Level Set based Shape Prior and Deep Learning for Image Segmentation

Yongming Han, Shuheng Zhang, Zhiqing Geng*
College of Information Science & Technology, Beijing University of Chemical Technology, Beijing 100029, China;
Engineering Research Center of Intelligent PSE, Ministry of Education in China, Beijing 100029, China

***Abstract*—Deep convolutional neural network (DCNN) can effectively extract the hidden patterns in images and learn realistic image priors from the training set. And fully convolutional networks (FCNs) have achieved state-of-the-art performance in the task of the image segmentation. However, these methods have the disadvantages of the noise, boundary roughness and no prior shape. Therefore, this paper proposes a level set with the deep prior method for the image segmentation based on the priors learned by FCNs from the training set. The output of the FCNs is treated as a probability map and the global affine transformation (GAT) is used to obtain the optimal affine transformation of the intrinsic prior shape at a specific image. And then, the level set method is used to integrate the information of the original image, the probability map and the corrected prior shape to achieve the image segmentation. Compared with the traditional level set method for images of simple scenes, the proposed method combines the traditional level set method with FCNs and corrected prior shape to solve the disadvantage of FCNs, making the level set method applicable to the image segmentation of complex scenes. Finally, a series of experiments with Portrait data set are used to verify the effectiveness of the proposed method. The experimental results show that the proposed method can obtain more accurate segmentation results than the traditional FCNs.**

***Index Terms*—Image Segmentation, Deep Learning, Level Set, Deep Prior, Global Affine Transformation.**

I. Introduction

IMAGE segmentation has been a hot research direction for the past decades which accurately mark the desired regions in the image [1]. And the deep learning has refreshed the research methods of many problems and becomes a watershed of ”traditional” methods and ”deep” methods [2]. Similarly, these methods for the image segmentation are also divided into traditional methods and deep learning methods.

Most of the traditional methods for the image segmentation are based on the global or local statistics information of a single image. And the image segmentation process for the target regions is mainly based on the pre-defined statistical assumptions. Threshold-based methods determined adaptively the threshold grayscale based on the global or local grayscale histogram of images [3]. And edge-based methods relied on ”the regions in the image uniquely determine the edges”. And the edges was detected by the edges detection methods to determine the segmentation regions [4]. Region-based methods obtained the final segmentation regions by dividing and merging similar regions [5], [6]. Graph-based methods used the graph structure to represent images, and then cut connections of the graph to achieve the image segmentation indirectly [7]. The mean shift method mapped all points to the high-dim feature space, and divided the regions based on the mean shift clustering [8]. Active contour-based methods modeled the target contour explicitly [9] or implicitly [10], and the target contours were obtained by minimizing a energy function. However, these traditional methods are viewed as unsupervised methods, which are based on artificially defined patterns rather than patterns learned from labeled segmentation results.

Different from the traditional methods, deep learning methods are more inclined to find the patterns in images through the training set. Deep neural networks (DNNs) have powerful ability to represent high-level features, so the image segmentation task in DNNs is extended to the semantic segmentation by segmenting regions with the complex high-level semantic information. Deep Learning methods for the image segmentation predict the category of each pixel in images. Fully convolutional networks (FCNs) replace fully connected layers in the network with convolutional layers to accomplish this dense prediction [11]. However, deep learning methods for the image segmentation have the disadvantages of the noise, boundary roughness and no prior shape [12], [13]. Therefore, the improved FCN bsed on Conditional Random Fields (CRFs) is used to solve the problems of noisy and imprecise at boundaries [14], [15].

Deep convolutional networks (DCNNs) can effectively extract the hidden patterns in images and learn realistic image priors from a large number of example images [16]. The prior obtained by the deep learning was used to initialize the surface of the level set as the shape prior in the iterative process [17], [18]. However, the prior knowledge generated by the deep leaning is also rough and imprecise. And the inherent shape prior of the target do not be taken into consideration.

Based on the shape prior representing the intrinsic shape of the target, this paper proposes a level set with deep prior method for the image segmentation based on the priors learned by FCNs. The shape prior is adjusted with the affine transformation to fit a specific image by Global Affine Transformation (GAT). And then the information of the original image, the probability map and the corrected shape prior are combined based on the level set method to obtain the segmentation results. Finally, a series

*Corresponding author.
E-mail address: gengzhiqiang@mail.buct.edu.cn (Zhiqiang Geng).

of experiments with Portrait data set [19] are used to verify the effectiveness of the proposed method. The experimental results show that the proposed method can obtain more accurate segmentation result than the traditional FCNs.

The rest of this paper is organized as follows. In Section II, the proposed method is described, and the original image, the probability map and the corrected shape prior is introduced in detail. The Portrait data set and the experimental results are shown in Section III. In section IV, some discussion of the image segmentation and the deep learning based on the level let method are described. Finally, the conclusion and future work are given in Section V.

II. The Level Set with the Deep Prior Method

In order to improve the performance of FCNs and complete objective segmentation of the complex sense by using the level set method, the level set with the deep prior method is proposed as shown in Fig. 1. The output of FCNs is taken as a

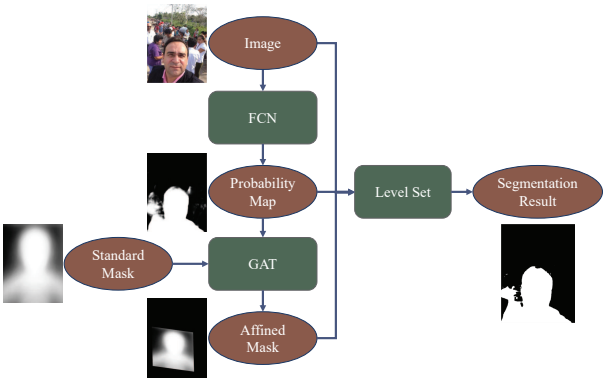


Fig. 1: The flow diagram of the proposed method.

probability map that each pixel belongs to a different category. The segmentation shape represented by the probability map is noisy, but it still retains a large part of the correct segmentation. Therefore, an optimal affine transformation of the standard shape mask (the shape prior) of the image can be obtained based on the "probability" shape with the GAT method. Finally, the image, the probability map and the affine mask are used as the input of the level set method to implement the image segmentation.

A. Deep Learning for the Probability Map

Deep Learning as an end-to-end method has achieved many excellent performances in semantic segmentation tasks. Because each pixel in the picture needs to be classified, the traditional network structure is useless. Long et al. proposed the FCN to replace the fully connected layer in the traditional convolution neural network (CNN) for classification into a convolution layer and realize the prediction of pixel-to-pixel [11]. Then many semantic image segmentation frameworks based the FCN are used in different segmentation tasks. The structure of the FCNs is shown in Fig. 2. There are several different types of layers in FCNs.

Convolution Layer: The convolution is often used to detect features in a picture. And many convolution-based feature extraction operators are proposed, such as edge detection [20] and corner detection [21]. Unlike traditional convolution kernels that are pre-designed, convolution kernels in CNNs are learnable and iteratively updated by the Back Propagation (BP) algorithm based on the training set. In terms of weights sharing, each feature map shares a convolution kernel. And in terms of the feature extraction, each convolution kernel extracts a pattern in each corresponding feature map of the previous layer. Therefore, the layered convolution structure makes the CNN having powerful feature extraction capabilities. As a result, many feature-based methods have been proposed to improve the CNN performance by improving the convolution kernel [22] or the network architecture [23].

ReLU Layer: The ReLU is just a no-linear activation function: $f(x) = \max(0, x)$. This activation function can be regarded as a threshold function. And the output of irrelevant nodes is suppressed to obtain a sparse output.

Pooling Layer: The Pooling layer represents multiple adjacent points in the feature map as a single point by using the maximum or the mean method to greatly reduce the size of the feature map. However, due to the transitivity of features in the network structure of the CNN [24], a point in a small feature map represents a large number of points in the original picture, which makes the information be dispersed in different small feature maps. Meanwhile, the information of location and shape details in the picture are weakened. Because, the image has the characteristics of pixel aggregation and the image information can be expressed in a compressed manner, the Pooling layer is effective in the CNN.

Deconvolution Layer: The Deconvolution layer is used to restore the size of the feature map reduced by the Pooling layer, making them the same size as the original image. It implements a learnable upsampling process by transposing the convolution kernels [25].

Softmax Layer: The Softmax Layer converts the output of the FCN to a form of probability through the softmax function as shown in Eq. (1).

$$p_{x,y}(c = i) = \frac{\exp(h_i(x, y))}{\sum_j \exp(h_j(x, y))} \tag{1}$$

where, $h_i(x, y)$ represents the value of the i-th feature map of the final output of FCN at the position (x, y), and $p_{x,y}(c = i)$ represents the probability of the pixel at (x, y) belonging to category $c = i$. In this paper, the output of the softmax layer is treated as a probability map to represent the probability of each pixel.

From the perspective of probability, the FCN is equivalent to a probability estimator, which is used to estimate the probability of category of the pixel at (x, y). The probability of an image I is expressed by Eq. (2).

$$p_{x,y}(c = i \mid I) \tag{2}$$

The distribution is a multinomial distribution with the experiment number 1, so the probability $O_{x,y}(c)$ of category of the ground truth at each pixel (x, y) is 1. the value of $O_{x,y}(c)$ is 1 when c is the true category, or the value is 0. Therefore,

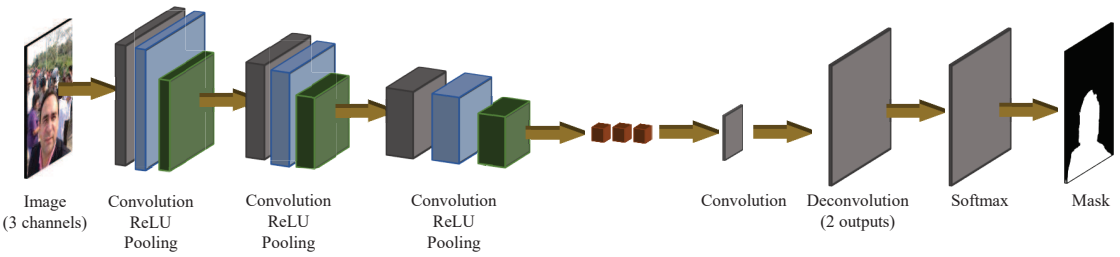


Fig. 2: The structure of FCNs.

the minimizing cross entropy method is used to minimize the difference between the estimated probability and the true probability to train the FCN [26]. And then the loss function is calculated through Eq. (3).

$$\mathcal{L} = \sum_{x,y} \sum_{i} p_{x,y}(c = i \mid I) \log O_{x,y}(c = i \mid I) \tag{3}$$

At the point of receptive fields [27], the output of the FCN at each location in the picture is only related to the receptive field of this location. So the probability model above can be expressed as $p_{x,y}(c = i \mid \mathcal{R}_{x,y})$, and $\mathcal{R}_{x,y}$ represents the receptive field of output at (x, y). Therefore, points with similar receptive fields will be predicted to the same category. Similarly, the similar regions in the original image are also predicted as the same category. But the output of the FCN is easily affected by the similar noise area. Some Segmentation results of FCNs

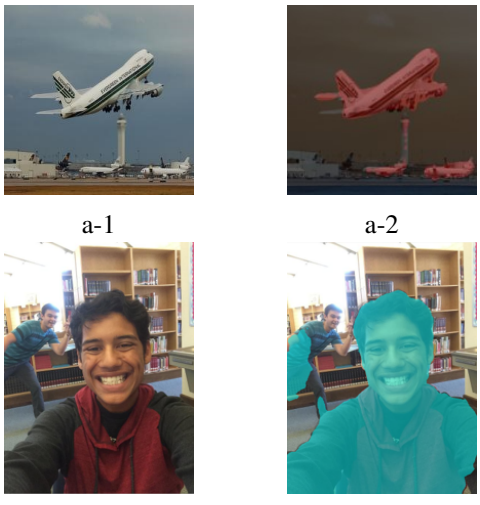


Fig. 3: Some segmentation result of FCNs.

are shown in Fig. 3 (The left is the original image and the right is the result of the segmentation by FCNs). The two pictures a-1 and a-2 in Fig. 3 are the example of the Pascal VOC data set, which is segmented by FCN8s [11]. It can be seen that the white beacon in the center of the picture is incorrectly predicted as an aircraft. The two pictures b-1 and b-2 in Fig. 3 are from a portrait data set, and segmented by the PortraitFCN [19] which is the retraining of FCN8s in the Portrait data set. From these two segmentation results, we can get some shortcomings of FCNs in segmentation tasks.

- First, the segmentation boundary cannot be accurately adapted to the target boundary.
- Second, the segmentation boundary is rough and noisy.
- Finally, the prior information of the target shape is not considered.

Therefore, FCNs combined with Conditional Random Fields (CRF) [15] or Markov Random Field (MRF) [28] are proposed to solve the first and second problems, and achieved a good performance. In order to solve the total three problems at the same time, the FCNs with level set method is proposed in this paper. After training on the training set, FCNs can extract features of the target and learn patterns of the target, and these capabilities are represented in the probability map. Therefore, the probability map preserves the information of the segmentation target based on the pixels in the receptive field. For example, in the Portrait data set, the probability map can represent the probability that each pixel belongs to a person. Although there is a certain probability which is incorrect, it still keeps most of the correct predictions, even the correct patterns information. Based on these properties of the probability map, it is possible to use the method of combining the probability map with the shape priors and the level set method for the semantic segmentation. And then how to get the optimal affine transformation of the standard shape prior using the GAT is introduced in the following subsection.

B. Global Affine Transformation for Shape Prior Corrected

There is only one mean mask of the portrait in the portrait data set [19], which is called the standard shape prior in this paper as shown in Fig. 4. In this subsection, the standard shape

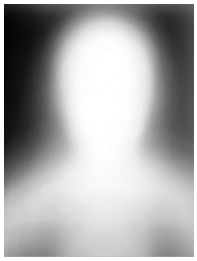


Fig. 4: The standard shape mask

prior is transformed to the best position of the picture, based on the probability map obtained in the previous subsection. The transformation is assumed as the affine transformation which only contains translation, scale, rotation, flip and shear,

and does not change the basic geometry of the original shape. Since the probability map is similar to the real shape, and the optimal affine transformation of the standard shape prior can be obtained based on the probability map.

The GAT is used to obtain the optimal affine transformation. The GAT was first used to find the optimal affine transformation between handwritten characters [29] [30]. It accepts a set of contour coordinates of the shape, so the probability map and the standard shape need to be converted to contours. The sample contour of the probability map and the standard shape prior is shown in Fig. 5. Let the original (probability map) contour set

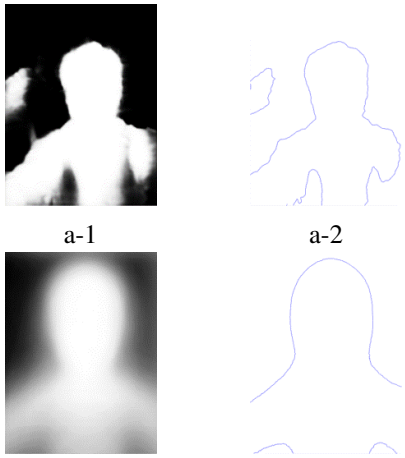


Fig. 5: The sample contour of the probability map (a-1, a-2 of Fig. 3. b-1) and standard shape prior (b-1, b-2).

is S and the target (standard shape prior) contour set is R:

$$\begin{aligned} S &= \{s_1, s_2, \cdots, s_i, \cdots, s_m\} \\ R &= \{r_1, r_2, \cdots, r_j, \cdots, r_n\} \end{aligned}$$

where s_i is the coordinate vector of the i-th point of contour S and r_j is the coordinate vector of the j-th point of contour R. s_i can be transformed into a new coordinate vector $\hat{s}_i$ by a affine transformation:

$$\hat{s}_i = As_i + b \tag{4}$$

where A is a 2×2 matrix, and b is a 2×1 vector. The transformed contour is represented by $\hat{S}$.

$$\hat{S} = \{\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_i, \cdots, \hat{s}_m\}$$

In order to get the optimal transformation, the distance between S and $\hat{S}$ is minimized, and the distance $\mathcal{D}$ is defined as follows:

$$\mathcal{D} = \frac{1}{2}\left[\frac{1}{m}\sum_{i}^{m} \min_{j} \|\hat{s}_i - r_j\|^2 + \frac{1}{n}\sum_{j}^{n} \min_{i} \|\hat{s}_i - r_j\|^2\right] \tag{5}$$

Based on A and b, the distance $\mathcal{D}$ is minimized.

$$A, b = \arg\min_{A,b} \ \mathcal{D} \tag{6}$$

There are two $\min$ operations in Eq. (5), so A and b cannot be directly calculated by Eq. (6). A computable model based on weighted least-squares criterion is proposed to solve this combinatorial optimization problem [29]. The objective function Φ is defined as

$$\Phi = \frac{1}{2}\left[\frac{1}{m}\sum_{i}^{m}\sum_{j}^{n} \mu_{ij}(D)\,\|\hat{s}_i - r_j\|^2 + \frac{1}{n}\sum_{j}^{n}\sum_{i}^{m} \nu_{ji}(D)\,\|\hat{s}_i - r_j\|^2\right] \tag{7}$$

$$\mu_{ij}(D) = \exp\left[-\frac{\|s_i - r_j\|^2 - \min_{k} \|s_i - r_k\|^2}{D}\right]$$

$$\nu_{ji}(D) = \exp\left[-\frac{\|s_i - r_j\|^2 - \min_{k} \|s_k - r_j\|^2}{D}\right]$$

$$D = \frac{1}{2}\left[\frac{1}{m}\sum_{i}^{m} \min_{j} \|s_i - r_j\|^2 + \frac{1}{n}\sum_{j}^{n} \min_{i} \|s_i - r_j\|^2\right] \tag{8}$$

The $\min$ operation is replaced with weighted summations using Gaussian functions of $\mu_{ij}(D)$ and $\nu_{ji}(D)$. The shortest distance between all points in S and R is considered in Eq. (8), so this method is called the "Global" Affine Transformation. And Eq. (8) can be rewritten as follows,

$$\Phi = \frac{1}{2}\sum_{i}^{m}\sum_{j}^{n} \rho_{ij}(D)\,\|\hat{s}_i - r_j\|^2 \tag{9}$$

$$\rho_{ij}(D) = \frac{\mu_{ij}(D)}{m} + \frac{\nu_{ji}(D)}{n}$$

Thus, the minimization of distance Φ can be obtained by solving the following equations:

$$\begin{cases} \frac{\partial \Phi}{\partial A} = & \sum_{i}^{m}\sum_{j}^{n} \rho_{ij}(D) s_i (As_i + b - r_j)^T & = 0 \\ \frac{\partial \Phi}{\partial b} = & \sum_{i}^{m}\sum_{j}^{n} \rho_{ij}(D)(As_i + b - r_j) & = 0 \end{cases} \tag{10}$$

Gaussian elimination can be used to solve Eq. (10), and there are 6 unknown parameters in total. The GAT is used to find the optimal transformation within the neighborhoods of each point, so the final optimal affine transformation can be obtained through iterative methods, and the detailed proof is given by Toru Wakahara [29].

Every point in contour S is disposed by the above affine transformation, and the optimal affine transformation can also be applied in the image [31]. And the transformation matrix is defined as follows:

$$T = \begin{bmatrix} A_{1,1} & A_{2,1} & 0 \\ A_{1,2} & A_{2,2} & 0 \\ b_1 & b_2 & 1 \end{bmatrix}$$

Then, the transformation T can be applied in the standard shape prior image with interpolation to get the shape prior corrected. Some results of the GAT in Portrait data set are shown in Fig. 6. From Fig. 6, we can see that even if the probability map is noisy, the GAT can still fit the shape expressed by the probability map. And the transformed shape prior shape is already very close to the shape of the probability map at 1 iter, so the GAT is effective.

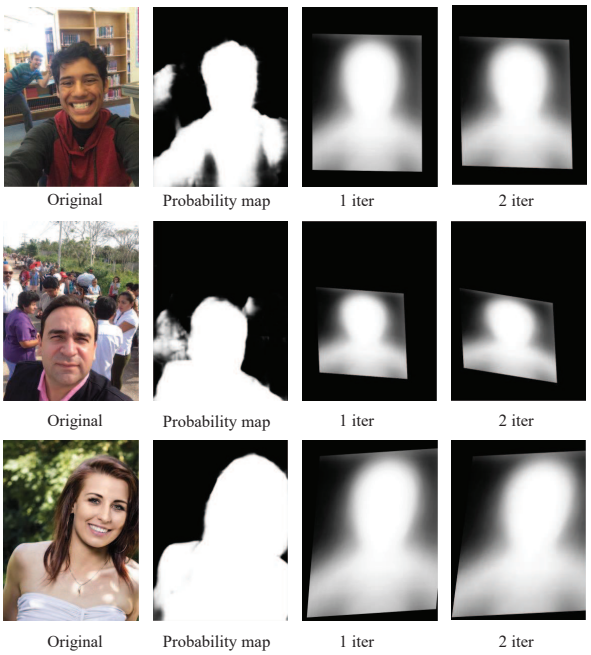


Fig. 6: Some results of the GAT in Portrait data set.

C. Level Set with Deep Prior Method for Image Segmentation

The level set method for the image segmentation is a region-based active contour model, which is a variational method based on the energy minimization to evolve the level set [32]. The level set is denoted by ϕ and the zero level $\phi = 0$ is regarded as the contour of target, and the region of $\phi < 0$ is regarded as the target region. The level set ϕ can be viewed as a potential function that represents the strength of each point in the image. Therefore, the level set ϕ can be treated as a kind of probability density, indicating the probability met by the point [33]. Moreover, the energy objective function of the level set method can be combined with many energy-based probability models to expand its capabilities [34].

The most classic level set method is the *CV* model proposed by Chan and Vese [10], and the energy function of the contour is obtained by Eq. (11).

$$\begin{aligned}\mathcal{F}(C, c1, c2) =& \lambda_1 \int_{\text{outside}(C)} |I(\mathbf{x}) - c_1|^2 \, \mathrm{d}\mathbf{x} \\ &+ \lambda_2 \int_{\text{inside}(C)} |I(\mathbf{x}) - c_2|^2 \, \mathrm{d}\mathbf{x} \\ &+ \nu \, |C|\end{aligned} \tag{11}$$

where outside(C) and inside(C) represent the regions outside and inside the contour C, respectively. $\mathbf{x}$ is a 2 dim vector that represents the image. $|C|$ represents length of the contour C. c_1 and c_2 are the statistics of the pixels outside and inside the contour, respectively. The energy function can be rewritten as the form of the level set function ϕ:

$$\begin{aligned}\mathcal{F}(\phi, c_1, c_2) =& \lambda_1 \int |I(\mathbf{x}) - c_1|^2 \, H(\phi) \mathrm{d}\mathbf{x} \\ &+ \lambda_2 \int |I(\mathbf{x}) - c_2|^2 \, (1 - H(\phi)) \mathrm{d}\mathbf{x} \\ &+ \nu \int |\nabla H(\phi)| \, \mathrm{d}\mathbf{x}\end{aligned} \tag{12}$$

where $H(z)$ is Heaviside function that indicates the regions of outside and inside represented by the level set function.

$$H(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases} \tag{13}$$

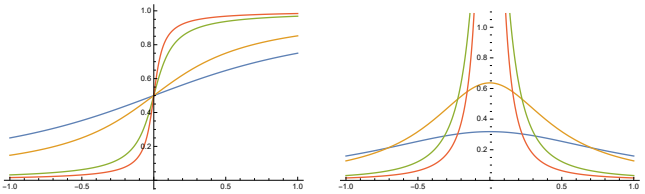
In order to introduce energy-based variational methods, the function H is smoothed to make it possible to get gradients.

$$H_\epsilon(z) = \frac{1}{2}\left[1 + \frac{2}{\pi} \arctan\left(\frac{z}{\epsilon}\right)\right] \tag{14}$$

So the derivative of H can be approximated by the derivative of H_ϵ based on Eq. (15).

$$\delta_\epsilon(z) = H'_\epsilon(z) = \frac{1}{\pi} \frac{\epsilon}{\epsilon^2 + z^2} \tag{15}$$

And the curves of H and δ functions of different ϵ are shown in Fig. 7.



$H_\epsilon(z)$ $\quad$ $\delta_\epsilon(z)$

Fig. 7: The function figures of different $\epsilon \in \{1, 0.5, 0.1, 0.05\}$.

The level set method has a very flexible definition of the energy function, so it can be convenient to combine the original image, the probability map and the corrected shape prior information into an energy function. The energy function is defined as the following.

$$\mathcal{F}(\phi) = \mathcal{E}_{\text{img}} + \mathcal{E}_{\text{shape}} + \mathcal{E}_{\text{edge}} + \mathcal{E}_{\text{reg}} \tag{16}$$

All the items in Eq. (16) will be described in detail as follows.

1) $\mathcal{E}_{img}$: The first item refers to a improved *CV* model proposed by Li [35], which introduces a weight function. And each point uses the weighted statistics of its neighbors as a reference value. And the probability map is added in this item, $\mathcal{E}_{\text{img}}$ is defined as

$$\begin{aligned}\mathcal{E}_{\text{img}} =& \sum_{i=1}^{2} \lambda_i \int P_i(\mathbf{x}) \\ &\cdot \left(\int K_\sigma(\mathbf{x} - \mathbf{y}) \, |I(\mathbf{y}) - f_i(\mathbf{x})|^2 \, M_i(\phi(\mathbf{y})) \mathrm{d}\mathbf{y} \right) \mathrm{d}\mathbf{x}\end{aligned} \tag{17}$$

where $P_1(\mathbf{x})$ and $P_2(\mathbf{x})$ are probability maps and represent the probability of background and target at point $\mathbf{x}$, respectively.

$M_1(\phi) = H(\phi)$ and $M_2(\phi) = 1 - H(\phi)$ represent the regions of outside and inside contour C, respectively. $K(u)$ represents a symmetrical weighted function and is used to weight neighbors of each point. The Gaussian kernel is often chosen as the weighted function,

$$K_\sigma(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{|u|^2}{2\sigma^2}} \tag{18}$$

$f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ represent the weighted statics at point $\mathbf{x}$ of outside and inside. In the iterative process of each step, it is necessary to fix ϕ first, and get f_i at this time. Fix ϕ, $f_i(\mathbf{x})$ can be obtained by minimizing the functional $\mathcal{E}_{\text{img}}$ in Eq. (17) based on the *Euler* equation.

$$f_i(\mathbf{x}) = \frac{\int K_\sigma(\mathbf{x}-\mathbf{y}) M_i(\phi(\mathbf{y})) I(\mathbf{y}) \mathrm{d}\mathbf{y}}{\int K_\sigma(\mathbf{x}-\mathbf{y}) M_i(\phi(\mathbf{y})) \mathrm{d}\mathbf{y}} \tag{19}$$

Since the integral can be converted into the convolution. And Eq. (19) can be written as

$$f_i(\mathbf{x}) = \frac{K_\sigma * [M_i^\epsilon(\phi(\mathbf{x})) I(\mathbf{x})]}{K_\sigma * M_i^\epsilon(\phi(\mathbf{x}))} \tag{20}$$

Then, fix f_1 and f_2, the partial derivative needs to minimize the energy functional $\mathcal{E}_{\text{img}}$.

$$\begin{aligned} \frac{\partial \mathcal{E}_{\text{img}}}{\partial \phi} &= \delta_\epsilon(\phi) \left(\lambda_1 P_1 e_1 - \lambda_2 P_2 e_2\right) \\ e_i(\mathbf{x}) &= \int K_\sigma(\mathbf{y}-\mathbf{x}) \left|I(\mathbf{x}) - f_i(\mathbf{y})\right|^2 \mathrm{d}\mathbf{y} \end{aligned} \tag{21}$$

Similarly, e_i can also be converted into the convolution, and the Eq. (21) needs to be expanded as the following:

$$\begin{aligned} e_i(\mathbf{x}) =& I^2(\mathbf{x}) \cdot [K_\sigma(\mathbf{x}) * \mathbf{1}] \\ & - 2I(\mathbf{x}) \cdot [K_\sigma(\mathbf{x}) * f_i(\mathbf{x})] \\ & + K_\sigma(\mathbf{x}) * f_i^2(\mathbf{x}) \end{aligned} \tag{22}$$

where $\mathbf{1}$ is a all-1 matrix. Since the interval $[-2\sigma, 2\sigma]$ already contains more than 95% in the Gaussian kernel function, and the size of convolution kernel of K_σ can be set to $4\sigma + 1$.

2) $\mathcal{E}_{shape}$: The corrected shape prior guarantees that the final segmentation result is similar to the shape prior. And the energy function is defined as

$$\mathcal{E}_{\text{shape}} = \sum_{i=1}^{2} \pi_i \int S_i(\mathbf{x}) M_i(\phi(\mathbf{x})) \mathrm{d}\mathbf{x} \tag{23}$$

where S_1 is the corrected shape prior, and $S_2 = 1 - S_1$. The energy function is equivalent to calculate the difference between the segmentation shape and the prior shape. Minimizing $\mathcal{E}_{\text{shape}}$ makes the segmentation shape as close as possible to the prior shape, and the partial derivative is calculated by Eq. (24).

$$\frac{\partial \mathcal{E}_{\text{shape}}}{\partial \phi} = \delta_\epsilon(\phi)(\pi_1 S_1 - \pi_2 S_2) \tag{24}$$

3) $\mathcal{E}_{edge}$: As in Eq. (12), this functional energy is used to calculate the length of the segmentation contour, which ensures the segmentation contour is smooth.

$$\begin{aligned} \mathcal{E}_{\text{edge}} &= \nu \int \left|\nabla H_\epsilon(\phi(\mathbf{x}))\right| \mathrm{d}\mathbf{x} \\ &= \nu \int \delta_\epsilon(\phi) |\nabla \phi| \mathrm{d}\mathbf{x} \end{aligned} \tag{25}$$

$$\frac{\partial \mathcal{E}_{\text{edge}}}{\partial \phi} = -\nu \delta_\epsilon(\phi) \nabla \times \frac{\nabla \phi}{|\nabla \phi|} \tag{26}$$

4) $\mathcal{E}_{reg}$: This $\mathcal{E}_{\text{reg}}$ is obtained based on the *Signed Distance Function*, which guarantees the basic shape of the level set method during the iterative process [36], [37].

$$\begin{aligned} \mathcal{E}_{\text{reg}} &= \mu \int \frac{1}{2} \left(|\nabla \phi(\mathbf{x})| - 1\right)^2 \mathrm{d}\mathbf{x} \\ \frac{\partial \mathcal{E}_{\text{reg}}}{\partial \phi} &= -\mu \left\{ \nabla^2 \phi - \nabla \times \frac{\nabla \phi}{|\nabla \phi|} \right\} \end{aligned} \tag{27}$$

Finally, the minimum energy functional $\mathcal{F}$ can be obtained though the steady state solution of the gradient flow equations with Eq. (21), Eq. (24), Eq. (26) and Eq. (27).

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= -\frac{\partial \mathcal{F}}{\partial \phi} \\ &= -\left(\frac{\partial \mathcal{E}_{\text{img}}}{\partial \phi} + \frac{\partial \mathcal{E}_{\text{shape}}}{\partial \phi} + \frac{\partial \mathcal{E}_{\text{edge}}}{\partial \phi} + \frac{\partial \mathcal{E}_{\text{reg}}}{\partial \phi} \right) \end{aligned} \tag{28}$$

The function ϕ is calculated iteratively.

$$\phi^t = \phi^{t-1} + \Delta t \cdot \frac{\partial \phi}{\partial t} \tag{29}$$

The result of the contour evolution process of the proposed level set with the deep prior for the image segmentation is shown in Fig. 10. The level set method has a flexible energy functional form, making it more convenient to integrate more useful information. This proposed level set with the deep prior method makes up their respective disadvantages of FCNs and the level set method.

III. Experiments and Results

A. Data Set

The portrait data set is used for the image stylization, which segment the portrait in the selfie for the style conversion. It is a simple single-target portrait segmentation problem that is applied in the segmentation of complex scenes. Some images

Fig. 8: Some images and ground truth of the Portrait data set.

and ground truth of the portrait data set are shown in Fig. 8. There are 1800 portrait images in total, each of which is automatically scaled and cropped to 600×800, so every image is a standard portrait. The 1800 labeled images data set are split into a 1500 image training data set and a 300 image testing data set by Shen et al. [19]. Because the images in the Portrait data set is labeled with the *Photoshop* quick selection, there is some noise in ground truth as shown in Fig. 9.

Fig. 9: Some noise in the Portrait data set.

B. Evaluation Measure

There is only one kind of the segmentation target in this image segmentation task, and only the regions of the target need to be marked. So we only need to calculate the difference between the output regions of models and the regions of ground truth. The standard metric Interaction-over-Union (IoU) is selected to represent the segmentation error.

$$\text{IoU} = \frac{\text{Area}(\text{Output} \cap \text{Ground Truth})}{\text{Area}(\text{Output} \cup \text{Ground Truth})} \tag{30}$$

Eq. (30) is calculated by dividing the intersection area and the union area of the region of output and ground truth. Finally, the mean IoU of 300 testing images is used to verify the performance of different models.

C. Implementation

All FCNs are trained with Caffe [38], and all parameters are given by Shen et al. [19]. The probability map is obtained from PortraitFCN and PortraitFCNplus, which are trained with the Portrait data set. It can be seen from Fig. 6 that it is already very close to the shape of the probability map at the second iteration, so the number of GAT iterations is set to 2. The level set function ϕ is initialized with the probability map, and extended to intervals $[-200, 200]$. The same parameters are used in all testing images, and the parameters are listed int Table I.

TABLE I: Parameters of the proposed method.

Parameter	Value	Reference
ϵ	0.5	Eq. (14)
σ	4	Eq. (18)
λ_1	20	Eq. (21)
λ_2	20	Eq. (21)
π_1	2*500	Eq. (24)
π_2	2*500	Eq. (24)
ν	0.5*255*255	Eq. (26)
μ	1.0	Eq. (27)
Δt	0.2	Eq. (29)

D. Result and Analysis

In this paper, we mainly compare with PortraitFCN and PortraitFCNplus [19], which is equivalent to an attempt at the image segmentation task with combining of the level set method and FCNs. FCN8s is a CNN structure proposed by Long et al. [11], and trained with the Pascal VOC data set [39]. There are 21 different classes in the Pascal VOC data set, but the Portrait data set used in this paper has 1 class, so only the person class in FCN8s is used. The PortraitFCN is the retrain of FCN8s at Portrait data set. The PortraitFCNplus expands the 3-channel of the original image into 6-channel based on the PortraitFCN, adding the mean mask and normalized x and y. The mean mask is shown in Fig. 4. In this paper, the output of PortraitFCN and PortraitFCNplus are selected as the probability maps respectively to verify the performance of the proposed method.

Finally, the performance comparison of different models at Portrait data set is shown in Table II, and the contour evolution process of proposed level set method with PortraitFCN at Portrait data set is shown in Fig. 10. The picture below in Fig. 10 is the evolution process. The blue curve is the contour by the probability map with PortraitFCN, and the red one is the proposed level set method.

TABLE II: Performance comparison of different models at Portrait data set.

Methods	Mean IoU
FCN(Person Class)	73.09%
PortraitFCN	94.20%
PortraitFCN + Proposed	95.17%
PortraitFCNplus	95.91%
PortraitFCNplus + Proposed	95.74%

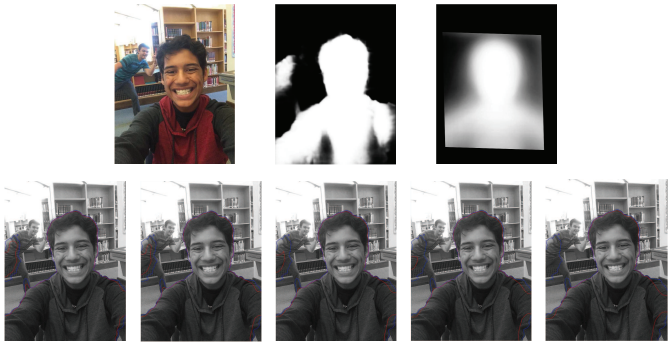


Fig. 10: The contour evolution process of proposed method at Portrait data set.

It can be seen from Table II that the proposed method gets great improvement with PortraitFCN, but some weakening with PortraitFCNplus. As shown in Fig. 6, the output of PortraitFCN have some shortcomings of noisy, non-smooth, imprecise at boundary and no shape prior in Subsection II-A. The proposed method solves the problem of the shape prior, so it is effective with PortraitFCN. However, with PortraitFCNplus, the mean mask has been added in the training process and achieved a great performance at most of the pictures. Because of the imprecise of the corrected shape prior, the original probability map information would be disturbed during the evolution of the level set, which degrades the performance. It can be seen from Fig. 10 that the regions far from the corrected prior shape would be erased during the iterative process. But the region in the lower right corner of the image is considered to have a low probability, so it is also erased.

E. The Results of Different Reference Information

In this subsection, a series of experiments are conducted to verify the availability of different reference information on the segmentation result. An image form the Portrait data set is used to analyze the result. And the image, the probability map from PortraitFCN and the corrected shape prior are shown in Fig. 11.

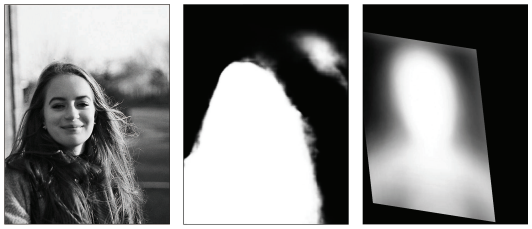


Fig. 11: The image, the probability map and the corrected shape prior for experiments.

1) Experiment 1: The reference information selected in Experiment 1 uses the original image and the probability map information in $\mathcal{E}_{\text{img}}$ like Subsection II-C, so that the regions that are close to each other in pixels and satisfy the probability map are grouped together. And $\mathcal{E}_{\text{shape}}$ just uses the corrected prior shape, which ensures the similarity of the final segmentation and the corrected shape prior. The contour evolution process of Experiment 1 is shown in Fig. 12. From the experiment

Fig. 12: The contour evolution process of Experiment 1.

result, it can be seen that the final segmentation result is more similar to the corrected prior shape. However, the corrected prior is imprecise, so the result of Experiment 1 is inaccurate in the segmentation details.

2) Experiment 2: There are some differences between experiment 1 and experiment 2. $\mathcal{E}_{\text{img}}$ only uses the original image information to capture the boundaries in the original picture.

$$\mathcal{E}_{\text{img}} = \sum_{i=1}^{2} \lambda_i \cdot \int \left(\int K_\sigma(\mathbf{x}-\mathbf{y}) \left| I(\mathbf{y}) - f_i(\mathbf{x}) \right|^2 M_i(\phi(\mathbf{y})) \mathrm{d}\mathbf{y} \right) \mathrm{d}\mathbf{x} \tag{31}$$

And the product of the probability map and the corrected prior shape can be obtained based on $Q_i(\mathbf{x}) = P_i(\mathbf{x}) \cdot Smooth(S_i(\mathbf{x}))$ as shown in Fig. 13. $Smooth(S_i(\mathbf{x}))$ is the average smooth of the corrected prior shape. Then, $\mathcal{E}_{\text{shape}}$ is redefined by Eq. (32).

$$\mathcal{E}_{\text{shape}} = \sum_{i=1}^{2} \pi_i \int Q_i(\mathbf{x}) M_i(\phi(\mathbf{x})) \mathrm{d}\mathbf{x} \tag{32}$$

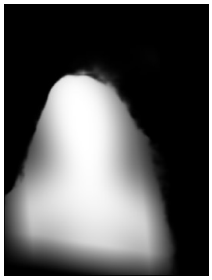


Fig. 13: The product of the probability map and the corrected prior shape.

Experiment 2 uses the intersection of the probability map and the corrected shape prior as the target shape regions. And then the boundary of target is located according to the information of the original image. The contour evolution process of Experiment 2 is shown in Fig. 14. It can be seen from

Fig. 14: The contour evolution process of Experiment 2.

Fig. 14 that the reference information method in Experiment 2 can more accurately locate the boundary of the target.

IV. Discussion

From the experimental results in subsection III-D, when there is enough information, the level set method can achieve a precise segmentation result. However, when the scene of the picture is more complex, the level set method cannot achieve the desired segmentation result. Therefore, some **pattern recognition** methods are proposed to bring more information about the target. For example, the regions of the target, and the probability that each pixel belongs to the target category. In the process of the contour evolution, the level set method relies on the statistic of pixels inside and outside the contour with fixed ϕ at this time, such as the mean, the weighted mean and the probability model about regions [40]. And the level set method is more like an integration method of the information, which uses the energy functional minimization principle to construct a potential function. Moreover, this potential function can be set as the probability function, which is disposed by probabilities and Bayesian methods.

The GAT is used to find the optimal affine transformation of the standard prior shape based on the probability map. But it is based on the contour, and the probability map expresses the regions. Therefore, when the results of FCNs are poor, the GAT will fail to match. The failure result of the GAT is shown in Fig. 15.

Even if the pixels are very similar, the imprecise probability map and the corrected prior shape still have a large effect, causing that similar pixels are split apart. But we can draw on

Fig. 15: The failure result of the GAT.

the idea of superpixels [6]. Most of those methods about the image segmentation can be summarized as how to extract and use the information in images. Therefore the most important issue is to build the dynamic hierarchical structured representation of images.

V. Conclusion

This paper proposes a novel level set segmentation method based on the priors learned by FCNs. The output of the FCNs is treated as a probability map to represent the probability of the target pattern. Based on the pattern prior learned from the training set by FCNs, the inherent prior shape is adjusted for the specific image. The proposed level set with the deep prior method can integrate the information of the original image, the probability map and the corrected prior shape for the image segmentation. Finally, through some experiments based on the Portrait data set, the effectiveness of the proposed method is verified. The experimental results show that compared with the traditional FCNs, the proposed method can obtain more accurate segmentation results.

In our future work, we will improve the GAT by the region-based method to find the local optimal affine transformation. Moreover, we will construct the multi-objective matching method that enables the proposed method to handle more complex scenes and tasks.

Acknowledgment

This research was partly funded by National Natural Science Foundation of China (61673046 and 61603025), the Natural Science Foundation of Beijing, China (4162045) and the National Key Research and Development Program of China (2017YFC1601800).

References

[1] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern recognition*, vol. 26, no. 9, pp. 1277–1294, 1993.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[3] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic imaging*, vol. 13, no. 1, pp. 146–166, 2004.

[4] N. Senthilkumaran and R. Rajesh, "Edge detection techniques for image segmentation–a survey of soft computing approaches," *International journal of recent trends in engineering*, vol. 1, no. 2, pp. 250–254, 2009.

[5] H. T. Nguyen, M. Worring, and R. Van Den Boomgaard, "Watersnakes: Energy-driven watershed segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, pp. 330–342, 2003.

[6] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[7] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.

[8] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[9] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

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

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[12] 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.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[15] 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.

[16] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," *arXiv:1711.10925*, 2017.

[17] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *IEEE CVPR*, 2017.

[18] M. Tang, S. Valipour, Z. Zhang, D. Cobzas, and M. Jagersand, "A deep level set method for image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 126–134.

[19] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, and I. Sachs, "Automatic portrait segmentation for image stylization," in *Computer Graphics Forum*, vol. 35, no. 2. Wiley Online Library, 2016, pp. 93–102.

[20] J. Canny, "A computational approach to edge detection," in *Readings in Computer Vision*. Elsevier, 1987, pp. 184–203.

[21] C. Harris and M. Stephens, "A combined corner and edge detector." in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.

[22] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4961–4970.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 2921–2929.

[25] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2018–2025.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[27] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[28] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1377–1385.

[29] T. Wakahara and K. Odaka, "Adaptive normalization of handwritten characters using global/local affine transformation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 12, pp. 1332–1341, 1998.

[30] Z. Xiaona and J. Lianwen, "Hierarchical chinese character database based on global affine transformation," in *Control Conference, 2007. CCC 2007. Chinese*. IEEE, 2007, pp. 584–588.

[31] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.
[32] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: A level set approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 2, pp. 158–175, 1995.
[33] D. Cremers, F. R. Schmidt, and F. Barthel, "Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–6.
[34] F. Chen, H. Yu, R. Hu, and X. Zeng, "Deep learning shape priors for object segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1870–1877.
[35] C. Li, C.-Y. Kao, J. C. Gore, and Z. Ding, "Minimization of region-scalable fitting energy for image segmentation," *IEEE transactions on image processing*, vol. 17, no. 10, pp. 1940–1949, 2008.
[36] C. Li, C. Xu, C. Gui, and M. D. Fox, "Level set evolution without re-initialization: a new variational formulation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 430–436.
[37] ——, "Distance regularized level set evolution and its application to image segmentation," *IEEE transactions on image processing*, vol. 19, no. 12, pp. 3243–3254, 2010.
[38] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.
[39] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.
[40] P. Lin, C. Zheng, Y. Yang, F. Zhang, and X. Yan, "A probability model-based level set method for biomedical image segmentation," *Journal of X-Ray Science and Technology*, vol. 13, no. 3, pp. 117–127, 2005.