Multi-task Fully Convolutional Network for Brain Tumour Segmentation

Haocheng Shen(✉), Ruixuan Wang, Jianguo Zhang, and Stephen McKenna

Computing, School of Science and Engineering, University of Dundee, Dundee, UK
hyshen@dundee.ac.uk

Abstract. In this paper, a novel, multi-task fully convolutional network (FCN) architecture is proposed for automatic segmentation of brain tumour. The proposed network builds on the hierarchical relationship between tumour substructures with branch and leaf losses imposed and optimised simultaneously. The network takes multimodal MR images along with their symmetric-difference images as input and extracts multi-level contextual information, firstly by the branch losses which are then fed to the leaf loss in a combination stage. The model was evaluated on BRATS13 and BRATS15 datasets and results show that the proposed multi-task FCN outperforms single-task FCN on all sub-tasks. The method is among the most accurate available and its computational cost is relatively low at test time.

Keywords: Deep learning · Tumour segmentation · Multi-task learning

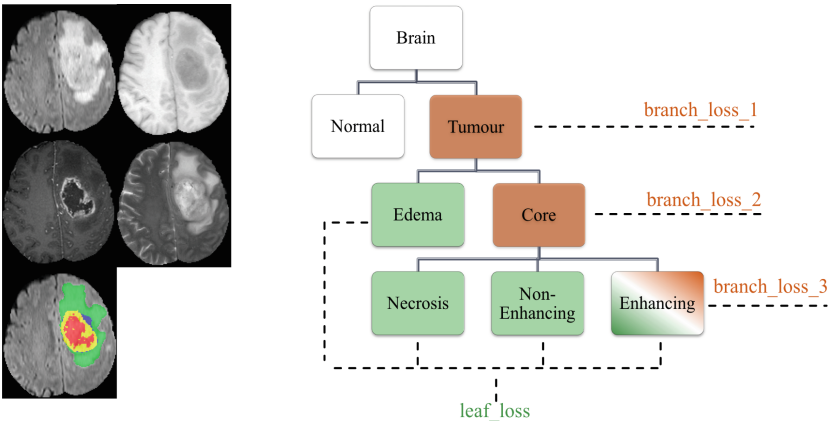
1 Introduction

Accurate localization of brain tumours in 3D MR images is clinically important for planning treatment, guiding surgery and monitoring the rehabilitation progress of patients. Unreliable segmentation risks potentially irreversible impact from surgery (e.g., difficulty in speaking fluently). Since manually segmenting brain tumour, particularly in 3D images, is a tedious and time-consuming process, computer-aided, automatic and reliable segmentation is desirable and would save clinicians' valuable time.

Among brain tumours, gliomas appear most frequently in adult patients [1] and can be graded as high grade (HG) or low grade (LG) according to aggressiveness. Due to the diversity of size, shape, location and appearance of gliomas, multimodal MRI is often used to enhance the ability to differentiate tumour and tumour substructures. Figure 1(a) shows a representative HG gliomas tumour and its sub-regions whose boundaries have been delineated by experts.

The automatic segmentation of glioma and its substructures is often formulated as a patch-level or voxel-level classification problem in which each (either 2D or 3D) patch or voxel in the 3D MR is classified as one type of substructure and the collection of all patches' or voxels' classifications generates the final,

M. Valdés Hernández and V. González-Castro (Eds.): MIUA 2017, CCIS 723, pp. 239–248, 2017.
DOI: 10.1007/978-3-319-60964-5_21



(a)Multimodal brain tumour images (b) Hierarchical labelling tree

Fig. 1. (a) An HG tumour in multimodal MRI. Flair, T1, T1c, T2 modalities (above) and expert delineation (below) showing: edema (green), necrosis (red), non-enhancing (blue), enhancing (yellow). (b) Hierarchical labelling tree of tumour tissues with corresponding *branch loss* and *leaf loss*. Note that the labels at the *leaves* (green blocks) are *mutually exclusive* whereas labels at the *branches* (brown blocks) are not. (Color figure online)

complete segmentation. While hand-crafted features and conditional random field (CRF) incorporating class-label smoothness terms have been adopted for the voxel-level classification [1,2], deep convolutional neural networks (CNNs), which have achieved substantial performance breakthroughs in several natural and medical image analysis benchmarks by automatically learning high-level discriminative feature representations, are not suprissingly achieving state-of-the-art results when applied to MRI brain tumour segmentation [3–5]. Specifically, Pereira et al. [3] trained a traditional 2D CNN as a patch-level classifier, and Havaei et al. [4] trained a 2D CNN to classify larger patches in a cascaded structure in order to capture both small and large-scale contextual information. Very recently, Kamnitsas et al. [5] trained a 3D CNN directly on 3D instead of 2D patches and considered global contextual features via an extra down-sampling path. Note that all these methods are *patch*-level classification.

Fully convolutional networks (FCNs) lack the fully connected layers often used for the last few layers in CNNs. FCNs have achieved promising results for natural image segmentation [9,10] as well as medical image segmentation [11–13]. In FCNs, up-sampling (de-)convolutional layers can be added on top of the traditional down-sampling convolutional layers in order to gain the same spatial size at the network output as at the input. Compared to CNNs applied to a sliding window on the input, FCNs can be applied to the whole input without using a sliding window and generate the classification result for each voxel (or pixel). Therefore, FCNs as voxel-level classifiers are more computationally efficient than traditional CNNs as patch-level classifiers.

In this paper, we propose a tree-structured, multi-task FCN model for brain tumour segmentation. The main contributions of our work are: (1) formulation and application of a tree-structured, multi-task FCN to multimodal brain tumour segmentation that implicitly encodes the hierarchical relationship of tumour sub-structures; (2) experiments providing evidence that the tree-structured, multi-task FCN can improve segmentation performance in all sub-tasks compared to single-task FCN on both BRATS13 and BRATS15 datasets; the proposed method is ranked top on the BRATS 2013 testing set and is more efficient than the closest competing methods.

2 Methodology

2.1 Hierarchical Labeling Tree

A tumour typically contains four sub-structures as shown in Fig. 1(a): edema (green), necrosis (red), non-enhancing (blue) and enhancing (yellow). We observe a hierarchical label relationship of tumour sub-regions, shown as a tree in Fig. 1(b). Specifically, the tree starts from a brain partitioned into non-tumour and tumour. The *complete* tumour normally consists of *edema* and *tumour core*. The *tumour core* can be further divided into *necrosis*, *non-enhancing* and *enhancing* parts. Finally, the leaves of the tree represent the five classes (including background) that are mutually exclusive (Fig. 1(b)). Encoding such a hierarchical relationship into an FCN framework can benefit tumour segmentation. For example, an enhancing part is always labeled as tumour core. We describe an FCN in a multi-task framework designed to implicitly encode the hierarchical relationship. In the following, we first describe a single-task FCN structure, upon which the proposed multi-task FCN is built.

2.2 Single-Task FCN

Our single-task FCN is a variant of FCN [9,12]. It includes a down-sampling path and three up-sampling paths, as shown in Fig. 2. The down-sampling path contains three convolutional blocks separated by max pooling (see yellow arrows in Fig. 2). Each block includes 2–3 convolutional layers similar to the VGG-16 network [6]. This down-sampling path extracts multi-scale features from low-level texture to higher-level context features. The three up-sampling paths are connected to the down-sampling path at different stages, i.e., at the last convolutional layer of each convolutional block in the downsampling path. Such a structure ensures that up-sampled feature maps are from different scales. The final feature maps in each of the three up-sampling paths (purple rectangles in Fig. 2) have the same spatial size as the input to the FCN and are concatenated before being fed to the final classification layer. ReLU activation functions and batch normalization are used after each convolutional layer. Note that the single-task FCN only considers separating the five classes at the leaf level in the hierarchical tree (i.e., a typical multi-class classification task). The efficacy of this single-task FCN was evaluated in [7].

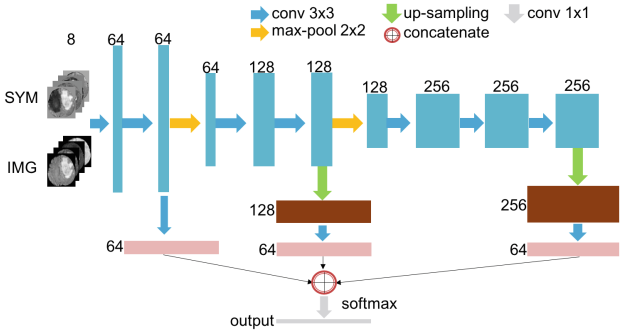


Fig. 2. Single-task FCN. Images and symmetry maps are concatenated as the input to the net [7]. Colored rectangles represent feature maps with numbers nearby being the number of feature maps. Best viewed in color. (Color figure online)

2.3 Multi-task FCN

The single-task FCN predicts the class label for each voxel. Although it can produce good probability maps its architecture ignores any hierarchical relationship shown in Fig. 1(b)). We design a multi-task FCN to implicitly encode such a relationship of tumour tissues labels. Specifically, there are two types of loss in our framework: *branch* loss and *leaf* loss (see Fig. 1(b)). The ground truth labels for branch loss (the brown blocks) are *hierarchical*, e.g., *complete tumour* contains *core* while *core* contains *enhancing* parts. On the other hand, the ground truth labels for leaf loss (the green blocks) are *mutually exclusive*. Note that the *enhancing* parts are involved in both branch loss and leaf loss. When designing a structure to match such a relationship, we also consider that the information flow runs from root to leaves. This implies that the branch loss will be applied earlier whilst leaf loss is the final layer.

The structure of the proposed multi-task FCN is illustrated in Fig. 3. We formulate the segmentation task within a multi-task learning framework, rather than treating it as a single voxel-wise classification problem. Three single-task FCNs with shared down-sampling path and three different up-sampling branches (the blue arrows in Fig. 3) are applied for three separate tasks: *complete tumour*, *tumour core* and *enhancing tumour* classification. Then, the outputs (i.e., probability maps) from the three branches are concatenated and fed to a block of two convolutional layers followed by the final softmax classification layer ('combination stage' in Fig. 3). The 'combination stage' task is a 5-class classification task whereas the others are binary classification tasks. Cross-entropy loss is used for each task. Therefore, the total loss in our proposed multi-task FCN is the sum of branch loss and leaf loss:

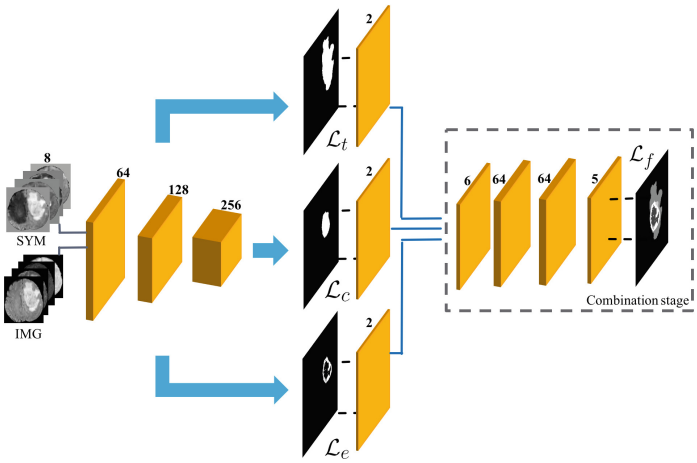


Fig. 3. The structure of multi-task FCN. The three up-sampling branches in the three FCNs are represented by blue arrows. Note that the upsampling paths are connected to the down-sampling path at different stage as described in Sect. 2.2 and Fig. 2. (Color figure online)

$$\begin{aligned}\mathcal{L}_{total}(w) &= \mathcal{L}_{leaf}(w_{leaf}) + \mathcal{L}_{branch}(w_{branch}) \\ &= \mathcal{L}_{leaf}(w_{leaf}) + \sum_{m\in\{t,c,e\}} \mathcal{L}_m(w_m) \\ &= -\sum_{m\in\{t,c,e,l\}} \sum_n \sum_i \log P_m(l_m(x_{n,i}); x_{n,i}, w_m)\end{aligned} \tag{1}$$

where $\{t, c, e, l\}$ are the tasks of *complete tumour*, *tumour core*, *enhancing core* and the leaf output by the final combination stage, respectively, and $w = \{w_t, w_c, w_e, w_l\}$ is the set of weight parameters in the multi-task FCN. $\mathcal{L}_m$ refers to the loss function of each task. $x_{n,i}$ is the i-th voxel in the n-th image used for training, and P_m refers to the predicted probability of the voxel $x_{n,i}$ belonging to class l_m.

In the proposed multi-task FCN, 2D slices from 3D MR volumes in axial view are used as part of the input to the network. In addition, since adding brain symmetry information has proved helpful for FCN based tumour segmentation [7], 'symmetric intensity difference' maps are combined with the original slices as input, resulting in 8 input channels to the network (see Figs. 2 and 3).

3 Evaluation

Our model was evaluated on BRATS13 and BRATS15 datasets. Each patient's data in the two datasets includes 4 modalities (T1, T1-contrast or T1c, T2, and

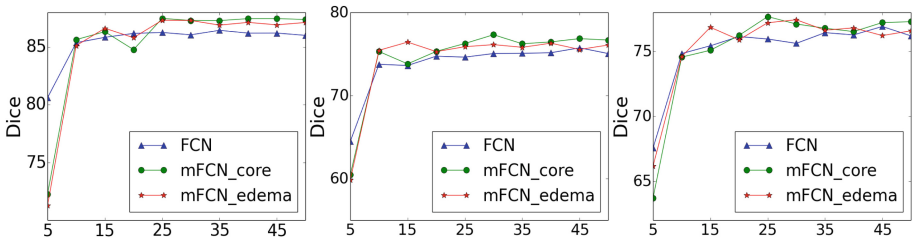


Fig. 4. Validation results of three models on BRATS13. From left to right: *Complete*, *Core* and *Enhancing* tumour task. The vertical axis is Dice while horizontal axis is the number of epochs.

Flair) which were skull-stripped and co-registered. BRATS13 contains 20 high-grade training data with known ground-truth segmentation maps and 10 high-grade testing data with ground-truth segmentation kept only by the BRATS13 organizer. (We do not use the 10 low-grade data; here we focus on high-grade tumour segmentation). For BRATS15, we used 220 released, annotated high-grade patients' images in the original training set for both training and testing. For each MR image, voxel intensities were normalised to have zero mean and unit standard deviation.

Quantitative evaluation was performed on three sub-tasks: (1) the *complete* tumour (including all four tumour sub-structures); (2) the tumour *core* (including all tumour sub-structures except "edema"); (3) the *enhancing* tumour region (including only the "enhancing tumour" sub-structure). For each sub-task, *Dice*, *Sensitivity* and *Positive Predictive Value (PPV)* were computed. Our network model was implemented in Keras with Theano as backend. The network was trained using the Adam optimizer with learning rate 0.001. The down-sampling path was initialized with VGG-16 weights [6] while up-sampling paths were initialized randomly using He's method [14].

3.1 Results on BRATS13 Dataset

A 5-fold cross validation was performed on the 20 high-grade training data in BRATS13. The training folds were augmented by scaling, rotating, and left-right flipping, resulting a dataset which was three times larger than the original one. Besides the proposed multi-task model, a variant of the proposed multi-task model was also evaluated by replacing the loss function of the *core* task with that of the *edema* task whose purpose is to segment edema. The motivation of evaluating such a variant model is from the fact that tumour *core* is a super-structure containing *enhancing*, *non-enhancing* and *necrotic* parts. These sub-structures are different in texture and appearance, e.g., in T1c (see Fig. 1) *enhancing* sub-structure shows hyper-intensity signal whereas *necrosis* has low-intensity signal. This causes large variability of *core* across patients which could be difficult for the network to model. In comparison, the texture and appearance of *edema* are relatively consistent across patients (e.g., hyper-intensity signal in

Flair). As a result, three models were evaluated on both validation set and test set: (1) single-task FCN (Fig. 2), denoted 'FCN' in the following; (2) the multi-task FCN with core task, denoted 'mFCN_core'; (3) the multi-task FCN with edema task, denoted 'mFCN_edema'.

With the validation set, Fig. 4 shows *Dice* values at every 5 epochs for each of the three models and for each of the three tasks. It can be observed that although at the starting points (e.g., the fifth epoch), mFCN_core and mFCN_edema have lower performance due to the extra parameters in the network, the highest *Dice* values achieved by mFCN_core and mFCN_edema are clearly higher than the highest *Dice* value achieved by the FCN in all the three tasks. Also, mFCN_core and mFCN_edema outperform the FCN in all three tasks at most training epochs, especially for mFCN_core. mFCN_edema gives competitive segmentation results in *complete* and *enhancing* tasks while it is slightly worse on the *core* task compared to mFCN_core, which indicates replacing *core* task by *edema* task might be unnecessary in this dataset. This could be partially due to the powerful capability of FCN to handle large appearance variability. However, mFCN_edema still outperforms FCN on all tasks, evidencing the efficacy of the tree-structured, multi-task FCN framework. The validation performances of both mFCN_core and mFCN_edema models were saturated or even decreased around 30 epochs. Therefore, models trained at 30 epochs were used for benchmarking on test data.

Further evaluation was performed on the 10 high-grade testing data (see Table 1). Here, all the 20 high-grade training data were used to train the models. The returned evaluation from the official organizer showed that both mFCN models are ranked higher than the FCN (Table 1). Due to the small size of the testing set, we observe marginal improvements in most tasks in terms of *Dice* while *Sensitivity* and PPV changed inversely (e.g., *Sensitivity* of mFCN increased over the FCN while PPV decreased a bit). Thus, we conducted a further evaluation by calculating F-scores which is the harmonic mean of *Sensitivity* and PPV for each model (see Table 3); mFCN outperformed FCN in all segmentation tasks and mFCN_core was the best on the *core* task. This conclusion is consistent with the results on the validation set.

Table 1 also shows that our proposed models are among the best of the state-of-the-art results on the BRATS13 testing set. Specifically, our models outperformed the best performers (Tustison et al. [2], Meier and Reza) from the BRATS13 challenge [1] as well as a semi-automatic method [8]. For CNN methods, our results are competitive with Pereira's et al. [3] and better than Havaei's et al. [4] while being roughly twice as fast in terms of average computational time (3 min compared to the 8 min reported by Pereira et al. [3]) due to the fast inference property of FCN. A direct comparison with 3D CNN [5] is not applicable as they did not report results on this dataset.

3.2 Results on BRATS15 Dataset

Here we randomly split 220 high-grade data in BRATS15 training set into three subsets at a ratio of 6:2:2, resulting in 132 training data, 44 validation data and 44 test data. No data augmentation was performed on this dataset. The performance curves are shown in Fig. 5.

Table 1. Comparison with the state-of-the-art on the testing set (ranked by VSD evaluation system [1])

Method	Dice			Positive Predictive Value			Sensitivity		
	Complete	Core	Enhancing	Complete	Core	Enhancing	Complete	Core	Enhancing
Pereira [3]	88	83	77	88	87	74	89	83	81
mFCN_core	88	83	75	86	85	70	91	85	83
mFCN_edema	88	82	76	85	82	72	92	85	82
FCN	87	82	75	85	87	72	89	79	80
Kwon [9]	88	83	72	92	90	74	84	78	72
Havaei [4]	88	79	73	89	79	68	87	79	80
Tustison [2]	87	78	74	85	74	69	89	88	83
Meier [1]	82	73	69	76	78	71	92	72	73
Reza [1]	83	72	72	82	81	70	86	69	76

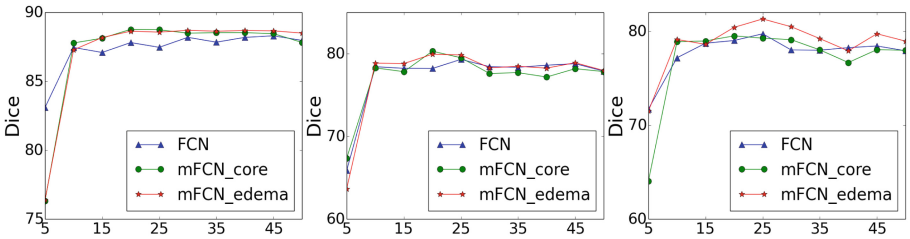


Fig. 5. Validation results of three models on BRATS15. From left to right: *Complete*, *Core* and *Enhancing* tumour task.

For the *Complete* task, both mFCN models outperform the baseline FCN. However, the mFCN_core model becomes overfitted more easily on the other two tasks. This may be due to the more powerful ability of the mFCN_core model to learn the larger appearance variability of the *Core* region in the training data, such that some of the largely varied *Core* region in the testing data may contain some new appearance or texture features which will not be well predicted by the over-trained mFCN_core model. For *Enhancing* task, mFCN_edema performs better than FCN_core and FCN, and its performance peaks at epoch 25.

On the 44 testing data, we trained the model for 25 epochs with the combination of training and validation set. The results of FCN and mFCN_edema are shown in Table 2. We found mFCN_core achieved the best results in all tasks in terms of *Dice* and *Sensitivity* as well as F1 score (see Table 3). This is consistent with the BRATS13 test result while contrary to the BRATS15 validation result where mFCN_edema seems to perform best. We might attribute this to several possible causes such as the relatively noisy ground truth in BRATS15, random initialization, unrepresentative epoch samplings or heterogeneity of data. Overall, from Table 3, we can conclude that both mFCN models appear to be better than the baseline FCN while mFCN_core is perhaps slightly better than mFCN_edema on this dataset.

Table 2. Performance on the BRATS15 44 testing set

Method	Dice			Positive Predictive Value			Sensitivity		
	Complete	Core	Enhancing	Complete	Core	Enhancing	Complete	Core	Enhancing
FCN	88.1	70.9	72.5	**92.2**	**82.7**	**79.7**	86.0	67.5	70.5
mFCN_edema	**88.5**	71.0	73.1	91.2	82.4	78.7	**87.5**	67.9	71.4
mFCN_core	**88.5**	**72.6**	**73.2**	91.1	81.3	78.2	**87.5**	**70.1**	**72.2**

Table 3. F-score on the BRATS13 and BRATS15 testing set

Method	BRATS13			BRATS15		
	Complete	Core	Enhancing	Complete	Core	Enhancing
FCN	87.0	82.8	75.8	89.0	74.3	74.8
mFCN_edema	**88.4**	83.5	**76.7**	**89.3**	74.5	74.9
mFCN_core	**88.4**	**85.0**	75.9	**89.3**	**75.3**	**75.1**

4 Conclusion

In this paper, we introduced a tree-structured, multi-task FCN for brain tumour segmentation. Our approach formulates and jointly learns the *Complete*, *Core* and *Enhancing* tumour segmentation tasks in a multi-task framework that implicitly encodes the hierarchical relationship of tumour subregions. This multi-task FCN achieved state-of-the-art results and improved segmentation in all sub-tasks on BRATS13 and BRATS15 datasets compared to the single-task FCN. Our method is among the top ranked methods and has relatively low computational cost. We would point out that the proposed multi-task network only takes the relationship between branch and leaf, and is one possible implementation of the tree in Fig. 1(b). However, the idea of imposing a loss at branch level is generic. Future work could include designing a structure to encode the hierarchy between branches.

References

1. Menze, B.H., et al.: The multimodal brain tumour image segmentation benchmark (BRATS). Med. Imaging **34**(10), 1993–2024 (2015)
2. Tustison, N.J., et al.: Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumour segmentation (simplified) with ANTsR. Neuroinformatics **13**(2), 209–225 (2015)
3. Pereira, S., et al.: Brain tumour segmentation using convolutional neural networks in MRI images. Med. Imaging **35**(5), 1240–1251 (2016)
4. Havaei, M., et al.: Brain tumour segmentation with deep neural networks. Med. Image Anal. **35**, 18–31 (2017)
5. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)

6. Simonyan, K., et al.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
7. Shen, H., et al.: Efficient symmetry-driven fully convolutional network for multimodal brain tumour segmentation (2017). Submitted to ICIP
8. Kwon, D., Shinohara, R.T., Akbari, H., Davatzikos, C.: Combining generative models for multifocal glioma segmentation and registration. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8673, pp. 763–770. Springer, Cham (2014). doi:10.1007/978-3-319-10404-1_95
9. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: CVPR 2015 (2015)
10. Chen, L.-C., et al.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint arXiv:1412.7062 (2014)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). doi:10.1007/978-3-319-24574-4_28
12. Chen, H., et al.: Deep contextual networks for neuronal structure segmentation. In: AAAI 2016 (2016)
13. Chen, H., et al.: DCAN: deep contour-aware networks for accurate gland segmentation. In: CVPR 2016 (2016)
14. He, K., et al.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: ICCV 2015 (2015)