Multi-level Activation for Segmentation of Hierarchically-Nested Classes

Marie Piraud[✉], Anjany Sekuboyina, and Björn H. Menze

Department of Computer Science,
Technische Universität München, Munich, Germany
`marie.piraud@tum.de`

Abstract. For many biological image segmentation tasks, including topological knowledge, such as the nesting of classes, can greatly improve results. However, most 'out-of-the-box' CNN models are still blind to such prior information. In this paper, we propose a novel approach to encode this information, through a multi-level activation layer and three compatible losses. We benchmark all of them on nuclei segmentation in bright-field microscopy cell images from the 2018 Data Science Bowl challenge, offering an exemplary segmentation task with cells and nested subcellular structures. Our scheme greatly speeds up learning, and outperforms standard multi-class classification with soft-max activation and a previously proposed method stemming from it, improving the Dice score significantly (p-values < 0.007). Our approach is conceptually simple, easy to implement and can be integrated in any CNN architecture. It can be generalized to a higher number of classes, with or without further relations of containment.

Keywords: Segmentation · Multiclass · Inclusion · Nested classes Class hierarchy

1 Introduction

For certain multi-class segmentation tasks, the classes have a hierarchical topological relation: one class is nested into another one, meaning that the set of pixels of the second class is spatially surrounded by pixels from the first one, as illustrated in Fig. 1(a). This is in the case of several important biological and medical image analysis tasks: anatomical structures are organized along the anatomical tree, tumors are often contained in one particular organ or anatomical structure, or intracellular features follow a specific organization within the cell. Informing the network about this type of structural relations between classes as a prior can significantly improve segmentation results, enabling the algorithm to focus on the hidden and unforseen features [1,2]. Convolutional Neural Networks

(CNNs), which have become the state-of-the-art for most image segmentation applications, have proven to be able to learn and encode very complex structures and relations between objects. However, very few CNN models are able to encode topological information as a prior.

In the literature, most of it predating the widespread use of CNNs, we distinguish three main avenues that have been pursued with that objective: (i) The first option is to use cascaded geometries [3], by training independently successive segmentation networks, the first for the surrounding class, and the second for the nested one. (ii) A second option can be to modify the loss term to penalize predictions which do not respect the expected topology, either by modifying the cross-entropy loss taking into account label-relations [2] or by integrating class relations through a specifically designed Wasserstein distance matrix in the Dice score loss [4]; both methods relying on soft-max activation. (iii) A third option is integrating label context via Conditional [5,6] and Markov [7] Random Fields that, although used as postprocessing routines in most application, can be integrated with deep learning architectures. All aforementioned methods however handle the nesting of classes in a rather indirect way – either in separate stages or through the loss, that often needs to be parametrized – and are therefore not optimally using information on class relations. Moreover, soft-max activation and cross-entropy loss assume that the classes are mutually-exclusive, as a pixel cannot belong to several classes at the same time, which does not make a natural basis for classes with hierarchical topological constraints. Applying such a standard method to segment nested-classes can lead to unreasonable results, with e.g. tumors detected outside of the organ of interest [3], or nuclei at the border of the cell (see Fig. 3), thereby limiting the quality of the results.

As a paradigm shift, we propose to consider the segmentation of hierarchically-nested classes as a generalized logistic regression problem by using a multi-level activation layer. This naturally and directly enforces the nesting of the classes, trading off neighbourhood constraints with local observations automatically and permits to segment all nested classes with a single output channel. This novel activation requires to move away from traditional cross-entropy loss, such that we introduce three adapted loss functions, and show that they all greatly speed up the learning process, and perform better than standard multi-class classification and the method from Ref. [2], on nuclei segmentation in bright-field microscopy images. We provide a second benchmark of our method on liver lesion segmentation in Computer Tomography (CT) images in the Supplementary Material. On top of being conceptually simple, the multi-level activation method is easy to implement, does not need parametrization and can be integrated in any CNN architecture.

2   Method

We start by describing our methodological contributions. We first introduce the new activation layer, and a matching thresholding scheme to infer the output segmentation map. We then propose three loss functions adapted to this activation, in Sect. 2.2.
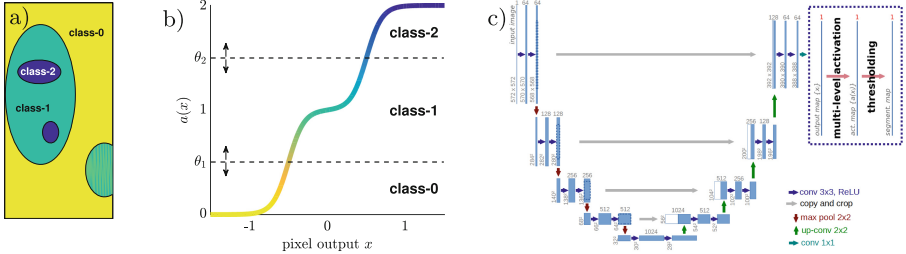
Fig. 1. *Illustration of the method.* (a) Sketch of 3 nested classes. (b) Corresponding multi-level activation [Eq. (1), with $h = 1$ and $\kappa = 10$]. (c) Multi-level activation block, which can be implemented on top of any segmentation architecture, here the U-Net [8].

2.1 Multi-level Activation Layer

Inspired by continuous regression, we propose a new multi-level activation layer, thereby generalizing logistic regression to hierarchically-nested classes [class-m $\subset$ class-$(m-1) \subset ... \subset$ class-1 $\subset$ class-0]. This activation function should have the same number of levels as the number of classes $m+1$, we therefore construct it from m equally-spaced sigmoids

$$a(x) = \sum_{n=1}^{m} \sigma \left(\kappa \left[x + h \left(n - \frac{m+1}{2} \right) \right] \right) , \tag{1}$$

where σ is the sigmoid function, κ its steepness and h the spacing between consecutive sigmoids. A similar activation function has also been introduced for unsupervised RGB image segmentation [9]. In the case of $m+1 = 3$ classes, which is illustrated in Fig. 1(a), it becomes a two-level sigmoid $a(x) = \sigma \left[\kappa(x + h/2) \right] + \sigma \left[\kappa(x - h/2) \right]$. This is illustrated in Fig. 1(b) for $h = 1$ and $\kappa = 10$, that we use in the following. This pixel-wise activation layer is designed to replace the soft-max layer of any CNN architecture, see Fig. 1(c), enabling the segmentation of nested classes with one output channel, inherently respecting their hierarchy. Note that it does not enforce the topology as a strict constraint in the segmentation map, but the hierarchy holds as long as the output map of the network remains smooth, which is the case if the resolution of the image is high enough.

Further generalizing logistic regression, we infer the output segmentation map from the activation map $a(x_i) \in [0, m]$ by setting m thresholds. For $m = 2$, class-0 is assigned to pixel i if $a(x_i) < \theta_1$, class-1 if $\theta_1 \le a(x_i) < \theta_2$, and class-2 if $\theta_2 \le a(x_i)$, see Fig. 1(b). The optimal values for $\theta_{\{1,2\}}$ can either be determined through validation or preset, e.g. to 0.5 and 1.5.

2.2 Loss Functions

Standard cross-entropy takes probability maps as input, and therefore cannot be used directly after multi-level activation, as $a(x) \in [0, m]$. To this end, we introduce different loss functions to accommodate this new activation, that are inspired both from regression and standard multi-class classification.
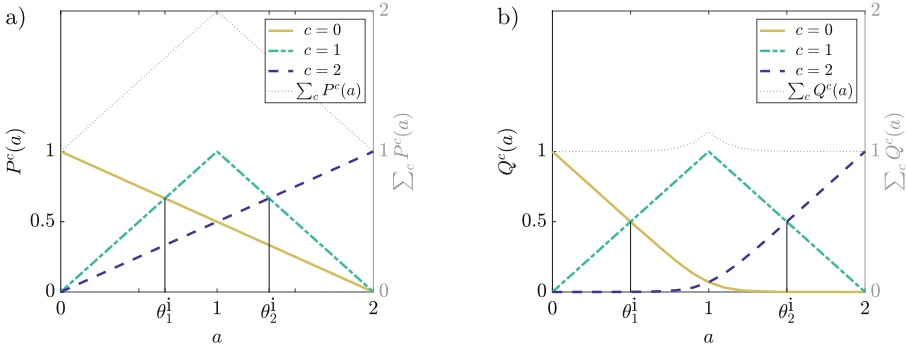
Fig. 2. *Class-'probabilities'.* Functions to map the output of the activation layer a to pseudo-probabilities. (a) $P^c(a)$ for the MCE loss, see Eq. (3), and (b) $Q^c(a)$ for the NCE loss, with $t = 10$, see Eq. (5).

Sum of Squared Error Loss. Considering the segmentation of nested classes as a regression problem, where the output map should be as close as possible to a layered-cake structure, we first propose the following Sum of Squared Error (SSE) loss

$$\mathcal{L}_{\text{SSE}} = -\frac{1}{N_{\text{tot}}} \sum_{\text{pixels } i} [a(x_i) - c_i]^2 \,, \tag{2}$$

where N_{tot} is the total number of pixels, x_i the CNN output for pixel i and $c_i \in \{0, 1, ..., m\}$ the corresponding target label, chosen consistently with Eq. (1).

Modified and Normalized Cross-Entropy Losses. Considering the problem from a multi-class classification perspective, we combine the multi-level activation with cross-entropy loss. We therefore need to map the activation $a(x) \in [0, m]$ to the interval $[0, 1]$ to mimic class-probabilities. For each class c, this mapping should peak at the target value c, which becomes the attractor during training. Focussing on the two-level case, we first propose the mapping

$$P^{c=0}(a) = 1 - a/2 \,,$$
$$P^{c=1}(a) = 1 - |1 - a| \,,$$
$$P^{c=2}(a) = a/2 \,. \tag{3}$$

Those pseudo-class-probabilities are illustrated in Fig. 2(a), but note that they are not strict probabilities, as they do not respect the addition rule ($\sum_c P^c(a) = 2 - |1 - a| \neq 1$). With this transformation, the activation map can be integrated into a Modified Cross-Entropy (MCE) loss

$$\mathcal{L}_{\text{MCE}} = -\frac{1}{N_{\text{tot}}} \sum_{\text{pixels } i} \sum_{\text{classes } c} \omega^c y_i^c \log\left(P^c[a(x_i)]\right) \,, \tag{4}$$

where $y_i^{c'} = 1$ for the ground-truth label c' of pixel i and $y_i^{c \neq c'} = 0$ otherwise. The $P^c(a)$ functions display different slopes, see Fig. 2(a), which biases the

training process towards class-1, that has a higher slope and will be favored in backpropagation. To compensate, we add the class-weights ω^c in Eq. (4), that are chosen to be proportional to the inverse of the number of pixels of each class in the training set, $\omega^c = N_{\text{tot}}/N_c$.

To make up for the drawbacks of MCE, we propose a second transformation

$$Q^{c=0}(a) = \text{s}(1 - a),$$
$$Q^{c=2}(a) = \text{s}(a - 1),\tag{5}$$

where $Q^{c=1}(a) = 1 - |1 - a|$ is unchanged, and we use the softplus function $\text{s}(x) = \frac{1}{t}\log\left(1 + e^{tx}\right)$, a smoothed version of the rectifier. Unlike in MCE, where the slopes were biased towards class-1, the $Q^c(a)$ functions, which are shown in Fig. 2(b) for $t = 10$, have the same slope around the maximum. Furthermore, they are asymptotically normalized, as we have $\sum_c Q^c(a) \to_{t\to\infty} 1$. From there we define the Normalized Cross-Entropy (NCE) loss

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{N_{\text{tot}}} \sum_{\text{pixels } i} \sum_{\text{classes } c} y_i^c \log\left(Q^c[a(x_i)]\right),\tag{6}$$

which leads to a balanced training.

The generalization to a higher number of nested classes is possible, and is presented explicitly for four classes in the Supplementary Material. We also present how MCE and NCE can be combined with standard cross-entropy, by the introduction of more output channels. For this reason, the use of those cross-entropy based losses, albeit counter-intuitive in the absence of soft-max activation and seemingly more convoluted than SSE, is of great interest: it enables the encoding of any hierarchical tree of topologically-nested and mutually-exclusive classes in a CNN.

3 Detection of Nuclei in Cells

Dataset and Training Strategy. We benchmark our method on the 2018 Data Science Bowl competition from Kaggle, whose challenge is to detect nuclei in cell images from different microscopy modalities. The simultaneous segmentation of cells and nuclei is an application of nested classes, as we have nuclei (class-2) $\subset$ cells (class-1) $\subset$ background (class-0). We select all bright-field microscopy images, on which the cells are clearly visible, and complement the available nuclei segmentation by a manual segmentation of the cell bodies (see Fig. 3). Another challenge from this competition is to deal with the limited number of training images in this modality ($N_{\text{im}} = 16$). We therefore rely on online data augmentation, and perform random flips, warping, rotations, translations and rescaling of the images at each training epoch. The images, resized to 512×512 pixels, are then fed into a U-Net like architecture [8]. We perform 4-fold cross-validation and a rotating testing scheme: we split the dataset into 4 subsets to perform cross-validation and further split each validation set into 4, for testing,
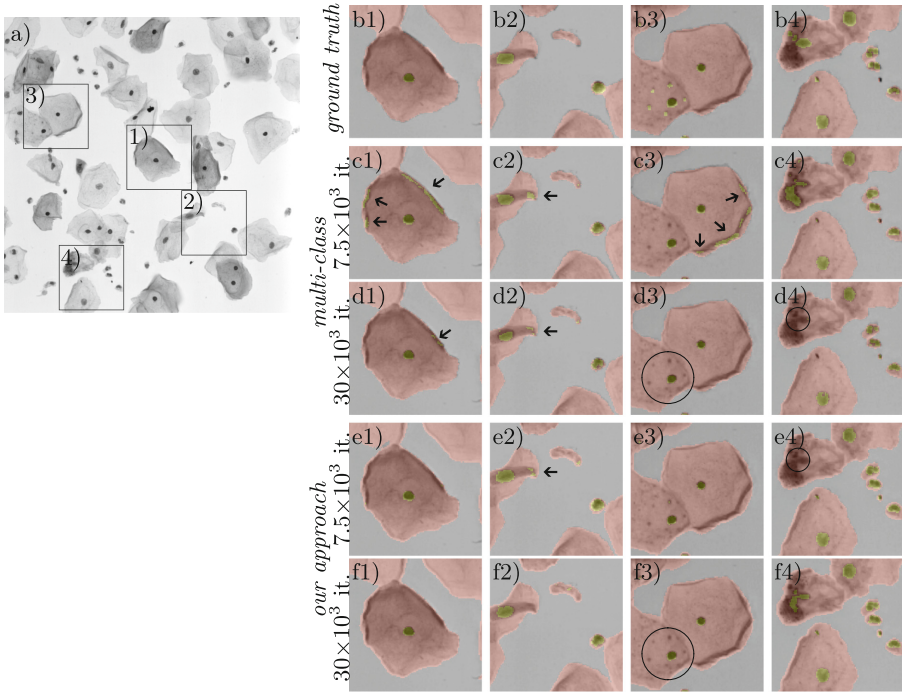
Fig. 3. *Examples of nuclei segmentation.* (a) Selected regions of a bright-field microscopy image are overlaid with (b) the ground truth segmentation, (c–d) results from standard multi-class segmentation after 7.5 and 30×10^3 iterations respectively, and (e–f) results from multi-level activation with MCE loss, our best performing method, after 7.5 and 30×10^3 iterations respectively. The cell- (respectively nuclei-) class is displayed in red (resp. green). Arrows indicate false positives and circles false negatives. (Color figure online)

such that we always train on 12 images, validate on 3 and test on one, in a 16-fold rotating fashion. For each fold, we select the threshold value θ_2 giving the best Dice score on the validation set and compute the scores of the test image after 30×10^3 iterations. Note that due to online data augmentation, the model does not tend to overfit.

Results. Qualitative results for standard multi-class segmentation and multi-level activation with MCE loss, after 7.5×10^3 iterations and after convergence, are presented in Fig. 3. Early on in the training process, multi-class segmentation outputs false positives for the nuclei class, in darker regions of the cells, even very close to the cell border, see Figs. 3(c1)–(c3). Those slowly disappear in the training process, as the CNN learns the relationship between classes, although some imperfections remain, see Figs. 3(d1)–(d2). This is strongly disfavored by multi-level activation, as the output map $\{x_i\}$ would have to oscillate on very short length scales, and we see that most such artefacts do not appear on Figs. 3(e)–(f).
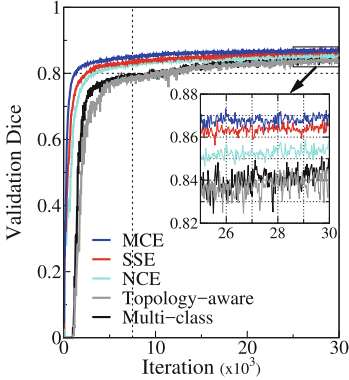
Fig. 4. *Scores for nuclei segmentation.* Mean validation Dice scores for the methods listed in the table (rows I, II and VI to VIII).

Table 1. *Scores for nuclei segmentation.* Mean test Dice scores for standard multi-class (row I), the 'topology-aware' loss [2] (row II) and multi-level activation (rows III-VIII). In rows III to V we use the threshold θ_2^i. In rows VI to VIII, θ_2 is selected during validation and we report its mean value. Grey cells highlight significant improvement over multi-class, i.e. p-values < 0.05 using the paired samples Wilcoxon test.

Model	θ_2	Test Dice
I - Multi-class	−	0.839 (0.055)
II - Topology-aware [2]	−	0.842 (0.055)
III - NCE, Eq. (6)	$\theta_2^i = 1.5$	0.841 (0.058)
IV - SSE, Eq. (2)	$\theta_2^i = 1.5$	0.863 (0.051)
V - MCE, Eq. (4)	$\theta_2^i = 4/3$	0.844 (0.049)
VI - NCE, Eq. (6)	1.31 (0.09)	0.853 (0.056)
VII - SSE, Eq. (2)	1.39 (0.08)	0.863 (0.053)
VIII - MCE, Eq. (4)	1.74 (0.05)	0.868 (0.051)

On the contrary, the spatial regularisation provided by multi-level activation can lead to false negatives at the beginning of the training process, as on Fig. 3(e4), which are then detected by the converged model, see Fig. 3(f4). The circles on Figs. 3(d3) and (f3) outline difficult cases which are missed by both methods.

We now compare quantitatively two soft-max based methods – standard multi-class segmentation and the 'topology-aware' method from Ref. [2] – with our multi-level activation layer, combined with the three losses introduced in Sect. 2.2. The validation Dice scores for the nuclei class are shown in Fig. 4 for each method of this benchmark. The corresponding test scores are reported in Table 1, for (i) a pre-determined value θ_2^i and (ii) the value of θ_2 selected during validation. In (i), we choose the *a priori* inferred value $\theta_2^i = 1.5$ for NCE and SSE, and $\theta_2^i = 4/3$ for MCE, the values where P^1 and P^2, respectively Q^1 and Q^2, intersect (see Fig. 2). While all methods perform comparatively well for cell segmentation, with mean test Dice scores ranging from 0.977 (0.01) to 0.979 (0.01), there are quantitative discrepancies in the nuclei segmentation. The two soft-max based methods perform on par: we report a mean test Dice score of 0.839 (0.055) for standard multi-class, which we consider as our baseline in the following, and of 0.842 (0.055) for the 'topology-aware' loss [2]. Indeed, multi-class is not plagued by the detection of nuclei outside cells, but rather outputs false positives near the border of the cell, as we have seen above, a problem which is not addressed by the method of Ref. [2].

In Fig. 4, we see that our three proposed methods converge much faster, and outperform the soft-max based ones during validation. Our methods indeed cross the 0.8 validation Dice score after 1.3 to 3.2×10^3 iterations, whereas the soft-max-based methods need more than three times as much iterations to achieve

this accuracy. Using θ_2^i, we improve the test Dice score by 0.2 and 0.5 for $\mathcal{L}_{\mathrm{NCE}}$ and $\mathcal{L}_{\mathrm{MCE}}$ and by 2.4 points with $\mathcal{L}_{\mathrm{SSE}}$. The paired samples Wilcoxon test gives a p-value of 0.0008 for $\mathcal{L}_{\mathrm{SSE}}$ vs multi-class, confirming the significance of this improvement. Threshold selection during validation improves the $\mathcal{L}_{\mathrm{NCE}}$ and $\mathcal{L}_{\mathrm{MCE}}$ results further, exceeding the multi-class score by 1.4 and 2.9 Dice points respectively (with p-values 0.007 and 0.001). $\mathcal{L}_{\mathrm{MCE}}$ therefore gives the best model of this benchmark. Note that using $\mathcal{L}_{\mathrm{MCE}}$ without reweighting [i.e. $\omega^c = 1$ in Eq. (4)] strongly undersegments nuclei, confirming that the training is then biased towards class-1, as anticipated in Sect. 2.2. Furthermore, with $\omega^c \propto 1/N_c$, the thresholds selected by validation, with mean $\theta_2^{\mathrm{mean}} = 1.74$, significantly differ from $\theta_2^i = 4/3$, indicating that class-2 might now be overfavored. We retrospectively verified that all other models do not benefit from the application of the weighting scheme $\omega^c \propto 1/N_c$.

Discussion. Our proposed multi-level activation layer, greatly speeds up learning and outperforms the soft-max based methods. It indeed permits to significantly improve the nuclei test Dice scores in all cases (with p-values < 0.007). This novel activation layer introduces thresholds in the multi-class classification context, which can be adjusted at validation time, leading to a significant performance gain for $\mathcal{L}_{\mathrm{NCE}}$ and $\mathcal{L}_{\mathrm{MCE}}$. But this is not the only benefit of our method, as all proposed loss functions outperform multi-class without threshold adjustment, significantly for $\mathcal{L}_{\mathrm{SSE}}$, proving that our regression-like method is better suited to the problem.

4 Conclusion

In this work, we have proposed a new paradigm for multi-class segmentation with topological constraints of inclusion. It consists in a novel multi-level activation layer and three matching loss functions, based on regression and cross-entropy loss. This scheme can be implemented in any network architecture, with minimal changes. We benchmarked our method on the segmentation of nuclei in bright-field microscopy images, giving significant improvement and speed-up over the soft-max based methods. In the Supplementary Material, we provide a second benchmark on liver lesions segmentation, in a larger and more imbalanced dataset, with the same conclusions. We expect those results to transfer to other tasks with nested classes, as nothing was handcrafted for the problems at stake. $\mathcal{L}_{\mathrm{MCE}}$ turned out to be the best performing loss in this paper, but the other losses also perform well, and might be better suited for different applications.

Informing the network on the relations between classes with the multi-level activation thus permits to train on less data, which is often crucial in biomedical applications. Finally, as shown in the Supplementary Material, the multi-level activation layer and the associated losses can be straightforwardly generalized to a deeper nesting hierarchy. We also show how $\mathcal{L}_{\mathrm{MCE}}$ and $\mathcal{L}_{\mathrm{NCE}}$ can be used alongside normal cross-entropy to segment nested classes together with further classes without topological prior. This enables the encoding of any tree of prior relations of containment between classes.

References

1. Nosrati, M.S., Hamarneh, G.: Local optimization based segmentation of spatially-recurring, multi-region objects with part configuration constraints. IEEE Trans. Med. Imaging **33**(9), 1845–1859 (2014)

2. BenTaieb, A., Hamarneh, G.: Topology aware fully convolutional networks for histology gland segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 460–468. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_53

3. Christ, P.F., et al.: Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 415–423. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_48

4. Fidon, L., Li, W., Garcia-Peraza-Herrera, L.C., Ekanayake, J., Kitchen, N., Ourselin, S., Vercauteren, T.: Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 64–76. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_6

5. Bauer, S., Tessier, J., Krieter, O., Nolte, L.-P., Reyes, M.: Integrated spatio-temporal segmentation of longitudinal brain tumor imaging studies. In: Menze, B., Langs, G., Montillo, A., Kelm, M., Müller, H., Tu, Z. (eds.) MCV 2013. LNCS, vol. 8331, pp. 74–83. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05530-5_8

6. Alberts, E., Charpiat, G., Tarabalka, Y., Huber, T., Weber, M.-A., Bauer, J., Zimmer, C., Menze, B.H.: A nonparametric growth model for brain tumor segmentation in longitudinal MR sequences. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Handels, H. (eds.) BrainLes 2015. LNCS, vol. 9556, pp. 69–79. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30858-6_7

7. Liu, Z., Li, X., Luo, P., et al.: Deep learning Markov random field for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **8828**(c), 1 (2017)

8. 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). https://doi.org/10.1007/978-3-319-24574-4_28

9. Bhattacharyya, S., Dutta, P., Nandi, P.K.: Multilevel activation functions for true color image segmentation using a self supervised parallel self organizing neural network (PSONN) architecture: a comparative study. Int. J. Comput. Sci. **1**(8), 2452–2464 (2007)