International Journal of Innovative
Computing, Information and Control
Volume **14**, Number **1**, February **2018**

ICIC International ©2018 ISSN 1349-4198
pp. 53–64

# SEGMENTATION OF NERVE ON ULTRASOUND IMAGES USING DEEP ADVERSARIAL NETWORK

Cong Liu[1,*], Feng Liu[1], Lang Wang[1], Longhua Ma[1] and Zhe-Ming Lu[1,2]

[1]Ningbo Institute of Technology
Zhejiang University
No. 1, South Qianhu Road, Yinzhou Dist., Ningbo 315100, P. R. China
*Corresponding author: 1210482congliu@tongji.edu.cn

[2]School of Aeronautics and Astronautics
Zhejiang University
No. 38, Zheda Road, Hangzhou 310027, P. R. China
zheminglu@zju.edu.cn

Received June 2017; revised October 2017

Abstract. *The nerve of brachial plexus controls the sensory of human upper limb. Accurately segmenting this nerve structure on ultrasound images is the premise for upper limb surgical anaesthesia. However, auto-segmentation is extremely difficult because it demands taking account of the global anatomical dependencies and organ elastic deformation. In this work, we develop a deep adversarial neural network to overcome these difficulties. Specifically, we firstly set up a segmentation network based on well-established deep neural network. Secondly, the anatomical dependencies are ensured by an discriminator network that assesses the segmentation quality and punishes the segmentation network accordingly. Thirdly, the elastic deformation and its byproduct, small object issue, are handled by deformation data augmentation and diluted convolutions respectively. Comparing our approach to estimates made by experts in brachial plexus diagnosis shows significant performance gain over state-of-the-art models.*

**Keywords:** Ultrasound image, Brachial plexus, Segmentation, Deep learning, Convolution neural networks, Diluted convolutions, Adversarial network

1. **Introduction.** Brachial plexus is the main sensory and motor nerve of the upper limb. Blocking brachial plexus can mitigate the pain for the surgery of upper limb [1]. Ultrasound is a noninvasive and real-time imaging technology which is widely used to guide the process of brachial plexus block [2]. Accurately segmenting brachial plexus in ultrasound images is a critical step in effectively inserting a patient's pain management catheter. Manual segmentation of brachial plexus is time-consuming and highly variable. Doctors are desperate for auto-segmentation to save the time and reduce variation. However, auto-segmentation of ultrasound is extremely difficult for a number of reasons including low image quality, anatomically inadequate images, edge blur and indistinguishable characterization.

Most existing methods perform auto-segmentation without annotated images (unsupervised methods) by grouping pixels that are homogeneous in low-level features, such as color, edge or texture, into larger regions. However, these methods are limited to inferior performance due to their unsupervised nature. Recently, with the help of large-scale annotated images, supervised methods achieve supreme performance. Especially, deep convolutional neural networks (DCNNs) based methods [4] are successfully used in segmentation tasks [3, 5, 6]. The combination of powerful feature learning and fine-grain end-to-end training works very well in practice. However, as shown by Figure 1,

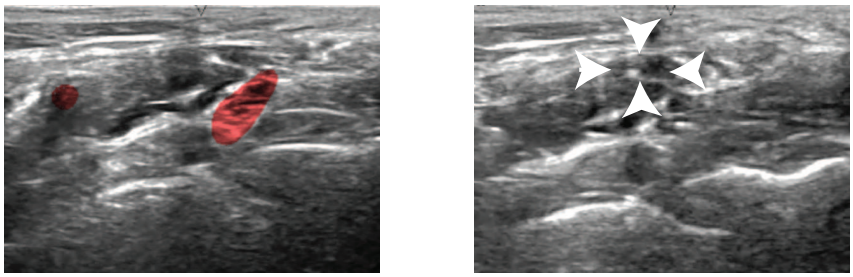

FIGURE 1. Two failed cases of brachial plexus segmentation using deep convolutional neural networks [3]. Left: superfluous brachial plexus (left red region) due to the lack of high order dependencies. Right: failed segmentation due to small brachial plexus region (surrounded by white arrows).

DCNNs-based methods show a couple of critical limitations on ultrasound segmentation task.

- Firstly, accurate segmentation usually requires anatomical contextual cues to reason the position of brachial plexus. The cues may include subclavian artery and other anatomical landmarks. However, the current DCNNs-based methods predict pixel label independently from each other and ignore the long range anatomical dependencies. As a result, the pixels that belong to the background may be mislabelled as brachial plexus, as shown in Figure 1 (left).
- Secondly, since human bodies are deformable structures, brachial plexus region may become smaller in some cases. Nevertheless, the current DCNNs-based methods will ignore or classify the small brachial plexus region as background, as shown in Figure 1 (right). This failure is caused by the reduction of image resolution happening at every layer of a standard DCNNs.

To overcome the first limitation, in this work, we develop a deep adversarial network that takes account of the long range anatomical dependencies and small object simultaneously. The deep adversarial network consists of two components: *segmentation network* and *discriminator network*. The segmentation network [3] is just one kind of DCNNs-based segmentation methods. Our main contribution is the introduction of the discriminator network which endows our model with the global anatomical consistency. The discriminator network tries to distinguish the label maps produced by the segmentation network from the ground-truth label maps. The segmentation network tries to produce label maps as real as possible to deceive the discriminator network. Since discriminator network makes decisions based on the entire label map, the global consistency will be enforced through the above adversarial training process. The adversarial network was originally designed by Goodfellow et al. [7] to generate perceptional pleasures images. However, here we use it as a regularizer to encourage the global dependencies.

We resolve the second limitation by using the dilated convolution proposed by Yu and Koltun [8]. The dilated convolution alleviates the resolution reduction issue by exponential expansion of the receptive field without loss of resolution or coverage. We then build the segmentation network out of multiple layers of diluted convolutions to improve the performance of small objection segmentation. Additionally, we use elastic transformation to augment the datasets, which enforces the deep adversarial network taking the deformable structures into account explicitly.

Our main contributions are summarized below: firstly, we present, to our best knowledge, the first application of adversarial network to ultrasound segmentation of brachial plexus; secondly, our approach is free from pixel-wise segmentation and small objection

issues; thirdly, our approach is tested and shows that it substantially outperforms the prior state-of-the-art approaches on challenging brachial plexus segmentation tasks.

The rest of this paper is organized as follows. We first review related works in Section 2 and describe the architecture of our networks in Section 3. Experimental results are demonstrated in Section 4. Finally, Section 5 concludes the paper.

2. **Related Works.** Most existing ultrasound image segmentation methods can be classified into two major categories: unsupervised methods and supervised methods. Unsupervised methods [1, 9] group the pixels that are homogeneous in low-level features (e.g., color, edge or texture) into large regions that belong to brachial plexus structure without any human intervention, i.e., annotated training examples. The unsupervised methods can perform segmentation with only one image presented. The supervised methods, by contrast, build up segmentation models with many annotated images. At the price of expensive labeling cost, supervised methods often achieve supreme performance for certain object-of-interest, whereas the unsupervised methods are often not consistent with the doctors clinical judgment.

Among the supervised methods, the state-of-the-art models following early work use DCNNs for this task by Grangier et al. [10] in 2009 and Farabet et al. [11] in 2013. Recently, fully convolutional networks (FCN) [3] have driven breakthrough on DCNNs-based segmentation. The FCN converts fully connected layers to convolution layers. This conversion allows the DCNNs to slide across pixels and predict pixel label. However, the long range dependencies are ignored, because FCN predicts pixel label independently from each other. To guarantee a consistent labelling of an image, recently several works exploit conditional random fields (CRFs) [12, 13, 14]. An alternative approach [15] uses a second CNN to learn data dependent pairwise terms. However, most of these works ensure the consistency through pairwise potentials which encourage neighbouring pixels to share the same label and omit the higher-order or global consistency. To address this issue, Pinheiro and Collobert [16] use recurrent networks to exploit the high-capacity trainable models, where each iteration maps the input image and current label map to a new label map.

Another issue of DCNNs-based methods is that they often involve a number of down-sampling layers. The aim of down-sampling layers is increasing the receptive field size, but the resolution of the output maps is reduced at the meanwhile. This poses a fundamental conflict between large receptive field and full-resolution output maps. To solve this problem, several works [3, 5, 6] propose using bi-linear interpolation, or learned up-sampling filters to upsample the output maps. Alternatively, Yu and Koltun [8] propose dilated convolutions to increase the receptive field size without losing resolution, Ronneberger et al. [6] propose skip connections to earlier high-resolution layers, and Zhou et al. [17] and Saxena and Verbeek [18] propose multi-resolution networks.

In comparison to these previous methods our work has the following merits: (i) unlike the methods [12, 13, 14, 16], our work handles the global consistency in an efficient way, because, once trained, it does not involve any higher-order CRF energy terms or recurrence in the model itself; (ii) unlike the methods [17, 18], our work supports exponential expansion of the receptive field without loss of resolution or coverage.

3. **Adversarial Network for Brachial Plexus Segmentation.** We describe the proposed method for brachial plexus segmentation in this section.

3.1. **Adversarial loss.** Core to our architecture is modifying the loss function of original segmentation network to adapt the adversarial training. The original loss function for the state-of-the-art segmentation network is cross-entropy function for multiclass classification which encourages the networks to predict correct class label at each pixel position. In

this framework's setting, only one class, brachial plexus, should be predicted. Therefore, the segmentation network actually uses binary cross-entropy as loss function. We denote the segmentation network as $S(\cdot)$ that produces class probability map $S(x) \in R^{H\times W}$ of size $H \times W$ for input brachial plexus image $x$ of size $H \times W$.

We propose using adversarial training to enforce global consistency by taking account of discriminator network $D(\cdot) \in [0, 1]$ that discriminates the predicting maps $S(x)$ produced by the segmentation network from ground-truth segmentation maps $y$. The discriminator network judges the entire or region predicting map and punishes the segmentation network if the map does not like the correct one. By this way, we hope the segmentation network can learn the high order consistency. Such consistency, for example, the contour of brachial plexus or whether the fraction of pixels in a region of a certain class exceeds a threshold, is not accessible by the standard pixel-wise binary cross-entropy loss function. Therefore, we propose adding an extra loss term $\ln[1 - D(S(x))]$ to standard loss function, i.e.,

$$\begin{aligned}\ell(\theta_S) = -\frac{1}{N}\sum_{n=1}^{N}\Bigg[\frac{1}{M}\sum_{m=1}^{M}[y_{nm}\ln(S(x_n)_m) + (1-y_{nm})\ln(1-S(x_n)_m)] \\ -\lambda\ln(1-D(S(x_n)))\Bigg]\end{aligned} \tag{1}$$

where $\theta_S$ denotes the parameters of segmentation network, $N$ denotes the sample number in a dataset, $m$ denotes a pixel in an image or a segmentation map, and there are $M$ pixels in the map. $\lambda$ is the weight balancing the pixel-wise standard loss and the adversarial loss such that both losses are on roughly the same scale. The training of the segmentation network minimizes the binary cross-entropy loss, while simultaneously degrading the performance of the adversarial networks. Therefore, the adversarial training encourages the segmentation network to produce segmentation maps that are hard to distinguish from ground-truth ones for the adversarial networks.

As with GAN, the loss function for discriminator network is defined as binary cross-entropy, i.e.,

$$\ell(\theta_D) = -\frac{1}{N}\sum_{n=1}^{N}[\ln(D(y_n)) + \ln(1 - D(S(x_n)))] \tag{2}$$

where $\theta_D$ denotes the parameters of discriminator network. The training of discriminator network maximizes the probability of assigning the correct label to both segmentation network outputs and the ground truth maps from training dataset.

3.2. **Adversarial training.** Minimizing loss functions Equation (2) and Equation (1) w.r.t $\theta_S$ and $\theta_D$ respectively can be considered segmentation network and discriminator network playing the following two-player minimax game with value function $V(\theta_S, \theta_D)$:

$$\begin{aligned}\min_{\theta_S}\max_{\theta_D} V(\theta_S,\theta_D) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{M}\Bigg[\sum_{m=1}^{M} -[y_{nm}\ln S(x_n)_m + (1-y_{nm})\ln(1-S(x_n)_m)] \\ +\lambda[\ln D(y_n) + \ln(1-D(S(x_n)))]\Bigg]\end{aligned} \tag{3}$$

As indicated by Goodfellow et al. [7], the term $\ln[1 - D(S(x))]$ in Equation (3) cannot provide sufficient gradient for training segmentation network, because this term will saturate at the beginning of the training that discriminator network has high confidence to discriminate produced maps from ground truth maps. Therefore, we replace minimizing $\ln[1 - D(S(x))]$ with maximizing $\ln[D(S(x))]$ which provides the same fixed point

of the dynamics but with much stronger gradients early in learning. We summarize the adversarial training in Algorithm 1.

**Algorithm 1:** Adversarial training algorithm

**Input** iteration T;
minibatch N;
discriminator update K;
**Result**: $\theta_S$
**while** $\theta_S$ *has not converged* **do**
  **for** $t = 0$ **to** $T$ **do**
    **for** $k = 0$ **to** $K$ **do**
      Sample minibatch of N ultrasound scans $x_n$ and corresponding ground truth masks $y_n$ from training dataset;
      Update the discriminator network by ascending its stochastic gradient:;
      $\bigtriangledown_{\theta_D} \frac{1}{N} \sum_{n=1}^{N} [-\ln(D(y_n)) - \ln D(S(x_n))]$
    **end**
    Update the segmentation network by descending its stochastic gradient:;
    $\bigtriangledown_{\theta_S} \frac{1}{N} \sum_{n=1}^{N} \Big[ \frac{1}{M} \sum_{m=1}^{M} [-y_{nm} \ln(S(x_n)_m) - (1 - y_{nm}) \ln(1 - S(x_n)_m)]$
    $+\lambda \ln(1 - D(S(x_n)))\Big]$
  **end**
**end**

As can be seen from Algorithm 1, the adversarial training algorithm consists of two stochastic gradient descents. On each iteration, a minibatch of $N$ ultrasound images and their corresponding labels are sampled from the training dataset. Then two gradient descent processes are made alternatively: one updating $\theta_D$ to reduce discriminator network loss $\ell(\theta_D)$ (i.e., Equation (1)) and one updating $\theta_S$ to reduce segmentation network loss $\ell(\theta_S)$ (i.e., Equation (2)). The gradient-based updates can use any standard gradient-based learning rule. We used momentum in the experiments. In practice, we suggest running $K$ steps of discriminator update before updating segmentation network for a more stable training process.

3.3. **Networks architectures.** For segmentation network, we use dilated convolution layers to systematically expand the the receptive field without losing resolution. The VGG-16 network module [19] is adopted as front end by removing the last two pooling and striding layers. The front end outputs feature maps at $64 \times 64$ resolution. The dilated layers module is plugged in the front end. It has 8 layers that apply $3 \times 3$ dilated convolution layers with different dilation factors: 1, 1, 2, 4, 8, 16, and 1. The factors are chosen to expand its receptive field to size $64 \times 64$ which should be consistent with the output size of the front end. The segmentation network architecture is illustrated in Figure 2.

For discriminator network, the input is ground truth label map or produced label map. The discriminator network tries to discriminate the former from the latter. The ground truth map is first converted to binary mask and then down-sampled to $64 \times 64$ resolution to match the output size of the segmentation network. We explore four architecture variants for the discriminator network. Their details are given in Section 4.4.

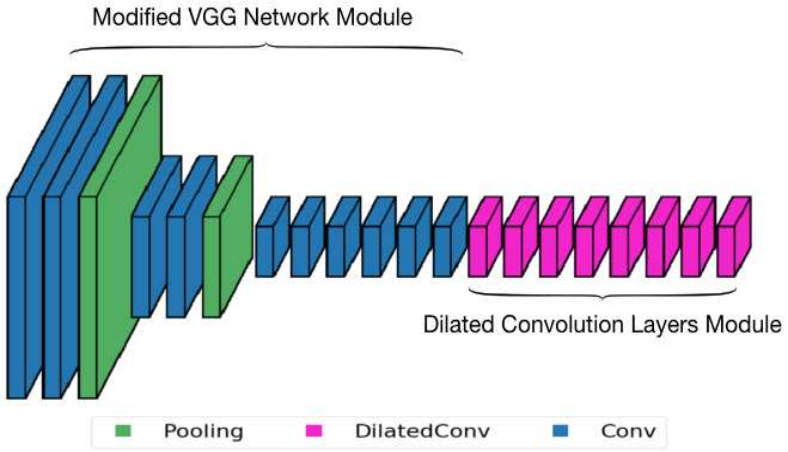


FIGURE 2. Segmentation network architecture

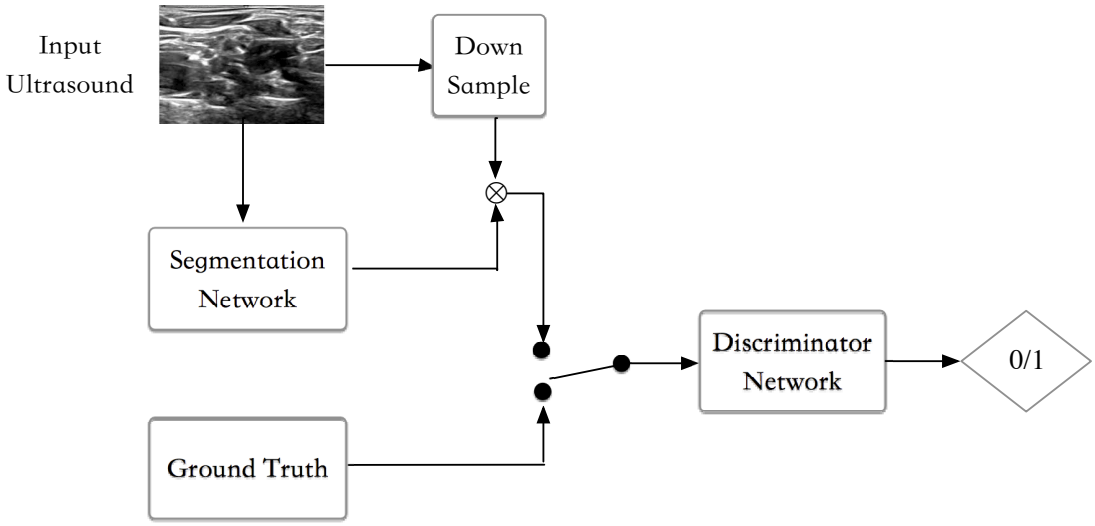


FIGURE 3. Adversarial network architecture

The adversarial network is illustrated in Figure 3, which consists of segmentation network and discriminator network. The ultrasound images are input to segmentation network to produce predicting label maps. The discriminator network tries to distinguish the label maps produced by the segmentation network from the ground-truth label maps. For better discrimination, we let the discriminator network conditioning on input image $x$, i.e., $D(x, S(x))$. Since the size of input images will be reduced by the segmentation network to $64 \times 64$, we need to down sample the input images to match this size. Three different conditioning architectures are explored. The first one is concatenating the feature map of the input image with the produced mask. The second one is multiplying the input image with each of the class probability maps (or ground truth) directly. The multiplicative interactions are designed to encode the relationships between the input image and produced mask [20], which can be viewed as a procedure of masking out non-brachial plexus region. The last one is replacing ground truth mask with distributions over the binary labels that put at least mass $\tau$ at the correct label, but are otherwise as similar as possible (in terms of KL divergence) to the distributions produced by the segmenting networks.

## 4. **Experiments.**

4.1. **Dataset.** Our segmentation module was trained on the dataset collected from ultrasound scans of brachial plexus. There are 5271 high-resolution ultrasound scans with annotations from doctors which enable large scale brachial plexus prediction for real clinic images. To evaluate the model performance, we collected another 509 scans and annotated them by a well trained physician. All ultrasound images are first pre-processed by subtracting the per-pixel mean. Training images are augmented with probability 0.5 by rotations drawn uniformly in $[-30, 30]$ degrees and elastic transformation by an affine factor drawn uniformly in $[1, 3]$ as described in [21], see Figure 4 for an example.

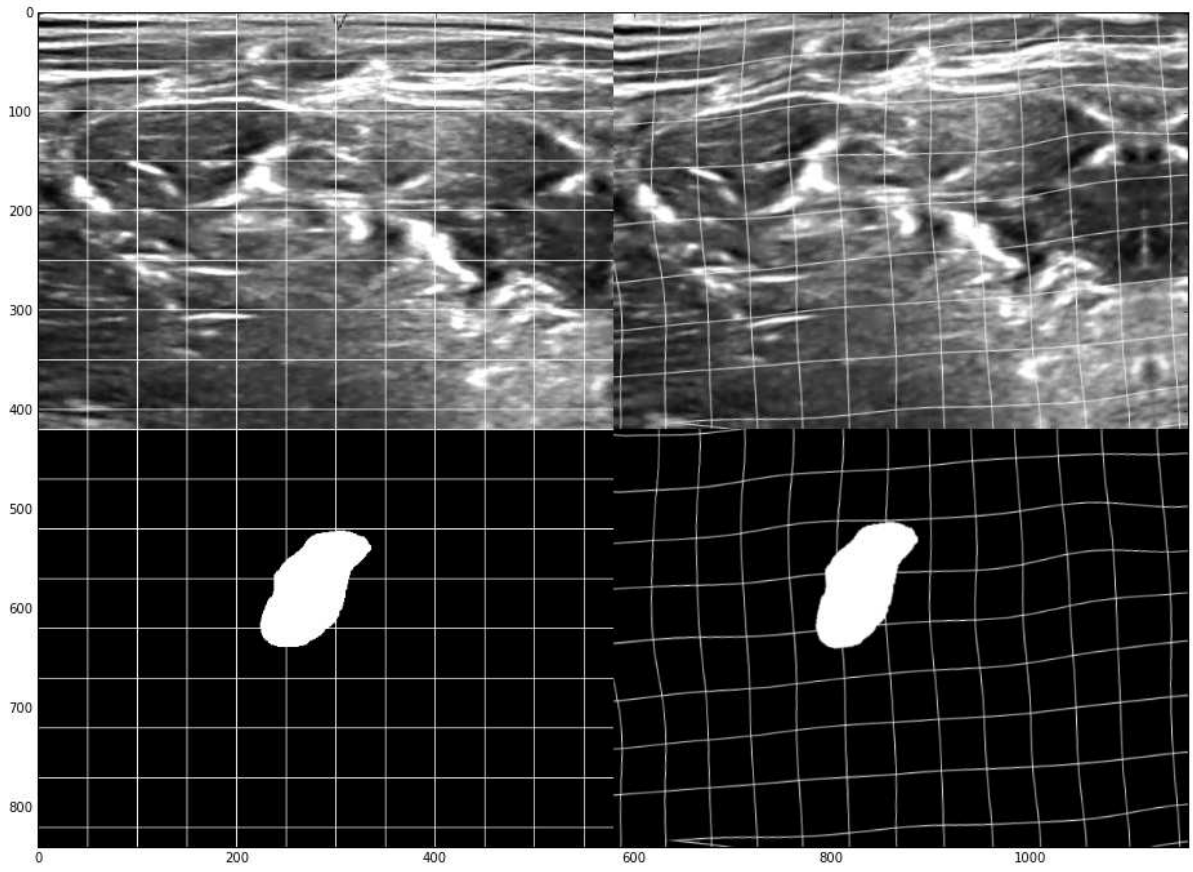


FIGURE 4. Elastic transformation for data augmentation

4.2. **Evaluation metrics.** The evaluation metric we used is standard intersection over union (IoU) as defined in [22], which can be used to compare the pixel-wise agreement between a predicted segmentation and its corresponding ground truth. The formula is given by:

$$\frac{S_{seg} \cap S_{gt}}{S_{seg} \cup S_{gt}} \tag{4}$$

where $S_{seg}$ is the predicted set of pixels and $S_{gt}$ is the ground truth. For the purpose of this paper, the contour quality of brachial plexus significantly contributes to the procedure of brachial plexus block. To take contour accuracy into consideration, we additionally use BF metric introduced by [23] to measure the boundary quality of brachial plexus which is based on the closest match between contour points in the prediction and the ground-truth segmentation. The tolerance factor $\theta$ decides whether a boundary point has a match or not. We used 0.75% of the image diagonal to set this factor.

4.3. **Training.** The segmentation network and discriminator network are trained alternatively from scratch. We also try to train the discriminator network in advance, and then use these trained networks to ensure global consistency when training segmentation network at the very beginning. However, the experimental result showed that training became rapidly unstable after just a few epochs. We then fall back Algorithm 1 to training both of them alternatively from scratch. This adversarial training was performed by stochastic gradient descent with minibatch size 100, learning rate $10^{-3}$, momentum 0.9 and adversarial loss weight $\lambda = 10^{-1}$. We have further divided the training set into ten folds, and the model was trained on all folds but one, which were used as validation set to choose hyper-parameter $\lambda$, learning rate and consequently the final model. The networks were trained for 60K iterations. The two losses defined in Equation (1) and Equation (2) will go to 0, as observed in Figure 5. The model converged when the discriminator is no longer able to differentiate and assign different values to segmentation and ground truth images.


FIGURE 5. The loss curve during the adversarial training for discriminator network (top) and segmentation network (bottom) respectively

4.4. **Comparison of architecture variants.** We explore the best discriminator network architecture in this experiment. Four architectures considered are:

- Arch 1: seven convolution layers with field-of-view size $34 \times 34$
- Arch 2: seven convolution layers with field-of-view size $34 \times 34$ and extra filter channels
- Arch 3: five convolution layers with field-of-view size $18 \times 18$
- Arch 4: five convolution layers with field-of-view size $18 \times 18$ and extra filter channels

Four different discriminator architectures with different fields-of-view and filter channels are explored here. The larger field-of-view size $34 \times 34$ is expected to be more effective to detect long-range label dependencies over larger regions, whereas the smaller field-of-view size $18 \times 18$ is expected to focus on more fine local details. The filter channels variants are designed to see if adding filter channels helps. The four architectures are further compared under concatenate, multiply and normalize variants. The experiment results are reported in Table 1. We notice that under the same field-of-view the parsimonious model (less filter channels) outperforms the lavish ones (more filter channels). This may suggest the over-fitting happened in the lavish networks. We also notice that deeper networks (larger field-of-view) show better results than the shallow ones (smaller field-of-view). We presume that this may attribute to the deeper networks with larger filed-of-view can better ensure the global consistency. Among the concatenation, multiplication and normalization variants, the multiplicative one is the most effective overall. This confirms the benefit of multiplicative interactions we discussed in Section 3.

TABLE 1. Performances comparison between different discriminator architectures

| | Concatenation | | Multiplication | | Normalization | |
|---|---|---|---|---|---|---|
| | mIOU | mBF | mIOU | mBF | mIOU | mBF |
| Arch.1 | 72.51 | **41.85** | 72.98 | 43.65 | **73.54** | 40.91 |
| Arch.2 | **72.89** | 39.91 | **73.29** | **43.82** | **73.54** | **43.37** |
| Arch.3 | 71.57 | 40.72 | 73.22 | 41.25 | 71.15 | 39.34 |
| Arch.4 | 71.62 | 40.94 | 71.33 | 39.59 | 70.71 | 40.26 |

4.5. **Effect of adversarial training.** To test if adversarial training can improve the quality of brachial plexus segmentation, we compare the segmentation results of the best networks architecture with and without adversarial training. We notice that while the adversarial training does not improve the performance significantly, it does improve the consistency between the produced segmentation and the ground truth from an anatomical correctness perspective. As illustrated in Figure 6, adding adversarial training makes the model take account of the high order context of the whole segmentation map which consequently provides a more plausible result. As presented in Table 2, we notice consistent gains for adversarial training, especially when performances are measured by mBF metric which takes the contour quality into account. Overall, this experiment confirms the benefit of adversarial training we discussed in Section 3.

4.6. **Comparison with state-of-the-art models.** To show the advantage of deep adversarial networks, we compare it to five state-of-the-art approaches. To make a fair comparison, we employ the same preprocessing pipeline such as mean subtract and data augment. The baseline in this comparison is the current best shallow model which takes the support vector machine with conditional random fields to combine unary prediction with pairwise dependencies [24]. The four other models are all deep neural networks based models including the full convolutional networks [3], the CNNs + CRF model [12], the

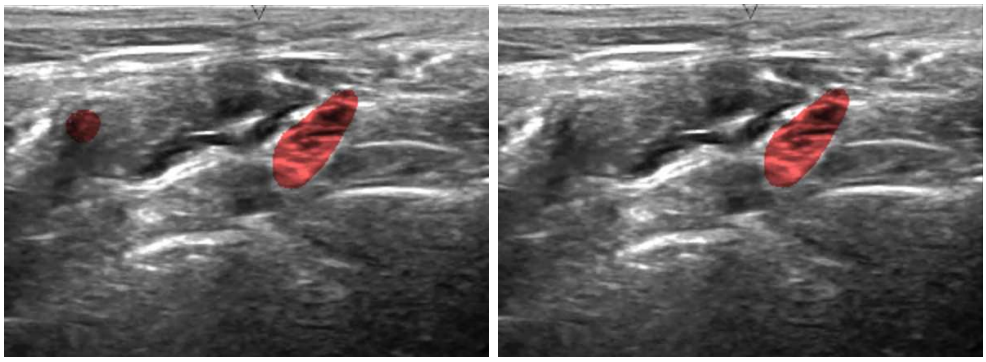

FIGURE 6. Comparison of segmentation with (left) and without (right) adversarial training

TABLE 2. Performances comparison for our model without or with adversarial training

| | mIOU | mBF |
|---|---|---|
| Without adversarial | 72.66 | 23.47 |
| with adversarial | **73.29** | **43.82** |

TABLE 3. Comparison with current models

| **Method** | **mIOU** |
|---|---|
| SVM + CRF [24] | 45.40 |
| Deep Adversarial Networks | 69.28 |
| Deep Adversarial Networks with Elastic Transformation | **73.29** |
| Full Convolutional Networks [3] | 57.18 |
| CNNs + CRF [12] | 69.30 |
| Two Convolutional Networks [15] | 71.57 |
| Convolutional Networks + Recurrent Networks [16] | 72.14 |

Two CNNs model [15] and the CNNs + RNNs model [16]. The results are reported in Table 3. The upper part shows results of the baseline. The middle part presents our models. The lower part reports the current best deep learning models.

As shown in Table 3, the proposed deep adversarial networks show consistent improvement over the baselines and four other deep models. This suggests the high order dependencies captured by the proposed model play a key role in this task. Additionally, we also note that the proposed method can be boosted significantly with the elastic transformation.

5. **Conclusion.** In this paper we described a novel approach for automatically segmenting brachial plexus structure from ultrasound images based on deep convolution neural networks. The major challenges we faced were the long range anatomical dependencies and elastic deformation. To overcome these challenges, we develop the networks that take account of the long range anatomical dependencies and full-resolution simultaneously. We find the results very promising: there is generally a good agreement between our approach's estimate and the expert's and in those cases with most disagreement, external expert opinion is in favor of the approaches result. We view our work here as a first step

toward dedicated deep learning based architecture for brachial plexus ultrasound segmentation. Future extension may include fully automating the analysis of medical imaging data across imaging modalities and 3D volume images.

**Acknowledgment.** This work is supported by National Nature Science Foundation of China under grant Nos. 61633019, 61272020; Zhejiang Provincial Natural Science Foundation under grant No. LZ15F030004; Ningbo Science and Technology Plan under grant No. 2014B82015, Ningbo Natural Science Foundation under grant No. 2015A610134; Open problem project of state key laboratory of industrial control technology under grant No. ICT170285.

## REFERENCES

[1] J. Van de Velde, J. Wouters, T. Vercauteren, W. De Gersem, F. Duprez, W. De Neve and T. Van Hoof, Morphometric atlas selection for automatic brachial plexus segmentation, *International Journal of Radiation Oncology* Biology* Physics*, vol.92, no.3, pp.691-698, 2015.

[2] J. C. Gadsden, J. J. Choi, E. Lin and A. Robinson, Opening injection pressure consistently detects needle-nerve contact during ultrasound-guided interscalene brachial plexus block, *The Journal of the American Society of Anesthesiologists*, vol.120, no.5, pp.1246-1253, 2014.

[3] E. Shelhamer, J. Long and T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.39, no.4, pp.640-651, 2017.

[4] A. Krizhevsky, I. Sutskever and G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, pp.1097-1105, 2012.

[5] H. Noh, S. Hong and B. Han, Learning deconvolution network for semantic segmentation, *Proc. of the IEEE International Conference on Computer Vision*, pp.1520-1528, 2015.

[6] O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp.234-241, 2015.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative adversarial nets, *Advances in Neural Information Processing Systems*, vol.27, pp.2672-2680, 2014.

[8] F. Yu and V. Koltun, *Multi-Scale Context Aggregation by Dilated Convolutions*, arXiv preprint arXiv:1511.07122, 2015.

[9] M. AwAN, B. A. DyER, J. Kalpathy-Cramer, E. Bongers, M. Dahele, J. Yang, G. V. Walker, N. G. Thaker, E. Holliday, A. J. Bishop et al., Auto-segmentation of the brachial plexus assessed with tactics – A software platform for rapid multiple-metric quantitative evaluation of contours, *Acta Oncologica*, vol.54, no.4, pp.562-566, 2015.

[10] D. Grangier, L. Bottou and R. Collobert, Deep convolutional networks for scene parsing, *ICML 2009 Deep Learning Workshop*, vol.3, 2009.

[11] C. Farabet, C. Couprie, L. Najman and Y. LeCun, Learning hierarchical features for scene labeling, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.35, no.8, pp.1915-1929, 2013.

[12] A. Arnab, S. Jayasumana, S. Zheng and P. H. Torr, Higher order conditional random fields in deep neural networks, *European Conference on Computer Vision*, pp.524-540, 2016.

[13] A. G. Schwing and R. Urtasun, *Fully Connected Deep Structured Networks*, arXiv preprint arXiv:1503.02351, 2015.

[14] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang and P. H. Torr, Conditional random fields as recurrent neural networks, *Proc. of the IEEE International Conference on Computer Vision*, pp.1529-1537, 2015.

[15] G. Lin, C. Shen, A. van den Hengel and I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3194-3203, 2016.

[16] P. H. Pinheiro and R. Collobert, Recurrent convolutional neural networks for scene labeling, *ICML*, pp.82-90, 2014.

[17] Y. Zhou, X. Hu and B. Zhang, Interlinked convolutional neural networks for face parsing, *International Symposium on Neural Networks*, pp.222-231, 2015.

[18] S. Saxena and J. Verbeek, Convolutional neural fabrics, *Advances in Neural Information Processing Systems*, pp.4053-4061, 2016.

[19] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv preprint arXiv:1409.1556, 2014.

[20] C. Liu, W. Xu, Q. Wu and G. Yang, Learning motion and content-dependent features with convolutions for action recognition, *Multimedia Tools and Applications*, vol.75, no.21, pp.13023-13039, 2016.

[21] P. Y. Simard, D. Steinkraus, J. C. Platt et al., Best practices for convolutional neural networks applied to visual document analysis, *ICDAR*, vol.3, pp.958-962, 2003.

[22] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn and A. Zisserman, The pascal visual object classes challenge: A retrospective, *International Journal of Computer Vision*, vol.111, no.1, pp.98-136, 2015.

[23] G. Csurka, D. Larlus, F. Perronnin and F. Meylan, What is a good evaluation measure for semantic segmentation?, *BMVC*, vol.27, 2013.

[24] A. Lucchi, Y. Li, K. Smith and P. Fua, Structured image segmentation using kernelized features, *Computer Vision – ECCV 2012*, pp.400-413, 2012.