Automatic Liver Segmentation Using U-Net with Wasserstein GANs

Yuki Enokiya, Yutaro Iwamoto, and Yen-Wei Chen
Information Science and Engineering, Ritsumeikan University, Shiga, Japan
Email: is0205xr@ed.ritsumei.ac.jp, yiwamoto@fc.ritsumei.ac.jp, chen@is.ritsumei.ac.jp

Xian-Hua Han
Faculty of Science, Yamaguchi University, Yamaguchi, Japan
Email: hanxhua@yamaguchi-u.ac.jp

***Abstract*—Automatic liver segmentation in CT images is an important step for computer-aided diagnosis and computer-aided hepatic surgery. Recently, though numerous methods based on deep learning such as U-Net have been proposed for automatic liver segmentation, it is still a challenging topic because of its low contrast and variations of liver shape. Additionally, limited training data for deep learning is another challenging problem. In this paper, we propose an automatic liver segmentation using U-Net with a Wasserstein Generative Adversarial Network (GAN). The Wasserstein GAN was used to improve U-Net's training, especially training with a small data set. We demonstrated that liver segmentation accuracy (Dice value) with 33 and 392 training data sets was improved from 88% to 92% and from 92% to 93%, respectively.**

***Index Terms*—liver, segmentation, deep learning, GAN, WGAN**

I. Introduction

The liver is an organ that is essential for vital body functions, such as its role in bile formation, nutrient storage, toxic decomposition, and production of certain blood components. However, it is also an organ that easily becomes severely ill, because of its scarce subjective symptoms for a disease. It is, therefore, called the "silent organ." CT or MRI is effective for an early detection and diagnosis of liver diseases. Additionally, recent advancements in the performance of these imaging techniques have led to the production of larger data volume. Therefore, physicians require computer assistance in assessing this large volume of obtained data. Automatic liver segmentation in CT images is an important step to achieve computer-aided diagnosis and to perform computer-aided hepatic surgeries [1].

To date, numerous methods for liver segmentation have been proposed, including level set [2], clustering [3], anatomic model-based methods [4], [5], and interactive methods [6]-[9]. Recently, deep learning has demonstrated a powerful ability in computer vision tasks, by automatically learning hierarchies of relevant features directly from the input data. Deep learning methods have been successfully applied to liver segmentation with the use of various networks, including Fully Convolutional Network (FCN) [10], [11] and U-Net [12], [13]. Because results obtained from FCN or U-Net segmentation results are neither perfect nor smooth, refinement has to be performed with Conditional Random Field (CRF) [14] or graph cut [12], [15]. An additional approach to refine the segmentation results is to use a Generative Adversarial Network (GAN) [16], [17]. In the GAN-based segmentation approaches, the generator is used to perform the segmentation task, whereas the discriminator is used to refine the training of the generator, which is exclusively used in the training. In the present study, we propose an automatic liver segmentation method based on U-Net with a Wasserstein GAN (WGAN). WGAN [18], [19] differs from GAN for its objective function. In WGAN, the Wasserstein distance is used as the objective function. Compared to conventional GAN, the learning of WGAN is more stable. We also propose to enlarge the generator result (segmented liver with surrounding region) as the input of the discriminator to improve the recall.

In this article, Section 2 describes related work, Section 3 describes the proposed method, Section 4 presents experimental results, and Section 5 details the discussion and conclusions.

II. Related Work

A. U-Net

U-Net is a network that is based on the principle of fully convolutional networks [12]. It is composed of an encoder for extracting features and a decoder for reconstructing images. Additionally, skip connection is used to combine low- and high-level features, enabling accurate localization. Such network architecture is often used for medical image analysis. Segmentation of a 3D structure, such as the liver, is performed by repeating a sequence of 2D slice segmentation. Because this approach does not include the context information along the z axis, the consistency among slices is lost.

Manuscript received August 10, 2018; revised December 6, 2018.

doi: 10.18178/joig.6.2.152-159

B. 3D U-Net

3D U-Net [13] is an extension of U-Net to maintain the 3D structure. Each U-Net layer is replaced with a 3D convolution and 3D max pooling. It successfully maintains continuity in the direction of the vertical axis by performing segmentation using volume sequence rather than slice sequence. In this respect, a 3D sequence is more desirable as it can capture all spatial contexts. However, when compared to 2D convolution, 3D convolution on the volume data is computationally more expensive. Commonly, 3D convolution has more parameters than 2D convolution, with a lower amount of learning data. This represents a significant problem in medical images with small training data.

C. Generative Adversarial Networks

GAN [20], [21] are a generation model composed of two networks: generator and discriminator. The purpose of GAN is to first learn a distribution close to the learning data and then generate data similar to the learning data. GAN is defined as a minimax game of generator G and discriminator D, as follows:

$$min_G\, max_D\, \mathbb{E}_{x\sim\mathbb{P}_r}[logD(x)] + \mathbb{E}_{z\sim\mathbb{P}_z}\left[log\left(1 - D\big(G(z)\big)\right)\right] \tag{1}$$

where $\mathbb{P}_r$ is the real data distribution and $\mathbb{P}_z$ represents the generated data distribution. z is a random number based on a uniform distribution.

The generator inputs a random number based on a uniform distribution and, to generate data, it conducts convolution and deconvolution. The discriminator inputs real or generated data as input, after assessing which one has to be input. Based on the discriminator's discrimination result, it updates parameters of the generator and the discriminator. Upon failure of the discriminator to discriminate the real data from generated data, it updates the discriminator's parameters. If it is determined that the discriminator's generated data is not real, the generator's parameters are updated. By repeating these steps, the generator can generate data similar to real data, and the discriminator can discriminate generated data from real data with a high rate of accuracy.

D. Image-to-Image Translation with Conditional Adversarial Networks

Image-to-Image Translation with Conditional Adversarial Networks (pix2pix) [22] is an image conversion method that uses GAN. The technique generates a pair of images by interpolation taking into consideration a relationship from one image by learning the relationship between two images. The difference between pix2pix and GAN is based on the fact that the input of the generator is not a random number but an image. Also, the discriminator uses patch GAN, developed based on Conditional GAN [23]. It discriminates each patch, not the whole image. Importantly, pix2pix can convert images with higher precision than traditional networks. Our proposed method is based on pix2pix.

E. Wasserstein GAN

WGAN [18], [19] is one of the GANs. WGAN differs from GAN for its objective function. There are problems with GAN, such as the vanishing gradient problem, as well as instable mode collapse and learning. WGAN has overcome these problems by using Wasserstein distance. WGAN is defined as a minimax game of generator G and discriminator D as follows:

$$\min_G \max_{D\in Đ} \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{z\sim\mathbb{P}_z}\big[D\big(G(z)\big)\big] \tag{2}$$

where $\mathbb{P}_r$ is real data distribution and $\mathbb{P}_z$ is generated data distribution. z is random number based on uniform distribution. D is a function of 1-Lipschitz. WGAN finds the Wasserstein distance of the input real image and the generated image.

As a result, the learning of WGAN is more stable than that of GAN.

III. Proposed Method

Fig. 1 describes the flow of our proposed method. First, we segment the CT image using the modified U-Net, which is used as a generator. Next, the segmentation result and the Ground Truth (GT) data (a manually segmented result) are input to the discriminator to judge if the generator result is real (1) or fake (0). We learn U-Net and discriminator based on the identification result and error obtained from U-Net.

A. Generator

Generators used in GAN generate data by inputting random numbers. However, in the present study, a liver segmentation image (a binary mask data) represents the output and a CT image represents the input. We used U-Net as the network architecture of generator. Of note, the configuration is changed based on Deep Convolutional GANs [13]. The change points are represented by the abolition of pooling layers; activation function of encoder to Leaky ReLU; activation function of decoder to ReLU; and introduction of Batch Normalization. Accuracy is not significantly affected by these changes, as shown in Table I by the number of training data 10 and 99. The generator consists of 19 convolution layers. Fig. 2 shows the network architecture of the generator. We called it a modified U-Net. The generator's parameter is updated based on cross entropy loss of the segmentation result and discriminator's discrimination result.

B. Discriminator

In the present study, two discriminators were used in our experiments. The first discriminator is a network based on DCGAN [24] with a loss function as (1). The second discriminator is based on WGAN, which has the same network structure and minimizes Wasserstein distance between the GT data and the generated data with a loss function as (2). The first discriminator is represented as GAN and the second discriminator is represented as WGAN. Fig. 3 describes the discriminator's network architecture. It is composed of 10 layers and has the same structure as the encoder of generator. Before inputting segmentation results or GT to the discriminator, CT data are clipped out with the segmentation result or GT (as shown in Fig. 4).

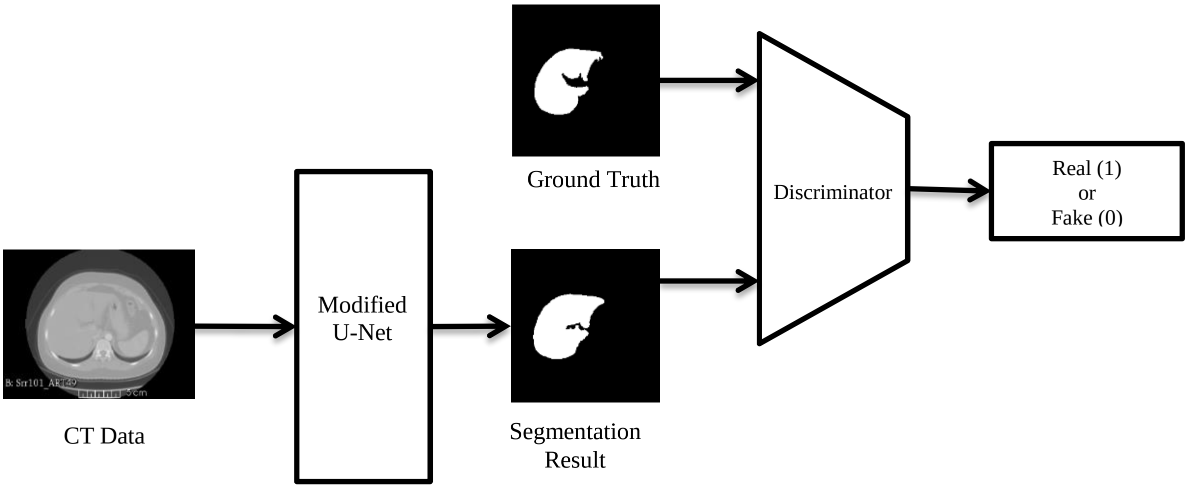

Figure 1. Flow of the proposed method. First, we segment the CT image using a modified U-Net, which is used as a generator. Next, the segmentation result and the ground truth (GT) data (a manually segmented result) are input in the discriminator to understand if the generator results are real (1) or fake (0). We learn U-Net and discriminator based on the identification result and error obtained from the modified U-Net. Generator is a modified U-Net and Discriminator is WGAN.

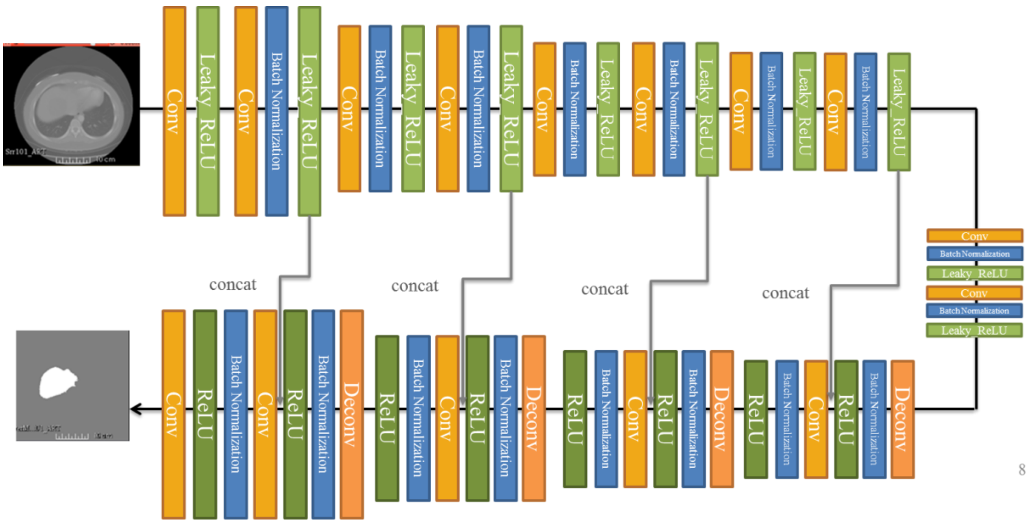

Figure 2. The architecture of the generator, which is modified U-Net. The modification is based on the contents of DCGAN. Compared to the original U-Net, our generator (the modified U-Net) introduced Leaky ReLU and Batch Normalization and deleted Max Pooling layers. The feature map channel and kernel size are the same as U-Net.

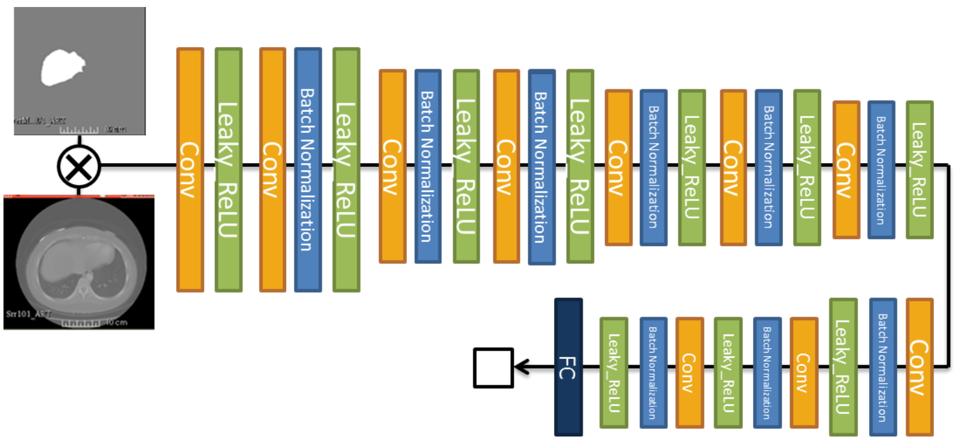

Figure 3. The architecture of the discriminator. This is same as encoder of the generator. The difference is that there is no decoder and Full Connection has been added.

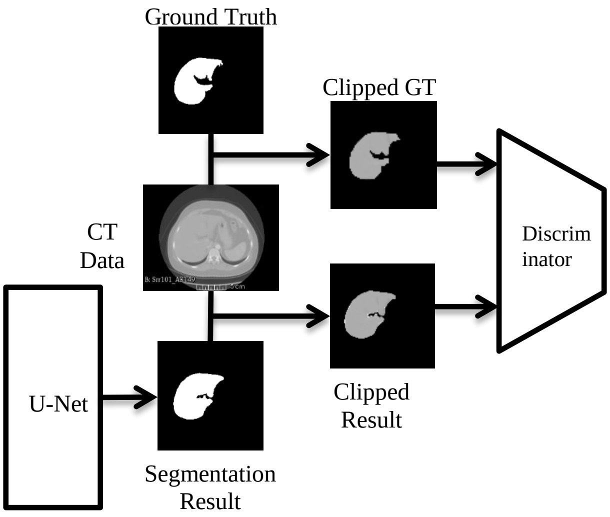

Figure 4. Clipped CT data image. CT data are clipped out by using segmentation result (mask image) or ground truth before being input into the discriminator.

IV. EXPERIMENTS

This section outlines an evaluation of our system. We trained our models with the use of our own dataset, provided by Zhejiang University Hospital. We tested our model by using 20 volumes of fully annotated 3DIRCADb dataset [25]. Additionally, we used RMSprop as an optimization for network learning. Learning rate was 0.005, and batch size was 3. These are generally used to perform generator and discriminator training. Table I, Fig. 5 and Fig. 6 outline the experimental results. 3D U-Net was not included in this article because of insufficient GPU memory. Experimentation under the same conditions as other methods was not possible because of higher capacity needed by the 3D convolution and 3D deconvolution. The quantitative evaluation metrics used in this study were dice, precision, and recall. Dice is comparable to F-score, which is the true positive number of the average size of two segmented areas. Precision is the proportion of what is actually positive, within data expected to be positive. Recall is the proportion of what is expected to be positive, within actually positive data. These measures are expressed as follows:

$$Dice = \frac{|T \cap P|}{(|T| + |P|)/2} \tag{3}$$

$$Precision = \frac{|T \cap P|}{|P|} \tag{4}$$

$$Recall = \frac{|T \cap P|}{|T|} \tag{5}$$

where T and P represent the GT region and the predicted region, respectively.

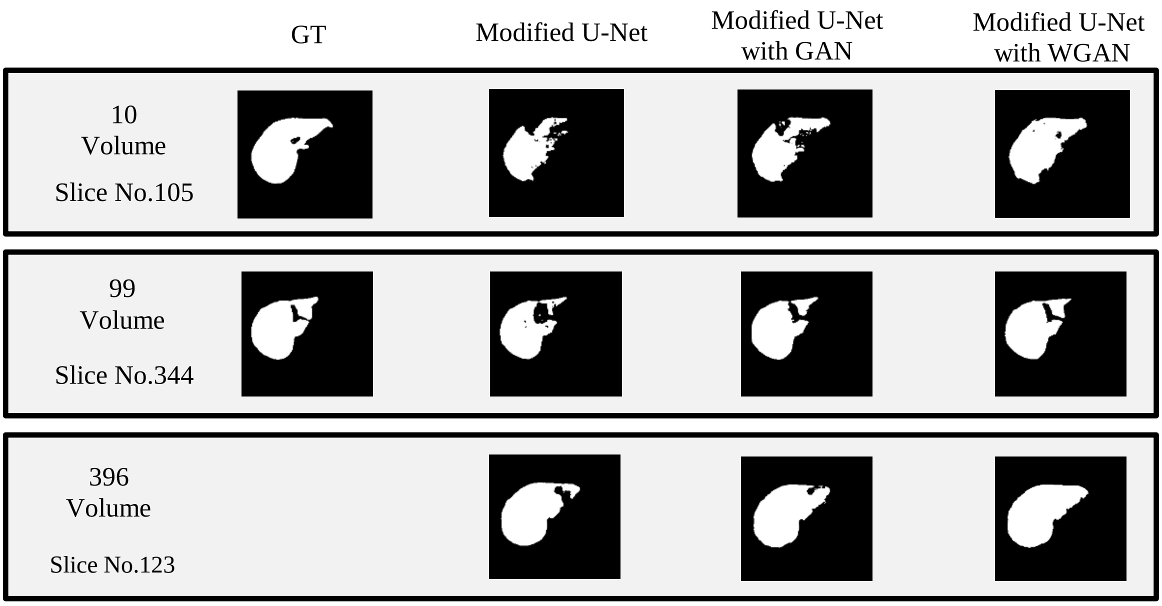

Figure 5. Examples of correctly segmentation result. First row is ground truth. Second row is the result of modified U-Net only. Third row is the result of modified U-Net with GAN's discriminator. Last row is the result of modified U-Net with WGAN's discriminator. By using GAN and WGAN, it has become possible to detect parts which could not be detected by modified U-Net only. This is in line with the rise of recall and dice shown in Table I.

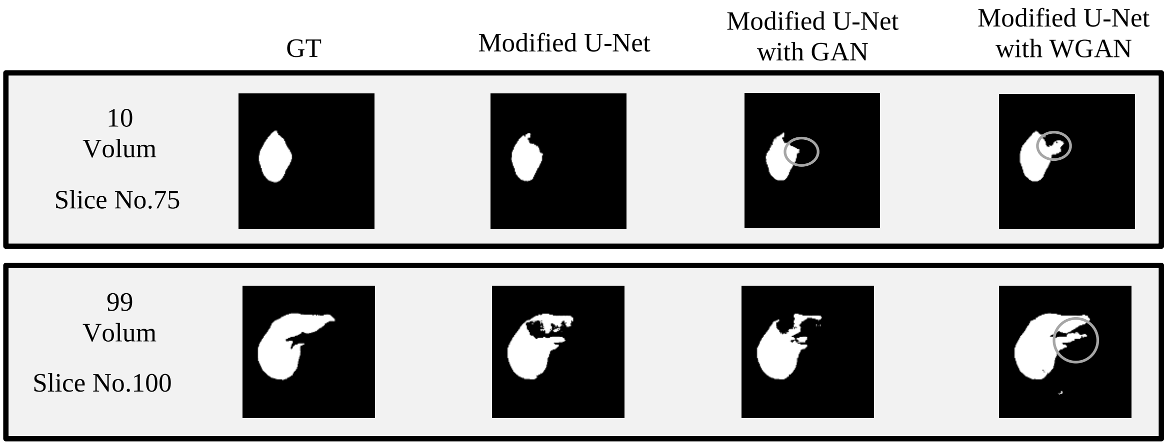

Figure 6. Examples of segmentation result with error. First row is ground truth. Second row is the result of modified U-Net only. Third row is the result of modified U-Net with GAN discriminator. Last row is the result of modified U-Net with WGAN discriminator. The circle represents the deteriorated part. Unnecessary parts have been detected with the use of GAN WGAN. As for precision, modified U-Net tends to be higher, in line with the results shown in Table I.

TABLE I. Comparison of Different Methods and Different Number of Training Data A

Number of training data	Method	Dice	Precision	Recall
10	U-Net	0.85	**0.95**	0.77
	Modified U-Net	0.85	**0.95**	0.76
	Modified U-Net + GAN	0.86	0.93	0.80
	Modified U-Net + WGAN	**0.90**	0.94	**0.86**
33	Modified U-Net	0.88	**0.95**	0.81
	Modified U-Net + GAN	0.89	0.94	0.85
	Modified U-Net + WGAN	**0.92**	0.94	**0.90**
99	U-Net	0.89	**0.95**	0.84
	Modified U-Net	0.89	0.94	0.85
	Modified U-Net + GAN	0.91	**0.95**	0.88
	Modified U-Net + WGAN	**0.92**	0.94	**0.91**
198	Modified U-Net	0.92	**0.96**	0.89
	Modified U-Net + GAN	**0.93**	0.95	**0.91**
	Modified U-Net + WGAN	**0.93**	**0.96**	**0.91**
396	Modified U-Net	0.93	0.96	0.90
	Modified U-Net + GAN	0.93	0.96	**0.91**
	Modified U-Net + WGAN	**0.94**	**0.97**	**0.91**

[a] "Number of training data" means total number of volumes of data used for learning, out of 396 volumes. The best result for each volume is shown in bold.

Modified U-Net with WGAN is the most accurate in several cases, compared to modified U-Net and modified U-Net with GAN. Although an improvement of the dice can be observed with an increase in the amount of data used for training, no significant change is seen once all the training data are used, even if each discriminator is used. Additionally, with a small amount of training data, a high rate of dice is observed with the use of the discriminator. A similar tendency to dice can be noted for recall. Based on these observations, an improvement in accuracy can be expected even when the discriminator is used, regardless of the use of a small amount of training data. Precision has the tendency of being higher for modified U-Net only, wand recall tends to lower. Based on these observations, we can infer that modified U-Net is making smaller prediction than it actually is. With the use of WGAN, it can be observed that dice and recall tend to be higher and precision tends to be lower, as opposed to modified U-Net. From these observations, it appears that modified U-Net with WGAN is making larger prediction than modified U-Net. However, in modified U-Net with WGAN, precision is higher than recall. Therefore, it can be implied that a smaller prediction than the actual one is being made. Segmentation using WGAN discriminator was effective with a small dataset. Therefore, we believe that it is effective also when using 3D convolution, because the latter uses fewer training datasets than 2D convolution.

Based on these findings, we thought of methods that would enable us to make larger predictions. Our goal was to judge with a wider range of information, by using a method that expands segmented mask images and inputs them into the discriminator. This approach is presented in Fig. 7. In this network, the segmentation's result is enlarged before inputting into the discriminator. Through this network, we aim at extending the mask image and considering the information of surrounding pixels.

In case of a small segmentation result, the liver should be in the expanded region. On the contrary, in case of a large segmentation, the boundary or other than the liver should be in the expanded region. By discriminating a wide region through this change, we expected that dice and recall will rise. Table II and Fig. 8 describe the experimental results.

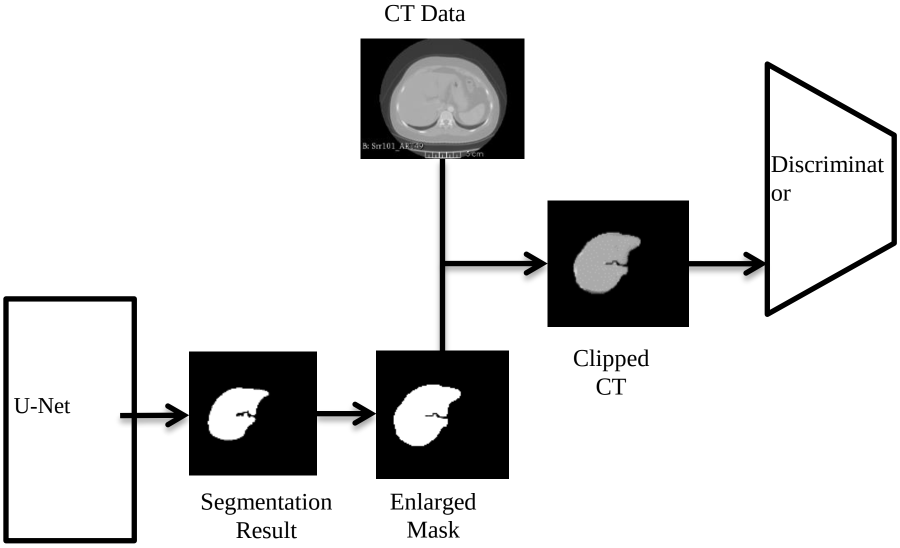

Figure 7. Network flow of mask expansion. In this network, the generator result (mask data) is enlarged before entering it into the discriminator. Although not shown here, ground truth follows the same processing. Other parts of the network are as in Figure 1.

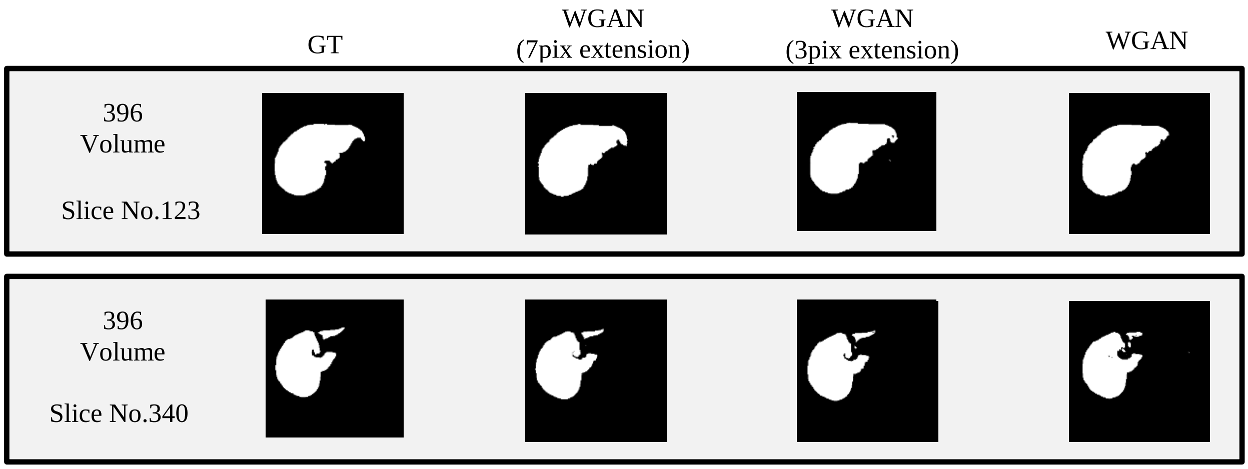

Figure 8. Result of segmentation using extended mask image. First row is ground truth. Second row is the result of using 7pix expanded mask. Third row is the result of 3pix expanded mask. Last row is the result of using no expanded mask. By enlarging the mask region, it is possible to detect apportion that has not been detected so far. However, expansion of the mask region results in blurring of the detection's outline.

TABLE II. MASK IMAGE ENLARGEMENTS COMPARISON[A]

Number of training data	Method	Dice	Precision	Recall
396	Modified U-Net + WGAN	**0.94**	**0.97**	0.91
	Modified U-Net + WGAN (expanded 3pix)	**0.94**	0.96	0.92
	Modified U-Net + WGAN (expanded 7pix)	**0.94**	0.95	**0.93**

[a]The best result of each metrics is indicated in bold.

When compared to modified U-Net with WGAN discriminator and 3pix enlarged, 7pix enlarged and no expanded, and no change in dice in either method was observed. However, we found that precision is the highest in no expanded and recall is the highest in 7pix enlarged. These results imply that, when the region of the mask is expanded, fail positive increase and fail negative decrease. Because we value Recall more than Precision as an evaluation metric, we can say that the accuracy has improved. When compared to modified U-Net only, there is a stronger tendency to increase recall and decrease precision. Additionally, as shown in Fig. 6, enlargement of the mask region allows the detection of apportion, which has not been detected so far. However, expansion of the mask region leads to a blurred outline of the detection result. Following an increase of the size to expand, this tendency becomes stronger. Interestingly, when using 7pix enlarged, we observed crushed details and differentiation becomes difficult. Between modified U-Net with WGAN and expanded 7pix, and between expanded 3pix and no expanded, we observed no change in dice. However, in the future with the continuation of mask region expansion, it is thought that following the stop of the increase of recall and the decrease of precision, dice decreases.

Finally, with localization of the discontinuous liver in the same slice, smaller liver detection rate is low. Fig. 9 describes such results. A smaller liver cannot be detected by either method. Because smaller liver is not detectable by the first U-Net segmentation, we can conclude that no change is present despite the use of the discriminator. To overcome this limitation, we believe that changes to the U-Net are more effective than those to the discriminator.

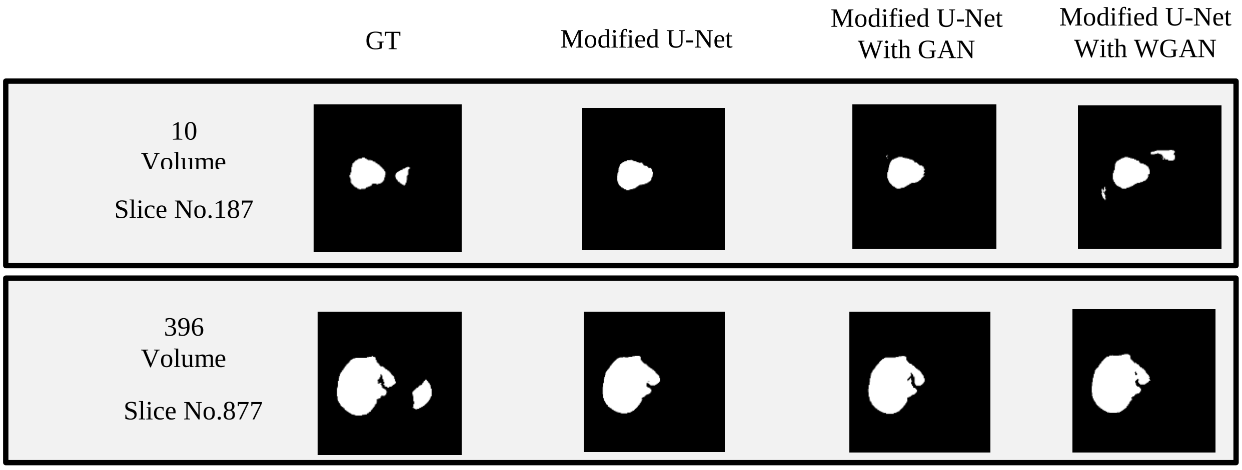

Figure 9. Result of segmentation methods and volumes. First row is ground truth. Second row is the result of modified U-Net only. Third row is the result of modified U-Net with GAN discriminator. Last row is the result of modified U-Net with WGAN discriminator. None are capable of detecting the smaller liver.

V. Conclusions

In the present article, we proposed a network that performs adversarial learning in segmentation of medical images. We confirmed an improvement of the dice value by about 3%–5% through the proposed adversarial training, with the use of a small training dataset. We observed an improvement of recall by about 2%, through extension of the region of the mask image as an input of the discriminator.

To date, in the presence of a discontinuous liver in the same slice, the detection rate of small liver is low. As a future task, we believe that we can address this issue by adjusting the weight of learning according to the liver's shape.

Acknowledgment

We thank Sir Run Run Shaw Hospital for providing medical data and helpful advice on this research. This work is supported in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant Nos. 18H03267, 18K18078, in part by Zhejiang Lab Program under the Grant No.2018DG0ZX01.

References

[1] M. Kaibori, *et al.*, "Novel liver visualization and surgical simulation system," *J. Gastrointest Surg.*, vol. 17, pp. 1422-1428, 2013.

[2] J. Lee, *et al.*, "Efficient liver segmentation using a level-set method with optimal detection of the initial liver boundary from level-set speed images," *Computer Methods and Programs in Biomedicine*, vol. 88, pp. 26-28, 2007.

[3] A. H. Foruzan, *et al.*, "Segmentation of liver in low-contrast images using K-means clustering and geodesic active contour algorithms," *IEICE Trans.*, vol. E96-D, pp. 798-807, 2013.

[4] T. Okada, *et al.*, "Automated segmentation of the liver from 3D CT images using probabilistic atlas and multi-level statistical shape model," *Academic Radiology*, vol. 63, pp. 1390-1403, 2008.

[5] C. Dong, *et al.*, "Segmentation of liver and spleen based on computational anatomy models," *Computers in Biology and Medicine*, vol. 67, pp. 146-160, 2015.

[6] Y. Boykov and M. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of object in N-D images," in *Proc. International Conference on Computer Vision*, 2001, pp. 105-112.

[7] T. Kitrungrotsakul, X. H. Han, and Y. W. Chen, "Liver segmentation using superpixel-based graph cuts and regions of shape constraints," in *Proc. IEEE International Conference on Image Processing*, 2015, pp. 388-QATc-141.

[8] C. Dong, *et al.*, "Simultaneous segmentation of multiple organs using random walks," *Journal of Information Processing Society of Japan*, vol. 24, pp. 320-329, 2016.

[9] Y. Yuan, *et al.*, "Hybrid method combining superpixel, random walk and active contour model for fast and accurate liver segmentation," *Computerized Medical Imaging and Graphics*, 2018 (in press).

[10] Q. Dou, *et al.*, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Proc. MICCAI*, 2016, pp. 149-157.

[11] T. Kitrungrotsakul, *et al.*, "Interactive liver segmentation in CT volumes using fully convolutional networks," in *Proc. IIMSS*, 2018, pp. 216-222.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234-241.

[13] O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*, 2016, pp. 424–432.

[14] P. F. Christ, *et al.*, "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields," in *Proc. MICCAI*, 2016, pp. 415-423.

[15] F. Lu, *et al.*, "Automatic 3D liver location and segmentation via convolutional neural network and graph cut," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, pp. 171-182, 2017.

[16] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proc. ICCV*, 2017, pp. 5688-5696.

[17] D. Yang, *et al.*, "Automatic liver segmentation using an adversarial image-to-image network," in *Proc. MICCAI*, 2017, pp. 507-515.
[18] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein GAN," arXiv: 1701.07875, 2017.
[19] I. Gulrajani, F. Ahamed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improver training of wasserstein GANs," arXiv: 1704.00028, 2017.
[20] I. Goodfellow, *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014.
[21] E. L. Denton, *et al.*, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Proc. NIPS*, 2015, pp. 1486-1494.
[22] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017.
[23] M. Mirza and S. Oshindero, "Conditional generative adversarial nets," in *Proc. CVPR*.
[24] A. Radford, L. Metz, and S. Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.
[25] L. Soler, *et al.* (2012). 3d image reconstruction for comparison of algorithm database: A patient-specific anatomical and medical image database. [Online]. Available: http://www-sop.inria.fr/geometrica/events/wam/abstract-ircad.pdf

Yuki Enokiya is a master student in information science and engineering at the Ritsmeikan University. He received his B.E. from Ritsumeikan University in 2017. His research interests are image processing, machine learning and medical image processing.

Yutaro Iwamoto is assistant professor in information science and engineering at the Ritsmeikan University. He received his B.E., M.E. and D.E. from Ritsumeikan University in 2011, 2013 and 2017, respectively. His research interests are image processing, machine learning and medical image processing.

Yen-Wei Chen is professor in information science and engineering at the Ritsmeikan University. He received his B.E., M.E. and D.E. from Kobe University and Osaka University in 1985, 1987 and 1990, respectively. His research interests are pattern recognition, image processing, machine learning and medical image processing.

Xian-Hua Han is associate professor in faculty of science at the Yamaguchi University. She received her B.E., M.E. and D.E. from Chongqing University, Shandong University and Ryukyu University in 1999, 2002 and 2005, respectively. Her research interests are pattern recognition, image processing, machine learning and super-resolution.