Neuroinformatics (2018) 16:383–392
https://doi.org/10.1007/s12021-018-9377-x

ORIGINAL ARTICLE

SegAN: Adversarial Network with Multi-scale L_1 Loss for Medical Image Segmentation

Yuan Xue[1] · Tao Xu[1] · Han Zhang[2] · L. Rodney Long[3] · Xiaolei Huang[1]

Published online: 3 May 2018

Abstract

Inspired by classic Generative Adversarial Networks (GANs), we propose a novel end-to-end adversarial neural network, called SegAN, for the task of medical image segmentation. Since image segmentation requires dense, pixel-level labeling, the single scalar real/fake output of a classic GAN's discriminator may be ineffective in producing stable and sufficient gradient feedback to the networks. Instead, we use a fully convolutional neural network as the segmentor to generate segmentation label maps, and propose a novel adversarial critic network with a multi-scale L_1 loss function to force the critic and segmentor to learn both global and local features that capture long- and short-range spatial relationships between pixels. In our SegAN framework, the segmentor and critic networks are trained in an alternating fashion in a min-max game: The critic is trained by maximizing a multi-scale loss function, while the segmentor is trained with only gradients passed along by the critic, with the aim to minimize the multi-scale loss function. We show that such a SegAN framework is more effective and stable for the segmentation task, and it leads to better performance than the state-of-the-art U-net segmentation method. We tested our SegAN method using datasets from the MICCAI BRATS brain tumor segmentation challenge. Extensive experimental results demonstrate the effectiveness of the proposed SegAN with multi-scale loss: on BRATS 2013 SegAN gives performance comparable to the state-of-the-art for whole tumor and tumor core segmentation while achieves better precision and sensitivity for Gd-enhance tumor core segmentation; on BRATS 2015 SegAN achieves better performance than the state-of-the-art in both dice score and precision.

Introduction

Advances in a wide range of medical imaging technologies have revolutionized how we view functional and pathological events in the body and define anatomical structures in which these events take place. X-ray, Computerized Axial Tomography (CAT), Magnetic Resonance Imaging (MRI), Ultrasound, Nuclear medicine, among other medical imaging technologies, enable 2D or tomographic 3D images to capture in-vivo structural and functional information inside the body for diagnosis, prognosis, treatment planning and other purposes.

One fundamental problem in medical image analysis is image segmentation, which identifies the boundaries of objects such as organs or abnormal regions (e.g. tumors) in images. Having the segmentation result makes it possible for shape analysis, detecting changes in volume, and planning for radiation therapy treatment. Since manual annotation of object boundaries can be very time-consuming and subjective, an accurate and reliable automated segmentation method is valuable for both clinical and research purposes.

Yuan Xue and Tao Xu are Co-first Authors.

✉ Yuan Xue
yux715@lehigh.edu

Tao Xu
tax313@lehigh.edu

Han Zhang
han.zhang@cs.rutgers.edu

L. Rodney Long
rlong@mail.nih.gov

Xiaolei Huang
xih206@lehigh.edu

[1] Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA

[2] Department of Computer Science, Rutgers University, Piscataway, NJ, USA

[3] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

In the literature of image processing and computer vision, various theoretical frameworks have been proposed for automatic segmentation. Traditional unsupervised methods such as thresholding (Otsu 1979), region growing (Adams and Bischof 1994), edge detection and grouping (Canny 1986), Markov Random Fields (MRFs) (Manjunath and Chellappa 1991), active contour models (Kass et al. 1988), Mumford-Shah functional based frame partition (Mumford and Shah 1989), level sets (Malladi et al. 1995), graph cut (Shi and Malik 2000), mean shift (Comaniciu and Meer 2002), and their extensions and integrations (Gooya et al. 2011; Lee et al. 2008; Lefohn et al. 2003) usually utilize constraints about image intensity or object appearance for segmentation. Supervised methods (Menze et al. 2015; Cobzas et al. 2007; Geremia et al. 2011; Wels et al. 2008; Ronneberger et al. 2015; Havaei et al. 2017), on the other hand, directly learn from labeled training samples, extract features and context information in order to perform a dense pixel (or voxel)-wise classification.

In recent years, Convolutional Neural Networks (CNNs) have been widely applied to visual recognition problems and are shown effective in learning a hierarchy of features at multiple scales from data. For pixel-wise image segmentation, CNNs have also achieved remarkable success. In Long et al. (2015), Long et al. proposed a fully convolutional network (FCNs) for semantic segmentation. The authors replaced fully connected layers in CNNs with convolutional layers to obtain a coarse label map, and then upsampled the label map with deconvolutional layers to get per pixel classification results. Noh et al. (2015) used an encoder-decoder structure to get more fine details about segmented objects. With multiple unpooling and deconvolutional layers in their architecture, they avoided the coarse-to-fine stage in Long et al. (2015). However, they still needed to ensemble with FCNs in their method to capture local dependencies between labels. Lin et al. (2016) combined Conditional Random Fields (CRFs) and CNNs to better explore spatial correlations between pixels.

In the domain of segmenting medical images, deep CNNs have also been applied with promising results. Ronneberger et al. (2015) presented a FCN, namely U-net, for segmenting neuronal structures in electron microscopic stacks. With the idea of skip-connection from Long et al. (2015), the U-net achieved very good performance and has since been applied to many different tasks such as image translation (Isola et al. 2016). In addition, Havaei et al. (2017) obtained good performance for medical image segmentation with their InputCascadeCNN. The InputCascadeCNN has image patches as inputs and uses a cascade of CNNs in which the output probabilities of a first-stage CNN are taken as additional inputs to a second-stage CNN. Pereira et al. (2016) applied deep CNNs with small kernels for brain tumor segmentation. They proposed different architectures for segmenting high grade and low grade tumors, respectively. Kamnitsas et al. (2017) proposed a 3D CNN using two pathways with inputs of different resolutions, and used 3D CRFs to refine their results.

Although these previous approaches using CNNs for segmentation have achieved promising results, they still have limitations. All above methods utilize a pixel-wise loss, such as softmax, in the last layer of their networks, which is insufficient to learn both local and global contextual relations between pixels. Hence they always need models such as CRFs (Chen et al. 2015) as an additional step of refinement to enforce spatial contiguity in the output label maps. Many previous methods (Havaei et al. 2017; Kamnitsas et al. 2017; Pereira et al. 2016) address this issue by training CNNs on image patches and using multi-scale, multi-path CNNs with different input resolutions or different CNN architectures. Using patches and multi-scale inputs could capture spatial context information to some extent. Nevertheless, the computational cost for patch training is very high and there is a trade-off between localization accuracy and the patch size. Instead of training on small image patches, current state-of-the-art CNN architectures such as U-net are trained on whole images or large image patches and use skip connections to combine hierarchical features for generating the label map. They have shown potential to implicitly learn some local dependencies between pixels. However, these methods are still limited by their pixel-wise loss function, which lacks the ability to enforce the learning of multi-scale spatial constraints directly in an end-to-end training process. Compared with patch training, an issue for CNNs trained on entire images is label or class imbalance. While patch training methods can sample a balanced number of patches from each class, the numbers of pixels belonging to different classes in whole-image training methods are usually imbalanced. To mitigate this problem, U-net uses a weighted cross-entropy loss to balance the class frequencies. However, the choice of weights in their loss function is task-specific and is hard to optimize. In contrast to the weighted loss in U-net, a general loss that could avoid class imbalance as well as extra hyper-parameters would be more desirable.

In this paper, we propose a novel end-to-end adversarial network architecture, called SegAN, with a multi-scale L_1 loss function, for semantic segmentation. We use the brain tumor segmentation application as an example to demonstrate our framework and the training process. Inspired by the original GAN (Goodfellow et al. 2014), the training procedure for SegAN is similar to a two-player min-max game in which a segmentor network (S) and a critic network (C) are trained in an alternating fashion to respectively minimize and maximize an objective function. However, there are several major differences between our

SegAN and the original GAN that make SegAN more suitable and effective for the task of image segmentation.

- In contrast to classic GAN with separate losses for its generator and discriminator, we propose a novel multi-scale loss function for both the segmentor and critic networks in SegAN. The critic is trained to maximize the novel multi-scale L_1 objective function that takes into account CNN feature differences between the predicted segmentation and the ground truth segmentation at multiple scales (i.e. at multiple layers).
- We use a fully convolutional neural network (FCN) as the segmentor S, which is trained with only gradients flowing through the critic, and with the objective of minimizing the same loss function as for the critic.
- Our SegAN is an end-to-end architecture trained on whole images, with no requirements for patches, or inputs of multiple resolutions.

Extensive experimental results demonstrate that the proposed SegAN achieves comparable or better results than the state-of-the-art CNN-based architectures including U-net. Further, by training the entire system end-to-end with back propagation and alternating the optimization of S and C using the same multi-scale loss function, SegAN can directly learn spatial pixel dependencies at multiple scales, therefore it produces smooth label maps directly without needing additional smoothing using CRFs.

The rest of this paper is organized as follows. "Methodology" introduces our SegAN architecture and methodology. Experimental results are presented in "Experiments". The discussion is in "Discussion". Finally, we conclude this paper in "Conclusions".

Methodology

As illustrated in Fig. 1, the proposed SegAN consists of two parts: the segmentor network S and the critic network C. The segmentor is a fully convolutional encoder-decoder network that generates a probability label map from input images. The critic network is fed with two inputs: original images masked by ground truth label maps, and original images masked by predicted label maps from S. The S and C networks are alternately trained in an adversarial fashion: the training of S aims to minimize our proposed multi-scale L_1 loss, while the training of C aims to maximize the same loss function.

Background on Generative Adversarial Networks

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have been successfully applied to many unsupervised or semi-supervised learning tasks (Salimans et al. 2016) and image generation tasks (Zhang et al. 2017).

The conventional GANs have an objective loss function defined as:

$$\min_{\theta_G}\max_{\theta_D}\mathcal{L}(\theta_G,\theta_D) = \mathbb{E}_{x\sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim P_z}\log(1-D(G(z)))]. \quad (1)$$

In this objective function, θ_G and θ_D represent the parameters for the generator G and discriminator D in GAN, respectively. x is a real image from an unknown distribution P_{data}, and z is a random input for the generator G, drawn from a probability distribution (such as Gaussion) P_z. The training procedure for the GAN framework is similar to a two-player adversarial min-max game. The generator G is trained to minimize the objective function, aiming to reproduce the true data distribution P_{data} and generate images that are difficult for the discriminator to differentiate from real images. Meanwhile, the discriminator D is trained to maximize the objective function, aiming to distinguish real images and synthetic images generated by G.

The Proposed Multi-scale L_1 Loss

Although both GANs and our proposed SegAN utilize the adversarial training process, they have different goals. Unlike conventional GANs which try to find the mapping function between two distributions P_{data} and P_z, SegAN aims at solving the mapping between input images and their correct segmentation masks (i.e. pixel-wise label maps). In our proposed SegAN, given a dataset with N training images x_n and corresponding ground truth label maps y_n, the multi-scale objective loss function $\mathcal{L}$ is defined as:

$$\min_{\theta_S}\max_{\theta_C}\mathcal{L}(\theta_S,\theta_C) = \frac{1}{N}\sum_{n=1}^{N}\ell_{\text{mae}}(f_C(x_n\circ S(x_n)),\, f_C(x_n\circ y_n)), \quad (2)$$

where ℓ_{mae} is the Mean Absolute Error (MAE) or L_1 distance; $x_n \circ S(x_n)$ is the input image masked by a segmentor-predicted label map (i.e., pixel-wise multiplication of predicted_label_map and original_image); $x_n \circ y_n$ is the input image masked by its ground truth label map (i.e., pixel-wise multiplication of ground_truth_label_map and original_image); and $f_C(x)$ represents the hierarchical features extracted from image x by the critic network. More specifically, the ℓ_{mae} function is defined as:

$$\ell_{\text{mae}}(f_C(x),\, f_C(x')) = \frac{1}{L}\sum_{i=1}^{L}||f_C^i(x) - f_C^i(x')||_1, \quad (3)$$

where L is the total number of layers (i.e. scales) in the critic network, and $f_C^i(x)$ is the extracted feature map of image x

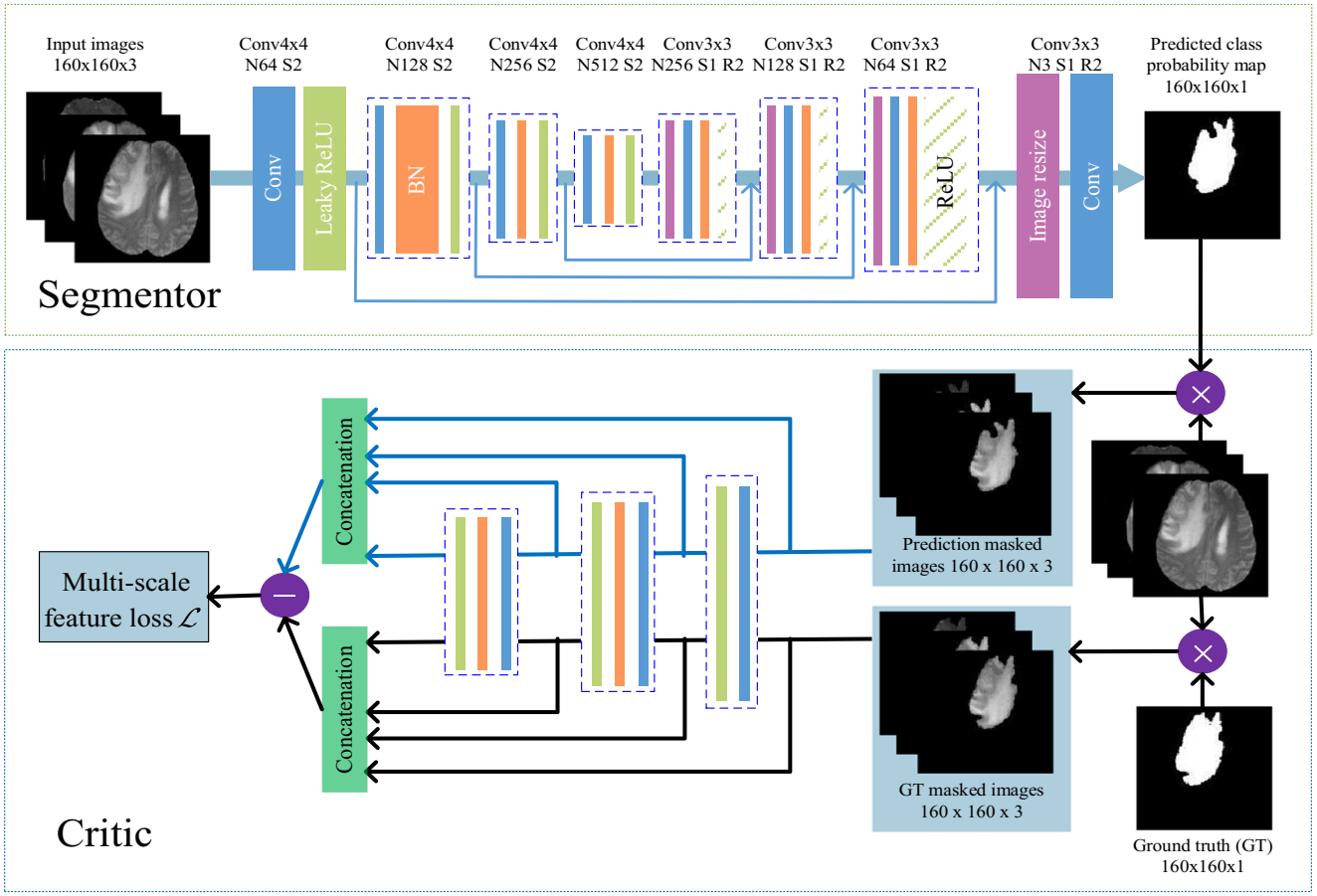

Fig. 1 The architecture of the proposed SegAN with segmentor and critic networks. In the segmentor network, 4×4 convolutional layers with stride 2 (S2) and the corresponding number of feature maps (e.g., N64) are used for encoding, while image resize layers with a factor of 2 (R2) and 3×3 convolutional layers with stride 1 are used for decoding. The critic network has the same structure, thus the same hyperparameters, as the first three blocks of the encoder in the segmentor. Masked images are calculated by pixel-wise multiplication of a label map and (the multiple channels of) an input image. Note that, although only one label map (for whole tumor segmentation) is illustrated here, multiple label maps (e.g. also for tumor core and Gd-enhanced tumor core) can be generated by the segmentor in one path. Best viewed in color

at the ith layer of C. Also, note that, to produce $x_n \circ S(x_n)$ and $x_n \circ y_n$, the label maps are directly applied on the input image so that the image masked by the ground truth label map yields an image with only the tumor part, and the goal of the segmentor is to predict a label map that can also mask out the tumor part.

SegAN Architecture

Segmentor We use a fully convolutional encoder-decoder structure for the segmentor S network. We use the convolutional layer with kernel size 4×4 and stride 2 for downsampling, and perform upsampling by image resize layer with a factor of 2 and convolutional layer with kernel size 3×3 stride 1. We also follow the U-net and add skip connections between corresponding layers in the encoder and the decoder.

Critic The critic C has the similar structure as the encoder in S. Hierarchical features are extracted from multiple layers of C and used to compute the multi-scale L_1 loss. This loss can capture long- and short-range spatial relations between pixels by using these hierarchical features, i.e., pixel-level features, low-level (e.g. superpixels) features, and middle-level (e.g. patches) features.

More details including activation layers (e.g., leaky ReLU), batch normalization layer and the number of feature maps used in each convolutional layers can be found in Fig. 1.

Training SegAN

The segmentor S and critic C in SegAN are trained by back-propagation from the proposed multi-scale L_1 loss. In an alternating fashion, we first fix S and train C for

one step using gradients computed from the loss function, and then fix C and train S for one step using gradients computed from the same loss function passed to S from C. As shown in Eq. 2, the training of S and C is like playing a min-max game: while S aims to minimize the multi-scale feature loss, C tries to maximize it. As training progresses, both the S and C networks become more and more powerful. And eventually, the segmentor will be able to produce predicted label maps that are very close to the ground truth as labeled by human experts. We also find that the S-predicted label maps are smoother and contain less noise than manually-obtained ground truth label maps.

We trained all networks using RMSProp solver with batch size 64 and learning rate 0.00002. We used a grid search method to select the best values for the number of up-sampling blocks and the number of down-sampling blocks for the segmentor (four, in both cases), and for the number of down-sampling blocks for the critic (three).

Proof of Training Stability and Convergence

Having introduced the multi-scale L_1 loss, we next prove that our training is stable and finally reaches an equilibrium. First, we introduce some notations.

Let $f : \mathcal{X} \to \mathcal{X}'$ be the mapping between an input medical image and its corresponding ground truth segmentation, where $\mathcal{X}$ represents the compact space of medical images[1] and $\mathcal{X}'$ represents the compact space of ground truth segmentations. We approximate this ground truth mapping f with a segmentor neural network $g_\theta : \mathcal{X} \to \mathcal{X}'$ parameterized by vector θ which takes an input image, and generates a segmentation result. Assume the best approximation to the ground truth mapping by a neural network is the network $g_{\hat{\theta}}$ with optimal parameter vector $\hat{\theta}$.

Second, we introduce a lemma about the Lipschitz continuity of either the segmentor or the critic neural network in our framework.

Lemma 1 *Let g_θ be a neural network parameterized by θ, and x be some input in space $\mathcal{X}$, then g_θ is Lipschitz continuous with a bounded Lipschitz constants $K(\theta)$ such that*

$$||g_\theta(x_1) - g_\theta(x_2)||_1 \leqslant K(\theta)(||x_1 - x_2||_1), \tag{4}$$

and for different parameters with same input we have

$$||g_{\theta_1}(x) - g_{\theta_2}(x)||_1 \leqslant K(x)||\theta_1 - \theta_2||_1, \tag{5}$$

[1] Although the pixel value ranges of medical images can vary, one can always normalize them to a certain value range such as [0,1], so it is compact.

Now we prove Lemma 1.

Proof Note that the neural network consists of several affine transformations and pointwise nonlinear activation functions such as leaky ReLU (see Fig. 1). All these functions are Lipschitz continuous because all their gradient magnitudes are within certain ranges. To prove Lemma 1, it's equivalent to prove the gradient magnitudes of g_θ with respect to x and θ are bounded. We start with a neural network with only one layer: $g_\theta(x) = A_1(W_1 x)$ where A_1 and W_1 represent the activation and weight matrix in the first layer. We have $\nabla_x g_\theta(x) = W_1 D_1$ where D_1 is the diagonal Jacobian of the activation, and we have $\nabla_\theta g_\theta(x) = D_1 x$ where θ represents the parameters in the first layer.

Then we consider the neural network with L layers. We apply the chain rule of the gradient and we have $\nabla_x g_\theta(x) = \prod_{k=1}^{L} W_k D_k$ where k represent the k-th layer of the network. Then we have

$$||\nabla_x g_\theta(x)||_1 = ||\prod_{k=1}^{L} W_k D_k||_1. \tag{6}$$

Due to the fact that all parameters and inputs are bounded, we have proved Eq. 4.

Let's denote the first i layers of the neural network by g^i (which is another neural network with less layers), we can compute the gradient with respect to the parameters in i-th layer as $\nabla_{\theta_i} g_\theta(x) = (\prod_{k=i+1}^{L} W_k D_k) D_i g^{i-1}(x)$. Then we sum parameters in all layers and get

$$\begin{aligned} ||\nabla_\theta g_\theta(x)||_1 &= \left\| \sum_{i=1}^{L} \left(\prod_{k=i+1}^{L} W_k D_k \right) D_i g^{i-1}(x) \right\|_1 \\ &\leqslant \sum_{i=1}^{L} \left\| \left(\left(\prod_{k=i+1}^{L} W_k D_k \right) D_i \right) g^{i-1}(x) \right\|_1 . \end{aligned} \tag{7}$$

Since we have proved that $g(x)$ is bounded, we finish the proof of Eq. 5. □

Based on Lemma 1, we then prove that our multi-scale loss is bounded and won't become arbitrarily large during the training, and it will finally converge.

Theorem 1 *Let $\mathcal{L}_t(x)$ denote the multi-scale loss of our SegAN at training time t for input image x, then there exists a small constant C so that*

$$\lim_{t \to +\infty} \mathbb{E}_{x \in \mathcal{X}} \mathcal{L}_t(x) \leqslant C. \tag{8}$$

Proof Let g and d represent the segmentor and critic neural network, θ and w be the parameter vector for the segmentor and critic, respectively. Without loss of generality, we omit

the masked input for the critic and rephrase Eqs. 2 and 3 as

$$\min_{\theta}\max_{w}\mathcal{L}_t = \mathbb{E}_{x\in\mathcal{X}}\frac{1}{L}\sum_{i=1}^{L}||d^i(g_\theta(x)) - d^i(g_{\hat{\theta}}(x))||_1, \quad (9)$$

recall that $g_{\hat{\theta}}$ is the ground truth segmentor network and d^i is the critic network with only first i layers. Let's firstly focus on the critic. To make sure our multi-scale loss won't become arbitrarily large, inspired by Arjovsky et al. (2017), we clamp the weights of our critic network to some certain range (e.g., $[-0.01, 0.01]$ for all dimensions of parameter) every time we update the weights through gradient descent. That is to say, we have a compact parameter space $\mathcal{W}$ such that all functions in the critic network are in a parameterized family of functions $\{d_w\}_{w\in\mathcal{W}}$. From Lemma 1, we know that $||d_w(x_1) - d_w(x_2)||_1 \leqslant K(w)(||x_1 - x_2||_1)$. Due to the fact that $\mathcal{W}$ is compact, we can find a maximum value for $K(w)$, K, and we have

$$||d(x_1) - d(x_2)||_1 \leqslant K||x_1 - x_2||_1. \quad (10)$$

Note that this constant K only depends on the space $\mathcal{W}$ and is irrelevant to individual weights, so it is true for any parameter vector w after we fix the vector space $\mathcal{W}$. Since Lemma 1 applies for the critic network with any number of layers, we have

$$\frac{1}{L}\sum_{i=1}^{L}||d^i(g_\theta(x)) - d^i(g_{\hat{\theta}}(x))||_1 \leqslant K||g_\theta(x) - g_{\hat{\theta}}(x)||_1. \quad (11)$$

Now let's move to the segmentor. According to Lemma 1, we have $||g_\theta(x) - g_{\hat{\theta}}(x)||_1 \leqslant K(x)||\theta - \hat{\theta}||_1$, then combined with Eq. 11 we have

$$\frac{1}{L}\sum_{i=1}^{L}||d^i(g_\theta(x)) - d^i(g_{\hat{\theta}}(x))||_1 \leqslant K(x)K||\theta - \hat{\theta}||_1. \quad (12)$$

We know $\mathcal{X}$ is compact, so there's a maximal value for $K(x)$ and it only depends on the difference between the ground truth parameter vector $\hat{\theta}$ and the parameter vector of the segmentor θ. Since we don't update weights in the segmentor when we update weights in the critic, there's an upper bound for $\mathcal{L}_t$ when we update the critic network and it won't be arbitrarily large during the min-max game.

When we update the parameters in the segmentor, we want to decrease the loss. This makes sense because smaller loss indicates smaller difference between $\hat{\theta}$ and θ. When $\theta \to \hat{\theta}$, $\mathcal{L}_t$ converges to zero because the upper bound of $\mathcal{L}$ becomes zero. However, we may not be able to find the global optimum for θ. Now let us denote a reachable local optimum for θ in the segmentor by θ_0, we will keep updating parameters in the segmentor through gradient descent and gradually approaches θ_0. Based on Eqs. 9 and 12, we denote the maximum of $K(x)$ by K' and have

$$\lim_{t\to+\infty}\mathcal{L}_t(x) \leqslant KK'||\hat{\theta} - \theta_0||_1 = C. \quad (13)$$

Since the constant C does not depend on input x, we have proved Theorem 1. □

Experiments

We evaluated our system on the fully-annotated MICCAI BRATS datasets (Menze et al. 2015). Specifically, we trained and validated our models using the BRATS 2015 training dataset, which consists of 220 high grade subjects and 54 low grade subjects with four modalities: T1, T1c, T2 and Flair. We randomly split the BRATS 2015 training data with the ratio 9 : 1 into a training set and a validation set. We did such split for the high grade and low grade subjects separately, and then re-combined the resulting sets for training and validation. Each subject in BRATS 2015 dataset is a 3D brain MRI volume with size $240\times240\times155$. We center cropped each subject into a subvolume of $180\times180\times128$, to remove the border black regions while still keep the entire brain regions. We did our final evaluation and comparison on the BRATS 2015 test set using the BRATS online evaluation system, which has *Dice*, *Precision* and *Sensitivity* as the evaluation metrics. The Dice score is identical to the F-score which normalizes the number of true positives to the average size of the two segmented regions:

$$\text{Dice} = \frac{2|P\cap T|}{|P| + |T|} \quad (14)$$

where P and T represent the predicted region and the ground truth region, respectively. Since the BRATS 2013 dataset is a subset of BRATS 2015, we also present our results on BRATS 2013 leaderboard set.

Due to the limitation of hardware memory and for the reason that brain images in the BRATS dataset are not perfectly aligned in the third dimension, we built a 2D SegAN network to generate the label map for each axial slice of a 3D volume and then restack these 2D label maps to produce the 3D label map for brain tumor. Since each subject was center cropped to be a $180\times180\times128$ volume, it yields 128 axial slices each with the size 180×180. These axial slices were further randomly cropped to size 160×160 during training for the purpose of data augmentation. They were center-cropped to size 160×160 during validation and testing.

For the number of blocks in the encoder and decoder of the segmentor and in the critic, we searched the values from 2 to 6. Increasing the number of blocks makes the architecture deeper but increases computational time. For

performance, increasing the number of blocks may improve accuracy levels but it is not guaranteed. Meanwhile, deeper networks can cause other issues such as overfitting and will make the network more difficult to train. Since we did not observe any obvious improvement on the validation dataset with deeper architectures, we ended up with using 4 blocks in the segmentor and the critic.

We used three modalities of these MRI images: T1c, T2, FLAIR. We did several experiments and found that using all four modalities gave almost the same results as using three modalities. Considering the computational cost and the limitation in GPU memory, we decided to use only these three modalities for most of our experiments. Corresponding slices of T1c, T2, FLAIR modalities are concatenated along the channel dimension and used as the multi-channel input to our SegAN model, as shown in Fig. 1. The segmentor of SegAN outputs label maps with the same dimensions as the input images. As required by the BRATS challenge (Menze et al. 2015), we did experiments with the objective to generate label maps for three types of tumor regions: *whole tumor*, *tumor core* and *Gd-enhanced tumor core*.

As for computational time, the training and testing of the networks were done on a workstation using two Intel Xeon E5-2623 CPUs and one Titan X Pascal GPU with 12G memory. The training time is around two days for the whole training set, and the testing time is about 3ms per image slice.

Choice of Components in SegAN Architecture

In this section, we compare different implementations of the proposed SegAN architecture and also evaluate the effectiveness of the proposed multi-scale L_1 loss on the BRATS validation set for the brain tumor segmentation task. Specifically, we compare the following implementations:

- **S1-1C.** A separate SegAN is built for every label class, i.e., one segmentor and one critic per label. In this case, we need three separate S1-1C to generate label maps for three classes. The output of each segmentor is a label probability map where its values represent the probability of a pixel belonging to that class. Note that, in the BRATS brain tumor segmentation application, a pixel may have high probabilities in multiple label maps, e.g. a pixel can belong to the class *whole tumor* and the class *tumor core*.
- **S3-1C:** A SegAN is built with one segmentor and one critic, where the segmentor generates three channels. Each output channel is the label probability map for one class, which produces three masked images. The masked images from all three channels are concatenated in the channel dimension and fed into the critic.
- **S3-3C.** A SegAN is built with one segmentor that generates three channels (i.e., three label probability maps), and three separate critics, one for each label class. The networks, one S and three Cs, are then trained end-to-end using the average loss computed from all three Cs.
- **S3-3C single-scale loss models.** For comparison, we also built two single-scale loss models: S3-3C-s0 and S3-3C-s3. S3-3C-s0 computes the loss using features from only the input layers (i.e., layer 0) of the critics, and S3-3C-s3 calculates the loss using features from only the output layers (i.e., layer 3) of the critics.

As shown in Fig. 2, models S1-1C and S3-3C give similar performance which is the best among all models. Since the computational cost for S1-1C is higher than S3-3C, S3-3C is more favorable and we use it to compare our SegAN model with other methods in "Choice of Components in SegAN Architecture". In contrast, while model S3-1C requiring the least computational cost, it sacrifices some performance; but by using the multi-scale loss, it still performs better than any of the two single-scale loss models especially for segmenting tumor core and Gd-enhanced tumor core regions.

Comparison to State-of-the-Art

In this subsection, we compare the proposed method, our S3-3C SegAN model, with other state-of-the-art methods on the BRATS 2013 Leaderboard (Havaei et al. 2017; Pereira et al. 2016) Test and the BRATS 2015 Test (Kamnitsas et al. 2017). We also implemented a U-net model (Ronneberger et al. 2015) for comparison. This U-net model has the exact same architecture as our SegAN segmentor except that the multi-scale SegAN loss is replaced with the softmax loss in the U-net. Table 1 gives all comparison results. From the table, one can see that our SegAN compares favorably to the existing state-of-the-art on BRATS 2013 while achieves

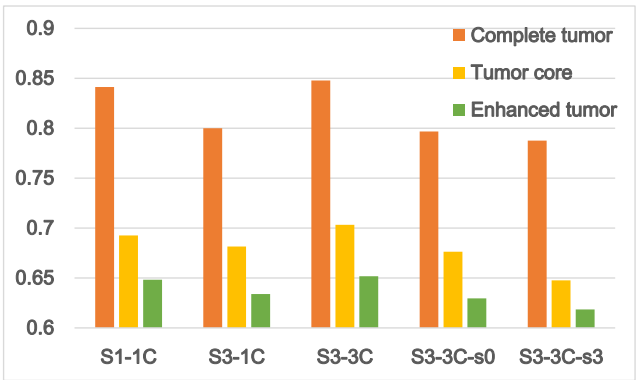

Fig. 2 Average dice scores of different architectures on BRATS validation set

Table 1 Comparison to previous methods and a baseline implementation of U-net with softmax loss for segmenting three classes of brain tumor regions: whole, core and Gd-enhanced (Enha.)

	Methods	Dice			Precision			Sensitivity		
		Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
BRATS 2013 Leaderboard	Havaei et al. (2017)	**0.84**	0.71	0.57	**0.88**	0.79	0.54	0.84	0.72	0.68
	Pereira et al. (2016)	**0.84**	**0.72**	0.62	0.85	**0.82**	0.60	**0.86**	**0.76**	0.68
	SegAN	**0.84**	0.70	**0.65**	0.87	0.80	**0.68**	0.83	0.74	**0.72**
BRATS 2015 Test	Kamnitsas et al. (2017)	**0.85**	0.67	0.63	0.85	**0.86**	0.63	**0.88**	0.60	**0.67**
	U-net	0.80	0.63	0.64	0.83	0.81	**0.78**	0.80	0.58	0.60
	SegAN	**0.85**	**0.70**	**0.66**	**0.92**	0.80	0.69	0.80	**0.65**	0.62

The bold for SegAN indicates our method, the bold for numbers indicates the best performance for each metric and for each dataset

better performance on BRATS 2015. Moreover, the dice scores of our SegAN outperform the U-net implementation for segmenting all three types of tumor regions, which demonstrates the superiority of our proposed adversarial training with multi-scale L_1 loss function compared with the conventional pixel-wise cross-entropy loss on the segmentation task.

Another observation is that our SegAN-produced label maps are smooth with little noise. Figure 3 illustrates some example results of our SegAN; in the figure, the segmented regions of the three classes (whole tumor, tumor core, and Gd-enhanced tumor core) are shown in yellow, blue, and red, respectively. Since the pixel-wise ground truth label maps can be noisy, we feel the fact that our approach achieved accuracy levels comparable to or better than the current state-of-the-art while having smoother results is a good indication of its effectiveness. Furthermore, SegAN is a very general framework that can be applied to other medical image segmentation tasks, even natural image segmentation tasks.

Discussion

To the best of our knowledge, our proposed SegAN is the first GAN-inspired framework adapted specifically for the segmentation task that produces superior segmentation accuracy. While there are very few works that apply adversarial learning to semantic segmentation, one such work that we found by Luc et al. (2016) used both the conventional adversarial loss of GAN and pixel-wise softmax loss against ground truth. They showed small but consistent gains on both the Stanford Background dataset and the PASCAL VOC 2012 dataset; the authors observed that pre-training only the adversarial network was unstable and suggested an alternating scheme for updating

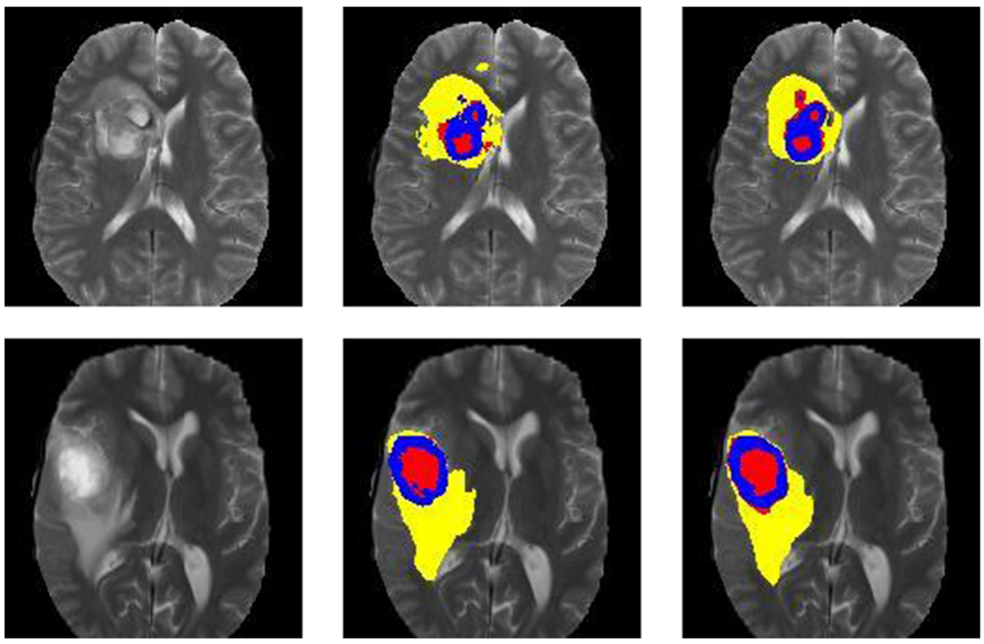

Fig. 3 Example results of our SegGAN (right) with corresponding T2 slices (left) and ground truth (middle) on BRATS validation set

the segmenting network's and the adversarial network's weights. We believe that the main reason contributing to the unstable training of their framework is: the conventional adversarial loss is based on a single scalar output by the discriminator that classifies a whole input image into real or fake category. When inputs to the discriminator are generated *vs.* ground truth dense pixel-wise label maps as in the segmentation task, the real/fake classification task is too easy for the discriminator and a trivial solution is found quickly. As a result, no sufficient gradients can flow through the discriminator to improve the training of generator. In comparison, our SegAN uses a multi-scale feature loss that measures the difference between generated segmentation and ground truth segmentation at multiple layers in the critic, forcing both the segmentor and critic to learn hierarchical features that capture long- and short-range spatial relationships between pixels. Using the same loss function for both S and C, the training of SegAN is end-to-end and stable.

In this work, the L_1 norm is adopted in our multi-scale loss function. We did also experiment with the L_2 norm. However, under our current network architecture and hyper-parameter setting, the adversarial training using the L_2-norm loss was unstable and we were unable to make it converge. Hence no meaning results were obtained from using the multi-scale loss defined with L_2 norm. Our speculation is that L_1 is less sensitive to outliers than L_2, and there is a higher probability of running into the exploding gradient problem with L_2. Although we used weight clipping to reduce the likelihood of gradient exploding, our theoretical proof is based on the L_1 loss and it is not guaranteed to work under L_2. Note that, we do not draw the conclusion that L_2 loss does not work for SegAN-like architectures based on current experimental observations. The potential of L_2 needs further investigation when we evaluate other variants of SegAN in our future work.

In our experiments, we tried several deeper architectures but did not observe any obvious improvement. In addition, deeper architectures are more difficult to train, especially for adversarial networks. We also did not observe improvement by utilizing pooling layers in SegAN. Hence, following the convention of GAN frameworks such as DCGAN (Radford et al. 2015), instead of pooling layers, we use convolutional layers with stride 2 to perform feature down-sampling in SegAN.

For SegAN, we considered several different options regarding the input to the critic network. Since the goal of the critic is to differentiate ground truth label maps from segmentor-predicted label maps, we have the option of directly feeding label maps to the critic, or the option of using label-map masked images as input to the critic. The first option turned out to be too easy for the critic network since the ground truth maps are strictly binary whereas the predicted label maps are not; this triviality of the task leads to unstable and failed training. Thus in this paper we adopt the latter option, which is to do a pixel-wise multiplication of a label map and the raw image and use the masked image (containing only the tumor part) as the input to the critic. This latter option has produced satisfactory results as demonstrated by our experiments.

From the comparison results, we can observe that our SegAN model still has some drawback when segmenting the core and Gd-enhanced regions. While SegAN can extract different levels of features, segmentation for relatively small regions such as core and Gd-enhanced may need more focus on pixel-level features. Thus, previous methods using pixel-level loss could have better performance than the proposed SegAN for segmenting these small regions under some circumstances. One possible improvement for future work can be using different network architectures for segmenting different types of regions. Another drawback that we observe is that, although our model can be easily extended to semantic segmentation tasks that have many label classes, the computational cost can be quite high when the number of classes is large. For instance, in a task with m different classes, to achieve best performance, we can build m S1-1C models (i.e. one segmentor and one critic per class) to generate segmentation masks for the m classes. However, a major limitation is that such a model would have high computational cost when m is very large. In our future work, we will investigate variants of the SegAN architecture in order to reduce computational cost without sacrificing accuracy.

Conclusions

In this paper, we propose a novel end-to-end Adversarial Network architecture, namely SegAN, with a new multi-scale loss for semantic segmentation. Experimental evaluation on the BRATS brain tumor segmentation dataset shows that the proposed multi-scale loss in an adversarial training framework is very effective and leads to more superior performance when compared with single-scale loss or the conventional pixel-wise softmax loss.

As a general framework, our SegAN is not limited to medical image segmentation applications. In our future work, we plan to investigate the potential of SegAN for general semantic segmentation tasks.

Information Sharing Statement

The publicly available BRATS dataset (RRID:SCR_016214) used in this paper can be found via the Google site link:

https://sites.google.com/site/braintumorsegmentation/home. The source code for our SegAN (RRID:SCR_016215) can be found via the GitHub link: https://github.com/YuanXue1993/SegAN.

Acknowledgements This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC), under Contract HHSN276201500692P.

References

Adams, R., & Bischof, L. (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(6), 641–647.

Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein gan. arXiv:170107875.

Canny, J. (1986). A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (6), 679–698.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K. (2015). Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*. arXiv:1412.7062.

Cobzas, D., Birkbeck, N., Schmidt, M., Jagersand, M. (2007). Murtha A (2007) 3d variational brain tumor segmentation using a high dimensional feature set. In *IEEE 11th international conference on computer vision. ICCV 2007* (pp. 1–8). IEEE.

Comaniciu, D., & Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619.

Geremia, E., Clatz, O., Menze, B. H., Konukoglu, E., Criminisi, A., Ayache, N. (2011). Spatial decision forests for ms lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, *57*(2), 378–390.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Gooya, A., Biros, G., Davatzikos, C. (2011). Deformable registration of glioma images using em algorithm and diffusion reaction modeling. *IEEE Transactions on Medical Imaging*, *30*(2), 375–390.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P. M., Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, *35*, 18–31.

Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. arXiv:161107004.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., Glocker, B. (2017). Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, *36*, 61–78.

Kass, M., Witkin, A., Terzopoulos, D. (1988). Snakes: active contour models. *International Journal of Computer Vision*, *1*(4), 321–331.

Lee, C. H., Wang, S., Murtha, A., Brown, M., Greiner, R. (2008). Segmenting brain tumors using pseudo–conditional random fields. In *Medical image computing and computer-assisted intervention–MICCAI 2008* (pp. 359–366).

Lefohn, A., Cates, J., Whitaker, R. (2003). Interactive, gpu-based level sets for 3d segmentation. In *Medical image computing and computer-assisted intervention-MICCAI 2003* (pp. 564–572).

Lin, G., Shen, C., van den Hengel, A., Reid, I. (2016). Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3194–3203).

Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Luc, P., Couprie, C., Chintala, S., Verbeek, J. (2016). Semantic segmentation using adversarial networks. arXiv:161108408.

Malladi, R., Sethian, J. A., Vemuri, B. C. (1995). Shape modeling with front propagation: a level set approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(2), 158–175.

Manjunath, B., & Chellappa, R. (1991). Unsupervised texture segmentation using markov random field models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(5), 478–482.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al (2015). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, *34*(10), 1993–2024.

Mumford, D., & Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, *42*(5), 577–685.

Noh, H., Hong, S., Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1520–1528).

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9* (1), 62–66.

Pereira, S., Pinto, A., Alves, V., Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in mri images. *IEEE Transactions on Medical Imaging*, *35*(5), 1240–1251.

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

Ronneberger, O., Fischer, P., Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention. Springer* (pp. 234–241).

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems* (pp. 2226–2234).

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Wels, M., Carneiro, G., Aplas, A., Huber, M., Hornegger, J., Comaniciu, D. (2008). A discriminative model-constrained graph cuts approach to fully automated pediatric brain tumor segmentation in 3-d mri. In *Medical image computing and computer-assisted intervention–MICCAI 2008* (pp. 67–75).

Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. IEEE Int. Conf. Comput. Vision (ICCV) 5907–5915.