Difficulty-Aware Attention Network with Confidence Learning for Medical Image Segmentation

Dong Nie,[1,2] Li Wang,[2] Lei Xiang,[2,3] Sihang Zhou,[2,4] Ehsan Adeli,[5] Dinggang Shen[2]
[1]Department of Computer Science, University of North Carolina at Chapel Hill, NC 27514, USA
[2]Department of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27514, USA
[3]Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[4]College of Computer, National University of Defense Technology, Changsha, China
[5]Stanford University, Stanford, CA 94305, USA

Abstract

Medical image segmentation is a key step for various applications, such as image-guided radiation therapy and diagnosis. Recently, deep neural networks provided promising solutions for automatic image segmentation; however, they often perform good on regular samples (i.e., easy-to-segment samples), since the datasets are dominated by easy and regular samples. For medical images, due to huge inter-subject variations or disease-specific effects on subjects, there exist several difficult-to-segment cases that are often overlooked by the previous works. To address this challenge, we propose a difficulty-aware deep segmentation network with confidence learning for end-to-end segmentation. The proposed framework has two main contributions: 1) Besides the segmentation network, we also propose a fully convolutional adversarial network for confidence learning to provide voxel-wise and region-wise confidence information for the segmentation network. We relax the adversarial learning to confidence learning by decreasing the priority of adversarial learning, so that we can avoid the training imbalance between generator and discriminator. 2) We propose a difficulty-aware attention mechanism to properly handle hard samples or hard regions considering structural information, which may go beyond the shortcomings of focal loss. We further propose a fusion module to selectively fuse the concatenated feature maps in encoder-decoder architectures. Experimental results on clinical and challenge datasets show that our proposed network can achieve state-of-the-art segmentation accuracy. Further analysis also indicates that each individual component of our proposed network contributes to the overall performance improvement.

Introduction

The recent development of deep learning has largely boosted the state-of-the-art segmentation methods (Long et al. 2015; Ronneberger et al. 2015). Among them, fully convolutional networks (FCN) (Long et al. 2015), a variant of convolutional neural networks (CNN), is a recent popular choice for semantic image segmentation in both computer vision and medical image fields (Long et al. 2015; Ronneberger et al. 2015; Yu et al. 2017; Pan et al. 2017; Yang et al. 2017; Xiao et al. 2017). FCN trains neural networks in an end-to-end fashion by directly optimizing intermediate feature layers, which makes it outperform the traditional methods that often regard the feature learning and segmentation as two separate tasks. UNet (Ronneberger et al. 2015), an evolutionary variant of FCN, has achieved excellent performance by effectively combining high-level and low-level features in the network architecture. Compared to FCN, UNet can improve the localization accuracy, especially near organ boundaries.

Though being effective in most cases, the above-mentioned deep segmentation networks cannot properly handle the hard-to-segment samples (or regions) since the training of the network is inclined to be dominated by the easy-to-segment samples. This easy-to-segment sample dominance phenomenon often occurs in medical image segmentation tasks due to the irregular distribution of some medical images which may be caused by the different abnormal degree of the lesion or the imaging factors, such as different vendor devices or imaging protocols.

Several works have been proposed in the literature to address the aforementioned challenges (Shrivastava, Gupta, and Girshick 2016; Lin et al. 2017; Zhou et al. 2017). To achieve better performance on hard-to-segment (or detect) samples, (Shrivastava, Gupta, and Girshick 2016) proposed a simple strategy to automatically select hard samples for further tuning the networks. To prevent the vast number of easy samples from overwhelming the networks during training, (Lin et al. 2017) proposed focal loss for detection and achieved promising results. In another work, (Zhou et al. 2017) introduced focal loss for the biomedical image segmentation. However, the focal loss has some shortcomings when applied to medical image segmentation due to its usage of predicted probability on the samples as the hard-or-easy evaluator which could neglect the structural information and also suffer from multi-category competition issues. We argue that the widely used adversarial learning strategies may contribute to building a better evaluator.

Adversarial learning, derived from the recent popular Generative Adversarial Network (GAN) (Goodfellow et al. 2014), has achieved great success in image generation and segmentation (Goodfellow et al. 2014; Kohl et al. 2017; Nie et al. 2017; Xue et al. 2018; Zhang et al. 2017; Zhu et al. 2018). The GAN framework consists of two competing networks: a generator and a discriminator, both of which are involved in an adversarial two-player game, in which the generator aims to learn the data distribution while the discriminator estimates the probability of a sample coming from the training data or the generator. It is shown that adversarial learning can help improve the segmentation accuracy (Moeskops et al. 2017; Kohl et al. 2017); however, it is challenging to train such a GAN framework due to the difficulty of balancing the generator and discriminator (i.e., since discriminator has an easier job compared to the generator, we may face vanishing gradient for the generator) (Goodfellow et al. 2014; Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017; Mao et al. 2017). Though various methods have been proposed to solve this problem (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017; Mao et al. 2017), this issue has been alleviated but still not solved (Mescheder, Geiger, and Nowozin 2018).

To overcome such issues, we propose a difficulty-aware attention mechanism based on confidence learning for medical image segmentation. Our framework is composed of two subnetworks: 1) segmentation network and 2) confidence network. Specifically, apart from the segmentation network, we propose a fully convolutional confidence learning scheme (i.e., using confidence network), which is inspired by the concept of adversarial learning, to learn how well the local regions are segmented (i.e., the confidence map generated by the confidence network can provide us the trustworthy and untrustworthy regions in the segmented label map from the segmentation network). Based on the confidence map, we propose a difficulty-aware attention mechanism to adaptively assign region-level and voxel-level importance for training the network. Since we can adopt a difficulty-aware mechanism to further train the segmentation network, the easy-sample dominance issue can be alleviated accordingly. Our proposed algorithm has been applied to several medical image segmentation tasks, such as prostate segmentation, which is critical for guiding both biopsy and cancer radiation therapy, and brain tissue segmentation, which can help diagnose the brain lesions. Experimental results indicate that our proposed algorithm can significantly improve the segmentation accuracy, compared to other state-of-the-art methods. In addition, our proposed *fully convolutional confidence learning* and *difficulty-aware attention mechanism* strategies are proved to be effective.

To summarize, we propose a novel difficulty-aware attention mechanism to overcome the limitations of training for FCN (or UNet) in medical image segmentation tasks. Specifically, our proposed method has two main contributions over FCN (or UNet):

1) We apply a fully convolutional adversarial network to provide voxel-wise and region-wise confidence information for the segmentation network. More importantly, we relax the adversarial learning to confidence learning, which can alleviate the training imbalance problem for the supervised generative adversarial network.

2) With confidence learning, we propose a difficulty-aware mechanism to largely alleviate the overwhelming effect of easy samples during training networks, which goes beyond the shortcomings of focal loss for medical image segmentation. Experiments on several clinical datasets and ablation studies demonstrate the effectiveness of our proposed method.

Method

As mentioned in the introduction, the proposed method consists of two sub-networks, i.e., 1) segmentation network (denoted as S) and 2) confidence network (denoted as D). The architecture of our proposed framework is presented in Fig. 1, in which we conduct the fully convolutional confidence learning to avoid the training imbalance of GAN and design the difficulty-aware mechanism to alleviate the easy-sample dominance issue for training the segmentation network.

To ease the description of the proposed algorithm, we first give the formal notation used throughout the paper. Given a labeled input image $\mathbf{X} \in R^{H\times W\times T}$ with corresponding ground-truth label map $\mathbf{Y} \in Z^{H\times W\times T}$, we encode it to one-hot format $\mathbf{P} \in R^{H\times W\times T\times C}$ (by converting the label map Y into C binary label maps with one-hot encoding), where C is the number of semantic categories in the dataset. The segmentation network outputs the class probability maps $\widehat{\mathbf{P}} \in R^{H\times W\times T\times C}$. The segmented label map can be obtained by $\widehat{\mathbf{Y}} = \arg\max \widehat{\mathbf{P}}$.

In the following subsections, we first introduce the segmentation network. Then, we describe the confidence network with fully convolutional adversarial learning, followed by the difficulty-aware attention mechanism. Finally, we describe the implementation details.

Segmentation Network

As shown in Fig. 1, the segmentation network can be any end-to-end segmentation network, such as FCN (Long et al. 2015), UNet (Ronneberger et al. 2015), VNet (Milletari et al. 2016), or DSResUNet (Yu et al. 2017) (a UNet-like structure with residual learning, element-wise addition of skip connection, and deep supervision). In this paper, we adopt an enhanced UNet as the segmentation network. Specifically, we replace all the convolutional layers but the last one with the residual modules (He et al. 2016), apply dilated residual module in the intermedia layers between encoder and decoder (the feature maps with the smallest size) (Yu, Koltun, and Funkhouser), utilize the transformation modules in the long skip connections (Nie et al. 2018), inject deep supervision at three scales in the decoder path (Merkow et al. 2016), and propose channel attention module to better fuse the concatenated information from lower layers and higher layers (Hu, Shen, and Sun).

Training segmentation network with hybrid loss: The class imbalance problem is usually serious in medical image segmentation tasks. To overcome it, we propose using a generalized multi-class Dice loss (Sudre et al. 2017) as the training loss for our segmentation network, as defined below in Eq. (1):

$$L_{Dice}(\mathbf{X},\mathbf{P};\theta_{\mathbf{S}}) = 1 - 2\frac{\sum_{c=1}^{C}\pi_c \sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T} P_{h,w,t,c}\widehat{P}_{h,w,t,c}}{\sum_{c=1}^{C}\pi_c \sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T} P_{h,w,t,c} + \widehat{P}_{h,w,t,c}}, \tag{1}$$

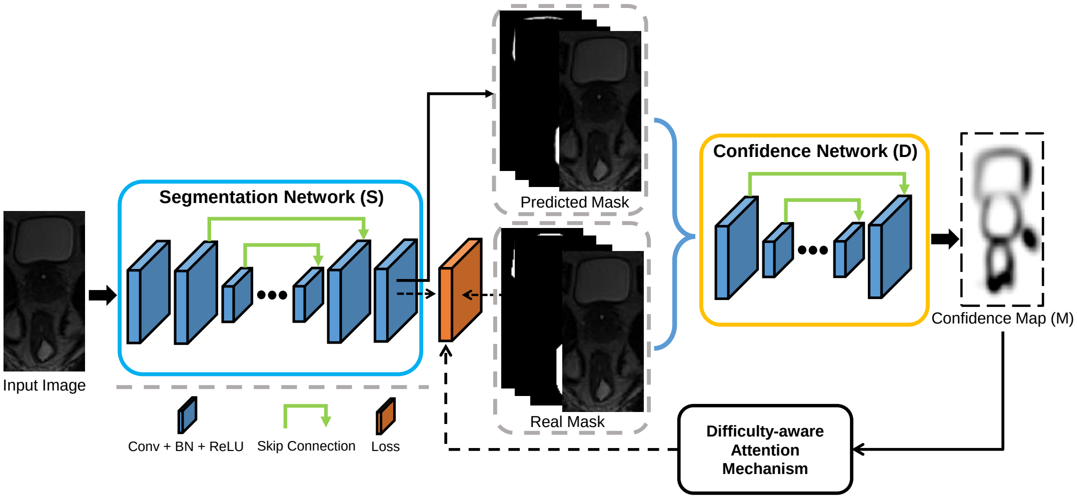

Figure 1: Illustration of the architecture of the proposed framework. This framework consists of a segmentation network (S), a confidence network (D), and the difficulty-aware attention mechanism. Note, we pursue *a perfect* D in this framework.

where π_c is the class balancing weight of category c, and $\theta_{\mathbf{S}}$ contains the parameters of segmentation network. We set $\pi_c = 1/\left(\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T} P_{h,w,t,c}\right)^2$. $\widehat{\mathbf{P}}$ is the predicted probability maps from the segmentation network: $\widehat{\mathbf{P}} = S\left(\mathbf{X}, \theta_{\mathbf{s}}\right)$.

Besides, we also use the multi-category cross entropy loss to form the voxel-wise measurement, as shown in Eq. (2):

$$L_{CE}\left(X, Y; \theta_S\right) = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T}\sum_{c=1}^{C} I\left\{Y_{h,w,t,c}\right\} \log \widehat{P}_{h,w,t,c} \tag{2}$$

To this end, the hybrid loss which leverages both losses for training the segmentation network can be concluded as in Eq. (3):

$$L_{Hyb} = L_{Dice} + L_{CE} \tag{3}$$

Fully Convolutional Confidence Learning

Adversarial learning has been shown to be effective in improving the segmentation network (Luc et al. 2016; Moeskops et al. 2017; Hung et al. 2018). More importantly, it can provide a better hard-easy sample evaluator with proper adjustment. Thus, we decide to incorporate adversarial learning in our architecture to further improve the segmentation network.

In the classical adversarial networks, the discriminator is mostly a CNN-based network with the output probability of an input image belonging to be the real (Sabokrou et al. 2018). Obviously, the conventional discriminator only provides a global confidence over the entire image domain, without providing any confidence at the local region, e.g., voxel-wise confidence. To address this issue, we propose using an FCN-based network to model the discriminator and name it as confidence network. The output of confidence network is called as confidence map (M) with size $H \times W \times T \times 1$, which indicates locally whether automatic segmentation is similar to the ground-truth segmentation. We argue that the confidence network can learn the structural information that can be used to regularize the output of segmentation network (Hung et al. 2018). In this paper, a simplified version of typical UNet (Ronneberger et al. 2015) is used as the architecture of confidence network. Specifically, to save memory, we only keep one convolutional layer at each stage and also half the number of feature maps in the convolution layers across the network except the last one.

However, non-convergence and model collapse issues usually occur in GAN's training, which are often explained as an imbalance between the discriminator and the generator. Though widely researched, this problem still exists in training GAN (Mescheder 2018; Mescheder, Geiger, and Nowozin 2018; Kodali et al. 2017). To avoid the imbalance, we relax the adversarial learning to confidence learning after analyzing the role of the discriminator in the GANs.

Relaxing the adversarial learning to confidence learning: We first analyze the role that the discriminator plays in traditional GANs. In classical GANs (Goodfellow et al. 2014; Radford, Metz, and Chintala 2015), the first role of the discriminator is to judge if the input is a real image or generated image by the generator. In other words, its goal is to determine how well the generated images look like the real image (for convenience, we denote it as confidence learning). Moreover, it provides adversarial learning to train the generator, and this is the most important role it plays since the GAN is to learn the distribution of the data through the generator and it is the sole source to provide supervision signals for training the generator. On the other hand, the discriminator (i.e., confidence network) in the proposed framework shown in Fig. 1 also has such two roles: to learn the confidence of how each local region is correctly segmented (i.e., confidence learning), and to provide adversarial learning to train the segmentation network (i.e., adversarial learning). However, there is the main difference between our proposed framework and the classical GANs, i.e., besides the adversarial learning from the discriminator, the segmentation network can provide strong supervision signals to train itself.

As mentioned before, the GAN framework has a big challenge of imbalance training. Since the discriminator is much easier to be perfectly trained than the generator and thus will result in few training signals for the generator from discriminator (Arjovsky, Chintala, and Bottou 2017), which will eventually lead to non-convergence and model collapse. However, it is a different situation in our case since the segmentation network has training supervision signals from itself which can provide continuous support to improve the segmentation network. To this end, we propose to relax adversarial learning to confidence learning for avoiding the training imbalance by adjusting the role of the discriminator: *we place the role of confidence learning prior to that of adversarial learning*. In other words, *we reformulate the original min-max game to a maximization of discriminator with a soft constraint over the generator*.

With this strategy, we can discover the difficulty degree of each local region of being segmented and can thus provide difficulty-aware information to guide the training of the segmentation network. To this end, the segmentation network can be further improved, which will in return boost the discriminator. As a result, the adversarial learning can be formulated as a *soft constraint* to work as a high-order potential regularization for the segmentation network.

Training the confidence network: The training objective of the confidence network is the summation of binary cross-entropy loss over the image domain, as shown in Eq. (4). Here, we use S and D to denote the segmentation and confidence networks, respectively.

$$L_D(\mathbf{X},\mathbf{P};\theta_\mathbf{D}) = L_{BCE}(D(\mathbf{P},\theta_\mathbf{D}),\mathbf{1}) + L_{BCE}(D(S(\mathbf{X}),\theta_\mathbf{D}),\mathbf{0}), \tag{4}$$

where

$$\begin{aligned} L_{BCE}\left(\widehat{\mathbf{Q}},\mathbf{Q}\right) = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T} Q_{h,w,t}\log\left(\widehat{Q}_{h,w,t}\right) \\ + (1-Q_{h,w,t})\log\left(1-\widehat{Q}_{h,w,t}\right) \end{aligned} \tag{5}$$

where $\mathbf{X}$ and $\mathbf{P}$ represent the input data and its corresponding manual label map (one-hot encoding format), respectively. $\theta_\mathbf{D}$ is network parameters for the confidence network.

Adversarial loss of the segmentation network: For segmentation network, besides the hybrid loss as defined in Eq. (3), there is another loss from D used as "variational" regularization term working as a soft constraint, which aims at enforcing higher-order consistency between ground-truth segmentation and automatic segmentation. In particular, the adversarial loss ("ADV") to improve S and fool D can be defined by Eq. (6):

$$L_{ADV}\left(\mathbf{X},\theta_\mathbf{S}\right) = L_{BCE}\left(D\left(S\left(\mathbf{X};\theta_\mathbf{S}\right)\right),\mathbf{1}\right) \tag{6}$$

Difficulty-Aware Attention Mechanism

Focal loss has been shown effective to alleviate the overwhelming effect of easy samples in many computer vision tasks, such as image detection and segmentation (Lin et al. 2017; Zhou et al. 2017). The success of focal loss can be attributed to its strategy that it tries to pay more attention on the recognized hard samples (regions) and less attention to the easy ones. The key point is how to recognize difficult samples (regions). Focal loss utilizes the predicted probability of a sample as the indicator of the difficulty degree, which may lead to some potential problems in medical image segmentation tasks. Firstly, training may be unstable due to the dominance of a certain class. Secondly, easy and hard samples may also have similar focal weights due to the potential multi-class competition. Thirdly, focal loss only provides voxel-level attention and ignores region-level attention. Lastly, only predicted mask may not really indicate the hard regions without considering the original input image of the segmentation network. These potential problems are mostly caused by the fact that the focal loss uses predicted probability from the segmentation network as the standard to determine whether it is a hard or easy sample. To overcome the above-mentioned problems, we would prefer *a more professional easy-or-hard representer*.

The previously described confidence learning provides us with a solution to better recognize the easy-or-hard samples. The confidence map produced by the confidence network contains the easy-or-hard information. Also, since confidence network is actually a binary classification model, it will avoid the multi-category competition issue. More importantly, the confidence map contains information from both the original input image and predicted probability mask, and thus it can provide structural information about the easy-or-hard samples (regions).

To this end, we propose a difficulty-aware attention mechanism to better represent the easy-or-hard information. Specifically, we design a difficulty-aware hybrid loss using region-level and voxel-level attentions from both predicted probability mask and confidence map.

At first, we propose an organ-level attention based generalized Dice loss to depict the region-level difficulty, which is shown in Eq. (7).

$$L_{FDice} = 1 - 2\frac{\sum_{c=1}^{C}\pi_c(1-dsc_c)^r\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T}P_{h,w,t,c}\widehat{P}_{h,w,t,c}}{\sum_{c=1}^{C}\pi_c(1-dsc_c)^r\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T}P_{h,w,t,c}+\widehat{P}_{h,w,t,c}} \tag{7}$$

where dsc_c is the average Dice similarity coefficient of a specific category c, e.g., a certain organ or tissue. γ is the organ-level attention parameter with a range of $[0,5]$. Following (Lin et al. 2017), we set γ to 2 in this paper.

The voxel-level difficulty-aware attention from the confidence map (M) is formulated (based on Eq. 2) in Eq. (8):

$$\begin{aligned} L_{FCE}\left(X,Y;\theta_S\right) = \\ -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T}\sum_{c=1}^{C} I\left\{Y_{h,w,t},c\right\}F_{h,w,t}\log\widehat{P}_{h,w,t,c} \end{aligned} \tag{8}$$

where

$$F = (1-M)^\beta \tag{9}$$

where β is the voxel-level attention parameter, and it follows the settings of γ as described above.

Now we can define the difficulty-aware attention mechanism with the hybrid loss as Eq. (10).

$$L_{DamHyb} = L_{FDice} + L_{FCE} \tag{10}$$

With the difficulty-aware hybrid loss in Eq. (10), we can pay more attention in the lower confidently (hard) segmented regions. Note, it is different from focal loss which is defined based on the probability map (P) from the segmentation network.

Total loss for segmentation network

By summing the above losses, the total loss to train the segmentation network can be defined by Eq. (11).

$$L_Seg = L_{DamHyb} + \lambda_1 L_{ADV} \tag{11}$$

where λ_1 is the scaling factor for the regularization term of adversarial learning. It is selected as a very small value (i.e., 0.005 in our case) since it works as soft constraint.

Implementation Details

Pytorch[1] is adopted to implement our proposed framework shown in Fig. 1. Since we desire a perfect discriminator (D), we do not adopt the traditionally used strategies to limit the D (Radford, Metz, and Chintala 2015). We adopt Adam algorithm to optimize the networks. The input size of the segmentation network is $64 \times 64 \times 16$. The network weights are initialized by the Xavier algorithm (Glorot and Bengio 2010) and weight decay is set to be 1e-4. For the network biases, we initialize them to 0. The learning rates for the segmentation network and the confidence network are both initialized to 5e-3, followed by decreasing the learning rate 2 times for the S, and 5 times for the D every 3 epochs during the training. A Titan X GPU server is utilized to train the networks.

Experiments

To evaluate the proposed method, we apply our algorithm on three different datasets. The first dataset is our own pelvic dataset and the other two are both publicly available challenge datasets which will be introduced in later subsections.

The pelvic dataset consists of 50 prostate cancer patients from a Cancer Hospital, each with one T2-weighted MR image and corresponding manually-labeled map by a medical expert. The images were acquired with 3T magnetic field strength, while different patients were scanned with different MR image scanners (i.e., Siemens Medical Systems and Philips Medical Systems). Under such a situation, the challenge for the segmentation task increases since both shape and appearance differences are large. The prostate, bladder, and rectum in all MRI scans have been manually segmented, which serve as the ground truth for evaluating our segmentation method. The image size is mostly $256 \times 256 \times (120 \sim 192)$, and the voxel size is mainly $1 \times 1 \times 1$ mm^3.

Five-fold cross-validation is used to evaluate our method. Specifically, in each fold of cross-validation, we randomly chose 35 subjects as the training set, 5 subjects as the validation set, and the remaining 10 subjects as the testing set. We use sliding windows to go through the whole MRI for prediction for a testing subject. Unless explicitly mentioned, all the reported performance by default is evaluated on the testing set. As for evaluation metrics, we utilize Dice Similarity Coefficient (DSC) and Average Surface Distance (ASD) to measure the agreement between the manually and automatically segmented label maps.

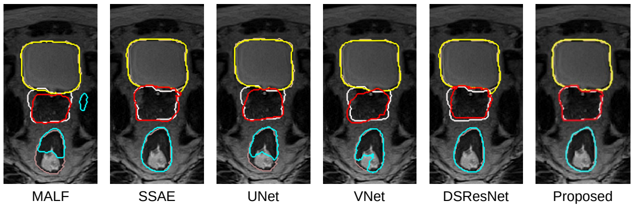

Figure 2: Pelvic organ segmentation results of a typical subject by different methods. Orange, silver and pink contours indicate the manual ground-truth segmentations, and yellow, red and cyan contours indicate automatic segmentations.

Comparison with state-of-the-art methods

To demonstrate the advantage of our proposed method, we compare our method with other five widely-used methods on the same dataset as shown in Table 1: 1) multi-atlas label fusion (MALF), 2) SSAE (Guo et al. 2016), 3) UNet (Ronneberger et al. 2015), 4) VNet (Milletari et al. 2016), and 5) DSResUNet (Yu et al. 2017). Also, we present the performance of our proposed method.

Table 1 quantitatively compares our method with the five state-of-the-art segmentation methods. We can see that our method achieves better accuracy than the five state-of-the-art methods in terms of both DSC and ASD, especially for the prostate and rectum which are believed more difficult to segment. The VNet works well in segmenting bladder and prostate, but it cannot work very well for rectum (which is often more challenging to segment due to the long and narrow shape). Compared to UNet, DSResUNet improves the accuracy by a large margin, indicating that residual learning and deep supervision bring performance gain. We also visualize some typical segmentation results in Fig. 2, which further show the superiority of our proposed method.

Impact of the Difficulty-aware Attention Mechanism

As mentioned in the Method Section, we propose an enhanced UNet with several widely used techniques injected, and we further propose a difficulty-aware attention mechanism to assign different importance for different samples (regions) so that the network can concentrate on hard-to-segment examples and thus avoid dominance by easy-to-segment samples. We visualize the performance comparison among the basic UNet, enhanced UNet (enUNet) and the one with difficulty-aware attention mechanism (enUNet+dam) in Fig. 3. (Note, we use the hybrid loss to

[1] https://github.com/pytorch/pytorch

Table 1: DSC and ASD on the pelvic dataset by different methods.

Method	DSC (%)			ASD (in mm)		
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
MALF	86.69(6.81)	79.28(8.72)	76.43(11.88)	1.641(.360)	2.791(.930)	3.210(2.112)
SSAE	91.75(3.10)	87.07(4.24)	86.38(4.41)	1.089(.231)	1.660(.490)	1.701(.412)
UNet	89.57(2.83)	82.22(5.88)	81.04(5.31)	1.214(.216)	1.917(.645)	2.186(0.850)
VNet	92.61(1.84)	86.40(3.61)	83.16(4.12)	1.023(.186)	1.725(.457)	1.969(.449)
DSResUNet	94.43(.90)	88.24(2.01)	86.91(3.24)	.914(.168)	1.586(.358)	1.586(.405)
Proposed	**97.48(.65)**	**92.11(1.70)**	**91.05(2.47)**	**.850(.146)**	**1.297(.276)**	**1.387(.346)**

Table 2: Quantitative comparison between our proposed method and other methods on the prostate challenge testing dataset.

Method	DSC (%)			ASD (in mm)			95HD			aRVD			Score(std)
	whole	base	apex	whole	base	apex	whole	base	apex	whole	base	apex	
pxl_mcg	91.23	89.07	88.54	1.60	1.76	1.57	4.47	4.48	3.64	2.08	-0.07	2.23	88.98(3.41)
Isensee	91.61	90.29	88.05	1.52	1.65	1.64	4.21	4.20	3.85	3.42	1.86	3.48	88.84(2.94)
whu_mlgroup(2)	91.42	89.41	88.51	1.54	1.79	1.57	4.21	4.88	3.82	5.27	4.00	6.43	88.72(4.36)
Proposed	90.12	88.95	87.71	1.84	1.73	1.68	5.36	4.43	3.99	4.99	2.19	6.65	88.28(3.02)
tbrosch	90.46	88.51	85.29	1.70	1.91	1.90	4.91	5.04	4.57	2.14	7.22	-4.93	87.24(4.46)
whu_mlgroup(1)	90.26	89.15	88.36	1.86	1.79	1.62	5.57	4.83	3.90	9.74	10.73	9.64	87.04(5.79)
AutoDenseSeg	90.14	88.09	86.79	1.83	1.94	1.79	5.36	5.13	4.32	4.53	5.19	2.05	87.19(4.25)
CUMED	89.43	86.42	86.81	1.95	2.13	1.74	5.54	5.41	4.29	6.95	11.04	15.18	86.65(4.42)
SCIRESU	90.24	88.98	83.30	1.74	1.81	2.11	4.93	4.51	5.34	6.01	8.18	-7.33	86.41 (3.49)
QUILL(M2)	88.81	87.39	85.46	1.97	2.01	1.91	5.29	5.07	4.35	6.97	4.76	5.85	85.93(4.97)

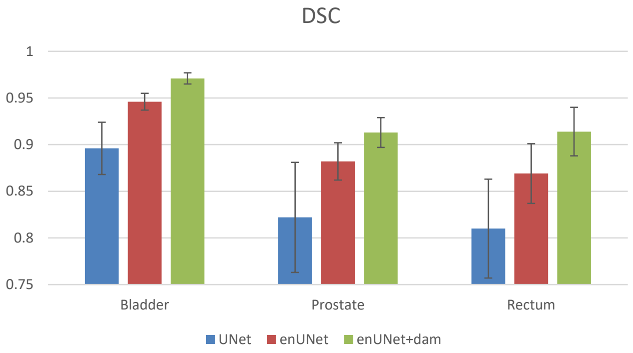

Figure 3: Average Dice ratios of different methods.

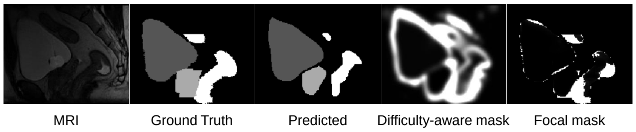

Figure 4: Visualization of the difficulty-aware mask and the focal mask, obtained after training the network for 5 epochs.

train UNet and enUNet). Actually, in our case, the widely used techniques injected to the basic UNet contribute most to the performance gain. The effectiveness of difficulty-aware attention mechanism is also confirmed by the improved performance as shown in Fig. 3. It is worth noting that our proposed difficulty-aware attention mechanism contributes more performance gain for prostate and rectum compared with the bladder. It is consistent with our assumption that difficulty-aware attention mechanism could pay more attention to difficult samples (regions) and thus can handle difficult samples (regions) much better.

Comparing with the Focal Loss

Since our proposed difficulty-aware attention mechanism is designed based on the focal loss, it is necessary to investigate the difference of the proposed module against focal loss for medical image segmentation.

To better understand the two strategies, we firstly visualize the difficulty-aware mask (i.e., $(1-M)$) and the focal mask (i.e., $\left(1-\widehat{P}\right)$) in Fig. 4. The focal mask mainly focuses on the regions with low predicted probability from segmentation network which needs more attention. Since it is directly related with predicted probability map, it can reflect the difficult regions more precisely in *voxel-level*. On the contrary, difficulty-aware mask reflects the difficulty regions in a more *structured* manner, in which it focuses more on the regions with lower confidence ratios from confidence network. The reason behind it is that we have a professional hard-or-easy recognizer: The D can represent the input containing both the predicted probability mask from segmentation network and the original input image by confidence learning so that we can have a more expert hard-or-easy representation, as expressed in Eq. (12):

$$M = D(\widehat{P} \cup X) \tag{12}$$

where $\cup$ denotes the concatenation operation.

We further conducted experiments with these different strategies to segment the prostate only, since the prostate

Table 3: Comparison of different strategies to segment prostate on the pelvic dataset in terms of DSC (%).

Method	Base	Middle	Apex
enUNet	86.70(4.91)	87.91(4.83)	83.92(5.87)
enUNet+Focal	88.24(4.53)	89.21(3.20)	86.83(4.90)
enUNet+Hybrid	88.12(4.19)	90.08(2.70)	86.71(5.47)
Proposed	89.41(3.68)	90.90(2.37)	88.21(4.14)

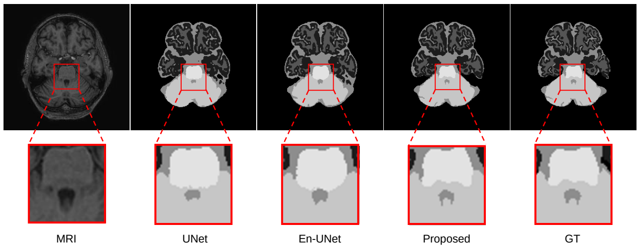

Figure 5: Comparison of segmentation results with different methods and the manual ground-truth on a sample subject.

is traditionally thought to be hard to segment. To make a fair comparison, we use the same architecture (enUNet) as the basis to conduct the experiments. Due to computational times, we only do a two-fold cross-validation for these comparison experiments. To better depict the difficult parts of the prostate, we partition the prostate into three parts: apex (first 1/3 of the prostate volume), base (last 1/3 of the prostate volume) and middle (the rest). The performance of the enUNet with different strategies is listed in Table 3.

As described in Table 3, the focal loss can help improve the performance, especially for the base and apex parts of the prostate, since it pays more attention to the hard voxels. The hybrid loss described in Eq. (3) can achieve similar performances with the focal loss since the hybrid loss can capture the organ structure as well as the voxel-level information. The proposed method (difficulty-aware attention mechanism) achieves the largest performance gain, since it can not only capture the difficult regions in a structured way but also absorb the advantage of the hybrid loss. This demonstrates that the proposed difficulty-aware attention mechanism can work better than the focal loss in medical image segmentation tasks.

Validation on MR Brain Challenge Dataset

We further validate our proposed method on MR Brain dataset[2]. This dataset contains 7 subjects, each with T1 MRI, Flair and manually labeled ground truth map. The task is to segment each voxel into one of the following (tissue) types: background, cortical gray matter (CGM), basal ganglia (BG), white matter (WM), WM lesion (WML), cerebrospinal fluid in the extracerebral space (CSF), ventricle (V), cerebellum (C), brain stem (BS), infarction, and other.

We conduct the experiment in a leave-one-out manner. We visualize one typical slice of a sample in Fig. 5 to make a qualitative comparison. The proposed method can capture better contour which is usually considered as hard regions compared with the UNet and enUNet; this again proves the effectiveness of our proposed method. The quantitative comparison (the proposed mechanism can improve the average performances by about 3.5% in terms of DSC) also indicates the success of the proposed modules.

Validation on Prostate Challenge Dataset

We also evaluate our proposed method on the prostate segmentation challenge dataset whose ground-truth label maps are hidden from the participants. The official evaluation metrics used in this challenge include the DSC, the average over the shortest distance between the boundary (surface) points of the volumes (ABD or ASD), the percentage of the absolute difference between the volumes (aRVD), and the 95% Hausdorff distance (95HD). It is worth noting that the organizers not only calculate the evaluation metrics on the whole prostate, but also on the apex and base parts of the prostate that are believed to be the most difficult regions for segmentation. In addition, an overall score (shown in the last column) combining the above-mentioned evaluation metrics is also provided to rank the submitted methods (please refer to (Litjens et al. 2014) for the details about the evaluation metrics).

The quantitative results of our method and our competitors are shown in Table 2. (Note, the results were directly obtained from the organizers). Currently, there are more than 150 teams successfully submitting their results and listed in the leaderboard. Note we only list top 10 teams in the Table for convenience, and please refer the whole leaderboard through this link[3]. Our proposed method ranks 4th in terms of the overall score among all the participants. It is worth noting that the top 3 methods all ensemble their results from different deep networks. In contrast, our submission is a single model as presented in this paper. More importantly, our proposed method presents a much lower standard deviation value compared to the other top 8 methods. (Note, the lowest standard deviation comes from the 2nd ranked team who ensembles results from 20 deep networks), which further indicates the effectiveness and robustness of our proposed method.

More importantly, our proposed method achieves a very competitive performance on the base and apex parts which are thought to be the most difficult segmented regions, and it further proves that our designed difficulty-aware attention mechanism indeed contributes to the gain of performance.

Conclusions

In this paper, we presented a novel difficulty-aware attention deep networks to segment medical images. Specifically, we proposed fully convolutional confidence learning to relax the adversarial learning so that we can largely alleviate the training imbalance between discriminator and generator, and the discriminator can thus provide wonderful confidence information. Based on that, difficulty-aware attention mechanism was proposed to effectively address the easy-to-segment sample dominance issue in a more structured way,

[2] http://mrbrains18.isi.uu.nl/

[3] https://promise12.grand-challenge.org/evaluation/results/

which goes beyond the shortcomings of focal loss for training medical image segmentation networks. By integrating these components into the framework, our proposed framework achieved significant improvement in terms of both accuracy and robustness on three datasets.

Acknowledgments. This work was supported by the National Institutes of Health under Grant R01 CA206100.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 249–256.

Goodfellow et al., I. 2014. Generative adversarial nets. In *NIPS*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *NIPS*, 5767–5777.

Guo et al., Y. 2016. Deformable mr prostate segmentation via deep feature learning and sparse patch matching. *IEEE TMI* 35:1077–1089.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hu, J.; Shen, L.; and Sun, G. Squeeze-and-excitation networks.

Hung, W.-C.; Tsai, Y.-H.; Liou, Y.-T.; Lin, Y.-Y.; and Yang, M.-H. 2018. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*.

Kodali, N.; Abernethy, J.; Hays, J.; and Kira, Z. 2017. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*.

Kohl et al., S. 2017. Adversarial networks for the detection of aggressive prostate cancer. *arXiv preprint arXiv:1702.08014*.

Lin et al., T.-Y. 2017. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*.

Litjens et al., G. 2014. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *MedIA* 18(2):359–373.

Long et al., J. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Luc, P.; Couprie, C.; Chintala, S.; and Verbeek, J. 2016. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Smolley, S. P. 2017. Least squares generative adversarial networks. In *ICCV*, 2813–2821. IEEE.

Merkow, J.; Marsden, A.; Kriegman, D.; and Tu, Z. 2016. Dense volume-to-volume vascular boundary detection. In *MICCAI*, 371–379. Springer.

Mescheder, L.; Geiger, A.; and Nowozin, S. 2018. Which training methods for gans do actually converge? In *ICML*, 3478–3487.

Mescheder, L. 2018. On the convergence properties of gan training. *arXiv preprint arXiv:1801.04406*.

Milletari et al., F. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 565–571. IEEE.

Moeskops et al., P. 2017. Adversarial training and dilated convolutions for brain mri segmentation. *arXiv preprint arXiv:1707.03195*.

Nie, D.; Trullo, R.; Lian, J.; Petitjean, C.; Ruan, S.; Wang, Q.; and Shen, D. 2017. Medical image synthesis with context-aware generative adversarial networks. In *MICCAI*, 417–425. Springer.

Nie, D.; Wang, L.; Adeli, E.; Lao, C.; Lin, W.; and Shen, D. 2018. 3-d fully convolutional networks for multimodal isointense infant brain image segmentation. *IEEE Transactions on Cybernetics*.

Pan, T.; Wang, B.; Ding, G.; and Yong, J.-H. 2017. Fully convolutional neural networks with full-scale-features for semantic segmentation.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Ronneberger et al., O. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Sabokrou, M.; Pourreza, M.; Fayyaz, M.; Entezari, R.; Fathy, M.; Gall, J.; and Adeli, E. 2018. Avid: Adversarial visual irregularity detection. *ACCV*.

Shrivastava, A.; Gupta, A.; and Girshick, R. 2016. Training region-based object detectors with online hard example mining. In *CVPR*, 761–769.

Sudre et al., C. H. 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *DLMIA*. Springer.

Xiao, H.; Wei, Y.; Liu, Y.; Zhang, M.; and Feng, J. 2017. Transferable semi-supervised semantic segmentation. *arXiv preprint arXiv:1711.06828*.

Xue, Y.; Xu, T.; Zhang, H.; Long, L. R.; and Huang, X. 2018. Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics* 1–10.

Yang, X.; Yu, L.; Wu, L.; Wang, Y.; Ni, D.; Qin, J.; and Heng, P.-A. 2017. Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images. In *AAAI*, 1633–1639.

Yu et al., L. 2017. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *AAAI*.

Yu, F.; Koltun, V.; and Funkhouser, T. A. Dilated residual networks.

Zhang, Y.; Yang, L.; Chen, J.; Fredericksen, M.; Hughes, D. P.; and Chen, D. Z. 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *MICCAI*, 408–416. Springer.

Zhou, X.-Y.; Shen, M.; Riga, C.; Yang, G.-Z.; and Lee, S.-L. 2017. Focal fcn: Towards small object segmentation with limited training data. *arXiv preprint arXiv:1711.01506*.

Zhu, W.; Xiang, X.; Tran, T. D.; Hager, G. D.; and Xie, X. 2018. Adversarial deep structured nets for mass segmentation from mammograms. In *ISBI*, 847–850. IEEE.