Interactive Image Segmentation with First Click Attention

Abstract

In the task of interactive image segmentation, users initially click one point to segment the main body of the target object and then provide more points on mislabeled regions iteratively for a precise segmentation. Existing methods treat all interaction points indiscriminately, ignoring the difference between the first click and the remaining ones. In this paper, we demonstrate the critical role of the first click about providing the location and main body information of the target object. A deep framework, named First Click Attention Network (FCA-Net), is proposed to make better use of the first click. In this network, the interactive segmentation result can be much improved with the following benefits: focus invariance, location guidance, and error-tolerant ability. We then put forward a click-based loss function and a structural integrity strategy for better segmentation effect. The visualized segmentation results and sufficient experiments on five datasets demonstrate the importance of the first click and the superiority of our FCA-Net.

1. Introduction

Interactive image segmentation aims to segment the instances of interest with minimal user input. It directly benefits many applications, *e.g*. image editing and medical imaging analysis . Recent years, with the popularization of data-driven deep learning techniques, the demand for mask-level annotations has increased dramatically in some fields, such as salient object detection , semantic segmentation , instance segmentation , camouflaged object detection , and image/video manipulation . Efficient interactive segmentation technologies are in urgent need to alleviate the annotating cost. Therefore, more and more researchers are carrying out extensive exploration in this field.

Many ways of interaction have been explored, such as bounding boxes , scribbles , and points . Drawing a bounding box as the interaction is a widely-used and convenient way. However, in most cases, users usually need to further correct the segmentation results which are not satisfactory enough. Therefore, the more practical approaches are based on interaction points or scribbles, which can further improve the segmentation result by iteratively marking the mislabeled areas. Compared with drawing scribbles, the clicking for points places less burden on users because it does not require a drag process. The typical interaction workflow of point-based methods, seen in figure, is as follows: Users first provide a positive point on the target object. According to the initial segmentation result, users further provide a positive point on the foreground or a negative point on the background, and the segmentation result is iteratively refined until it meets the users' requirement.

A mount of traditional and deep learning based methods have been explored in this direction. For most existing works, they use all interaction points indiscriminately to generate the final predictions. However, we observe that not all interaction points have the same segmentation effect. We collect the statistics of real human interactions over 2000 images on an interactive segmentation method , as shown in table. We find that the first click often plays an important role in interactive segmentation. The performance improvement of the first point is remarkable and the first click is usually close to the center of the target object. Combined with the workflow mentioned above, there is an intuitive observation that the first click is of importance and it can serve as a location indication and global information guidance for the target object. From figure, we can see that the object segmentation can obtain a fine initialization with the first click. On contrary, the goal of other interaction points is to achieve better segmentation based on the result of the first one. Thus, the first point is more conducive to obtaining the overall information of the object, while the other points focus on refinement. Based on the analyses mentioned above, we conjecture that specially treating the first click will benefit interactive segmentation.

In this paper, we are the first to treat these two kinds of points separately. We propose a First Click Attention Network (FCA-Net), where a simple path on the basic segmentation network is constructed for further verification. In our network, we use the first click as the side input to supervise the global segmentation. With the first click as an anchor for interactive segmentation, the location and main body information of the target object can be better guided. The prediction mask will

focus on the area around the first click and get a better result. For network training, we propose an improved loss function, which takes all clicks provided by users into consideration and focuses on these regions around clicks. We finally raise a new post-process strategy, whereby some small mispredicted areas could be feasibly removed, and the structural integrity of the segmented object could be maintained. We conduct comprehensive experiments and achieve the state-of-the-art performances on GrabCut , Berkeley , PASCAL VOC , DAVIS , and MSCOCO datasets. Results and analyses of comparative experiments prove the uniqueness of the first click and effectiveness of our proposed methods.

Our contributions can be summarized as follows:

- ▷ This is the first work to demonstrate the critical role of the first click. We also propose a FCA-Net, which is equipped with a simple yet effective module for utilizing the guidance information of the first click.
- ▷ We propose the click loss considering annotations of users and a structural integrity strategy, which is helpful in the task of interactive segmentation.
- ▷ The state-of-the-art results over five datasets demonstrate the importance of the first click and effectiveness of our FCA-Net, click loss function, and structural integrity strategy.

2. Related Work

In the early years, most traditional methods of interactive segmentation mainly made use of hand-crafted features. Some research methods such as paid much attention to the boundary properties. Approaches based on graphical models became more popular after , where the interactive segmentation task is modelled as a graph cut optimization problem and it can be efficiently solved by the well-known min-cut/max-flow algorithm . Among them a classic method based on graph cut called GrabCut was proposed in . It takes the Gaussian mixture model as the color model and the bounding box as input to simplify the segmentation process. Kim *et al*. improved the algorithm of random walk which is proposed in with restart. Kim *et al*. also introduced a new higher-order formulation, additionally imposing the soft label consistency constraint. Gulshan *et al*. and Bai *et al*. both applied geodesic distance for optimization in interactive image segmentation. Bai *et al*. provided an error-tolerant method, which allows users to have some wrong interactions. These methods based on low-level features cannot adapt to object segmentation in complex and variable scenes.

Neural networks have the ability to perceive complex global and local features. With the popularization of deep learning, more and more researches have been trying to apply neural networks to interactive segmentation. In recent years, Xu *et al*. first proposed a CNN-based model with some effective point sampling strategies for training in this field. Then, Liew *et al*. proposed a RIS-Net to capture regional information according to pairs of positive and negative points for local refinement. Song *et al*. applied reinforcement learning to make computers generate more potential interaction points. Scuna *et al*. utilized recurrent neural networks to get precise segmentation which can be represented as a polygon consisting of multiple points. Then Ling *et al*. improved the polygon-based method above with graph convolutional networks recently. Li *et al*. used neural networks to provide and select a more accurate choice to solve ambiguity situations in interactive segmentation. Maninis *et al*. provided a novel interactive way about extreme points for segmentation. Mahadevan *et al*. put forward an effective strategy of iterative training in this area. Hu *et al*. raised a two-stream fusion network for interactive segmentation. Jang and Kim offered a backpropagating refinement scheme to force each interaction point to have the correct segmentation result. Majumder and Yao made use of interaction points to generate special guidance maps as input to neural networks according to some other information, such as superpixel. All these methods have a commonality that they treat all interaction points indiscriminately in neural networks. However, we find and propose the uniqueness of the first point and take it as a special guidance in our network architecture.

3. Proposed Method

This section contains five parts. In Algorithm, we introduce our proposed FCA-Net, which treats the first point specially. In Algorithm, we describe the calculation process of proposed click loss, which is used to assist our interactive segmentation network to achieve better performance. In Algorithm, we explain the structural integrity strategy for postprocessing the prediction of FCA-Net. In Algorithm, we analyze some benefits of adopting our first click attention through some comparative examples. In the end, we will show the implementation details of our interaction point simulation strategies and the settings of training in Algorithm.

3.1. Network Architecture

The architecture of FCA-Net is shown in figure. To explain the validity of the first click better, we do not make too many changes to the widely-used network structure of the interactive segmentation. Instead, a simple additional module called first click attention module is added to the basic segmentation network. Therefore, FCA-Net can be split into a basic segmentation network and a first click attention module.

Basic Segmentation Network. Following , we employ the common FCN architecture, whose specific structure is similar to DeepLab v3+ . As shown in figure, it contains three parts: a backbone network, an Atrous Spatial Pyramid Pooling (ASPP) module, and a decoder module. We take ResNet101 as the backbone. We denote the features of the last four stages as $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$. To capture multi-scaled objects in interactive segmentation, we also adopt dilated convolutions on the last stage of ResNet101 instead of taking stride as 2. Thus, output stride of the backbone is 16. The input of backbone is the RGB image concatenated with two Gaussian maps of annotated positive points and annotated negative ones. The Gaussian map is calculated according to the Euclidean distance map, as shown in figure. The Gaussian radius in our experiments is set to 10.

For the ASPP module shown in figure, the input is concatenated features ($\mathbf{F}_4 \oplus \mathbf{F}_{FCA}$), where $\oplus$ means the concatenation operation and $\mathbf{F}_{FCA}$ means the output of the first click attention module. The concatenated features are fed into four dilated convolutional layers with different dilation sizes of 1, 6, 12, 18 and a global average pooling layer. Then the output features of the five branches are concatenated and fed into an extra convolutional layer. For the decoder module shown in figure, it takes the low-level features $\mathbf{F}_1$ and the output features of ASPP as input and use convolutional layers to generate the final prediction result. To supervise the prediction result, we design a click-based loss function to replace conventional binary cross entropy loss function. We call this general click loss, which is detailed in Algorithm.

First Click Attention Module. For utilizing the guidance information of the first click, we design a simple module alongside the basic segmentation network. It takes the low-level features $\mathbf{F}_1$ and the Gaussian map $\mathbf{M}_f$ centered on the first click as input. The concatenated features ($\mathbf{F}_1 \oplus \mathbf{M}_f$) are fed into six 3×3 convolution layers. On the first and fourth layers, we take stride as 2 to reduce the resolution. The number of channels is 256 in the first three layers, and 512 in the last three ones. Thus, the output features $\mathbf{F}_{FCA}$ are with 512 channels. It will be merged into the basic segmentation network before the ASPP module. In addition, we supervise the $\mathbf{F}_{FCA}$ with a first click loss, which focuses on these pixels around the first point. We will describe the details in Algorithm.

To better illustrate the effect of the first click attention, in figure, we visualize the predicted probability maps of the model with FCA (c-d) and without FCA (b). Note that, in these three tests (b-d), the coordinates of these positive points are exactly consistent. As can be seen in figure (b), in the absence of FCA, the two positive points share the same importance. By introducing the FCA (c-d), the model's attention shifts. In test (c) and (d), the order of the positive points marked by the user is different. We can see that no matter where it is, the first click attracts more attention, serving as a segmentation anchor, while the other points play an auxiliary role for detail repair. Compared with the equal treatment of interaction points, the introduction of the FCA makes the model work more in line with the real interaction behavior of users as discussed in Algorithm.

3.2. Click Loss

For a better explanation in the following parts, we define some symbols and operations here. All pixels are represented as $\mathcal{G}$. We use $\mathcal{G}_p$ and $\mathcal{G}_n$ to represent the sets of pixels in foreground and background according to the ground truth mask. $\mathcal{A}$ represents all the annotated points. $\mathcal{A}_p$ and $\mathcal{A}_n$ represent positive points and negative points respectively. We use $d(p_1, p_2)$ to represent the Euclidean distance between point p_1 and point p_2. And we use $\phi(p, \mathcal{S})$ to represent the shortest distance from one point p to another region $\mathcal{S}$, which is defined as:

For the task of binary segmentation, we usually use binary cross entropy (BCE) as a loss function to supervise the neural network. The loss function is beneficial to focus on the global segmentation quality. For the interactive segmentation task, we prefer to see that user interactions can play a guiding role. It is preferable that more accurate results at and around these interaction points, so we design a loss function based on user interactions to assist our FCA-Net with better performance.

The click loss can be thought of a kind of weighted binary cross entropy loss. Conventional binary cross entropy loss function can be formulated as follows: where x_p means the probability of point p in prediction mask and y_p means the label of point p in the ground truth mask (0 or 1).

Firstly, we define a function ψ to represent the distance weight between a point p and a set of annotated points $\mathcal{S}$ (*e.g.* $\mathcal{A}_p$ and $\mathcal{A}_n$), which is formulated as follows: where τ is the influence range of each annotated point.

For the loss function to supervise the final prediction, we propose a loss called general click loss ($\mathcal{L}_g$) which considers all clicks, which is formulated as follows: N is the number of all pixels. The weight in equation can be represented as follows: where α and β are used to adjust the range of loss.

For the loss function to supervise the output of FCA module, we use a special loss called first click loss ($\mathcal{L}_f$) which focuses on the area around the first point. It is formulated as follows: The weight in equation can be represented as follows: where a_f means the first point in $\mathcal{A}_p$.

In experiments, we choose τ at 100, α at 0.8, β at 2.0.

3.3. Structural Integrity Strategy

Through experiments, we find that the prediction masks of neural networks may contain some scattered regions of wrong results. In most cases, people will prefer to get the object mask which maintains structural integrity in the task of interactive segmentation. Therefore, we propose a strategy to maintain the structural integrity of the segmentation based on interaction points.

Normally, we take 0.5 as the threshold to get the final binarized mask from the output of neural networks. Let $\mathcal{P}$ represent these points which are predicted as foreground. We will postprocess these prediction areas according to the interaction points and get new $\mathcal{P}'$, which is formulated as follows: where $\sigma(p_1, p_2) = 1$ when there is an eight-connected path from point p_1 to point p_2. The structural integrity strategy can work in most cases. The effect of it can be seen in table.

3.4. Strength Analysis

Can the first click attention really improve the quality of segmentation? In this section, we will illustrate some benefits of joining the first click supervision by comparing some visual results in figure.

Focus Invariance. We know that all positive and negative points are equally important in most methods. They take all annotated points as input to generate the final result. These positive points except the first one are often clicked for repairing local details and may be close to the boundary of the target object. If the neural network treats these points equally as the first point, it will often result in a wrong segmentation. For example, in figure (a), we want to segment the table with a white tablecloth. The first click is near the center of the table. The other positive point is used to fix errors near the edge of the table. Without the guidance of the first point, the neural network will mistakenly segment the person in the image because it treats each point equally. With the help of our first click attention, there will be fewer wrong segmentations.

Location Guidance. Obviously, the first point guides the location of the target object. If there are multiple objects in the scene, there will be less error segmentation in local regions with the help of the first point. For example, in figure (b), we want to segment the left sheep. We click three negative points around the right sheep. Without the accurate understanding of the global location information, the network may be mistaken that there is a target object in the area surrounded by these negative points. This may cause some errors, such as the wrong prediction of the right sheep. With the first click attention, the prediction will focus on the location of the first click and get a better result.

Error-Tolerant Ability. In the process of interactive segmentation, it is inevitable that there will be some click errors, especially at the edge of the object or in the area where the background is similar to the foreground. For example, in figure (c), we want to segment the penguin. A positive point on the right near the boundary of the target object accidentally falls into the background area. We can see that this may cause serious segmentation errors, as shown in the right one of figure (c) if we do not use the first click attention. With the guidance of the first point, the influence of these error points will be greatly reduced.

3.5. Implementation Details

In this section, we will show some details on the training. Since user-annotations are unavailable in those segmentation datasets, we turn to take some strategies to simulate various interaction points as done in most papers, including general clicks and the first click. We will also introduce our training settings in this section.

General Click Simulation. For most clicks, we use the strategy similar to that in . The numbers of clicks in the foreground and background are determined randomly within [1, 10] and [0, 10], respectively. For positive points, they are chosen on the foreground, at least P_1 pixels away from the object boundaries and P_2 pixels from each other. We define $\mathcal{A}^*$ as the set of

these previous annotated points. A new positive point is selected from a candidate set $\mathcal{C}_p$, which can be expressed as follows: For negative points, they are chosen in the background, $N_1 \sim N_2$ pixels away from the object boundaries and N_3 pixels from each other. A new negative point is selected from a candidate set $\mathcal{C}_n$, which can be expressed as follows: In our experiments, we choose P_1 in $\{5, 10, 15, 20\}$, P_2 in $\{7, 10, 20\}$, N_1 in $\{15, 40, 60\}$, N_2 in $\{80\}$, N_3 in $\{10, 15, 25\}$.

First Click Simulation. The first click is always on the target object, and it is usually close to the object center. We thus use $\mathcal{E}(p)$ (called CD in table) to express the distance between the point p and the object center, which is formulated as follows: Here $\mathcal{E}(p)$ closer to 1 means that the first click point locates at a more central position of the object. In our experiments, we choose the point whose $\mathcal{E}(p)$ equals 1 in cropped training images as the first point. The Gaussian radius of it is set to three times that of general points.

Training Settings. We train the FCA-Net on 10582 training images of the augmented dataset (PASCAL VOC + SBD) which excludes the validation images of PASCAL VOC dataset. Actually, we can get 25832 instance-level images and corresponding masks for training. The input image is proportionally resized with its smaller side fixing to 512 pixels. Then, we take a random crop with 512 $\times$ 512 pixels guaranteeing that the cropped image contains at least a part of the object. We take the same iterative training strategy for clicks simulation. We take ResNet101 pre-trained on ImageNet as a backbone. We set the batch size to 8. We set the initial learning rate to 0.007 for ResNet and 0.07 for other parts and take stochastic gradient descent with 0.9 momentum for optimization. We adopt the polynomial learning rate decay for 30 epochs and constant learning rate for additional 3 epochs in the end. All the experiments are implemented with the PyTorch framework and run on a single NVIDIA Titan XP GPU.

4. Experiments

4.1. Evaluation Details

Datasets. We adopt the following widely used datasets for evaluation:

- **GrabCut :** The dataset contains 50 images and is used in most methods for interactive segmentation. Most of the images have obvious differences between foreground and background.
- **Berkeley :** The dataset contains 100 object masks on 96 images. There are some images that are difficult to segment in this dataset because of similar appearances in foreground and background.
- **PASCAL VOC :** We use the validation set in this dataset which contains 1449 images with 3427 instances. Thus, we take these instance-level object masks for validation. These objects are semantically consistent with the data used for training.
- **MSCOCO :** The dataset contains objects of 80 categories. We divide this dataset into MSCOCO (seen) and MSCOCO (unseen) and sample 10 images per category for evaluation as done in .
- **DAVIS :** The dataset is for video object segmentation. It contains 50 videos whose ground truth masks are of high quality. We sample the same 10% frames as for evaluation.

Metrics. Following , we employ the Mean Intersection Over Union (mIoU) as a metric. We also take a robot user to simulate clicks in the evaluation. Specifically, the first point will undoubtedly be a positive point to guide the segmenting of the target object. We will get a prediction mask based on annotated points. Then the next point will be placed in the center of the largest error region. We plot the curves of the mIoU and the number of clicks for comparing the performance of each method on the fixed interactions. We adopt the Mean Number Of Clicks (mNoC) as an evaluation metric, which reflects the average interactions for obtaining a certain IoU threshold on each sample of a dataset. The selection of IoU thresholds is different for each dataset and the default maximum number of clicks is limited to 20 for each sample. These settings above are consistent with the previous works.

Inference Time. We test the inference speed on the Intel i7-8700K 3.70GHz CPU and a single NVIDIA Titan XP GPU. It takes about 0.07 second for each click on a 512 $\times$ 512 image. The speed is fast enough to meet the needs of real-time interaction.

4.2. Comparison with the State-of-the-Art

We compare our results with other existing methods, including graph cut (GC) , growcut (GRC) , random walk (RW) , geodesic matting (GM) , Euclidean star convexity (ESC) , geodesic star convexity (GSC) , deep object selection (DOS) , regional image segmentation (RIS) , latent diversity based segmentation (LD) , backpropagating refinement scheme (BRS) , and content-aware multi-level guidance (CMG) . Some scores are from .

figure illustrates the mIoU of each method on the different number of clicks. The curves of FCA-Net are plotted on the results without structural integrity strategy for post-process. We can see that the curves of our method are superior to other methods after the first point in most cases. This is in line with our expectations. With the first point as the main body and location guidance, the prediction of neural networks will contain fewer error regions. Thus, the FCA-Net can produce more accurate results.

table shows the mNoC metric on six sets of five datasets. Our FCA-Net reaches the state-of-the-art in five datasets. With a structural integrity strategy to postprocess the result, the performance will be further improved. We do not make too many changes on the network architecture and just set up a simple first click attention module. However, the improvement of performance is significant, which also indirectly reflects the unique validity of the first point.

4.3. Ablation Study

To further verify our contributions, we conduct ablation study on the validation set of PASCAL VOC and Berkeley. We take the basic segmentation network as the baseline (No.1), and gradually equip the strategies mentioned in this paper (No.2-5). The ablated results of the mean number of clicks (mNoC) are shown in table. Comparing No.2 with the baseline, adding the FCA module, we find that the performance is dramatically improved with 0.55 and 0.52 alleviating of mNoC. This improvement is in line with our expectation that the guidance information of the first click is utilized more effectively by introducing the FCA. Comparing No.3 with No.2, we see that the click loss proposed in this paper brings a considerable effect improvement. We also employ the same iterative training strategy , which improves the effect of the final model to a certain extent. Since the proposed FCA-Net is only a simple implementation to explore the critical role of the first click, we do not modify the widely-used framework too much; thus, in practice, we can get better results by replacing stronger backbones or more sophisticated designs. For instance, in No.5, we take Res2Net to replace ResNet as the backbone, which further improves the accuracy. Finally, we use the proposed structural integrity strategy to postprocess results and show its reliable improvements in table.

4.4. Limitation Analysis

In this section, we discuss the possible limitations of our FCA-Net in some special cases. As shown in figure (a), due to the strong location prior provided by the first click, our FCA-Net is not good at segmenting multiple instances in an image at the same time. Fortunately, in real-world applications, the limitation can be alleviated by annotating for each instance object with its own first click. In the figure (b-c), we show two interesting scenes, where the center of these instances may not be clicked by users due to the structure or occlusion. In these cases, the locating guidance may deviate from the center. It will sometimes lead to unsatisfactory segmentation results, for which the users have to add more points for repairing.

5. Conclusions

In this paper, we explore and demonstrate the importance of the first click for interactive segmentation. We propose a FCA-Net, which adds a simple module on the basic segmentation network to shift more attention to the first click. We also raise an effective click-based loss function for our FCA-Net and a new strategy to maintain the integrity of prediction masks. The state-of-the-art performances over 5 datasets show the importance of the first click and superiority of our methods.

Acknowledgements

This research was supported by Major Project for New Generation of AI under Grant No. 2018AAA0100400, NSFC (61922046, 61972216), the national youth talent support program, and Tianjin Natural Science Foundation (17JCJQJC43700, 18JCYBJC41300). Shao-Ping Lu is the corresponding author of the paper.