Computational Visual Media
DOI 10.1007/s41095-xxx-xxxx-x Vol. x, No. x, mouth year, xx–xx

Article Type (Research)

Sequential Interactive Image Segmentation

Zheng Lin, Zhao Zhang, Zi-Yue Zhu, Deng-Ping Fan (✉), Xia-Lei Liu

Abstract Interactive image segmentation (IIS) has served as a vital technique in obtaining pixel-level annotations. In many cases, the target objects share similar semantics, such as semantic segmentation, instance segmentation, and human parsing tasks. The previous object representations, user interactions, and prediction masks can provide proper priors for the current annotation along with the annotating. However, IIS methods neglect the connection and these newly obtained cues. In this paper, we formulate a sequential interactive image segmentation (SIIS) task for minimizing user interaction cost on sequence data; meanwhile, we bring a practical solution with two pertinent designs. The first is a novel interaction mode. When annotating a new sample, our method can automatically propose an initial click proposal based on previous annotating. It dramatically helps reduce the interaction burden on users. The second is a densely online optimization strategy. To reduce the semantic gap in annotating specific targets, we further optimize the model with the dense supervision from previously labeled samples. Experiments demonstrate the effectiveness of our methods and the importance of the proposed SIIS task.

Keywords Interactive Segmentation, User Interaction, Object Segmentation.

1 Introduction

In the field of image editing and data annotating, it is crucial to obtain a high-quality pixel-level mask with minimal labor and time costs. Therefore, the community has invested a lot of attention in interactive image segmentation (IIS) technology, by which users participate in the segmentation process and provide interactive information iteratively to get good masks. To reduce the burden of users, researches on IIS mainly focus on two principles. The first is to carefully design the interaction mode [20, 27, 37, 39, 49, 52], so that users can provide more information with minimal interaction cost. The second is to carefully design the back-end algorithm [21, 29, 31, 34, 36, 38, 42] to maximize the use of information provided by users.

In practice, users often annotate multiple related images, such as images with the same categories in a semantic segmentation task and the same substructure in the human/scene parsing task. Meanwhile, in the inference stage, the IIS model will obtain an almost exact mask, unlike the uncertain one in most computer vision tasks. All the above observations enlighten us to think whether we can use the previous annotation information to assist the current one, as illustrated in Fig. 1. However, the idea is largely neglected by current IIS methods that deal with each image independently without considering the useful priors in previous annotations. A recent work [26] first tried to regard interaction segmentation as a sequence task and optimize parameters through user clicks. This

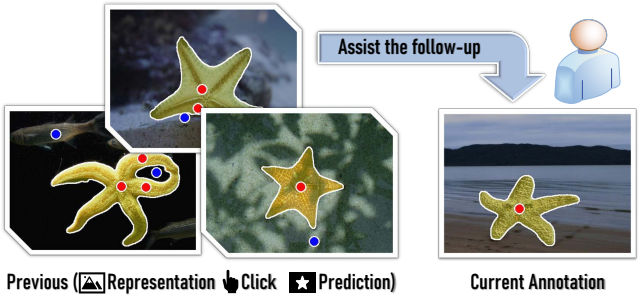


Fig. 1 Motivation of sequential interactive image segmentation (SIIS). In the process of annotating, the previous object representations, user interactions, and final prediction masks can provide assistance for the current interaction and prediction.

TKLNDST, College of Computer Science, Nankai University, Tianjin, China (E-mail: {frazer.linzheng, zzhang, zhuziyue, dengpingfan}@mail.nankai.edu.cn, xialei@nankai.edu.cn).
Corresponding author: Deng-Ping Fan.
Manuscript received: 2022-02-11; accepted: 20xx-xx-xx.

work only takes use of the click information and adopts the incomplete mask as regularization. We have taken a step further on its basis and propose exploring the interaction logic level in this sequence task. Moreover, we propose obtaining an accurate mask in the particular task to optimize the parameters better.

In this paper, focusing on the two investigating principles (interaction mode and back-end algorithm), we propose a systematic solution with two corresponding modules, *i.e.,* initial click proposal and online purification optimization, for the SIIS problem. In terms of interaction mode, we design a new interaction logic that greatly reduces user interaction burden with initial click proposal (ICP), as shown in Fig. 2 and Fig. 3. Specifically, ICP maintains a bank of initial click embeddings for a semantic target. When dealing with a new target, through similarity measurement, ICP will propose an initial click on the most likely position serving as a real interaction. If adopting the proposal, users can further interact for refinement; otherwise, directly correct the proposal with a new click. In terms of the back-end algorithm, we propose an online purification optimization (OPO) strategy for sequential interactive segmentation based on previous interactive results, as shown in Fig. 2. OPO keeps a group of parameters for each semantic target for narrowing the semantic gap. With increased user annotations, our pipeline will become more efficient for specific semantic targets.

Our contributions can be summarized as follows:

- The paper formulates the sequential interactive image segmentation (SIIS) task with two investigating principles, interaction mode and back-end algorithm.
- To improve the interaction efficiency, we design the initial click proposal (ICP) for SIIS to recommend the initial click instead of the real user input.
- To better utilize interaction cues, we raise the online purification optimization (OPO) to adapt the model to a specific semantic target using previous annotations.

2 Related Work

2.1 Image Interactive Segmentation

The field of interactive image segmentation has been explored for nearly two decades. Unlike the automatic segmentation like semantic segmentation [14], the present methods gradually consider human articipation [53]. Interactive segmentation is a typical example. It aims to segment the specified object through the user and mainly focuses on two study views. *1) interaction mode.* The research on interaction mode aims to make users provide the maximum information with the least interaction. Traditional methods mainly employ scribbles [3, 16, 23, 25, 45, 47] to denote background and foreground regions. In addition, many variants, such as cross-instance scribble [48], error-tolerant scribble [2], bounding box [41], and automatic border lasso [28, 40], are studied by the community. Recently, deep learning technology brings stronger perception, which makes lighter interaction modes possible. For example, user can directly click on the target object to select object and on background to erase error predictions [50]. Some other lighter modes, such as extreme points [39] and boundary clicks [20, 27], are investigated as well. As a novel way, the IOG [52] method, which combines the outside bounding box and an inside click, has also achieved excellent results. *2) back-end algorithm.* The research on algorithm logic aims to maximize the use of interactive information provided by users for accurate prediction. Traditional methods are mainly based on color features [5, 6, 15, 24]. Recently, methods based on convolutional neural network [50], recurrent neural network [1, 7], graph convolutional network [35], reinforcement learning [43] spring up in IIS task. The various architectures are also mentioned, *e.g.,* regional refinement block [30] and two-stream fusion [19]. Some researches [29, 31] try to solve the ambiguity in interactive segmentation. Besides, some important cues about interactive segmentation also receive attention, *e.g.,* training strategy [36], interaction map [4], and user intent [33, 34, 38].

2.2 Interactive Segmentation with Online Learning

Online learning has been used in many segmentation-related works [8, 51]. For the IIS task, user interactions can be used as a reference for prediction and as a supervision signal for fine-tuning models. BRS [21] first takes the idea into individual interactive segmentation. According to its assumption, the confidence value of the model's prediction at the user's click may not be high enough. Fortunately, the uncertainty brings the possibility for model optimization. BRS takes the mispredicted clicked pixels as punishment to fine-tune the input distance map, which is more specialized for the target object and ensures that the prediction can cover the interaction points well. f-BRS [42] improves the BRS by back-propagating the part of the model instead of the whole one. In this way, it makes the

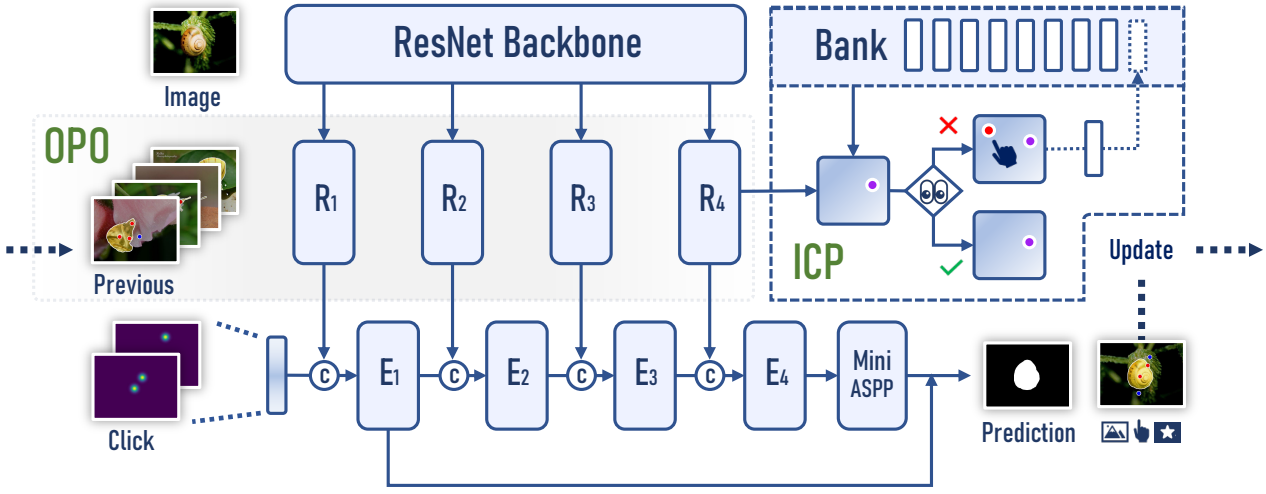


Fig. 2 The overall pipeline of the proposed sequential interactive segmentation method. ICP is the initial click proposal, detailed in Sec. 3.2. It aims to propose an initial click proposal for reducing user burden automatically. OPO is the online purification optimization, detailed in Sec. 3.3. It continuously optimizes specialized parameters for the semantic target according to the previous annotation masks to ensure a semantic adaptive proposal space and more efficient segmentation. The symbol © means a two convolutional layers. The red point "•" and blue point "•" mean the click in foreground and background, and the purple one "•" indicates the recommended click.

online training efficient again. Recently, Kontogianni *et al.* [26] preliminarily attempts to introduce the clicked position supervision to image sequence segmentation, which achieves promising results. It employs sparse supervision with only several positive and negative clicks and uses these incomplete masks for the regularization constraint in sparse optimization. However, the particularity of interactive segmentation lies in that the user will get complete masks after interacting with previous images. Our method is a step further and directly uses all previous predictions as dense supervision instead of several user clicks. Based on this, we propose OPO, which utilizes the previous final masks to assist the subsequent pictures.

3 Proposed Method

In this section, we introduce the proposed method in three parts. In Sec. 3.1, we introduce our modified DeepLab v3+ [9], which is specially designed for the sequential interactive segmentation. In Sec. 3.2, we describe the Initial Click Proposal (ICP), which provides the users the initial click proposal based on previous initial clicks in the image sequence. In Sec. 3.3, we propose the Online Purification Optimization (OPO), which optimizes the purification parameters in the modified DeepLab v3+ with online training based on the previous annotation masks.

3.1 Network Architecture

Interactive segmentation is an essentially particular task of object segmentation. For most previous click-based interactive segmentation works [26, 31, 34, 36, 42], they usually adopt DeepLab v3+ [9] as the segmentation network with 5 channels (the RGB image + positive/negative click maps) as input. The network architecture works well most time. However, there are two problems for the original architecture in the sequential interactive segmentation. Firstly, we need to utilize the feature correlation within images of a specific category. The traditional 5-channel-input does not meet our requirements because the annotation input will disturb the semantic feature. In other words, the correlation between interaction points will be significantly enhanced, and the semantic similarity will dramatically disappear. Secondly, in sequential interactive segmentation, the parameters need to be optimized continuously. We need to save the specific parameters for each category. Optimizing global parameters will significantly increase the burden of hardware storage (such as memory or video memory) in the real application environment. For the above reasons, we split the original architecture into two parts, the feature extraction part, and the interactive segmentation part, as shown in Fig. 2. We call it modified DeepLab v3+. For feature extraction, we also adopt ResNet-101 [18] with the output stride of

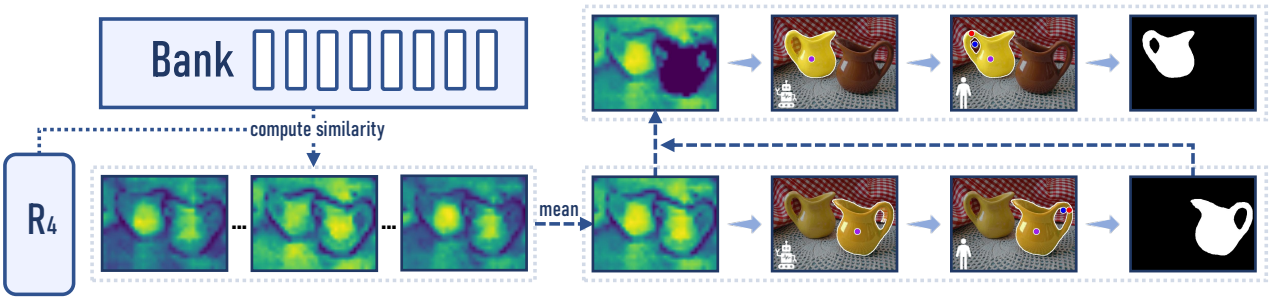


Fig. 3 Details of the initial click proposal (ICP). ICP proposes an initial click based on the confidence map, which is a mean of multiple similarity maps measured by the image feature and all embeddings in the bank from previous initial clicks. The proposal will act as a real click from users and conduct an initial segmentation. When the proposal is correct, the user can further refine it by providing more interactions. In the multi-target scene, the previous masks will be erased from the confidence map so that ICP can propose a new initial click for the next target.

16 as backbone. The features of the last four layers are with channels of $\{256, 512, 1024, 2048\}$. These features are fed into a simple purification module containing a few 1×1 convolutions to reduce and purify the feature for the specific category. The channel-reduced features are defined as $\{\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \mathbf{R}_4\}$, with channels of $\{32, 64, 128, 256\}$, which is only $\frac{1}{8}$ of the original ones. With the annotation guidance maps (click maps) and these channel-reduced features, we conduct a Mini-DeepLab v3+ module, whose architecture is similar to the original one. For the encoder module, the input is the annotation guidance map $\mathbf{E}_0$, which is two Gaussian maps based on points. The input features are gradually combined with channel-reduced features, which is formulated as:

$$\mathbf{E}_i = \mathcal{C}\left(\mathbf{R}_i \oplus \mathcal{D}\left(\mathbf{E}_{i-1}\right)\right), i \in \{1, \ldots, 4\}, \tag{1}$$

where $\mathcal{D}(\cdot)$ means down-sampling, $\oplus$ means feature concatenation, and the $\mathcal{C}$ means two convolutional layers with the kernel size as 3×3. For the output $\mathbf{E}_4$ of the mini-encoder module, it will be fed into a Mini-ASPP module. Different from the original one, the down-sampling is only $\frac{1}{4}$ instead of $\frac{1}{8}$. The output with 64 channels of the Mini-ASPP will finally be concatenated by the $\mathbf{E}_1$ and convoluted to the final predictions. Although the modified DeepLab v3+ has some additional parts based on the original version, this network is lighter than the original one due to reduced channels. Because of the separation of feature extraction and interactive segmentation, more sequential operations can be implemented, and we can better explore the sequential interactive segmentation, like the ICP and OPO.

3.2 Initial Click Proposal

In sequential interactive segmentation, how to reduce the burden of users in interaction logic is an important problem. We propose the initial click proposal (ICP) to maintain a click bank that records the feature vectors on the pixels where previous initial clicks are located for each category. It is initialized as an empty bank. When the user intends to do interactive segmentation for a specific category, the similarity between the feature vectors of all pixels and those of previous initial clicks in this bank will be calculated. For initial segmentation, the most similar pixel will be marked as a initial click proposal. If the user is not satisfied with the initial click proposal or the proposal is with the mistake, the user can select the initial point manually and continue the following interactive segmentation. After the user selects the initial click manually or adopts the initial click proposal, which is correct, the corresponding feature vector will be stored in the click bank.

How to choose the initial click proposal? We utilize the cosine similarity (shown as ϕ in Equ. (2)) to find the recommended point. Suppose that the target image is $\mathcal{T}$ and the image features we choose are defined as $\mathbf{F}$. $\mathbf{F}(p)$ means the feature vector in the corresponding pixel p. Then calculation for the recommended point $\hat{p}$ is formulated as:

$$\hat{p} = \underset{p_n \in (\mathcal{T} - \mathcal{I})}{\arg\max} \frac{\sum_{i=1}^{n-1} \phi\left(\mathbf{F}(p_n), \mathbf{F}(p_i)\right)}{n-1}, \tag{2}$$

$$\phi\left(\mathbf{a}, \mathbf{b}\right) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}, \tag{3}$$

where p_n mean the recommended point, $p_1 \ldots p_{n-1}$ means the previous initial clicks, and the $\mathcal{I}$ means the ignore mask, which is initially set to $\varnothing$ and used for the segmentation of multiple instances. In practice, we can do a mean filtering on the confidence map before selecting the maximum point to prevent the occasional extreme points. We choose the last layer $\mathbf{R}_4$ to generate channel-reduced features for the initial click

proposal. It will provide more semantic information. The ablation study about choosing features of different layers is also shown in Tab. 3.

Initial click proposal, in real scenarios, can be implemented in different interactions. For example, for the recommended point that users are not satisfied with, users can re-select the initial click through the middle mouse button, and continue to use the left and right mouse buttons for interactive segmentation. For multi-semantic annotation, ICP will maintain a click bank of initial clicks for each category. When users select the semantic tags to be annotated, the corresponding initial click proposal will be generated and shown. If multiple instances of the same categories are in an image, ICP can also recommend numerous instances. As shown in the Fig. 3, every time the user has finished an annotation $\mathcal{A}$ for instance, the ignore mask $\mathcal{I}$ will be updated to $\mathcal{I} \cup \mathcal{A}$. The following recommended points will not be repeated with the previous ones.

3.3 Online Purification Optimization

As the first exploration of online training based on previous annotation results, we adopt the most concise training settings, which is similar to that when training the baseline model. The core difference is that for the training of the baseline model, the number of foreground and background points is within $[1,\ 10]$ and $[0,\ 10]$, respectively, to simulate user operations better. For the online training, they are within $[1,\ 5]$ and $[0,\ 5]$ to reduce the impact of interaction points. The batch size is set to 8, and we train for four iterations after a complete interactive segmentation. The images and instances are selected from all the annotated ones. In the process of online training, we also use stochastic gradient descent for the optimization, and the learning rate is fixed to 5×10^{-3}. Different from other online learning methods, we only optimize a small set of the parameters in the purification module, which is called online purification optimization, as shown in Fig. 2. The purification module is composed of multiple 1×1 convolutions, and its parameters play the role of extracting the original features. For the sequential interactive segmentation, the features that each category depends on are often different. Through this purification module, the features are regrouped and integrated, and the parameters of the purification module are changed through online learning. It is called a purification module because it is conducive to extracting parameters for specific categories from the original complex image features. Before the module, the features are impure because they represent all kinds of image features. After this module, the reduced feature can better represent this specific kind of object. As shown in Fig. 6, the segmentation performance of the initial clicks reflects the parameters that have fit the corresponding characteristics of the category.

Every time a user completes an instance segmentation, a round of online learning of that category will be performed in real scenarios. For each category, the segmentation system will save a set of parameters of the purification module. Because the parameters are extremely small, as shown in Tab. 7, the storage space required is small, and it will not cause a burden. When users select the semantic tags to be annotated, the corresponding parameters will be adopted. The task using online learning usually faces the problem that it is challenging to run in real-time. But the task of interactive segmentation does not have this kind of problem at all. In most cases, the time used for thinking is much longer than that used for computer processing during interactive segmentation. So as long as you take a specific interval δ, and use the model trained by the first $(n - \delta)$ samples when segmenting the n th object, you can fully achieve the effect of simultaneous user interactions. Generally, as long as δ is greater than 1, it can meet the real-time requirement.

4 Experiments

4.1 Settings

Datasets. *Augmented PASCAL VOC* [11, 17], a widely used semantic segmentation dataset across 20 categories. Like some previous works, we use the training set (25832 instances) for training and the validation set (3427 instances) for testing. *COCO* [32], a large-scale dataset provided by Microsoft. We take three settings for testing. For comparison with individual interactive segmentation, we adopt the same setting in [34]. For comparison with sequential interactive segmentation, we adopt the same setting in [26], including COCO (Unseen 6k), COCO (Donut, Bench, Umbrella, Bed). *CoSOD3k* [12, 13] is a dataset for co-salient object detection which has abundant categories. We selected the whole set with 4874 instances across 160 categories for the test. *CoCA* [54] is another dataset for co-salient object detection which has special categories. These categories are not typical and appear in other datasets, ideal for studying independent semantic tasks. We selected the whole set with 2143 instances across 80 special categories

#	ICP	OPO	PASCAL @85%	PASCAL @90%	COCO @85%	COCO @90%	CoSOD3k @85%	CoSOD3k @90%	CoCA @85%	CoCA @90%	Fashionpedia @85%	Fashionpedia @90%	LeedsButterfly @85%	LeedsButterfly @90%
(a)			3.18	4.07	5.81	8.25	4.86	7.24	6.81	9.61	13.87	16.47	2.16	2.89
(b)	✓		2.82	3.70	5.40	7.85	4.34	6.70	6.50	9.32	13.68	16.30	1.37	2.07
(c)		✓	3.11	3.98	5.36	7.88	4.47	6.82	5.88	8.76	11.34	14.29	1.25	1.48
(d)	✓	✓	2.74	3.60	4.98	7.51	3.93	6.29	5.57	8.48	11.17	14.14	0.29	0.52

Tab. 1 Core ablation study about NoC metric (@85% and @90%) on all six datasets with initial click proposal (ICP) and online purification optimization (OPO) proposed in our pipeline. The lower value means the better performance.

for testing. *Fashionpedia* [22] is a dataset about fashion images. We use the 8781 part masks across 46 categories in the validation set for testing. We adopt it for exploring segmenting object parts in sequential interactive segmentation. *LeedsButterfly* [46], a dataset that contains 832 images of butterflies.

Metrics. For evaluating the interactive segmentation, we take the same metric like that in most interactive segmentation works. A robot user is adopted, selecting the next point in the center of the largest error region. The mean Number of Clicks (NoC) indicates the average number of clicks in the interactive process until each instance reaches the specified Intersection over Union (IoU) score (represented as @XX%). The lower value means the better performance. It is worth mentioning that the value of NoC when using online training is the mean value of 5 experiments.

Implementation Details. We take ResNet-101 [18] pre-trained on ImageNet [10] as a backbone. We set the batch size to 8 and train for 30 epochs. We use the binary cross entropy loss function in baseline training. We adopt the exponential learning rate decay strategy with the initial learning rate of 7×10^{-3} and gamma of 0.95 for each epoch. For parameters optimization, we take stochastic gradient descent with a momentum of 0.9 and weight decay of 5×10^{-4}. We crop and resize images to 384×384 with random flip and random clip augmentation. For annotation simulation, we use a similar strategy in [34] and take the same iterative training strategy in [36]. All the experiments are implemented with the PyTorch [44] framework and run on a single NVIDIA Titan XP GPU.

4.2 Ablation Study & Discussion

We have conducted sufficient ablation experiments with our core issues on the six selected datasets. Tab. 1 shows the NoC metric on different target thresholds on all datasets. Observing from the overall data, no matter which dataset, no matter which target threshold (@85% or 90%), our ICP and OPO can play a role in improving the performance, which fully proves the effectiveness of the proposed methods. This section will analyze the effects of the two core modules, ICP and OPO, on different types of datasets, taking the data with @85% as an example.

For the validation dataset in PASCAL, whose categories are the same as the training set, the parameters in the purification module have been fully fitted to these seen categories. We can find that the ICP is highly effective, with 11.37% improvement. However, the improvement is quite limited with OPO. This is also reasonable because the fully fitted features are more suitable for providing the click proposal, while it is difficult to improve these parameters through online training with limited samples with seen categories. COCO and CoSOD3k have a small number of categories that overlap with the training set, and the classes in CoCA are unique. These three datasets are rich in categories, but the number of each class is limited, The improvement is 7.04%, 10.61%, 4.67% with ICP, and 7.70%, 7.97%, 13.66% with OPO, respectively. Combining ICP and OPO, the performance improvement can reach 14.32%, 19.09%, and 18.21%. These data fully reflect that our method can bring noticeable improvement even if there are few samples for each category. Fashionpedia is the most difficult because the segmentation targets are fashion parts, while the training samples are all instances, especially a whole human body. The improvement is only 1.33% with ICP, but it can achieve 18.25% with OPO. We speculate that this phenomenon is that the neural network has a high probability of treating the human body as a unified category so that the feature similarity will cause mismatches for these partial clothing, accessories, etc. But parameter optimization is still helpful for improving the performance of such part objects. For the LeedsButterfly, which only contains several butterflies, the improvement brought by ICP and OPO is significant. The OPO brings 42.04% improvement compared to baseline. For the ICP equipped with the baseline, the improvement

Name	Params (M)	FLOPs (G)	SPC (s)
DeepLab v3+	59.345	50.149	0.024
Ours	45.743	32.364	0.016

Tab. 2 Network comparison between our modified DeepLab v3+ and the original version. (See Q1)

Dataset	R1	R2	R3	R4	R-MS
CoSOD3k	7.00	7.04	6.84	6.70	6.64
CoCA	9.43	9.46	9.37	9.32	9.23

Tab. 3 The NoC (@90%) when using features in different layers in the backbone network for ICP module. "R-MS" means to use multi-scale features.The lower value means the better performance. (See Q2)

#	Setting	C1	C2	C3	C4	All
(a)	Provide Initial Click	1.62	1.64	1.28	1.48	1.51
(b)	Judge Positive Sample	0.48	0.46	0.47	0.43	0.46
	Judge Negative Sample	0.58	0.56	0.54	0.55	0.56

Tab. 4 The user study of the ICP module. (See Q3)

Dataset	Ours	Full	Foreground
CoSOD3k	4.34 / 6.70	4.77 / 7.15	4.50 / 6.86
CoCA	6.50 / 9.32	6.77 / 9.56	6.55 / 9.38

Tab. 5 The NoC (@85%/@90%) when adopting a random click from full image or the foreground in the ICP module.The lower value means the better performance. (See Q4)

reaches 36.62%, and the ΔNoC achieves 0.79. After adding the OPO, the advance of ICP is more significant, with 76.52%, and the ΔNoC achieves 0.96, whose maximum is 1.0. It reflects that the features obtained after parameters optimization are more suitable for the initial click proposal. In other words, the OPO can assist ICP to get better performance. The NoC metric is only 0.29 with ICP and OPO. That means that this framework can complete satisfactory annotations under approximately semi-automatic interaction, which significantly reduces the burden on annotators.

Here are some additional ablation experiments and discussions with some questions:

Q1: What is the difference between this network and the original DeepLab v3+? Tab. 2 shows the primary metrics of the two network architectures, including the number of parameters (Params), the floating-point operations per second (FLOPs), and seconds per click (SPC). We can see that the modified one is relatively lighter. It is worth mentioning that this does not mean that our network is better than the original one, but because of the unique design of our ICP and OPO, we have to adopt such a change.

Q2: How about selecting features in another layer of the backbone network for ICP module? Tab. 3 explores this situation when using features in other layers of the purification module. We can find that the performance is best with the $\mathbf{R}_4$ features, and the next is $\mathbf{R}_3$, $\mathbf{R}_1$, $\mathbf{R}_2$. This is consistent with our intuition; using the highest-level feature information is more conducive to the initial click proposal. We also carry out an additional experiment with multi-scale features for initial click proposal. We can find that the performance can be further improved.

Q3: Can the ICP really save time for users to reduce interaction burdon? As shown in Tab. 4, we conduct a user study for the proposed ICP module to verify its effectiveness. 40 images with 4 categories in COCO [26, 32] dataset are selected as the study set. Half of the data provide the correct recommendation point for each category, and the other half is the opposite. We invited 20 volunteers for our user study. They were asked to complete two tests. (a) One is to find and click on the corresponding category of objects in the provided random pictures. (b) The other is to judge whether the recommendation point is correct with provided random pictures and corresponding recommendation point. From the table, We can find that the time to judge the wrong recommendation point is more than to judge the correct one. Both are less than the time to click on the object directly. This reflects that the ICP module can save users' interaction time in practical applications.

Q4: How does the quality of initial clicks affect the final segmentation? We carry out additional experiments for initial click proposals with random clicks. We choose two random strategies for comparison. One is to select the random click from the full image. Another is to replace the initial click proposal with a random point on the object of this category when the proposal is correct. Results of the mean value from five experiments are shown in Tab. 5. We find that the performance will decrease if the random click is selected in the full image or the foreground. Because the ICP module is mainly used to locate objects of this category, the performance degradation is relatively minor when selecting from the foreground compared to the full image.

Method	PASCAL @85%	COCO @85%	CoSOD3k @90%	CoCA @90%	Fashionpedia @85%	LeedsButterfly @90%
CVPR - DOS [50]	6.88	9.07	11.04	13.04	16.27	5.32
ICCV - RIS [30]	5.12	N/A	N/A	N/A	N/A	N/A
CVPR - LD [29]	N/A	7.86	8.73	11.94	16.41	3.66
BMVC - ITIS [36]	3.80	6.51	8.67	11.42	16.77	3.43
ICCV - MS [31]	3.88	N/A	N/A	N/A	N/A	N/A
CVPR - BRS [21]	N/A	5.16	N/A	N/A	N/A	N/A
CVPR - CMG [38]	3.62	5.90	N/A	N/A	N/A	N/A
CVPR - FCA [34]	2.98	5.28	6.31	9.51	13.31	2.44
CVPR - f-BRS [42]	N/A	5.75	6.93	9.46	14.40	2.86
Ours	2.74	4.98	6.29	8.48	11.17	0.52

Tab. 6 Comparison of the NoC metric between our solution and other methods. The lower value means the better performance.

Name	Params (M)	SPB (s)	CoSOD3k	CoCA
Purification	0.697	0.098	6.82	8.76
Global	45.743	0.277	6.67	8.38

Tab. 7 Comparison between optimizing purification parameters and global parameters. The last two columns is NoC@90%. The lower value means the better performance.. (See Q5)

Method	Donut	Bench	Umbrella	Bed	Unseen 6k
CA [26]	6.50	13.30	10.20	5.00	9.30
Ours	5.65	12.56	9.63	4.56	9.18

Tab. 8 Comparison of the NoC@85% metric between our solution for sequential interactive segmentation with another sequence-based work [26] on the same five sets of COCO dataset. The lower value means the better performance.

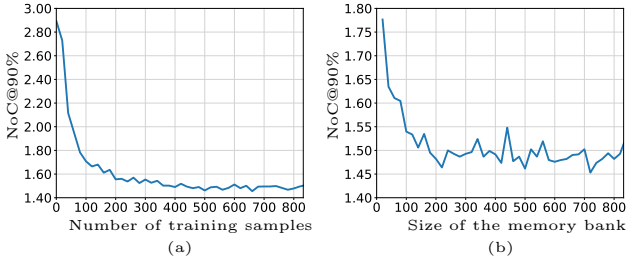


Fig. 4 Performance trends when increasing the training data (a) and the memory size of online training (b). The lower value means the better performance. (See Q6, Q7)

Q5: Why do we only optimize the purification parameters? Tab. 7 compares optimizing purification parameters with global parameters. We find that the purification module's parameter amount (Params) is tiny compared to the whole network, even less than 2% of it. However, the performance gap is not significant. For sequential interactive segmentation, it is necessary to save unique parameters for each category. Such a small parameter amount is undoubtedly appropriate. The optimization speed, indicated by Second Per Batch, brought by a small number of parameters is faster, which is also helpful to the task.

Q6: Will the performance improves with the increase of online training data? Fig. 4 (a) illustrates the NoC metric trends on LeedsButterfly when stopping online training after a specified number (abscissa) of samples. With the increase of online training data, the performance is continuously improving, and it reaches a stable state later.

Q7: Should the method require to access the whole training dataset during online learning? The method does not need to access the whole training dataset. We can set a memory bank, and the training samples will always be chosen in this bank. Fig. 4 (b) illustrates the NoC metric trends on LeedsButterfly when adopting the different sizes of memory banks. We can see that the performance will be affected if the bank size is too small. However, after a specific size, it is stable and suitable for practical applications.

4.3 Comparison with State-of-the-Arts

In Tab. 6 and Fig. 6, we compare the quantitative and qualitative results between our method and other state-of-the-arts. Here, we will further elaborate on the inference process of "B+OPO". When the user labels an image, the OPO module will switch to the specific parameters of the working category. Users will constantly provide a foreground and background clicks for annotation. After each interaction, the information of the image and clicks will be input into our network to generate the corresponding mask. According to the generated mask, users can continue to add the

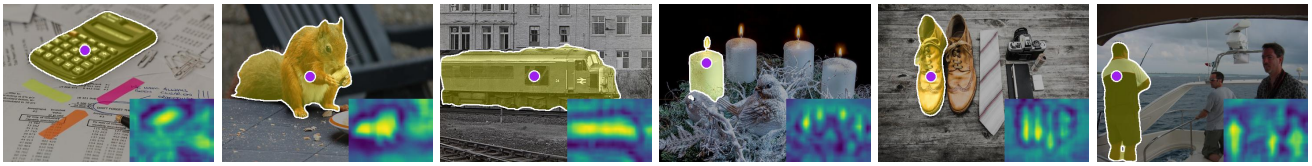


Fig. 5 The qualitative results of the ICP with the click proposal, confidence map, and prediction.

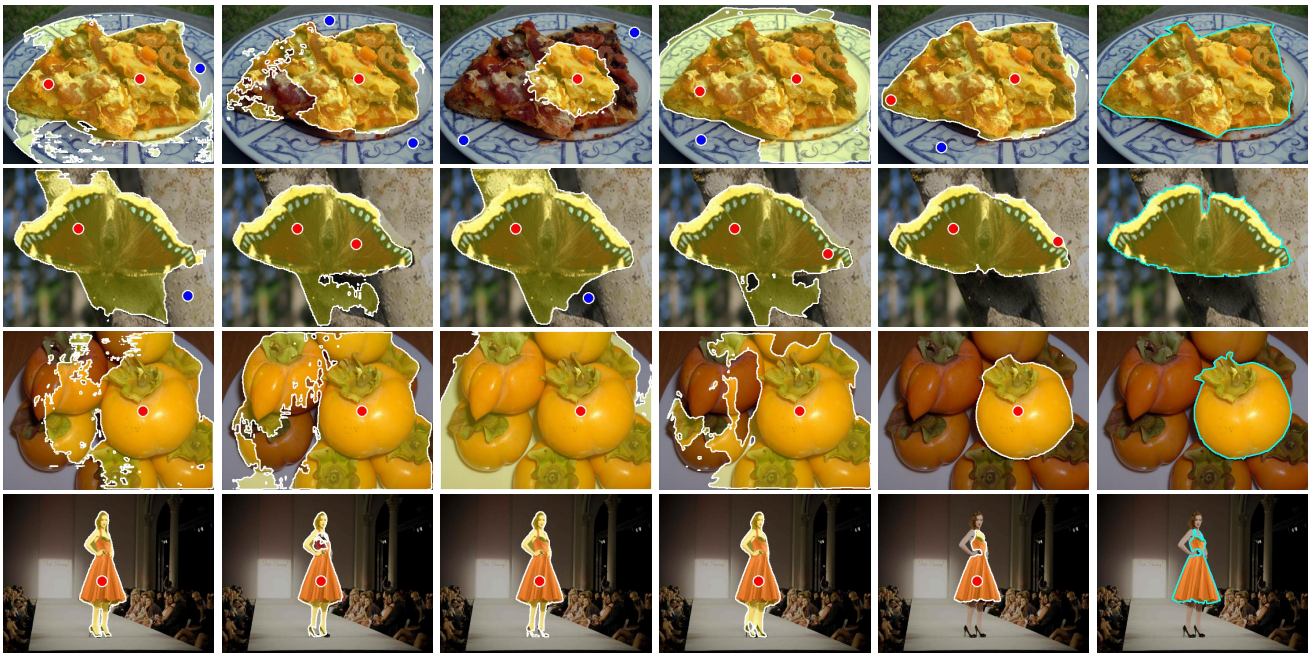


Fig. 6 The qualitative results of the OPO compared to other methods and the baseline (B).

next click for refinement until the mask meets users' needs. The online training phase will be carried out when a new image is annotated completely. The newly annotated and some previous images will be randomly chosen for an online training step. As the previously annotated masks have satisfied the user, they can naturally serve as ground truth labels for supervision. We random simulate clicks in these images according to corresponding annotated masks. Then images and these clicks will be fed into the network like the standard training phase and generate the predictions for optimization. We employ binary cross-entropy loss between predictions and previously annotated masks. It is worth mentioning that, during online training, only the parameters of the OPO module will be optimized through stochastic gradient descent, while the parameters of other modules are fixed. These parameters will become more suitable for this category. As the process continues, the performance of labeling certain objects will become better and better.

Comparison with Individual IIS. Tab. 6 shows the NoC metric of IIS methods in these datasets with rich categories. Due to the different emphasis of individual and sequential interactive segmentation, we provide the performance only for an intuitive comparison. These methods are carefully designed and focus on IIS. Our method mainly addresses the problem of sequential interactive segmentation and makes some modifications to the basic network. Its performance is comparable to or even surpasses these cutting-edge methods. It reflects that regarding interactive segmentation as a sequential process is beneficial.

Comparison with Sequential IIS. As shown in Tab. 8, we also compare the only state-of-the-art method [26] related to sequential interactive segmentation. This method utilizes correction clicks for online training, and our approach goes a step further on it and can achieve better results. Because our method focuses more on semantic objects, we compare most semantic data for our experiments. Through more

dense supervision, we make the model more sensitive to this category of objects, so as to perform better.

Qualitative Results. Fig. 5 and Fig. 6 show some results of the proposed methods. From Fig. 5, we can find that the initial click proposals are precisely located in the interior of objects with different categories, such as animals and vehicles, which can reduce the interaction burden on users. Fig. 6 shows the segmentation results compared to the baseline and other methods with 1, 2, and 3 clicks. With the optimized parameters, our method can get more accurate results at the same number of interaction points, whether it's an instance (the first three rows) or the object part (the fourth row). This can also be beneficial for users to segment the target from the scene with multiple instances (3rd row).

5 Conclusion

In this paper, we formulate the task of sequential image interactive segmentation (SIIS). To solve SIIS, we systematically explore it from the views of interaction mode and back-end algorithm. Specifically, for the interaction logic, we design the initial click proposal (ICP), which utilizes the previous semantic embeddings of the target object to recommend an initial click proposal serving as the real-input one for the current annotation. We put forward the online purification optimization (OPO) for the algorithm logic, which fine-tunes model parameters to a target category using previous accurate annotations. Extensive experiments prove the importance of sequential image interactive segmentation and the effectiveness of our method.

6 Conflicts of Interests

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

[1] D. Acuna, H. Ling, A. Kar, and S. Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 859–868, 2018.

[2] J. Bai and X. Wu. Error-tolerant scribbles based interactive image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 392–399, 2014.

[3] X. Bai and G. Sapiro. Geodesic matting: A framework for fast interactive image and video segmentation and matting. *Int. J. Comput. Vis.*, 82(2):113–132, 2009.

[4] R. Benenson, S. Popov, and V. Ferrari. Large-scale interactive object segmentation with human annotators. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11700–11709, 2019.

[5] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE T. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, 2004.

[6] Y. Y. Boykov and M.-P. Jolly. Interactive grrrraph cuts for optimal boundary & region segmentation of objects in nd images. In *Int. Conf. Comput. Vis.*, pages 105–112, 2001.

[7] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler. Annotating object instances with a polygon-rnn. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5230–5238, 2017.

[8] F. Cermelli, M. Mancini, S. R. Bulo, E. Ricci, and B. Caputo. Modeling the background for incremental learning in semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9233–9242, 2020.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis.*, pages 801–818, 2018.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009.

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

[12] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen. Re-thinking co-salient object detection. *IEEE T. Pattern Anal. Mach. Intell.*, 2021.

[13] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng. Taking a deeper look at co-salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2919–2929, 2020.

[14] L. Gong, Y. Zhang, Y. Zhang, Y. Yang, and W. Xu. Erroneous pixel prediction for semantic image segmentation. *Computational Visual Media*, 8(1):165–175, 2022.

[15] L. Grady. Random walks for image segmentation. *IEEE T. Pattern Anal. Mach. Intell.*, 28(11):1768–1783, 2006.

[16] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3129–3136, 2010.

[17] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *Int. Conf. Comput. Vis.*, pages 991–998, 2011.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[19] Y. Hu, A. Soltoggio, R. Lock, and S. Carter. A fully convolutional two-stream fusion network for interactive image segmentation. *Neural Netw.*, 109:31–42, 2019.

[20] S. D. Jain and K. Grauman. Click carving: Interactive object segmentation in images and videos with point clicks. *Int. J. Comput. Vis.*, 127(9):1321–1344, 2019.

[21] W.-D. Jang and C.-S. Kim. Interactive image segmentation via backpropagating refinement scheme. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5297–5306, 2019.

[22] M. Jia, M. Shi, M. Sirotenko, Y. Cui, C. Cardie, B. Hariharan, H. Adam, and S. Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *Eur. Conf. Comput. Vis.*, pages 316–332, 2020.

[23] M. Jian and C. Jung. Interactive image segmentation using adaptive constraint propagation. *IEEE T. Image Process.*, 25(3):1301–1311, 2016.

[24] T. H. Kim, K. M. Lee, and S. U. Lee. Generative image segmentation using random walks with restart. In *Eur. Conf. Comput. Vis.*, pages 264–275, 2008.

[25] T. H. Kim, K. M. Lee, and S. U. Lee. Nonparametric higher-order learning for interactive segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3201–3208, 2010.

[26] T. Kontogianni, M. Gygli, J. Uijlings, and V. Ferrari. Continuous adaptation for interactive object segmentation by learning from corrections. In *Eur. Conf. Comput. Vis.*, pages 579–596, 2020.

[27] H. Le, L. Mai, B. Price, S. Cohen, H. Jin, and F. Liu. Interactive boundary prediction for object selection. In *Eur. Conf. Comput. Vis.*, pages 18–33, 2018.

[28] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum. Lazy snapping. *ACM T. Graph.*, 23(3):303–308, 2004.

[29] Z. Li, Q. Chen, and V. Koltun. Interactive image segmentation with latent diversity. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 577–585, 2018.

[30] J. Liew, Y. Wei, W. Xiong, S.-H. Ong, and J. Feng. Regional interactive image segmentation networks. In *Int. Conf. Comput. Vis.*, pages 2746–2754, 2017.

[31] J. H. Liew, S. Cohen, B. Price, L. Mai, S.-H. Ong, and J. Feng. Multiseg: Semantically meaningful, scale-diverse segmentations from minimal user input. In *Int. Conf. Comput. Vis.*, pages 662–670, 2019.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014.

[33] Z. Lin, Z.-P. Duan, Z. Zhang, C.-L. Guo, and M.-M. Cheng. Focuscut: Diving into a focus view in interactive segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[34] Z. Lin, Z. Zhang, L.-Z. Chen, M.-M. Cheng, and S.-P. Lu. Interactive image segmentation with first click attention. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13339–13348, 2020.

[35] H. Ling, J. Gao, A. Kar, W. Chen, and S. Fidler. Fast interactive object annotation with curve-gcn. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5257–5266, 2019.

[36] S. Mahadevan, P. Voigtlaender, and B. Leibe. Iteratively trained interactive segmentation. In *Brit. Mach. Vis. Conf.*, 2018.

[37] S. Majumder, A. Rai, A. Khurana, and A. Yao. Two-in-one refinement for interactive segmentation. In *Brit. Mach. Vis. Conf.*, 2020.

[38] S. Majumder and A. Yao. Content-aware multi-level guidance for interactive instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11602–11611, 2019.

[39] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. Deep extreme cut: From extreme points to object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 616–625, 2018.

[40] E. N. Mortensen and W. A. Barrett. Intelligent scissors for image composition. In *Annual Conf. on Comput. Graph. and Intera. Tech.*, pages 191–198, 1995.

[41] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM T. Graph.*, 23(3):309–314, 2004.

[42] K. Sofiiuk, I. Petrov, O. Barinova, and A. Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In

IEEE Conf. Comput. Vis. Pattern Recog., pages 8623–8632, 2020.

[43] G. Song, H. Myeong, and K. Mu Lee. Seednet: Automatic seed generation with deep reinforcement learning for robust interactive segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1760–1768, 2018.

[44] B. Steiner, Z. DeVito, S. Chintala, S. Gross, A. Paszke, F. Massa, A. Lerer, G. Chanan, Z. Lin, E. Yang, et al. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, volume 32, pages 8024–8035, 2019.

[45] V. Vezhnevets and V. Konouchine. Growcut: Interactive multi-label nd image segmentation by cellular automata. *Proc. of Graph.*, 1(4):150–156, 2005.

[46] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *Brit. Mach. Vis. Conf.*, 2009.

[47] T. Wang, J. Yang, Z. Ji, and Q. Sun. Probabilistic diffusion for interactive image segmentation. *IEEE T. Image Process.*, 28(1):330–342, 2018.

[48] J. Wu, Y. Zhao, J.-Y. Zhu, S. Luo, and Z. Tu. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 256–263, 2014.

[49] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang. Deep grabcut for object selection. In *Brit. Mach. Vis. Conf.*, 2017.

[50] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang. Deep interactive object selection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 373–381, 2016.

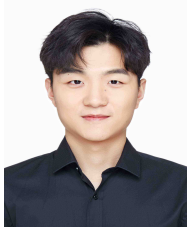
[51] C.-B. Zhang, J. Xiao, X. Liu, Y. Chen, and M.-M. Cheng. Representation compensation networks for continual semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[52] S. Zhang, J. H. Liew, Y. Wei, S. Wei, and Y. Zhao. Interactive object segmentation with inside-outside guidance. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12234–12244, 2020.

[53] X. Zhang, L. Wang, J. Xie, and P. Zhu. Human-in-the-loop image segmentation and annotation. *Science China Information Sciences*, 63(11):1–3, 2020.

[54] Z. Zhang, W. Jin, J. Xu, and M.-M. Cheng. Gradient-induced co-saliency detection. In *Eur. Conf. Comput. Vis.*, pages 455–472, 2020.

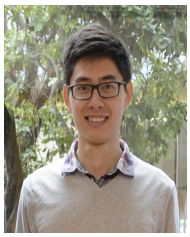
Zheng Lin is currently a Ph.D. candidate with College of Computer Science, Nankai University, under the supervision of Prof. Ming-Ming Cheng. His research interests include deep learning, computer graphics, and computer vision.



Zhao Zhang received the bachelor's degree from Yangzhou University in 2019. He is currently pursuing the master's degree with Media Computing Laboratory, Nankai University, under the supervision of Prof. Ming-Ming Cheng. His research interest includes deep learning and computer vision.

Zi-Yue Zhu is currently a master student from the College of Computer Science at Nankai University, under the supervision of Prof. Ming-Ming Cheng. His research interests include deep learning and computer vision.

Deng-Ping Fan received his PhD degree from Nankai University in 2019. From 2019-2021, he was a research scientist and team lead of IIAI-CV&Med in Inception Institute of Artificial Intelligence (IIAI). He has published about 30+ top journals and conference papers such as TPAMI, TIP, IJCV, TMI, CVPR, ICCV, etc. His research interests include computer vision, deep learning, and saliency detection. He was recognized as the CVPR 2019 outstanding reviewer with a special mention award, the CVPR 2020 outstanding reviewer, the ECCV 2020 high-quality reviewer, and the CVPR 2021 outstanding reviewer. He served as a senior program committee (SPC) member of IJCAI 2021.



Xia-Lei Liu is currently an associate professor at Nankai University. Before that, He was a postdoctoral researcher at the University of Edinburgh. He received his Ph.D. degrees from the Autonomous University of Barcelona in 2019, supervised by Prof. Joost van de Weijer and Prof. Andrew D. Bagdanov. He works in the field of computer vision and machine learning. His research interests include continual learning, self-supervised learning, few-shot learning, long-tailed learning, and many applications (classification, detection, segmentation, crowd counting, image quality assessment, etc).