ADVERSARIAL LEARNING WITH MULTI-SCALE LOSS FOR SKIN LESION SEGMENTATION

Yuan Xue, Tao Xu and Xiaolei Huang

Lehigh University
Department of Computer Science and Engineering
Bethlehem, PA

ABSTRACT

Inspired by classic Generative Adversarial Networks (GAN), 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 with new activation function in the last layer 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. We show that such a SegAN framework is more effective in the segmentation task and more stable to train, and it outperforms current state-of-the-art segmentation methods in the ISBI International Skin Imaging Collaboration (ISIC) 2017 challenge, Part I Lesion Segmentation.

***Index Terms*—** skin lesion segmentation, deep convolutional neural networks, adversarial training

1. INTRODUCTION

Convolutional Neural Networks (CNNs) have been widely applied to visual recognition problems in recent years, and they are shown effective in learning a hierarchy of features at multiple scales from data. For pixel-wise semantic segmentation, CNNs have also achieved remarkable success. In [1], Long *et al.* first proposed a fully convolutional networks (FCNs) for semantic segmentation. The authors replaced conventional 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.* [2] 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 [1]. However, they still needed to ensemble with FCNs in their method to capture local dependencies between labels. Lin *et al.* [3] combined Conditional Random Fields (CRFs) and CNNs to better explore spatial correlations between pixels, but they also needed to implement a dense CRF to refine their CNN output.

In the field of medical image analysis, deep CNNs have also been applied with promising results. Esteva *et al.* [4] fine-tuned a pre-trained deep CNN on a very large dataset and achieved higher accuracy than human dermatologists for skin cancer classification. For segmentation, Ronneberger *et al.* [5] presented a FCN, namely U-net, for segmenting neuronal structures in electron microscopic stacks. With the idea of skip-connection from [1], the U-net achieved very good performance and has since been applied to many different tasks such as image translation [6].

Although previous approaches using CNNs for segmentation have achieved promising results, they still have limitations. One challenge is how to learn both local and global contextual relations between pixels. Most methods utilize a pixel-wise loss such as cross entropy in the last layer of their networks and are trained on small image patches, they often need models such as CRFs [7] as a refinement to enforce spatial contiguity in the output label maps. Instead of training on patches, current state-of-the-art CNN architectures such as U-net [5] are trained on whole images or large image patches and use skip connections to combine hierarchical features for generating the label map. However, pixel-wise loss functions still restrict their ability to learn multi-scale spatial constraints directly in the end-to-end training process.

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. Inspired by the original GAN [8], 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. The main novel contributions of our proposed SegAN are three fold.

- We propose a novel multi-scale loss function for both

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.

the segmentor and critic networks. Our 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 with residual blocks/skip connections and new activation function 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, or further smoothing of the predicted label maps such as using CRFs.

Extensive experimental results demonstrate that the proposed SegAN achieves better results than other state-of-the-art CNN-based architectures, on the ISIC 2017 sub-challenge on skin lesion segmentation. The rest of this paper is organized as follows. Section 2 introduces our SegAN architecture and methodology. Experimental results are presented in Section 3. Finally, we conclude this paper in Section 4.

2. METHODOLOGY

As illustrated in Figure 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 probability label maps from input images. The task for the segmentor S is to generate the segmentation mask corresponding to an input image; the task for the critic network C is to distinguish two types of inputs: original images masked by ground truth label maps, and original images masked by predicted label maps from S. During an adversarial training process, the critic forces the segmentor to learn to generate more accurate segmentation results for training images. During testing, only the segmentor S is utilized to generate the predicted label map for a test image.

The S and C networks are alternately trained by back-propagation 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. More specifically, 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 another step using gradients computed from the same loss function passed from C to S. As training progresses, both S and C 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.

2.1. The multi-scale L_1 loss

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:

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

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 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:

$$\ell_{\text{mae}}(f_C(x), f_C(x')) = \frac{1}{L}\sum_{i=1}^{L}||f_C^i(x) - f_C^i(x')||_1 \ , \tag{2}$$

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 at the ith layer of C.

To make it harder for the critic to distinguish between predicted label maps and ground truth label maps, we want the output of segmentor S to be as close to discrete 0/1 output as possible. To this end, we replace the conventionally used sigmoid/softmax activation in the segmentor's last layer with an adaptive logistic activation function, which provides a smooth approximation to the hard step function:

$$f(z) = \frac{1}{1+e^{-z/k}} \ , \tag{3}$$

where k controls the steepness of the curve. The initial value for k is set to be 1 which makes it equivalent to the sigmoid function to provide sufficient gradient information. As it becomes smaller, this logistic function becomes sharper and more closely approximates the unit step function to generate 0/1 like output.

2.2. SegAN architecture

Segmentor. We use a fully convolutional encoder-decoder structure for the segmentor S network. The general principles of designing the segmentor in SegAN are: use a deep enough encoder to extract sufficient features from the input; use a decoder with relatively large convolution kernel to get a larger reception field and incorporate more spatial information; add skip connections such as residual blocks and concatenation between encoder and decoder to connect different level of features as well as overcome overfitting.

In the encoder, we use the convolutional layer with kernel size $7, 5, 4$ and stride 2 for downsampling to extract features from the input images. In the decoder, we perform upsampling by image resize layer with a factor of 2 and global convolutional layer [9] with kernel size $11, 9, 7$ and stride 1 to construct segmentation mask from features extracted by

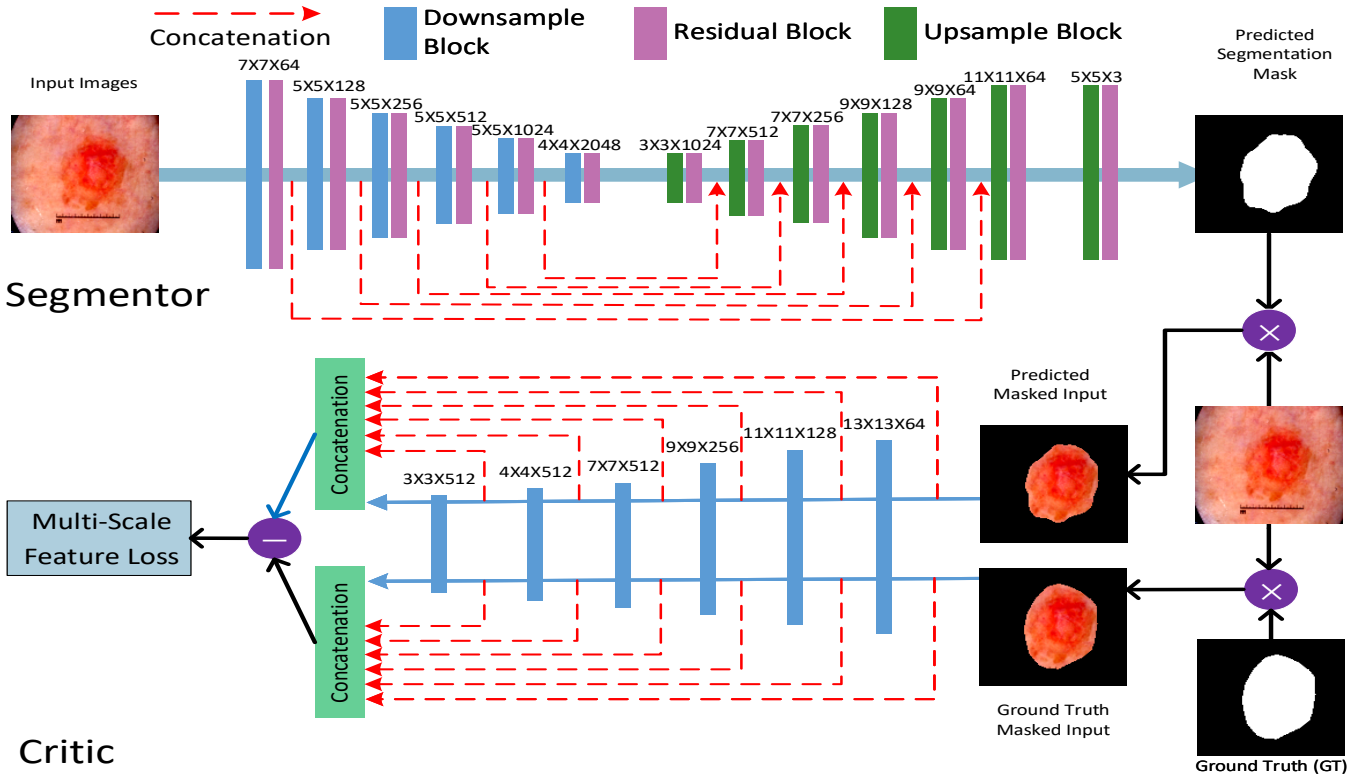

Fig. 1. The architecture of the proposed SegAN with segmentor and critic networks. For kernel size greater than 7, we use global convolution instead of normal convolution. We use bilinear image resize with a factor of 2 in each upsampling block for decoding. Masked images are calculated by pixel-wise multiplication of a label map and (multiple channels of) the corresponding input image. Note that, although only one label map (binary segmentation mask) is illustrated here, multiple label maps (i.e., multi-class labels) can be generated by the segmentor in one path. Best viewed in color.

encoder. We also add a residual block after each upsampling/downsampling block which consists of one 1×1 conv, one 3×3 conv, then followed by another 1×1 conv. Same as the U-net [5], we add skip connections between corresponding layers in the encoder and the decoder. The last layer in S is our proposed adaptive logistic activation function.

Critic. The critic C also uses global convolution to get a larger reception field with fewer parameters. It has the similar structure as the decoder in S, but with direction in reverse and without residual blocks. 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.

Note that batch normalizations are used in all blocks after convolution except the first downsampling block in both S and C; leaky ReLU is used in all downsampling blocks and regular ReLU is used in upsampling blocks after batch normalization. More details can be found in Figure 1.

2.3. Comparison with related Adversarial Networks

To the best of our knowledge, our proposed SegAN is the first adversarial training framework adapted specifically for the segmentation task that produces superior segmentation accuracy. While conventional GANs as widely used adversarial networks have been successfully applied to many unsupervised learning tasks (e.g., image synthesis [10]) and semi-supervised classification [11], there are very few works that apply adversarial learning to semantic segmentation. One such work that we found by Luc *et al.* [12] used both the conventional adversarial loss of GAN and pixel-wise cross entropy loss against ground truth. However, it is well known that the adversarial training can be "massively unstable" [13]. We also tried conventional GAN loss in segmentation and the training was indeed very unstable. We believe that the main reason contributing to the potential unstable training of previous frameworks is that: the conventional adversarial loss is based on a single scalar output by the discriminator (critic in our paper) 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 (segmentor in our paper).

In comparison, our SegAN uses a multi-scale feature loss

along with an adaptive logistic activation 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. The proof for the stability of our SegAN can be found in [14].

3. EXPERIMENTS

We evaluated our system on the ISIC dataset for ISBI 2017 Challenge on Skin Lesion Analysis Towards Melanoma Detection [15]. Since SegAN is a segmentation framework, we only focus on the Skin Lesion Segmentation sub-challenge. Specifically, we trained and validated our models using the fully annotated ISIC 2017 training dataset, which consists of 2000 dermoscopic images and the corresponding lesion masks. Since these training images have various sizes but most of them have an aspect ratio of roughly 4/3 , we first resize an input image to size 180(width)$\times 135$(height). Then we further randomly crop the input to size 128×128 during training for the purpose of data augmentation. We also randomly flip the input images horizontally and vertically with probability 0.5. ColorJitter including randomly changing the brightness, contrast, saturation and hue values of the input image is also applied for data augmentation. We did our final evaluation and comparison on the ISIC 2017 test set, which contains 600 dermoscopic images, using the evaluation metrics *Accuracy*, *Dice* and *Jaccard*.

We choose the input size to be 128×128 for the consideration of training speed and memory usage. With a smaller image size, we could have a larger batch size and faster convergence rate. We did not add drop-out during the training since the residual blocks in our segmentor have already suppressed over-fitting by having lots of weight sharing among multiple sub-networks for different levels of features [16].

We train all networks using the Adam optimizer[17] with batch size 25. The initial learning rate is set to 0.003, and a linear learning rate schedule is implemented. After each iteration for training C, the weights of our critic network are clamped to some certain range (e.g., [0.05, 0.05] for all dimensions of parameter, refer to [14] for more details). Recall that k is the steepness of our logistic function, k is annealed by a decay rate of 0.9 every 25 epochs until it reaches a certain value. We apply a thresholding to generate the final binary segmentation mask. We used a grid search method to select the optimal values for the final k and threshold value, which is 0.3 and 0.4, respectively. Replacing sigmoid with our proposed logistic function makes training converge faster and brings us more than 1% improvement in Jaccard score. The final result is generated by the ensemble of three SegANs and no further post-processing is applied.

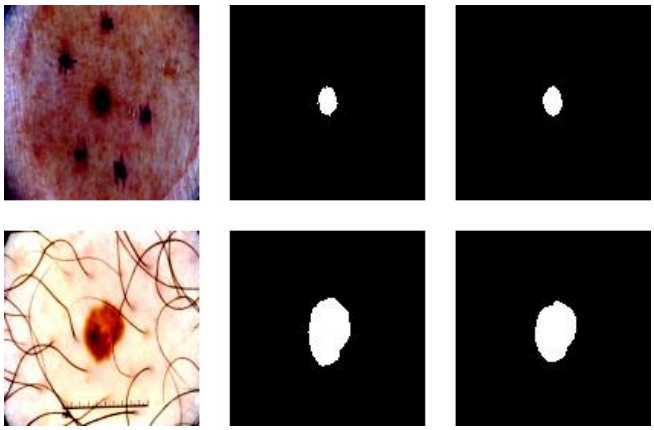

Fig. 2. Example results of our SegAN (right) with corresponding input (left) and ground truth (middle) on ISIC 2017 test set.

3.1. Comparison to state-of-the-art

In this subsection, we compare the proposed SegAN model, with other state-of-the-art methods on the ISIC 2017 Leaderboard. Table 1 gives all comparison results. Note that the ranking in the leaderboard is based on Jarccard index and we list top 3 of the leaderboard. From the table, one can see that our SegAN outperforms all previous state-of-the-art methods and boosts the highest Jarccard index from 76.5 to 78.5. Considering the small gap between these previous top 3 methods and the fact that they have produced very good segmentations, our SegAN method shows improvement by a significant margin. Another important observation is that the SegAN-produced label maps have very smooth boundary. Figure 2 illustrates some example results of our SegAN. We can see that although input images are noisy, SegAN still provides impressive segmentation results.

Table 1. Comparison to top 3 methods on ISIC 2017 leaderboard

Methods	Accuracy	Jaccard	Dice
Bi *et al.* [18]	93.4	76.0	84.4
Berseth [19]	93.2	76.2	84.7
Yuan [20]	93.4	76.5	84.9
SegAN	**94.1**	**78.5**	**86.7**

4. 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 ISIC skin lesion 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 other methods. Further, SegAN not only improves segmentation accuracy, it but also does not suffer from unstable training like other adversarial learning frameworks. In our future work, we plan to investigate the potential of SegAN for other semantic segmentation tasks.

5. REFERENCES

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[2] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[3] Guosheng Lin, Chunhua Shen, Anton van den Hengel, and Ian Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3194–3203.

[4] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[9] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," *arXiv preprint arXiv:1703.02719*, 2017.

[10] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2226–2234.

[12] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.

[13] Martin Arjovsky and Léon Bottou, "Towards principled methods for training generative adversarial networks," *arXiv preprint arXiv:1701.04862*, 2017.

[14] Yuan Xue, Tao Xu, Han Zhang, Rodney Long, and Xiaolei Huang, "Segan: Adversarial network with multi-scale l_1 loss for medical image segmentation," *arXiv preprint arXiv:1706.01805*, 2017.

[15] M. Emre Celebi Brian Helba Michael A. Marchetti Stephen W. Dusza Aadi Kalloo Konstantinos Liopyris Nabin Mishra Harald Kittler Allan Halpern Noel C. F. Codella, David Gutman, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1710.05006*, 2017.

[16] Zifeng Wu, Chunhua Shen, and Anton van den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *CoRR*, vol. abs/1611.10080, 2016.

[17] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] Lei Bi, Jinman Kim, Euijoon Ahn, and Dagan Feng, "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks," *arXiv preprint arXiv:1703.04197*, 2017.

[19] Matt Berseth, "Isic 2017-skin lesion analysis towards melanoma detection," *arXiv preprint arXiv:1703.00523*, 2017.

[20] Yading Yuan and Yeh-Chi Lo, "Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks," *arXiv preprint arXiv:1709.09780*, 2017.