Level Set based Shape Prior and Deep Learning for Image Segmentation

Shuheng Zhang
E-mail: zsh_o@qq.com

***Abstract*—Abstract**

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

I. Introduction

THIS is Introduction. Describe the data set, the features of the data set, and explain why the data set is used, combining the features of the level set method.

II. Proposed Method

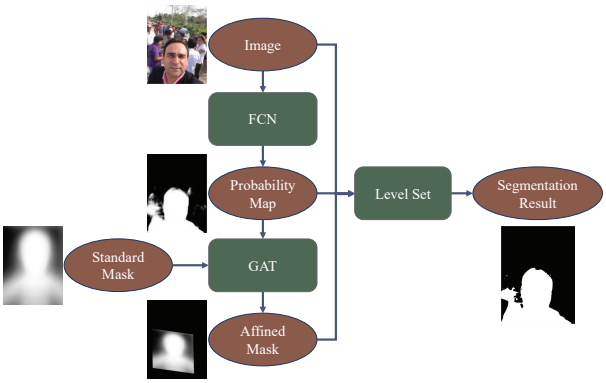


Fig. 1: The flow diagram of the proposed method

In order to improve the performance of FCNs and make Level Set method be able to complete objective segmentation of complex sense, this paper proposed a method to integrate the advantages of Deep Learning and Level Set method. An overview of the proposed method is shown in Fig. 1. In this method, the output of FCNs is taken as a 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. So, an optimal affine transformation of the standard shape mask (shape prior) of the image can be obtained based on the "probability" shape with GAT (Global Affine Transformation) method. Finally, image, probability map and affine mask are used as the input of Level Set method to implement image segmentation. The following subsections show the details of this proposed method.

Thanks

A. Deep Learning for the Probability Map

Deep Learning as a end-to-end method has achieved many excellent performance on semantic segmentation tasks. In this task, each pixel in the picture needs to be classified, so the traditional network structure cannot be applied. Long et al. proposed the Fully Convolutional Network (FCN) that replaced the fully connected layer in the traditional convolution neural network for classification into a convolution layer, and realized the prediction of pixel-to-pixel [1]. Then many semantic image segmentation frameworks based FCN were proposed for different segmentation tasks. The structure of the FCNs is shown in Fig. 2. There are several different types of layers in FCNs:

Convolution Layer Convolution is often used to detect features in a picture, and many convolution-based feature extraction operators are proposed, such as edge detection [2], corner detection [3], etc. Unlike traditional convolution kernels that are pre-designed, convolution kernels in Convolution Neural Networks are learnable, and iteratively update 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 feature extraction, each convolution kernel extracts a pattern in each corresponding feature map of the previous layer. Therefore, the layered convolution structure makes convolution neural networks (CNN) have powerful feature extraction capabilities. As a result, many feature-based method have been proposed to improve CNN performance by improving the convolution kernel [4] or network architecture [5].

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, the output of irrelevant nodes is suppressed, and it is easier to obtain a sparse output.

Pooling Layer The Pooling layer represents multiple adjacent points in the feature map as a single point, using maximum, mean, etc., greatly reducing the size of the feature map. However, due to the transitivity of features in the network structure of CNN [6], 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, thus weakening the information of location, shape details and so on in the picture. However, the image data has the characteristics of pixel aggregation, so the Pooling layer is effective in the image task.

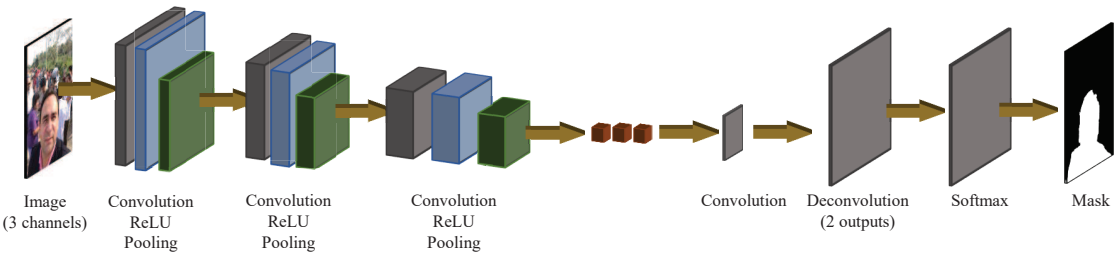


Fig. 2: The structure of FCNs

Deconvolution Layer The Deconvolution layer is used to restore the size of the feature map reduced by the Pooling layer, makes them the same size as the original image. It implements a learnable upsampling process by transposing the convolution kernels [7].

Softmax Layer The Softmax Layer converts the output of the FCN to a form of probability by the function:

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

$h_i(x,y)$: the value of the i-th feature map of the final output of FCN at the position (x,y), $p_{x,y}(c=i)$: the probability of the pixel at (x,y) belongs 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 pixel at (x,y). Given a image I, the form is as follows,

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

The distribution is a multinomial distribution with experiment number 1, so that, the probability of category of the ground truth at each pixel (x,y) is 1. It can be expressed as $O_{x,y}(c)$, only 1 when c is the true category, the rest is 0. Therefore, the method of minimizing cross entropy is used to minimize the difference between the estimated probability and the true probability to train the FCN [8]. So, the loss function is bellow:

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

At the point of receptive fields [9], the output of 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 $p_{x,y}(c=i \mid \mathcal{R}_{x,y})$, $\mathcal{R}_{x,y}$ 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 to be the same category. So that the output of the FCN is easily affected by the similar noise area. Some Segmentation result is of FCNs is show in Fig. 3. The above two pictures are an example of the Pascal VOC data set, which is segmented by FCN8s [1]. We can see that the white beacon in the center of the picture is incorrectly predicted to be an aircraft. The two pictures below are from a portrait data set, and segmented by PortraitFCN [10] which is the retraining of FCN8s in the Portrait data set.

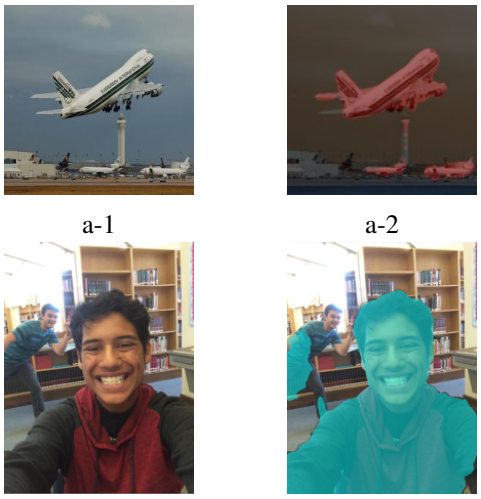


Fig. 3: Some segmentation result of FCNs
The left is the original image and the right is the result of the segmentation by FCNs.

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.

FCNs combined with Conditional Random Fields (CRF) [11] or Markov Random Field (MRF) [12] were 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, this paper combines FCNs with Level Set method. It should be noted that this method is based on the fact that FCNs as feature extraction and pattern recognition can achieve an effective performance. In other words, after training on the training set, FCNs extracted features of the target and learned 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 that it 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 probability map with shape priors and Level Set method for semantic segmentation. Next, how to get the optimal affine transformation of standard shape prior using Global Affine Transformation (GAT) is introduced in the following subsection.

B. Global Affine Transformation for Shape Prior Corrected

There is only one mean mask of portrait in the portrait data set [10], which is called the standard shape prior in this paper, and it is show in Fig. 4. In this subsection, the

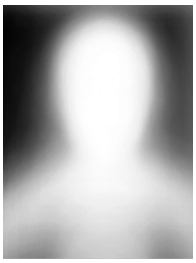


Fig. 4: The standard shape mask

standard shape prior is transformed to the best position of the picture, based on the probability map obtained in the previous subsection. In this paper, it is assumed that the transformation is affine transformation. The affine transformation 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, the optimal affine transformation of the standard shape prior can be obtained from the probability map.

In this paper, The Global Affine Transformation (GAT) is used to accomplish this function. The GAT was first used to find the optimal affine transformation between handwritten characters [13] [14]. 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 show in Fig. 5. Let the original (probability map) contour set is S and 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:

$$\hat{s}_i = As_i + b \tag{1}$$

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

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

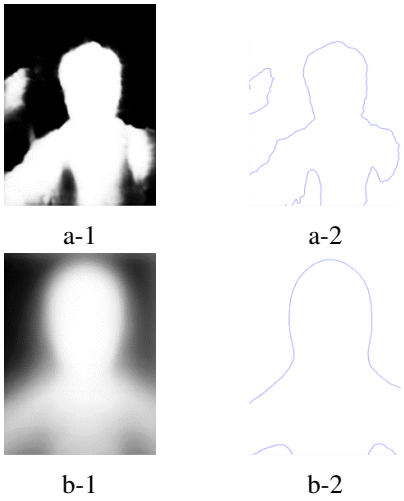


Fig. 5: The sample contour of the probability map (a-1, a-2 of Fig. 3. b-1) and standard shape prior (b-1, b-2).

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

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

We need to find A and b to minimize the distance $\mathcal{D}$,

$$A, b = \underset{A,b}{arg\,min} \;\; \mathcal{D} \tag{3}$$

There are two $\min$ operations in Eq. 2, so A and b cannot be directly calculated by Eq. 3. A computable model based on weighted least-squares criterion is proposed to solve this combinatorial optimization problem [13]. The objective function Φ is defined as

$$\begin{aligned} \Phi = \frac{1}{2}\Bigg[\frac{1}{m}\sum_{i}^{m}\sum_{j}^{n} \mu_{ij}(D) \left\|\hat{s}_i - r_j\right\|^2 \\ + \frac{1}{n}\sum_{j}^{n}\sum_{i}^{m} \nu_{ji}(D) \left\|\hat{s}_i - r_j\right\|^2\Bigg] \end{aligned} \tag{4}$$

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

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

$$D = \frac{1}{2}\left[\frac{1}{m}\sum_{i}^{m} \underset{j}{min} \left\|s_i - r_j\right\|^2 + \frac{1}{n}\sum_{j}^{n} \underset{i}{min} \left\|s_i - r_j\right\|^2\right] \tag{5}$$

The $\min$ operation is replaced with weighted summations using Gaussian functions of $\mu_{ij}(D)$ and $\nu_{ji}(D)$. Eq. 5 considers the shortest distance between all points of S and R, so it is called

the "Global" Affine Transformation. Eq. 5 can be rewritten as follows:

$$\Phi = \frac{1}{2}\sum_{i}^{m}\sum_{j}^{n}\rho_{ij}(D)\left\|\hat{s}_i - r_j\right\|^2 \tag{6}$$

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

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

$$\frac{\partial\Phi}{\partial A} = \sum_{i}^{m}\sum_{j}^{n}\rho_{ij}(D)s_i(As_i + b - r_j)^T = 0 \tag{7}$$

$$\frac{\partial\Phi}{\partial b} = \sum_{i}^{m}\sum_{j}^{n}\rho_{ij}(D)(As_i + b - r_j) = 0 \tag{8}$$

Gaussian elimination can be used to solve this equations, and there are 6 unknown parameters in total. The GAT can be regarded as finding the optimal transformation within the neighborhoods of each point, so final optimal affine transformation can be obtained through iterative methods, and the detailed proof is given by Toru Wakahara [13].

The above affine transformation is for every point in contour S, and the optimal affine transformation can also be applied to image [15]. And the transformation matrix is defined as follows:

$$T = \begin{bmatrix} A_{1,1} & A_{2,1} & 0 \\ A_{1,2} & A_{2,2} & 0 \\ b_1 & b_2 & 1 \end{bmatrix} \tag{9}$$

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

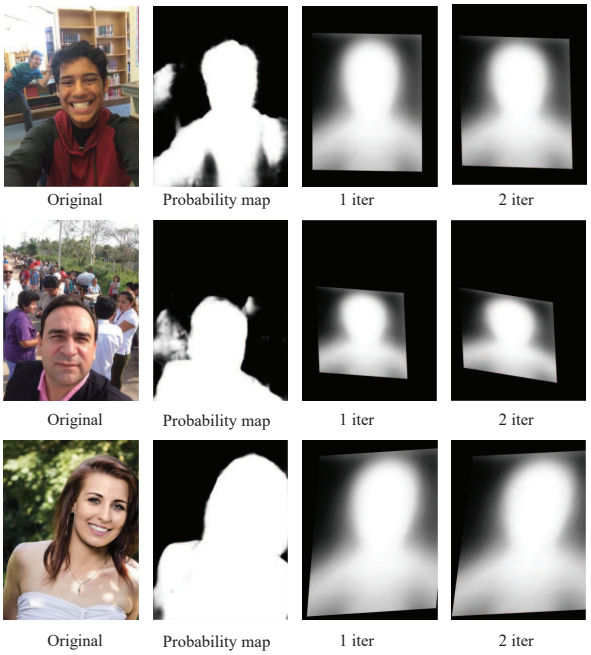


Fig. 6: Some results of the GAT in Portrait data set.

At this point, the probability map and the corrected shape prior have all been obtained. In the next subsection, the method based on Level Set will be introduced how to implement image segmentation based on original image, the probability map and the corrected shape prior.

C. Level Set Method for Image Segmentation

Level Set method for image segmentation is a region-based active contour models. It is a variational method based on energy minimization to evolve the level set [16]. The level set is denoted by ϕ and the zero level $\phi = 0$ is regarded as the contour of target, 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 that the point meets [17]. Moreover, its energy objective function can be combined with many energy-based probability models to expand its capabilities [18].

The most classic level set method is the *CV* model proposed by Chan and Vese [19], and its energy function of contour is as follows:

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

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. This energy function can be rewritten as the form of level set function ϕ:

$$\begin{aligned}\mathcal{F}(\phi, c_1, c_2) =& \lambda_1 \int |I(\mathbf{x}) - c_1|^2 H(\phi)\mathrm{d}\mathbf{x} \\ &+ \lambda_2 \int |I(\mathbf{x}) - c_2|^2 (1 - H(\phi))\mathrm{d}\mathbf{x} \\ &+ \nu \int |\nabla H(\phi)| \,\mathrm{d}\mathbf{x}\end{aligned} \tag{10}$$

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

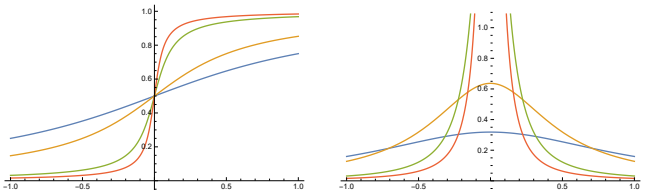
$$H(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases}$$

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

$$H_\epsilon(z) = \frac{1}{2}\left[1 + \frac{2}{\pi}\arctan\left(\frac{z}{\epsilon}\right)\right] \tag{11}$$

And the derivative of H_ϵ is

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



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

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

The figure of H and δ function of different ϵ is shown in Fig. 7.

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

$$\mathcal{F}(\phi) = \mathcal{E}_{\text{img}} + \mathcal{E}_{\text{shape}} + \mathcal{E}_{\text{edge}} + \mathcal{E}_{\text{reg}} \tag{12}$$

All the items in Eq. 12 will be described in detail below.

1) $\mathcal{E}_{img}$: The first item refers to a improved *CV* model proposed by Li [20], which introduces a weight function such that 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 follow:

$$\begin{aligned}\mathcal{E}_{\text{img}} = &\sum_{i=1}^{2} \lambda_i \int P_i(\mathbf{x}) \\ &\cdot \left(\int K_\sigma(\mathbf{x}-\mathbf{y}) \left| I(\mathbf{y}) - f_i(\mathbf{x}) \right|^2 M_i(\phi(\mathbf{y})) \mathrm{d}\mathbf{y} \right) \mathrm{d}\mathbf{x}\end{aligned} \tag{13}$$

Where $P_1(\mathbf{x})$, $P_2(\mathbf{x})$ 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 proposed by Li, and is used to weight neighbors of each point. The Gaussian kernel is often chosen sa the weighted function:

$$K_\sigma(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{|u|^2}{2\sigma^2}} \tag{14}$$

$f_1(\mathbf{x})$ 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 fist, and get f_i at this time. Fix ϕ, minimize functional $\mathcal{E}_{\text{img}}$ in Eq. 13 based on the *Euler* equation, we can obtain $f_i(\mathbf{x})$:

$$f_i(\mathbf{x}) = \frac{\int K_\sigma(\mathbf{x}-\mathbf{y}) M_i(\phi(\mathbf{y})) I(\mathbf{y}) \mathrm{d}\mathbf{y}}{\int K_\sigma(\mathbf{x}-\mathbf{y}) M_i(\phi(\mathbf{y})) \mathrm{d}\mathbf{y}} \tag{15}$$

Since the integral can be converted into convolution, Eq. 15 can be written as

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

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

$$\begin{aligned}\frac{\partial \mathcal{E}_{\text{img}}}{\partial \phi} &= \delta_\epsilon(\phi) \left(\lambda_1 P_1 e_1 - \lambda_2 P_2 e_2 \right) \\ e_i(\mathbf{x}) &= \int K_\sigma(\mathbf{y}-\mathbf{x}) \left| I(\mathbf{x}) - f_i(\mathbf{y}) \right|^2 \mathrm{d}\mathbf{y}\end{aligned} \tag{16}$$

Similarly, e_i can also be converted into convolution, but first the formula needs to be expanded

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

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, the size of convolution kernel of K_σ can be set to $4\sigma + 1$.

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

$$\mathcal{E}_{\text{shape}} = \sum_{i=1}^{2} \pi_i \int S_i(\mathbf{x}) M_i(\phi(\mathbf{x})) \mathrm{d}\mathbf{x} \tag{17}$$

Where S_1 is the corrected shape prior, and $S_2 = 1 - S_1$. The energy function is equivalent to calculating 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

$$\frac{\partial \mathcal{E}_{\text{shape}}}{\partial \phi} = \delta_\epsilon(\phi)(\pi_1 S_1 - \pi_2 S_2) \tag{18}$$

3) $\mathcal{E}_{edge}$: As with Eq. 10, this functional energy is used to calculate the length of segmentation contour, which ensures the segmentation contour is smooth.

$$\begin{aligned}\mathcal{E}_{\text{edge}} &= \nu \int \left| \nabla H_\epsilon(\phi(\mathbf{x})) \right| \mathrm{d}\mathbf{x} \\ &= \nu \int \delta_\epsilon(\phi) |\nabla \phi| \mathrm{d}\mathbf{x}\end{aligned} \tag{19}$$

$$\frac{\partial \mathcal{E}_{\text{edge}}}{\partial \phi} = -\nu \delta_\epsilon(\phi) \nabla \times \frac{\nabla \phi}{|\nabla \phi|} \tag{20}$$

4) $\mathcal{E}_{reg}$: This item is a regularization item based *Signed Distance Function*, it guarantees the basic shape of the level set method during iteration [21] [22].

$$\begin{aligned}\mathcal{E}_{\text{reg}} &= \mu \int \frac{1}{2} \left(|\nabla \phi(\mathbf{x})| - 1 \right)^2 \mathrm{d}\mathbf{x} \\ \frac{\partial \mathcal{E}_{\text{reg}}}{\partial \phi} &= -\mu \left\{ \nabla^2 \phi - \nabla \times \frac{\nabla \phi}{|\nabla \phi|} \right\}\end{aligned} \tag{21}$$

Finally, the minimum energy functional $\mathcal{F}$ can be find though the steady state solution of the gradient flow equation. And combining Eq. 16, Eq. 18, Eq. 20 and Eq. 21, we can get

$$\begin{aligned}\frac{\partial \phi}{\partial t} &= -\frac{\partial \mathcal{F}}{\partial \phi} \\ &= -\left(\frac{\partial \mathcal{E}_{\text{img}}}{\partial \phi} + \frac{\partial \mathcal{E}_{\text{shape}}}{\partial \phi} + \frac{\partial \mathcal{E}_{\text{edge}}}{\partial \phi} + \frac{\partial \mathcal{E}_{\text{reg}}}{\partial \phi} \right)\end{aligned} \tag{22}$$

The function ϕ is calculated iteratively

$$\phi^t = \phi^{t-1} + \Delta t \cdot \frac{\partial \phi}{\partial t} \tag{23}$$

A result of contour evolution process of the proposed Level Set method is shown in Fig. 10. The Level Set method has a flexible energy functional form, making it more convenient to integrate more useful information. In this paper, we integrate the original image, the probability map and the corrected shape prior information, and finally use the Level Set method for image segmentation. This method make up their respective disadvantages to some extent.

III. Experiments and Results

A. Data Set

Portrait data set is used in this paper. The portrait data set is used for image stylization, which is to segment the portrait in the selfie for style conversion. It is a simple single-target portrait segmentation problem that is suitable for experiments with the level set method to try to apply it the segmentation of complex scenes. Some images and ground truth of portrait

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

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 was split into a 1500 image training data set and a 300 image testing data set by Shen et al. [10], and this division is also used in this paper. Because the images in Portrait data set is labeled with Photoshop quick selection, there is some noise in ground truth, as shown in Fig. 9.

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

The traditional Level Set method is suitable for image of simple scenes, and the proposed method in this paper combines it with FCNs and shape prior to make it possible to be applied to image segmentation of complex scenes.

B. Evaluation Measure

There is only one kind of segmentation target in this image segmentation task, only the regions of 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.

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

It is calculated by dividing the intersection area and the union area of the region of model output and ground truth . Finally, the mean IoU of 300 testing images is used to verify the performance of models.

C. Implementation

All FCNs are trained with Caffe [23], and with all parameters given by Shen et al. [10]. The probability map is obtained from PortraitFCN and PortraitFCNplus which are trained with 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. [1]

TABLE I: Parameters of the proposed method.

Parameter	Value	Reference
ϵ	0.5	Eq. 11
σ	4	Eq. 14
λ_1	20	Eq. 16
λ_2	20	Eq. 16
π_1	2*500	Eq. 18
π_2	2*500	Eq. 18
ν	0.5*255*255	Eq. 20
μ	1.0	Eq. 21
Δt	0.2	Eq. 23

D. Result and Analysis

In this paper, we mainly compare with Shen's method, which is equivalent to an attempt at image segmentation task with the combination of Level Set method and FCNs. FCN8s is a CNN structure proposed by J Long et al. [1], and trained with the Pascal VOC data set [24]. There are 21 different classes in the Pascal VOC data set, but the Portrait data set used in this paper has noly 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

[1]The codes are available at https://github.com/zsh965866221/LevelSet-ShapePrior-DeepLearning

Mean Mask and Normalized x and y. The Mean Mask is shown is Fig. 4. In this paper, we select the output of PortraitFCN and PortraitFCNplus as the probability maps respectively to verify the performance of the proposed Level Set method. Finally, the performance comparison of different models at Portrait data set is show in Table II, and the contour evolution process of proposed Level Set method at Portrait data set is shown in Fig. 10.

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%

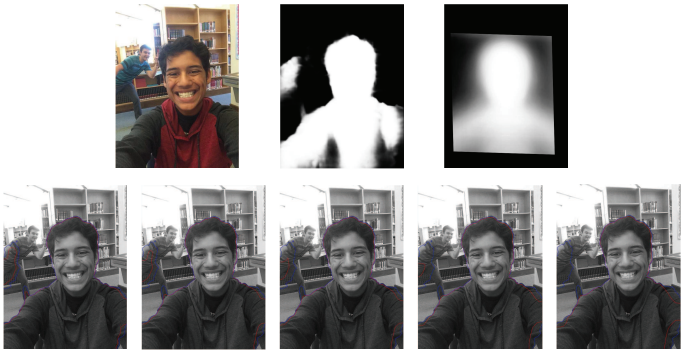


Fig. 10: The contour evolution process of proposed Level Set method at Portrait data set. The picture below is the evolution process. the blue curve is the contour by the probability map, and the red is the proposed Level Set method.

We can see from Table II that, the proposed method has some improvement with PortraitFCN, but some weakening with PortraitFCNplus. As shown in Fig. 6, the output of PortraitFCN have some shortcomings list in Item. II-A, such as noisy, non-smooth, imprecise at boundary and no shape prior. The proposed method mainly to solve the problem of shape prior, so it is effective with PortraitFCN. However, with PortraitFCNplus, the mean mask has been added in the training process, and it has 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, that degrades the performance. We can see from Fig. 10, 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 is conducted to verify the effect of different reference information on the segmentation result. We select an image form Portrait data set to explain, and the image, the probability map from PortraitFCN and the corrected shape prior are shown in Fig. 11.

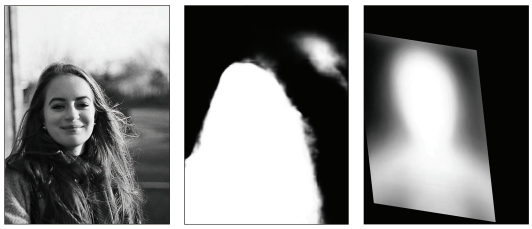


Fig. 11: The image, the probability map and the corrected shape prior for experiments.

1) Experiment 1*:* The reference information selected in Experiment 1 is just like the Subsection II-C above. It uses the original image and the probability map information in $\mathcal{E}_{\text{img}}$, 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, it 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

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

the experiment result, we can see that the final segmentation result is more similar to the corrected prior shape.

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

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

And $\mathcal{E}_{\text{shape}}$ uses the product of the probability map and the corrected prior shape $Q_i(\mathbf{x}) = P_i(\mathbf{x}) \cdot Smooth(S_i(\mathbf{x}))$, $Smooth(S_i(\mathbf{x}))$ is the average smooth of the corrected prior shape. The result of product is shown in Fig. 13. Now, $\mathcal{E}_{\text{shape}}$

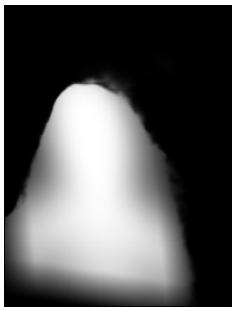


Fig. 13: The product of the probability map and the corrected prior shape.

is redefined as follows

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

Experiment 2 is equivalent to using 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

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

that this method works better in this image, and it can more accurately locate the boundary of the target.

IV. Discussion

From the experimental results in subsection III-D, we can see that 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, and some **pattern recognition** methods are needed to bring more information about the target, for example, where is the target regions, and the probability that each pixel belongs to the target category. In the process of contour evolution, it relies on the statistic of pixels inside and outside the contour with fixed ϕ at this time, such as mean [19], weighted mean [20], and probability model about regions [25]. The Level set method is more like an integration of information, using energy functional minimization principle to construct a potential function based on those information. And this potential function can have a higher meaning, such as probability, so that we can add probabilities and Bayesian methods.

In this paper, 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, so when the result of FCNs are poor, the GAT will fail to match. The failure result of the GAT is shown in Fig. 15.

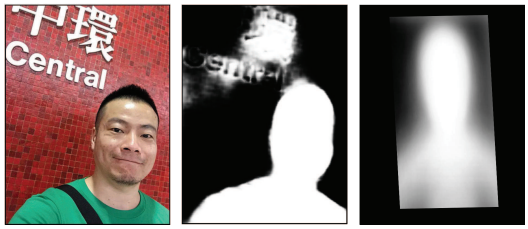


Fig. 15: The failure result of the GAT.

Even if the pixels are very similar, the imprecise probability map and corrected prior shape still have a large effect, causing similar pixels are split apart. But we can draw on the idea of superpixels [26]. Most of the methods can be summarized as how to extract and use the information in images, so the most important issue is to build the dynamic hierarchial structured representation of images.

V. Conclusion and Future Work

Conclusion and Future work

Acknowledgment

Acknowledgment

References

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[16] 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.

[17] 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.

[18] 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.

[19] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.
[20] 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.
[21] 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.
[22] ——, "Distance regularized level set evolution and its application to image segmentation," *IEEE transactions on image processing*, vol. 19, no. 12, pp. 3243–3254, 2010.
[23] 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.
[24] 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.
[25] 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.
[26] 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.