DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks

Martin Rajchl,* Matthew C. H. Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A. Rutherford, Joseph V. Hajnal, Bernhard Kainz, and Daniel Rueckert

***Abstract*— In this paper, we propose *DeepCut*, a method to obtain pixelwise object segmentations given an image dataset labelled weak annotations, in our case bounding boxes. It extends the approach of the well-known *GrabCut* [1] method to include machine learning by training a neural network classifier from bounding box annotations. We formulate the problem as an energy minimisation problem over a densely-connected conditional random field and iteratively update the training targets to obtain pixelwise object segmentations. Additionally, we propose variants of the *DeepCut* method and compare those to a naïve approach to CNN training under weak supervision. We test its applicability to solve brain and lung segmentation problems on a challenging fetal magnetic resonance dataset and obtain encouraging results in terms of accuracy.**

***Index Terms*— Bounding box, convolutional neural networks, DeepCut, image segmentation, machine learning, weak annotations.**

I. Introduction

MANY modern medical image analysis methods that are based on machine learning rely on large amounts of annotations to properly cover the variability in the data (*e.g.* due to pose, presence of a pathology, *etc*). However, the effort for a single rater to annotate a large training set is often not feasible. To address this problem, recent studies employ forms of weak annotations (*e.g.* image-level tags, bounding boxes or scribbles) to reduce the annotation effort and aim to obtain comparably accurate results as to under full supervision (*i.e.* using pixelwise annotations) [2]–[4].

User-provided bounding boxes are a simple and popular form of annotation and have been extensively used in the field of computer vision to initialise object segmentation methods [1], [5]. Bounding boxes have advantages over other forms of annotations (*e.g.* scribbles or brush strokes [6]–[8]), as they allow to spatially constrain the problem (*i.e.* ideally, the object is unique to the bounding box region and fully contained in it). In a practical sense, bounding boxes can be defined via two corner coordinates, allowing fast placement (approximately 15 times faster than pixelwise segmentations [9]) and lightweight storage of the information. Considering the required interaction effort and the amount of provided information, these properties qualify bounding boxes as preferred weak annotation for image analysis methods.

In segmentation studies [1], [5], [10], bounding boxes are employed as both initialisation and spatial constraints for the segmentation problem. The above approaches model the image appearance (*i.e.*, colours or greyscale intensity) and impose smoothness constraints upon the segmentation results for each image. However, given an image database and corresponding annotations, we can assume that objects share common shape and appearance information, which can be learned (*i.e.* instead of direct image-by-image object segmentation, a common model can be learned for the all images in the database). This is particularly interesting for segmentation problems on medical images, where typically an entire cohort is to be analysed for a specific organ or region, exhibiting large class similarity in terms of shape and appearance.

In this paper, we propose combining a neural network model with an iterative graphical optimisation approach to recover pixelwise object segmentations from an image database with corresponding bounding box annotations. The idea builds on top of the popular GrabCut [1] method, where an intensity appearance model is iteratively fitted to a region and subsequently regularised to obtain a segmentation. Similarly to this, the proposed *DeepCut* method iteratively updates the training targets (*i.e.* the class associated with a voxel location, described by an image patch) learned by a convolutional neural network (CNN) model and employs a fully connected conditional random field (CRF) to regularise the segmentation.

Manuscript received June 5, 2016; revised October 15, 2016; accepted October 18, 2016. Date of publication November 9, 2016; date of current version February 1, 2017. This work was supported by Wellcome Trust and EPSRC IEH award [102431] for the iFIND project and the Developing Human Connectome Project, which is funded through a Synergy Grant by the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement number 319456. This research was also supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. *Asterisk indicates corresponding author.*

*M. Rajchl is with the Department of Computing, Imperial College London, London, SW7 2AZ, U.K. (e-mail: m.rajchl@imperial.ac.uk).

M. C. H. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, B. Kainz, and D. Rueckert are with the Department of Computing, Imperial College London, London, SW7 2AZ, U.K.

M. Damodaram is with Queen Charlotte's Fetal Medicine Department, Hammersmith Hospital and Imperial College London, London, W12 0HS, U.K.

M. A. Rutherford and J. V. Hajnal are with the Department of Biomedical Engineering, King's College London, London, WC2R 2LS, U.K.

Digital Object Identifier 10.1109/TMI.2016.2621185

The approach is formulated in a generic form and thus can be readily applied to any object or image modality. We briefly review recent advancements in the following section to put this approach into the context of the current state-of-the-art and highlight our contributions.

A. Related work

Graphical energy minimisation techniques are popular methods for regularised image segmentation due to inherent optimality guarantees and computational efficiency [11], [12]. They have been extensively used in the optimisation of interactive [6], [8], [13], [14] and fully automated segmentation methods [15]–[18].

An iterative optimisation of such energy functionals allows to address more complex problems, such as the pixelwise segmentation from bounding box annotations. The popular GrabCut method [1] iteratively updates parameters of a Gaussian mixture model (GMM) and optimises the resulting energy with a graph cut. Lempitsky *et al.* [5] extended this method by adding a topological prior to prevent excessive shrinking of the region and Cheng et al. [10] improved upon its performance by employing a fast fully connected CRF [19]. Similar approaches include the time-implicit propagation of levelsets via continuous max-flow [20], [21], iterative graph cuts [22] and the use of the expectation-maximisation (EM) algorithm [2], [23], [24].

Similarly to the above mentioned segmentation approaches, learning-based methods have been recently investigated to exploit the advantages of weak annotations, primarily to reduce the effort of establishing training data. In contrast to learning under full supervision (*i.e.* using pixelwise annotations), weakly supervised methods aim to learn from image-level tags, partial labels, bounding boxes, *etc.* and infer pixelwise segmentations.

Recently, several multiple instance learning (MIL) techniques were investigated, particularly when images could potentially contain multiple objects. Cinbis *et al.* [25] proposed a multi-fold MIL method to obtain segmentations from image level tags. Vezhnevets and Buhmann [26] addressed the problem with a Texton Forest [27] framework, extending it to MIL. With the re-emerging of convolutional neural networks [28], [29], MIL methods have been proposed to exploit such methods [30], [31]. However MIL-based methods, even when including additional modules under weak supervision [31], have not been able to achieve comparable accuracy to fully supervised methods [2]. Latest developments using CNN learning with weakly supervised data have shown remarkable improvements in accuracy. Schlegl et al. [4] parse clinical reports to associate findings and their locations with optical coherence tomography images and further obtain segmentations of the reported pathology. Dai et al. [3] iterate between updating region proposals in bounding boxes and model training. Papandreou et al. [2] formulate an Expectation-Maximization (EM) algorithm [23] to iteratively update the training targets. Both of the latter methods were able to achieve comparable accuracy to those under full supervision.

B. Contributions

In this paper, we build upon the ideas of GrabCut [1], a well-known object segmentation method employed on single images. We extend the basic idea with recent advances in CNN modelling and propose *DeepCut*, a method to recover semantic segmentations given a database of images with corresponding bounding boxes. For this purpose, we formulate an iterative energy minimisation problem defined over a densely connected conditional random field (CRF) [19] and use it to update the parameters of a CNN model. We compare the proposed method against a fully supervised (CNN_{FS}) and a naïve approach to weakly supervised segmentation ($\text{CNN}_{\text{naïve}}$), to obtain upper and lower accuracy bounds for a given segmentation problem. Further, we examine the effect of region initialisation on the proposed method by providing a pre-segmentation within the bounding box, (DC_{PS}). Finally, we compare all methods in their segmentation accuracy using a highly heterogeneous and challenging dataset of fetal magnetic resonance images (MRI) and further evaluate the performance to GrabCut [1], as an external method to this framework.

II. Methods

DeepCut falls into a class of iterative optimisation methods, most closely related to the well-known GrabCut method [1]. There are two key stages to both algorithms: model estimation and label update.

GrabCut uses a GMM to parametrise the colour distributions of the foreground and background. In the model estimation stage the parameters Θ for the GMM are computed based on the current label assignment f for each pixel i. At the label update stage the pixels are relabelled based on the new model. *DeepCut* replaces the GMM with a Neural Network model and the graph cut solver from [11] with [19] on a densely-connected graph. In contrast to [1], and rather than recomputing our model, we make use of transfer learning [32] and reinitialise the CNN with the parameters of the last iteration. Similarly, [2] and [3] describe methods to iteratively update f, however only employ regularisation as a post-processing step during testing.

A. Segmentation via Iterative Energy Optimisation

Let us consider problems using energy functionals over graphs as described in [11]. We seek a labelling f for each pixel i, minimising

$$E(f) = \sum_i \psi_u(f_i) + \sum_{i<j} \psi_p(f_i, f_j), \tag{1}$$

where $\psi_u(f_i)$ serves as unary data consistency term, measuring the fit of the label f at each pixel i, given the data. Additionally, the pairwise regularisation term $\psi_p(f_i, f_j)$ penalises label differences for any two pixel locations i and j. Typically, pairwise regularisation terms have the form of

$$\psi_p(f_i, f_j) \propto \exp\left(-\frac{(I_i - I_j)^2}{2\theta_\beta}\right), \tag{2}$$

and enforce contrast-sensitive smoothness penalties between the intensity vectors I_i and I_j [6], [17]. We can minimise the

energy in Eq. (1) using a densely-connected CRF [19], where we replace the pairwise term with

$$\psi_p(f_i) = g(i, j)[f_i \neq f_j], \tag{3}$$

consisting of two penalty terms modelling appearance (4a) and smoothness (4b) between the locations p_i and p_j:

$$g(i, j) = \omega_1 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) \tag{4a}$$

$$+ \omega_2 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right). \tag{4b}$$

The relative contributions of the two penalties is weighted by the regularisation parameters ω_1 and ω_2 and the degrees of spatial proximity and similarity are controlled by θ_α and θ_β, respectively [19].

The unary potential is computed independently for each pixel i by a data model with the parameters Θ that produces a distribution y_i over the label assignment given an input image or patch x and is defined as the negative log-likelihood of this probability:

$$\psi_u(f_i) = -\log P(y_i|x, \Theta). \tag{5}$$

In contrast to [1], where the unary term is computed from a GMM of the observed colour or intensity vector, we employ a CNN with the parameters Θ. We describe the network architecture in Section II-B in detail.

B. Convolutional Neural Network Model

The CNN is a hierarchically structured feed-forward neural network comprising of least one convolutional layer [28], [29]. Additionally, max-pooling layers downsample the input information to learn object representations at different scales. Downstream of several convolutional and max-pool layers is typically a layer of densely-connected neurons, reducing the output to a desired number of classes [4]. Our CNN is a typical feed-forward neural network model which consist of convolutional layers (feature extraction), max-pooling layers (shift/scale invariance) and dense layers (classification stage). The network architecture used in this study is depicted in Fig. 1.

1) Input & Output Space: Given a database of size N containing images $I = \{I_1, ..., I_N\}$ and corresponding bounding boxes $B = \{B_1, ..., B_N\}$, we attempt to learn pixelwise segmentations of the objects depicted in I and constrained to B. For this purpose, we employ a CNN with parameters Θ to classify image patches centred around a voxel location i into foreground and background.

We describe each voxel location $i \in I_n$ as a 3D patch of size $p_x \times p_y \times p_z$, centred around X. Each patch X is associated with an integer value $Y = \{0, 1\}$, representing background and foreground class of the centre voxel at i, respectively. The patch X serves as input to the network, which aims to predict Y. Because of anticipated motion artefacts between fetal MR slices, we decided to emphasise the in-plane context of our training patches, using a patch size of 33x33x3, rather than cubic patch dimensions.

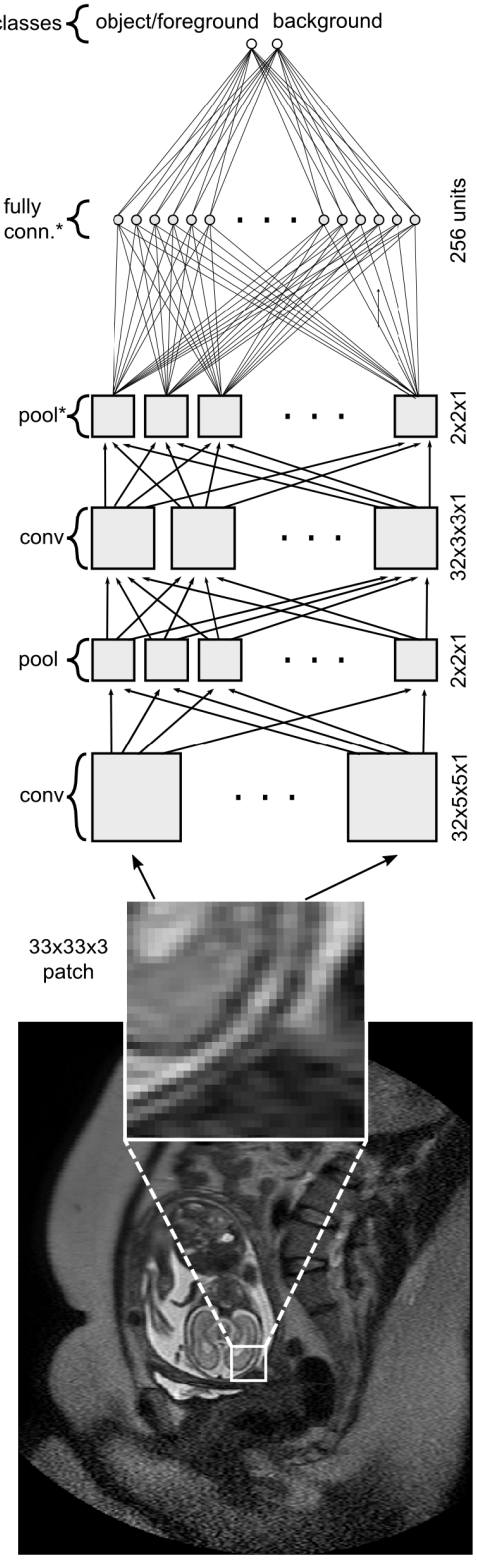


Fig. 1. CNN architecture with convolutional (conv), max-pooling (pool) layers, and fully connected layers for foreground/background classification. Layer, which inputs are subjected to 50% dropout [33] are marked with *.

2) Network Configuration: For the purpose of this study, we designed a simple CNN inspired by the well-known *LeNet* architecture [34]. While the proposed approach can employ other networks, we chose this configuration, as it is easily understandable and simple to reproduce. The CNN consists of

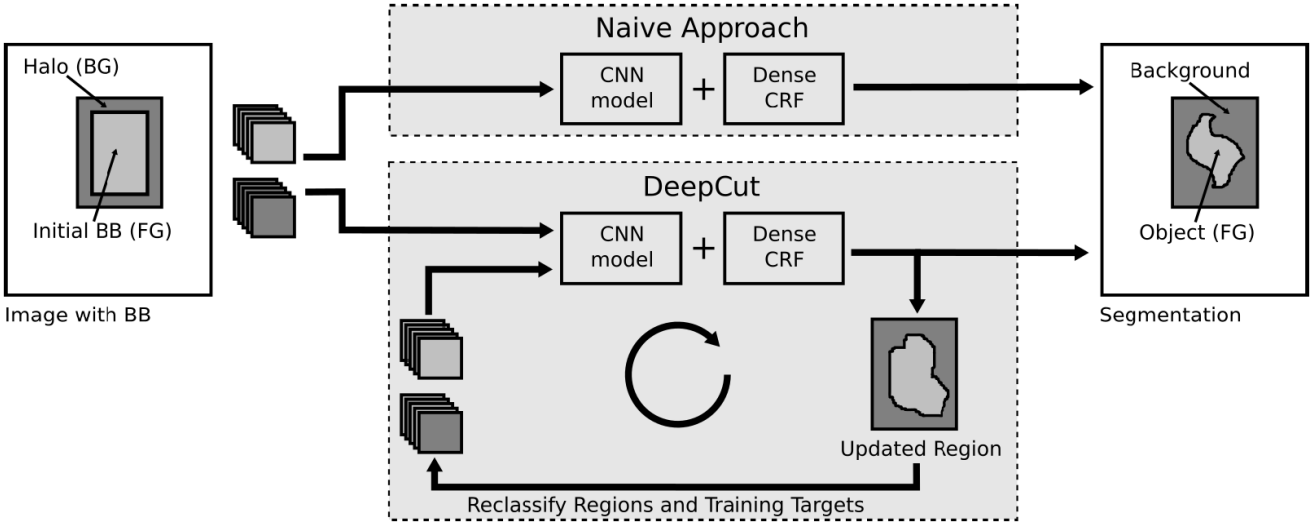


Fig. 2. Naïve CNN learning versus the proposed *DeepCut* approach, iteratively updating the learning target classes for input patches.

two sets of convolutional (*conv*) followed by rectifying linear units [35] and subsequent max-pooling (*pool*) layers. Since a convolution reduces the layer output by half the kernel size, we pad its output with zeros to preserve the size of the input tensor. The two serial *conv/pool* layers are attached to a layer with densely-connected (*dense*) neurons and an output layer with neurons associated with foreground and background. For regularisation purposes, we add dropout to both inputs of the *dense* and output layer, randomly sparsifying the signal and reducing the potential of over-fitting [33]. Figure 1 depicts the specific configuration, which has been fixed for all experiments in this study.

3) Training & Optimisation: To avoid a class imbalance, we sample an equal number of $K = 10^5$ patches X_k from the training database for each class Y_k (i.e. foreground and background). All CNN weights are initialised by sampling from a Gaussian distribution. We employ an adaptive gradient descent optimisation (ADAGRAD) [36] of mini-batches using a constant learning rate $\eta = 0.015$ for a fixed number of epochs. This has the advantage of adapting the learning rate locally for each feature, thus exhibiting more robust behaviour and faster convergence. The loss function is defined as the categorical cross-entropy between the true and coding distributions, respectively. For *DeepCut* training, during each iteration we update the training target regions within B (see Fig. 2) according to Sec. II-A after a fixed amount of epochs. Note, that this is only required during training. Inference during testing does not differ from standard feed forward networks.

4) Data Augmentation: The training set undergoes data augmentation for better generalisation of the learned features and to prevent over-fitting to the training data. For this purpose, we add a Gaussian-distributed intensity offset to each patch with the standard deviation σ and randomly flip the patch across the spatial dimensions to increase the variation in the training data.

C. Naïve Learning Approach

If we assume that the patches describing the object constrained to the bounding box B are unique, we can attempt to classify patches $X_i \in B$ into object and background, by using patches sampled from the bounding boxes as foreground targets. To obtain background targets, we can extend B to a halo region H, solely containing background voxels. A naïve approach to segmentation would be to assume that all $X_i \in B$ belong to the object and all $X_i \in H$ to the background. However, since the region B contains false positive locations (*i.e.*, the object does not fully extend to the entire bounding box region, but is merely a subset of it), we will introduce errors into the model, which will impact the accuracy of the final segmentation. To ensure a fair comparison of the segmentation results, post-processing includes regularisation with a densely-connected CRF [19]. This naïve approach ($\text{CNN}_{\text{naïve}}$) is depicted in Figure 2.

D. DeepCut

In order to develop a better approach compared to Section II-C, we start training the CNN model Θ with patches X sampled from B and H for foreground and background, respectively. In contrast to the naïve approach, we interrupt the training of Θ after a fixed number of epochs and update the classes Y for all voxels in $B = R_{FG} \cup R_{BG}$, via inference and subsequent CRF regularisation, according to Section II-A. We continue training with the updated targets and reinitialise the CNN with Θ.

E. Region Initialisation

While methods such as *DeepCut*, GrabCut [1] or others [2], [3] rely on (approximately) globally optimal solvers (*i.e.*, [11] and [19], respectively), the iterative nature of the algorithm will be limited to finding local optima. Thus the resulting segmentation is dependent on the initial regions R_{FG} and R_{BG}. Papandreou et al. [2] propose performing a pre-segmentation within B to initialise R_{FG} and R_{BG} closer to the object and observed large improvements in accuracy of the final segmentation. Similar initialisation can be done with the proposed *DeepCut* method and will be examined in Section III. In our experiments, we distinguish the variants of *DeepCut* initialised with bounding boxes and pre-segmentations with DC_{BB} and DC_{PS}, respectively.

III. Experiments

A. Image Data

For all experiments, we used the database in [37], consisting of MR images of 55 fetal subjects. The images were acquired on 1.5T MR scanner using a T2-weighted ssFSE sequence (scanning parameters: TR 1000 ms, TE 98ms, 4 mm slice thickness, 0.4 mm slice gap). Most of the images contain motion artefacts that are typical for such acquisitions. The imaged population consists of 30 healthy subjects and 25 subjects with intrauterine growth restriction (IUGR) and their gestational age ranged from 20 to 30 weeks. For all images, the brain and the lung regions have been manually segmented by an expert rater. We want to emphasise that the brain segmentations are in fact not tissue segmentations, but whole brains similar to the data used in [38], [39].

B. Preprocessing & Generation of Bounding Boxes

All images underwent bias field correction [40] and normalisation of the image intensities to zero mean and unit standard deviation computed from the bounding box. Bounding boxes B were generated from manual segmentations by computing the maximum extent of the segmentation and enlarging it by 5 voxels for each slice. The background halo regions H were created by extending B by 20 voxels.

C. Comparative Methods

To compare the performance of the proposed *DeepCut* approach, we fix the CNN architecture (see Section II-B), preprocessing and CRF parameters for all learning-based methods. We can consider the $\text{CNN}_{\text{naïve}}$ a lower bound in terms of accuracy performance. Alternatively, we train the CNN directly under full supervision (*i.e.* from pixel-wise segmentations, CNN_{FS}), which can be considered an upper accuracy bound given model complexity and data. We assess both the performance of the proposed *DeepCut* initialised by bounding boxes (DC_{BB}) or via a GrabCut pre-segmentation (DC_{PS}) as suggested in Sec. II-E. Lastly, we state results from the GrabCut (GC) method [1] for a performance comparison external to the proposed framework. During testing of all compared methods, we infer a segmentation into B and evaluate the result against manual segmentations.

D. Experimental Setup, Evaluation & Parameter Selection

We performed 5-fold cross-validation of randomly selected healthy and IGUR subjects and fixed the training and testing databases for all compared methods. The resulting segmentations are evaluated in their overlap with expert manual segmentations using the Dice Similarity Coefficient, measuring the mean overlap between two regions A and B:

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \tag{6}$$

Without any pre-selection, three randomly selected subjects were left out of the evaluation experiments and used to tune the GrabCut MRF regularisation weight and the CRF regularisation parameters in (4b) and (4a) via random permutations of the parameter combinations. The total amount of epochs used in both *DC* variants was set as maximum epochs for the comparative CNNs. The average training loss was computed to confirm convergence of all methods. An overview of all employed settings can be found in Tab. I.

TABLE I
DeepCut Parameters

Parameter	Value
Convolutional Neural Network	
Patch size (p_x x p_y x p_z)	33 x 33 x 3
Learning rate η	0.015
N_{Epochs} (Brain)	500
N_{Epochs} (Lungs)	250
N_{Epochs} per DeepCut iteration	50
N_{Batch}	10^5 patches per Epoch
$N_{Mini-batch}$	5000 patches
σ	0.1
Densely-connected CRF	
ω_1, ω_2	5.0
θ_α	10.0
θ_β (Brain)	20.0
θ_β (Lungs)	0.1
θ_γ (Brain)	1.0
θ_γ (Lungs)	0.1
$N_{Iterations}$	5
GrabCut (see [1])	
γ (Brain)	2.5
γ (Lungs)	1.0

TABLE II
Numerical Accuracy Results for Fetal Brain Segmentation. All Measurements Are Reported as Mean DSC [%] Across All Test Subjects

	BB	GC [1]	$\text{CNN}_{\text{naïve}}$	DC_{BB}	DC_{PS}	CNN_{FS}
mean	63.0	80.7	74.0	86.6	90.3	94.1
std.	4.5	4.9	4.5	4.7	5.4	4.1

E. Implementation Details & Hardware

We implemented the CNN as shown in Fig. 1 with *Lasagne*[1] and *Theano*[2] [41]. All experiments were run on Ubuntu 14.04 machines with 256 GB memory and a single Tesla K80 (NVIDIA Corp., Santa Clara, CA) with 12GB of memory. We used the CRF implementation in [42] to solve Eq. (1), according to [19].

IV. Results

Example segmentation results of all compared methods and initialisations can be found in Figures 3 and 4. We observe comparable agreement of the proposed DC_{PS} and a fully supervised CNN_{FS} for the brain region. Generally, an increase of segmentation accuracy can be seen in brain and lungs with increasing sophistication of the learning-based methods.

[1] http://lasagne.readthedocs.org/
[2] http://deeplearning.net/software/theano/

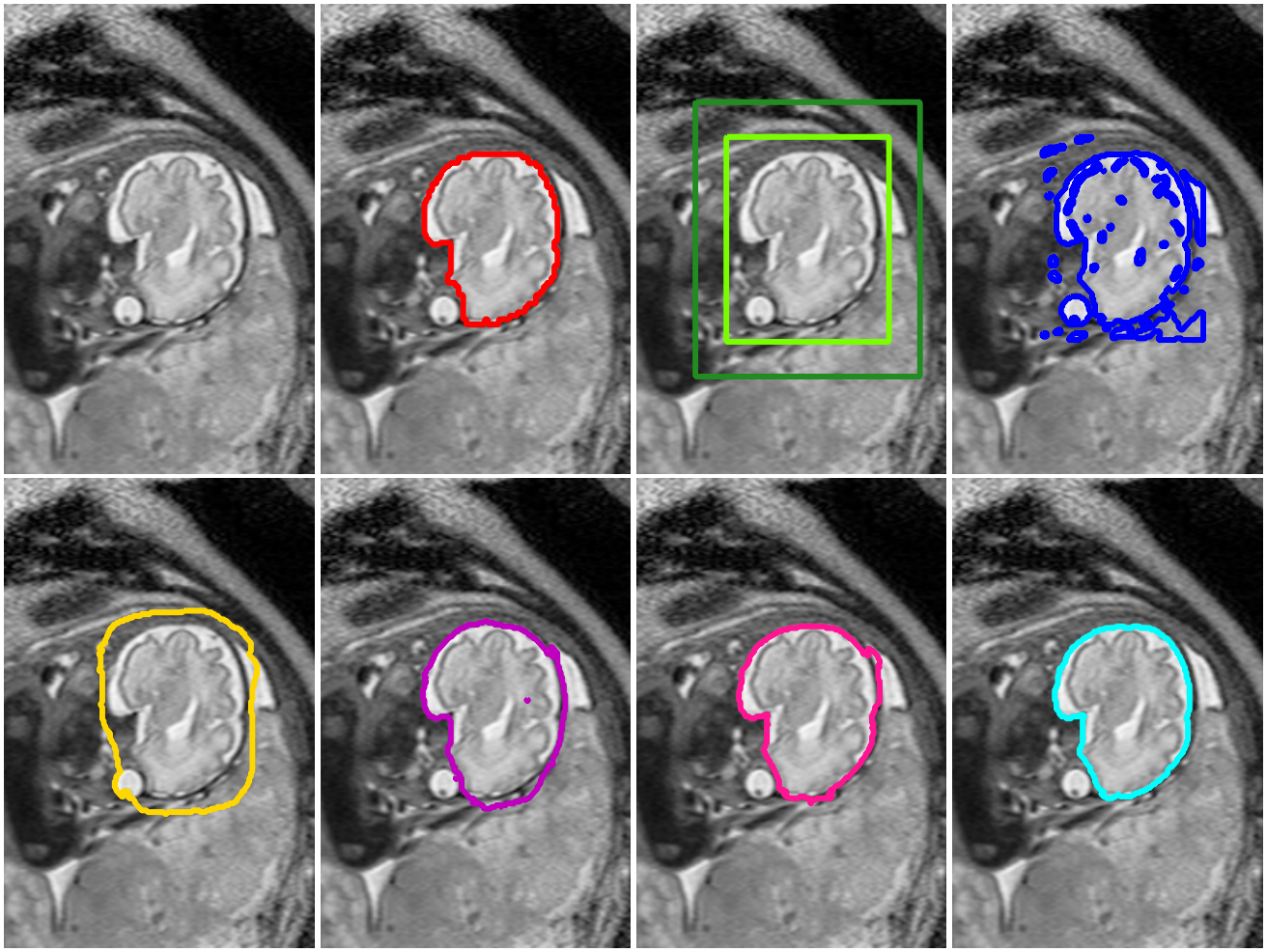


Fig. 3. Example brain segmentation results for all compared methods: Top row (from left to right): (1) original image (2) manual segmentation (red), (3) initial bounding box B with halo H, (4) GrabCut [1] (GC, blue). Bottom Row: (5) naïve learning approach ($CNN_{naïve}$, yellow), (6) DeepCut from bounding boxes (DC_{BB}, purple), (7) DeepCut from pre-segmentation (DC_{PS}, pink) and (8) fully supervised CNN segmentation (CNN_{FS}, cyan).

TABLE III
NUMERICAL ACCURACY RESULTS FOR FETAL LUNGS SEGMENTATION. ALL MEASUREMENTS ARE REPORTED AS MEAN DSC [%] ACROSS ALL TEST SUBJECTS

	BB	GC [1]	$CNN_{naïve}$	DC_{BB}	DC_{PS}	CNN_{FS}
mean	47.0	58.6	61.1	70.0	74.9	82.9
std.	4.1	19.0	6.4	8.1	6.7	10.0

A. Naïve Learning Approach versus DeepCut

Comparison of methods directly learning from bounding boxes (i.e. $CNN_{naïve}$ and DC_{BB}), demonstrate that the iterative target update of the proposed *DeepCut* method results in large improvements in accuracy for both the brain and the lungs (see Fig. 3 and 4, Tab. II and III). Numerically, we obtain an increase of 12.6% and 8.9% in terms of average DSC for the brain and lungs, respectively.

B. Initialisation with Pre-segmentations

Further, when a pre-segmentation instead of bounding boxes is used to initialise the *DeepCut* method, the mean DSC is improved by another 3.7% for the brain and 4.9% for the lungs. This can be seen in the example segmentation in Fig. 3 and 4, where the DC_{PS} segmentation for both organs is visually closer to those of the fully supervised CNN_{FS}.

C. Comparison with GrabCut

While the comparative GrabCut method performs well for the brain (DSC 80.7±4.9%), we observe less robust behaviour in the lungs (DSC $58.6 \pm 19.0\%$). GrabCut outperforms the $CNN_{naïve}$ for brain segmentation, however the presence of large outliers results in a lower mean accuracy in the lung regions. Several segmentations present with DSC $< 20\%$, indicating that the GrabCut was not able to detect an object in some cases.

D. DeepCut versus Fully Supervised Learning

We halted training and evaluated intermediate accuracy results of the proposed *DeepCut* variants over iterations. For both DC_{BB} and DC_{PS}, accuracy increases after each iteration (see Fig. 6) and approaches the upper bound of fully supervised training (CNN_{FS}). Most importantly, both proposed *DeepCut* methods present with higher average accuracy than the naïve approach, however the accuracy remains

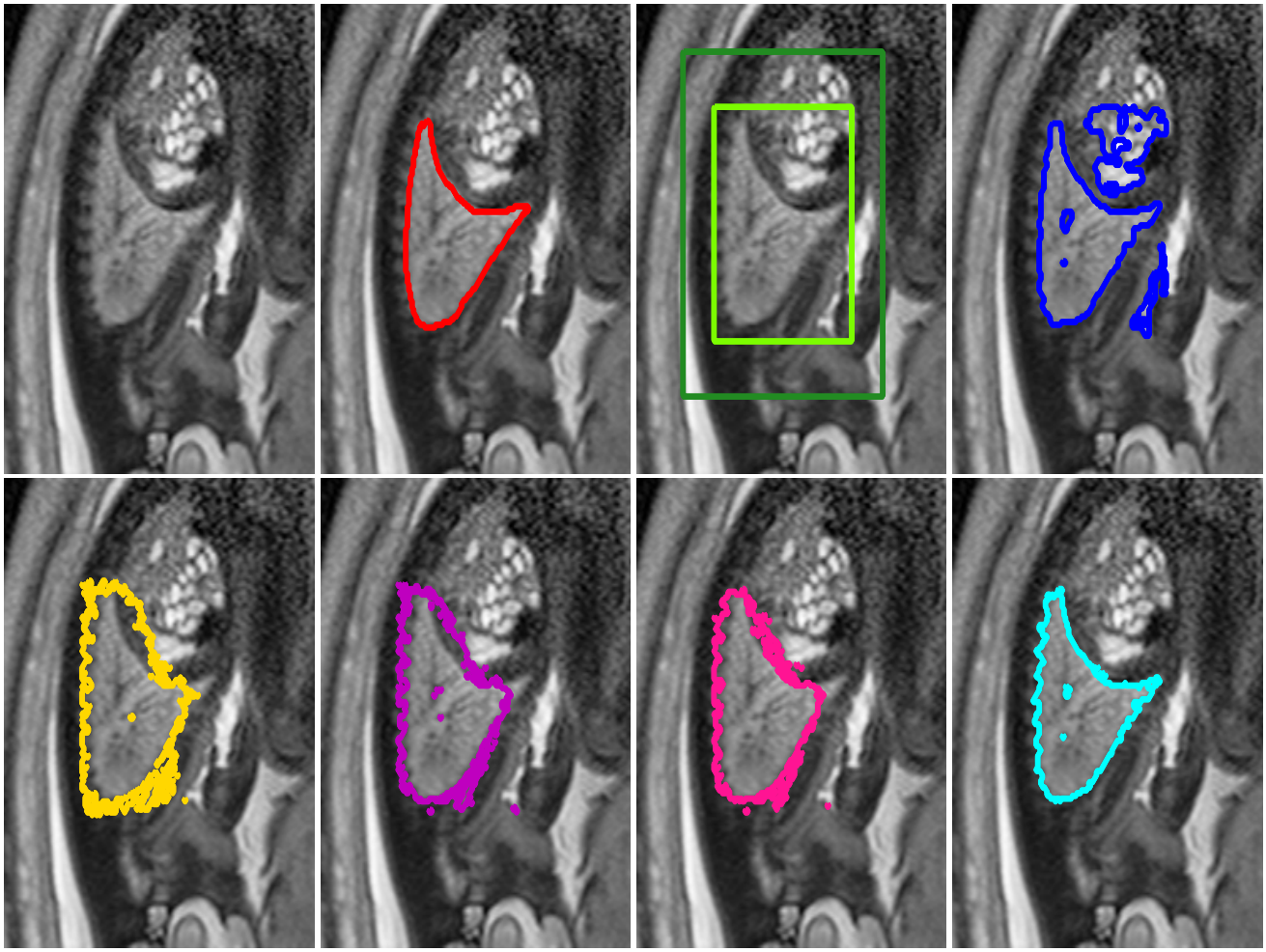
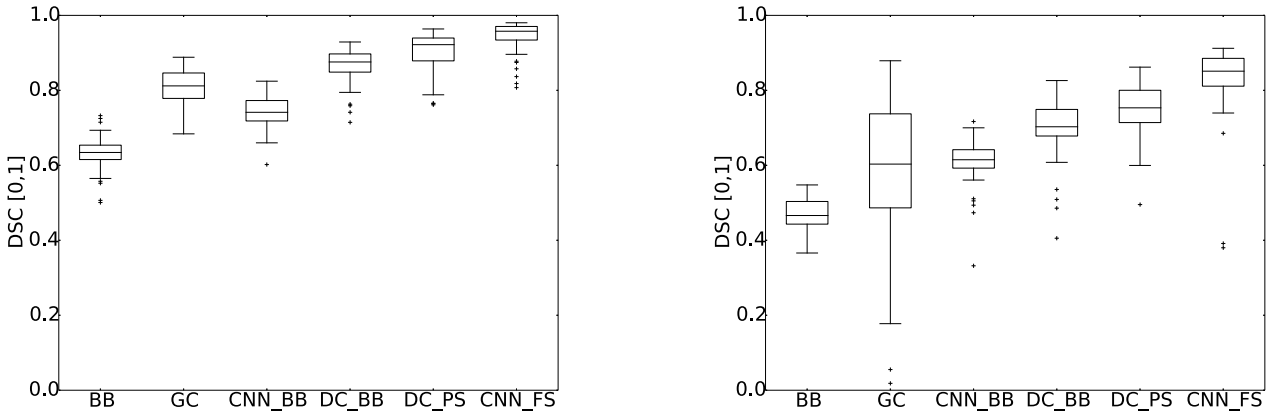


Fig. 4. Example lung segmentation results for all compared methods: Top row (from left to right): (1) original image (2) manual segmentation (red), (3) initial bounding box B with halo H, (4) GrabCut [1] (GC, blue). Bottom Row: (5) naïve learning approach ($CNN_{naïve}$, yellow), (6) DeepCut from bounding boxes (DC_{BB}, purple), (7) DeepCut from pre-segmentation (DC_{PS}, pink) and (8) fully supervised CNN segmentation (CNN_{FS}, cyan).



(a) Brain (b) Lungs

Fig. 5. Comparative accuracy results for the segmentation of the fetal brain (a) and lungs (b) for all methods: Initial bounding boxes (BB), GrabCut [1] (GC), naïve CNN $CNN_{naïve}$ learning approach from bounding boxes (CNN_{BB}), DeepCut initialised from bounding boxes (DC_{BB}), DeepCut initialised via pre-segmentation (DC_{PS}) and a fully supervised learning approach from manual segmentations (CNN_{FS}) as upper bound for this network architecture.

lower than CNN_{FS}. For reference, in Fig. 6 the mean (black) and standard deviation (gray) of $CNN_{naïve}$ and CNN_{FS} are also shown.

V. Discussion

The proposed *DeepCut* allows for obtaining pixelwise segmentations from an image database with bounding box

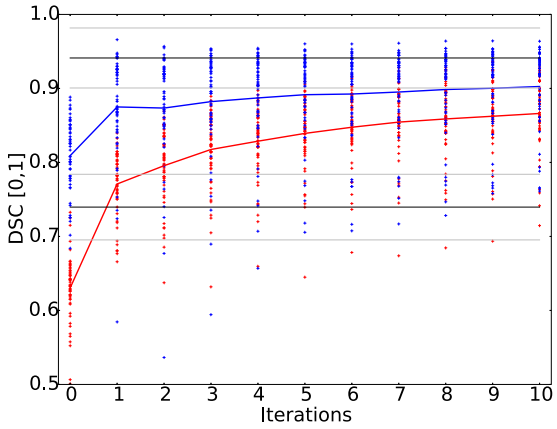


Fig. 6. Accuracy improvement in terms of DSC over *DeepCut* iterations in case of fetal brain segmentations. DeepCut initialisation with bounding boxes (DC_{BB}) (red) versus initialisation with pre-segmentation (DC_{PS}) (blue) in context with lower ($CNN_{naïve}$) and upper (CNN_{FS}) accuracy bound, depicted with mean (black) and standard deviations (grey).

annotations. Its general formulation allows for readily porting it to other segmentation problems and its use of CNN models avoids feature engineering as required by other learning-based methods.

A. Image Data

We deliberately chose a database exhibiting large variation in the imaged anatomy (*e.g.* the arbitrary position of the fetal body in the uterus, the extended gestational age range of 20-38 weeks or the presence of growth restriction (IGUR)) to test if a simple network configuration suffices for object segmentation problems constrained to bounding boxes. By restricting learning background patches from the halo H, we avoid learning features for the entire image domain, allowing for faster training.

B. Comparison with Related Studies

Comparing fully supervised learning (CNN_{FS}) qualitatively, we obtain an increased mean accuracy (94.1% DSC) over Keraudren et al. [38] (93.0% DSC) and large improvements over Taleb et al. [39] (84.2% DSC) for fetal brain segmentations. However, these methods [38], [39] are highly problem-specific solutions, which are applied to the *entire* image domain, making direct comparisons difficult. The only conclusion drawn from this comparison is roughly what accuracy range can be expected for automated fetal brain segmentation methods. In this sense, the generally applicable DC_{PS} method yields similar accuracy (90.3% DSC) by employing bounding box annotations, potentially placed 15x faster than pixel-wise annotations [2], [9]. This is qualitatively comparable to a recent study employing state-of-the-art CNNs under full (92.7%) and weak (90.3%) supervision to perform segmentations on full MR volumes [43]. *DeepCut* is closely related to the optimisation principle of GrabCut [1] and its use of CNNs allow the data-driven extraction of features, resulting in improved robustness. Similar iterative CNN training approaches have recently been proposed, employing region proposals [3] and foreground boosting [2] of intermediate scores to update target regions. However, these do not include regularisation when updating targets. This might be less problematic for the segmentation on natural images, but might be required to obtain smooth regions from noisy medical images.

C. Differences in Brain and Lung Segmentation Performance

For all internally compared methods, we observe a higher accuracy for brain segmentation results than for those of the lungs. There are several contributing factors to these differences, affecting all compared methods similarly. The regular shape of the brain can be better approximated with a bounding box than the lungs, which is underlined by the higher mean overlap of BB (refer to Tab. II and III, respectively). This introduces a lower amount of false positive initial targets, facilitating training the CNN. Secondly, the contrast of the background is higher in the brain, as it is often surrounded by hyper-intense amniotic fluid or hypo-intense muscular tissue. We can observe this in the tuned CRF parameters θ_β, penalising intensity differences (see Tab. I), to automatically tune to a lower value than the brain. Additionally, this can experimentally be observed with the GrabCut (GC) method, which heavily relies on intensity differences between the object and the adjacent background, performing worse than $CNN_{naïve}$ (*c.f.*, Fig. 5 (a) and (b), and Tab. II and III).

D. Effect of Initialisation on DeepCut performance

As shown in Fig. 6, we observe an increase in segmentation accuracy, when initialising the *DeepCut* with a pre-segmentation, rather than a bounding box. Papandreou et al. [2] reported a similar increase in accuracy when initialising their EM-based algorithm. Although in both approaches, the accuracy steadily increases with the number of epochs, the methods converge to different optima. This is due to the locally optimal nature of this iterative method and other iterative optimisation schemes, such as levelsets [21], [44] or iterated graphical methods [1], [22], [45] even when employing an (approximately) globally optimal solver such as [11], [19], [21]. Potential improvements might include to update the targets with a higher frequency than in this study (*c.f.*, Tab. I, N_{Epochs} per *DeepCut* iteration) or entertaining the notion, that an optimal set of CRF regularisation parameters exists at each iteration. However, tuning for the latter might be computationally expensive and thus of little practical value. Recent advances of expressing the employed CRF [19] as a recurrent neural network [46], might be a solution for back-to-back training of the θ parameters involved in (4a) and (4b) at each *DeepCut* iteration.

E. Internal Comparative Experiments

For both lungs and brain segmentation, we report a large improvement in accuracy with *DeepCut* variants over a naïve learning approach (*c.f.*, Fig. 5 and Tab. II and III). As in Section II-E, we suggest to initialise *DeepCut* with a pre-segmentation, reducing the amount of false positive targets for the initial training. A closer initialisation leads to a performance improvement, even if the pre-segmentation is not accurate (*c.f.*, Fig. 5 (b), where there is a remarkable

improvement from DC_{BB} to DC_{PS}). Methodological improvements from DC do not necessarily effect all test subjects equally as we observe a lower precision for both DC methods compared to the naïve CNN approach. However, an increased minimum accuracy for *DeepCut* on both lungs and brain segmentations can be observed with increasing sophistication of the method. Further, the higher precision of all learning-based methods underline the robust performance compared to image segmentation methods, such as GrabCut. One explanation is that learning a model of the object from a collection of images is favourable to fitting a model (*e.g.* a GMM) to a single image, as done in many object segmentation methods. If desired, the model can be adjusted in depth to cover a wider variation of appearance and scales, as those employed in [2], [3]. However, when the objects exhibit a large class similarity in terms of shape and appearance, simple CNN architectures could suffice for most *medical* image segmentation problems.

F. Limitations

The default CRF parameters θ and ω in [19] (see Tab. I) were not appropriate for medical images and required tuning (*c.f.* Sec. III-D). This required manual segmentations to find approximate and working parameter sets. However, these might be modality and problem dependent, requiring adjustment when translating *DeepCut* to new segmentation problems. We also applied this tuning procedure to adapt the parameters of the GrabCut method to ensure a fair comparison of the results. Further, the time required for training of one epoch was approximately 9 minutes and inference during testing (including CRF) was less than 8 minutes for the largest MR volume. This is relatively high for a simple network architecture and can be largely reduced by employing a fully-convolutional network and smaller mini-batch sizes as shown in [43].

G. Conclusions

We proposed *DeepCut*, a new method to obtain pixelwise segmentations, given a database of bounding box annotations and studied variants employing an iterative dense CRF formulation and convolutional neural network models. *DeepCut* is able to segment both the fetal brain and lungs from an image database of large variation in the anatomy and is readily applicable to similar problems on medical images. The proposed method performs well in terms of accuracy compared to a model trained under full supervision and simultaneously greatly reduces the annotation effort required for analysis.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Tesla K40 GPU used for this research.

References

[1] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[2] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. (2015). "Weakly- and semi-supervised learning of a DCNN for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1502.02734

[3] J. Dai, K. He, and J. Sun. (2015). "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation." [Online]. Available: https://arxiv.org/abs/1503.01640

[4] T. Schlegl, S. M. Waldstein, W.-D. Vogl, U. Schmidt-Erfurth, and G. Langs, "Predicting semantic descriptions from medical images with convolutional neural networks," in *Information Processing in Medical Imaging*. New York, NY, USA: Springer, 2015, pp. 437–448.

[5] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 277–284.

[6] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1. Jul. 2001, pp. 105–112.

[7] M. Rajchl *et al.*, "Interactive hierarchical-flow segmentation of scar tissue from late-enhancement cardiac MR images," *IEEE Trans. Med. Imag.*, vol. 33, no. 1, pp. 159–172, Jan. 2014.

[8] J. S. Baxter, M. Rajchl, T. M. Peters, and E. C. Chen, "Optimization-based interactive segmentation interface for multi-region problems," *Proc. SPIE*, vol. 9413, p. 94133T, Apr. 2015.

[9] T.-Y. Lin *et al.*, "Microsoft Coco: Common objects in context," in *Computer Vision—ECCV*. New York, NY, USA: Springer, 2014, pp. 740–755.

[10] M.-M. Cheng, V. A. Prisacariu, S. Zheng, P. H. Torr, and C. Rother, "Densecut: Densely connected CRFs for realtime grabcut," *Comput. Graph. Forum*, vol. 34, no. 7, pp. 193–201, 2015.

[11] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[12] N. Komodakis, G. Tziritas, and N. Paragios, "Fast, approximately optimal solutions for single and dynamic MRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[13] D. Freedman and T. Zhang, "Interactive graph cut based segmentation with shape priors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2005, pp. 755–762.

[14] B. L. Price, B. Morse, and S. Cohen, "Geodesic graph cut for interactive image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3161–3168.

[15] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, and S. Yan, "Semantic segmentation without annotating segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2176–2183.

[16] R. Wolz *et al.*, "Measurement of hippocampal atrophy using 4D graph-cut segmentation: Application to ADNI," *NeuroImage*, vol. 52, no. 1, pp. 109–118, 2010.

[17] L. M. Koch, M. Rajchl, T. Tong, J. Passerat-Palmbach, P. Aljabar, and D. Rueckert, "Multi-atlas segmentation as a graph labelling problem: Application to partially annotated atlas data," in *Information Processing in Medical Imaging*. New York, NY, USA: Springer, 2015, pp. 221–232.

[18] M. Rajchl *et al.*, "Hierarchical max-flow segmentation framework for multi-atlas segmentation with Kohonen self-organizing map based Gaussian mixture modeling," *Med. Image Anal.*, vol. 27, pp. 45–56, Jan. 2016.

[19] P. Krähenbühl and V. Koltun. (2012). "Efficient inference in fully connected CRFs with Gaussian edge potentials." [Online]. Available: https://arxiv.org/abs/1210

[20] J. Yuan *et al.*, "Jointly segmenting prostate zones in 3D MRIs by globally optimized coupled level-sets," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Berlin, Germany: Springer, 2013, pp. 12–25.

[21] M. Rajchl *et al.*, "Variational time-implicit multiphase level-sets," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. New York, NY, USA: Springer, 2015, pp. 278–291.

[22] C. Nambakhsh *et al.*, "Left ventricle segmentation in MRI via convex relaxed distribution matching," *Med. Image Anal.*, vol. 17, no. 8, pp. 1010–1024, 2013.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, 1977, pp. 1–38.

[24] T. Kapur, W. E. L. Grimson, W. M. Wells, and R. Kikinis, "Segmentation of brain tissue from magnetic resonance images," *Med. Image Anal.*, vol. 1, no. 2, pp. 109–127, 1996.

[25] R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold MIL training for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2409–2416.

[26] A. Vezhnevets and J. M. Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3249–3256.

[27] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Jun. 2008, pp. 1–8.
[28] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3642–3649.
[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[30] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. (2014). "Fully convolutional multi-class multiple instance learning." [Online]. Available: https://arxiv.org/abs/1412.7144
[31] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1713–1721.
[32] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.
[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[34] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.
[35] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Aistats*, vol. 15, no. 106, p. 275, 2011.
[36] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.
[37] M. S. Damodaram *et al.*, "Foetal volumetry using magnetic resonance imaging in intrauterine growth restriction," *Early Human Develop.*, vol. 88, pp. S35–S40, Mar. 2012.
[38] K. Keraudren *et al.*, "Automated fetal brain segmentation from 2-D MRI slices for motion correction," *NeuroImage*, vol. 101, pp. 633–643, Nov. 2014.
[39] Y. Taleb, M. Schweitzer, C. Studholme, M. Koob, J.-L. Dietemann, and F. Rousseau, "Automatic template-based brain extraction in fetal MR images," Tech. Rep., 2013.
[40] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, Feb. 1998.
[41] F. Bastien *et al.* (2012). "Theano: New features and speed improvements." [Online]. Available: https://arxiv.org/abs/1211.5590
[42] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert, and B. Glocker, "Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI," in *Proc. Ischemic Stroke Lesion Segmentation*, 2015, p. 13.
[43] M. Rajchl *et al.* (2016). "Learning under distributed weak supervision." [Online]. Available: https://arxiv.org/abs/1606.01100
[44] E. Ukwatta, J. Yuan, M. Rajchl, W. Qiu, D. Tessier, and A. Fenster, "3-D carotid multi-region MRI segmentation by globally optimal evolution of coupled surfaces," *IEEE Trans. Med. Imag.*, vol. 32, no. 4, pp. 770–785, Apr. 2013.
[45] I. Ben Ayed, H.-M. Chen, K. Punithakumar, I. Ross, and S. Li, "Graph cut segmentation with a global constraint: Recovering region distribution via a bound of the Bhattacharyya measure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3288–3295.
[46] S. Zheng *et al.* (2015). "Conditional random fields as recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1502.03240