Published in final edited form as:

Int J Comput Vis. 2020 November ; 128(10-11): 2494–2513. doi:10.1007/s11263-020-01321-2.

Adversarial Confidence Learning for Medical Image Segmentation and Synthesis

Dong Nie[1,2], **Dinggang Shen**[2,3]

[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]Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea

Abstract

Generative adversarial networks (GAN) are widely used in medical image analysis tasks, such as medical image segmentation and synthesis. In these works, adversarial learning is directly applied to the original supervised segmentation (synthesis) networks. The usage of adversarial learning is effective in improving visual perception performance since adversarial learning works as realistic regularization for supervised generators. However, the quantitative performance often cannot improve as much as the qualitative performance, and it can even become worse in some cases. In this paper, we explore how we can take better advantage of adversarial learning in supervised segmentation (synthesis) models and propose an adversarial confidence learning framework to better model these problems. We analyze the roles of discriminator in the classic GANs and compare them with those in supervised adversarial systems. Based on this analysis, we propose adversarial confidence learning, i.e., besides the adversarial learning for emphasizing visual perception, we use the confidence information provided by the adversarial network to enhance the design of supervised segmentation (synthesis) network. In particular, we propose using a fully convolutional adversarial network for confidence learning to provide voxel-wise and region-wise confidence information for the segmentation (synthesis) network. With these settings, we propose a difficulty-aware attention mechanism to properly handle hard samples or regions by taking structural information into consideration so that we can better deal with the irregular distribution of medical data. Furthermore, we investigate the loss functions of various GANs and propose using the binary cross entropy loss to train the proposed adversarial system so that we can retain the unlimited modeling capacity of the discriminator. Experimental results on clinical and challenge datasets show that our proposed network can achieve state-of-the-art segmentation (synthesis) accuracy. Further analysis also indicates that adversarial confidence learning can both improve the visual perception performance and the quantitative performance.

Publisher's Disclaimer: This Author Accepted Manuscript is a PDF file of an unedited peer-reviewed manuscript that has been accepted for publication but has not been copyedited or corrected. The official version of record that is published in the journal is kept up to date and so may therefore differ from this version.

Keywords

Adversarial Confidence Learning; Medical Image Analysis; Segmentation; Image Synthesis

1 Introduction

Generative Adversarial Network (GAN) [4] is currently a very popular and successful unsupervised model that can generate samples following an implicit distribution. 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. Adversarial learning, derived from GAN [4], has been widely applied to the supervised models (such as segmentation and generation models) with purpose of enhancing models' capacity and achieved great success in image generation and segmentation [4, 35, 9, 42, 58, 63, 68]. Many works have demonstrated that adversarial learning can contribute to generate much more perceptive images or videos [4, 35, 42, 27], in which the generation can even fool human. It is also shown that adversarial learning can help improve the segmentation performance, for instance, fixing the obvious segmentation errors [8, 9, 45]. However, the performance gain brought by adversarial learning is usually inconsistent (or limited) across different metrics, for instance, the generated images are becoming much more realistic, while the performance in terms of quantitative metric cannot have an obvious improvement and may even become worse [27, 28, 45]. Moreover, it is quite 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 problem of vanishing gradient for the generator) [4, 12, 20, 34]. Though various methods have been proposed to solve this problem [12, 20, 34], this issue has been alleviated but still not solved [38, 33]. Besides, mode collapse phenomenon occurs quite often in practice when training GAN systems [39].

To alleviate such issues, we first conduct an analysis for the roles of discriminators in the classic GANs and make a comparison with those in supervised adversarial systems. In particular, the generator (G) is updated only through the gradients from the discriminator (D); in traditional supervised adversarial systems, the generator is trained by supervised loss and the adversarial loss. However, in both cases, the balance between G and D should be well kept during training, and the mode collapse issue should be avoided. Moreover, though the adversarial learning could significantly improve the visual perception, it cannot improve quantitative performance at a similar ratio [27, 28]. Based on this analysis, we propose another learning scheme, adversarial confidence learning, to replace the adversarial learning in the supervised adversarial systems, i.e., besides using the adversarial gradient as in the classic GANs, we also rely on the confidence information provided by the discriminator to improve the supervised generator. Since the training of generator largely relies on the confidence information provided by the discriminator, our learning scheme can reduce the dependency for the adversarial learning and can thus alleviate (or avoid) the common issues in training GANs. In particular, we propose a difficulty-aware attention mechanism based on confidence learning for medical image segmentation (synthesis). Specifically, apart from the segmentation network, we propose a fully convolutional adversarial network to work as confidence network to learn how well the local regions are segmented or synthesized (i.e., the confidence map generated by the confidence network can provide us the trustworthy and untrustworthy regions in the segmented (synthesized) label map from the segmentation (synthesis) network). Based on the confidence map, we propose a difficulty-aware attention mechanism to adaptively assign region-level and voxel-level importance to improve the design of the supervised segmentation (synthesis) network. Since we can adopt a difficulty-aware mechanism to further enhance the segmentation network, the easy-sample dominance issue can be alleviated accordingly. In our proposed framework, the visual perception performance gain is mainly coming from the adversarial learning, and the quantitative performance improvement is mainly from the improved design of the supervised generator. As a consequence, our proposed system is no longer vulnerable to training imbalance between generator and discriminator. Besides, we also investigate the loss functions for the confidence network, i.e., to guarantee a powerful discriminator, we do a survey on various objective functions for the adversarial learning and further propose using binary cross entropy loss as in the original classic GAN, instead of using the widely adopted Wasserstein distance [12]. To this end, our proposed adversarial confidence learning framework can take better advantage of adversarial learning but avoid or alleviate the drawbacks of the adversarial learning. Our proposed algorithm has been applied to several medical image segmentation and synthesis tasks, such as prostate segmentation, which is critical for guiding both biopsy and cancer radiation therapy, and brain tumor image cross-modality synthesis, which can help diagnose the brain lesions. Experimental results indicate that our proposed algorithm can improve not only the visual perception performance but also the quantitative segmentation (synthesis) 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 an adversarial confidence learning framework to overcome the limitations of adversarial learning in supervised models, such as adversarial learning for deep network based medical image segmentation (synthesis) tasks. Specifically, our proposed method has three main contributions over the adversarial supervised systems:

1. We conduct an analysis for the roles of discriminator in the classic GANs and make a comparison with those in the supervised adversarial systems. We further conduct experiments to certify that adversarial learning is effective in providing realistic constraint. With the analysis and experiments, we propose using a fully convolutional adversarial network as a discriminator to provide voxel-wise and region-wise confidence information for the segmentation (synthesis) network. We argue the confidence information estimated by our proposed method is a better choice than the traditionally used predicted softmax probabilities.

2. With dense confidence information, we propose a difficulty-aware mechanism to alleviate the overwhelming effect of easy samples during the phase of training the networks. Since our attention mechanism considers the structure and neighborhood information, our method can mitigate shortcomings of focal loss for medical image segmentation (synthesis). Experiments on several clinical

datasets and ablation studies demonstrate the effectiveness of our proposed method.

3. We further explore the loss functions for the confidence network. We argue that the binary cross-entropy loss is a reasonable choice because it could provide a powerful discriminator.

A preliminary version of this work has been presented at a conference [46], but we extend this work in a large degree. Herein, we (i) propose to analyze the roles of discriminators in classic GANs and compare with those in supervised adversarial learning systems, (ii) analyze and certify the raw adversarial learning work as realistic regularization in supervised models (we have asked medical experts to choose the desired segmentations and synthesized image to validate the realistic regularization), (iii) extend the adversarial confidence learning framework to medical image synthesis task and have achieved great success in lesion medical image synthesis, (iv) analyze the selection of loss functions for the discriminator in our proposed framework, and (v) include more literature overview and additional discussions that are not in the conference publication.

2 Related Work

GAN:

GANs [4] are efficient unsupervised models that can generate samples following an implicit distribution given a set of data. Radford et al. [50] explored unsupervised learning with CNN and introduced a class of CNNs called as deep convolutional GANs (DCGANs) with certain architectural constraints to be strong candidates for unsupervised learning. To solve the gradient vanishing issues in GANs, Arjovsky et al. [12] proposed Wasserstein distance as a metric to measure how close the generated distribution and the real distribution are. Many other works are also proposed to solve or alleviate this problem [33, 34, 49]. Metz [39] proposed the unrolled GAN to stabilize the training of GAN and mitigate the mode collapse phenomenon.

Adversarial learning has been also widely extended to supervised models, such as image segmentation and synthesis. Isola et al. [27] used conditional GAN on image-to-image translation problems, in which, a L_1 or L_2 loss is also applied to train the generator and this supervised loss can guarantee the correspondence between input modality and output modality, while the adversarial loss contributes to generate realistic style images. Nie et al. [42] proposed to use adversarial learning together with a L_1 or L_2 loss and gradient different loss to generate realistic-like CT images from MRI. In particular, adversarial learning has also been employed for data augmentation to help training neural networks [13, 59]. For example, Chaitanya et al. [13] proposed a generative model, which was optimized towards the task, to synthesize the new training examples. Xue et al. [59] utilized conditional GANs to synthesize realistic cervical histopathological images for cervical intraepithelial neoplasia grading of histopathological images. However, the performance in terms of quantitative metrics become even worse when using adversarial learning. Luc et al. [32] proposed to use adversarial learning to help segmentation tasks, and the authors argue that adversarial learning works as a regularization to enforce higher-order spatial consistency among different classes since adversarial learning can assess the joint configuration of many label variables, which is beyond the capacity of cross-entropy loss. Adversarial learning is also used to improve the medical image segmentation [58, 45, 54]; however, the performance gain by adversarial learning is actually limited to fix the obvious segmentation errors.

Medical Image Segmentation:

The recent development of deep learning has largely boosted the state-of-the-art segmentation methods [5, 7]. Among them, fully convolutional networks (FCN) [5], a variant of convolutional neural networks (CNN), is a recent popular choice for semantic image segmentation in both computer vision and medical image fields [5, 7, 6, 48, 61, 57]. 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 feature learning and segmentation as two separate tasks. UNet [7], an evolutionary variant of FCN, has achieved excellent performance for medical image segmentation, 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. Lin et al. [31] introduced a generic multi-path refinement network with carefully designed encoder/decoder modules to increase the capacity of U-shape network. Chen et al. [16] proposed using atrous separable convolution to enhance the encoder-decoder networks for semantic segmentation. Chen et al. [15] proposed to use contour-aware knowledge to help accurately segment gland images. Zhu et al. [67] introduced a boundary-weighted domain adaptive network to accurately delineate the boundaries of MRI prostate. Apart from the architecture exploration, some works are also proposed to enhance the UNet [51, 47] with the idea of applying attention mechanism for better feature learning.

Medical Image Synthesis:

It is often quite challenging directly synthesize high-quality demanded medical modality images. Convolutional neural network (CNN) provides a new way for learning highly non-linear relationships because of employing multiple-layer mapping [21, 24, 43, 65, 56, 17]. For example, Huang et al. [24] proposed to simultaneously conduct super-resolution and cross-modality medical image synthesis by the weakly-supervised joint convolutional sparse coding. Nie et al. [43] proposed supervised adversarial learning framework with gradient difference loss to synthesize CT from MRI. Costa et al. [17] applied end-to-end adversarial learning for retinal image synthesis. Although the training of the above-mentioned image synthesis methods could achieve very good visual perceptive performances in most cases, they cannot obtain reasonable quantitative results in certain situations, especially in tumor (lesion) regions. This is because 1) adversarial learning imposes an additional objective function to enforce the entire realism, and it will probably affect the optimization of the direct voxel-level reconstruction which is usually the basis of the quantitative metric; 2) training of the network tends to be dominated by the majority of samples/regions that are easy to synthesize, i.e., normal tissue regions, while ignoring the minority of tumor/lesion regions although they are the most important biomarkers in clinical diagnosis.

Recently, combination of synthesis and segmentation with an end-to-end design has been investigated to enhance both synthesis and segmentation. For example, Zhang et al. [64]

proposed a generic cycle-consistent cross-modality synthesis approach by ensuring consistent anatomical structures and could thus improve the segmentation accuracy. Huo et al. [26] presented an end-to-end synthesis and segmentation network to simultaneously synthesize CT image from unpaired MRI and also segment CT splenomegaly without using manual labels on CT. Chen et al. [14] proposed a unsupervised domain adaptation algorithm by designing semantic-aware GAN to work on the domain-shift-related segmentation tasks.

Easy-Sample Dominance Issue:

The easy-to-segment sample dominance phenomenon often occurs in deep learning based medical image analysis tasks due to the irregular distribution of medical images, which is usually attributed to the different abnormal degree of the lesion or the imaging factors, such as different imaging protocols or vendor devices. Several works have been proposed in the literature to address this problem [53, 10, 66]. To achieve better performance on hard-to-segment (or detect) samples, Shrivastava et al. [53] proposed a simple strategy to automatically select hard samples for further tuning the networks. Kumar et al. [29] made use of hard example mining technique to develop an incremental learning framework that can adapt to new medical data while retaining existing knowledge. To prevent the vast number of easy samples from overwhelming the networks during training, Lin et al. [10] designed focal loss for detection and achieved promising results. In another work [66], the authors introduced to directly apply 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 may also suffer from multi-category competition issues.

3 Method

We first present analysis for the roles of discriminator which is the basis our proposed adversarial confidence learning. Then, we introduce the components of our proposed framework one by one with an example of medical image segmentation. Finally, we also extend the adversarial confidence learning framework to lesion image synthesis.

3.1 Analyzing Role of Discriminator

To take better advantage of adversarial learning, we analyze the roles of discriminators in GAN systems and compare them between classic GAN and supervised adversarial learning system. Fig. 1(a) and (b) illustrate these two typical architectures.

In classic GAN, there are two roles of the D: 1) distinguishing the real image v from the generated image u; 2) providing adversarial loss to train the G. In this unsupervised system, the training signal for G only comes from the D network, as a consequence, the generated u does not necessarily correspond to v but follow an implicit distribution of $\{v\}$. Similarly, the supervised adversarial learning system shown in Fig. 1(b), D also has the same roles. However, since G also benefits from the supervised loss from y besides the adversarial loss from D, the generated image y has a spatial match with the ground-truth image y. In other words, the G in supervised adversarial learning system in Fig. 1(b) does not rely on D as much as that in classic GANs.

Some research papers [32, 27, 45] figure out that adversarial learning in supervised models (segmentation and synthesis) work as high-order spatial consistency regularization to improve the supervised model since the traditional supervised losses (i.e., cross entropy loss for segmentation and L_p loss for synthesis) for G only consider pixel-level correspondence but ignore image-level (or pairwise) match. With such adversarial learning, the qualitative performances (mainly visual perception) usually becomes better, while it cannot produce the same level of contribution to quantitative performance gain (the quantitative performances even degenerated in many cases) [27, 45].

In this study, we hope to take better advantage of adversarial learning so that we can synchronously improve both the visual perception and the quantitative performance. We propose adversarial confidence learning to achieve this goal. As shown in Fig. 1(c), we retain the adversarial learning by adopting a fully convolutional (dense) discriminator, and also develop confidence learning to enhance the design of the supervised generator. Specifically, we propose a fully convolutional adversarial framework as shown in Fig. 2: 1) we adopt a full convolutional discriminator for local adversarial learning and also learn dense confidence information (i.e., confidence network D); 2) with the well-learned confidence map, we propose difficulty-aware mechanism to improve the design of the supervised loss of the generator for medical image segmentation and synthesis (i.e., base generator network S).

To ease the description of the proposed algorithm, we 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 output map (segmentation or output modality) $\mathbf{Y} \in Z^{H\times W\times T}$. For segmentation map, we encode it to one-hot format $\mathbf{P} \in R^{H\times W\times T\times C}$ 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 base generator 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 $\widehat{\mathbf{Y}} = \arg\max \widehat{\mathbf{P}}$.

3.2 Base Generator Network for Segmentation

Since segmentation and synthesis shares the same characteristic as dense prediction, the base generator network for segmentation (synthesis) can be any end-to-end dense prediction network (as shown in Fig. 2), such as FCN [5, 41], UNet [7, 18], VNet [2], or DSResUNet [6] (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 adopt residual learning [23] to improve the original UNet, i.e., we replace all the convolutional layers but the last one with the residual modules. We also apply dilated residual module in the intermedia layers between encoder and decoder (the feature maps with the smallest size) [62]. Deep supervision is injected at three scales in the decoder path [37].

Training Segmentation Network: Category imbalance is usually a serious problem in medical image segmentation tasks. To address it, we propose using a generalized multi-class Dice loss [1] to train the segmentation network, as defined below in Eq. 1:

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

where π_c is the class balancing weight for 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: $\widehat{\mathbf{P}} = S(\mathbf{X}, \theta_{\mathbf{s}})$.

In addition, the multi-category cross entropy loss is also adopted to form the voxel-wise measurement, as shown in Eq. 2:

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

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

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

where λ is a non-negative weighting coefficient.

We have conducted a line search ($\lambda \in \{0.25, 0.5, 1, 1.25, 1.5, 2\}$) for the value of the weighting coefficient. We finally select $\lambda = 0.7$ according to the validation experiments.

3.3 Fully Convolutional Adversarial Confidence Learning

Adversarial learning has been shown to be effective in improving visual perception performance for segmentation and synthesis tasks [32, 8, 25, 50, 43]. In the classic adversarial networks, the discriminator is mostly a CNN-based network with the output probability of an input image belonging to be the real [52]. Obviously, the conventional discriminator only provides a global confidence over the entire image domain, without providing local confidence in the dense map, i.e., voxel-wise confidence information. To provide a dense confidence map, we propose using a UNet-based network to model the discriminator and name it as confidence network for convenience. The output of confidence network (denoted as confidence map (M) with size $H \times W \times T$) indicates whether automatic segmentation (generated image) is similar to the ground-truth segmentation (real image) in a voxel-wise manner [30]. We argue that the confidence network can learn the structural information that can be used to regularize the output of base dense prediction network [25, 30]. Also, this local discriminator can mitigate the gradient vanishing issue to some degree [22, 40]. In this paper, a simplified UNet [7] (i.e., we half the number of feature maps for all the convolutional layers except the last one and replace pooling with strided convolution) is adopted to implement the confidence network.

Training the Confidence Network: The training objective of the confidence network is the summation of binary cross-entropy loss over the entire 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 **X** and **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 as Realistic Regularization: For segmentation network, the above-mentioned hybrid loss as defined in Eq. 3 mainly targets at bringing voxel-level or organ-level match between ground-truth segmentation and automatic segmentation. However, it cannot evaluate the match of the two segmentation's in an overall sense. As a result, we propose using an adversarial loss term from D to work as a *realistic regularization*, which aims at enforcing higher-order spatial consistency between ground-truth segmentation and automatic segmentation in an implicit manner. In particular, the adversarial loss ("ADV") to improve S and fool D can be defined by Eq. 6:

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

3.4 Difficulty-Aware Attention Mechanism

Focal loss is a common choice for alleviating the overwhelming effect of easy samples in many computer vision tasks, such as image detection and segmentation [10, 66]. The success of focal loss can be attributed to its strategy that pays more attention on the recognized hard samples (regions) and less attention to the easy samples. Thus, the critical point for such a methods is how to recognize the confident (or difficult) samples (regions). Vu et al. [55] utilized the entropy maps to represent the confidence information to help unsupervised domain adaptation in semantic segmentation. Similarly, focal loss utilizes the predicted probability of a sample to indicate the difficulty degree of this sample, 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-category competition. Thirdly, only considering information from the predicted mask may not really indicate the hard regions since it can ignore structural information without considering the original input image of the segmentation network. Most importantly, the predicted probability (i.e., $P(Y = \hat{y} \mid X, \theta_S)$, where $\hat{y}$ is the predicted category via argmax operation) is just *the probability of estimating* *the sample to be the predicted category*, which means, the confidence value is inclined to be higher than it should be because of the maximum operation (i.e., argmax operation to select the category with the highest probability) if we use the probability as the confidence information. 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 argue that *a more professional easy-or-hard representer* is needed.

The previously described confidence learning provides us with a solution to better recognize the easy-or-hard samples. Firstly, the confidence value produced by the confidence network contains the easy-or-hard information (i.e., $P(\hat{y} = y^* \mid X \cup \hat{y}, \theta_D)$, where y^* is the ground-truth category and P indicates *the probability of the predicted category to be the ground-truth category.* Thus, it is more suited to work as the confidence information than the softmax probability. Also, it can avoid the issue of pushing the confidence value bigger). Also, since confidence network is actually a binary classification model which will not suffer from the multi-category competition of the focal loss in many cases. More importantly, the confidence map contains abstract information from both the original input image and the predicted probability mask, which makes it be able to 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 for segmentation using region-level and voxel-level attentions from both predicted probability mask and confidence map. We also propose difficulty-aware mechanism for lesion medical image synthesis.

Difficult-aware Attention based Segmentation Loss: First, we propose an organ-level attention based generalized Dice loss to depict the region-level difficulty, as shown in Eq. 7.

$$L_{F\ Dice(X,P;\theta_S)} = \frac{1}{C}\sum_{c=1}^{C}(1-dsc_c)^{\alpha}\left(1 - \frac{\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T}P_{h,w,t,c}\hat{P}_{h,w,t,c}}{\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T}\left(P_{h,w,t,c}+\hat{P}_{h,w,t,c}\right)}\right) \tag{7}$$

where dsc_c is the average Dice similarity coefficient of a specific category c, e.g., a certain organ or tissue. α is the organ-level attention parameter with a range of [0, 5]. Following [10], we set α 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:

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

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.

$$L_{DamHyb} = L_{F\ Dice} + L_{FCE} \tag{10}$$

With the difficulty-aware hybrid loss in Eq. 10, we can pay more attention in the less 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.

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

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. In this perspective, the adversarial loss term can be viewed as “variational” regularization term to guarantee the overall realism of the automatic segmentation.

3.5 Adversarial Confidence Learning for Cross-Modality Lesion Image Synthesis

Adversarial learning has been widely used for cross-modality medical image synthesis due to its capacity to generate realistic images [42, 56, 60]. However, the quantitative performance cannot be improved as much as qualitative improvement (it can even become worse with adversarial learning in many cases). Especially, for the irregular regions, such as lesion or tumor regions, both the visual perception and the quantitative performance need further improvement even with the conventional adversarial learning. To achieve this goal, we propose to use the similar framework shown in Fig. 2 for cross-modality lesion image synthesis.

Basic L_p Loss for Reconstruction: As mentioned in the Introduction section, typically an L_1/L_2 loss is conventionally used to train the typical cross-modality synthesis network as shown in Eq. 12.

$$L_G(X, Y) = \|Y - G(X)\|^p \tag{12}$$

where Y is the ground-truth target image, and $G(X)$ is the generated target image from the source image X by the Generator network G and p is 1 or 2.

Realistic Regularization with Adversarial Learning: To produce realistic target modality images, we adopt Eq. 6 to work as an regularization term. This realistic regularization term drives the objective function of image synthesis to consider the realistic effect in an entire view instead of only optimizing towards the minimal reconstruction error in voxel (pixel) level.

Difficult-aware Attention based L_p Loss: Due to the inhomogeneous characteristics and irregular distribution of the medical images, certain region of the images are usually more difficult to well synthesize. As a consequence, it is quite desired to build a model that can better model the hard-to-prediction regions. Since the local discriminator could provide the dense confidence information about how well each region is synthesized, we can thus pay more attention on the hard-to-predict regions (e.g., lesion regions) so that these regions can be better modeled. To this end, we propose using the above-mentioned adversarial difficulty-aware attention mechanism to better represent the easy-or-hard information. Specifically, we design a difficulty-aware L_1/L_2 loss using region-level attentions from the adversarial local confidence map.

The voxel-level difficulty-aware attention from the confidence map (M) is formulated (based on Eq. 12) in Eq. 13:

$$L_{AttG}(X, Y) = F \odot \|Y - G(X)\|^p \tag{13}$$

where ⊙ is the element-wise multiplication and

$$F = (1 - M)^\beta \tag{14}$$

where β is the voxel-level attention parameter. Note, F here works as a scaling factor, which largely suppresses the contribution of easy-to-synthesize regions to the training loss and emphasizes the hard-to-synthesize regions.

With the difficult-region-aware L_1/L_2 loss in Eq. 13, we can pay more attention in the less confidently (i.e., hard-to-predict) regions and thus better model them (e.g., tumor or lesion regions). As a consequence, this adversarial difficulty-region-aware attention mechanism provides an opportunity to use voxel-wise focal loss in regression context.

Total Loss for Training Generator: To this end, the total loss for training generator includes the attention based Li L_1/L_2 loss, and the local adversarial loss, which can be summarized below Eq. 15.

$$L_G = L_{AttG} + \lambda_1 L_{ADV} \tag{15}$$

In this study, the balance coefficient (λ_1) is selected at 0.005. The above training loss could encourage G to generate target images with voxel-wise correspondence to real target image. At the same time, the generated image will be constrained to be as realistic as possible so that it can fool the discriminator.

3.6 Discussion for Selection of Adversarial Loss Functions

There are many well designed loss functions proposed for the GANs [33]. Among them, the widely used loss functions are classic GAN [4], NSGAN [4], WGAN [12], WGANGP [20], LSGAN [34], respectively.

Table 1 presents the basic discriminator and generator loss functions. Since the classic GAN does not impose any prior on the data distribution, its implicit assumption is that GAN could generate samples from any data distributions. To achieve such an effect, the classic GAN implicitly assumes that their discriminator has infinite modeling capacity which can distinguish the distributional consistency between generated and real samples [4, 49]. The non-saturate GAN (NSGAN) also has such an assumption but it instead utilizes a non-saturating loss to generate better gradient signal for the generator. To alleviate the gradient vanishing issues in classic GAN and NSGAN, the authors of WGAN [12] proposed using Earth-Move distance to build their GAN models with Lipschitz regularity. Similar works are proposed in [20, 34]. The ideas of these works are actually to limit the infinite modeling capacity of the D so that we can mitigate the gradient vanishing issue.

In our study, apart from improving the visual perception with the adversarial learning, we also hope to improve quantitative performance with the supervised generator by utilizing the confidence information from the D. To achieve such a goal, we have to retain the infinite modeling capacity of D. Moreover, it is critical in our adversarial confidence learning system to obtain the probability information indicating whether the local region is well predicted or not. However, WGAN, WGANGP and LSGAN cannot directly provide such probability information. Furthermore, NSGAN is suggested to be more stable than GAN [4]. Therefore, we select NSGAN as the loss function to train our supervised adversarial learning system.

3.7 Implementation Details

Pytorch[1] is adopted to implement our proposed framework shown in Fig. 2. Part of the codes are released in the github repositories [2] and [3]. Since we desire a perfect discriminator (D), we do not adopt the traditionally used strategies to limit the D [50]. We adopt Adam algorithm to optimize the networks. For segmentation tasks, the input size of the segmentation network is 64 × 64 × 16 and the batch size is set to be 6. The network weights are initialized by the Xavier algorithm [19] and weight decay is set to be 1e-4. For the network biases, we initialize them to 0. The learning rates for the generator 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 until smaller than 1e-7. For synthesis tasks, the input size is set as 240 × 240 × 5 with a batch size 16. The network weights are also initialized by the Xavier algorithm [19] and weight decay is set to be 1e-4. The network biases are initialized to 0. The learning rates for the generator network and the confidence network are both initialized to 5e-4, followed by decreasing the learning rate 2 times for the S, and 5 times for the D every 3 epochs during the training until smaller than 5e-7. Then we use SGD as optimal solver to continue the training until the loss cannot

[1] https://github.com/pytorch/pytorch
[2] https://github.com/ginobilinie/medSynthesisV1
[3] https://github.com/ginobilinie/medSegmentation

decrease any more. A Titan X GPU server is utilized to train the networks. For all tasks, we preprocess the data by $\frac{x-\mu}{\sigma}$, where μ and σ are the mean and standard deviation for the training dataset, respectively.

4 Experiments for Segmentation Tasks

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

The pelvic dataset consists of 50 prostate cancer patients from a Cancer Hospital, each with one T2-weighted MR image and its 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 × 256 × (120 ~ 192), and the voxel size is mainly 1 × 1 × 1 mm^3.

Five-fold cross-validation is used to evaluate our method. Specifically, we randomly and evenly partition the data set to five folds (with each fold having 10 subjects). In each fold of cross-validation, we select the one fold (that has not been used as testing set before) as the testing set. Accordingly, the rest four parts are used as training set and validation set (note, we randomly sample 5 samples out of these 40 subjects for validation). In this manner, each subject is actually involved in the testing set once after the five-fold cross-validation. It is worth noting that the model training and selection are only based on the training and validation sets, i.e., the test set is never used to change or select the model. We use sliding windows to go through the whole MRI for prediction for each testing subject, in which the stride is stride is set as 16 × 16 × 8. Unless explicitly mentioned, all the reported performance by default is evaluated on the testing set. As for evaluation metrics, we utilize the mean Dice Similarity Coefficient (DSC) and the Average Surface Distance (ASD) to measure the agreement between the manual and automatic label maps.

4.1 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 2: 1) multi-atlas label fusion (MALF), 2) SSAE [11], 3) UNet [7], 4) enUNet, 5) VNet [2], and 6) DSResUNet [6]. Also, we present the performance of our proposed method.

We have performed Wilcoxson signed-rank test to validate whether the improvement of our proposed method, compared to previous methods, is significant or not. The experimental results in Table 3 demonstrate statistical significant improvements ($p < 0.05$ by Wilcoxon signed-rank test) of our proposed method over all the compared methods in terms of DSC. As for ASD. As for ASD, the proposed method has achieved significant improvements towards most of the compared methods.

We visualize some typical segmentation results in Fig. 3, which can show the superiority of our proposed method, especially for the hard-to-segment regions, i.e., prostate and rectum. We also present the quantitative comparison of our method with the five state-of-the-art segmentation methods in Table 2. 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. In contrast, 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). DSResUNet presents good performance in the bladder and rectum regions, but it cannot well model the prostate region which is the most difficult but important region. More importantly, thanks to the adversarial confidence learning framework, the quantitative performance gain can now align with the visual perception improvement.

4.2 Impact of the Realistic Regularization

To investigate how adversarial learning helps the segmentation model, we visually check two typical subjects in Fig. 4. In Fig. 4(a), enUNet gives similar (or a little bit better) segmentation results as the enUNet with adversarial learning (i.e., enUNet+globalD and enUNet+localD), which means adversarial learning can only provide very subtle improvement even if enUNet has already produced very similar organs to the manual ground truth. In Fig. 4(b), we can clearly see that adversarial learning has corrected obvious errors in the segmented organs by enUNet. To summarize, adversarial learning can supply reasonable visual perception improvement, especially when the results from the segmentation model have obvious structural errors.

On the other hand, with the global adversarial learning, the DSC values are only improved by 0.2%, 0.2% and 0.1% in average for bladder, prostate and rectum, respectively. With local adversarial learning, the Dice ratios are improved by 0.1%, 0.2%, 0.2% for these three organs, respectively. The experimental results indicate that adversarial learning based realistic regularization can contribute to performance gain but in a subtle manner which does not correspond to the visual perception improvement. We argue that, adversarial learning provides a way of minimizing the “variational” loss by enforcing higher-order consistency between ground-truth segmentations and automatic segmentations. As a result, the visual perception performance can be improved in a larger degree, due to such emphasis on image-level similarity. While the quantitative performance, i.e., *DSC*, is actually included by the original objective function of the segmentation network, it still cannot benefit as much as the visual perception performance.

Besides, qualitative and quantitative performance gains indicate that both local and global adversarial learning contribute very similarly to the performance gains if only considering the effect of adversarial learning.

To further explore the effectiveness of the realistic regularization, we ask a physician to select the segmentations from UNet and UNet with realistic regularization (Note, the physician does not know which method produced the segmentations beforehand). About 65% of segmentations chosen by the physician are those segmented by UNet with realistic regularization, which validates that realistic regularization can improve visual perception for medical image segmentation.

4.3 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, enUNet and the one with difficulty-aware attention mechanism (enUNet+dam) in Fig. 5. (Note, we use the hybrid loss to train UNet and enUNet). Actually, in our case, the enhancement for the UNet with certain modules as introduced before contribute most to the performance gain. The effectiveness of difficulty-aware attention mechanism is also confirmed by the improved performance as shown in Fig. 5. 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.

Comparison with the Focal Loss: Since our proposed difficulty-aware attention mechanism is designed based on the focal loss, it is interesting to explore the difference of the proposed module against focal loss for medical image segmentation.

To better understand the two strategies, we first visualize the difficulty-aware mask (i.e., (1 − M)) and the focal mask (i.e., $\left(1 - \hat{P}\right)$) in Fig. 6. 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. 16:

$$M = D(\hat{P} \cup X) \qquad (16)$$

where ∪ denotes the concatenation operation.

We further conducted experiments with these different strategies to segment the prostate only, since the prostate 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 5.

As described in Table 5, 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.

4.4 Validation on Prostate Challenge Dataset

We have also evaluated our proposed method on the prostate segmentation challenge dataset[4]. The ground-truth label maps for the testing set 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. Besides, an overall score (shown as the last column in Table 4) combining the above-mentioned evaluation metrics is also provided to rank the submitted methods (please refer to [3] for the details about the evaluation metrics).

The quantitative results of our method and our competitors are shown in Table 4. (Note, the results were directly obtained from the organizers based on our submission in Sep. 2018). There were more than 150 teams successfully submitting their results and being listed in the leaderboard at that time. Note we only list top 10 teams in the Table for convenience, and please refer the entire leaderboard through this link[5]. Our proposed method ranks k^{th} in terms of the overall score among all the 150 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 minimum standard deviation comes from the 2nd ranked team who has assembled results from 20 segmentation networks), which further indicates the effectiveness and robustness of our proposed method.

It is interesting to note that our proposed method achieves a very competitive performance on the base and apex parts which are usually thought to be the most difficult segmented regions, and it further validates that our proposed difficulty-aware attention mechanism indeed contributes to the performance gain.

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

5 Experiments on Synthesis Tasks

We choose the BRATS dataset to evaluate our proposed method, which is a publicly available dataset of MRI from brain tumor patients [36]. A total of 354 pairs of T1 MRI and T2 MRI were assembled, where 200 subjects were used for training and 60 for validation, and the rest 94 for testing.

To demonstrate the advantage of our proposed method in terms of synthesis accuracy, we compare it with four widely-used approaches: atlas-based, FCN, UNet [21], UNet with CNN-based global adversarial learning (UNet+GlobalD or AdUNet) [44], and UNet with FCN-based local adversarial learning (UNet+LocalD) [30]. For fair comparison, we use the UNet to work as the image synthesis network (generator).

5.1 Impact of Realistic Regularization

To explore the contribution of realistic regularization, we conduct experiments comparing between the UNet, UNet+GlobalD and UNet+LocalD on the BRATS dataset. The visual comparison is shown in Fig. 8. Obviously, both the global and local adversarial learning can largely improve the visual perception performance. We also ask a radiologist to select the synthesized T2 MRI by UNet and UNet with realistic regularization (Note, the radiologist does not know which method produced the images beforehand). Interestingly, the synthesized T2 MRI chosen by the radiologist are almost generated by UNet with realistic regularization. This fact is a strong proof that realistic regularization can make visual perception improvement for medical image synthesis.

The PSNR values are 26.2dB, 25.9dB and 26.0dB in average for these three methods, respectively. Note that these results are achieved with the ordinary L_1 loss for the generator. Compared to UNet, the adversarial learning seems not able to improve the quantitative performance. This is consistent with the objective functions of these three methods, since UNet only optimizes towards minimizing the L_1 loss which actually directly maximizes the PSNR, while UNet with adversarial learning is also constrained by the realistic regularization.

5.2 Impact of Difficult-Region-Aware Attention Mechanism

To show the impact of our proposed difficult-region-aware attention mechanism, we first conduct experiments to compare the performance for cases with/without this mechanism on the BRATS dataset. The experimental results indicate that the performance could be improved by 0.8dB in terms of PSNR using our proposed attention mechanism. To further investigate the impact of our proposed mechanism, we focus on evaluating the synthesis performances only on tumor regions. By using the manually segmented tumor regions in this database, we compute PSNR on tumor regions of testing set, obtaining 1.2dB performance gain in average.

We also visualize results in Fig. 8. We can clearly see that the generated image by using our proposed difficult-region-aware attention mechanism (i.e., ‘UNet+LocalD+Attention’) could recover much more details, compared to the results without using our proposed mechanism (i.e., ‘UNet+LocalD’), especially for the tumor regions.

Besides the above-mentioned qualitative comparison, a quantitative measurement is designed to further investigate the realistic effect of adversarial learning. Specifically, 100 slices are randomly sampled from the ground-truth images. Then the corresponding 100 slices are also sampled from the corresponding synthetic images by UNet and UNet +GlobalD, respectively. Next, two selection games are designed:

1. A radiologist is asked to select the ‘real’ slice between ground-truth slice and the synthetic slice by UNet;
2. The same radiologist is also asked to select ‘real’ slice between the ground-truth slice and the synthetic slice by UNet+LocalD (Note, the radiologist does not know the ground-truth images beforehand);
3. The same radiologist is further asked to select ‘real’ slice between the ground-truth slice and the synthetic slice by UNet+LocalD+Attention.

As a result, 12% of the UNet based synthetic slices are chosen to be the real ones by the radiologist; in other words, 12% of the synthetic slices could confuse the expert (i.e., the confusion rate is 12%). As for UNet+LocalD, 31% of the synthetic slices are chosen the by the radiologist (i.e., the confusion rate is 31%). In contrast, 38% of the synthetic slices by our proposed method (UNet+LocalD+Attention) could confuse the expert. As a result, three facts can be observed from this experiment. a) The synthetic slices cannot work as well as the ground-truth slices (since all of the confusion rates are below 50%); b) The adversarial learning can largely improve the visual effect of synthetic MRI, which indicates the capability of adversarial learning works as a realistic regularization for medical image synthesis by improving visual perception; c) With attention mechanism, visual perception can become much better due to its improvement in the details.

To better understand why the difficult-region-aware mechanism works, we also analyze the confidence map generated by the local discriminator (i.e., LocalD). We find that, initially, the tumor regions are evaluated to be poorly synthesized as indicated by local confidence, and thus more attention is paid to tumor regions in later training of the generator network. In the end of training, tumor regions can also be better synthesized.

5.3 More Evaluation for the Cross-Modality Image Synthesis Quality

In the above paragraph, we have evaluated the quality of the generated images with a global metric (such as PSNR) and a self-defined visual perception metric. In a medical setting, however, it is important to asses if the image is anatomically correct. To evaluate medical applicability, we apply segmentation algorithm on both generated images and real images. In particular, we train a UNet [7]) in order to segment the brain tumor MRI into enhancing tumor (ET), tumor core (TC), and whole tumor (TW). We then evaluate the Dice similarity score of the segmentation maps obtained using the original image and the generated image, respectively, as inputs to the network. We show a slice of the segmentation maps in Fig. 9, and also show Dice scores in Table 6. The results show that the synthetic T2 MRI produces segmentation maps very close to those by the real T2 MRI, in terms of Dice scores. We further conducted segmentation experiments using T1 MRI, T1 MRI together with ground-truth T2 MRI and T1 MRI together with synthetic T2 MRI as input to the network, respectively. The reported experimental results in Table 6 indicate that combining T1 and T2 MRI (i.e., ground-truth T2 MRI or synthetic T2 MRI) could largely outperform the results obtained by directly segmenting the T1 MRI. These results imply that the synthetic images have high quality and are also applicable to image segmentation.

5.4 Comparison with Other Methods

To qualitatively compare the image synthesis results by different methods, we show synthetic target image, along with real target image, in Fig. 10. We can see that the proposed algorithm can better preserve the continuity, coalition and smoothness in the synthetic results, since it uses both global and local adversarial learning constraints in the image patch. More importantly, the tumor region of generated T1 MRI can recover much more details than other methods, and thus looks much closer to the real T2 MRI compared to all other methods. We argue that this is due to the difficult-region-aware attention mechanism which reweight more on the recognized hard-to-synthesis regions, i.e., tumor regions.

We also quantitatively compare the predicted results in Table 7, in terms of both PSNR and MAE. Our proposed method outperforms all other competing methods in both metrics. It is worth noting that the quantitative performance cannot improve much with only adversarial learning (it may even become worse), while our method can improve both the quantitative performance and the qualitative performance due to characteristic of the proposed adversarial confidence learning.

Fig. 10(a) shows synthesis results on BRATS dataset (with brain tumors) by different methods. It can be seen that our result is more consistent with the real T2 MRI (right).

To show the generalization ability of our proposed method, we also evaluate it on another brain dataset for synthesizing CT from MRI. To save time, we conduct the experiments in an four-fold cross validation manner. Fig. 10(b) shows CT synthesis results by different methods, and Table 8 gives quantitative comparison results. It is clear that our proposed method can work better than the state-of-the-art methods, demonstrating the good generalization of our proposed method to other datasets for other image synthesis tasks.

6 Discussion

6.1 Limitations

The adversarial confidence learning framework can be applied to many dense prediction tasks with simple adjustment, such as medical image segmentation and synthesis. In particular, the difficulty-aware attention mechanism relies on the proposed dense confidence network, and thus it is limited to work together with the adversarial confidence learning framework. Moreover, if we have to make this mechanism work well, we need to learn a sufficiently good confidence map. Similarly, the confidence-aware semi-supervised learning also needs to go with the adversarial confidence learning, and it depends heavily on the well-learned confidence map. Since, in both scenarios, the dense confidence map is the key indicator to determine which regions are well-segmented and which regions are poorly-segmented.

6.2 Future Work

We have pointed out that adversarial learning could not increase the quantitative performance gain as much as the visual perception improvement. To achieve a synchronous performance increment, we propose a adversarial confidence learning framework, followed by difficulty-aware attention mechanism and confidence-aware semi-supervised learning. Though this framework can indeed improve quantitative performance and retain qualitative performance, we believe there should be another more elegant solution to this problem. The essence of inconsistency performance of the adversarial learning is the inconsistency between the objective function of the generator and the adversarial loss. It will be beneficial to investigate more about adversarial loss and further develop a compatible adversarial learning system which could not only improve the visual perception as the traditional adversarial learning but also increase the quantitative performance gain.

7 Conclusions

We have investigated the roles of discriminator in the classic GANs and compared them with those in supervised adversarial learning systems. With the analysis and experiments, we certify that adversarial learning in supervised models actually works as realistic regularization, which aims at constraining the outputs of generator to be as real as possible in an entire view. To align the quantitative performance with the visual perception performance, we propose an adversarial confidence learning framework to take better advantage of adversarial learning for medical image segmentation and synthesis. The proposed difficulty-aware attention mechanism uses the confidence information from the dense output of the discriminator to enhance the design of the objective function of the supervised generator so that we can better handle the hard-to-segment (or hard-to-synthesize) regions. The experiments on the clinical datasets validate that our proposed framework can improve the quantitative performance and visual perception for both medical image segmentation and synthesis models.

Acknowledgments.

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

References

1. Sudre CHet al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: DLMIA. Springer (2017)
2. Milletari Fet al. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571. IEEE (2016)
3. Litjens Get al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. MedIA 18(2), 359–373 (2014)
4. Goodfellow Iet al. Generative adversarial nets. In: NIPS (2014)
5. Long Jet al. Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
6. Yu Let al. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In: AAAI (2017)
7. Ronneberger Oet al. U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)

8. Moeskops Pet al. Adversarial training and dilated convolutions for brain mri segmentation. arXiv preprint arXiv:1707.03195 (2017)

9. Kohl Set al. Adversarial networks for the detection of aggressive prostate cancer. arXiv preprint arXiv:1702.08014 (2017)

10. Lin TYet al. Focal loss for dense object detection. arXiv preprint arXiv:1708.02002 (2017)

11. Guo Yet al. Deformable mr prostate segmentation via deep feature learning and sparse patch matching. IEEE TMI 35, 1077–1089 (2016)

12. Arjovsky M, Chintala S, Bottou L: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)

13. Chaitanya K, Karani N, Baumgartner CF, Becker A, Donati O, Konukoglu E: Semi-supervised and task-driven data augmentation. In: International Conference on Information Processing in Medical Imaging. pp. 29–41. Springer (2019)

14. Chen C, Dou Q, Chen H, Heng PA: Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In: International Workshop on Machine Learning in Medical Imaging. pp. 143–151. Springer (2018)

15. Chen H, Qi X, Yu L, Heng PA: Dean: Deep contour-aware networks for accurate gland segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2487–2496 (2016)

16. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 801–818 (2018)

17. Costa P, Galdran A, Meyer MI, Niemeijer M, Abramoff M, Mendonca AM, Campilho A: End-to-end adversarial retinal image synthesis. IEEE transactions on medical imaging 37(3), 781–791 (2018) [PubMed: 28981409]

18. Dong C, Loy CC, He K, Tang X: Image super-resolution using deep convolutional networks. IEEE TPAMI 38(2), 295–307 (2016)

19. Glorot X, Bengio Y: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS. pp. 249–256 (2010)

20. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC: Improved training of wasserstein gans. In: NIPS. pp. 5767–5777 (2017)

21. Han X: Mr-based synthetic ct generation using a deep convolutional neural network method. Medical Physics 44(4), 1408–1419 (2017) [PubMed: 28192624]

22. Hardy C, Merrer EL, Sericola B: Md-gan: Multi-discriminator generative adversarial networks for distributed datasets. arXiv preprint arXiv:1811.03850 (2018)

23. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)

24. Huang Y, Shao L, Frangi AF: Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. arXiv preprint arXiv:1705.02596 (2017)

25. Hung WC, Tsai YH, Liou YT, Lin YY, Yang MH: Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv:1802.07934 (2018)

26. Huo Y, Xu Z, Bao S, Assad A, Abramson RG, Landman BA: Adversarial synthesis learning enables segmentation without target modality ground truth. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 1217–1220. IEEE (2018)

27. Isola P, Zhu JY, Zhou T, Efros AA: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)

28. Johnson J, Alahi A, Fei-Fei L: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)

29. Kumar P, Srivastava MM: Example mining for incremental learning in medical imaging. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 48–51. IEEE (2018)

30. Li C, Wand M: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: European Conference on Computer Vision. pp. 702–716. Springer (2016)

31. Lin G, Milan A, Shen C, Reid I: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1925–1934 (2017)

32. Luc P, Couprie C, Chintala S, Verbeek J: Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 (2016)

33. Lucic M, Kurach K, Michalski M, Gelly S, Bousquet O: Are gans created equal? a large-scale study. In: Advances in neural information processing systems. pp. 700–709 (2018)

34. Mao X, Li Q, Xie H, Lau RY, Wang Z, Smolley SP: Least squares generative adversarial networks. In: ICCV. pp. 2813–2821. IEEE (2017)

35. Mathieu M, Couprie C, LeCun Y: Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440 (2015)

36. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE TMI 34(10), 1993 (2015)

37. Merkow J, Marsden A, Kriegman D, Tu Z: Dense volume-to-volume vascular boundary detection. In: MICCAI. pp. 371–379. Springer (2016)

38. Mescheder L, Geiger A, Nowozin S: Which training methods for gans do actually converge? In: ICML. pp. 3478–3487 (2018)

39. Metz L, Poole B, Pfau D, Sohl-Dickstein J: Unrolled generative adversarial networks. arXiv preprint arXiv:1611.02163 (2016)

40. Mordido G, Yang H, Meinel C: Dropout-gan: Learning from a dynamic ensemble of discriminators. arXiv preprint arXiv:1807.11346 (2018)

41. Nie D, Cao X, Gao Y, Wang L, Shen D: Estimating ct image from mri data using 3d fully convolutional networks. In: DLMIA. pp. 170–178. Springer (2016)

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

43. Nie D, Trullo R, Lian J, Petitjean C, Ruan S, Wang Q, Shen D: Medical image synthesis with context-aware generative adversarial networks. In: MICCAI (2017)

44. Nie D, Trullo R, Lian J, Wang L, Petitjean C, Ruan S, Wang Q, Shen D : Medical image synthesis with deep convolutional adversarial networks. IEEE Transactions on Biomedical Engineering 65(12), 2720–2730 (2018) [PubMed: 29993445]

45. Nie D, Wang L, Gao Y, Lian J, Shen D: Strainet: Spatially varying stochastic residual adversarial networks for mri pelvic organ segmentation. IEEE transactions on neural networks and learning systems (2018)

46. Nie D, Wang L, Xiang L, Zhou S, Ehsan A, Shen D: Difficulty-aware attention network with confidence learning for medical image segmentation. In: AAAI (2019)

47. Oktay O, et al. Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)

48. Pan T, Wang B, Ding G, Yong JH: Fully convolutional neural networks with full-scale-features for semantic segmentation. (2017)

49. Qi GJ: Loss-sensitive generative adversarial networks on lipschitz densities. arXiv preprint arXiv:1701.06264 (2017)

50. Radford A, Metz L, Chintala S: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

51. Roy AG, et al. Concurrent spatial and channel ‘squeeze & excitation’ in fully convolutional networks. In: MICCAI

52. Sabokrou M, Pourreza M, Fayyaz M, Entezari R, Fathy M, Gall J, Adeli D: Avid: Adversarial visual irregularity detection. ACCV (2018)

53. Shrivastava A, Gupta A, Girshick R: Training region-based object detectors with online hard example mining. In: CVPR. pp. 761–769 (2016)

54. Tsai YH, Hung WC, Schulter S, Sohn K, Yang MH, Chandraker M: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2018)

55. Vu TH, Jain H, Bucher M, Cord M, Perez P: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2019)
56. Wolterink JM, et al. Generative adversarial networks for noise reduction in low-dose ct. TMI 36(12)
57. Xiao H, Wei Y, Liu Y, Zhang M, Feng J: Transferable semi-supervised semantic segmentation. arXiv preprint arXiv:1711.06828 (2017)
58. Xue Y, Xu T, Zhang H, Long LR, Huang X: Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. Neuroinformatics pp. 1–10 (2018) [PubMed: 29353340]
59. Xue Y, Zhou Q, Ye J, Long LR, Antani S, Cornwell C, Xue Z, Huang X: Synthetic augmentation and feature-based filtering for improved cervical histopathology image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 387–396. Springer (2019)
60. Yang Q, Yan P, Zhang Y, Yu H, Shi Y, Mou X, Kalra MK, Zhang Y, Sun L, Wang G: Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. IEEE transactions on medical imaging 37(6), 1348–1357 (2018) [PubMed: 29870364]
61. Yang X, Yu L, Wu L, Wang Y, Ni D, Qin J, Heng PA: Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images. In: AAAI. pp. 1633–1639 (2017)
62. Yu F, Koltun V, Funkhouser TA: Dilated residual networks.
63. Zhang Y, Yang L, Chen J, Fredericksen M, Hughes DP, Chen DZ: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: MICCAI. pp. 408–416. Springer (2017)
64. Zhang Z, Yang L, Zheng Y: Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9242–9251 (2018)
65. Zhao C, et al. A deep learning based anti-aliasing self super-resolution algorithm for mri. In: MICCAI. pp. 100–108. Springer (2018)
66. Zhou XY, Shen M, Riga C, Yang GZ, Lee SL: Focal fcn: Towards small object segmentation with limited training data. arXiv preprint arXiv:1711.01506 (2017)
67. Zhu Q, et al. Boundary-weighted domain adaptive neural network for prostate mr image segmentation. arXiv preprint arXiv:1902.08128 (2019)
68. Zhu W, Xiang X, Tran TD, Hager GD, Xie X: Adversarial deep structured nets for mass segmentation from mammograms. In: ISBI. pp. 847–850. IEEE (2018)

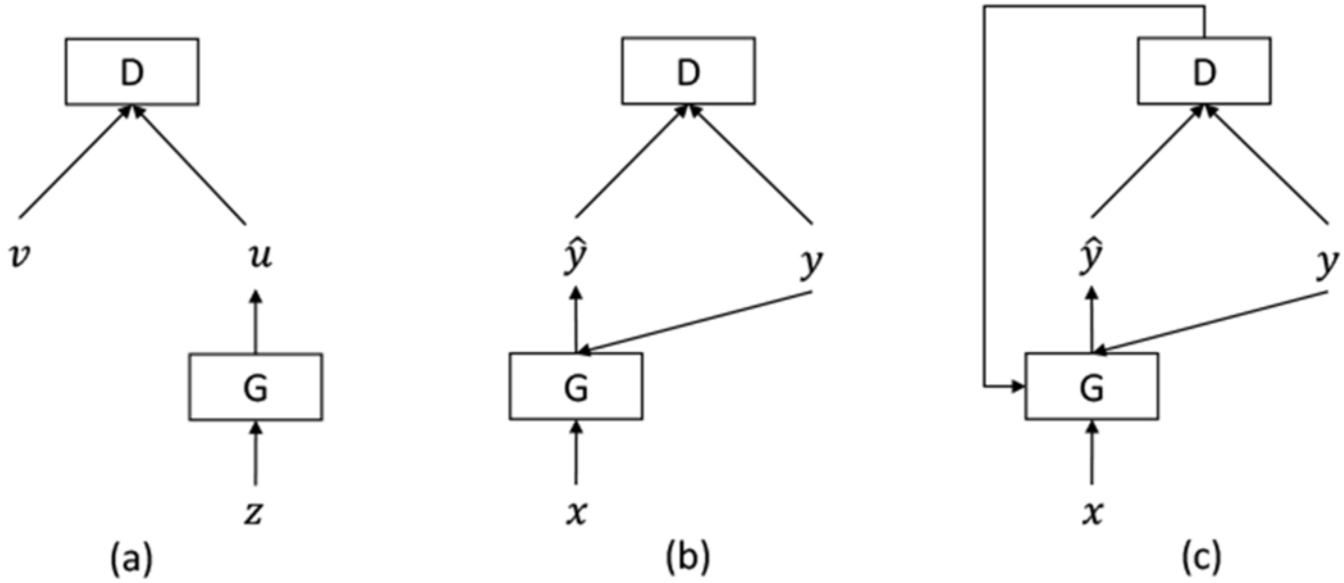

Fig. 1.

Illustration of classic GAN and supervised adversarial learning system, (a) shows a typical classic GAN, where z is an input signal following a certain distribution, u is the generated image, and v is the real image, (b) depicts a typical supervised adversarial learning system, where x is the input modality, $\hat{y}$ is the generated image, and y is the corresponding ground truth image, (c) introduces our proposed adversarial confidence learning framework which retains the adversarial learning and imposes confidence learning to enhance the supervised generator.

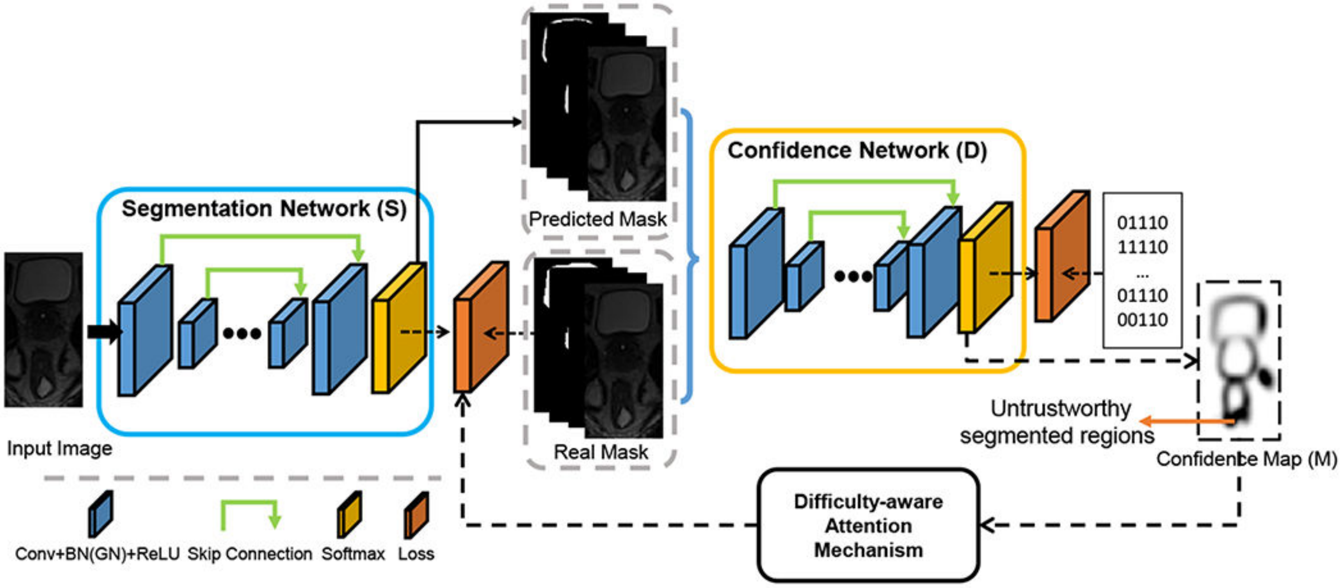

Fig. 2.

Illustration of the architecture of the proposed framework by taking the segmentation task as an example (although this framework can be also adapted to the synthesis task). This framework consists of a segmentation network (*S*), a confidence network (*D*), and the difficulty-aware attention mechanism. In this framework, we pursue *a perfect D* so that we can obtain the confidence map (*M*) to guide the supervised training of the *S*, which means we can inject confidence learning besides the adversarial learning. 0 means the lowest confidence for the prediction of voxel (the predicted category for the voxel is not consistent with the ground-truth category at all) and 1 means the highest confidence for the prediction of the voxel (the predicted category is fully consistent with the ground-truth category).

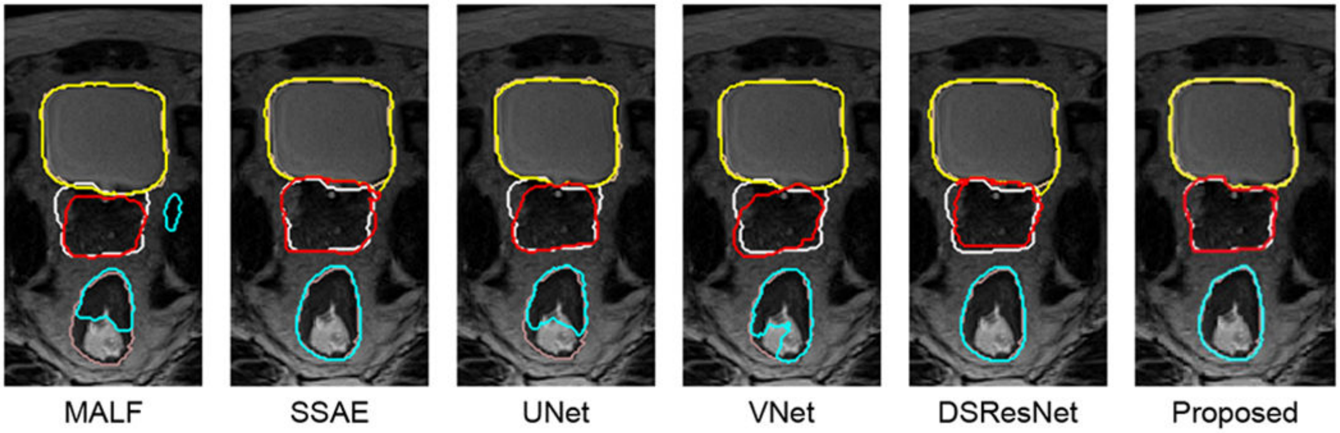

Fig. 3.

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.

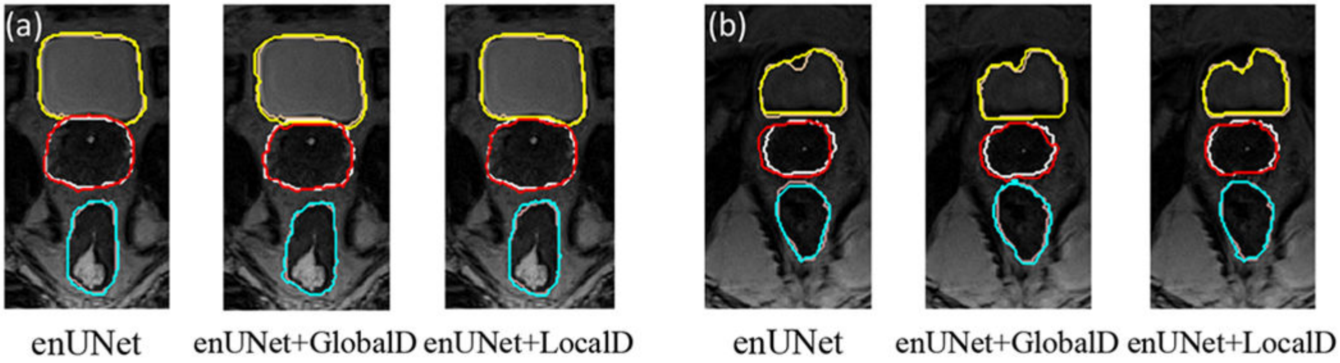

Fig. 4.

Visual inspection of segmentation improvements by adversarial learning on two different cases. Here, enUNet means our proposed network without adversarial learning, enUNet +GlobalD means the proposed segmentation network with global adversarial learning, and enUNet+LocalD means the one with dense adversarial learning. In (a), adversarial learning does not help much, as enUNet already gives good results. In (b), both local and global adversarial learning can obviously help correct the segmented organs obviously, due to large segmentation errors by enUNet.

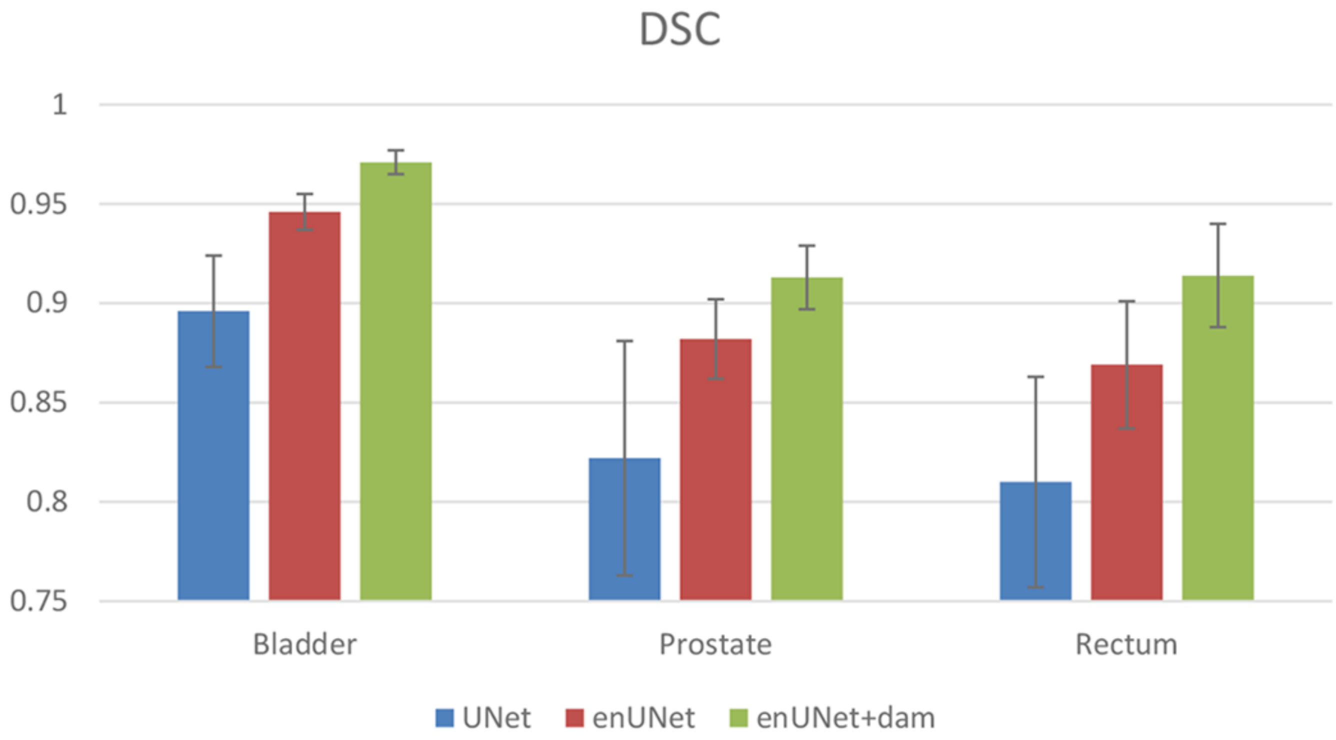

Fig. 5.

Average Dice ratios of different methods.

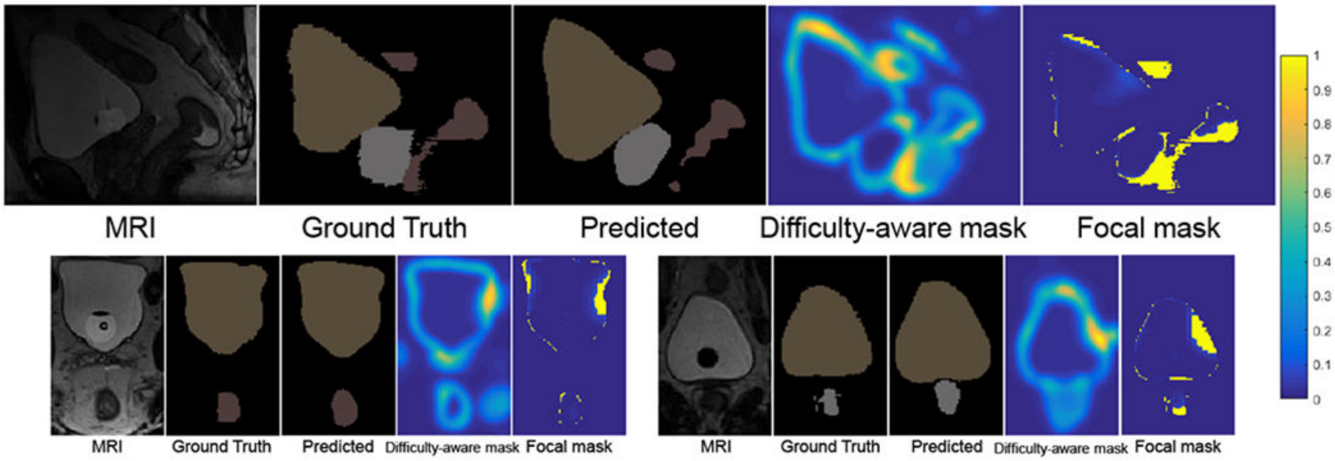

Fig. 6.

Visualization of the difficulty-aware mask and the focal mask, obtained after training the network for 5 epochs. The first row shows sagittal view, and the second row shows both axial and coronal views. In the results of ground-truth and predicted segmentations, orange, silver and pink indicate the bladder, prostate and rectum, respectively. We also use different colors to code the difficulty-aware mask and the focal mask, with the green indicating high-confident regions and the yellow indicating low-confident regions).

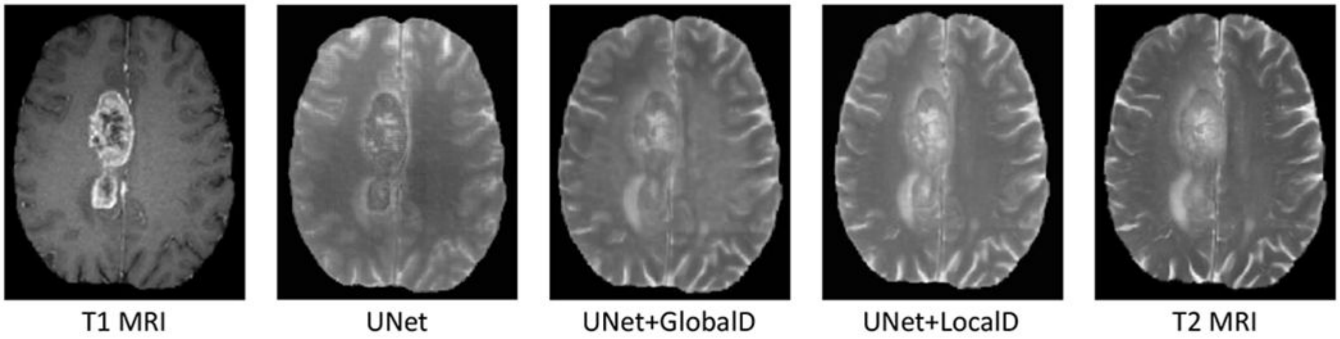

Fig. 7.

Visual evaluation of the effect of the realistic regularization. Using the proposed realistic regularization, the respective results (third column) looks more similar to the real target T2 MRI (fourth column), compared to the case without using the mechanism.

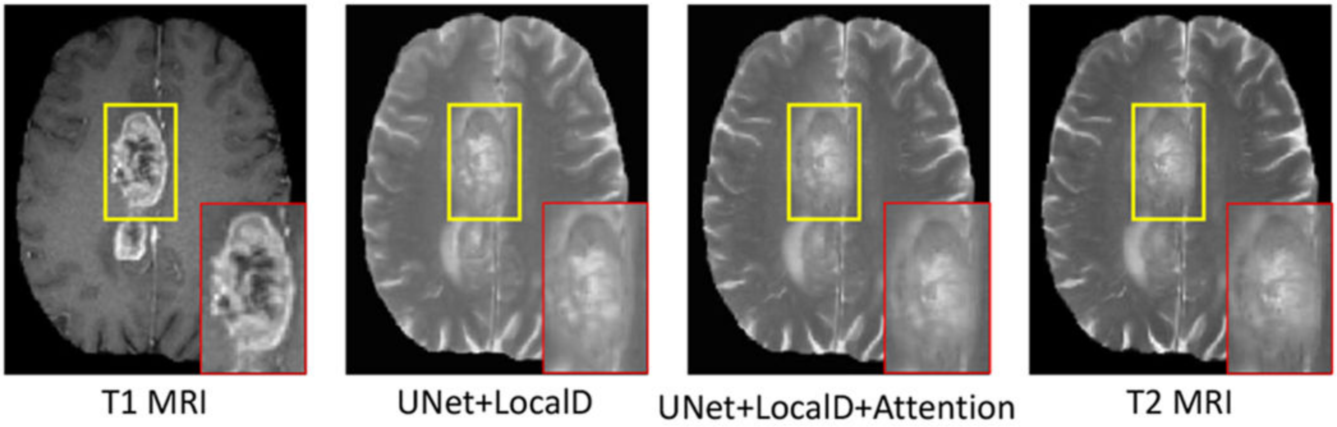

Fig. 8.

Visual evaluation of our proposed difficult-region-aware attention mechanism. Using our proposed mechanism, the respective results (third column) is more similar to the real target T2 MRI (fourth column), compared to the case without using our proposed mechanism (second column).

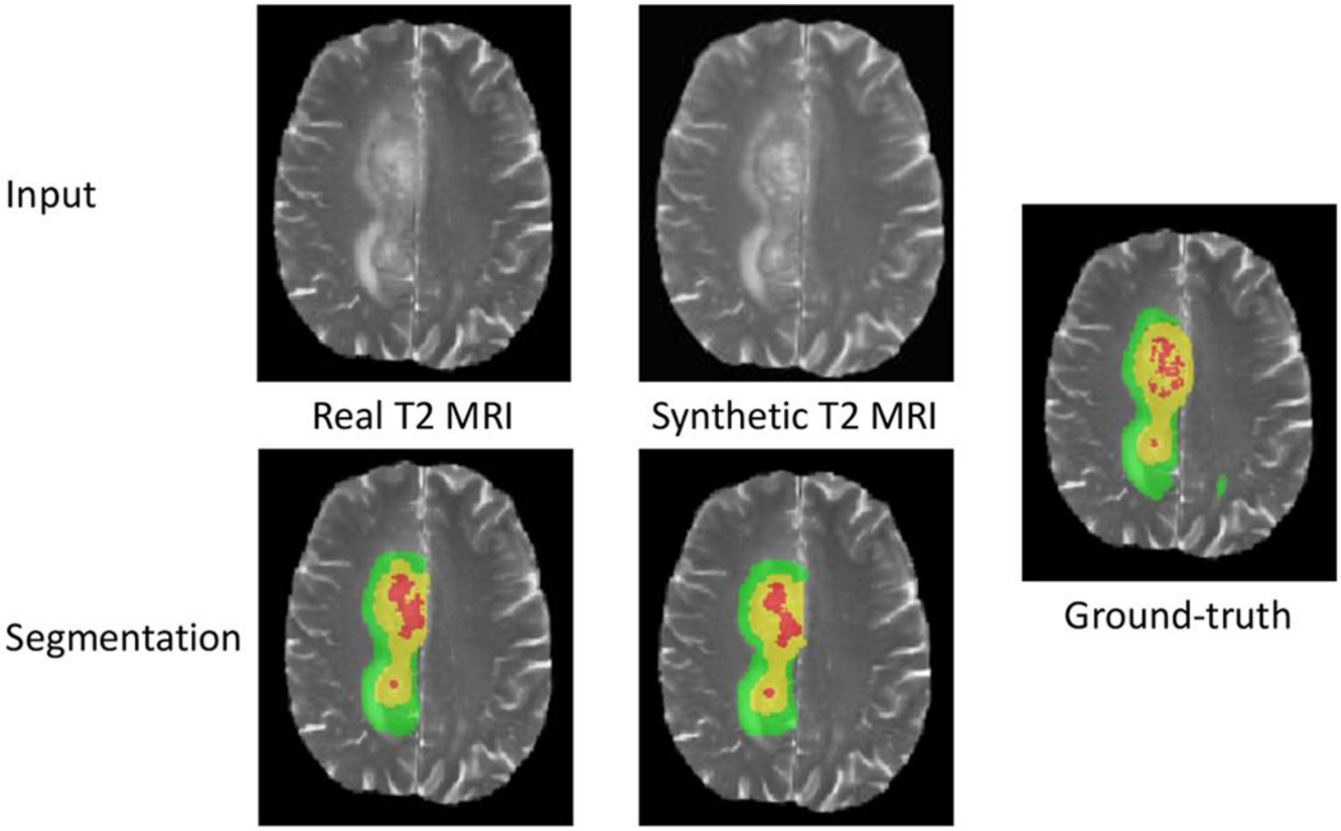

Fig. 9.

Visual comparison of segmentation results for a typical subject by using different input data (real T2 MRI and synthetic T2 MRI).

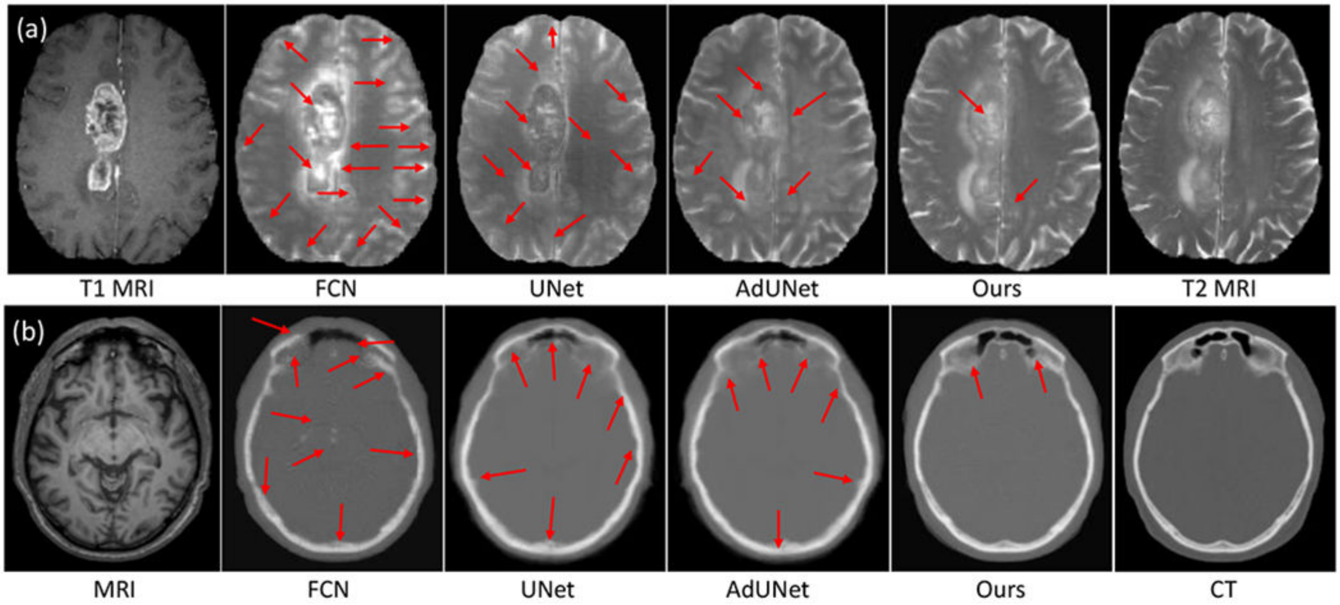

Fig. 10.

Visual comparison of results by different methods for two cases of application: (a) T1 MRI to T2 MRI synthesis, and (b) MRI to CT synthesis. Red arrows indicate poorly-synthesized regions.

Table 1.

Loss functions in the adversarial learning system: GAN, NSGAN, WGAN, WGANGP and LSGAN.

GAN	Loss for Discriminator	Loss for Generator
GAN	$-E_y \log(D(y)) - E_{\hat{y}} \log(1 - D(\hat{y}))$	$E_{\hat{y}} \log(1 - D(\hat{y}))$
NSGAN	$-E_y \log(D(y)) - E_{\hat{y}} \log(1 - D(\hat{y}))$	$-E_{\hat{y}} \log(D(\hat{y}))$
WGAN	$E_{\hat{y}} D(\hat{y}) - E_y D(y)$	$-E_{\hat{y}} D(\hat{y})$
WGANGP	$L_D^{WGAN} + \lambda E_{\hat{y}}(\parallel \nabla D(\alpha y + (1 - \alpha\hat{y})) \parallel_2 - 1)^2$	$-E_{\hat{y}}(D(\hat{y})$
LSGAN	$E_{\hat{y}} D(\hat{y})^2 - E_y (D(y) - 1)^2$	$-E_{\hat{y}}(D(\hat{y} - 1))^2$

Table 2.

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)
enUNet	94.62(.98)	88.17(2.17)	86.87(3.35)	.907(.182)	1.611(.366)	1.602(0.447)
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	**97.68(.67)**	**92.23(l.69)**	**91.07(2.45)**	**.848(.147)**	**1.301(.275)**	**1.380(.34)**

Table 3.

P-Values by performing Wilcoxon signed-rank test between our proposed method and all the compared methods for both DSC and ASD values on the pelvic dataset.

Method	DSC (%)			ASD (in mm)		
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
MALF	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
SSAE	< 0.01	< 0.05	< 0.01	< 0.05	< 0.05	< 0.05
UNet	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
enUNet	< 0.01	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
VNet	< 0.01	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01
DSResUNet	< 0.01	< 0.05	< 0.05	< 0.05	< 0.05	< 0.10

Table 4.

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)

Table 5.

Comparison of different strategies in segmenting 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	88.25(4.14)	90.11(2.67)	86.67(5.46)
Proposed	89.52(3.59)	90.97(2.35)	88.20(4.16)

Table 6.

Performance of segmentation on the Brats dataset in terms of Dice Index and its corresponding standard deviation.

Input	ET	WT	TC
Synthetic T2 MRI	67.84(3.18)	85.80(1.65)	71.45(2.48)
Ground-truth T2 MRI	68.10(3.26)	86.03(1.66)	72.08(2.39)
T1 MRI	69.83(2.84)	86.67(1.92)	71.85(2.18)
T1 MRI + Synthetic T2 MRI	72.42(2.97)	87.01(1.86)	73.23(2.25)
T1 MRI + Ground-truth T2 MRI	72.85(2.92)	87.76(1.88)	73.52(2.33)

Table 7.

Average MAE and PSNR on 94 testing subjects from the BRATS dataset.

Method	**MAE**	**PSNR**
FCN	34.5(8.6)	25.0(2.3)
UNet	28.8(6.9)	26.2(1.8)
pix2pix	30.2(6.8)	25.8(2.1)
AdUNet	29.4(5.7)	26.0(**1.5**)
Ours	**27.3**(5.2)	**26.9(1.6)**

Table 8.

Average MAE and PSNR on 16 subjects from the brain dataset.

Method	MAE	PSNR
FCN	24.4(15.1)	22.7(3.2)
UNet	21.8(12.8)	26.7(2.1)
pix2pix	22.3(11.5)	26.4(1.8)
AdUNet	21.9(11.3)	26.8(**1.7)**
Ours	**20.8(l0.8)**	**27.3**(1.8)