

ScribbleSeg: Scribble-based Interactive Image Segmentation

Xi Chen¹ Yau Shing Jonathan Cheung¹ Ser-Nam Lim² Hengshuang Zhao¹
¹The University of Hong Kong ²Meta AI

Abstract

*Interactive segmentation enables users to extract masks by providing simple annotations to indicate the target, such as boxes, clicks, or scribbles. Among these interaction formats, scribbles are the most flexible as they can be of arbitrary shapes and sizes. This enables scribbles to provide more indications of the target object. However, previous works mainly focus on click-based configuration, and the scribble-based setting is rarely explored. In this work, we attempt to formulate a standard protocol for scribble-based interactive segmentation. Basically, we design diversified strategies to simulate scribbles for training, propose a deterministic scribble generator for evaluation, and construct a challenging benchmark. Besides, we build a strong framework **ScribbleSeg**, consisting of a Prototype Adaption Module (PAM) and a Corrective Refine Module (CRM), for the task. Extensive experiments show that ScribbleSeg performs notably better than previous click-based methods. We hope this could serve as a more powerful and general solution for interactive segmentation. Our code will be made available.*

1. Introduction

Interactive image segmentation requires users to indicate the target by providing simple annotations such as boxes, scribbles, and clicks. Compared with traditional annotation tools like the lasso or brush, interactive models could largely reduce the time and cost of creating masks, which is especially important in the era of big data.

Demonstrations of common forms of interactions are shown in Fig. 1. Among them, we claim that drawing scribbles is the most flexible and practical way to indicate the foreground and background regions. As shown in Fig. 1 (c), although boxes could indicate the size and rough location of the target, they could not make further indications inside the rectangle. As in (b), though clicks could provide in-depth annotations, a small number of clicks is unable to indicate the shape and size of the object accurately. Thus, click-based models often require extensive interactions, especially when annotating large and complicated ob-



Figure 1. Comparisons of different interaction forms for interactive image segmentation. Foreground scribbles/clicks are marked in green and background scribbles/clicks in blue.

jects. Scribbles, on the other hand, have the combined advantages of boxes and clicks. Long scribbles can accurately indicate the shape and size of the target, while short scribbles can make detailed corrections. Scribbles are therefore regarded as an extension of clicks as they encode more information about the user’s intention.

Although drawing scribbles is more practical and favorable, this topic is rarely discussed by researchers. The settings for the few existing works [3, 2, 32, 1] vary greatly. They use different dataset, scribble-simulation methods, evaluation metrics, and does not provide code. This makes it hard to make comparisons and hinders the development of scribble-based interactive segmentation. In contrast, click-based interactive segmentation is flourishing with booming works [34, 19, 24, 31, 22, 16, 30, 6, 7]. We believe a core reason is that DIOS [34] (CVPR’16) formulated a training pipeline and evaluation protocol, thus other researchers could follow the standard setting and focus on specific points to improve the performance.

In this work, we attempt to reference the successes of click-based methods to reformulate the task of scribble-based interactive segmentation. However, there exist gaps between these two tasks, and in our exploration, we tackle the following challenges:

How to get diversified scribbles for training? Clicks could simply be represented by a pair of coordinates, but scribbles have arbitrary shapes and complicated representations. Previous scribble-based works majorly use feature similarities to guide the segmentation, thus they [3, 2, 32] do not need training scribbles. [1] simply links randomly sampled points to simulate scribbles, but is too naive to cover different real-world user interactions.

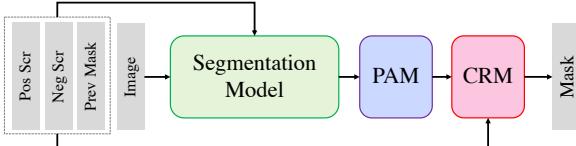


Figure 2. The pipeline of ScribbleSeg. ‘PAM’ denotes Prototype Adaption Module, and ‘CRM’ means Corrective Refine Module.

In this paper, we design multiple meta-simulators to simulate different kinds of annotating behavior and make compositions to ensure the diversity of training samplers. We also adapt the iterative training strategy [24] to add scribbles on the error regions of the last prediction.

How to fairly evaluate the model? During evaluation, we design a deterministic simulator to generate the scribble according to the shape and size of the given mask. We use this method to add positive/negative scribbles automatically on the max difference region between the ground truth and the predicted masks. Thus, similar to click-based settings, we could measure the Number of Interactions required to reach the target IOU. This protocol provides a unified benchmark for different types of interactions, and enables us to compare the performance among models with different interaction forms. Besides, we construct a benchmark based on ADE20K [35] to evaluate the model’s ability in diversified scenarios and categories.

How to fully utilize the indications in scribbles? Scribbles could be regarded as an extension of clicks, thus many designs from click-based methods could be transferred. Differently, scribbles contain more indications than clicks that could be explored. Thus, we first build a vanilla pipeline adapting from click-based methods and make specific designs considering the characteristics of scribbles. As in Fig. 2. We first represent positive and negative scribbles into two binary masks. Then, we feed the 3-channel image, along with two scribble masks, and the previous prediction masks into a segmentation model. This method could be regarded as a vanilla solution. Starting from this baseline, we add two components to further improve its performance.

As scribbles cover more pixels than clicks, they could not only provide the location priors but also the appearance indications (the scribble-covered regions), thus, we develop a Prototype Adaption Module (PAM) to update the final projection kernel according to the user-provided scribbles. Besides, to produce high-quality masks, we design a Corrective Refine Module (CRM), which takes the prediction of the segmentation model as input to estimate the probable error region and make corrections for the details.

Our contribution could be summarized in three folds: 1) We reformulate the task of scribble-based interactive segmentation and provide a standard train/validation protocol and benchmark. 2) We propose ScribbleSeg, which shows strong performance for scribble-based interactive segmen-

tation. 3) We design PAM and CRM, which are simple and effective modules for interactive segmentation.

2. Related Work

Interactive image segmentation with click. Click-based interactive segmentation methods aim to obtain masks of the targeted objects with reference to user-provided clicks. Early methods [9, 4, 10, 18] focused on optimization-based solutions. DIOS [34] was the first deep learning method that proposed embedding positive and negative clicks into distance maps, then stacking them together with the image as input to the network. BRS [16] proposed an online optimization scheme for interactive segmentation, and f-BRS [30] sped it up by optimizing only the auxiliary variables of the network. Later methods [6, 22, 23] have also employed a similar model architecture and provided further improvements. RITM [31] improved model performance by taking the previous mask along with the click maps and image as input. FocalClick [7] performed prediction and update in localized areas, and has improved the model’s efficiency and mask refinement performance. These works all follow the train/val protocol proposed by DIOS [34], and report the performance with the same metrics, thus a good research community is formed. However, the biggest disadvantage of click-based methods is that clicks embed little information, and the model, therefore, requires extensive annotations to segment objects with complicated shapes.

Interactive image segmentation with scribbles. Compared to click-based segmentation, scribble-based interactive segmentation has a lot fewer methods proposed.

Early works [20, 15, 3, 17] used graph constraints, energy functions, or Gabor filters to deal with scribbles. DeepIGeoS [32] encoded scribbles with geodesic distance transforms and performed mask refinement with it. [1] allowed the sharing of scribble annotations across multiple object regions. [2] leveraged the appearance similarity to propagate scribble information to other regions. In the current field, however, there has yet to be a standard training and validation protocol proposed. Researchers would use IOU [1], Dice Coefficient [3], and annotating time [2] as metrics and report evaluation result on different benchmarks.

Scribble-based video object segmentation. DAVIS-2018 [5] provide a track for scribble-based video object segmentation, which aims to produce masks for object annotated in all frames of a video, and use scribbles to make target indications and corrections. Although using scribbles to refine masks is an important step in this task, DAVIS-2018 [5] only cares about the performance on the whole video sequence, and therefore previous works [26, 27, 14, 13, 8] often use a simple module to deal with scribbles and focus on the information propagation between frames.

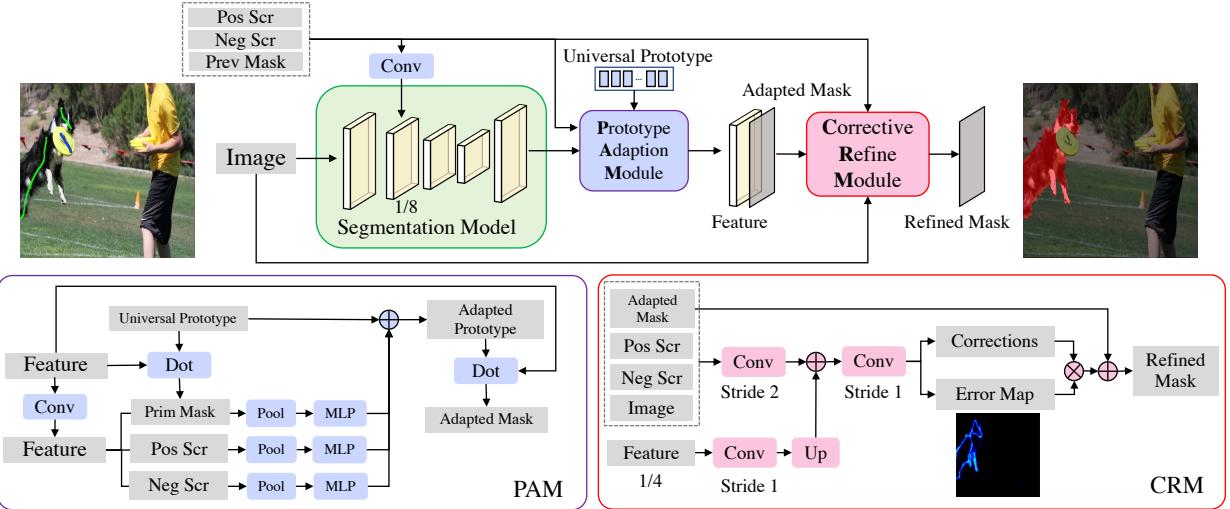


Figure 3. The demonstration for the pipeline of ScribbleSeg. We take the image, the scribble maps, and the previous mask as inputs to extract the mask of the target object. The positive and negative scribbles are marked in green and blue. **PAM** denotes Prototype Adaption Module. **CRM** means Corrective Refine Module. The detailed structure of KAM and CRM can be found at the bottom part of the figure.

3. Method

We first give an overall introduction for the task of scribble-based interactive segmentation and survey the current status of the community in Sec. 3.1. Then, we elaborate on the detailed model structure of ScribbleSeg in Sec. 3.2. Afterward, we introduce the training protocol in Sec. 3.3. In Sec. 3.4, we describe the evaluation method and our constructed benchmark for evaluation.

3.1. Task Overview

Task definition. To extract a target mask, the user sequentially draws positive/negative scribbles to add/remove mask regions, and the model returns the predicted mask after receiving new interactions. For each interaction period, only one additional scribble could be provided.

Existing works. We survey the scribble-based interactive segmentation methods in Tab. 1. Most of the existing works use different protocols (domain, data, metrics), which makes it hard to make comparisons and hinders the development of the community. However, as scribbles contain more information than clicks, it has the potential to be a more promising choice for interactive segmentation. Thus, the community needs a standard framework to unleash the potential of scribble-based interactive segmentation.

	Train Data	Train Scribbles	Test Data	Test Scribbles	Metrics
[1]	COCO	Linked points	COCO	Linked points	IoU per Scribble
[2]	COCO	None	COCO	Human Drawn	Annotation Time
[3]	None	None	GrabCut	Random Pixels	Label Accuracy
[32]	Medical	None	Medical	Random Pixels	Dice Score

Table 1. Survey for existing works of scribble-based interactive segmentation. Their configurations differ from each other.

3.2. Framework for ScribbleSeg

As shown in Fig. 3, we first feed the image and interaction maps into a segmentation model. This is a commonly used baseline solution transferred from click-based methods [31]. To improve its performance, we analyze the characteristics of using scribbles as interactions to design novel components. Accordingly, we develop two modules:

Prototype Adaption. Scribbles often cover more pixels than clicks, thus besides providing position/shape priors like clicks, the scribble-covered regions could better indicate the target appearance and semantics. According to this property, we propose to use the scribble-covered regions to enhance the target extraction procedure.

Masks extraction could be understood as a correlation between the projected features and learned prototypes (the last FC layer for the segmentation model). Traditional semantic segmentation models learn fixed prototypes for each category. However, interactive segmentation models learn only one prototype for the universal targets. As the segmentation target could be any object, stuff, or part, it is challenging to use a single prototype to represent the diverse target.

We propose the Prototype Adaption Module (PAM), which dynamically adapts the universal prototype (the last projection kernel) by interacting with the scribble-covered features. Thus, the parameters of the segmentation model become image-specific and scribble-specific.

As shown in the left bottom part of Fig. 3, we first use the universal prototype as the convolution kernel to generate a primitive mask M_{prim} via dot production with the final feature map. This prototype is initialized with learnable parameters. Afterward, we use the two user-provided scribble maps, and the primitive mask prediction to refine the mask.

feature map into three embeddings. Then, we project those embeddings using MLP layers and add them to the original prototype. As the labels of the scribble-marked regions are known, they contain more cues that indicate the user’s intentions. The primitive mask could also give a global representation for the segmentation target, which helps construct a dynamic prototype with high-consistency representations. Finally, we use this adapted prototype to predict the adapted mask, and note it as M_{ada} .

Corrective Refine. The scribbles could also provide indications for refining the segmentation details. We propose Corrective Refine Module (CRM) to make modifications on the M_{ada} predicted by PAM. As demonstrated in the right bottom of Fig. 3, the first branch of CRM concatenates the predicted mask with the scribble maps and the original image to extract the features with fine details. The second branch fuses the detached features from the segmentation model. The features in CRM are kept at the resolution of stride-2 to preserve the fine details. Afterwards, we predict an error map M_{error} and a correction map M_{corr} . The error map is supervised with the difference between the ground truth and the M_{ada} . The final prediction is a combination of the M_{ada} and M_{corr} in the predicted error region M_{error} . The structure of CRM is inspired by the refiner of FocalClick [7]. The core differences are, we detach the feature of the segmentation model, and we use the error map to guide the detail correction process. The effectiveness of these modifications would be verified in the experiments.

Training losses. The principle supervision is for the final refined prediction, for which we use normalized focal loss (NFL) [31], and note it as \mathcal{L}_{ref} . Besides, we also use NFL to supervise the primitive and adapted masks produced by KAM, and note them as \mathcal{L}_{prim} , and \mathcal{L}_{ada} . In addition, the error map of CRM is supervised with binary cross-entropy loss, \mathcal{L}_{error} . The total loss is a combination of these losses, where α, β, γ all equal 0.4.

$$\mathcal{L}_{total} = \mathcal{L}_{ref} + \alpha\mathcal{L}_{prim} + \beta\mathcal{L}_{ada} + \gamma\mathcal{L}_{error} \quad (1)$$

3.3. Training Scribble Simulation

Meta-simulators. The naive methods [1, 3, 32] use random dilated points to simulate scribbles, which could not satisfy the diversity. Additionally, we develop multiple meta-simulators to generate various scribbles. As demonstrated in Fig. 4, the bezier scribble uses bezier function to draw curves within the mask regions; the axial scribble calculates the media axis of the given mask; the boundary scribble draws lines along with the mask boundary. For the stroke thickness, we randomly choose values from 3 to 7.

Scribble composition. We combine these four strategies to generate diversified scribbles during training with carefully tuned ratios, some of the results could be found in

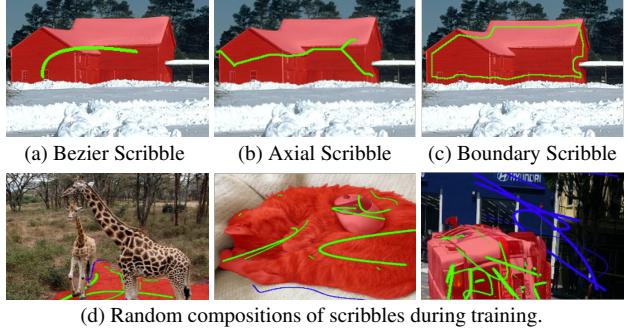


Figure 4. Demonstrations of the meta scribble simulators. We combine these strategies to generate scribbles during training.



Figure 5. The procedure of deterministic scribble generation. The true positives, false negatives, and false positives of the segmentation result are marked in red, green, and blue respectively.

Fig. 4 (d). When starting from a void previous mask, we generate positive scribbles in the foreground regions, and negative scribbles in the background.

Iterative sampling. To better simulate the practical usage, we add scribbles iteratively. When starting from a previous mask, we first calculate the False Positive (FP) and False Negative (FN) regions and generate negative and positive scribbles accordingly. We develop two strategies to simulate the flawed masks. The first one is applying the iterative training schema [24]. Another strategy is to exert random perturbations on the ground truth masks, where we use random dilation, erosion, translation, and local erasing. We combine these two strategies during training and make experiments to find the best combination ratio.

3.4. Evaluation Method

In this section, we introduce our evaluation protocol that could compare different methods fairly and automatically.

Revisiting click-based evaluation. DIOS [34] proposes the evaluation methods that followed by almost all previous click segmentation works [30, 16, 22, 31, 31, 6, 7]. In this setting, the simulated clicks are added sequentially on the center of the maximum error regions for the previous prediction. For example, the first positive click is added at the center of the ground truth mask. Then, we calculate the error regions of the prediction and extract the maximum connected area. Afterwards, an additional positive/negative click would be placed at the center of this maximum connected area. Thus, clicks would be added automatically until the prediction reaches the target IOU, or when the number of clicks reaches the limitation. In this way, we could report NoC85/90 (the average Number of Clicks required to

Algorithm 1 Deterministic Scribble Simulator

```
1: max\_mask = max(error\_mask)
2: skel\_mask = MEDIALAXIS(max\_mask)
3: Graph = RADIUSNEIGHBOURGRAPH(skel\_mask)
4: for subgraph ∈ CONNECTED(Graph) do
5:   while true do
6:     cycle = FINDCYCLE(subgraph)
7:     if cycle == None then break
8:     else REMOVECYCLE(subgraph, cycle)
9:     end if
10:    end while
11:   end for
12:   distance = []
13:   for v ∈ Graph.nodes() do
14:     max\_path = SHORTESTPATH(Graph, v)
15:     distance.append(max\_path)
16:   end for
17:   longest_path = max(distance)
18:   scribble = BEZIERCURVE(longest_path)
```

reach the IOU 85/90%), and NoF²⁰85/90 (the Numbers of Failures to reach IOU 85/90% within 20 clicks).

Deterministic scribble-simulator. We attempt to extend the click-based protocol into a more general form for scribbles. The challenge is that clicks could simply be added at the center of the error region, but not for scribbles, as they have various shapes, which introduces randomness.

Accordingly, We develop a **deterministic** scribble simulator that could simulate human-like scribbles according to the shape and size of the given mask. The pseudo-code for the scribble generation process is shown in Alg. 1, and the demonstration is given in Fig. 5. Similar to the click-based protocol, we first calculate the max error regions. Then, we compute the medial axis for the largest error mask to obtain the skeleton of the objects. Afterwards, we transform the skeleton mask into a radius neighbor graph, where the neighborhood of a vertex is points at a distance less than the radius from it. Then we divide the graph into connected components sub-graphs and remove its cycles. Finally, we will create a Bezier curve with the points in the graph’s longest path. The thickness of the stroke is set to a fixed value (3 as default).

With this deterministic scribble simulator, we could iteratively add scribbles on the FP or FN regions. Thus, we generalize the NoC metric for clicks to NoI (Number of Interactions), and report NoI85/90, NoF²⁰85/90.

Evaluation benchmark. To perform comparisons with previous methods, we first evaluate ScribbleSeg on the benchmarks [29, 25, 12, 28] used by click-based models. However, they have the following disadvantages: GrabCut [29] and Berkeley [25] only have 50 and 100 test images respectively, which could not provide convincing results. SBD [12] contains 2802 samples, but the mask an-

notations are coarse, thus often causing inconsistent results, which has been discussed in [6, 7]. DAVIS [28] contains 345 annotations, but the targets are all saliency objects.

In general, the benchmark above only contains relatively easy cases for the salient object. However, a good annotation tool should be able to deal with large-scope categories in complex scenes. Therefore, we additionally use ADE20K [35] to evaluate our model, which covers both things and stuff for 150 categories. We use the panoptic format annotation for the ADE20K validation split. For each category, we randomly pick 5 samples (we take the max number if there are fewer than 5 samples) and we have obtained 246 samples for stuff and 499 samples for things.

4. Experiments

4.1. Experiment Configurations

Implementation details. The segmentation model in ScribbleSeg could be an arbitrary semantic segmentation network. In this work, we choose the SegFormer [33] for experiments following previous SOTAs [7] of click-based segmentation. The input size of ScribbleSeg is kept as 384×384 during training and inference.

During inference, we apply the zoom-in strategy proposed in [30]. Concretely, starting from the second stroke of scribble, we calculate the external box according to the previous mask and the current scribbles and expand it with a ratio of 1.4. Then, we crop the model input according to this expanded box and resize it to 384×384 . This allows the model to focus on the target region, which is especially effective when the target is small in the image.

Training configurations. Following previous works of click-based segmentation [7, 31], we train our model on the combined dataset of COCO [21] and LVIS [11]. For data augmentation, we use random flip and random resize with a scale ratio from 0.75 to 1.4. We take 3,000 images for each epoch and train ScribbleSeg with 150 epochs. The initial learning rate is set as 0.0005, and we add two lr decay with the ratio of 0.1 at the epoch of 110 and 130. For the optimizer, we pick ADAM with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

4.2. Quantitative Analysis

We first give a quantitative analysis for ScribbleSeg on our newly constructed ADE20K benchmark. Using the proposed evaluation protocol, we evaluate the model performance with NoI@IoU, which means the average Number of Interactions required to reach the target IoU.

As explained in Sec. 3.1, previous works are not suitable for comparison. Hence, we reproduce some representative baselines for comparison in Tab. 2. In row 1, we train a similarity-based model like [2], which shows poor performance as it could not deal with the fine details. Row 2 corresponds to using the click-based solution [31] to deal with

	Method	Train Interaction	ADE-Stuff			ADE-Thing			ADE-Full		
			NoI 85	NoI 90	NoF 90	NoI 85	NoI 90	NoF 90	NoI 85	NoI 90	NoF 90
1	AppearanceSim [2]	None	9.61	15.34	92	10.01	12.12	186	9.87	13.18	278
2	RITM-scribble [31]	Click points	7.20	10.33	76	7.91	10.82	166	7.73	10.65	242
3	RITM-scribble [31]	Linked points [1]	6.11	7.92	58	6.99	8.74	131	6.69	8.47	189
4	ScribbleBase-B0	Composed scribbles	4.89	7.14	41	5.61	8.19	111	5.37	7.84	152
5	ScribbleSeg-B0	Composed scribbles	4.77	6.90	41	5.08	7.63	105	4.97	7.38	146
6	ScribbleSeg-B3	Composed scribbles	4.41	6.57	39	5.06	7.46	109	4.84	7.16	148

Table 2. Comparison results on our proposed ADE20K benchmark. We report the performance for stuff and thing categories respectively. All models are trained on COCO+LVIS. ‘NoI 85/90’ denotes the average Number of Interactions required the get IoU of 85/90%. ‘B0/B3’ denotes using SegFormer-B0/B3 [33] as the segmentation model.

Pred : Pertub	Berkeley		DAVIS		ADE-full	
	NoI 90	NoF 90	NoI 90	NoF 90	NoI 90	NoF 90
0 : 0	2.30	3	6.02	64	8.95	179
1 : 0	2.06	1	5.46	61	8.61	176
1 : 0.2	1.79	0	5.21	52	8.01	156
1 : 0.4	1.77	0	5.18	51	7.98	153
1 : 0.6	1.81	0	5.33	54	8.12	160

Table 3. Comparison results of using different ratios of predicted masks (generated by iterative training) and perturbed ground truth mask to simulate the previous mask.

Max Number	Berkeley		DAVIS		ADE-full	
	NoI 90	NoF 90	NoI 90	NoF 90	NoI 90	NoF 90
8	1.92	1	5.34	56	8.33	164
12	1.81	0	5.17	51	8.11	158
16	1.77	0	5.18	51	7.98	153
20	1.76	0	5.23	52	8.01	155

Table 4. Comparisons for the maximum number of strokes for the simulated scribbles during training.

Proportions	Berkeley		DAVIS		ADE-full	
	NoI 90	NoF 90	NoI 90	NoF 90	NoI 90	NoF 90
0	1.74	0	5.26	52	8.12	155
0.2	1.77	0	5.18	51	7.98	153
0.4	1.79	0	5.29	53	8.09	153

Table 5. We set the Bezier curve as the principle strategy for scribble simulation, and analyze the proportions for the axial and boundary scribble during training.

scribbles. We find that, although clicks could be regarded as short scribbles, directly using click-based models could not get satisfactory results. In row 3, we follow IFIS [1] to simulate scribbles via linking randomly sampled points. This strategy brings improvements compared to using click disks. However, it still shows a big gap compared with our work which is demonstrated in the second part.

Row 4 denotes the baseline version without PAM and CRM, and row 5 and 6 show the full ScribbleSeg using SegFormer-B0/B3 [33] as the segmentation model. It could be observed that ScribbleSeg achieves significantly better performance than its counterparts. On stuff categories, the advantage of ScribbleSeg is even larger. This is because scribbles can provide more indications compared to clicks in stuff categories that cover a relatively big region of arbitrary shapes and complicated semantics.

Scribble Type	Berkeley		DAVIS		ADE-full	
	NoI 90	NoF 90	NoI 90	NoF 90	NoI 90	NoF 90
Allow Error	2.01	1	5.45	57	8.98	183
Clean Boundary	1.77	0	5.18	51	7.98	153
Protect Boundary	1.66	0	5.14	53	7.84	152

Table 6. Different strategies for dealing with the simulated scribbles that are near the boundaries of the ground truth mask.

4.3. Ablations Studies

After verifying our promising performance, in this section, we dive into the details of our framework, including the basic settings of scribble-simulation, and the two novel components: Prototype Adaption Module (PAM) and Corrective Refine Module (CRM).

We first make an analysis of the basic settings to explore what makes a strong baseline for scribble-based interactive segmentation. We use a vanilla model without PAM and CRM and mainly focus on exploring the strategies of simulating the previous masks and the scribbles during training.

Flawed masks simulation. We first analyze the simulation of previous masks. In Tab. 3, we explore the combined ratio for two kinds of previous mask generation methods introduced in Sec. 3.3. ‘Pred’ denotes the predicted masks of the current model, which is generated by iterative training [24]. ‘Pertub’ means the perturbed mask of the ground truth. The results show that the previous masks are important for interactive segmentation, and we choose the combination ratio that achieves the best performance.

Scribble simulation. Afterwards, we make explorations for the scribble simulation strategies. In Tab. 4, we report the performance of using different numbers of scribble strokes. During training, we set the maximum number of strokes, and randomly pick a stroke number with a probability decay of 0.8, which means that the probability of choosing n strokes is 0.8 of $n - 1$.

In Tab. 5, we tune the combined ratio for the three types of scribbles introduced in Sec. 3.3. We use the Bezier curve as the principle strategy, and tune the ratios of axial and boundary scribbles. As the Bezier curve is the closest to human-drawn scribbles, a higher portion results in better performance. At the same time, a small portion of the axial

Method	Berkeley		DAVIS		ADE-full	
	NoI 90	NoF 90	NoI 90	NoF 90	NoI 90	NoF 90
StrongBase	1.66	0	5.14	53	7.84	152
+PAM	1.68	0	4.98	52	7.52	152
+CRM	1.53	0	4.76	51	7.46	148
+PAM+CRM	1.49	0	4.68	50	7.38	146

Table 7. Ablation studies for our novel components.

Method	Berkeley		DAVIS		ADE-full	
	NoI 90	NoF 90	NoI 90	NoF 90	NoI 90	NoF 90
StrongBase	1.66	0	5.14	53	7.84	152
+Scribble Pool	1.66	0	5.01	51	7.60	151
+Mask Pool	1.64	0	5.06	53	7.63	153
+Full PAM	1.68	0	4.98	52	7.52	152

Table 8. Ablation studies for the details of PAM.

Method	Berkeley		DAVIS		ADE-full	
	NoI 90	NoF 90	NoI 90	NoF 90	NoI 90	NoF 90
StrongBase	1.66	0	5.14	53	7.84	152
+Refiner [7]	1.61	0	5.01	52	7.66	153
+CRM w/o Detach	1.55	0	4.95	52	7.64	151
+CRM w/o Error Map	1.52	0	4.87	52	7.65	152
+CRM w/o Scribbles	1.58	0	4.93	51	7.56	150
+Full CRM	1.53	0	4.76	51	7.46	148

Table 9. Ablation studies for the details of CRM.

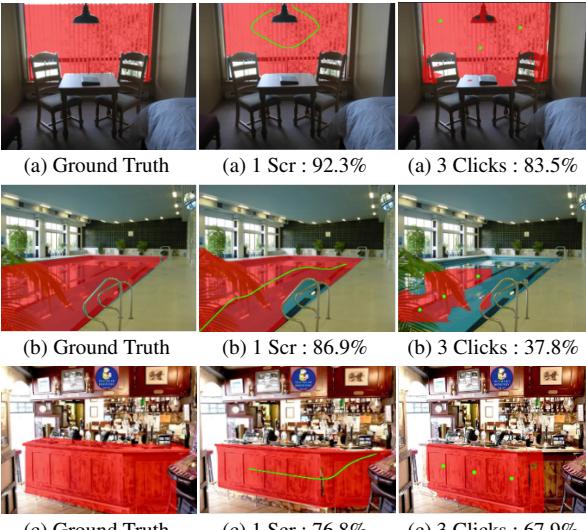


Figure 6. Comparisons of scribble-based method (ScribbleSeg) and click-based method (FocalClick [7]).

and boundary scribbles could increase the training diversity, which is also beneficial.

Tab. 6 shows that dealing with scribbles in boundary regions is also important. ‘Allow Error’ means allowing the simulated scribbles to slightly exceed the ground truth masks. ‘Clean Boundary’ denotes removing the exceeded parts of simulated scribbles. ‘Protect Boundary’ describes eroding the ground truth mask as the target region to simulate scribbles, and is proven to be effective.

After tuning the settings explored above, we get a strong baseline for scribble-based interactive segmentation, which already displays great performance according to Tab. 10. In the next section, we add PAM and CRM to make further improvements and analyze the details of these two modules. In Tab. 7, we show that PAM and CRM could enhance the

Method	Berkeley [25]		SBD [12]		DAVIS [28]	
	NoI 90	NoI 85	NoI 90	NoI 85	NoI 90	NoI 85
f-BRS-B-hr32 [30]	2.44	4.37	7.26	5.17	6.50	
RITM-hr18s [31]	2.60	4.04	6.48	4.70	5.98	
RITM-hr32 [31]	2.10	3.59	5.71	4.11	5.34	
FocalClick-hr18s-S2 [7]	2.66	4.43	6.79	3.90	5.25	
FocalClick-B0-S2 [7]	2.27	4.56	6.86	4.04	5.49	
FocalClick-B3-S2 [7]	1.92	3.53	5.59	3.61	4.90	
ScribbleBase-B0	1.66	2.18	4.50	3.67	5.14	
ScribbleSeg-B0	1.49	2.56	4.21	3.29	4.68	
ScribbleSeg-B3	1.35	2.42	3.99	3.10	4.45	

Table 10. Evaluation results on Berkeley, SBD and DAVIS datasets. ‘NoI 85/90’ denotes the average Number of Interactions (clicks or scribbles) required to get IoU of 85/90%.

performance of the strong baseline independently, and combining both of them could make further improvements.

Prototype Adaption. PAM gathers information from the scribble-marked regions and the mask regions to update the projection kernel. This assists ScribbleSeg in making more consistent predictions. PAM is composed of mask-pooling and scribble-pooling-guided prototype adaption. The results in Tab. 8 demonstrate that both of these two modules are effective.

Corrective Refine. CRM makes detailed refinement in the predicted error regions. In Tab. 9, we make analyses for the different implementations. The results show that the error map is important for CRM, as it enables CRM to focus on the fine details. Detaching the feature and masks from the previous stage also brings improvements.

4.4. Comparisons with Click-based Methods

Considering that our evaluation protocol introduced in Sec. 3.4 is compatible with click-based methods in a general form, we could directly compare our ScribbleSeg with previous click-based solutions. We first compare ScribbleSeg with the SOTA solutions for the click-based setting to show the benefits of using scribbles as the interaction form and prove the effectiveness of our method.

Quatitative comparisons. In Tab. 10, we list the click-based SOTA methods, and use our generalized metrics introduced in Sec. 3.4 to perform comparisons for our ScribbleSeg. We also report the performance of our strong baseline without PAM and CRM. The results show that our baseline already surpasses all click-based methods. This reflects the superiority of using scribbles as the interaction format. With PAM and CRM, the full version ScribbleSeg gets steady improvements.

Qualitative results. In Fig. 6, we compare ScribbleSeg with the click-based SOTA method FocalClick [7]. The results show that when given only one stroke of scribble, ScribbleSeg could outperform FocalClick with 3 clicks. It

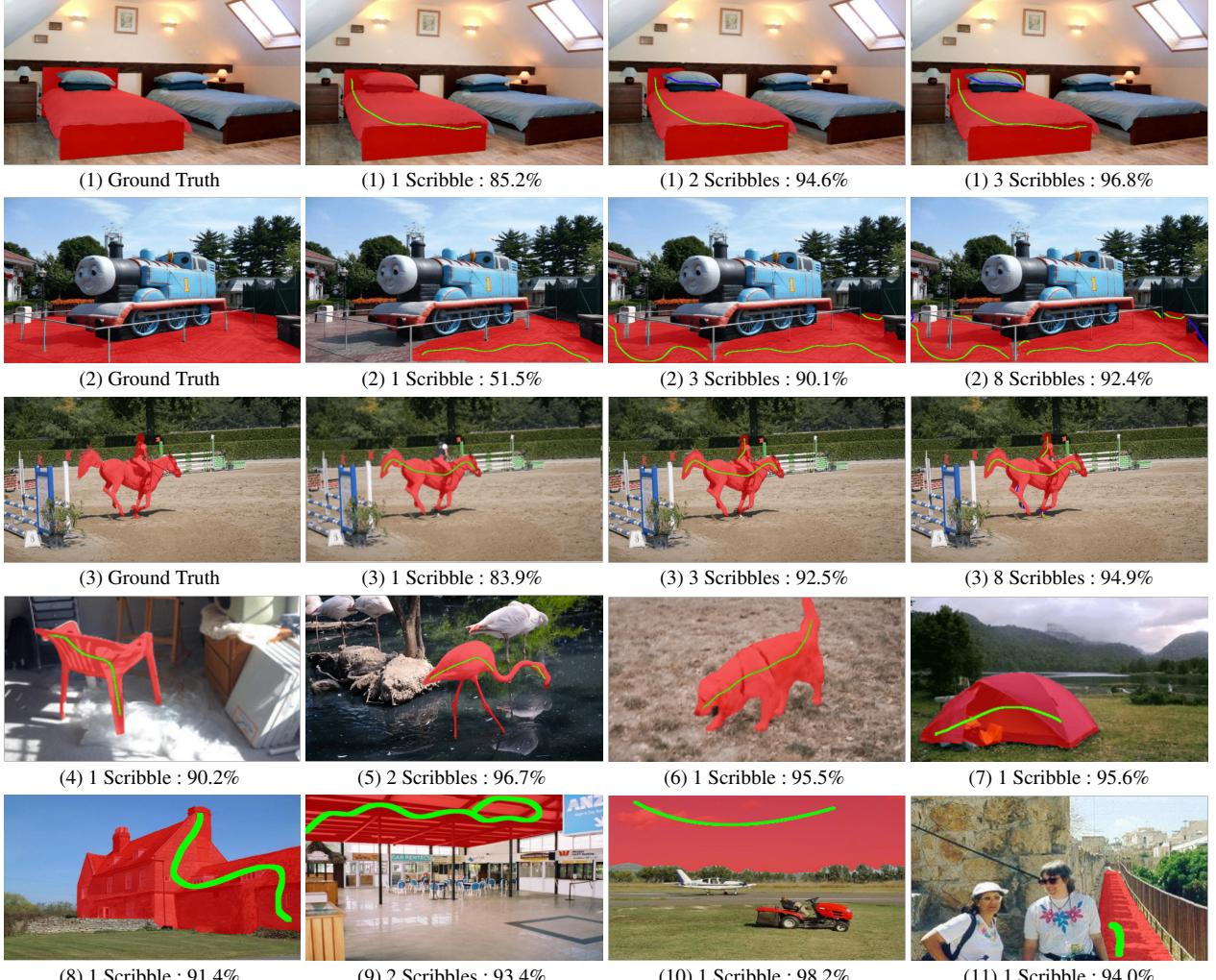


Figure 7. Visualization results for ScribbleSeg-B3 on ADE20K [35] and DAVIS [28]. The numbers of scribbles and the IOU is marked below each image. The positive and negative scribbles are marked in green and blue. Demos 1-7 show the deterministic scribbles, which demonstrate our automatic evaluation procedure. Demos 8-11 show the user customer scribbles.

is clear that scribbles could provide significantly more indications than clicks. We hope ScribbleSeg could serve as a preferred choice for interactive segmentation.

4.5. Qualitative Results

The qualitative results for ScribbleSeg are demonstrated in Fig. 7, where we use SegFormer-B3 [33] as the segmentation model and make predictions on DAVIS [28] and ADE20K [35] benchmarks.

Evaluation procedure. In 1-4 rows, we show the evaluation procedure for sequentially added scribbles with the deterministic simulator. Results show that ScribbleSeg performs well on both things and stuff across diverse scenes.

User customer scribbles. The examples in the demo 8-11 show the user given scribbles with arbitrary shapes and thicknesses. It demonstrates the robustness and generalization ability of ScribbleSeg for different interactions.

5. Limitation

Although ScribbleSeg shows great performance across different benchmarks, this version of the model would only be applicable to natural images as it is only trained on COCO [21] and LVIS [11]. If we want to use it on other domains like industrial defects and medical images, we have to collect data from the target domain and finetune the model.

6. Conclusion

We are the first to formally address the task of scribble-based interactive image segmentation. We have constructed the standard train/val protocol and propose ScribbleSeg. Our method shows clear advantages compared with previous click-based models. We hope this work could serve as the baseline and assist the community in making further explorations on scribble-based interactive image segmentation and developing more powerful mask annotation tools.

References

- [1] Eirikur Agustsson, Jasper R. R. Uijlings, and Vittorio Ferrari. Interactive full image segmentation by considering all regions jointly. In *CVPR*, 2019. 1, 2, 3, 4, 6
- [2] Mykhaylo Andriluka, Stefano Pellegrini, Stefan Popov, and Vittorio Ferrari. Efficient full image interactive segmentation by leveraging within-image appearance similarity. *arXiv:2007.08173*, 2020. 1, 2, 3, 5, 6
- [3] Junjie Bai and Xiaodong Wu. Error-tolerant scribbles based interactive image segmentation. In *CVPR*, 2014. 1, 2, 3, 4
- [4] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *ICCV*, 2001. 2
- [5] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv:1803.00557*, 2018. 2
- [6] Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. Conditional diffusion for interactive segmentation. In *ICCV*, 2021. 1, 2, 4, 5
- [7] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *CVPR*, 2022. 1, 2, 4, 5, 7
- [8] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, 2021. 2
- [9] Leo Grady. Random walks for image segmentation. *TPAMI*, 2006. 2
- [10] Varun Gulshan, Carsten Rother, Antonio Criminisi, Andrew Blake, and Andrew Zisserman. Geodesic star convexity for interactive image segmentation. In *CVPR*, 2010. 2
- [11] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 5, 8
- [12] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 5, 7
- [13] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Interactive video object segmentation using global and local transfer modules. In *ECCV*, 2020. 2
- [14] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Guided interactive video object segmentation using reliability-based attention maps. In *CVPR*, 2021. 2
- [15] Enhua Wu Hong Li, Wen Wu. Robust interactive image segmentation via graph-based manifold ranking. In *Computational Visual Media*, 2015. 2
- [16] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *CVPR*, 2019. 1, 2, 4
- [17] Majeed Kassis and Jihad El-Sana. Scribble based interactive page layout segmentation using gabor filter. In *ICFHR*, 2016. 2
- [18] Tae Hoon Kim, Kyoung Mu Lee, and Sang Uk Lee. Non-parametric higher-order learning for interactive segmentation. In *CVPR*, 2010. 2
- [19] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In *CVPR*, 2018. 1
- [20] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016. 2
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5, 8
- [22] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive image segmentation with first click attention. In *CVPR*, 2020. 1, 2, 4
- [23] Qin Liu, Meng Zheng, Benjamin Planche, Srikrishna Karanam, Terrence Chen, Marc Niethammer, and Ziyu Wu. Pseudoclick: Interactive image segmentation with click imitation. *arXiv:2207.05282*, 2022. 2
- [24] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. In *BMVC*, 2018. 1, 2, 4, 6
- [25] Kevin McGuinness and Noel E O’connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 2010. 5, 7
- [26] Naveen Shankar Nagaraja, Frank R. Schmidt, and Thomas Brox. Video segmentation with just a few strokes. In *ICCV*, 2015. 2
- [27] Seoung Oh, Joon-Young Lee, Ning Xu, and Seon Kim. Fast user-guided video object segmentation by interaction-and-propagation networks. In *CVPR*, 2019. 2
- [28] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 5, 7, 8
- [29] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *TOG*, 2004. 5
- [30] Konstantin Sofiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *CVPR*, 2020. 1, 2, 4, 5, 7
- [31] Konstantin Sofiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *ICIP*, 2022. 1, 2, 3, 4, 5, 6, 7
- [32] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *TPAMI*, 2018. 1, 2, 3, 4
- [33] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 5, 6, 8
- [34] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *CVPR*, 2016. 1, 2, 4
- [35] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 2, 5, 8