Applied Intelligence
https://doi.org/10.1007/s10489-020-02114-3

Adaptive fusion with multi-scale features for interactive image segmentation

Zongyuan Ding[1] · Tao Wang[1] · Quansen Sun[1] · Hongyuan Wang[2]

Accepted: 2 December 2020

Abstract
Multi-scale features are usually utilized to improve the performance of interactive image segmentation, however, they have varying leverages over the result of segmentation, for example, thinner segmentation results could be achieved by pixel-level features, but sensitive to image noise, and superpixel-level features could provide the semantic perception of the object, but easily lead to over-segmentations. Therefore, we propose an interactive image segmentation algorithm by adaptive fusion with multi-scale features (AFMSF). It intends on combining the multi-scale information adaptively for the segmentation via learning the influence coefficients of multi-scale features. First, multi-scale superpixel layers are generated by controlling the size of superpixels. Based on features of this multi-scale information, the similarity matrices and label priors with pixel-superpixel levels are then obtained. A fusion with diffusion strategy is designed to build the energy function by combining these multi-cues. Finally, the influence coefficient of each scale and the labeling are updated with each other until convergence. The algorithm we proposed is robust to diverse circumstances of objects, the experimental results on public interactive image segmentation datasets Graz, LHI, and MSRC validate the superior performance of the proposed method.

Keywords Interactive image segmentation · Semi-supervised learning · Multi-scale features · Fusion with diffusion

1 Introduction

As an important task in computer vision, image segmentation divides images into regions with a robust internal consistency based on the low-level features (such as grayscale, color, texture, shape, etc.) [1]. The quality of the segmentation has a direct impact on image analysis, recognition, and other high-level image analysis [2]. Image segmentation approaches can be categorized into full supervised methods [3–5], semi-supervised methods [6–10], and unsupervised methods [11, 12]. The deep learning based algorithm, a typical full supervised method, has a strong ability to learn high-level semantic features of images, and many state-of-the-art works emerge in recent years [3–5]. However, the high-quality performance of these approaches generally depends on plentiful images, and it is trivial to label the whole training set. Besides, users may have distinct desired targets for the same image, so the parameters of the deep learning model need to be finetuned to meet the demands of specific users, which is time-consuming. Though no human efforts are required in the unsupervised methods, their results generally have a shortage of accuracy and generalization since no label prior is provided. To tackle this dilemma, semi-supervised (interactive) methods have been developed, in which only a small amount of user interaction is required to guide the better segmentation and produce the user desired targets.

Many interactive segmentation algorithms have been developed, generally categorized to pixel-based methods [7, 9, 13, 14], superpixel-based methods [15, 16], and pixel-superpixel combination methods [17, 18]. Typical pixel-based approaches include GraphCut [7], GrabCut [13], and Random Walk [9]. GraphCut utilizes statistical histograms

✉ Tao Wang
wangtaoatnjust@163.com

✉ Quansen Sun
sunquansen@njust.edu.cn

Zongyuan Ding
dzyha2011@163.com

Hongyuan Wang
hywang@cczu.edu.cn

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, No.200, Xiao Ling Wei Street, Nanjing 210094, China

[2] School of Information Science and Engineering, Changzhou University, No.1, Ge Hu Zhong Lu, Changzhou 213164, Jiangsu, China

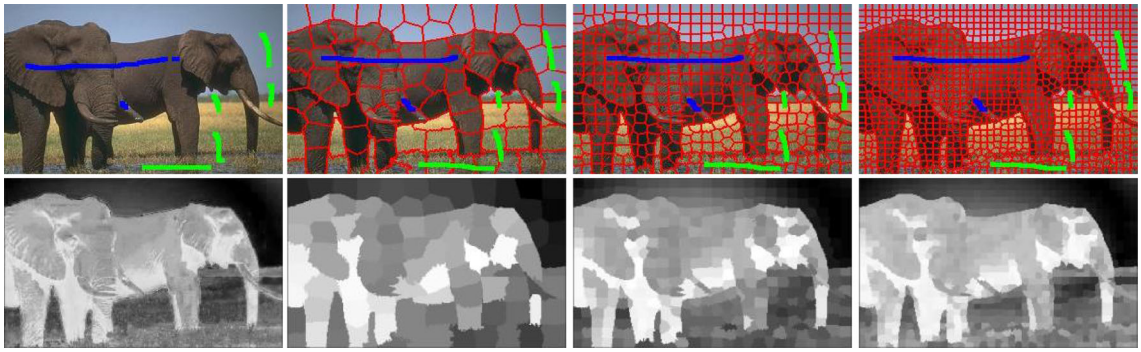
Published online: 12 January 2021

to measure the distances between pixels and labels. The energy function is built with unary and pairwise potentials for considering the regional and boundary properties of the segmentation. Solving the energy function is equivalent to minimizing a specific s/t graph [7]. GrabCut is an iterative expansion of graph cut, and the user needs to specify a rough bounding box to extract the target. Instead of the statistical histograms, GrabCut utilizes the Gaussian mixture model (GMM) to calculate the label prior, and the GMM parameters are updated in every iterative step. In Random Walk model, the probabilities that each pixel reaches to the seeds can be computed, and the final segmentation can be obtained by calculating the maximal probability. Since pixel-level information is too local to capture more abundant image cues, it is hard for pixel-based methods to extract consistent targets, and their results generally contain many small isolated points. Furthermore, they also cannot work well when the object and background have semblable appearance. Apart from utilizing pixels, some superpixel-based methods [15, 16] have been developed, which restricts pixels within the same superpixel sharing a label. Compared with pixels, superpixels can better promote connectivity of segmentations. However, superpixels are not always consistent with the image boundaries [18], which easily leads to over-segmentation results. To solve the aforementioned issues, pixel-superpixel combination methods have been proposed [17, 18], in which both the geometrical adjacency and the regional connectivity constraints can be imposed for the segmentation. Besides, multiple superpixels produced with different unsupervised algorithms [11, 19] are utilized to improve the segmentation performance. Nevertheless, they still have the following two drawbacks: first, generated multiple superpixels generally lack scale variability, some superpixels in different layers are overlapping. Furthermore, the scale of each superpixel is difficult to control, while the superpixel generated algorithm(simple linear iterative clustering, SLIC) [20], which generates superpixel by clustering the pixels in a certain patch of the image, used in this work contains rich scale information. It violates the purpose of using multiple superpixels to obtain image information as rich as possible. Second, the multi-layer information (including the pixel-layer and multiple superpixel-layers) is treated equally in these methods. However, as shown in Fig. 1, information at different scales may play different roles for the segmentation. For example, as shown in Fig. 1a, d, finer-scale features can preserve image details more easily, but lack of semantics, which easily leads to under-segmentations. Stronger "objectness" can be produced by coarser-scale features (Fig. 1b-c), but easy to lose details, which often leads to over-segmentations. Therefore, it is a key issue to adaptively fuse multi-scale information for the segmentation.

The Euclidean distance is generally utilized to compute the pairwise similarities [7, 13, 17]. Whereas the shape of the data manifold is complex, and the intrinsic manifold structure is easily omitted in the Euclidean space. Many diffusion-based algorithms are proposed to solve the above issues [21–25]. The local affinity can be gradually propagated in the iterative diffusion process. Graph Transduction [25] replaces the Euclidean distance with the geodesic path to estimate the shape manifold. Self-diffusion [23] conducts a diffusion process by spreading the similarity coefficients within the intrinsic manifold of data. Tensor graph diffusion(TPG) [22] constructs a novel graph via the tensor product of the initial graph with itself. In image segmentation, the popular Random Walk [9] can also be regarded as the diffusion approach. Recently, many interactive Random Walk extended methods have been proposed [26–28]. Laplacian coordinates (LC) [27] minimizes the mean distances and keeps label anisotropic spread. Sub-Markov Random Walk (SMRW) [26] improves the segmentation performance with introduced label prior, particularly for thin and elongated objects. Normalized Random Walk (NRW) [28] designs a degree-aware term to measure the node centrality of every neighborhood and weighted each neighborhood's importance to the diffusion process. However, only single scale information is diffused in the above approaches, which is not sufficient to characterize the abundant image structures. Therefore, fusion with multi-scale information should be explored to further boost the segmentation performance.

Inspired by the binary affinity matrix fusion with diffusion in image retrieval [29–31], this paper proposes a multi-scale information-based unary label adaptive fusion with diffusion algorithm for interactive image segmentation. Figure 2 illustrates the workflow of the proposed algorithm. As described above, both coarse-level and fine-level features are essential for the segmentation. Therefore, the motivation of this paper is trying to adaptively fuse multi-scale information from coarse to fine. To keep scale variability, the simple linear iterative clustering (SLIC) algorithm [20] is first utilized to produce multi-layer superpixels. Coarse-to-fine features can be obtained by uniformly controlling the number of regions. Then the relationships, such as similarity matrix and label prior, between image elements (pixels or superpixels) and labels can be extracted from each scale layer (shown in Fig. 2). A fusion with diffusion strategy is developed to build the energy function via adaptively fusing these multi-layer relationships. The labeling and influence coefficient are iteratively updated until convergence, which leads to an optimal combination scheme for the segmentation.

The detailed contributions of our work are concluded as: (1) multi-scale (pixel level and multi-superpixel levels) information are explored for segmentation, (2) fusion with diffusion (multiple affinity matrices and label prior) framework is constructed for the interactive image segmentation, (3) diffusion optimization strategy is designed to iteratively update the labeling and the influence coefficient of each scale information.



(a) Pixel layer (b) Superpixel layer-100 (c) Superpixel layer-500 (d) Superpixel layer-1000

Fig. 1 The priors produced by different layers(the first row shows the seeds on the different layers, blue lines represent foreground and green lines represent background). **a** the prior produced by pixel layer **b**-**d** the priors produced by superpixel layer with scale 100, 500, 1000. **a** Pixel layer. **b** Superpixel layer-100. **c** Superpixel layer-500. **d** Superpixel-1000.

The rest of this paper is organized as: In Section 2, we will give a detailed introduction and solution method of the proposed algorithm. In Section 3, we will present comprehensive experimental comparisons and analyses on several datasets. Section 4 will give the conclusion of this paper.

2 Proposed method

We propose an interactive image segmentation algorithm using adaptive fusion with multi-scale features (AFMSF). As described in Fig. 2, a few seeds for the foreground (blue line) and background (green line) need to be marked first, and then the SLIC algorithm is utilized to produce multi-superpixel layers. For each scale layer, the label prior can be effectively measured based on the seed information, and the affinity matrix can be computed based on the pairwise similarities between elements (pixels or superpixels). A fusion with diffusion strategy is finally utilized to construct the energy function by combining the above multi-scale relationships adaptively.

2.1 Label diffusion

Given a label set L and an image pixel set P, where N represents the number of the pixels of the image, the goal of image segmentation is to assign one label $l \in L$ to each pixel $p_i \in P$. We need to calculate the similarities among all pairwise pixels to propagate the relationship between labeled pixels and unlabeled pixels. To unify the similarity matrices of all scale layers into an energy function, we use the features of the superpixel corresponding to the pixel p_i and p_j to calculate the similarity $w_{ij}^{(h)}$ of the h^{th} superpixel layer. The similarity coefficient $w_{ij}^{(h)}$ of the h^{th} scale layer between pixels p_i and p_j always defined as following Gaussian function:

$$w_{ij}^{(h)} = \begin{cases} \exp\left(-\delta \left\| c_i^{(h)} - c_j^{(h)} \right\|_2\right), if\, p_j \in \Psi_i \\ 0, otherwise \end{cases} \quad (1)$$

where $c_i^{(h)}$ represents the feature at pixel p_i if $h = 1$, or represents the feature of superpixel corresponding to the pixel p_i if $2 \le h \le H$, H is the number of different scale layers, δ is a controlling constant, and Ψ_i represents the neighborhood set of p_i.

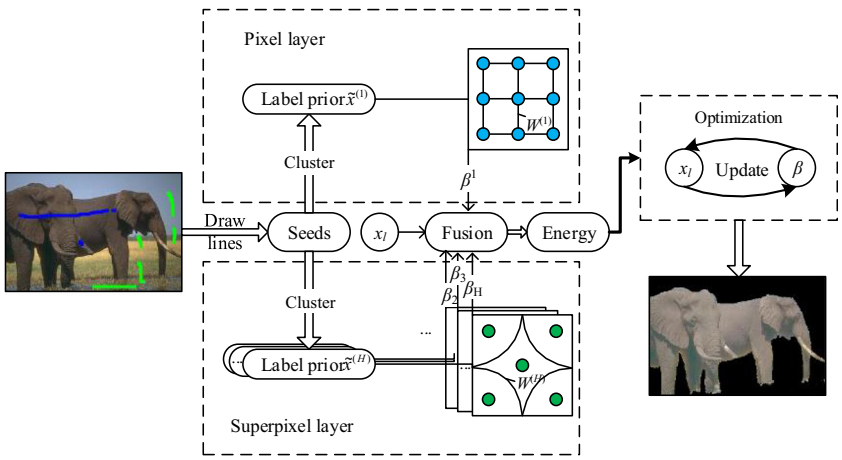


Fig. 2 The workflow of our method(blue lines denote the foreground and green lines denote background). $\tilde{x}^{(h)}$ represents the label prior of the h^{th} layer, $W^{(h)}$ is the similarity matrix of the h^{th} layer, β_h is the weight coefficient of the h^{th} layer, x_l is the probability vector of all pixels belongs to label l. Fusion means to weightly fuse similarity matrix and inintial probability vector. Energy represents the energy function of our proposed method, as summaried in Eq. (9)

Seeds contain the prior knowledge of each label. The label prior can be calculated from the seed information. First, we exploit the k-means algorithm [32] to cluster the seeds of different scale layers and K cluster centers $S = \left[s_k^{(h)}\right]_{K\times 1}$ of the h^{th} scale layer can be obtained, where K denotes the number of clusters. The label prior with the h^{th} scale-layer feature can be obtained as:

$$\widetilde{x}^{(h)} = \max_{k=1,2,\ldots,K}\left(\exp\left(-dis\left(c^{(h)}, s_k^{(h)}\right)\right)\right) \tag{2}$$

where $dis\left(c^{(h)}, s_k^{(h)}\right) \in [0,1]$ is the normalized Euclidean distance between the feature vector $c^{(h)} = \left[c_i^{(h)}\right]_{N\times 1}$ and cluster center $s_k^{(h)}$. It needs to be noted that the label prior is obtained by finding the maximum value by column in Eq. (2).

The conventional method estimates the likelihood of the pixel being part of the label $l \in L$ by minimizing the energy function [24, 33]:

$$\begin{aligned} \mathrm{F}(x_l) &= \frac{1}{2}\sum_{i,j=1}^{N} w_{ij}\left(x_{il}-x_{jl}\right)^2 + \lambda\sum_{i=1}^{N} d_{ii}\left(x_{il}-\widetilde{x}_i\right)^2 \\ &= x_l^T L x_l + \lambda\left(x_l-\widetilde{x}\right)^T D\left(x_l-\widetilde{x}\right) \end{aligned} \tag{3}$$

where $x_l = [x_{il}]_{N\times 1}$ is the probability vector of all pixels belonging to label l, and x_{il} stands for the probability of pixel x_i being part of label l. λ is a balance factor that adjusts the significance of two terms. The $L = D\text{-}W$ is Laplace matrix of pixel layer, $D = diag([d_{11}, \ldots, d_{NN}])$, and $d_{ii} = \sum_{j=1}^{N} w_{ij}$, where W is the similarity matrix of the pixel layer, i.e., $W = W^{(1)} = \left[w_{ij}^{(1)}\right]_{N\times N}$, and $\widetilde{x}$ means the label prior of the pixel layer, i.e., $\widetilde{x} = \widetilde{x}^{(1)}$.

The second row of the Eq. (3) is the matrix form of the first row. The first term of Eq. (3) constricts the relationships of the boundary, which relieves the weak boundary problem [9]. And the latter term of the Eq. (3) is the regional term that makes the segmentation result has strong internal consistency.

Through minimizing the energy function in the Eq. (3), it has the closed-form solution [24, 33]:

$$x_l = (1-\alpha)(I-\alpha P)^{-1}\widetilde{x} \tag{4}$$

where I is an identity matrix, $P = D^{-1}W$, $\alpha = 1/(\lambda+1)$.

2.2 Label diffusion by average fusion with multi-scale features

The solution of the Eq. (4) is equivalent to the diffusion process [24, 33]:

$$x_l^{(t)} = \alpha P x_l^{(t-1)} + (1-\alpha)\widetilde{x} \tag{5}$$

where t is the iterative step, when $t \to \infty$, the diffusion process has the same form as the Eq. (4).

Pixel-based methods lack regional connectivity so that the segmentation may contain some isolated points, while superpixel-based methods have better regional connectivity. Therefore, fusion with multi-scale features (pixel-level and superpixel-level) can effectively boost the result of segmentation. There are two average fusion strategies: Naïve early fusion and Naïve late fusion.

With the idea of naïve early fusion, the label priors of different layers need to be fused first, and then the fusion label prior is propagated to other unlabeled pixels. The label probabilities can be calculated as:

$$x_l = (1-\alpha)(I-\alpha P)^{-1}\frac{1}{H}\sum_{h=1}^{H}\widetilde{x}^{(h)} \tag{6}$$

In contrast, under the idea of naïve late fusion, label priors of different layers are propagated to other unlabeled pixels, and then label probabilities are merged. So the label probabilities are:

$$x_l = \frac{1}{H}\sum_{h=1}^{H}(1-\alpha)(I-\alpha P)^{-1}\widetilde{x}^{(h)} \tag{7}$$

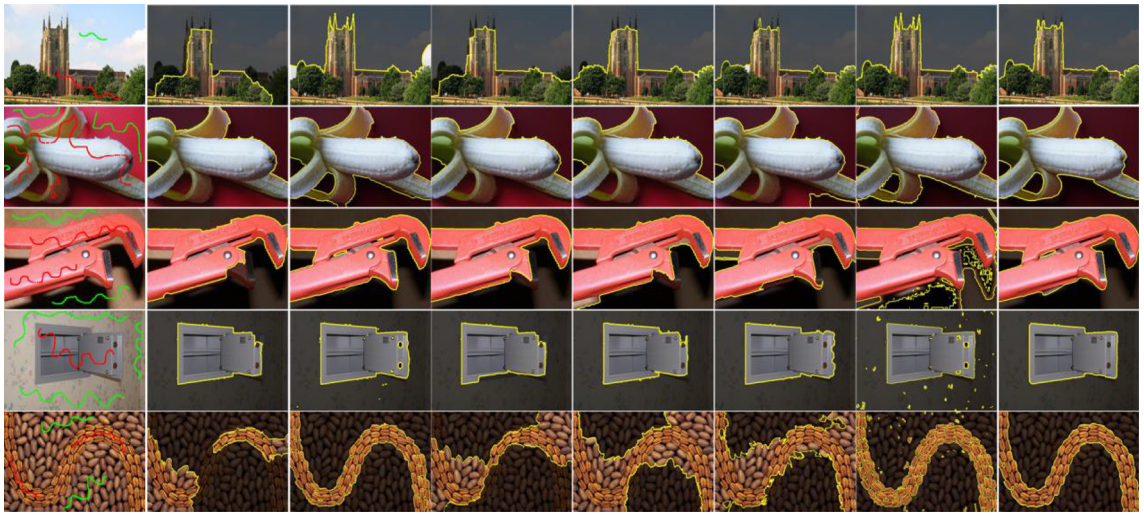
2.3 Label diffusion by adaptive fusion with multi-scale features

Though better performance is obtained to some extent by average fusion multi-scale features, however, as illustrated in Fig. 1, the contribution of each layer is different, so it is inappropriate to treat each layer equally. To solve the above issue, we adaptively fuse multi-scale features for image segmentation. Based on the Eq. (3), we assign different weight to each scale layer and modify the energy function to:

$$\begin{aligned} &\mathrm{F}(x_l,\beta) = \frac{1}{2}\sum_{h=1}^{H}\beta_h\sum_{i,j=1}^{N} w_{ij}^{(h)}\left(x_{il}-x_{jl}\right)^2 + \lambda\sum_{h=1}^{H}\beta_h\sum_{i=1}^{N} d_{ii}^{(h)}\left(x_{il}-\widetilde{x}_i^{(h)}\right) + \frac{1}{2}\gamma\|\beta\|_2^2 \\ &s.t. 0\le\beta_h\le 1,\ \sum_{h=1}^{H}\beta_h = 1 \end{aligned} \tag{8}$$

where β_h is the weight coefficient of the h^{th} scale layer, $d_{ii}^{(h)} = \sum_{j=1}^{N} w_{ij}^{(h)}$, $\gamma > 0$ is a regularization factor.

For the regional term of the Eq. (8), if the i^{th} element (pixel or superpixel) has a strong correlation with other elements (pixels or superpixels) in eight-neighborhood, the $d_{ii}^{(h)}$ is large, the distance between the label probability vector x_l and the label prior $\widetilde{x}^{(h)}$ should be small, which makes the labels have durable internal consistency. To avoid the solution prefers to one layer while omits the influence of other layers, the regularization term $\frac{1}{2}\gamma\|\beta\|_2^2$ is used for the weight coefficients



(a) Original (b) RW (c) SMRW (d) LC (e) NRW (f) TPG (g) GrabCut (h) AFMSF

Fig. 3 Segmentation results of compared methods on Graz dataset. **a** Test images with seeds (red: foreground, green: background); **b**-**h** segmentation results of RW, SMRW, LC, NRW, TPG, GrabCut, and our method. **a** Original. **b** RW. **c** SMRW. **d** LC. **e** NRW. **f** TPG. **g** GrabCut. **h** AFMSF.

$\beta = [\beta_h]_{H\times 1}$ to be solved. Besides, we need to constrict the sum of weight coefficients to be one, and any weight should be not greater than one and not less than zero. The energy function can be transformed into matrix form as the following equation:

$$\begin{aligned} \mathrm{F}(x_l,\beta) &= \sum_{h=1}^{H}\beta_h x_l^{\mathrm{T}} L^{(h)} x_l + \lambda \sum_{h=1}^{H}\beta_h \left(x_l - \widetilde{x}^{(h)}\right)^{T} D^{(h)}\left(x_l - \widetilde{x}^{(h)}\right) + \frac{1}{2}\gamma\|\beta\|_2^2 \\ s.t. 0 \le \beta_h \le 1, &\sum_{h=1}^{H}\beta_h = 1 \end{aligned} \tag{9}$$

Compared with the average fusion with diffusion that only uses the similarity matrix of the pixel layer, we fuse the similarity matrix of the pixel layer and superpixel layers with multiple scales, which can improve the performance of segmentation for different images.

Noted: the energy function in the Eq. (9) is similar to the energy function described in [31]. Though the proposed fusion strategy is inspired by this work [31] (also described in the Introduction section), the difference compared with [31] can be described as: first, from the category of tasks, our work attempts to solve the task of image segmentation, instead of the image retrieval [31]; second, from the model motivation, this paper focuses on combining different scale information (from pixel-level to multi-scale superpixel-level) to improve the segmentation performance, and the fusion with diffusion strategy is a tool for adaptive fusion, third, from the model construction, this paper focuses on the unary label diffusion with the similarity matrix and label prior (different scale) fusion instead of the binary affinity diffusion [31].

2.4 Optimization of the energy function

For our proposed method, there are two variables to optimize, it is non-convex for these two variables, but it is convex for one of them by fixing another one variable. So we can decompose the energy problem into two sub-problems for optimization.

Optimize x_l by fixing β:

As described above, because the solution is a vector, not a tensor, we can get a closed-form solution for x_l. The variable β is viewed as a constant, so the derivative is independent on β, we can remove the third term and simplify the energy function as:

$$\mathrm{F}(x_l) = \sum_{h=1}^{H}\beta_h x_l^{\mathrm{T}} L^{(h)} x_l + \lambda \sum_{h=1}^{H}\beta_h \left(x_l - \widetilde{x}^{(h)}\right)^{T} D^{(h)}\left(x_l - \widetilde{x}^{(h)}\right) \tag{10}$$

It is a convex function regarding variable x_l, by taking the derivative of x_l, we can get:

$$\frac{\partial \mathrm{F}(x_l)}{\partial x_l} = \left(\sum_{h=1}^{H}\beta_h L^{(h)} + \lambda \sum_{h=1}^{H}\beta_h D^{(h)}\right)x_l - \lambda \sum_{h=1}^{H}\beta_h D^{(h)}\widetilde{x}^{(h)} \tag{11}$$

Setting the Eq. (11) to 0, we have:

$$x_l = \left(\sum_{h=1}^{H}\beta_h L^{(h)} + \lambda \sum_{h=1}^{H}\beta_h D^{(h)}\right)^{-1}\left(\lambda \sum_{h=1}^{H}\beta_h D^{(h)}\widetilde{x}^{(h)}\right) \tag{12}$$

which can be simplified as:

$$x_l = (1-\alpha)\left(I - \alpha\overline{P}\right)^{-1}\overline{x}_{fusion} \tag{13}$$

where, $\overline{P} = \left(\sum_{h=1}^{H} \beta_h D^{(h)}\right)^{-1} \left(\sum_{h=1}^{H} \beta_h W^{(h)}\right)$, $\bar{x}_{fusion} = \left(\sum_{h=1}^{H} \beta_h D^{(h)}\right)^{-1} \sum_{h=1}^{H} \beta_h D^{(h)} \tilde{x}^{(h)}$. The $\bar{x}_{fusion}$ is tantamount to the regularized fusion with prior label probabilities of different layers.

Because the construction of similarity matrix W in each layer only considers the relationship of eight-neighborhood, so the matrix P is sparse, using the '\' operation of Matlab can accelerate inverse operation significantly because of QR decomposition application.

Optimize β by fixing x_l:

By fixing x_l, the energy function can be transformed as:

$$\begin{aligned} &\mathrm{F}(\beta) = \sum_{h=1}^{H} \beta_h M^{(h)} + \frac{1}{2}\gamma \|\beta\|_2^2 \\ &s.t. 0 \le \beta_h \le 1, \sum_{h=1}^{H} \beta_h = 1 \end{aligned} \tag{14}$$

we solve it by using coordinate descent. Through fixing other components, we optimize two components β_i and β_j iteratively. The equality constraint $\sum_{h=1}^{H} \beta_h = 1$ ought to be taken into consideration first, thus the optimal value of these two components can be derived as:

$$\begin{cases} \beta_i^* = \dfrac{\gamma(\beta_i + \beta_j) + \left(M^{(j)} - M^{(i)}\right)}{2\gamma} \\ \beta_j^* = \beta_i + \beta_j - \beta_i^* \end{cases} \tag{15}$$

The detailed derivation process of Eq. (15) is shown in the appendix. To guarantee the inequality constraints $0 \le \beta_h \le 1$, we clip the value of β_i^* and β_j^* which violates the inequality constraint using the following rules:

$$\begin{cases} \beta_i^* = 0, if \gamma(\beta_i + \beta_j) + \left(M^{(j)} - M^{(i)}\right) < 0 \\ \beta_j^* = 0, if \beta_i + \beta_j - \beta_i^* < 0 \end{cases} \tag{16}$$

The optimization strategy above is Sequential Minimal Optimization (SMO) [34] algorithm, we decompose the energy function into two sub-problems by fixing one variable and optimizing another variable, until convergence. The convergence of the optimization of energy function can be guaranteed as described in [31]. The pseudocode of the adaptive label fusion with diffusion is presented in Algorithm 1. The initial value of β_h is $1/H$. Notice that all layers contain pixel-layer and superpixel-layer, i.e., $H \ge 2$.

Algorithm 1: Adaptive Fusion with Multi-Scale Features

Input:

the number of scale-layers H, λ, γ

Output:

The probability vectors x_l of foreground or background

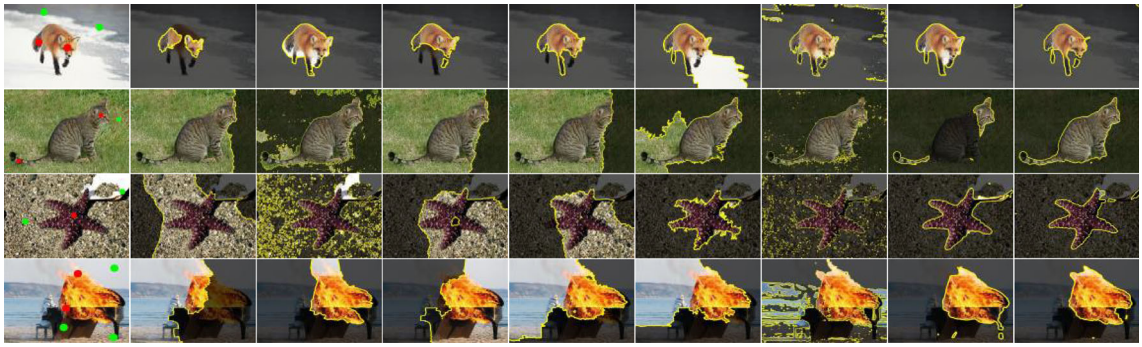
1. Calculate similarity matrices $\{W^{(h)}\}_{h=1}^{H}$ using Eq. (1)
2. Calculate priors $\{\tilde{x}^{(h)}\}_{h=1}^{H}$ using Eq. (2)
3. Initialize the weight $\beta_h = 1/H$
4. **repeat**
5. Compute x_l using Eq. (13)
6. Compute β using Eq. (15) and Eq. (16)
7. **until** *convergence*
8. **return** x_l

Table 1 The mIoU and VoI values of comparison methods on the Graz dataset

Method	RW	SMRW	LC	NRW	TPG	GrabCut	AFMSF
mIoU(%)	70.28	77.31	73.36	78.36	72.31	71.46	79.61
VoI	0.672	0.511	0.611	0.554	0.558	0.576	0.502

3 Experiments

We gave a thorough qualitative and quantitative experiment comparison of several popular interactive segmentation algorithms. The public Graz dataset [35], LHI dataset [36], and MSRC dataset [13] are employed in our work. The compared



(a) Original (b) RW (c) SMRW (d) LC (e) NRW (f) TPG (g) GrabCut (h) BRS (i) AFMSF

Fig. 4 Segmentation results of compared methods on the LHI dataset. **a** Test images with three kinds of seeds (red: foreground, green: background); **b-i** segmentation results of RW, SMRW, LC, NRW, TPG, GrabCut, BRS and Our method

algorithms are Random Walk (RW) [9], Sub-Markov Random Walk(SMRW) [26], Laplacian Coordinates(LC) [27], Normalized Random Walk(NRW) [28], Tensor graph diffusion(TPG) [22] and GrabCut [13]. Besides, to evaluate the efficacy of the proposed adaptive diffusion strategy, we also compared our method with naïve early fusion and naïve late fusion diffusion methods.

We fixed the parameters $\alpha = 0.99$, $\gamma = 100$, $H = 6$ through cross-validation (Section 3.5 introduces the specific ways of selecting these parameters), the simple linear iterative clustering (SLIC) algorithm (as describing in introduction) [20] is utilized to produce superpixels, the number of superpixel layers is set as six which corresponds to 1 K, 3 K, 5 K, 7 K, and 9 K superpixels, respectively (1 K = 1000, and so on), and LAB color space feature is selected as the feature of image in this paper.

For quantitative evaluation, we chose variation of information (VoI) [37], ranges in [0,∞), to evaluate the segmentation results, which stands for the information content of each region and how much information one region supplies to the other. It represents a more precise segmentation with a smaller value. Except for VoI, the intersection over union index (IoU), also known as the Jaccard index, was selected in this paper. Unlike the accuracy, the IoU can take into account the misclassified pixels. The higher the IoU is, the better the segmentation results fit the target, rather than covering a large area but not fitting the target. For the output segmentation P and the ground-truth mask T, IoU is calculated as:

$$\mathrm{IoU} = \frac{|\mathcal{P} \cap \mathcal{T}|}{|\mathcal{P} \cup \mathcal{T}|}$$

Table 2 The mIoU and VoI values of comparison methods on the LHI dataset

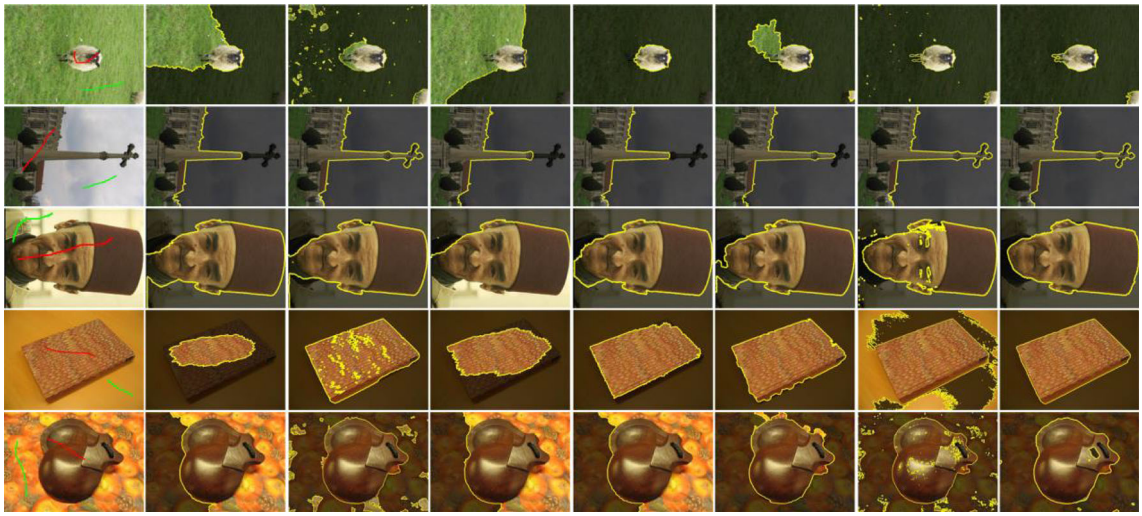
Method	RW	SMRW	LC	NRW	TPG	GrabCut	AFMSF
mIoU (%)	37.97	56.47	48.96	47.57	52.82	49.73	62.70
VoI	1.039	0.929	1.302	0.962	1.024	1.054	0.795

where ∩ means intersection of two sets, and ∪ means union of two sets. Therefore, IoU (or Jaccard index) is obtained by the quotient of the intersection and union of the predicted segmentation and the groundtruth.

3.1 Results on Graz dataset

Graz dataset [35] is especially collected to test the interactive image segmentation algorithms, and it consists of 262 seed and ground-truth pairs from 158 natural images. Images with binary labels (foreground/background) are chosen to evaluate the algorithms. Figure 3 shows the experimental results of the methods to be compared, where Fig. 3a illustrates the test images with scribbles and Fig. 3b-h show the qualitative segmentation results of RW [9], SMRW [26], LC [27], NRW [28], TPG [22], GrabCut [13] and the proposed method, respectively. With limited scribbles, our method can achieve better results than the other six methods. Notice that, as shown in the 3rd-5th rows, the results of GrabCut contain some isolated points, this is owing to reducing the cues of foreground and background caused by modifying the interactions from bounding box to scribble. The colors of the target and background of the fifth image are similar, and our method can segment the object well while RW algorithm has weak boundary problem due to the lacking of superpixel features. However, as shown in the first row, the many slender regions are not well segmented in the proposed method due to the influence of multi-superpixel features, this is because multi-layer superpixel fusion participates in segmentation, weakening the effect of pixel layer on segmentation. To overcome the aforementioned limitation, deep features can be introduced to segment the object in future works.

For the quantitative comparison, Table 1 shows the mean IoU (mIoU) and VoI of the different algorithms. The best value is emphasized in bold, it can be concluded that our method achieves better performance (the highest mIoU and smallest VoI) than other algorithms, which verifies the superiority of the proposed adaptive fusion with diffusion model.



(a) Original (b) RW (c) SMRW (d) LC (e) NRW (f) TPG (g) GrabCut (h) AFMSF

Fig. 5 Segmentation results of compared methods on the MSRC dataset. **a** Test images with seeds (red: foreground, green: background); **b-h** segmentation results of RW, SMRW, LC, NRW, TPG, GrabCut, and our method

3.2 Results on LHI dataset

LHI dataset [36] is also particularly built to evaluate the interactive segmentation algorithms. Different from Graz dataset, three kinds of user annotations are provided in LHI, and we conducted experiments based on seeds of click points to test the algorithm robustness to less interaction information (It needs to be noted that the seed of click point is strengthened into a circle with a radius of 10). Figure 4 lists the comparison results of different methods, and we added a set of segmentation results of the BRS (Backpropagating Refinement Scheme) algorithm [38], which is a deep learning segmentation based on click point interaction. Consistent with the results in the Graz dataset, it is clear to see that the results of GrabCut also lack internal consistency for each image. By comparison, our method is robust to less interaction information. Although there are few little miss-segmented patches of our method in the test images, on the whole, better segmentation results are obtained. From the segmentation results in the last two rows, the proposed method still achieves satisfactory results of images with more complex backgrounds owing to the strategy of adaptive fusion with multi-scale features. For the results of the BRS, we can find that the result can achieve better boundary alignment, although some segmentation results are not satisfying. Note the image of the cat, the result of the BRS algorithm is trapped in the localization when less interaction information is supplied, but our method can relieve the problem of localization.

Table 2 summarizes the mIoU and VoI values of the comparison algorithms on the LHI dataset. Because the BRS algorithm, weak in segmentation by supplying the seeds at once, is designed for the rectification mechanism, we did not show the quantitative comparison for the BRS and other algorithms. It can be concluded that the proposed algorithm acquires the highest mIoU and lowest VoI compared with other methods (at least 10% of the mIoU improvement), indicating that even a small amount of interaction of the proposed algorithm can achieve satisfactory segmentations.

3.3 Results on MSRC dataset

The popular MSRC dataset [13] contains 50 test images with weak boundary problems. The qualitative experiment results are illustrated in Fig. 5, and Fig. 5a represents the original test images with interactive marks and Fig. 5b-h represent the segmentation results of comparison methods and the proposed algorithm. The segmentation results of our method fit the object boundary well and have better regional consistency. From

Table 3 The mIoU and VoI values of comparison methods on the MSRC dataset

Method	RW	SMRW	LC	NRW	TPG	GrabCut	AFMSF
mIoU(%)	51.89	58.88	52.01	58.21	66.99	66.96	73.71
VoI	0.965	0.838	1.017	0.9	0.71	0.689	0.543

Table 4 The mIoU(%) of our algorithm and degraded algorithms on three datasets

Method	Graz	LHI	MSRC
Without fusion	76.47	58.73	69.93
Naïve early fusion	77.95	61.47	72.39
Naïve late fusion	77.95	61.47	72.39
AFMSF	79.61	62.70	73.71

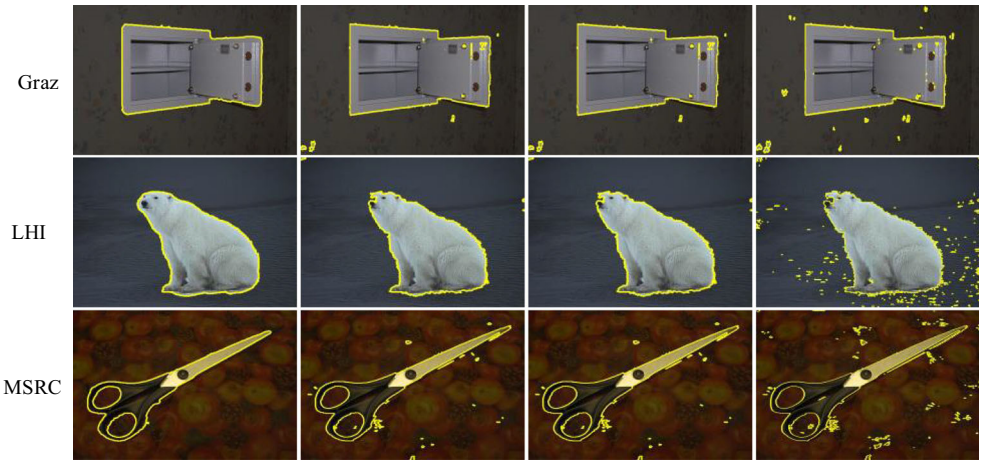


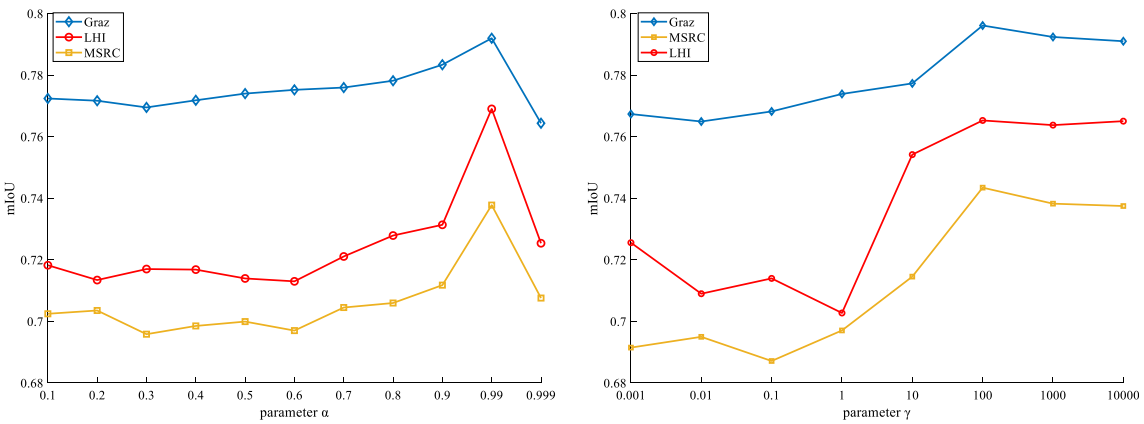
Fig. 6 Segmentation results of our method and degraded methods on three datasets

images in rows 1st to 3rd, the boundaries of the objects are not obvious, i.e., they have weak boundary problems, and our method can segment the boundary well. Because pixel-level features have drawbacks in regional connectivity, they lead to many isolated points in the segmentation results of GrabCut and SMRW. Besides, the segmentation results of RW, LC, and NRW include significant unsegmented regions, so it is essential to fuse the multi-scale features adaptively in that the large scale information can retain more semantic information as illustrated in Fig. 1b. The color of the foreground and background are similar in the last two rows of samples, and the segmentation results of our method are still superior to other methods.

The quantitative experiments were shown in Table 3, which lists the mIoU and VoI of the different algorithms on MSRC. It can be seen that the mIoU of the proposed algorithm is significantly larger than the other algorithms (at least 7% improvement), and the VoI of our method is far smaller than that of other algorithms, which signifies the superior performance of the proposed algorithm.

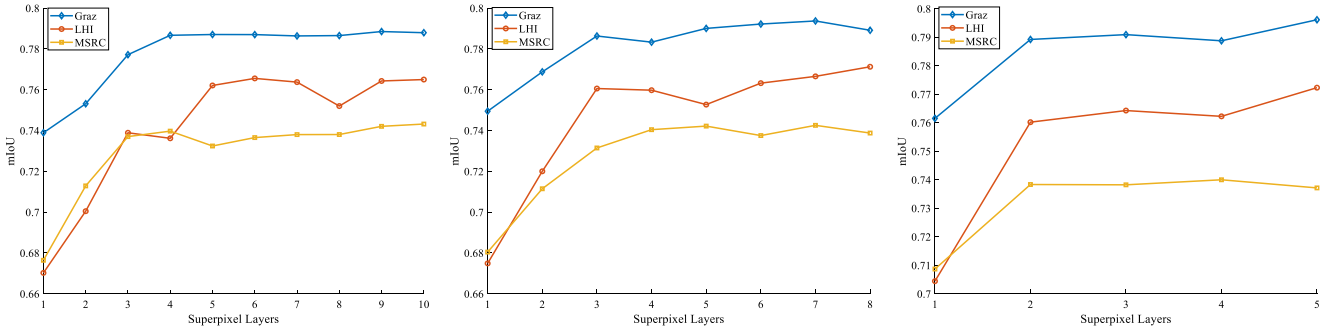
3.4 Ablation study

We analyzed the efficacy of each composition in our method, and two ablation studies were performed on datasets Graz, LHI, and MSRC. First, we removed the adaptive weighting module and adopted the average fusion method to fuse multi-scale layers. As described in Section 2, there are two average fusion strategies: naïve early fusion and naïve late fusion, i.e., solutions of Eqs. (6) and (7), so we adopted these two strategies for ablation study. Second, we removed the fusion strategy and only use pixel level for interactive segmentation [33], i.e., solution of Eq. (4). Table 4 summarizes the mIoU of the proposed method, early fusion strategy, late fusion strategy, and without fusion strategy [33]. It is plain to see that the mIoU of our method is larger than other methods. Furthermore, we can find that the mIoU of early fusion and late fusion are the same. Actually, from Eqs. (6) and (7), it is evident that these two fusion strategies should theoretically achieve the same results. Therefore, adaptive



(a) mIoU-α (b) mIoU-γ

Fig. 7 The mIoUs of our algorithm on Graz, LHI, and MSRC by varying values of different parameters



(a) dense scale (b) sparse scale (c) sparser scale

Fig. 8 The mIoUs of our algorithm on three datasets by varying densities of scale information. **a**: $H = 11$, the numbers of superpixels correspond 0.5 K, 1 K, 1.5 K, 2 K, 2.5 K, 3 K, 3.5 K, 4 K, 4.5 K, 5 K. **b** $H = 9$, the numbers of superpixels correspond 1 K, 2 K, 3 K, 4 K, 5 K, 6 K, 7 K, 8 K. **c** $H = 6$, the numbers of superpixels correspond 1 K, 3 K, 5 K, 7 K, 9 K

fusion with multi-scale features strategy takes a key part in boosting the performance of segmentation.

Same to the above experiments, we need to analyze the qualitative experiment about the ablation study. The segmentation results of the proposed algorithm and degraded methods on datasets Graz, LHI, and MSRC are shown in Fig. 6, and we can discover that results of naïve early/late fusion strategy and without-fusion strategy contain more noise than our method, verifying the effectiveness of our approach. In particular, as seen in the second and fourth images, the segmentation results of early/late fusion and without fusion are devoid of smoothing the boundary of the object.

3.5 Parameter settings

Three parameters require adjustment for the proposed algorithm: α, γ, and the number of superpixel layers. We first conducted experiments by varying parameters α and γ on three datasets (in LHI dataset, annotations of lines are utilized). By fixing $\gamma = 100$, and $H = 6$ (the number of superpixels from the second layer to the seventh layer are 1 K, 3 K, 5 K, 7 K, and 9 K respectively), Fig. 7a shows the mIoU curves while varying parameter α on datasets Graz, LHI, and MSRC. It can be seen that the highest mIoU are obtained when $\alpha = 0.99$ (i.e., $\lambda \approx 0.01$) on these three datasets, which means energy function focus more on boundary term while regional consistency can be guaranteed by adaptive fusing multi-scale features. As illustrated in Fig. 7b, we find that it is detrimental to the segmentation results if γ is small than 100, this is because smaller γ might make the weight coefficients tending to focus on one specific layer and neglect the influence of other layers. Moreover, from the Eq. (15), we can conclude that the segmentation results will be similar to the average fusion strategy based methods if γ is large enough.

For the number of layers, Fig. 8a-c show the experiments according to different scale densities($\alpha = 0.99$ and $\gamma = 100$ are fixed). First, the number of superpixels of each scale is densely distributed, and we performed the experiments with $H = 11$, numbers of superpixels correspond 0.5 K, 1 K, 1.5 K, 2 K, 2.5 K, 3 K, 3.5 K, 4 K, 4.5 K, 5 K. As is shown in Fig. 8a, the value of mIoU converges early (or even drops), and it is difficult to achieve better accuracy. When we set more layers, the time complexity of our algorithm will increase significantly. Second, we experimented with our algorithm with sparse superpixels of each scale (numbers of superpixels correspond 1 K, 2 K, 3 K, 4 K, 5 K, 6 K, 7 K, 8 K). The mIoU has a slight improvement when more layers are merged (shown in Fig. 8b). Finally, the more sparse density of scale information is fused in the algorithm, as illustrated in Fig. 8c. We utilized six layers to conduct the experiments (numbers of superpixels from the second layer to the sixth layer are 1 K, 3 K, 5 K, 7 K, 9 K), we find it can achieve a balance between accuracy and efficiency when we use six superpixel layers.

Table 5 Average runtimes(s) of compared algorithms on datasets Graz, LHI, MSRC

Dataset	RW	SMRW	LC	NRW	TPG	GrabCut	AFMSF
Graz	0.99	2.17	1.74	6.48	1.71	0.78	3.95
LHI	1.06	2.73	2.01	8.09	1.78	0.56	4.28
MSRC	0.79	2.15	1.64	6.56	1.58	0.49	4.55

3.6 Runtime analysis

Table 5 lists the average runtime of RW, SMRW, LC, NRW, TPG, GrabCut, and our method on three datasets (not including superpixel generating time of our method). We performed experiments with MATLAB R2018b on an Intel(R) Core i5–8400 CPU @ 2.80GHz machine with 8G RAM. We find that the runtime of RW and GrabCut is quite low on these three

datasets, while the runtime of our method is only lower than NRW. We can conclude that it is time-consuming for fusion with multi-scale information contains the computation of $\sum_{h=1}^{H}\beta_h D^h$. Since the matrix W^h is sparse, so the computation of the inverse matrix is quietly efficient.

4 Conclusion

We proposed an interactive image segmentation algorithm via adaptive fusion with multi-scale features in this paper. This method considers the importance of the multi-scale features comprehensively. For different images, we can learn different weight coefficients for different scale layers, making the algorithm more robust to different challenges. The experimental results show the evidence that our methods can get a more accurate mask of the object and has an excellent inhibitory effect on noise. However, our method is limited by computation complexity. Therefore, we will pay attention to improving the speed and accuracy of the algorithm simultaneously in the future.

Acknowledgments This work was supported in part by the Natural Science Foundation of Jiangsu Province, China, under Grant BK20180458, in part by the National Science Foundation of China under Grants 61802188, 61673220, 61976028, and 61572085, in part by Project funded by China Postdoctoral Science Foundation under Grant 2020 M681530.

Appendix

Derivation of Eq. (15)

In each iteration process, we only optimize two components β_i and β_j whilst let others be constant, so the Eq. (14) can be written as:

$$\mathrm{F}(\beta_i,\beta_j)=\beta_i M^{(i)}+\beta_j M^{(j)}+\frac{1}{2}\gamma\left(\beta_i^2+\beta_j^2\right)+\mathrm{C} \tag{17}$$

Where C refers to a constant. Because of the equation constraint $\sum_{h=1}^{H}\beta_h=1$, we assume $\beta_i+\beta_j=z$, so $\beta_i=z-\beta_j$. By this assumption, we can obtain:

$$\begin{aligned} F(\beta_i) &= \beta_i M^{(i)}+(z-\beta_i)M^{(i)}+\frac{1}{2}\gamma\left[\beta_i^2+(z-\beta_i)^2\right] \\ &= \left(M^{(i)}-M^{(j)}\right)\beta_i+\frac{1}{2}\gamma\left(2\beta_i^2-2z\beta_i+z^2\right)+zM^{(i)} \\ &= \gamma\beta_i^2+\left(M^{(i)}-M^{(j)}-z\gamma\right)\beta_i+\left(\frac{1}{2}\gamma z^2+zM^{(i)}\right) \end{aligned} \tag{18}$$

It is obvious that objective function (18) with variable β_i is convex, so the optimal value of β_i can be got by differentiating the objective function and making the derivative equal to 0:

$$\begin{aligned} &2\gamma\beta_i+\left(M^{(i)}-M^{(j)}-z\gamma\right)=0 \\ &\Rightarrow 2\gamma\beta_i=z\gamma+M^{(j)}-M^{(i)} \\ &\Rightarrow \beta_i^*=\frac{z\gamma+\left(M^{(j)}-M^{(i)}\right)}{2\gamma} \end{aligned} \tag{19}$$

Via taking $\beta_i+\beta_j=z$ into consideration, we have the final updating form of β_i and β_j:

$$\begin{cases} \beta_i^*=\dfrac{\gamma\left(\beta_i+\beta_j\right)+\left(M^{(j)}-M^{(i)}\right)}{2\gamma} \\ \beta_j^*=\beta_i+\beta_j-\beta_i^* \end{cases} \tag{20}$$

References

1. Kang W, Yang Q, Liang R (2009) The comparative research on image segmentation algorithms. In: International Workshop on Education Technology and Computer Science. pp 703–707
2. Xia Y, Ji Z, Zhang Y (2016) Brain MRI image segmentation based on learning local variational Gaussian mixture models. Neurocomputing 204:189–197
3. Ghosh S, Das N, Das I, Maulik U (2019) Understanding deep learning techniques for image segmentation. ACM Computing Surveys (CSUR) 52(4):1–35
4. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: International Conference on Medical image computing and computer assisted intervention. pp 234–241
5. Shelhamer E, Long J, Darrell T (2017) Fully Convolutional Networks for Semantic Segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3431–3440
6. Bai X, Sapiro G A (2007) Geodesic framework for fast interactive image and video segmentation and matting. In: 2007 IEEE 11th International Conference on Computer Vision. pp 1–8
7. Boykov YY, Jolly M-P (2001) Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In: Proceedings 8th IEEE international conference on computer vision. pp 105–112
8. Cremers D, Rousson M, Deriche R (2007) A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. Int J Comput Vis 72(2):195–215
9. Grady L (2006) Random walks for image segmentation. IEEE Trans Pattern Anal Mach Intell 28:1768–1783
10. Mortensen EN, Barrett WA (1998) Interactive segmentation with intelligent scissors. Graph Model Image Process 60(5):349–384
11. Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24(5): 603–619
12. Ji Z, Xia Y, Chen Q, Sun Q, Xia D, Feng DD (2012) Fuzzy c-means clustering with weighted image patch for image segmentation. Appl Soft Comput 12(6):1659–1667
13. Rother C, Kolmogorov V, Blake A (2004) " GrabCut" interactive foreground extraction using iterated graph cuts. ACM Trans Graph (TOG) 23(3):309–314
14. Freedman D, Zhang T (2005) Interactive graph cut based segmentation with shape priors. In: 2005 IEEE computer society conference on computer vision and pattern recognition. pp 755–762

15. Shen J, Du Y, Wang W, Li X (2014) Lazy random walks for superpixel segmentation. IEEE Trans Image Process 23(4):1451–1462
16. Li Y, Sun J, Tang C-K, Shum H-Y (2004) Lazy snapping. ACM Trans Graph (ToG) 23(3):303–308
17. Wang T, Ji Z, Sun Q, Chen Q, Jing X-Y (2016) Interactive multilabel image segmentation via robust multilayer graph constraints. IEEE Trans Multimed 18(12):2358–2371
18. Kim TH, Lee KM, Lee SU (2010) Nonparametric higher-order learning for interactive segmentation. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp 3201–3208
19. Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905
20. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S (2012) SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 34(11):2274–2282
21. Kallel I K, Almouahed S, Solaiman B, Bossé É (2018) An iterative possibilistic knowledge diffusion approach for blind medical image segmentation. Pattern Recognit 78:182–197
22. Yang X, Prasad L, Latecki LJ (2012) Affinity learning with diffusion on tensor product graph. IEEE Trans Pattern Anal Mach Intell 35(1):28–38
23. Wang B, Tu Z (2012) Affinity learning via self-diffusion for image segmentation and clustering. In: 2012 IEEE conference on computer vision and pattern recognition. pp 2312–2319
24. Donoser M, Bischof H (2013) Diffusion processes for retrieval revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1320–1327
25. Bai X, Yang X, Latecki LJ, Liu W, Tu Z (2009) Learning context-sensitive shape similarity by graph transduction. IEEE Trans Pattern Anal Mach Intell 32(5):861–874
26. Dong X, Shen J, Shao L, Van Gool L (2015) Sub-Markov random walk for image segmentation. IEEE Trans Image Process 25(2):516–527
27. Casaca W, Gustavo Nonato L, Taubin G (2014) Laplacian coordinates for seeded image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp 384–391
28. Bampis CG, Maragos P, Bovik AC (2016) Graph-driven diffusion and random walk schemes for image segmentation. IEEE Trans Image Process 26(1):35–50
29. Pedronette DCG, Torres RdS (2016) Rank diffusion for context-based image retrieval. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. pp 321–325
30. Luo L, Shen C, Zhang C, van den Hengel A (2013) Shape similarity analysis by self-tuning locally constrained mixed-diffusion. IEEE Trans Multimed 15(5):1174–1183
31. Bai S, Zhou Z, Wang J, Bai X, Latecki LJ, Tian Q (2018) Automatic ensemble diffusion for 3d shape and image retrieval. IEEE Trans Image Process 28(1):88–101
32. Krishna K, Murty MN (1999) Genetic K-means algorithm. IEEE Trans Syst Man Cybern Part B (Cybern) 29(3):433–439
33. Wang T, Yang J, Ji Z, Sun Q (2018) Probabilistic diffusion for interactive image segmentation. IEEE Trans Image Process 28(1):330–342
34. Yang Q, Li C, Guo J (2019) Multi-order information for working set selection of sequential minimal optimization. In: the 22nd International Conference on Artificial Intelligence and Statistics. pp 3264–3272
35. Santner J, Pock T, Bischof H (2010) Interactive multi-label segmentation. In: Asian conference on computer vision. pp 397–410
36. Yao B, Yang X, Zhu S-C (2007) Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. In: International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. pp 169–183
37. Meilă M (2005) Comparing clusterings: an axiomatic view. In: Proceedings of the 22nd international conference on Machine learning. pp 577–584
38. Jang W, Kim C (2019) Interactive Image Segmentation via Backpropagating Refinement Scheme. Comput Vis Pattern Recognit 5297–5306

Zongyuan Ding received the M.S. in the School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, China, in 2018. And he is currently pursuing the Ph.D. in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research includes image segmentation, pattern recognition.

Tao Wang received the B.E. and Ph.D. in the School of Computer Science and Engineering, Nanjing University of Science and Technology(NUST), Nanjing, China, in 2012 and 2017, respectively. He is currently an associate professor with the School of Computer Science and Engineering, NUST. His current research interests include image processing, medical imaging, and pattern recognition