PhraseClick: Toward Achieving Flexible Interactive Segmentation by Phrase and Click

Henghui Ding[1], Scott Cohen[2], Brian Price[2], and Xudong Jiang[1]

[1] Nanyang Technological University, Singapore
[2] Adobe Research, USA
{ding0093, exdjiang}@ntu.edu.sg, {scohen, bprice}@adobe.com

Abstract. Existing interactive object segmentation methods mainly take spatial interactions such as bounding boxes or clicks as input. However, these interactions do not contain information about explicit attributes of the target-of-interest and thus cannot quickly specify what the selected object exactly is, especially when there are diverse scales of candidate objects or the target-of-interest contains multiple objects. Therefore, excessive user interactions are often required to reach desirable results. On the other hand, in existing approaches attribute information of objects is often not well utilized in interactive segmentation. We propose to employ phrase expressions as another interaction input to infer the attributes of target object. In this way, we can 1) leverage spatial clicks to locate the target object and 2) utilize semantic phrases to qualify the attributes of the target object. Specifically, the phrase expressions focus on "what" the target object is and the spatial clicks are in charge of "where" the target object is, which together help to accurately segment the target-of-interest with smaller number of interactions. Moreover, the proposed approach is flexible in terms of interaction modes and can efficiently handle complex scenarios by leveraging the strengths of each type of input. Our multi-modal phrase+click approach achieves new state-of-the-art performance on interactive segmentation. To the best of our knowledge, this is the first work to leverage both clicks and phrases for interactive segmentation.

Keywords: Interactive segmentation, click, phrase, flexible, attribute

1 Introduction

Interactive object segmentation (or interactive object selection) aims to accurately segment the image into foreground and background given a minimal amount of user interactive inputs. It allows user to gradually refine the prediction with further interaction inputs if any mistakes are made in prediction. These inputs usually come in the form of user clicks/strokes [5, 28, 34, 41, 51, 58] or bounding boxes [16, 29, 49, 57]. This form of input gives hard constraints regarding the location of the object of interest. With the recent advances in deep learning such methods can now often select familiar objects with a small amount of input. Alternately, systems have been proposed that instead use language-based input

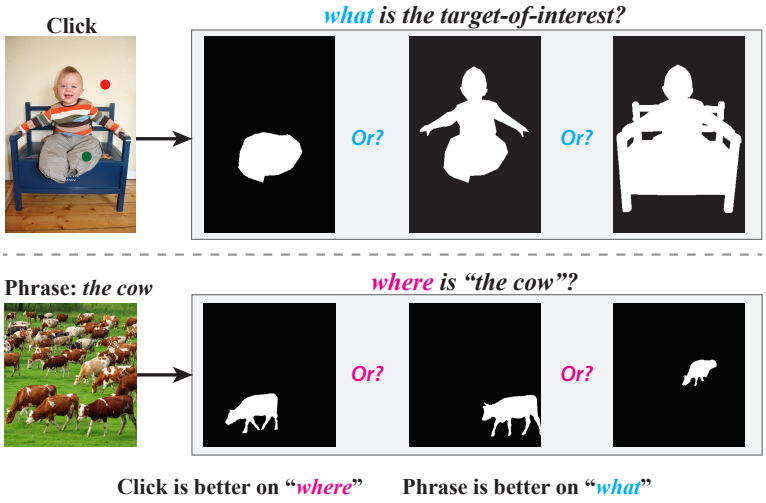

Fig. 1. (Best viewed in color) As interaction inputs, both the click and phrase have their advantages and disadvantages. Here we show ambiguities that exist when using only click or phrase interactions. We propose to utilize both interaction types to make selection flexible and robust to handle complex scenarios.

to drive the selection [24, 35, 50, 61]. Natural language phrases can be used by a neural network to infer high-level attribute information about what the object of interest looks like that can then be used to select the objects.

While great strides have been made in interactive selection, each of these interaction approaches may still fall short and require additional and excessive user interaction. For example, click-based methods are required to infer the target object given only spatial constraints and usually are trained to select entire objects. However, the region of user interest may instead be an object part or a combination of multiple objects. For example, the first row in Fig. 1, the click on the boy may indicate the trousers, the boy, or the whole foreground (boy and chair). This leads to ambiguities that must be overcame with additional user inputs, which directly runs counter to the goal of minimizing user interaction. Click-based methods also generally assume accurate input, but with mobile devices it can be difficult for users to accurately click on objects, especially given that the users finger is occluding the object of interest. It remains a significant challenge to accurately segment a target-of-interest with a few clicks.

On the other hand, language-driven segmentation methods learn the overall appearance of objects and must infer their location. A language phrase naturally and easily overcomes ambiguities such as whether the target is an object, object part, or collection of objects. A phrase can also provide rough spatial information (the woman on the left). Besides, for mobile devices like smartphones, speech is a natural and desired interface and easier than precise touching on a small phone screen. However, in many cases an object name and rough location is

not sufficient to produce a desired result. For example, in images where there are multiple objects with similar appearance (an image of dozens of cows, see Fig. 1) it can be very difficult to articulate using language a single target instance. In such cases, directly clicking on the object is much easier for a user. It can also be difficult to verbally articulate some required corrections that do not correspond to an entire semantically-meaningful region. Further, due to the long tail distribution of objects in images, it is difficult to obtain labels and training data for all possible objects.

The strengths and weaknesses of click-based and phrase-based inputs are complementary with clicks giving hard spatial constraints and phrases giving high-level attribute information. An effective combination of these inputs may reduce the amount of user interactions needed for accurately selecting objects of interest. Given this observation, we propose to build a versatile interactive segmentation network that accepts both clicks and phrases as interaction input. We use a convolutional neural network (CNN) to process the input image and clicks, and employ the bi-directional LSTM (bi-LSTM) to encode the phrases and infer language features. To bridge click-based spatial constraints with phrase-based attribute constraints, we introduce a novel attribute guided feature attention module to effectively integrate language and vision features together. Our approach can better handle complex scenarios via utilizing advantages of these two interactions.

The main contributions of this paper can be summarized as follows: 1) To the best of our knowledge, this is the first work to leverage both clicks and phrases for interactive segmentation. The proposed approach allows the user the flexibility of using the interaction method that is most suitable for a given task, making the system more practical than past approaches. 2) We propose an attribute guided feature attention module to bring clicks and phrases information together. It extracts discriminative attribute clues from interactions and integrate the vision and language attribute clues. 3) Extensive experimental results have demonstrated that phrase expressions are indeed effective at boosting the performance of interactive segmentation, especially in some complex scenarios and with few clicks.

2 Related Work

2.1 Interactive Object Segmentation

Many interactive segmentation methods have been proposed in the past decades using many different interaction types such as bounding boxes [16, 29, 49, 57], contours [1, 6, 26, 45], strokes [5, 31, 51, 53] and clicks [25, 28, 34, 40, 41, 58]. There are also some language-based segmentation methods [24, 35, 61], but they only provide an initial result and cannot further refine the result to correct mistakes.

Early methods rely on low-level features, such as color similarity or boundary properties [26, 45]. For example, [5, 31, 49] adopt graphical models, [18] employs random walker and [3, 11, 47] are based on geodesic approaches. However, low-level features are not robust and hence excessive user interactions are required

for these methods to achieve desirable segmentation results. Recently, thanks to the great success of convolution neural networks (CNN) [36, 44, 55, 56, 63–70], CNN-based interactive segmentation methods have achieved exciting progress. For example, Xu et al. [58] concatenate Euclidean distance maps to user foreground and background clicks with the image as input to a fully convolutional network (FCN) [39]. Hu et al. [25] employ a two-stream network to deal with image and user clicks separately to make the user interaction impact the result more directly. Le et al. [28] introduce boundary clicks to perform object selection. Besides the clicks, bounding boxes have also been used in CNN-based methods [57]. As a variant to using bounding boxes, Papadopoulos et al. [46] propose extreme points and Maninis et al. [41] use extreme points to generate Gaussian heatmap and crop the image to achieve instance segmentation. Agustsson et al. [2] further segment all regions jointly with extreme points and scribbles. As these methods receive only spatial constraints, they cannot provide high-level attribute information to the method and thus may require additional user input to overcome this drawback in challenging cases. Different from these methods, our approach receives both spatial constraints and high-level attribute information.

2.2 Semantic/Instance Segmentation

Semantic segmentation and instance segmentation are closely related with interactive segmentation. Driven by the significant success of CNN, many deep-learning-based works have been proposed for semantic segmentation [8, 12–15, 22, 52, 71] and instance segmentation [7, 20, 23, 37]. For example, Long et al. [39] propose the FCN to train the segmentation network end-to-end. He et al. [20] propose to add an instance-level segmentation mask branch on the top of Faster R-CNN [48]. However, it is not reasonable to directly transform semantic/instance segmentation to interactive segmentation [58]. Interactive object segmentation methods respond to user's inputs instead of predefined labels, thus the ability to segment any unseen objects is required and this is impractical in current segmentation approaches. In this work, we modify the semantic segmentation framework [9] to have it accept interaction inputs, and re-train the modified framework on interactive object segmentation datasets.

2.3 Referring Expression Comprehension

Referring expression comprehension methods [27, 35, 38, 54, 60–62] detect a specific object in an image given a referring expression. Some comprehension methods can be used to segment the referential object [24, 35, 59, 61]. However, these methods only compute an initial segmentation and cannot further correct the segmentation mask, which is required for interactive segmentation if any mistakes are made. For example, MAttNet [61] depends on the instance segmentation results of Mask R-CNN [20]. MAttNet first scores and ranks the similarity between phrase embedding and instance objects generated by Mask R-CNN, and then outputs the mask of best matched object. These methods cannot

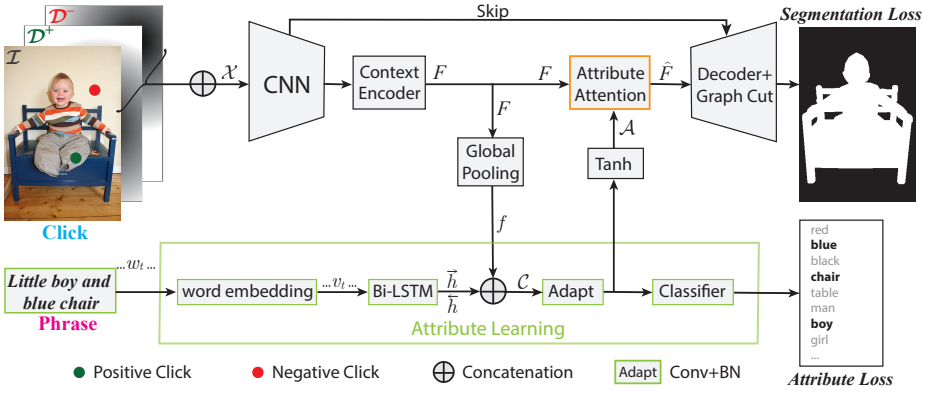

Fig. 2. The proposed approach accepts clicks and phrases as interaction input, but it is *not* necessary to enter both types of interactions at the same time.

further improve the segmentation mask even with new additional phrase input. Thus, they cannot meet the requirement of practical application for interactive segmentation. In contrast, our approach enables users to add interactive information until the segmentation results meets the users' requirements.

3 Approach

In this work, we propose to build a versatile interactive segmentation network that can take both clicks and phrases as interaction input. Compared with previous approaches that accept only one type of interaction, the proposed approach 1) is more flexible in terms of interaction ways and 2) can better handle complex scenarios via utilizing advantages of these two interactions.

3.1 Network Architecture

The overall architecture of our approach is shown in Fig. 2, the proposed approach accepts both clicks and phrases as interaction input. We use the ResNet-101 [21] based DeepLabv3+ [9] as the backbone of vision part. The clicks are transformed to distance maps and concatenated with original image to form a 5-channel input for the CNN as in [58]. For the language part, the phrase is processed by a word-to-vector model and then a bi-directional LSTM to extract the language clues. During testing, we employ graph cut [5] as a post-processing tool to refine the final segmentation mask.

3.2 Click Interaction

Click-based input is commonly used to provide location information and give spatial constraints about a target of interest. We also use clicks to provide spatial information about the object.

Click Input Transformation. Click information is usually introduced at the beginning of a network. As in [58], the user provides a set of sequential clicks to segment the target region. The interactions contain positive clicks $\mathcal{C}^+$ on the target region and negative clicks $\mathcal{C}^-$ on "background". We employ a Euclidean distance transformation to transform the positive clicks $\mathcal{C}^+$ and negative clicks $\mathcal{C}^-$ to two Euclidean distance maps $\mathcal{D}^+$ and $\mathcal{D}^-$. These two distance maps are truncated to the same spatial size as the original input image $\mathcal{I}$. We concatenate the 3 channels of the input image $\mathcal{I}$ and the two distance maps $(\mathcal{D}^+, \mathcal{D}^-)$ to form the 5 channel input:

$$\mathcal{X} = \mathcal{I} \oplus \mathcal{D}^+ \oplus \mathcal{D}^- \tag{1}$$

where $\oplus$ denotes concatenation and $\mathcal{X}$ is the 5-channel input to the network.

Click Simulation Protocol. We follow the simulation strategies proposed in [58] to form a set of synthetic positive and negative clicks. Furthermore, to encourage the model to learn to generate correct prediction from ambiguous clicks, we follow [33] and introduce another sampling strategy that samples clicks on the overlapped foreground and overlapped background of different object masks, for example, the positive and negative clicks in Fig. 2.

3.3 Phrase Interaction

While clicks as a kind of spatial interaction can indicate the position of target object, such spatial interactions cannot explicitly express the semantic attributes of a target object. A language phrase naturally and easily expresses attributes of objects and is commonly used in referring segmentation [10, 24, 27, 54, 60–62]. However, referring segmentation methods only compute an initial segmentation mask and cannot further improve the segmentation mask even with new additional phrase input, which does not agree with the goal of interactive segmentation.

Bringing Click and Phrase Together. In this work, we explore using phrase expressions as another interaction input to express the attributes of a target-of-interest and quickly narrow the range of candidates, which can assist the click interaction process and decrease the user interaction times. For regions that are difficult to articulate, such as small mistakes that need correction or objects that do not have easily identifiable attributes that will separate them from surrounding objects, clicks provide a strong spatial constraint to supplement the attribute information.

Phrase Expression Annotation. The interactive segmentation datasets have no phrase annotation. To train the proposed network, corresponding phrase annotation for each segmentation mask is required. Therefore, we annotate phrase expressions for every image in Grabcut [49] and Berkeley [43], some examples are shown in Fig. 3.

3.4 Attribute Feature Attention Module

To uniformly integrate click and phrase interactions in a single network, we propose an attribute attention module. CNN features contain rich attributes

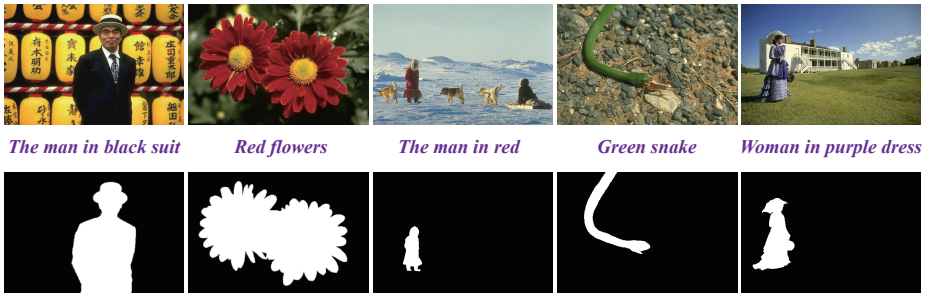

Fig. 3. Examples of our annotated phrase-segmentation pairs for Grabcut [49] and Berkeley [43].

information like color or shape, which are preserved in different feature channels [10]. Based on this observation, we explore to infer attribute clues of target-of-interest from vision features of $\mathcal{X}$ (in Eq. 1) and language features of phrases. The attribute clues of the target object are changed with user's interaction, *i.e.* phrases and clicks in this work. We employ the attribute clues to compute channel-wise attribute attention, and then leverage this attribute attention to emphasize some specific channels and suppress others in order to help the network to determine the object of interest.

First, to process language input, we use a word-to-vector model to embed each word $w_t \in \{w_1, ..., w_T\}$, $v_t = word2vec(w_t)$. Next, we employ the bi-directional LSTM (bi-LSTM) to encode the holistic context of phrase expression:

$$\overrightarrow{h}_t = \overrightarrow{LSTM}(v_t, \overrightarrow{h}_{t-1}), \quad \overleftarrow{h}_t = \overleftarrow{LSTM}(v_t, \overleftarrow{h}_{t+1}) \tag{2}$$

Besides the context from phrase expression, we also extract visual context vector from CNN features, $f = pooling(F)$, where *pooling* is a global average pooling operation, F with size of $H \times W \times \#C$ is the high-level feature of CNN, f with size of $1 \times 1 \times \#C$ is visual context vector that preserves visual clues in channels. Then we concatenate the visual context vector and the bi-directional context vectors of phrase to generate final attribute context representation:

$$\mathcal{C} = \overrightarrow{h} \oplus \overleftarrow{h} \oplus f \tag{3}$$

where $\overrightarrow{h} = (\overrightarrow{h}_1, ..., \overrightarrow{h}_t, ..., \overrightarrow{h}_T)$ and $\overleftarrow{h} = (\overleftarrow{h}_1, ..., \overleftarrow{h}_t, ..., \overleftarrow{h}_T)$, and $\mathcal{C}$ is the attribute context vector. Next we adapt the attribute context vector to make it have the same number of channels ($\#C = 512$ in our experiments) with features in CNN, and normalize the values to range of $[-1, 1]$ by:

$$\mathcal{A} = tanh(Adapt(\mathcal{C}, \Theta)) \tag{4}$$

where $tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$, *Adapt* is `Conv+BN` and Θ is its learnable parameters. Now we get the attribute attention weight $\mathcal{A}$, in which different

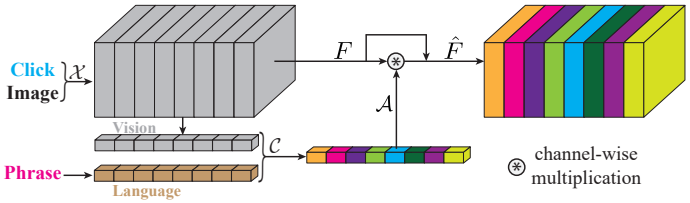

Fig. 4. Attribute Attention: It emphasizes feature channels that have larger response in semantic attribute learning, which is based on the phrase interaction, click interaction and visual patterns.

attribute channels have different attention weights in range of $[-1, 1]$. Finally, as shown in Fig. 4, we apply the attribute attention weights on feature maps in CNN:

$$\hat{F} = \beta \mathcal{A} * F + F \tag{5}$$

where β is a learnable weight and $*$ denotes channel-wise multiplication. To have our approach more flexible in terms of interaction methods, the phrase interaction is not compulsory. When there is no phrase interaction input, the $\overrightarrow{h}$ and $\overleftarrow{h}$ are set to 0.

The attribute attention module helps to bridge the clicks and phrases information together. These two interaction inputs work together to reduce the interaction times for improved performance.

3.5 Training Loss

The proposed approach is a unified framework that brings vision part and language part together. To train such a network end-to-end and balance the effect of two different interactions, we carefully design its loss function. First, we use the binary cross-entropy loss for the segmentation branch training. Given the segmentation ground truth t and the segmentation branch output s, the segmentation loss is:

$$\mathcal{L}_m = -tlog(\sigma(s)) - (1-t)log(1-\sigma(s)) \tag{6}$$

where $\sigma(x) = 1/(1+e^{-x})$ is sigmoid function.

Additionally, in order to help the network optimize the attribute learning process and catch meaningful visual and phrasal attributes, we introduce an attribute loss for the attribute learning module. We first generate the attribute label $\{a_i\}_{i=1}^{N}$ following [27], $a_i \in \{0, 1\}$ indicates whether the i-th attribute word exists in the input phrase. Denoting the output of attribute learning module as p_i, the objective function for attribute module is defined as:

$$\mathcal{L}_a = \sum_{i=1}^{N} w_i(a_i log(\sigma(p_i)) + (1-a_i)log(1-\sigma(p_i)) \tag{7}$$

Interaction Inputs	Phrase	Click	Graph Cut	IoU
Phrase-only	✓			50.98
Click-only		✓		77.93
PhraseClick	✓	✓		83.94
PhraseClick	✓	✓	✓	85.02

Table 1. Ablation study of the proposed approach on testA of RefeCOCO.

where w_i is a weighting factor to address the unbalance of different attributes, and a_i indicates whether the *i-th* attribute word exists in the input phrase. The final loss for the proposed network is $\mathcal{L} = \omega_m \mathcal{L}_m + \omega_a \mathcal{L}_a$, where ω_m and ω_a are used to balance the contributions of the segmentation and attribute loss.

4 Experiments

4.1 Implementation Details

The network and all the experiments are implemented based on the public Pytorch platform. We use ResNet-101 [21] based DeepLabv3+ [9] as the backbone of the segmentation branch. The first convolution layer is modified to $5 \times 64 \times 7 \times 7$ to deal with the 5-channel input. The proposed attribute attention module is placed after ASPP. The parameters of newly added layers are randomly initialized from a Gaussian distribution with standard variance of 10^{-2}. The network is trained by SGD with batch size set to 10. For batch processing, we resize the inputs to 512×512 pixels during training. We adopt random horizontal flipping to augment the training data. The learning rate is set to 1×10^{-8} and the parameters of the new layers are trained with a higher learning rate 1×10^{-7}. Momentum and weight decay are fixed to 0.99 and 1×10^{-4} respectively. We empirically set ω_m and ω_a to 1 and 10. The network is evaluated with Intersection-over-Union (**IoU**) [39]. To jointly train the attribute learning branch and segmentation branch, we train our network on RefCOCO [27] first, which contains referring phrase expressions and segmentation masks for every instance objects in each of 19994 images. Then, for interactive segmentation datasets like PASCAL VOC, we take the name of categories as the phrase input, like person and car. We annotate phrase expressions for the Grabcut and Berkeley datasets, as shown in Fig. 3.

4.2 Ablation Studies

We conduct the ablation studies on RefCOCO [27], as shown in Table 1. First, to verify the effectiveness of the proposed attribute attention module, we discard the click interaction and only take the phrase as input. With *Phrase-only* input, our network can achieve an acceptable IoU of 50.98%. We also display

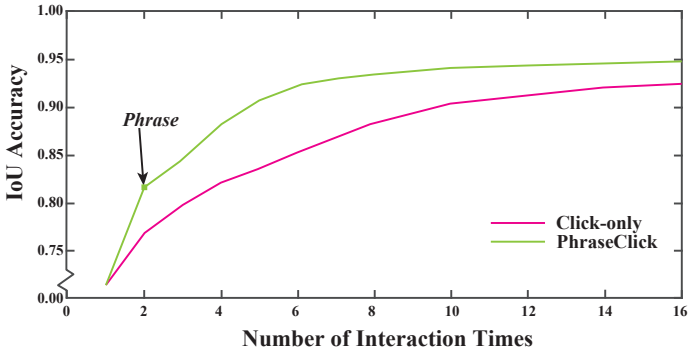

Fig. 5. Segmentation accuracy with different number of interaction times.

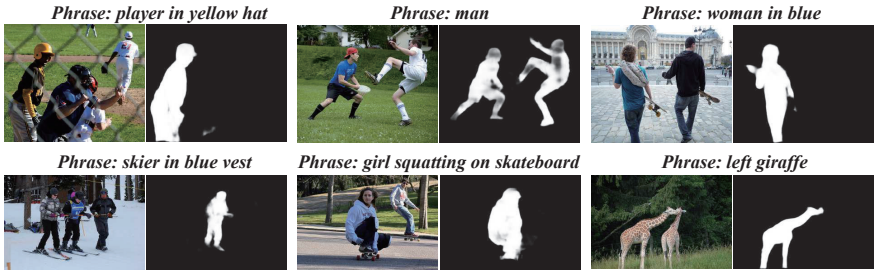

Fig. 6. Examples of segmentation score maps (soft value from 0 to 1) of *Phrase-only*.

some visualized segmentation score maps in Fig. 6. Both the quantitative and qualitative results demonstrate that the attribute attention module does help the network to select the target-of-interest according to referential phrase expression.

We also conduct a *Click-only* experiment in which the the $\overrightarrow{h}$ and $\overleftarrow{h}$ in Eq. (3) are set to 0. This achieves 77.93% IoU by 3 clicks, which is much higher than phrase only. It shows the advantage of click interaction for interactive object segmentation. Next, we verify whether the phrase interaction can speed up the click-based interactive segmentation process and improve the segmentation performance. To this end, we first compare the segmentation performance of *Click-only vs*. *PhraseClick*. As shown in Table 1, compared with *Click-only* without phrase, *PhraseClick* achieves 6.01% better performance on IoU. This demonstrates that the phrase expression and attribute attention module visibly enhances the performance of click-based interactive segmentation. Note, in Table 1, the interaction inputs for *PhraseClick* are one phrase and two clicks (one positive and one negative). For fair comparison, there is one more positive click for *Click-only* than *PhraseClick* in Table 1, *i.e*. the number of interaction inputs for *Click-only* and *PhraseClick* is the same. Next, we show the segmentation performance with different number of interaction times in Fig. 5. We start with

Methods	PASCAL VOC (85% IoU)	GrabCut (90% IoU)	Berkeley (90% IoU)
GraphCut [5]	15.06	11.10	14.33
Geodesic Matting [4]	14.75	12.44	15.96
Random Walker [18]	11.37	12.30	14.02
Geodesic Convexity [19]	11.73	8.38	12.57
iFCN [58]	6.88	6.04	8.65
RIS-Net [34]	5.12	5.00	6.03
ITIS [40]	3.80	5.60	-
LDN [32]	-	4.79	-
FCTSFN [25]	4.58	3.76	6.49
PhraseClick (Ours)	**3.12**	**2.06**	**3.26**

Table 2. Comparison with previous state-of-the-art interactive object segmentation methods. We demonstrate the number of interactions that each method needed to reach a certain segmentation performance.

an initial positive click, then gradually add a phrase/click as the next interaction according to the current segmentation prediction. As shown in Fig. 5, to reach a specific IoU performance, the network taking the *PhraseClick* as interaction inputs is faster than the network that only takes *Click* input. This shows the phrase expression helps to speed up the interaction process and improve the segmentation performance. During the inference, we employ the graph cut [5] as post-processing method to refine the segmentation mask, which improve the performance by 1.08% IoU. The performance gain is brought by the refinement in boundary regions.

4.3 Results on Benchmarks

Interactive Object Segmentation

We first compare the proposed approach with previous state-of-the-art interactive segmentation methods. Interactive object segmentation enables users to add interactive information until the desired selection is reached. We conduct the interactive segmentation experiments on three public benchmarks with instance object annotations: PASCAL VOC [17], Grabcut [49], Berkeley [43].

To evaluate the performance of interactive object segmentation, the standard process is: 1) starting with an initial positive click at the center of the target-of-interest, the network generates an initial segmentation mask for the target object; 2) then a succeeding positive/negative click is added at the center of the largest wrongly predicted piece to add/remove this piece; 3) the 2nd step is repeated again and again until the desirable accuracy is reached or the maximum number (set to 20 as [58]) of clicks is reached. Different from existing interactive segmentation methods that only use clicks as input, the proposed approach accepts both the clicks and phrases as interaction input. Therefore, to evaluate

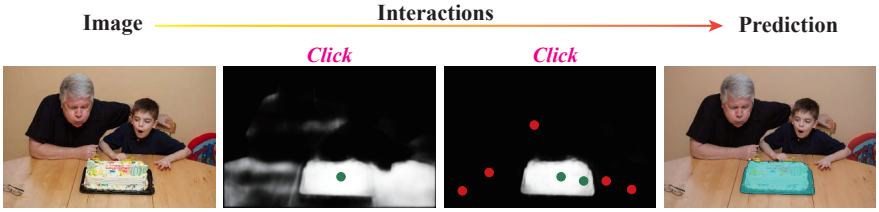

Fig. 7. Existing click-based methods' interaction process: click first, then gradually refine with more clicks.

Methods	User#1	User#2	User#3	User#4	User#5	User#6	User#7	Average
Click-only	4.38	4.41	4.65	4.52	4.33	4.59	4.57	4.49
PhraseClick	3.02	3.13	3.17	3.14	3.00	3.07	3.12	**3.09**

Table 3. User study. *PhraseClick* requires less interactions than *Click-only*.

our approach and compare with existing state-of-the-art works, we input the phrase for the target-of-interest after the initial click in the first step, which together are counted as two interaction times. Then the subsequent interactions only use clicks. Table 2 shows the number of interactions that each approach requires to reach a certain IoU. We achieve new state-of-the-art performance on the three benchmarks, which shows the advantage of the proposed approach.

Furthermore, comparing to existing methods that accept a single type of interaction (as shown in Fig. 7), the proposed approach is more flexible in terms of interaction and can better handle complex scenarios via utilizing advantages of both clicks and phrases. Although we fix the order of phrase input to be the second interaction after the initial positive click in Table 2, this is just for convenience of evaluation. In practice, users can input the phrase at any time and in any order, *e.g.* phrase first and then refine with clicks. We show some interaction examples in Fig. 8. In the first row in Fig. 8, we start with a phrase to get the initial segmentation prediction, then refine the initial prediction with clicks. It demonstrates the advantages of clicks at locating "where" the object is when there are multiple objects with the similar appearance. In the second row in Fig. 8, phrases are employed to disambiguate the confused results of clicks, which shows the advantages of phrases at describing "what" the object of interest is. The first two rows show how the integration of clicks and phrases help to improve the performance of interactive segmentation and better handle complex scenarios. Our model can also take only clicks similar to previous click-based segmentation methods, as shown in the last row in Fig. 8, which shows the abilities of clicks to correct mistakes and refine details.

To justify the robustness and effectiveness of the proposed PhraseClick with different real users, we conduct a user study in Table 3. We randomly choose 50 images from PASCAL VOC, RefCOCO, and Berkeley for this testing. Seven

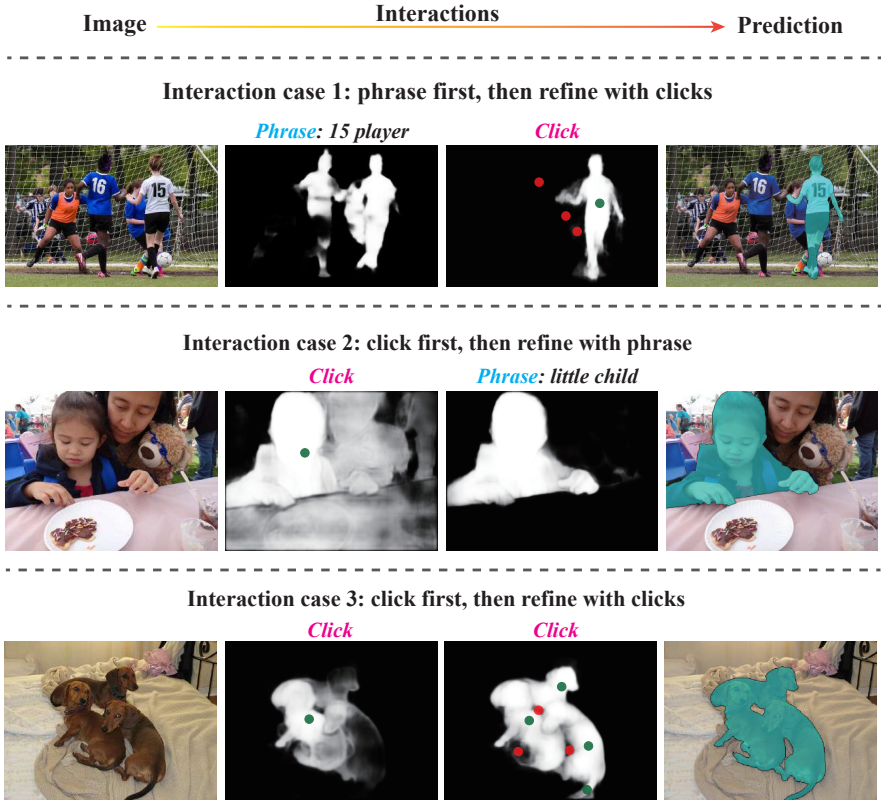

Fig. 8. Our interaction process is more flexible. The user can choose either click or phrase in each interaction step, making the system more practical than past approaches. Furthermore, our approach can better handle complex scenarios via utilizing advantages of these two interactions.

users participated in this study. We record the number of interactions they required to reach 85% IoU segmentation performance. As shown in Table 3, our *PhraseClick* requires less interactions than *Click-only*. In practical applications, the amount of time to input a phrase depends greatly on implementation and the input device. Given a strong voice-to-text algorithm, the time to input a phrase could be close to click. In our specific case, we do not have a sophisticated voice-to-text system, so we allow the user to type in the desired phrase. This will be slower than using voice and likely slower than clicks.

Referential Object Segmentation

Although the proposed network is designed for interactive object segmentation, it can also be used for referential object segmentation. Existing referential object segmentation methods only compute an initial segmentation and cannot further correct the mistakes, thus they cannot meet the requirement of practical

Methods	RMI [35]	DMN [42]	RNN [30]	MAttNet [61]	CMSA [60]	*Phrase*	*PhraseClick*
IoU	45.18	49.78	55.33	56.51	58.32	50.42	**80.02**

Table 4. Comparison with state-of-the-art referring segmentation methods.

# of interactions	Phrase*1	Phrase*1+Click*1	Phrase*1+Click*2
DMN(5-channel) [42]	49.78	60.32	62.08
RNN(5-channel) [30]	55.33	63.41	65.02
CMSA(5-channel) [60]	**58.32**	67.10	70.02
PhraseClick(Ours)	50.42	**80.02**	**84.56**

Table 5. Comparison with our modified referring segmentation methods.

application for interactive segmentation. The proposed approach can gradually refine the segmentation result until it meets the user's requirement.

We compare with referential object segmentation methods on the validation set of RefCOCO [27]. As shown in Table 4, although our *Phrase* is proposed as part of a network design that includes clicking, it achieves competitive results with phrase-only methods. With only one click (*PhraseClick*), the proposed approach significantly outperforms the state-of-the-arts by a very large margin, illustrating the utility of adding clicks to phrase-based selection. For fair comparison, we also transform some referring segmentation methods into interactive segmentation by replacing their RGB input with the same five-channel input our method uses (Sec. 3.2) and compare their performance with different number of interactions, as shown in Table 5. Our approach significantly outperform others as interaction times increase, showing that a naive transformation of referential object segmentation methods is not as effective as our model that integrates the phrase information in an attribute guided feature attention module.

5 Conclusion

In this work, we propose an interactive segmentation network that can take either clicks or phrases or both as interaction input, which utilizes the complementary merits of these two interactions. Besides the commonly used spatial constraints like clicks, we introduce phrase expression as another interaction input to infer semantic attributes and propose an attribute attention module to integrate such attributes information into the network. Specifically, the language phrases focuses on "what" the target-of-interest is and the spatial clicks are in charge of "where" the target-of-interest is. Extensive experimental results have shown that the proposed approach is flexible in terms of interaction and can handle complex scenarios well.

References

1. Acuna, D., Ling, H., Kar, A., Fidler, S.: Efficient interactive annotation of segmentation datasets with polygon-rnn++. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 859–868 (2018)
2. Agustsson, E., Uijlings, J.R., Ferrari, V.: Interactive full image segmentation by considering all regions jointly. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 11622–11631 (2019)
3. Bai, X., Sapiro, G.: A geodesic framework for fast interactive image and video segmentation and matting. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)
4. Bai, X., Sapiro, G.: Geodesic matting: A framework for fast interactive image and video segmentation and matting. International journal of computer vision **82**(2), 113–132 (2009)
5. Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In: IEEE international conference on computer vision. vol. 1, pp. 105–112. IEEE (2001)
6. Castrejon, L., Kundu, K., Urtasun, R., Fidler, S.: Annotating object instances with a polygon-rnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5230–5238 (2017)
7. Chen, L.C., Hermans, A., Papandreou, G., Schroff, F., Wang, P., Adam, H.: Masklab: Instance segmentation by refining object detection with semantic and direction features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4013–4022 (2018)
8. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv:1606.00915 (2016)
9. 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). pp. 801–818 (2018)
10. Chen, Y.W., Tsai, Y.H., Wang, T., Lin, Y.Y., Yang, M.H.: Referring expression object segmentation with caption-aware consistency. arXiv preprint arXiv:1910.04748 (2019)
11. Criminisi, A., Sharp, T., Blake, A.: Geos: Geodesic image segmentation. In: European Conference on Computer Vision. pp. 99–112. Springer (2008)
12. Ding, H., Jiang, X., Liu, A.Q., Thalmann, N.M., Wang, G.: Boundary-aware feature propagation for scene segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6819–6829 (2019)
13. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Context contrasted feature and gated multi-scale aggregation for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2393–2402 (June 2018)
14. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Semantic correlation promoted shape-variant context for segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8885–8894 (June 2019)
15. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Semantic segmentation with context encoding and multi-path decoding. IEEE Transactions on Image Processing **29**, 3520–3533 (2020)
16. Dutt Jain, S., Grauman, K.: Predicting sufficient annotation strength for interactive foreground segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (2013)

17. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2) (2010)
18. Grady, L.: Random walks for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(11), 1768–1783 (2006)
19. Gulshan, V., Rother, C., Criminisi, A., Blake, A., Zisserman, A.: Geodesic star convexity for interactive image segmentation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3129–3136. IEEE (2010)
20. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (2016)
22. Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., Perazzi, F.: Temporally distributed networks for fast video semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8818–8827 (2020)
23. Hu, R., Dollár, P., He, K., Darrell, T., Girshick, R.: Learning to segment every thing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4233–4241 (2018)
24. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: European Conference on Computer Vision. pp. 108–124. Springer (2016)
25. Hu, Y., Soltoggio, A., Lock, R., Carter, S.: A fully convolutional two-stream fusion network for interactive image segmentation. Neural Networks **109** (2019)
26. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International journal of computer vision **1**(4) (1988)
27. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014)
28. Le, H., Mai, L., Price, B., Cohen, S., Jin, H., Liu, F.: Interactive boundary prediction for object selection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 18–33 (2018)
29. Lempitsky, V.S., Kohli, P., Rother, C., Sharp, T.: Image segmentation with a bounding box prior. In: ICCV. vol. 76 (2009)
30. Li, R., Li, K., Kuo, Y.C., Shu, M., Qi, X., Shen, X., Jia, J.: Referring image segmentation via recurrent refinement networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2018)
31. Li, Y., Sun, J., Tang, C.K., Shum, H.Y.: Lazy snapping. ACM Transactions on Graphics (ToG) (2004)
32. Li, Z., Chen, Q., Koltun, V.: Interactive image segmentation with latent diversity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 577–585 (2018)
33. Liew, J.H., Cohen, S., Price, B., Mai, L., Ong, S.H., Feng, J.: Multiseg: Semantically meaningful, scale-diverse segmentations from minimal user input. In: The IEEE International Conference on Computer Vision (2019)
34. Liew, J., Wei, Y., Xiong, W., Ong, S.H., Feng, J.: Regional interactive image segmentation networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2746–2754. IEEE (2017)
35. Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1271–1280 (2017)

36. Liu, J., Ding, H., Shahroudy, A., Duan, L.Y., Jiang, X., Wang, G., Kot, A.C.: Feature boosting network for 3d pose estimation. IEEE transactions on pattern analysis and machine intelligence **42**(2), 494–501 (2020)
37. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8759–8768 (2018)
38. Liu, X., Wang, Z., Shao, J., Wang, X., Li, H.: Improving referring expression grounding with cross-modal attention-guided erasing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
39. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
40. Mahadevan, S., Voigtlaender, P., Leibe, B.: Iteratively trained interactive segmentation. BMVC (2018)
41. Maninis, K.K., Caelles, S., Pont-Tuset, J., Van Gool, L.: Deep extreme cut: From extreme points to object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 616–625 (2018)
42. Margffoy-Tuay, E., Pérez, J.C., Botero, E., Arbeláez, P.: Dynamic multimodal instance segmentation guided by natural language queries. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 630–645 (2018)
43. McGuinness, K., Oconnor, N.E.: A comparative evaluation of interactive segmentation algorithms. Pattern Recognition **43**(2), 434–444 (2010)
44. Mei, J., Wu, Z., Chen, X., Qiao, Y., Ding, H., Jiang, X.: Deepdeblur: text image recovery from blur to sharp. Multimedia Tools and Applications **78**(13), 18869–18885 (2019)
45. Mortensen, E.N., Barrett, W.A.: Intelligent scissors for image composition. In: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques. ACM (1995)
46. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: IEEE International Conference on Computer Vision. pp. 4930–4939 (2017)
47. Price, B.L., Morse, B., Cohen, S.: Geodesic graph cut for interactive image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3161–3168. IEEE (2010)
48. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
49. Rother, C., Kolmogorov, V., Blake, A.: "grabcu": interactive foreground extraction using iterated graph cuts. ACM transactions on graphics (TOG) **23**(3), 309–314 (2004)
50. Rupprecht, C., Laina, I., Navab, N., Hager, G.D., Tombari, F.: Guide me: Interacting with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8551–8561 (2018)
51. Shi, J., Malik, J.: Normalized cuts and image segmentation. Departmental Papers (CIS) p. 107 (2000)
52. Shuai, B., Ding, H., Liu, T., Wang, G., Jiang, X.: Toward achieving robust low-level and high-level scene parsing. IEEE Transactions on Image Processing **28**(3), 1378–1390 (2018)
53. Vezhnevets, V., Konouchine, V.: Growcut: Interactive multi-label nd image segmentation by cellular automata. In: proc. of Graphicon. vol. 1, pp. 150–156. Citeseer (2005)

54. Wang, P., Wu, Q., Cao, J., Shen, C., Gao, L., Hengel, A.v.d.: Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
55. Wang, X., Ding, H., Jiang, X.: Dermoscopic image segmentation through the enhanced high-level parsing and class weighted loss. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 245–249. IEEE (2019)
56. Wang, X., Jiang, X., Ding, H., Liu, J.: Bi-directional dermoscopic feature learning and multi-scale consistent decision fusion for skin lesion segmentation. IEEE Transactions on Image Processing **29**, 3039–3051 (2019)
57. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.: Deep grabcut for object selection. In: BMVC (2017)
58. 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. pp. 373–381 (2016)
59. Ye, L., Liu, Z., Wang, Y.: Dual convolutional lstm network for referring image segmentation. IEEE Transactions on Multimedia (2020)
60. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10502–10511 (2019)
61. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1307–1315 (2018)
62. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision. pp. 69–85. Springer (2016)
63. Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., Lu, H.: High-resolution image inpainting with iterative confidence feedback and guided upsampling. In: European Conference on Computer Vision. Springer (2020)
64. Zeng, Y., Lu, H., Zhang, L., Feng, M., Borji, A.: Learning to promote saliency detectors. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
65. Zeng, Y., Zhuge, Y., Lu, H., Zhang, L.: Joint learning of saliency detection and weakly supervised semantic segmentation. In: IEEE International Conference on Computer Vision (2019)
66. Zeng, Y., Zhuge, Y., Lu, H., Zhang, L., Qian, M., Yu, Y.: Multi-source weak supervision for saliency detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
67. Zhang, L., Dai, J., Lu, H., He, Y.: A bi-directional message passing model for salient object detection. In: CVPR (2018)
68. Zhang, L., Lin, Z., Zhang, J., Lu, H., He, Y.: Fast video object segmentation via dynamic targeting network. In: ICCV (2019)
69. Zhang, L., Zhang, J., Lin, Z., Lu, H., He, Y.: Capsal: Leveraging captioning to boost semantics for salient object detection. In: CVPR (2019)
70. Zhang, L., Zhang, J., Lin, Z., Mech, R., Lu, H., He, Y.: Unsupervised video object segmentation with joint hotspot tracking. In: ECCV (2020)
71. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)