symmetry

MDPI

Article

Interactive Image Segmentation Based on Feature-Aware Attention

Jinsheng Sun [1], Xiaojuan Ban [1,2,*], Bing Han [3], Xueyuan Yang [1] and Chao Yao [4]

[1] Institute of Artificial Intelligence, University of Science and Technology Beijing, Beijing 100083, China
[2] School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China
[3] Shunde Graduate School, University of Science and Technology Beijing, Foshan 528399, China
[4] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
* Correspondence: banxj@ustb.edu.cn

Abstract: Interactive segmentation is a technique for picking objects of interest in images according to users' input interactions. Some recent works take the users' interactive input to guide the deep neural network training, where the users' click information is utilized as weak-supervised information. However, limited by the learning capability of the model, this structure does not accurately represent the user's interaction intention. In this work, we propose a multi-click interactive segmentation solution for employing human intention to refine the segmentation results. We propose a coarse segmentation network to extract semantic information and generate rough results. Then, we designed a feature-aware attention module according to the symmetry of user intention and image semantic information. Finally, we establish a refinement module to combine the feature-aware results with coarse masks to generate precise intentional segmentation. Furthermore, the feature-aware module is trained as a plug-and-play tool, which can be embedded into most deep image segmentation models for exploiting users' click information in the training process. We conduct experiments on five common datasets (SBD, GrabCut, DAVIS, Berkeley, MS COCO) and the results prove our attention module can improve the performance of image segmentation networks.

Keywords: interactive segmentation; feature-aware; attention; human-computer interaction

Citation: Sun, J.; Ban, X.; Han, B.; Yang, X.; Yao, C. Interactive Image Segmentation Based on Feature-Aware Attention. *Symmetry* **2022**, *14*, 2396. https://doi.org/10.3390/sym14112396

Academic Editors: João Ruivo Paulo, Cristina P. Santos and Gabriel Pires

Received: 16 October 2022
Accepted: 3 November 2022
Published: 12 November 2022

1. Introduction

Interactive image segmentation is a critical component of advanced image editing software. It provides interactive inputs like strokes, bounding boxes, and clicks to let users choose objects of interest. The segmented results can be useful for various activities, such as localized editing and image/video composition. Most existing works focus on combining interaction information as weak labels for training segmentation models, without considering the users' impact. As a result, emphasizing the importance of human intention in the interactive segmentation process is an important issue.

Interactive segmentation has been a topic of research for decades, with early algorithms relying primarily on low-level hand-crafted features to create algorithms or models that work admirably on simple images [1–5]. However, several characteristics, such as lighting levels, angles, and postures, make these features less resilient, limiting the performance of segmentation. An inadequate availability of interaction information is especially an issue. Users must exert a substantial interactive effort to get satisfactory outcomes.

Thanks to the outstanding performance in computer vision tasks, deep learning techniques are rapidly being utilized more widely in image segmentation. Some of the latest works based on the click-point information have explored how to properly embed into deep models. Deep interactive object selection (DOS) [6] is the first deep solution to the interactive segmentation problem. To supervise the global segmentation, Majumder and Yao [7] transformed user clicks into content-aware guidance maps to exploit more hierarchical

information. Ding et al. [8] used the early-late fusion strategy to construct a click encoding method based on deep semantic information. These deep learning methods mainly try to use the interactive information as weak-supervise labels to train the semantic segmentation modules. The weak-supervised information guides the learning of the position of semantic features; however, a majority of potential information that humans can easily be aware of, such as semantic attributes and semantic relevance, etc., are eliminated. Therefore, the existing algorithms cannot establish accordant relationships between interaction information and semantic features.

Alternatively, in order to effectively combine local and global information for many computer vision applications, attention processes have been extensively researched in deep CNNs, such as object detection [9], classification [10], and segmentation [11]. By avoiding the use of multiple similar feature maps and highlighting salient features that are useful for a given task, attention enables the network to focus on the most relevant features without the need for additional supervision, in contrast to standard multi-scale features fusion approaches that compress an entire image into a static representation. Lin et al. [12] employed the first click attention module to better use the first click. Attention modules have also helped semantic segmentation networks, improving models for pixel-wise identification tasks [13,14].

This paper explores a feature-aware attention mechanism to simulate the human intention in the interactive image segmentation task. The potential semantic relevance is learned from the users' click action. First, a pre-trained image segmentation network is used in the proposed method to generate coarse segmentation results. To refine the semantic features, we propose a progressive fusion strategy to integrate the multi-click information to learn the relevant semantic information. Finally, we design a refinement network to achieve precise segmentation. We conduct thorough experiments on five well-known datasets, and the experimental outcomes show the usefulness and adaptability of the proposed architecture.

2. Related Work

Early interactive image segmentation methods typically consider boundary attributes or use graphical methods. For instance, Grady [4] calculated the likelihood that a random walker at each unlabeled pixel should first arrive at one of the labeled pixels by rewriting the task as a combinatorial Dirichlet problem. Based on weighted geodesic distances, Bai [15] divided each pixel into the foreground and background. Freedman [2] used shape priors to further improve performance. However, all of these traditional methods typically rely on hand-crafted features. As a result, these models frequently produce unacceptable results when the background is complicated, or there are different lighting conditions.

In recent years, with the success of deep learning in several computer vision tasks, image segmentation has also achieved impressive results [8,12,16–19]. It impacts the majority of interactive image segmentation models [20]. Weakly supervised segmentation methods revamp coarse segmentation masks from weaker supervision information. The following sources of supervision have been investigated for semantic/instance segmentation: image level label only [21], point clicks [6,22], boxes only [23–25], scribbles only [22,26], boxes plus clicks [24]. Regardless of the form of interaction, interactive segmentation optimizes or changes the segmentation result by capturing the user intent in the interaction information. Therefore, the model structure of these methods is almost the same; the main difference is the interaction point fusion module.

Most interactive segmentation networks focus on employing the user's input information into segmentation networks [27]. The first CNN-based model [16] introduced some effective point sampling strategies. Liew et al. [28] captured the local regional information surrounding user input for local refinement. Acuna et al. [29] used a recurrent architecture to learn the precise segmentation, which could be represented as a polygon consisting of multiple points. Ling et al. [30] predicted all vertices simultaneously using a graph convolution and parameterized objects with either polygons or splines. Islam et al. [31] proposed

a label refinement network that predicted segmentation labels in a coarse-to-fine fashion. Yang et al. [32] integrated both pixel-level and instance-level embedding to implement the match between the reference and the predicted sequence, which could make the model robust to various object scales. Lin et al. [12] used both image and users' input information to produce the segmentation results. However, the input information represented via distance transforms is not learnable. Majumder and Yao [7] transformed user clicks into content-aware guidance maps to exploit more hierarchical information. Jang and Kim [31] altered the backpropagation scheme to improve model performance. Nevertheless, these fusion strategies treat intention clicks as weak-labeled information via the concatenation of the semantic and click embeddings. Limited by the learning ability of the model, this strategy can not precisely estimate the interaction intention.

Besides the click information, Ding et al. [33] used phrase expressions as an additional interaction input to estimate the features of the target object. Kontogianni et al. [34] used corrections to adapt the model parameters to a specific object and its background, or shifted distributions. Sofiiuk et al. [31] proposed f-BRS that solves an optimization problem concerning auxiliary variables instead of the network inputs to reduce computing cost. In [35], the previous step mask is concatenated with the click maps to refine the segmentation iteratively.

Even though attention mechanisms are becoming popular in many vision problems, the literature on interactive image segmentation with attention remains scarce, with simple attention modules. Wang et al. [20] employed attention modules at multiple resolutions to combine local deep attention features (DAF) with global context for prostate segmentation on Ultrasound images. To capture long-range dependencies, local and global features are combined in a simple attention module, which contains three convolutional layers followed by a softmax function to create the attention map. A similar attention module, composed of two convolutional layers followed by a softmax, is integrated with a hierarchical aggregation framework integrated into UNet for left atrial segmentation.

3. Proposed Method

We propose a novel network for interactive segmentation. The core idea is to extract the guidance map by combining the click information with the input image for image segmentation. Specifically, given an image X, the segmentation is to generate a coarse result, denoted as I_c. We integrate the multi-scale feature Fsa from the backbone E. DeepLabV3+ [36] is adopted as our segmentation network following [8,12]. Then, a feature-aware attention module combines the interactive points sets (S_p, S_n), where p represents positive clicks while n represents negative clicks, the multi-scale feature map Fsa, and a guidance map I_g, which denotes the difference between the human intention and automatic segmentation results. Finally, a refinement network, denoted as R generates the final segmentation results using the coarse segmentation result and the guidance map.

3.1. Coarse Segmentation Module

The task of interactive segmentation is very similar to instance or semantic segmentation in terms of the network architecture. The key difference is in the user input: its main aspects are the encoding and processing of the encoded input inside the network. Therefore, it is reasonable to rely on time-tested state-of-the-art segmentation networks and focus on interaction-specific parts. Following [31], we adopt the DeepLabV3+ [36] semantic segmentation architectures as a backbone.

Three components comprise the coarse segmentation network: an encoder, an ASPP module, and a decoder. Following [36], we establish ResNet with removing the fully connected layer as the encoder network, denoted as E. Let F_0, F_1, F_2, F_3, F_4 be the five layers of features, which are used to calculate the guidance map with the click points. To capture multi-scale context information due to excessive feature downsampling, the striding of the last two layers (F_3, F_4) are replaced with atrous convolution (dilation = 2). The input of the encoder is X, while the output of the decoder is the coarse segmentation results I_c.

3.2. Feature-Aware Attention Module

To better utilize the user interaction intent to refine the image segmentation results, based on the symmetry of user clicks and high-dimensional features of images, we use the attention mechanism to extract the user segmentation intent from the click information. In order to learn potential semantic relevance from users' click actions to refine the segmentation, we follow the multi-scale strategy recently used in [11,20]. Feature integrated multiple scales are denoted as F_s, where s indicates the level in the backbone (Figure 1). Bilinear interpolation is used for upsampling features to a similar resolution because they arrive at different resolutions for each level. Then, SF_s from all the scales are concatenated, forming a tensor that is concatenated to create a common multi-scale feature map IF_s and fed into the guided attention modules with click set S to generate the attention features G_s.

$$G_s = CGA(IF_s, S) \tag{1}$$

where CGA represents each click-guided attention module. In the following sections, we detail how the attentive features G_s are obtained.

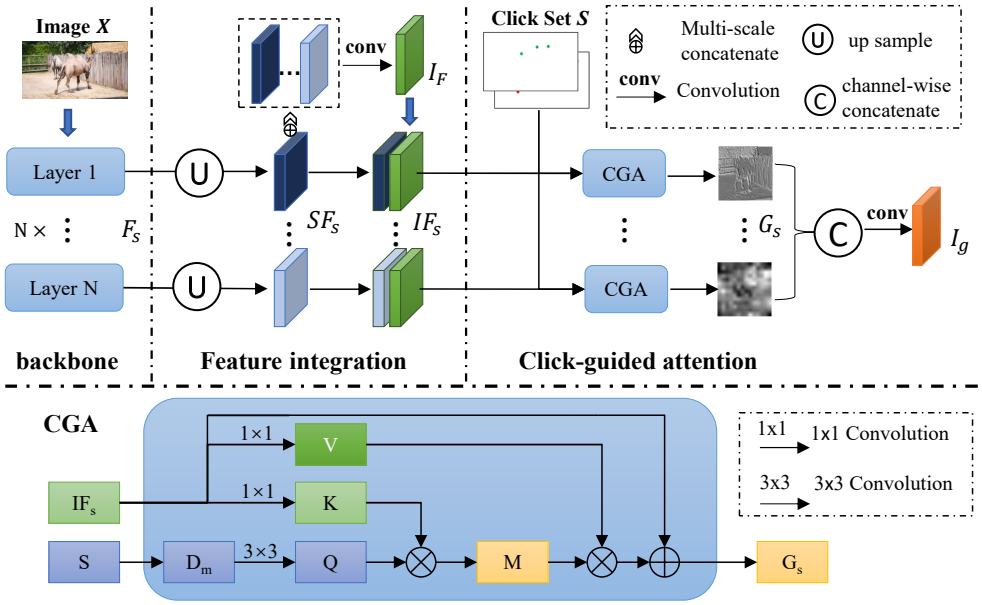

Figure 1. Overview of the proposed feature-aware attention module. Dense local features are firstly extracted from the backbone's N Layers, denoted as F_s. The multi-scale features are integrated to form a feature tensor IF_s.Then each of the feature maps at different scales is integrated with this new multi-scale feature map and fed into the guided attention modules to generate the attention features G_S. All the attention features are concatenated to generate the guidance map I_G.

The key to interactive image segmentation is to classify each pixel point of the image via making good use of prior knowledge (interaction information). Only position-guided with lower-level characteristics, Euclidean distance maps, Gaussian maps, and disks with modest radii provide semantic information about the target. As a result, it is not easy to thoroughly learn the relationship between interactive and semantic information.

In order to estimate the segmentation intention, we focus on multi-scale features and develop feature-aware attention. Specifically, the proposed pixel distance is a trainable variance due to the distance between the foreground and background pixel features being different in different images, so a fixed pixel distance could bring a certain error. The

trainable pixel distance allows the network to learn the difference and get the most suitable pixel distance. The pixel feature distance is defined as follows:

$$D_0 = \min_{q \in S^0} \left\| e_p - e_q \right\|_2 \tag{2}$$

$$D_1 = \min_{q \in S^1} \left\| e_p - e_q \right\|_2 \tag{3}$$

p is the pixel of the input image, and q is the point of the user click sets. D_1 is the positive click distance, D_0 is the negative click distance.

The feature distance is calculated by matching the interaction points with the feature maps acquired by the scale transformation module. The likelihood that the pixel belongs to the foreground will be higher if the distance between the pixel feature and the positive click is smaller; otherwise, the probability value representing this point as the background will be higher. And the probability that the pixel belongs to the foreground, denoted as P_{fd}, is defined as follows:

$$P_{fd} = \frac{1}{1 + e^{(D_1 - D_0)}} \tag{4}$$

It should be noted that the difference is calculated not only on a pixel point in the image representation but also on the region near the pixel point. Moreover, the region is determined in multiple scale dimensions.

Then we use attention mechanisms to learn user interaction intention from sets of interaction points and image features. Feature interaction IF_s and click map D_m are normalized and then transformed into three feature spaces V, K and Q to calculate the attention between multi-scale feature information and click sets. Here, the learned weight metrics of IF_s are implemented as 1×1 convolutions and the metrics of D_m are implemented as 3×3 convolution as in [37].

We argue that the feature-aware attention is the target scale more accurately, achieving significantly improved segmentation performance with limited interactions.

3.3. Refinement Network

Coarse segmentation results and a guidance feature map, which are calculated from coarse image segmentation and multi-scale interactive feature matching, are further utilized to generate the final segmentation mask. As show in Figure 2, the process can be formulated as:

$$\hat{M} = R(concat(I_C, I_G, I_F)) \tag{5}$$

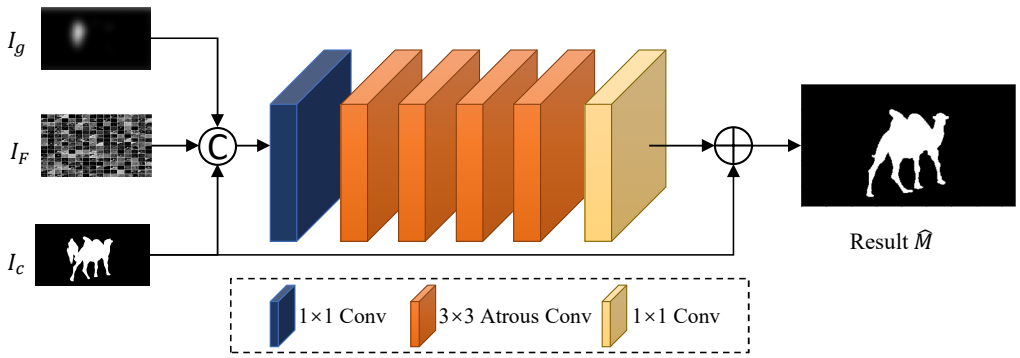

Figure 2. Overview of the proposed refinement module, which takes in concatenation of the coarse masks I_c, semantic feature I_F, and feature-aware guided map I_G and outputs precise mask result $\hat{M}$.

Specifically, we first concatenate coarse segmentation results, denoted as I_C and guidance feature map I_G in the channels. Additionally, the high-level feature map in the

segmentation networks contains essential semantic information about the image, so we further append the high-level feature map I_F to the refinement module input.

Our refinement module R contains 6 convolution layers (see Table 1), each followed by batch normalization and a ReLU. The motivation behind this model is that the striding makes it easy to capture long range information from the concatenating channels. Consequently, the refinement module R receives the $H \times W \times (c + 11)$ input and produces the better refined segmentation mask $\hat{M} \in \mathcal{R}^{H \times W \times 3}$. We impose the reconstruction loss to enforce the $\hat{M}$ to be close to the ground-truth mask M, which is denoted by:

$$L_{rec} = \left\| M - \hat{M} \right\|_1 \tag{6}$$

Table 1. Refinement module has 6 layers. Atrous convolution with rate > 1 is applied in the middle 4 layers.

Layer	**1**	**2**	**3**	**4**	**5**	**6**
Convolution	1×1	3×3	3×3	3×3	3×3	1×1
Dilation	1	2	4	8	16	1

4. Experiments

We extensively evaluated the proposed approach by conducting ablation studies for all sufficient parts of the method, exploring the convergence properties with increasing clicks, and comparing our method with current state-of-the-art works.

4.1. Datasets

We employed the Semantic Boundaries Dataset(SBD) [38] for training. It provides a wide-range domain and high-quality boundaries, including 8498 training and 2820 test images. It is a supplement to Pascal VOC that uses the same graphics but has more thorough annotations. SBD offers binary object segmentation masks for every object in the Pascal VOC [39] challenge's training and validation sets.

We evaluated the performance of our method on the following widely used datasets for interactive segmentation with instance-level annotations. The summary of datasets (including the publication years, class number, total instances number, total images number and the resolution) can be found in Table 2.

- GrabCut [3]: The dataset contains 50 images and the segmentation masks of the respective scene objects.
- DAVIS [40]: This dataset is introduced for the evaluation of video segmentation datasets. We use the subset of 345 randomly sampled frames of video sequences that are introduced in [41] for evaluation.
- Berkeley [42]: One hundred photos with a single foreground object make up this dataset. The photos in this dataset contain numerous characteristics that make image segmentation challenging, such as poor foreground-background contrast or a heavily textured backdrop.
- MS COCO [43]: With 80 distinct object categories, this dataset is a sizable image segmentation dataset. For evaluation, we sample 800 object instances from the validation set of COCO 2017 following the implemenation of [31]. Specifically, we sample 10 unique instances from each of the 80 categories in MS COCO.

Table 2. Summary of common datasets for interactive image segmentation task.

Dataset	Year	Classes	Instances	Images	Resolution
SBD [38]	2011	20	26,843	11,355	variable
GrabCut [3]	2004	-	one object each	50	variable
DAVIS [40]	2016	4	one object each	345	640 × 480
Berkeley [42]	2010	-	100	96	variable
MS COCO [43]	2014	80	800	800	variable

4.2. Experimental Settings

As the backbone of the coarse segmentation network for the model test in this study, we select two networks, VGG19 [44] and ResNet101 [45]. We utilized SBD as the training dataset, resizing each image to 320 × 320 pixels and augmenting it with random rotation, random flipping horizontally, and random Gaussian blur. We use the iterative sampling approach to create n consecutive clicks simultaneously for all of the images in each minibatch, with the batch size being set to 8 and $1 \leq n \leq 10$. The backbone network of the coarse segmentation model uses the weight parameters pre-trained on Imagenet [46].

In addition, the supervised outcomes of the interactive nonlocal block and the network's final outputs underwent the calculation and minimization of the binary cross entropy loss (BCEL).

We train the network for 50 epochs using the Adam optimizer with a weight decay of 10^{-5}, where the learning rates for the parameters of the pretraining model (i.e., the backbone) and for other components of the network are 10^{-4} and 10^{-3}, respectively.

On a configuration with a single NVIDIA Tesla V100-PCI-E-16G, an Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz CPU, and 16 GB RAM, we set up all of our tests using the PyTorch [47] framework. The time needed to execute inference for each image is roughly 0.05 seconds per click, which is sufficient for real-time segmentation.

4.3. Evaluation Metric

We performed the evaluation using the standard Number of Clicks (NoC) measure, reporting the number of clicks required to achieve the predefined Intersection over Union (IoU) threshold between predicted and ground truth masks. We denoted NoC with the IoU threshold set to 90% as NoC@90. To generate clicks during the evaluation procedure, we followed the strategy used in [6].

4.4. Comparison Results

We contrasted our approach with alternative approaches from both qualitative and quantitative perspectives. The compared techniques are mainly divided into two groups: deep interactive image segmentation algorithms, such as DOS [6], Regional image segmentation (RIS) [28], content-aware multi-level guidance (CAG) [7], latent diversity-based segmentation (LD) [41], backpropagating refinement scheme (BRS) [48], and feature-based backpropagation algorithms(F-BRS) [31]; Traditional interactive segmentation algorithms based on handcrafted features, such as GraphCut (GC) [3], random walker (RW) [4], and geodesic star convexity (GSC) [49].

The mNoC values needed by various algorithms on the five datasets are listed in Table 3, where mNoC@x is the quantity of clicks needed for the algorithm to reach x% mIoU. The number of clicks needed to get at the desired mIoU value is higher for any deep interactive image segmentation method than it is for any traditional interactive image segmentation methodology, suggesting that deep features are more generalizable than constructed features. On five datasets, the mNoC values of our method are at least 0.05 (COCO) lower than those of the competing methods except f-BRS on Berkeley (0.44 higher). Our strategy significantly improved DAVIS, as demonstrated. We discovered that the proposed interactive strategy significantly improved the performance of our network with different backbones.

Table 3. Segmentation results of different models on five datasets.

Method	GrabCut mNoc@90	Berkeley mNoc@90	SBD mNoc@85	DAVIS mNoc@85	COCO mNoc@85
GC [3]	10	14.22	13.6	15.13	18.53
RW [4]	13.77	14.02	12.22	16.71	14.10
GSC [49]	9.12	12.57	12.69	15.35	14.08
DOS [6]	6.08	8.65	9.22	9.03	8.31
LD [41]	4.79	-	7.41	5.05	-
RIS [28]	5.00	-	6.03	-	5.98
CAG [7]	3.58	5.6	-	-	5.4
BRS [48]	3.60	5.08	6.59	5.58	-
f-BRS [31]	2.98	4.34	5.06	5.04	-
Ours (VGG19)	2.89	5.16	5.32	4.58	5.79
Ours (ResNet101)	2.43	4.78	4.89	4.23	5.35

For each approach on the five datasets, curves of the mIoU values versus the number of clicks are shown in Figure 3. The area under the curve (AuC), which is standardized to [0, 1], is the value in the legend. From Figure 3, it is intuitively clear that our method is faster and more accurate than other ways. As seen in the GrabCut, SBD, and COCO curves, a higher AuC value indicated that our method is more accurate than the other methods for the majority of clicks.

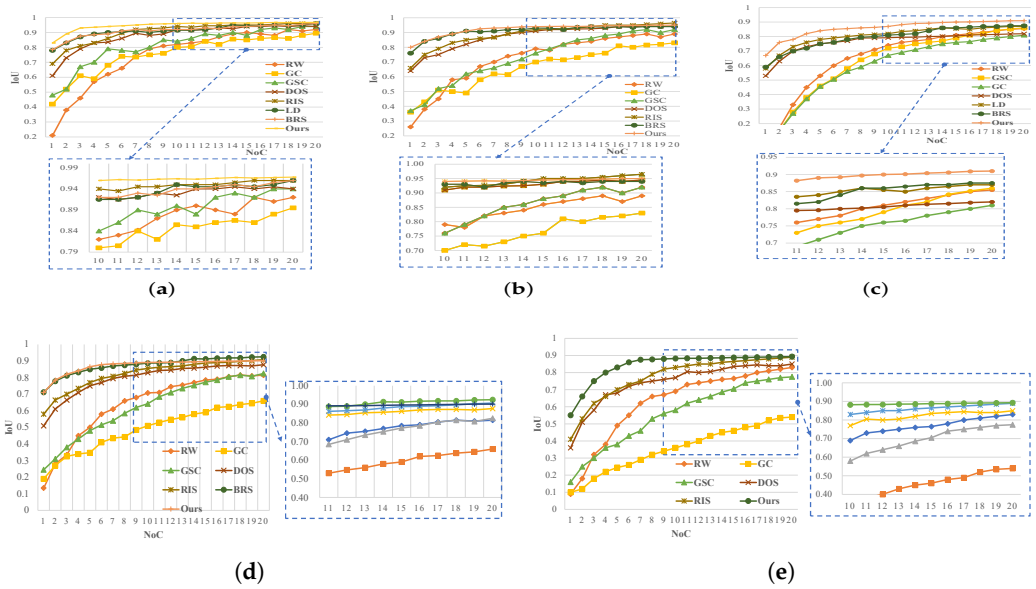

Figure 3. The curves of IoU based on the number of clicks on five datasets. (**a**) GrabCut; (**b**) DAVIS; (**c**) Berkeley; (**d**) SBD; (**e**) MS COCO.

The qualitative comparison results between our method and other methods are exhibited in Figure 4. To obtain similar accuracy, we drew lines as the interactions for the RW [4]. However, even if scribbled lines could provide more priors, the segmentation results obtained by using handcrafted features are still not as detailed and accurate as the results obtained by utilizing deep learning algorithms. We took LD [41] and fBRS [31] as examples of deep interactive image segmentation methods. We also observed that our method can produce superior regional segmentation results with fewer interactions. Other methods, on the other hand, necessitated more interactions, but their mIoU values are still inferior to ours.

Figure 4 illustrates the results of a qualitative comparison between our method and other methods. We create lines as the RW's interactions to get equivalent accuracy. Even

while scrawled lines could offer more priors, deep learning algorithms nevertheless produce segmentation results that are less precise and detailed than those produced by handwritten features. As illustrations of deep interactive image segmentation techniques, we use LD and BRS. We also notice that our approach requires fewer contacts while producing more substantial regional segmentation outcomes. While some approaches required more interactions, their mIoU values are still lower than ours. Furthermore, our method obtained outstanding performance for a specific object selected among several targets with only a few clicks, demonstrating the superiority of semantic guiding with interaction information.

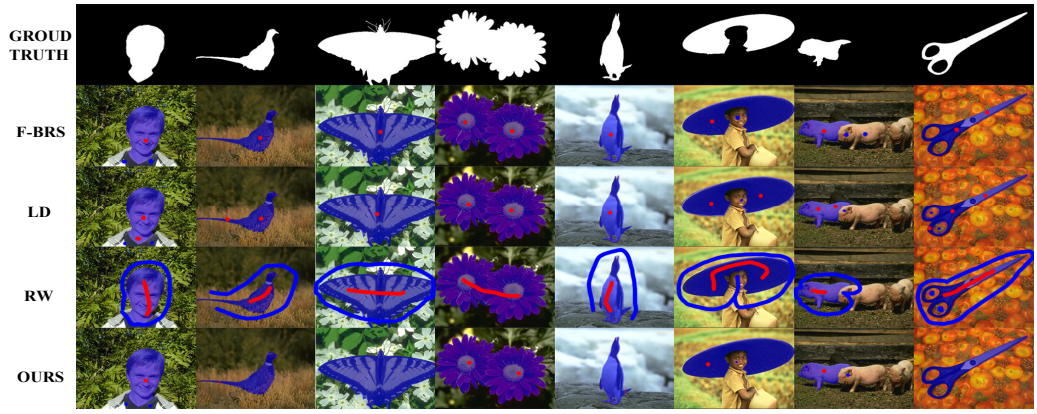

Figure 4. Comparison of segmentation results of different models.

4.5. Ablation Study

Network architecture ablations. We carried out a number of ablation experiments to confirm the viability of our approach. Table 4 shows the mNoC values that are produced by our algorithm on the various datasets when removing the refinement module. Our network also used three backbones to further investigate the flexibility of the feature-aware attention module.

As shown in Table 4, deleting the refining module gradually increases the amount of clicks needed to attain the desired mIoU. The largest difference in clicks between the results and the whole model is 0.74, indicating that the removal of the refining block has an effect on network performance.

Table 4. Evaluation for ablation experiments.

Settings	Backbone	GrabCut	Berkeley
Full	VGG19	2.89	5.16
	ResNet50	2.50	4.97
	ResNet101	2.43	4.78
w/o RF	VGG19	3.32	5.90
	ResNet50	3.08	5.63
	ResNet101	2.99	5.42

Non-interactive comparison. Different from other methods, our method is also capable of obtaining a segmentation result without providing interaction points, so as to reduce the complexity of user interaction to the clearly identifiable foreground targets. The comparison results are shown in Table 5, where the baseline column is the result of the coarse segmentation method without interaction points. The coarse segmentation result in the COCO dataset is poor at 0.54 due to many categories, and the best coarse segmentation result in GrabCut dataset is 0.81. Our method revises and improves the initial segmentation result. The Berkeley dataset and the COCO dataset showed a large increase on the first click, with an increase in IoU of 0.7. All datasets exceeded 0.8 on the third click, with GrabCut reaching 0.93.

Table 5. Segmentation accuracy is improved with the increase of interaction points.

Datasets	Baseline	1st Click	2nd Click	3rd Click
Grabcut	0.81	0.83	0.89	0.93
SBD	0.7	0.72	0.81	0.83
DAVIS	0.69	0.72	0.83	0.87
berkeley	0.73	0.8	0.84	0.87
COCO	0.54	0.61	0.72	0.81

The visualization effect of our method and the non-interactive segmentation method (Baseline) are compared, as shown in Figure 5. A total of 3 groups of images are taken out for experimentation, respectively, (a), (b) and (c). Each group includes four columns of images. The third column represents the detection effect of DeepLabV3+ of the non-interactive segmentation network, and the fourth column represents the segmentation effect of the interactive image segmentation method proposed in this paper. Our method is able to improve the accuracy of segmentation results by user clicks in comparison to non-interactive segmentation models. Depending on the user's intention, the segmentation result may be a combination of multiple targets or a partial region of a target, which is challenging to define in a supervised manner.

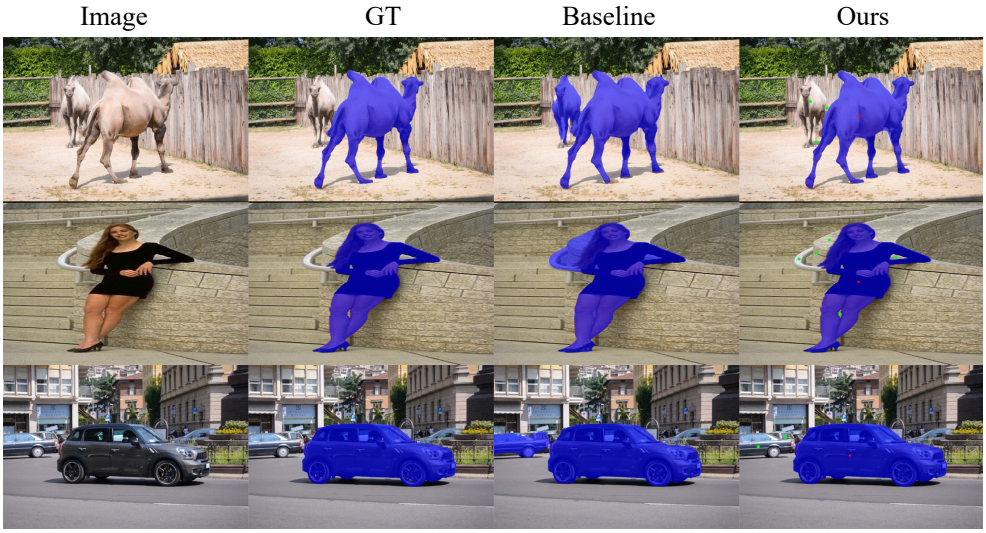

Figure 5. The visualization effect of our method and the non-interactive segmentation method.

5. Conclusions

In this paper, we propose a deep interactive image segmentation network, where a feature-aware attention module is utilized to integrate the human-click information with semantic features. The designed module is plug-and-play for most deep image segmentation networks, which prompts deep models to employ users' input information to refine the segmentation result. The experimental results prove our proposed module can improve the performance of image segmentation networks and refine the segmented objects. Our method obtains great performance for a specific object selected among several targets with only a few clicks. The performance of our network with various backbones is considerably enhanced since the proposed interactive method is flexible for most networks. Since our interaction module is considered as an optimization guide for segmentation results, the segmentation model can be executed normally without this guide information and gives a fairly good result. Therefore, for simple segmentation targets, our approach can effectively reduce unnecessary user interactions. In the future, we will further explore the interaction intention understanding model based on the attention mechanism, in order to be able to predict the user's next interaction intention based on the accurate simulation

of the user's segmentation intention. Further, we hope to extend our research area to 3D segmentation and explore faster and more effective interactive image segmentation models

Author Contributions: Conceptualization, X.B.; Data curation, B.H. and X.Y.; Writing—original draft, J.S.; Writing—review & editing, C.Y. All authors read and approved the final manuscript.

Funding: This research is supported by the National Key Research and Development Program of China (2019YFC0605301), National Science Foundation of China (61873299), Science Foundation of Guangdong (2021A1515012285), Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB (2021BH002), Fundamental Research Funds for the Central Universities of China (FRF-IDRY-20-038).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Boykov, Y.Y.; Jolly, M.P. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV) 2001, Vancouver, BC, Canada, 7–14 July 2001; Volume 1, pp. 105–112.
2. Freedman, D.; Zhang, T. Interactive graph cut based segmentation with shape priors. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 755–762.
3. Rother, C.; Kolmogorov, V.; Blake, A. "GrabCut" interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* **2004**, *23*, 309–314. [CrossRef]
4. Grady, L. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1768–1783. [PubMed]
5. Kass, M.; Witkin, A.; Terzopoulos, D. Snakes: Active contour models. *Int. J. Comput. Vis.* **1988**, *1*, 321–331. [CrossRef]
6. Xu, N.; Price, B.; Cohen, S.; Yang, J.; Huang, T.S. Deep interactive object selection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 373–381.
7. Majumder, S.; Yao, A. Content-aware multi-level guidance for interactive instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 11602–11611.
8. Ding, Z.; Wang, T.; Sun, Q.; Chen, F. Rethinking click embedding for deep interactive image segmentation. *IEEE Trans. Ind. Inform.* **2022**. [CrossRef]
9. Chen, S.; Tan, X.; Wang, B.; Hu, X. Reverse attention for salient object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018; pp. 234–250.
10. Li, K.; Wu, Z.; Peng, K.C.; Ernst, J.; Fu, Y. Tell me where to look: Guided attention inference network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 9215–9223.
11. Sinha, A.; Dolz, J. Multi-scale self-guided attention for medical image segmentation. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 121–130. [CrossRef] [PubMed]
12. Lin, Z.; Zhang, Z.; Chen, L.Z.; Cheng, M.M.; Lu, S.P. Interactive image segmentation with first click attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 13339–13348.
13. Chen, L.C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649.
14. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018; pp. 267–283.
15. Bai, X.; Sapiro, G. Geodesic matting: A framework for fast interactive image and video segmentation and matting. *Int. J. Comput. Vis.* **2009**, *82*, 113–132. [CrossRef]
16. Feng, J.; Price, B.; Cohen, S.; Chang, S.F. Interactive segmentation on rgbd images via cue selection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 156–164.
17. Chen, X.; Zhao, Z.; Zhang, Y.; Duan, M.; Qi, D.; Zhao, H. FocalClick: Towards Practical Interactive Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1300–1309.
18. Liu, Q.; Zheng, M.; Planche, B.; Karanam, S.; Chen, T.; Niethammer, M.; Wu, Z. PseudoClick: Interactive Image Segmentation with Click Imitation. *arXiv* **2022**, arXiv:2207.05282.
19. Kontogianni, T.; Celikkan, E.; Tang, S.; Schindler, K. Interactive Object Segmentation in 3D Point Clouds. *arXiv* **2022**, arXiv:2204.07183.
20. Wang, Y.; Deng, Z.; Hu, X.; Zhu, L.; Yang, X.; Xu, X.; Heng, P.A.; Ni, D. Deep attentional features for prostate segmentation in ultrasound. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018; Springer: Cham, Switzerland, 2018; pp. 523–530.

21. Zhou, Y.; Zhu, Y.; Ye, Q.; Qiu, Q.; Jiao, J. Weakly supervised instance segmentation using class peak response. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3791–3800.
22. Bearman, A.; Russakovsky, O.; Ferrari, V.; Fei-Fei, L. What's the point: Semantic segmentation with point supervision. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 549–565.
23. Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; Schiele, B. Simple does it: Weakly supervised instance and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 876–885.
24. Xu, N.; Price, B.; Cohen, S.; Yang, J.; Huang, T. Deep grabcut for object selection. *arXiv* **2017**, arXiv:1707.00243.
25. Dai, J.; He, K.; Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1635–1643.
26. Lin, D.; Dai, J.; Jia, J.; He, K.; Sun, J. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3159–3167.
27. Xu, C.; Dong, B.; Stier, N.; McCully, C.; Howell, D.A.; Sen, P.; Höllerer, T. Interactive Segmentation and Visualization for Tiny Objects in Multi-megapixel Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–23 June 2022; pp. 21447–21452.
28. Liew, J.; Wei, Y.; Xiong, W.; Ong, S.H.; Feng, J. Regional interactive image segmentation networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2746–2754.
29. Acuna, D.; Ling, H.; Kar, A.; Fidler, S. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 859–868.
30. Ling, H.; Gao, J.; Kar, A.; Chen, W.; Fidler, S. Fast interactive object annotation with curve-gcn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5257–5266.
31. Sofiiuk, K.; Petrov, I.; Barinova, O.; Konushin, A. f-brs: Rethinking backpropagating refinement for interactive segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 8623–8632.
32. Yang, Z.; Wei, Y.; Yang, Y. Collaborative video object segmentation by foreground-background integration. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 332–348.
33. Ding, H.; Cohen, S.; Price, B.; Jiang, X. Phraseclick: Toward achieving flexible interactive segmentation by phrase and click. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 417–435.
34. Kontogianni, T.; Gygli, M.; Uijlings, J.; Ferrari, V. Continuous adaptation for interactive object segmentation by learning from corrections. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 579–596.
35. Sofiiuk, K.; Petrov, I.A.; Konushin, A. Reviving iterative training with mask guidance for interactive segmentation. *arXiv* **2021**, arXiv:2102.06583.
36. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
37. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
38. Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; Malik, J. Semantic contours from inverse detectors. In Proceedings of the 2011 International Conference On Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 991–998.
39. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
40. Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 724–732.
41. Li, Z.; Chen, Q.; Koltun, V. Interactive image segmentation with latent diversity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 577–585.
42. McGuinness, K.; O'connor, N.E. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognit.* **2010**, *43*, 434–444. [CrossRef]
43. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.
44. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
46. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
47. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026–8037.
48. Jang, W.D.; Kim, C.S. Interactive image segmentation via backpropagating refinement scheme. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5297–5306.
49. Gulshan, V.; Rother, C.; Criminisi, A.; Blake, A.; Zisserman, A. Geodesic star convexity for interactive image segmentation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3129–3136.