RACE-net: A Recurrent Neural Network for Biomedical Image Segmentation

Arunava Chakravarty, *Student Member, IEEE*, and Jayanthi Sivaswamy,*Member, IEEE*

***Abstract*— The level set based deformable models (LDM) are commonly used for medical image segmentation. However, they rely on a handcrafted curve evolution velocity that needs to be adapted for each segmentation task. The Convolutional Neural Networks (CNN) address this issue by learning robust features in a supervised end-to-end manner. However, CNNs employ millions of network parameters which require a large amount of data during training to prevent over-fitting and increases the memory requirement and computation time during testing. Moreover, since CNNs pose segmentation as a region-based pixel labeling, they cannot explicitly model the high-level dependencies between the points on the object boundary to preserve its overall shape, smoothness or the regional homogeneity within and outside the boundary.**

We present a Recurrent Neural Network (RNN) based solution called the RACE-net to address the above issues. RACE-net models a generalized LDM evolving under a constant and mean curvature velocity. At each time-step, the curve evolution velocities are approximated using a feed-forward architecture inspired by the multi-scale image pyramid. RACE-net allows the curve evolution velocities to be learned in an end-to-end manner while minimizing the number of network parameters, computation time and memory requirements. The RACE-net was validated on three different segmentation tasks: Optic disc and cup in color fundus images, cell nuclei in histopathological images and the left atrium in cardiac MRI volumes. Assessment on public datasets was seen to yield high Dice values between 0.87 and 0.97 which illustrates its utility as a generic, off-the-shelf architecture for biomedical segmentation.

***Index Terms*— biomedical segmentation, deep learning, deformable model, RNN.**

I. Introduction

The segmentation of anatomical structures in medical images plays an important role in the diagnosis and treatment of diseases. It is frequently used to visualize organs, extract quantitative clinical measurements, define the region of interest to localize pathologies and aid in the surgical or radiation treatment planning. Accurate segmentation algorithms can save the time and effort of medical experts and minimize the intra and inter-subject variability involved in manual segmentation. However, this task is often challenging due to the lack of sufficient contrast between the anatomy of interest and its background, large variability in its shape, variations in image quality across subjects or scanners and the presence of noise and non-uniform illumination.

Motivated by the recent success of deep learning in biomedical image segmentation, *we propose a deep Recurrent Neural Network (RNN) architecture, called the RACE-net (Recurrent Active Contour Evolution Network) which is inspired from the level set based deformable models (LDM)*. Given a rough localization of the anatomical structure to be segmented, RACE-net iteratively evolves it using a combination of a constant and a mean curvature velocity which are learned from a set of training images in an end-to-end manner. A Python implementation of the proposed method is available for public research use at `https://www.dropbox.com/s/71b4wybceapeamb/RACE_net_codes.zip?dl=0`. The key contributions of this paper are: a) LDM is formulated as a novel RNN architecture; b) An appropriate loss function is introduced to overcome the problems related to the re-initialization of the level set function in LDMs and the vanishing gradients during the training of the RNN; c) The constant and mean curvature velocities are modeled by a novel feed-forward architecture inspired from the multi-scale image pyramid; d) The effectiveness of the method is assessed on a variety of structural segmentation tasks, namely, the optic disc and cup in color fundus images, cell nuclei in histopathological images and the left atrium in 3D cardiac MRI Volumes.

II. Related Work

The use of a fixed protocol and view during image acquisition and the overall similarity in the anatomical structures across subjects often allow for a rough localization of the structure of interest using simple image processing techniques and/or spatial priors. However, an accurate segmentation at a sub-pixel level is often essential for proper diagnosis. Traditionally, the LDMs (also known as the geometric or implicit active contour models) have been widely used for this task [1] in which an initial curve (or a surface in 3D) is iteratively evolved by a curve evolution velocity until it converges onto the desired boundary. Typically, the curve evolution velocity comprises image dependent and regularization terms. The image dependent terms drive the curve towards the desired boundary. On the other hand, the regularization terms such as the boundary length minimization [2] preserves its

This research was supported in part by the doctoral fellowship provided by Tata Consultancy Services (TCS) under their Research Scholarshp Program.

Arunava Chakravarty and Jayanthi Sivaswamy are with the Center for Visual Information Technology, International Institute of Information Technology, Hyderabad, India (e-mail: arunava.chakravarty@research.iiit.ac.in; jsivaswamy@iiit.ac.in).

smoothness, while the balloon force [3] is used to drive the evolving curve into the concave regions in the object boundary.

The LDMs provide an Energy Minimization framework (see section III-A) which allows an explicit modelling of the high level constraints on the boundary such as the trade-off between the intensity discontinuity [2], [4] at the boundary and its smoothness, or the regional homogeneity within and outside it [5]. However, LDMs cannot be used "off-the-shelf" as the curve evolution velocity has to be specifically handcrafted for each segmentation task, failing which, the curve can take a large number of iterations to converge, get entrapped near noise or spurious edges, or smooth out the sharp corners of the object of interest.

Recently, the Convolutional Neural Networks (CNNs) have been widely applied to various medical image segmentation tasks such as the extraction of neuronal structures in 2D electron microscopy [6], [7], prostate segmentation in MRI volumes [8] and, the optic disc and vessel segmentation in fundus images [9]. CNNs learn a hierarchical feature representation of the images from large datasets in a supervised end-to-end manner, eliminating the need for handcrafted features. In the sliding window approach [6], [10], [11] a CNN is employed to classify the pixels into foreground or background at a patch level followed by a Conditional Random Field (CRF) [10] or LDM [11] based post-processing to incorporate the global context. Alternatively, the pixel labeling for the entire image is obtained in a single step by employing a deconvolutional architecture which consists of a "contraction" followed by an "expansion" stage [7], [8], [12]. The contracting stage has a series of convolutional layers with pooling layers in between which successively reduce the resolution of the feature channels. The expansion stage has one convolution layer corresponding to each contraction layer with upsampling operations in between to recover the original image resolution. Additional skip connections are also employed to directly connect the corresponding contraction and expansion layers.

CNNs lack an explicit way to capture high level, long range label dependencies between similar pixels and the spatial and appearance consistency of the segmentation labels [13]. This can result in poor object dilineation, non-sharp boundaries and small spurious regions in the segmentation output. Few RNN architectures have been explored to address these issues. RNNs generally employ a simple recurrent unit (RU) which comprises a single neuron with a feedback connection. Gated variants of the RUs such as LSTMs [14], [15] are commonly used to overcome the problem of vanishing gradients encountered during training over large time-steps.

The multi-dimensional RNN (MDRNN) was explored in [16] to segment perimysium in skeletal muscle microscopy images. In MDRNN, the RUs are connected in a grid-like fashion with as many recurrent connections as there are spatial dimensions in the image. At each step, the RU receives the current pixel as input and the hidden states of the pixels explored in each direction in the previous step through feedback connections, thus recursively gathering information about all other pixels in the image. PyraMiD-LSTM [17] modified the toplology of the feedback connections in MDRNNs to improve its computational efficiency and applied it to segment neuronal electron microscopy images and MRI brain Volumes.

Some Deep learning architectures have also attempted to combine RNN and CNNs in a *time-distributed* manner where the external input to the RNN at each time-step was provided by the output of the CNN. Such architectures have been employed in [14], [15] to segment the intra-retinal tissue layers in OCT and 3D neuronal structures in electron microscopy images respectively. Both [14], [15] employed bi-directional RNNs to capture the context from both the past as well as the future time-steps. Moreover, multiple layers of RNNs were stacked together to obtain a deep RNN architecture. Alternatively, CNN and RNNs can also be integrated by replacing the RU by a CNN within the RNN architecture. Such an architecture was employed in [13] to formulate the mean-field based inference of a CRF as a RNN. The individual iterations of the mean-field algorithm was modeled as a CNN and the multiple mean-field iterations were implemented by using a simple feedback connection from the output of the entire CNN to its input, resulting in a RNN.

In comparison to CNNs, RNNs have received relatively less attention in the field of biomedical image segmentation. In this work, we explore a RNN architecture similar to [13] which combines the advantages of both LDM and CNN. In comparison to the LDMs, RACE-net learns the level set curve evolution velocity in an end-to-end manner. An attractive feature of the RACE-net is that the evolving curve was empirically found to converge onto the desired boundary over large distances in very few (5-7) time steps on a diverse set of applications. In contrast, LDMs either fail to converge over large distances or require a large number of iterations.

In comparison to the CNNs, the RACE-net offers two main advantages. First, it provides a boundary-based alternative to the existing region-based pixel labeling approach of the CNNs. The RACE-net models a generalized level set curve evolution (see Section III-A). Consequently, it can explicitly learn the high level dependencies between the points on the object boundary to preserve its overall shape and smoothness. At the same time, it can also maintain an optimal trade-off between the boundary discontinuity and the regional homogeneity of the structure in a convolutional feature space (which is learned in an end-to-end manner). Secondly, RACE-net can have a very complex network structure while utilizing very few learnable network parameters in comparison to the CNNs. Fewer network parameters serve to a) reduce the risk of overfitting on small training datasets which is particularly useful in the medical domain where obtaining the ground truth markings is expensive; b) It reduces the computation time and memory requirements of the pre-trained architecture allowing it to be deployed on systems with limited resources [18].

III. Method

We begin by obtaining a generalized level set equation for a LDM in Section III-A to model it as a RNN. Next, we present a Feedforward neural network (FFNN) architecture which is designed to model each time-step of the curve evolution. The FFNN consists of a customized layer to compute the normal and curvature of the level set function (Section III-B) and a novel CNN architecture to model the evolution

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2852635, IEEE Journal of Biomedical and Health Informatics

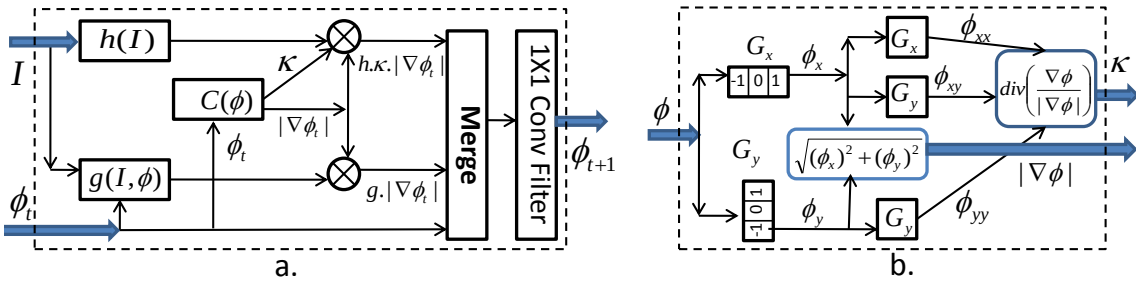

Fig. 1. **a.** The Feedforward neural network (FFNN) architecture that models the evolution of the level set function ϕ in a single time step t. **b.** The details of the customized $C(\phi)$ module used within the FFNN in Fig. a. to compute the normal and curvature of ϕ.

velocities (Section III-B.1). Finally, we show how the entire curve evolution is modeled as a RNN in Section III-C with an appropriate loss function (Section III-C.1) for training.

A. A generalized PDE for curve evolution

Let I represent the image to be segmented. Consider a family of 2D planar curves (the discussion can be easily generalized to 3D surfaces) represented using an arc length parameterization as $C(s) = \{(x(s), y(s))|0 \leq s \leq l\}$ where $(x(s), y(s))$ are points on C and l denotes the length of the curve. The evolution of C, can be modeled by the Partial Differential Equation (PDE) $\frac{\partial}{\partial t}C(s,t) = V.\vec{n}(s,t)$, where t is the temporal parameter, $\vec{n}(s,t)$ represents the normal vector to C at $(x(s), y(s))$ and V is a velocity function defined over the spatial support of I. In order to define V in a meaningful manner, an Energy functional $E(C(s))$ is defined such that its minimum corresponds to the object boundary. $E(C(s))$ is minimized using functional gradient descent by iteratively evolving $C(s)$ in the *negative direction of the Euler-Lagrange* of E, i.e., $V = -\frac{d}{dC}E(C(s))$. Typically, $E(C(s))$ is composed of a linear combination of boundary and region-based cost terms such that

$$\underset{C}{\operatorname{argmin}}\ E(C(s)) = \gamma_1 . \oint_C h(x,y)ds + \gamma_2 . \iint_{R_c} q(x,y)dA, \tag{1}$$

where $h(x,y)$ and $q(x,y)$ are functions defined over the spatial support of I; γ_1 and γ_2 are the relative weights. The boundary cost term minimizes the line integral of $h(x,y)$ along the curve C. Both $h(x,y)$ and $q(x,y)$ can be composed of a weighted sum of multiple functions. For instance, $h(x,y)$ is often composed of an edge indicator function which is inversely proportional to the image gradient magnitude [2], [3] to drive C towards the edges. Setting $h(x,y) = 1$ as a constant scalar function serves to minimize the curve length and hence commonly used as a regularization term to smooth the boundary [3], [5].

The second term helps minimize the regional cost as it is the area integral of $q(x,y)$ in the region enclosed by C. Assigning $q(x,y) = \lambda_{in}.|I - c_{in}|^2 - \lambda_{out}.|I - c_{out}|^2$ leads to the Chan-Vese model [5], where λ_{in}, λ_{out} are the scalar weights. The terms c_{in} and c_{out} represent the mean intensities of the object and the background regions respectively, at each time step of the curve evolution. Setting $q(x,y) = 1(or\ -1)$ serves to minimize (or maximize) the area enclosed by C and leads to the balloon force [3] which defines the default tendency of the curve to contract (or expand) in the absence of nearby edges in the image and speeds up the curve evolution.

The Euler-Lagrange for the boundary and the regional cost terms in eq. 1 are given by $\gamma_1.\{\langle\nabla h, \vec{n}\rangle + h.\kappa\}\vec{n}$ and $\gamma_2.q.\vec{n}$ respectively, where κ represents the curvature at $C(s)$ (See [19] and the Supplementary material [1] for details). Thus, the curve evolution which minimizes eq.1 is given by

$$\frac{\partial}{\partial t}C(s,t) = -\{\gamma_2.g.\vec{n} + \gamma_1.h.\kappa.\vec{n}\}, \tag{2}$$

where $g = \left(\frac{\gamma_1}{\gamma_2}.\langle\nabla h, \vec{n}\rangle + q\right)$. The curve evolution under the velocities $g.\vec{n}$ and $h.\kappa.\vec{n}$ are known as the constant flow and the mean curvature flow respectively.

During implementation, C is often represented using level sets due to its ability to handle topological changes such as the merging or splitting of the boundary. The evolving object boundary $C(t)$ is represented by the zero level set $C(t) = \{(x,y)|\phi(x,y;t) = 0\}$ of a higher dimensional scalar function $\phi(x,y;t)$ which is defined as the signed distance function. By convention, in this work we define $\phi(x,y;t)$ to be positive inside and negative outside $C(t)$. The curvature and the normal can be shown to be $\kappa = div\left(\frac{\nabla\phi}{|\nabla\phi|}\right)$ and $\vec{n} = -\frac{\nabla\phi}{|\nabla\phi|}$ in the level set representation [20]. Substituting these values in eq. 2 results in the equivalent level set equation

$$\frac{\partial\phi}{\partial t} = \alpha_1.g(I,\phi).|\nabla\phi| + \alpha_2.h(I).\kappa|\nabla\phi|, \tag{3}$$

where $\alpha_1 = -\gamma_2$ and $\alpha_2 = -\gamma_1$ are the scalar weights. The sign (positive or negative) of α_1 and α_2 determines the direction (inward or outward) of the curve evolution. We propose to model eq. 3 with a RNN architecture. Each time-step of the curve evolution is modeled as a FFNN by approximating the $g(I,\phi)$ and $h(I)$ functions by two separate CNNs. To simplify the network architecture, the $g(I,\phi)$ is learned directly and not restricted to the form $\left(\frac{\gamma_1}{\gamma_2}.\langle\nabla h, \vec{n}\rangle + q\right)$. This leads to a more generalized curve evolution model.

B. Single time-step of the Curve Evolution

Let ϕ_t denote the evolving level set function at a time step t. The evolution of ϕ_t to ϕ_{t+1} is modeled as a FFNN depicted in Fig. 1 a. Both $g(I,\phi)$ and $h(I)$ are modeled using CNNs

[1] available at `https://researchweb.iiit.ac.in/~arunava.chakravarty/supplementary_material_RACE_net.pdf`.

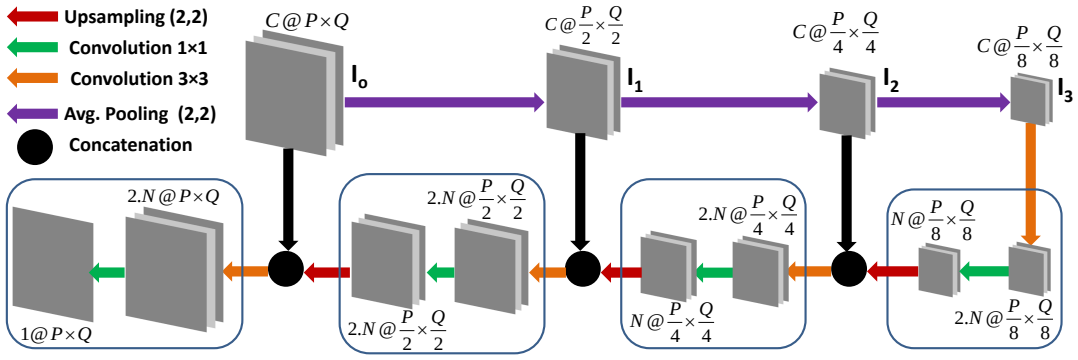

Fig. 2. The proposed CNN architecture to model $\boldsymbol{h(I)}$ and $\boldsymbol{g(I,\phi)}$. This figure is best viewed in color.

which output a single channel feature map. We propose a customized network layer architecture $C(\phi)$ to compute the $|\nabla\phi|$ and κ from a given ϕ. The first and second terms in eq.3, namely, $g(I,\phi).\mid\nabla\phi\mid$ and $h(I).\kappa.\mid\nabla\phi\mid$, are obtained using a pixel-wise multiplication of $g(I,\phi)$ and $h(I)$ with the appropriate outputs from $C(\phi)$. They are concatenated with the input ϕ_t to obtain a three channel feature map. A (1×1) convolution filter (across the three channels) with a linear activation function is applied to obtain the updated ϕ_{t+1}. A detailed depiction of $C(\phi)$ is shown in Fig. 1 b. The two convolutional filters G_x, G_y are used to obtain the first and second order derivatives of ϕ. We *freeze* the weights of G_x, G_y to the one-dimensional gradient filter kernels depicted in Fig. 1 b to ensure that they are not modified by the backpropagation algorithm during training. Thereafter, the curvature $\kappa = \frac{\phi_{xx}.\phi_y^2 - 2.\phi_x.\phi_y.\phi_{x,y} + \phi_{yy}.\phi_x^2}{(\phi_x^2+\phi_y^2)^{3/2}}$ and $\mid\nabla\phi\mid$ are computed by two custom layers defined using the Theano library [21].

1) Network Architecture of g(I, ϕ) *and* h(I)*:* $g(I,\phi)$ and $h(I)$ are modeled using a similar CNN architecture but with separate network weights and inputs. While I is used as an input for $h(I)$, both I and ϕ_t are concatenated to obtain a multi-channel input for $g(I,\phi)$. Minimizing the number of network parameters can reduce the memory requirements and help prevent over-fitting. This is achieved by designing a novel CNN architecture depicted in Fig. 2 which is inspired from the multi-scale image pyramid used in image processing. In Fig. 2, $\{I_l \mid 0 \le l \le 3\}$ represents a four level image pyramid. I_0 is the input to the CNN, while the successive scales I_l are obtained by applying *average pooling* in a 2×2 neighborhood on I_{l-1}. The segmentation proceeds from a coarse to fine scale from $l=3$ to 0. The output feature channels from the coarse scale I_l is upsampled and concatenated with the image in the successive finer scale I_{l-1} to obtain the input for the convolutional layers in the $l-1^{th}$ scale. At the coarsest scale (l=3), only I_3 is used for the input.

The processing at each scale is done in two convolutional layers. The first convolutional layer (depicted by orange arrows) has $2.N$ filters to capture the patterns in the local neighborhood. The filter's receptive field size S_l is decreased from the coarse (l=3) to the finest scale (l=0), ie., $S_l \ge S_{l-1}$. The second convolutional layer in each scale (depicted by green arrows) has N, 1×1 ($\times1$ in 3D) filters to reduce the dimensionality of its input feature channels from $2.N$ to N thereby *reducing the number of learnable filter weights in the next scale by half*. This is because the number of network parameters in a convolution layer is the product of filter kernel size and the number of input feature channels. Moreover, it adds an additional non-linearity to the network. An exception is the last layer which has only one (instead of N) filter to ensure a single feature channel as the final CNN output. Thus the overall network size is controlled by a single tunable hyper-parameter N. All convolutional filters use the Rectified Linear Unit as the activation function. An exception is the final 1×1 convolutional filter at $l=0$ which uses a linear activation to allow the RACE-net to learn arbitrary updates for ϕ in both positive (curve expands) or negative (curve contracts) direction. This was empirically found to be crucial in improving the segmentation performance.

The proposed architecture can be visualized as a modification to the U-net [7] where the encoder network is replaced by an image pyramid representation resulting in a drastic reduction of the network parameters.

C. Curve Evolution as a RNN

The FFNN in Fig. 1 a models one iteration of the curve evolution. At a time-step t, it takes the input image I and the current level set function ϕ_t as input to compute the evolved level set function ϕ_{t+1}. Let the function $f_\theta : \phi\times I \rightarrow \phi$ represent the complex, non-linear operation performed by the FFNN where θ denotes its learnable network parameters.

If ϕ_t evolves over a maximum of T time-steps during segmentation, then θ should be learned such that the final ϕ_T converges onto the desired boundary. We propose to do this by defining a RNN that can model the recursive equation: $\phi_{t+1} = f_\theta(\phi_t, I), 0 \le t \le T$. This is achieved by introducing a recurrent feedback connection which provides the output of f_θ as a feedback input to itself in the next iteration (see Fig. 3 a) for T time-steps. The feedback connection allows us to model all iterations of the curve-evolution within a RNN architecture which we refer to as the RACE-net. It is different from the most commonly used RNN architecture ([22] for example) where the output of an individual *neuron* is fed back onto itself to update its hidden state. In contrast, in the RACE-net architecture, the output of *an entire feed-forward*

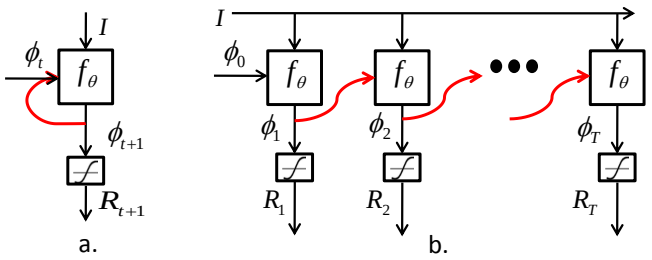

Fig. 3. **a.** Curve Evolution as a RNN. **b.** The unrolling of the RNN in a. over time. The recurrent feedback connections are depicted in red.

network is fed back to its input in the next time-step. An example of a similar RNN architecture can be found in [13] which modeled an iterative mean-field approximation based algorithm for CRF inference as a RNN. Each iteration of the update of the marginal probabilities of a fully connected CRF was modeled by a CNN and multiple iterations of the update were modeled by adding a feedback connection (similar to our method) from the output of the CNN to its input.

The RACE-net depicted in Fig. 3 a has the same network parameters θ as the FFNN in Fig. 1 a used to model a single time-step. However, the key difference between them is the feedback connection in Fig. 3 from the output ϕ_{t+1} to its input ϕ_t in the next time-step. It allows the RACE-net to be trained in an end-to-end manner by defining a loss function over the entire curve evolution spanning multiple (T) time-steps. The initial level set function ϕ_0 is the input to the RACE-net, ϕ_T is its final output, while $\phi_t, 1 \leq t \leq T-1$ are the feature maps computed by f_θ in the intermediate time-steps of the RACE-net. A simple feedback connection without any sophisticated gated architecture (in contrast to the LSTMs) was employed in the RACE-net and auxiliary loss functions were incorporated at each individual time-step of the RNN during training to overcome the vanishing gradients problem.

A suitable loss function must maximize the regional overlap between the ground truth (GT) segmentation mask and the segmentation estimated by the RACE-net. Hence, the level set function ϕ_t is converted into a segmentation mask R_t within the the RACE-net architecture itself. Since ϕ_t is positive inside and negative outside the boundary curve, the binary segmentation mask of the foreground region can be obtained by thresholding $\phi_t > 0$. This hard thresholding operation is equivalent to applying the step function as an activation function to ϕ_t. Since step function is non-differentiable, it is approximated by a sigmoid function to derive a softmap that closely approximates a binary image, ie., $R_t = Sig(\phi_t)$. Both ϕ_t and R_t are the outputs of the RACE-net. While ϕ_t is fed back to provide the input to f_θ for the next time-step, R_t is an additional auxilliary output used to define the loss function for training. The final binary segmentation mask can be obtained by thresholding $R_T > 0.5$.

By *unrolling* the RNN through time, RACE-net can be viewed as a cascade of T feed-forward networks (f_θ) with identical parameters θ for each time-step (see Fig. 3 b). Thus, by choosing an appropriate T, an arbitrarily deep network can be obtained while keeping the number of shared network parameters constant. θ is learned in an end-to-end manner across the T time-steps using Backpropagation through time (BPTT) [23] which is logically equivalent to a simple backpropagation over the unrolled architecture of the RACE-net.

While g is a function of both I and ϕ, h is a function of I alone (see eq. 3). Thus, during implementation, h is computed only once before the feedback connection and used as an additional input at each time-step, to improve the computational efficiency.

1) Loss Function: In order to define an appropriate loss function to train the RACE-net, we need a metric to measure the regional overlap between the GT and the estimated segmentation masks. The Dice coefficient was used for this purpose. Let Y denote the binary ground truth (GT) segmentation mask for a training image I. Without any loss of generality, let R denote the segmentation mask obtained from the RACE-net for I at a particular time-step. The Dice coefficient between R and Y is defined as

$$Dice(R,Y) = \frac{2.|R \cap Y|}{|R| + |Y|} = \frac{2\sum_i r_i.y_i}{(\sum_i r_i + \sum_i y_i)}, \quad (4)$$

where r_i and y_i represents the value at the i^{th} pixel/voxel in R and Y respectively and the summations run over all the pixels/voxels in I. The Dice coefficient is differentiable with respect to r_i yielding the gradient

$$\frac{\partial}{\partial r_i}\frac{2\sum_i r_i.y_i}{(\sum_i r_i + \sum_i y_i)} = 2.\frac{y_i.(\sum_i r_i + \sum_i y_i) - (\sum_i r_i.y_i)}{(\sum_i r_i + \sum_i y_i)^2}. \quad (5)$$

The RACE-net should be trained to minimize the loss $S(R_T, Y) = 1 - Dice(R_T, Y)$ where $R_T = Sig(\phi_T)$ is the final segmentation mask obtained after the T time-steps. Though an iterative evolution of ϕ_t is commonly employed in the implementation of the LDMs, they lead to numerical instabilities and require ϕ_t to be re-initialized to a signed distance function (SDF) after every few iterations. Since $|\nabla\phi_t| = 1$ is a property of the SDF [24], an additional regularization term $|\nabla\phi_t - 1|_2^2$ is added to the loss function at each time-step to ensure that the ϕ_t remains close to a SDF during the curve evolution.

The RNN architectures are also susceptible to vanishing gradients. The gradient of the loss function at the final time-step T becomes extremely small when *backpropagated through time* to the earlier time-steps, thereby inhibiting the network's ability to learn curve evolutions across a long distance. Hence, an auxiliary loss function is added at each time-step to ensure that the network in time step t, receives multiple gradients backpropagated from the loss functions of all the following time-steps. Though the GT for the intermediate R_t is not available, we assume that the dice coefficient should successively increase with respect to Y as the curve evolves. Therefore, the auxiliary loss functions are defined as $\lambda_t S(R_t, Y)$ where λ_t represents the relative weightage of these terms such that $\lambda_t < \lambda_{t+1}$. Thus, the total loss L for the RNN is defined as

$$L = \sum_{t=1}^{T} \left\{ \lambda_t S(R_t, Y) + \beta.|\nabla\phi_t - 1|_2^2 \right\}. \quad (6)$$

In our experiments, $\lambda_t = 10^{\{\lfloor -T/2 \rfloor + t\}}$ was decayed in powers of 10 (earlier time-steps had lower weights for $S(R_t, Y)$)

while the relative weight $\beta = \lambda_1$ was fixed across all time-steps. The RACE-net is trained by minimizing L using BPTT [23] with the ADAM algorithm [25] to adapt the learning rate.

IV. Results

The proposed method was evaluated on three different segmentation tasks encompassing a wide range of anatomical structures and imaging modalities to demonstrate its robustness and generalizability. The specific tasks considered were the segmentation of the Optic Disc and Cup in color fundus images, cell nuclei in histopathology images and the left atrium in cardiac MRI volumes. The left atrium segmentation was performed in 3D. The RACE-net was implemented for both 2D and 3D images in the Keras API [26] using Theano as the backend on a 12GB Nvidia Titan X GPU and Intel i7 processor. The details of each experiment is presented below.

A. Optic Disc and Cup Segmentation

Fundus photography is used to capture 2D color images of the retina which forms the interior surface of the eye. In fundus images, the Optic disc (OD) appears as a bright, roughly elliptical structure with a central depression called the Optic Cup (OC). Glaucoma is a chronic optic neuropathy which leads to the loss of retinal nerve fibers and results in the enlargement of the OC with respect to the OD. Hence, an accurate segmentation of the OD and OC is clinically important. OD segmentation suffers from vessel occlusion, indistinct gradient at boundaries and the presence of abnormalities such as Peripapillary atrophy. OC segmentation is more challenging as it is primarily characterized by the 3D depth information which is unavailable in the 2D color fundus images [27].

The proposed method was evaluated on the public DRISHTI-GS1 dataset [28] which consists of 50 training and 51 test images with ground truth (GT) OD and OC segmentation masks. The two structures were segmented sequentially using two different RACE-net architectures. At first, a square region of interest (ROI) (depicted in the first columns of Fig. 4 b, i-iv) was extracted from the entire fundus image using a method based on [27] which employed an intensity thresholding followed by a circular hough transform on the edge map extracted from the fundus image. The initial level set function was initialized as the largest circle within the ROI. Thereafter, the result of the OD segmentation was used to initialize the OC segmentation. The network parameters were fixed at N=6(6) and T=6(5) for OD (OC). The filter sizes S_l at each scale were fixed at 9×9, 7×7, 5×5 and 3×3 respectively from the coarse ($l = 3$) to the fine ($l = 0$) scale in both the architectures, resulting in a total of $26,322$ learnable parameters for each network. The number of network parameters is independent of T since they are shared across the time steps. The training dataset was augmented by applying horizontal, vertical translations and vertical flipping to each image. During testing, the RACE-net completed OD and OC segmentation at an average of 0.21 and 0.19 seconds per image respectively, using a GPU.

The qualitative results depicted in Fig. 4 demonstrate the strength of the RACE-net. The curve evolution for OD and OC in Fig. 4 a indicates that the network architecture is able to learn to evolve inwards, stopping at the desired boundary while preserving the boundary smoothness. The OC boundary does not correspond to a sharp intensity gradient demonstrating the strength of the method over traditional edge based LDMs.

TABLE I
SEGMENTATION PERFORMANCE ON DRISHTI-GS1 TEST SET. (MEAN/STANDARD DEVIATION)

Method	Optic Disc		Optic Cup	
	Dice	BLE (pixels)	Dice	BLE
Vessel Bend [29]	0.96/0.02	8.93/2.96	0.77/0.20	30.51/24.80
Graph cut prior [34]	0.94/0.06	14.74/15.66	0.77/0.16	26.70/16.67
Multiview [30]	0.96/0.02	8.93/2.96	0.79/0.18	25.28/18.00
Superpixel [31]	0.95/0.02	9.38/5.75	0.80/0.14	22.04/12.57
Joint OD-OC [27]	**0.97/0.02**	6.61/3.55	0.83/0.15	18.61/13.02
CNN, Zilly et. al. [32]	94.7	9.4	0.83	16.5
CNN, U-net [7]	0.96/0.02	7.23/4.51	0.85/0.10	19.53/13.98
CNN, Sevastopolsky [33]	–	–	0.85	–
Proposed	**0.97/0.02**	**6.06/3.84**	**0.87/0.09**	**16.13/7.63**

The quantitative performance has been reported in Table I. In addition to the Dice coefficient, the average Boundary Localization Error (BLE) [27], [28] has also been reported which measures the average distance (in pixels) between the GT and the estimated boundary points. It is defined as $BLE = \frac{1}{n}\sum_{\theta_1}^{\theta_n} |r_{gt}^{\theta} - r_{est}^{\theta}|$, where r_{est}^{θ} and r_{gt}^{θ} denote the Euclidean distances (in the radial direction) of the estimated and the GT boundary points from the centroid of the GT respectively, at orientation θ. The average BLE is computed based on the localisation errors at $n = 24$ equi-spaced boundary points. The desirable value for BLE should be close to 0 pixels. The RACE-net outperforms several existing methods. Both [29] and [30] employ a modified Chan-Vese based LDM to segment the OD while a supervised super-pixel feature classification based method is employed in [31] for both OD and OC segmentation. On the task of OD segmentation, the proposed architecture shows marginal improvement with respect to these methods and is at par with the recently proposed MRF based joint OD and OC segmentation method in [27].

Since, the depth information is important in defining the OC boundary, [29] detects the bend in the blood vessels as they enter the OC, [30] employs stereo image pairs, while [27] employs a coupled sparse dictionary based regression to obtain the depth estimates. In comparison, RACE-net outperforms these methods while directly using the raw RGB color channels as input without the need for explicit depth computation. The RACE-net also outperforms the existing CNN architectures proposed in [32], the U-net [7] and its modified version in [33] which was used to segment the OC .

B. Cell Nuclei segmentation

The segmentation of the cell nuclei in hematoxylin and eosin ($H\&E$) stained histopathological images plays an important role in the detection of breast cancer. The cell nuclei appear as roughly elliptical purple blobs surrounded by a pink cytoplasm with the connective tissues appearing as wispy pink filaments (see Fig. 5). The challenges here include a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2852635, IEEE Journal of Biomedical and Health Informatics

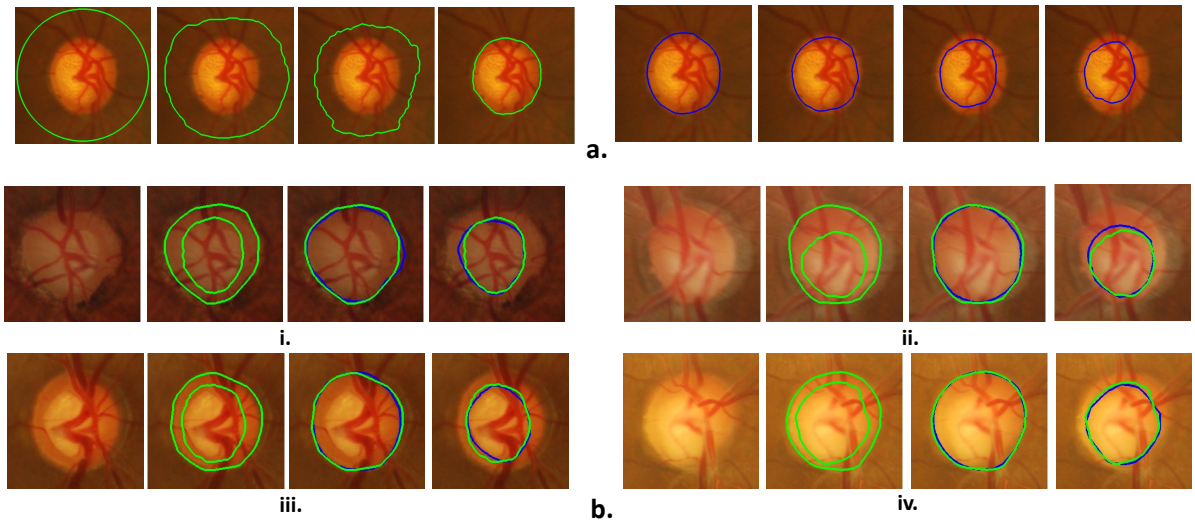

Fig. 4. **a.** depicts the curve evolution of RACE-net over the intermediate time-steps, T=0, 2, 4, 6 for the OD boundary (in green) and over T=0, 1, 3, 5 for the OC boundary (in blue) from left to right. **Fig. b** depicts four sample results for OD and OC boundaries. In each subimage **i-iv**, the cropped region of interest is depicted in column 1, the result of the RACE-net for the OD and OC boundaries (in green) are depicted in column 2 followed by a comparison with the Ground truth markings (in blue) for OD (column 3) and OC (column 4). This figure is best viewed in color.

large background clutter, intensity inhomogeneities, artifacts introduced during the slide preparation, staining or imaging and the variations in the morphology across a large number of cells in the image. The task unlike the previous case study involves the segmentation of multiple structures in each image.

The proposed method was evaluated on 58 images consisting of 26 malignant and 32 benign cases provided in the UCSB Bio-segmentation Benchmark dataset [35]. Each image comprises a 200×200 ROI along with a pixel-level binary mask provided as the GT. For a direct comparison with the recent work in [36], similar experimental setup and evaluation metrics were employed. The dataset was randomly divided into 24 images for training, 6 for validation and 28 for testing respectively. The network hypermeters were set to $N = 5$ and $T = 6$. The filter sizes were fixed to 9×9, 7×7, 5×5 and 3×3 respectively from the coarse to the fine scale. The training dataset was augmented by horizontal and vertical flipping, and rotating the images by $\pm 45^{\circ}$. During testing, segmentation with the RACE-net took an average of 0.04 seconds using a GPU to process each image.

TABLE II
CELL NUCLEI SEGMENTATION PERFORMANCE ON THE UCSB BIO-SEGMENTATION BENCHMARK DATASET

Method	Precision(%)	Recall(%)	Accuracy(%)	F1-measure
FCM [37]	71.63	84.86	88.84	0.7768
Watershed based [38]	70.26	87.01	88.56	0.7775
MI [39]	-	-	89.55	0.7733
DRLSE [24]	88.24	74.87		0.8042
CNN+SR [36]	82.41	86.04	92.45	0.8393
CNN [40]	-	-	86.88	-
CNN, U-net [7]	81.41	85.84	92.21	0.8333
Proposed	**85.29**	**88.38**	**93.82**	**0.8661**

Raw RGB image channels were used as the input without any preprocessing and the border of the entire ROI was provided as the intitial boundary. Qualitative results depicted in Fig. 5 b demonstrate the strength of the proposed architecture. RACE-net effectively handles the topological changes associated with the splitting of the boundary curve to capture multiple cell nuclei. It is also capable of converging onto the desired boundary over large distances thereby eliminating the need for any preprocessing step for an accurate initialization. Fig. 5 a depicts the evolution of the level set curve across the time-steps. In the early time steps (T = 1-3), the level set curve progresses inwards, roughly segmenting the cell nuclei from the outer to the inner regions, while the final time-steps refine the object boundaries.

Results of quantitative assessment are reported in Table II. The results show that the RACE-net outperforms the traditional image processing based methods employing Fuzzy C-means clustering [37], semi-supervised multi-image model (MI) [39] and the marker controlled watershed [38]. It also achieves a 6% improvement in F1-measure against a deformable model [24] with distance regularized level set evolution (DRLSE) that employs handcrafted velocity terms. Unlike the proposed method, DRLSE failed to converge onto the cell boundaries over a large distance and Otsu thresholding was employed to initialize the level set. A comparison was also done against the existing CNN-based methods which have generally outperformed the traditional methods. The proposed method was found to perform better in this case as well. Moreover, while RACE-net employs the raw RGB images as input, the method in [36] employs a computationally expensive sparse coding based preprocessing step to roughly remove the background and accentuate the nuclei regions.

C. Left Atrium Segmentation

The accurate segmentation of the Left Atrium (LA) and the proximal pulmonary veins (PPVs) in cardiac MRI volumes plays an important role in the ablation therapy planning for

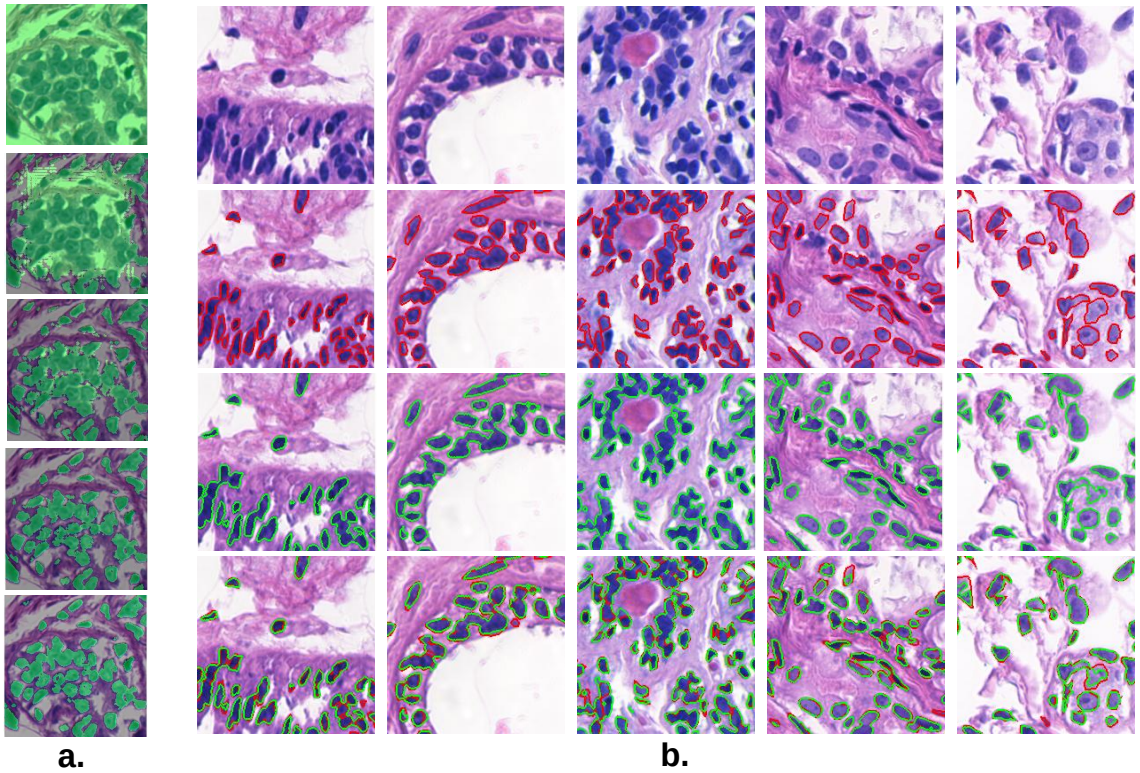

Fig. 5. **a.** The curve evolution over time steps T=0,2,3,4,6 (from top to bottom). **Fig. b.** depicts the qualitative results of the RACE-net: $\mathbf{1}^{st}$ **row:** input image, $\mathbf{2}^{nd}$ **row:** Ground truth markings (in red), $\mathbf{3}^{rd}$ **row:** result of RACE-NET (in green), $\mathbf{4}^{th}$ **row:** the result of RACE-net (in green) and the Ground truth markings (in red) are overlapped for comparison. This figure is best viewed in color.

the treatment of Atrial fibrillation, automated quantification of LA fibrosis and the construction of biophysical cardiac models. However, this task is non-trivial because of the large inter-subject morphological variations in the anatomy, overlapping image intensities between LA and the surrounding tissue, very thin myocardial walls and difficulty in demarcating LA from the left ventricle due to the different opening positions and low visibility of the mitral valve leaflets [41].

The proposed method was evaluated on the public STACOM 2013 left atrium segmentation challenge dataset [41], [42]. It consists of 10 training and 20 test cardiac MRI volumes. The binary segmentation masks for the LA and PPVs are provided for the training images, while an evaluation code written in ITK is provided for the test set. The MRI volumes were acquired using a 1.5T Philips Achieva scanner with an in-plane resolution of $1.25 \times 1.25\ mm^2$ and a slice thickness of $2.7\ mm$. The PPVs were defined as the segment of the pulmonary vein up to the first vessel branching or a maximum of 10 mm. We refer to [41] for further details of the protocols used in the acquisition of the images and GT.

Since all MRI volumes were acquired using the same view and resolution, the spatial priors on the anatomical structures were exploited to reduce the computational and memory requirements. An ROI was defined as a cuboid of dimensions $72 \times 104 \times 116$ centered at the image coordinates $(36, 52, 56)$ in each volume that contained the left atrium. Since the volumes were not registered, a relatively large ROI (see Fig. 6 c) had to be defined to ensure that the left atrium lies within it across all the images. On an average, the left atrium constituted only 8.05% of the ROI volume and appeared at varying spatial locations within the ROI. The level set surface ϕ_0 was initialized in a simple manner at the border of the ROI.

The proposed RACE-net architecture was extended to 3D by employing 3D convolutional filters within the CNNs that modeled the constant and mean curvature evolution velocities. The customized layer depicted in Fig. 1 b was also modified (written in Theano library [21]) to compute the gradient magnitude and the curvature by considering the x, y as well as the z directions. The network hyperparameters were fixed to $N = 8$ and $T = 7$. The convolutional filter sizes were fixed to $7 \times 7 \times 7$, $5 \times 5 \times 5$, $3 \times 3 \times 3$ and $3 \times 3 \times 3$ respectively from the coarse to fine scale.

The training dataset was augmented by applying random translations along the three directions and modifying the image intensity at a patch level by adapting the PCA based technique employed in [43]. In this method, the training volumes were divided into $3 \times 3 \times 3$ patches and projected onto a 27 dimensional PCA basis. Thereafter, a random Gaussian noise with zero mean and a standard deviation of 0.1 was added to each basis coefficient and the image patches were reconstructed.

The qualitative results of the proposed method is depicted in Fig. 6. A 3D rendering of the segmentation results in the intermediate time steps is presented in Fig. 6 a to visualize the surface evolution. In the first time step $T = 1$ itself, the initial surface can be seen to jump over a large distance from the border of the ROI to provide a moderately good localization of the left atrium. Thereafter, the shape of the structure is iteratively refined with large changes in the early iterations

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2852635, IEEE Journal of Biomedical and Health Informatics

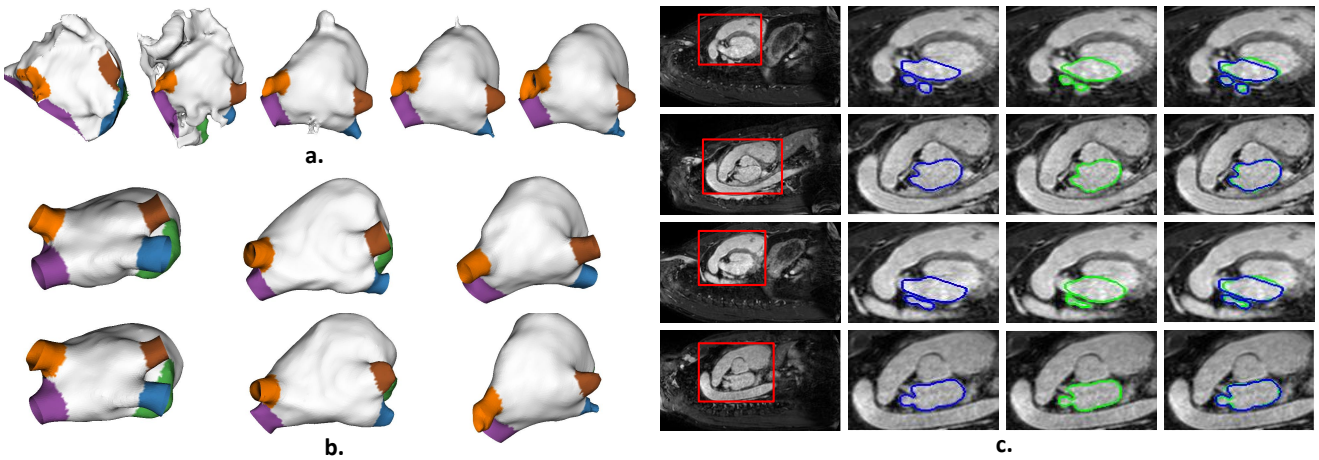

Fig. 6. **a.** The curve evolution over time steps T=1,2,3,5 and 7 (from left to right). **b.** The 3D surface rendering of the segmented left atrium: $\mathbf{1}^{st}$ row is the Ground truth and second row depicts the corresponding segmentation by the RACE-net. **c.** Qualitative results for the individual slices depicted in the $\mathbf{1}^{st}$ column are provided. Only a small region marked by the red bounding box has been magnified for better visibility of the result. $\mathbf{2}^{nd}$ column depicts the Ground Truth markings in blue, $\mathbf{3}^{rd}$ column depicts the result of RACE-NET depicted in green. The Ground truth and the result of the RACE-net are overlapped for comparison in the $\mathbf{4}^{th}$ column. This image is best viewed in color.

(T=2,3) followed by fine refinements in the later time steps. In Fig. 6 b, sample results are presented for 3 MRI test volumes where the 3D rendering of the proposed method is compared against the GT. These results demonstrate the ability of the method to handle sharp corners and protrusions in the object boundary due to the PPVs. The qualitative results on individual MRI slices depicted in Fig. 6 c indicate the ability of RACE-net to learn topological changes (splitting of the surface) near the PPV openings.

TABLE III
3D LEFT ATRIUM SEGMENTATION PERFORMANCE ON THE STACOM 2013 CHALLENGE DATASET (MEAN/STANDARD DEVIATION)

Method	Left Atrium		PPVs	
	Dice	S2S (mm)	Dice	S2S (mm)
LUB_SSM	0.77/0.13	3.63/1.83	0.28/0.21	4.73/4.65
INRIA	0.78/0.26	3.66/4.59	0.42/0.28	4.66/7.74
LUB_SRG	0.84/0.12	2.67/1.75	0.51/0.31	2.80/2.63
TLEMCEN	0.85/0.07	2.40/0.96	0.37/0.30	3.68/3.20
LTSI_VSRG	0.87/0.03	2.22/0.32	0.48/0.19	2.54/0.54
LTSI_VRG	0.91/0.05	1.68/0.90	0.65/0.17	1.95/1.04
UCL_1C	0.94/0.02	1.09/0.31	0.61/0.23	1.62/0.61
3D U-net [12]	0.90/0.04	1.84/0.75	0.63/0.19	1.97/0.99
RACE-net	0.91/0.04	1.73/0.72	0.61/0.25	1.90/1.01
RACE-net ensemble	0.92/0.03	1.49/0.56	0.67/0.24	1.66/0.79
Human Expert 2	0.95/0.05	0.88/0.78	0.83/0.11	1.02/0.64

The quantitative assessment of results are presented in Table III. The evaluation code provided by the STACOM 2013 left atrium segmentation challenge [42] was employed to compute the Dice coefficient and the surface-to-surface distance (S2S) in mm [41]. A good segmentation is characterized by a high value for the Dice coefficient and low value for the S2S metric. The S2S metric is more sensitive to the local variations in the boundary as compared to dice which measures the global regional overlap. The proposed method was benchmarked against the performance of the other competing methods in the STACOM 2013 left atrium segmentation challenge [41] as well as the 3D U-net architecture [12].

The overall performance of the RACE-net is comparable to that of the 3D U-net. The RACE-net performed slighty better on the left atrium body both in terms of Dice and the S2S metrics. In case of the PPV, 3D U-net performed slightly better in terms of Dice while RACE-net performed slightly better in terms of the S2S. However, RACE-net provides significant gains in terms of the computational and memory requirements. We refer to Section V for the details.

In comparison to the participating methods in the STACOM 2013 left atrium segmentation challenge, our method achieved a dice of 0.91 for the left atrium body performing comparably to the second best performing method $LTSI\_VRG$, with UCL_1C achieving the best performance with a dice of 0.94. The PPVs offer a greater challenge for segmentation. Again the proposed method is the second best with a dice of 0.61 performing comparably to UCL_1C but below the performance of the $LTSI\_VRG$ method with a dice of 0.65. The biggest strength of the RACE-net over these methods is its computation speed; 1.72 seconds to process each MRI volume on an average using a GPU which is orders of magnitude lower than that reported by UCL_1C and $LTSI\_VRG$, namely 1200 and 3100 seconds respectively.

Incidentally, both UCL_1C and $LTSI\_VRG$ employed a multi-atlas approach. In these methods, multiple segmentation results were obtained by registering each atlas template to the test image followed by a majority voting scheme to obtain the final segmentation. Inspired from these methods, we evaluated an additional strategy which we refer to as the *RACE-net ensemble*. In this scheme, the data augmentation techniques used during the training were also applied to each test volume to obtain three additional images. The segmentation of all the four images were obtained using the RACE-net. Thereafter, the segmentation masks of the augmented images were brought back to the original co-ordinate space by reversing the translations employed during their construction. The average of the four segmentation masks was computed and thresholded at 0.5 to obtain the final binary segmentation map. Though

the *RACE-net ensemble* increases the computational time, it significantly improves the robustness of the method. While there is a slight improvement of 1% for the LA body in terms of dice, the performance of the PPV improves from 0.61 to 0.67 which outperforms all the existing methods.

V. Discussion

A. Performance

The results presented in the previous section show that the performance of the RACE-net architecture is comparable to the state of the art on a wide range of segmentation tasks involving different anatomical structures and imaging modality, outperforming the existing methods in most cases. RACE-net closely approximates the level set based active contour models within a deep learning framework. In the OD and OC segmentation task, it learned to extract the OC boundary though it is not characterized by a sharp intensity gradient and preserved the boundary smoothness constraints for both structures. On the task of cell nuclei segmentation, RACE-net demonstrated its ability to allow topological changes by allowing the boundary to split in order to capture multiple cell nuclei. RACE-net was also extended to 3D for left atrium segmentation, where the method showed its ability to converge to the desired boundary over large distances in only a few time-steps. A recent, unpublished work [44] has explored an alternative way to incorporate CNNs within the active contour model framework to evolve a curve over small distances (30 pixels). It uses a parametric representation for the boundary and hence cannot handle topological changes. Furthermore, the velocity vectors at each boundary point are predicted independently at a local image patch level, requiring a handcrafted regularization step after each iteration to preserve the boundary smoothness.

The performance of RACE-net is summarized in Table IV. The BLE and the S2S were used as the Boundary error metrics for the OD, OC and the left atrium segmentation tasks respectively. A paired T-test was done to compare the mean performance of the proposed method against that of the U-net. On the task of the OD, cell nuclei and the left atrium segmentation, the p-value is < 0.05 for both the dice and the Boundary error metric indicating a statistically significant improvement in performance. However, in case of the OC and PPV, though our method's performance is marginally better than the U-net, the difference is not statistically significant.

TABLE IV
Statistics for the Dice and Boundary Error of the proposed RACE-net on various tasks.

	min	max	mean	median	standard deviation	p value vs. U-net
Dice						
Optic Disc	0.88	0.99	0.97	0.98	0.02	0.0042
Optic Cup	0.51	0.95	0.87	0.89	0.09	0.1016
Cell nuclei	0.84	0.91	0.87	0.87	0.02	<0.0001
Left Atrium	0.86	0.96	0.92	0.92	0.03	0.0239
PPV	0.20	0.88	0.67	0.70	0.16	0.4033
Boundary Error (pixels)						
Optic Disc	2.48	24.60	6.06	4.84	3.84	0.0095
Optic Cup	6.53	37.32	16.13	14.13	7.63	0.1537
Cell nuclei	–	–	–	–	–	–
Left Atrium	0.70	2.66	1.49	1.47	0.56	0.0236
PPV	0.87	3.58	1.66	1.41	0.79	0.1373

B. Failure cases

The qualitative examples of a few failure cases of the RACE-net are depicted in Fig. 7. Additionally, the results obtained using the U-net has been provided for comparison. Although RACE-net performs very well for OD segmentation in general, the failure case shown in Fig. 7 a is an exception which has the minimum dice (of 0.87) across all the test images in the DRISHTI-GS1 dataset. In this case, the RACE-net over-estimates the OD boundary and is attracted towards the thick blood vessels in the Inferior-nasal (bottom-right) quadrant. On the other hand, the U-net underestimates it and is attracted towards the pallor boundary in the inferior (bottom) sector. The pallor is a pale yellow region within the OD with maximum color contrast that lies close to the OC. The possible reason for failure in this case could be due to the poor contrast and the lack of a strong gradient at the OD boundary in the inferior sector.

OC segmentation from fundus images is a challenging task and is primarily guided by the intensity gradient at the pallor boundary and the bends in the thin blood vessels within the OD. However, the image depicted in Fig 7 b doesnot have a distinct pallor. Though the result of the RACE-net closely follows the GT in the inferior sector, it may have been attracted to the bend in the thin vessel in the superior (top-right) sector.

A failure case for the cell nuclei segmentation in histopathological images is depicted in In Fig. 7 c. In this case, the performance of RACE-net is slightly better than the that of the U-net (dice of 0.84 in comparison to 0.81). However, both the methods exhibit a tendency to miss the boundaries between the adjacent or overlapping cell nuclei.

In case of the left atrium segmentation, the errors in segmentation typically occur in slices near the PPV openings which are characterized by sharp corners and topological changes in the boundaries. Two such examples of failure cases are provided in Fig. 7 d. The example in the first row depicts a case where RACE-net under-estimates the boundary of the larger connected component and fails to detect the smaller one. In this case, there is a lack of contrast and distinct intensity gradient between the foreground and background regions. Another example (second row in Fig. 7 d) depicts a case of overestimation where instead of two connected components, RACE-net smoothes out the boundary resulting in a single larger region encompassing both.

C. Computational time and network size

RACE-net is ideal for deployment on systems with limited resources such as low memory and the unavailability of a GPU. Typically, the existing CNNs employ millions of parameters. For eg., the U-net architecture described in [7] employs $34,513,345$ parameters. However, the memory bandwith and storage requirements of a deep learning system is directly proportional to the number of network parameters. Therefore, the network parameters in RACE-net were kept to a minimum

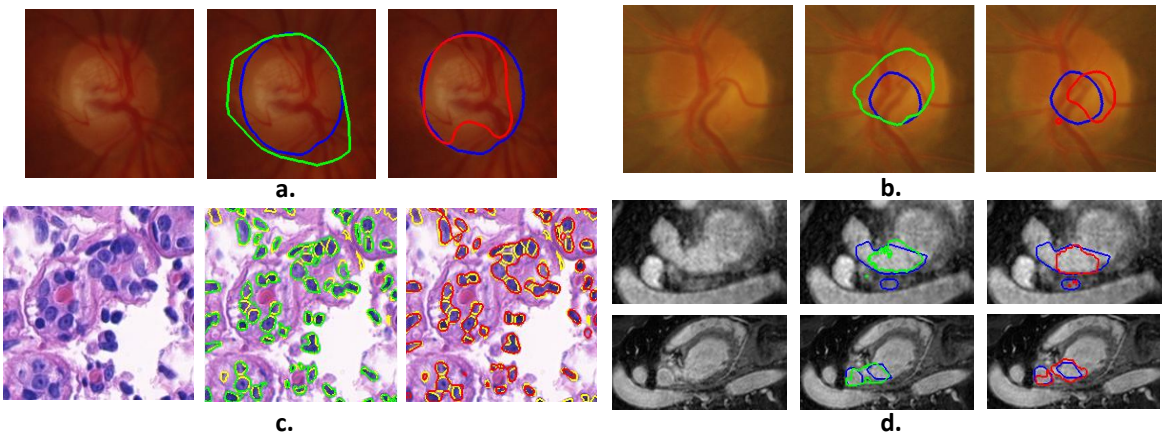

Fig. 7. Examples of failure cases for: **a.** Optic disc, **b.** Optic cup, **c.** cell nuclei and **d.** left atrium. The Ground truth is marked in blue in **a.**, **b**, **d** and yellow in **c**. The results of RACE-net are marked in green in each case. The results obtained using U-net (marked in red in each case) are provided for comparison. This image is best viewed in color.

TABLE V
AVERAGE TIME (IN SECONDS) TO SEGMENT EACH IMAGE USING CPU.

	Optic Disc	Optic Cup	Cell Nuclei	Left atrium
U-net	5.88	7.79	1.92	24.13
RACE-net	0.74	0.60	0.15	7.74

by applying two strategies. First, $g(I, \phi)$ and $h(I)$ functions in Fig. 1 were approximated by a novel CNN architecture inspired from the multi-scale image pyramid that employed very few parameters. Secondly, each time-step of the RACE-net shares the same network parameters thereby allowing for arbitarily complex architectures by adjusting T while keeping the number of shared network parameters constant. As a result, the RACE-net architecture for the cell nuclei segmentation (described in Section IV-B) employs only $8,394$ parameters. Similarly, for the OC segmentation task, the modified U-net architecure in [33] employed about 25 times the number of network parameters in the RACE-net ([33] employed $660,000$ compared to the $26,322$ network parameters in the RACE-net). On the 3D left atrium segmentation task, the RACE-net architecture detailed in Section IV-C employed $71,862$ learnable netwrork parameters as compared to the 3D U-net [12] with $10,003,401$ parameters.

RACE-net runs moderately fast even in the absence of a GPU. To demonstrate this, the average segmentation time per image of the RACE-net is compared against the U-net [7] in Table V. Both the networks were tested on all the three segmentation tasks without using a GPU. The 3D extension of the U-net in [12] was used for the left atrium segmentation. The results clearly illustrate the advantage of our method with a 3 to 12 times speedup across the different tasks. The low run time can be attributed to the fact that each time-step of the RACE-net is modeled using a very small FFNN that employs few computations. Moreover, the RACE-net only requires a few time-steps (5-7) to converge and uses a simple feedback connection instead of the computationally expensive gated architecture used in LSTMs [16] which further improves the computational efficiency. During training, the vanishing gradients problem is handled by using auxiliary loss functions at each time-step.

VI. CONCLUSION

In this paper, a novel RNN architecture (RACE-net) was proposed for biomedical image segmentation which models the object boundaries as an evolving level set curve. Each time-step of the curve evolution is modeled using a FFNN that employs a customized layer to compute the normal and curvature of the level set function and a novel CNN architecture is used to model the curve evolution velocities. The size of the CNN is determined by a single hyper-parameter N which controls the number of convolutional filters at each scale. The recurrent feedback connections are used to model the T time steps of the curve evolution.

Since RACE-net can be viewed as a cascade of feedforward networks with shared network weights, an arbitarily deep network can be obtained to model complex long range, high-level dependencies by adjusting T while keeping the number of shared network parameters constant. Overfitting is avoided with fewer network parameters which is also critical for the deployment of RACE-net on devices with limited memory and computational resources. The entire RACE-net is trained in a supervised, end-to-end manner. An appropriate loss function was defined as a weighted sum of intermediate dice coefficients at each time-step to mitigate the vanishing gradients problem and a regularization term on the level set function was also incorporated to ensure its numerical stability.

Consistent performance of RACE-net on a diverse set of applications indicates its utility as a generic, off-the-shelf architecture for biomedical segmentation. Since RACE-net is based on the deformable models, it has the potential to incorporate explicit shape priors which presents a promising direction for work in the future.

REFERENCES

[1] T. McInerney and D. Terzopoulos, "Deformable models in medical image analysis: a survey," *Medical image analysis*, vol. 1, no. 2, pp. 91–108, 1996.

[2] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.

[3] L. D. Cohen and I. Cohen, "Finite-element methods for active contour models and balloons for 2-d and 3-d images," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 11, pp. 1131–1147, 1993.

[4] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Transactions on image processing*, vol. 7, no. 3, pp. 359–369, 1998.

[5] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.

[6] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

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

[8] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 565–571.

[9] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Deep retinal image understanding," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 140–148.

[10] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[11] K. H. Cha, L. Hadjiiski, R. K. Samala, H.-P. Chan, E. M. Caoili, and R. H. Cohan, "Urinary bladder segmentation in ct urography using deep-learning convolutional neural network and level sets," *Medical physics*, vol. 43, no. 4, pp. 1882–1896, 2016.

[12] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.

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

[14] K. Gopinath, S. B. Rangrej, and J. Sivaswamy, "A deep learning framework for segmentation of retinal layers from oct images," in *Fourth Asian Conference on Pattern Recognition*, 2017, p. to appear.

[15] J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen, "Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation," in *Advances in Neural Information Processing Systems*, 2016, pp. 3036–3044.

[16] Y. Xie, Z. Zhang, M. Sapkota, and L. Yang, "Spatial clockwork recurrent neural network for muscle perimysium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 185–193.

[17] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber, "Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation," in *Advances in Neural Information Processing Systems*, 2015, pp. 2998–3006.

[18] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 126–136, 2018.

[19] A. Mitiche and I. B. Ayed, *Variational and level set methods in image segmentation*. Springer Science & Business Media, 2010, vol. 5, pg. 17–19.

[20] G. Sapiro, *Geometric partial differential equations and image analysis*. Cambridge university press, 2006, pg. 74–76.

[21] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

[22] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[23] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[24] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE transactions on image processing*, vol. 19, no. 12, pp. 3243–3254, 2010.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint:1412.6980*, 2014.

[26] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.

[27] A. Chakravarty and J. Sivaswamy, "Joint optic disc and cup boundary extraction from monocular fundus images," *Computer Methods and Programs in Biomedicine*, vol. 147, pp. 51 – 61, 2017.

[28] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish *et al.*, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomedical Imaging Data Papers*, vol. 2, no. 1, p. 1004, 2015.

[29] G. D. Joshi, J. Sivaswamy, and S. Krishnadas, "Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment," *IEEE Transactions on Medical Imaging*, vol. 30, pp. 1192–1205, 2011.

[30] G. Joshi, J. Sivaswamy, and S. Krishnadas, "Depth discontinuity-based cup segmentation from multiview color retinal images," *IEEE Transactions on IT in Biomedicine*, vol. 59, no. 6, pp. 1523–1531, 2012.

[31] J. Cheng, J. Liu, Y. Xu, F. Yin, D. W. K. Wong, N.-M. Tan, D. Tao, C.-Y. Cheng, T. Aung, and T. Y. Wong, "Superpixel classification based optic disc and optic cup segmentation for glaucoma screening," *IEEE Transactions on Medical Imaging*, vol. 32, no. 6, pp. 1019–1032, 2013.

[32] J. G. Zilly, J. M. Buhmann, and D. Mahapatra, "Boosting convolutional filters with entropy sampling for optic cup and disc image segmentation from fundus images," in *International Workshop on Machine Learning in Medical Imaging*, 2015, pp. 136–143.

[33] A. Sevastopolsky, "Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network," *Pattern Recognition and Image Analysis*, vol. 27, no. 3, pp. 618–624, Jul 2017.

[34] Y. Zheng, D. Stambolian, J. OBrien, and J. C. Gee, "Optic disc and cup segmentation from color fundus photograph using graph cut with priors," in *Medical Image Computing and Computer-Assisted Intervention MICCAI*, 2013, pp. 75–82.

[35] E. D. Gelasca, B. Obara, D. Fedorov, K. Kvilekval, and B. Manjunath, "A biosegmentation benchmark for evaluation of bioimage analysis methods," *BMC bioinformatics*, vol. 10, no. 1, p. 368, 2009.

[36] X. Pan, L. Li, H. Yang, Z. Liu, J. Yang, L. Zhao, and Y. Fan, "Accurate segmentation of nuclei in pathological images via sparse reconstruction and deep convolutional networks," *Neurocomputing*, vol. 229, pp. 88 – 99, 2017.

[37] J. Tang, I. N. Mat, M. Schwarzfischer, and E. Chng, "A fuzzy-c-means-clustering approach: Quantifying chromatin pattern of non-neoplastic cervical squamous cells." *PLoS ONE*, vol. 10, no. 11, 2015.

[38] F. Buggenthin, C. Marr, M. Schwarzfischer, P. S. Hoppe, O. Hilsenbeck, T. Schroeder, and F. J. Theis, "An automatic method for robust and fast cell detection in bright field images from high-throughput microscopy," *BMC Bioinformatics*, vol. 14, no. 1, p. 297, Oct 2013.

[39] Y. N. Law, H. K. Lee, M. K. Ng, and A. M. Yip, "A semisupervised segmentation model for collections of images," *IEEE transactions on Image processing*, vol. 21, no. 6, pp. 2955–2968, 2012.

[40] N. Hatipoglu and G. Bilgin, "Classification of histopathological images using convolutional neural network," in *Image Processing Theory, Tools and Applications (IPTA), 2014 4th International Conference on*, 2014, pp. 1–6.

[41] C. Tobon-Gomez, A. J. Geers, J. Peters, J. Weese, K. Pinto, R. Karim, M. Ammar, A. Daoudi, J. Margeta, Z. Sandoval *et al.*, "Benchmark for algorithms segmenting the left atrium from 3d ct and mri datasets," *IEEE transactions on medical imaging*, vol. 34, no. 7, pp. 1460–1473, 2015.

[42] C. Toboz-Gomez", *Left Atrium Segmentation Challenge*, 2012. [Online]. Available: http://www.cardiacatlas.org/challenges/left-atrium-segmentation-challenge/

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[44] C. Rupprecht, E. Huaroc, M. Baust, and N. Navab, "Deep active contours," *arXiv preprint arXiv:1607.05074*, 2016.