis because it can produce images with sufficient resolution to enable observation of single cells on an organ. This enables more detailed analysis than with micro-CT and MRI imaging.

The aim of this study is to develop a method for segmenting the mouse heart region in 3D histological microscopy images. Analysis of the 3D structure of organs is performed by slicing a histological tissue and then capturing images of the slices by a microscopic scanner. To reconstruct the entire 3D region of an organ, the target organ region (*i.e.,* heart) is segmented in each slice, and then these segmentation results are aggregated to make the 3D structure.

Such segmentation has three main difficulties. First, the boundary between the heart and adjacent regions is ambiguous, as shown by the middle and right images in Fig. 1, in which the difference in texture is insignificant. It is very difficult for an untrained technician to segment such images. Second, the sliced images of mouse embryos have various appearances due to differences in the slice angle and staining, as shown by the images on the left in Fig. 1, which the shape and color greatly differ. Third, annotation for 3D segmentation is time-consuming and can only be done by trained biological experts. Therefore, the amount of training data for a machine learning approach is limited.

These difficulties limit the applicability of supervised learning algorithms. Although CNN [1] has been shown to outperform traditional computer vision techniques in various segmentation tasks, CNN is not well suited for the histological segmentation task since it requires a large amount of training data to learn the various types of appearances for accurate prediction.

Another approach to segmenting regions is to use additional rough annotation (scribbles) for the test data. For example, in graph-cut segmentation, scribbles indicating the heart region and other regions are added to the test data, and then an energy function is optimized by using the scribbles as seeds. The energy function consists of a data term and a smoothness term. How to design the data term for accurate region prediction is problematic due to the limitation.

Contribution: The main contribution of this work is proposing a segmentation method that effectively uses scribble annotations as part of the test data, not only for graph-cut segmentation but also for metric learning-based feature representation. The key feature of our approach is that the feature space is retrained to represent the discriminative features in the test data. This is important because the image appearance (feature) produced by CNN of the test data may differ slightly from that of the training data due to the various appearances and the small amount of training data. Metric learning is used to retrain the representation of the feature space by using a small amount of additional labeled data from the test data so that the distances between samples with the same class label are reduced, while those between samples with different class labels are enlarged. This facilitates pixel classification. We use the features produced by metric learning as the data term of the energy function in the 3D graph-cut segmentation, enabling it to perform highly accurate 3D segmentation. The proposed method was

[1]Department of Advanced Information Technology, Kyushu University, Fukuoka City, Japan bise@ait.kyushu-u.ac.jp
[2]Division of Embryology, National Institute for Basic Biology, Okazaki, Aichi, Japan

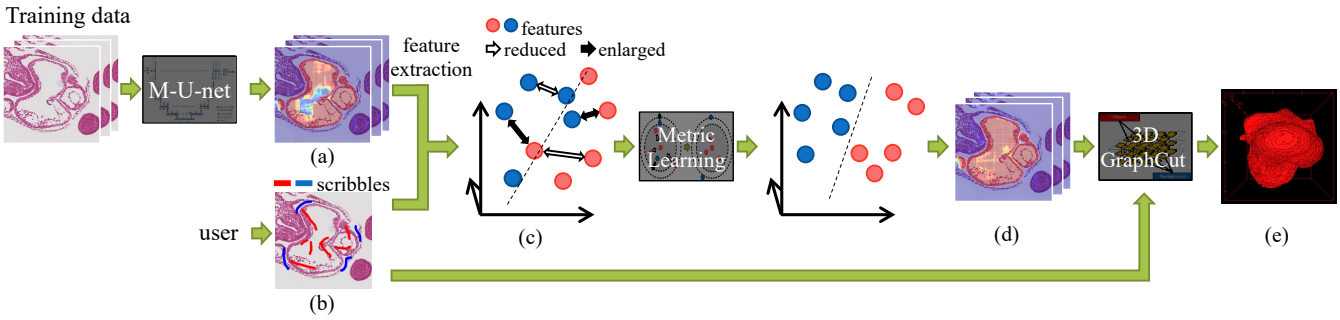

Fig. 2. Overview of proposed method: (a) probability heat-map predicted by M-Unet, (b) examples of scribbles, (c) feature space produced by M-Unet, (d) feature map retrained using discriminative deep metric learning, (e) 3D segmentation results.

evaluated in a heart segmentation task using actual sliced histological images of early mouse embryos. It achieved the best performance compared with three other methods.

II. Related work

Several methods for segmenting an organ in a mouse embryo image obtained by micro-CT or micro-MRI imaging have been proposed. Zamyadi *et al.* [2] used an automated intensity-based group-wise registration approach that optimizes the transformation parameters used to register mouse embryos in MRI image. Wong *et al.* [3] proposed an atlas-based segmentation method for CT-images. These methods cannot be directly applied to sliced histological images since our target data have various slice angles.

More recently, CNN-based segmentation methods have been proposed and shown to outperform traditional computer vision techniques in various tasks. In particular, U-net [4] has been widely used for medical image segmentation. Many segmentation CNNs have been proposed that incorporate a graphical model [5][6][7][8], spatial pyramid pooling [5][9][10], dilated convolution [11], or multi-scale inputs (i.e., image pyramid) [5][12][13][14] for deep learning. Tokunaga *et al.* [15] proposed multi-field-of-view U-net (M-Unet), which uses the context of multi-magnifications. As discussed in the introduction, such supervised learning methods are limited due to the small amount of training data and the various appearances of the images.

Methods combining a CNN and graph-cut segmentation have been proposed. Lu *et al.* [16] used a probability heat-map predicted by a 3D-CNN as a data term for 3D graph-cut segmentation to segment the liver in CT images. Ma *et al.* proposed an integrated method to classify nasopharyngeal cancers. In these methods, additional annotations are used only for graph-cut seeds, not for feature representation. Unlike these methods, our segmentation method effectively uses scribble annotations not only graph-cut segmentation but also metric learning-based feature representation.

III. 3D segmentation by scribbles-based metric learning

Fig. 2 shows an overview of the proposed method, in which the proposed method first estimates the probability heat-map of the heart region for each slice and then 3D segmentation is performed by 3D-graphcut. First, the regions of the heart are estimated using M-Unet [15] for each slice, where the M-Unet is trained using training data. Fig. 2(a) shows the probability heat-map for the heart region. The output of the layer immediately before the output layer can be used as features for each pixel (Fig. 2(c)). The features of the positive and negative samples may not be distinguished by a linear function (one-by-one convolution) due to the small amount of training data and the various appearances of the images, *i.e.*, the prediction includes incorrect regions. For several slices, a user adds scribbles (Fig. 2(b)) for both foreground and background regions to fix the incorrect regions in the probability heat-map. Next, the metric learning is trained using both the original training data and the data from the scribbles to facilitate pixel classification. Using the retrained feature map produced by metric learning improves the probability heat-map of the heart region (Fig. 2(d)). Finally, 3D graph-cut segments the 3D structure of the heart region (Fig. 2(e)), where the probability heat-map from metric learning is used as the data term and the scribbles are used as foreground and background seeds. The details of these steps are explained in the following sections.

A. Multi-field-of-view Unet

In general, a histological image is much larger than a natural image, so an entire image cannot be input to a network due to the GPU memory size limitation. Instead, small patches are clipped from the image and used as inputs to the CNN. Under the memory size limitation, if we use high-resolution images for a patch, the field-of-view is too small to discriminate the heart region. The field-of-view can be broadened by downsampling, but this reduces the spatial resolution.

To address this trade-off problem, we use M-Unet [15]. M-Unet aggregates contextual information from multiple magnification images by aggregating several segmentation networks that are trained using different-field-of-view images. The probability heat-maps from patch images are integrated to create a complete map.

Fig. 3 shows the overview of the training process. To integrate M-Unet and the network for metric learning, the M-Unet is first trained individually and then the network for metric learning is trained using the output features of the M-Unet. In the first step, M-Unet is trained using the training

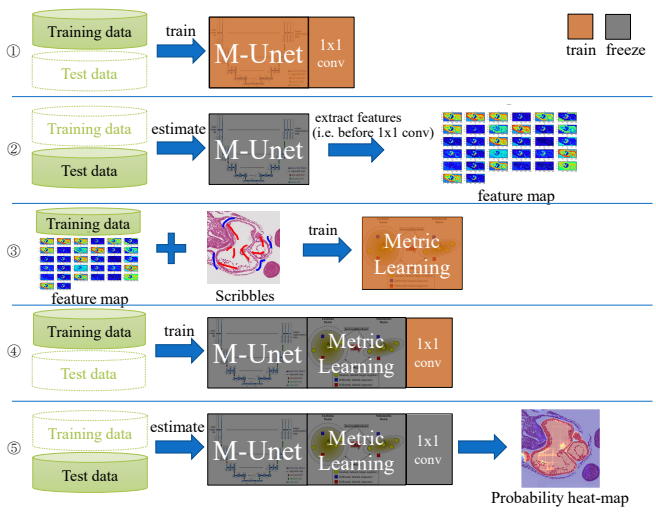

Fig. 3. The overview of the training process. 1) Training M-Unet, 2) extracting features using trained M-Unet, 3) annotating scribbles and training metric learning network, 4) training discriminating layer (1-by-1 conv) with fixing other layers, 5) estimating the porobability heat-map.

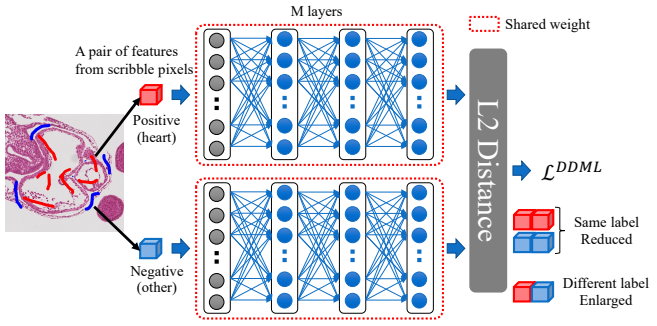

Fig. 4. Overview of DDML used for feature representation learning. A pair of pixel features, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, are mapped into same feature subspace by using fully connected layers; distance between their outputs is computed at the rightest layer. In this example, distance between positive sample (red) and negative sample (blue) is enlarged, and this is used for training network parameters.

data so that the last layer (*i.e.*, 1-by-1 conv) produces the probability heat-map of the heart region. In M-Unet, the last 1-by-1 conv layer can be considered as a linear function that inputs the image features and outputs the heat-map, in which the feature maps can be obtained from the layer immediately before the one-by-one. On the second step, the method estimates the probability map and the feature maps for each sliced image in the test data by using the trained M-Unet.

B. Scribbles for discriminative deep metric learning

The entire probability heat-map produced by M-Unet is overlaid on the original image, and the overlaid images are shown to the user, who fixes the incorrect regions by adding annotations (scribbles) as shown by the 3rd step in Fig. 3. We here note that the annotations are added for only several slices in the entire 3D data.

These scribbles and the original training data are used to train metric learning to learn a good distance metric for the test image so that the distance between the same-label pairs (*i.e.*, {heart, heart} or {other, other}) is reduced and that of the different-label pairs (*i.e.*, {heart, other}) is enlarged as much as possible. Discriminative deep metric learning (DDML) [17] is used to train a deep neural network that learns a set of hierarchical nonlinear transformations to project pixel pairs into the same feature subspace so that discriminative information can be exploited in the deep network.

The DDML consists of fully connected layers to compute the representations of a pixel pair by passing them through multiple layers of nonlinear transformations, as shown in Fig. 4, in which the number of layers is M, and the DDML is represented as function f. A pair of pixel features $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be represented as $f(\boldsymbol{x}_i)$ and $f(\boldsymbol{x}_j)$ at the output layer once they have passed through the deep network. Their distance can be defined by computing the squared Euclidean:

$$d_f^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\|_2^2 . \tag{1}$$

The DDML method aims to find a nonlinear function f so that $d_f^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is smaller than threshold τ in the transformed space if the labels of $\boldsymbol{x}_i$, and $\boldsymbol{x}_j$ are the same ($l_{ij} = 1$) and larger than τ if their labels are different. Therefore, the loss function of DDML is formulated as:

$$\mathcal{L}^{\mathrm{DDML}} = \frac{1}{2}\sum_{i,j} g\left(1 - \ell_{ij}\left(\tau - d_f^2\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right)\right) \tag{2}$$

$$+ \frac{\lambda}{2}\sum_{m=1}^{M}\left(\left\|\mathbf{W}^{(m)}\right\|_F^2 + \left\|\mathbf{b}^{(m)}\right\|_2^2\right), \tag{3}$$

where $g(z) = \frac{1}{\beta}\log\left(1 + \exp(\beta z)\right)$ is the generalized logistic loss function [18], where β is a sharpness parameter. The second term is a regularization term for weights W and bias $\mathbf{b}$, where λ is a hyperparameter. The optimization problem can be solved by backpropagation.

The trained DDML is used to obtain the transformed features. Since the outputs of the DDML is the feature maps, we have to estimate the probability heat-map from the feature maps. Therefore, the one-by-one convolution layer (discriminating layer) is added to after the last layer, and then the layer is fine-tuned with freezing the other layers (M-Unet and DDML layer) as shown by the 4th step in Fig. 3. The trained DDML and the discriminating layer is used to obtain the probability heat-map for all sliced data, as shown by the 5th step in Fig. 3. The set of sliced data can be treated as 3D volume data in the next step. We define the integrated layers consisting of the M-Unet, DDML, and discriminating layers as function h, and the output of this function takes a value from 0 to 1 for each voxel.

C. 3D graph-cut

To obtain the final 3D segmentation results, we use 3D graph-cut that optimizes the following energy function:

$$E(L|X,H) = \sum_{v\in\mathcal{V}} D_v(L|H) + \gamma \sum_{(v,v')\in\mathcal{N}} B_{v,v'}(L|X), \tag{4}$$

$$D_v(L|H) = |l_v - h_v|, \tag{5}$$

$$B_{v,v'}(L|X) = |l_v - l'_v|\left[\exp\left\{-\frac{(x_v - x_{v'})^2}{\sigma^2}\right\} + \delta\right], \tag{6}$$

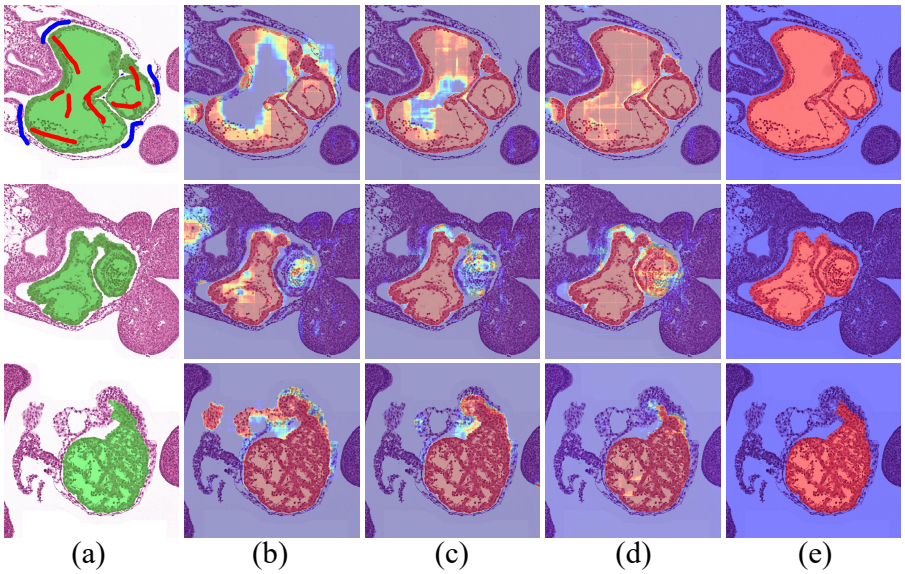

Fig. 5. Example segmentation results obtained by each method: (a) ground-truth, (b) U-net, (c) M-Unet, (d) M-Unet+DDML, (e) M-Unet+DDML+Graph-cut (proposed). Top row shows examples in which the slice is annotated. Middle and bottom rows show eamples that the slice is not annotated.

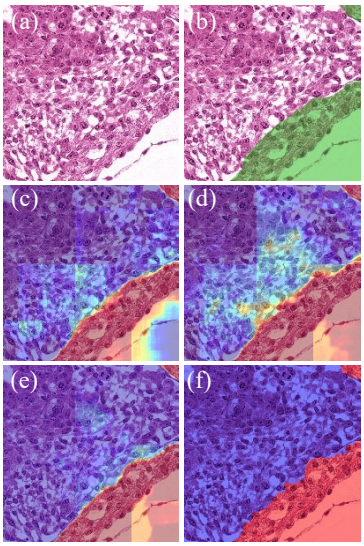

Fig. 6. Example of difficult case: (a) original image, (b) ground-truth, (c) U-net, (d) M-Unet, (e) M-Unet+DDML, (f) proposed.

TABLE I
AVERAGE OF EVALUATION METRICS BY METHOD.

	Precision	Recall	F1	IoU
U-net	0.810	0.824	0.810	0.681
M-Unet	0.841	0.960	0.889	0.807
M-Unet+DDML	0.874	0.959	0.910	0.841
M-Unet+DDML+GC	**0.983**	**0.981**	**0.982**	**0.965**

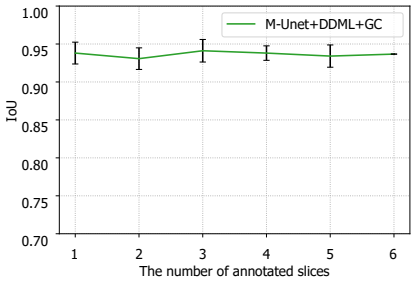

Fig. 7. Average IoU when the number of slices that were annotated changes, where the horizontal axis is the number of annotated slices, and the vertical axis is the average of IoU. Error bars indicate the standard deviation.

where $\mathcal{V}$ is the set of voxels in the 3D volume, $X = \{x_v | v \in \mathcal{V}\}$ is the set of voxel values, and x_v is the intensity of the v-th voxel. $L = \{l_v | v \in \mathcal{V}\}$ is the set of binary labels $l_v \in \{0, 1\}$ to be optimized. If l_v is 1, v is in the heart region; otherwise v is in another region. The first term, $D_v(L|H)$, is a data term that penalizes the energy if labels L differ from $H = \{h_v | v \in \mathcal{V}\}$, where h_v is the v-th voxel value of the probabilities predicted by function h (*i.e.*, the output of our entire network consisting of M-Unet, DDML, and the discriminating layer). The second term, $B_{v,v'}(L)$, is a pairwise term that penalizes the energy if the labels of neighboring voxels differ when these intensities take similar values. For this pairwise term, we use the intensity values of the original image since three-quarters of the boundary have edges. γ, σ and δ are hyperparameters that control the spatial smoothness of the output labels, and these values are optimized using validation data. This optimization problem can be solved by using max-flow/min-cut optimization. Fig. 2(e) shows an example of the 3D segmentation results.

IV. EVALUATION

A. Dataset

To evaluate the proposed method, we first sliced mouse embryos, captured an image of each slice by a microscopic scanner, and placed the images in six-volume data. Then, images in each data were aligned to create 3D volume data. After that, the color of the images in each data is normalized using a color normalization method proposed by Vahadane *et al.* [19]. Each data volume included from 130 to 250 z-slices depending on the size of the mouse. An annotated heart region appeared in about 80 slices for each volume data, where parts of the heart cannot be segmented even by biological experts. The size of a sliced image is 5700×10000 pixels. For the ground-truth, biological experts manually segmented the heart regions for each slice in the data. Two volume data were used for the training, two for validation, and two were used for testing.

As the additional annotation (scribbles) in the test data, every 10 sliced were picked up (*i.e.*, 4 to 6 slices per volume), and then those were annotated using scribbles by a user, with four or five curves for the heart region (red scribbles) and for the other regions (blue scribbles). To evaluate the robustness for scribbles, we made three different scribbles for each data. The top row image in Fig. 5(a) shows an example of the scribbles overlaid on an original image.

B. Parameter setup

To train the proposed network that integrating M-Unet, DDML and the discriminating layer, we used the Adam

[20] for optimization, with a batch size of 8, and a learning rate of 0.001. For initialization, we used M-Unet pre-trained using other pathological images. The hyperparameters were optimized using validation data and then fixed for all test data.

C. Segmentation Accuracy

As an ablation study, we evaluated the proposed method (M-Unet + DDML + Graph-cut) compared with U-net [4], M-Unet [15], and M-Unet + DDML. Note that the last method is a part of the proposed method. We evaluated three metrics; the mean of the precision, recall, F1-score (F1), the intersection over the union (IoU), which have been widely used for evaluating segmentation [21].

Fig. 5 shows examples of the segmentation images predicted by each method. Looking at the top row, we see that U-net had many false negatives inside the heart region, and it had some false positives on the left side of the heart. M-Unet had fewer false negatives inside the heart region, but there were still some false negatives and a false positive on the left side. DDML had even fewer false negatives, but the false positives remained. Finally, the proposed method (M-Unet+DDML+Graph-cut) reduced the number of false positives. Looking at the middle and bottom rows, we see similar results.

Fig. 6 shows an example of a difficult case in which the boundary of the heart region is ambiguous. The DDML reduced the false positives compared with U-net and M-Unet, and thus the proposed method successfully segmented the regions.

Table I summarizes the results of quantitative experiments. Three different scribble patterns were added for each volume, the averages of evaluation metrics were described in the table. The results were similar to those for qualitative evaluation. DDML improved three metrics (Precision, F1, and IoU) compared with M-Unet, and graph-cut further improved it. Thus, the proposed method achieved the best accuracy in all metrics, in which our method outperformed single networks significantly. Furthermore, this rank order of accuracy was the same for all test data and scribble patterns.

To demonstrate that our method robustly works with the variation of scribbles, we evaluated the IoU when the number of slices that were annotated changes from 1 to 6. In each condition, three types of scribble patterns were prepared for the evaluation, and the average IoU of them was computed. As shown in Fig. 7, our method performed well even though the number of annotated slices is very few. It indicates that the scribbles patterns and the number of scribbles are not so significant for the performance. We here note that the IoU slightly decreased when the number of them is 2. We consider that the scribble pattern may affect a little, but it's not significant.

V. CONCLUSIONS

Our proposed segmentation method using scribbles for metric learning can retrain the feature space for representing discriminative features in test data. Deep discriminative metric learning is trained so that the distances between samples with the same class label are reduced while those of samples with different class labels are enlarged. This facilitates pixel classification. Evaluation using a heart region segmentation task showed that the proposed method performs better than three other methods. In future work, we will apply the proposed method for various segmentation tasks, such as tumor region segmentation from pathological images, and we will further investigate the relationship between the scribble pattern and the accuracy.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP17K19402, JP16H06280, and JP18H05104.

REFERENCES

[1] A. Krizhevsky, S. Ilya, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[2] M. Zamyadi, R. M. Henkelman, S. Bhattacharya, J. E. Schneider, and J. G. Sled, "Registration of 3d mr images of the mouse embryos," in *International Society for Magnetic Resonance in Medicine*, vol. 16, 2014.

[3] J. P. L. Michael D. Wong, Yoshiro Maezawa and R. M. Henkelman, "Automated pipeline for anatomical phenotyping of mouse embryos using micro-ct," *Development*, vol. 141, pp. 2533–2541, 2009.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation."

[5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs."

[6] V. Jampani, M. Kiefel, and P. Gehler, "Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks."

[7] S. Zheng, S. Jayasumana, B. Romera-Paredes, and V. V. *et.al.*, "Conditional random fields as recurrent neural networks."

[8] R. Vemulapalli, O. Tuzel, M. Liu, and R. Chellappa, "Gaussian conditional random field network for semantic segmentation."

[9] L. Chen, G. Papandreou, and F. Schroff, "Rethinking atrous convolution for semantic image segmentation," in *arXiv*, 2017.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network."

[11] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations*, 2016.

[12] L. Chen, Y. Yang, J. Wang, W. Xu, and A. Yuille, "Attention to scale: Scaleaware semantic image segmentation."

[13] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture."

[14] G. Lin, C. Shen, A. Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation."

[15] H. Tokunaga, Y. Teramoto, A. Yoshizawa, and R. Bise, "Multi-field-of-view cnn for semantic segmentation in pathology," in *IEEE CVPR*, 2019. (in press).

[16] F. Lu, F. Wu, P. Hu, Z. Peng, and D. Kong, "Automatic 3d liver location and segmentation via convolutional neural network and graph cut," *Computer Assisted Radiology and Surgery*, vol. 12, no. 2, pp. 171–182, 2017.

[17] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild."

[18] A. Mignon and F. J. Pcca, "A new approach for distance learning from sparse pairwise constraints."

[19] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *Medical Imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[20] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization."

[21] G. Csurka, D. Larlus, and F. Perronnin, "What is a good evaluation measure for semantic segmentation?"