Variance-Insensitive and Target-Preserving Mask Refinement for Interactive Image Segmentation

Chaowei Fang[1], Ziyin Zhou[1], Junye Chen[2], Hanjing Su[3], Qingyao Wu[4], Guanbin Li[2,5]*

[1]School of Artificial Intelligence, Xidian University, Xi'an, China
[2]School of Computer Science and Engineering, Research Institute of Sun Yat-sen University in Shenzhen, Sun Yat-sen University, Guangzhou, China
[3]Tencent
[4]School of Software Engineering, South China University of Technology, Guangzhou, China
[5]GuangDong Province Key Laboratory of Information Security Technology

Abstract

Point-based interactive image segmentation can ease the burden of mask annotation in applications such as semantic segmentation and image editing. However, fully extracting the target mask with limited user inputs remains challenging. We introduce a novel method, Variance-Insensitive and Target-Preserving Mask Refinement to enhance segmentation quality with fewer user inputs. Regarding the last segmentation result as the initial mask, an iterative refinement process is commonly employed to continually enhance the initial mask. Nevertheless, conventional techniques suffer from sensitivity to the variance in the initial mask. To circumvent this problem, our proposed method incorporates a mask matching algorithm for ensuring consistent inferences from different types of initial masks. We also introduce a target-aware zooming algorithm to preserve object information during downsampling, balancing efficiency and accuracy. Experiments on GrabCut, Berkeley, SBD, and DAVIS datasets demonstrate our method's state-of-the-art performance in interactive image segmentation.

Introduction

Interactive image segmentation (IIS) serves as a prominent method for the extraction of binary masks corresponding to targeted objects, guided by user interaction cues. It holds substantial significance in diverse applications, ranging from easing the burden of data annotation in semantic segmentation to acting as essential components in image editing tasks such as image inpainting (Bertalmio et al. 2000). The field of IIS recognizes a variety of interaction cues, including points (Xu et al. 2016), bounding boxes (Yu et al. 2017), and scribbles (Lin et al. 2016). Within this study, our focus is drawn to the utilization of positive and negative click points (see Fig. 1). The central challenge lies in the generation of accurate object masks with minimum clicks.

The earliest point guided IIS method built upon deep neural networks (DNN) can be traced to (Xu et al. 2016). Subsequent developments have introduced iterative methods to increase the flexibility of model training. For instance, Mahadevan, Voigtlaender, and Leibe (2018) unveiled an iterative training framework that autonomously samples pseudo click points relative to error maps between predicted and ground-truth (GT) segmentations, thereby enhancing the adaptability of IIS. Similarly, Sofiiuk, Petrov, and Konushin (2022a) strategized an approach to iteratively refine the previously generated mask in the current interaction phase. Recent contributions such as (Lin et al. 2022) and (Chen et al. 2022) have expanded upon this foundation by employing a coarse-to-refine framework. This methodology initially executes a preliminary coarse segmentation from a global low-resolution perspective and subsequently refines the details

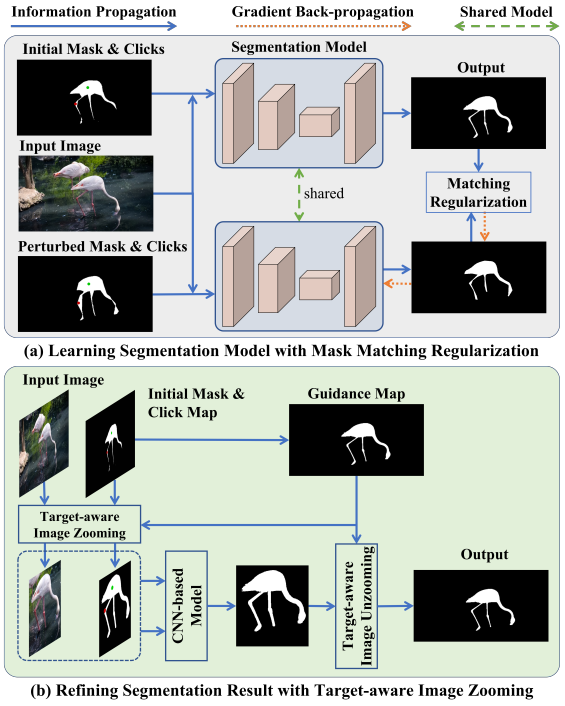

Figure 1: We enhance the interactive image segmentation model's robustness against initial mask fluctuations using mask matching regularization (a) and introduce a target-aware zooming operation (b) for image downsampling.

*Corresponding author.

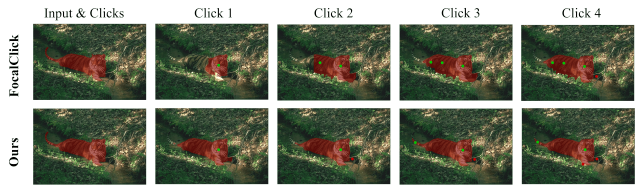

Figure 2: Segmentation results of FocalClick and our method across varying click numbers. Green and red points indicate foreground and background clicks, respectively.

from a localized high-resolution viewpoint.

Despite recent advancements, existing methods still necessitate a moderate number of interaction points to attain satisfactory performance. A significant issue with current techniques (Sofiiuk, Petrov, and Konushin 2022a; Lin et al. 2022; Chen et al. 2022), is their sensitivity to the fluctuation of the initial mask. During each training step, the initial mask is either set as fully zero or derived from the model's previous prediction. Consequently, models trained in this fashion exhibit a lack of robustness when addressing the fluctuation of initial masks during inference. The other prevailing challenge is that existing approaches typically rely on conventional downsampling techniques, such as bilinear interpolation, to enhance inference efficiency. This process inevitably leads to information loss, complicating the discrimination of the target object. As seen in Fig. 2, a state-of-the-art method FocalClick (Chen et al. 2022) still requires quite a few clicks to delineate the complete object.

In response to these challenges, we introduce a novel algorithm entitled *Variance-Insensitive and Target-preserving Mask Refinement*. Centered on a DNN model, our method predicts the segmentation map for an image, utilizing the click map and initial mask as supplementary inputs. To fortify the model against initial mask fluctuation, we propose a mask matching regularization strategy. This involves generating two initial mask variants: 1) utilizing the segmentation map from the first interaction step; 2) synthesizing a mask by distorting the GT mask. Drawing inspiration from the smoothness assumption (Bonaccorso 2017), we establish a regularization term between model predictions for the two initial mask variants (Fig. 1 (a)). This innovation allows greater flexibility in selecting the initial mask, diversifying beyond conventional methods like RITM (Sofiiuk, Petrov, and Konushin 2022b) and FocalClick (Chen et al. 2022). Furthermore, inspired by (Thavamani et al. 2023), we incorporate a Target-Aware Image Zooming (TAIZ) operation (Fig. 1 (b)), to mitigate information loss during image downsampling. Our TAIZ operation uniquely leverages the combination of the last segmentation map and click map to generate a re-sampling grid that redirects points outside the salient region to inside it. The efficacy of our approach is evident in Fig. 2, which illustrates superior results with even fewer clicks compared to FocalClick. Comprehensive experiments conducted on four public datasets, namely GrabCut (Rother, Kolmogorov, and Blake 2004), Berkeley (Martin et al. 2001), SBD (Hariharan et al. 2011), and DAVIS (Perazzi et al. 2016), demonstrate that our method sets a new benchmark for state-of-the-art performance.

Main contributions of this paper are summarized as below.

- We develop an innovative IIS framework with the mask matching regularization. This alleviates the model's sensitivity to variances in the initial mask.
- We introduce a target-aware image zooming operation which can maintain the intrinsic characteristics of the target object during the input image downsampling process.
- We conduct extensive experiments on GrabCut, Berkeley, SBD, and DAVIS datasets. The results affirm that our method significantly surpasses existing methods.

Related Work

Interactive Image Segmentation

Interactive image segmentation (IIS) has been a longstanding challenge in computer vision. The advent of deep learning in semantic segmentation led to its application in IIS by Xu et al. (2016), establishing a mainstream approach. Early deep learning-based IIS methods (Liew et al. 2017; Xu et al. 2016, 2017) overlooked the information contained in previously generated masks. Mahadevan, Voigtlaender, and Leibe (2018) recognized the importance of previous segmentation results as additional inputs, a concept subsequently adopted by many researchers (Lin et al. 2020; Chen et al. 2022; Wei, Zhang, and Yong 2023; Zhou et al. 2023).

Hao et al. (2021a) sought to fully utilize generated masks by implementing multi-stage feature fusion, while others performed a coarse segmentation and then incorporated additional modules to refine the coarse segmentation output (Chen et al. 2021; Hao et al. 2021b). However, such approaches may substantially increase the inference time. To accelerate the mask refinement process, Wei, Zhang, and Yong (2023) attempted to accelerate local refinement through similarity-driven updates, and Chen et al. (2022) exploited local refinement by focusing on specific regions.

However, existing methods suffer from the loss of critical visual information during the downsampling process, which is typically employed to ensure efficiency in inference. This loss is particularly detrimental to identifying intricate aspects of the target object such as boundaries and small-scale components. To address this challenge, we introduce a new target-aware image zooming (TAIZ) algorithm. Unlike traditional downsampling methods, TAIZ accentuates the content of the target object, thus offering a more nuanced understanding of the target object.

Consistency Regularization

In line with the smoothness hypothesis (Bonaccorso 2017), a model should exhibit robustness against variations in the input, meaning that the introduction of noise to input samples should not significantly affect the model's inference. This principle has inspired many semi-supervised learning techniques e.g. (Xie et al. 2020; Sajjadi, Javanmardi, and Tasdizen 2016). It also has important implications for IIS methods (Sofiiuk, Petrov, and Konushin 2022a) based on iterative

mask refinement. Such methods are sensitive to variations in the initial mask. To address this issue, we propose a mask matching regularization strategy, which enhances model robustness by enforcing the model to generate consistent predictions across different variations of the initial mask.

Methodology

This work addresses the problem of point-based interactive image segmentation. Given an image $I \in \mathbb{R}^{H\times W\times 3}$, we train a DNN model to extract the target object through T interaction steps. In addition to the image and cumulative click points, the inputs include the segmentation map obtained from the previous interaction step as the initial mask.

Framework Overview

As shown in Fig. 3, we follow (Chen et al. 2022) to construct the network architecture which is consisting of a coarse segmentation module and a local refinement module. The coarse segmentation module generates a segmentation map from the input image I, guided by an initial mask and click map. Let the initial mask be denoted by $M_0 \in \mathbb{R}^{H\times W}$, and the click map be represented as a two-dimensional disc map $D \in \mathbb{R}^{H\times W\times 2}$, indicating the positions of positive and negative clicks. The segmentation logit of the coarse segmentation module is defined as $O_1 \in \mathbb{R}^{H\times W}$, where $O_1 = \mathcal{F}_{coarse}(I, D, M_0)$ and $\mathcal{F}_{coarse}(\cdot)$ denotes the inference function of the coarse segmentation module. The coarse segmentation map M_1 is then obtained by thresholding O_1. The local refinement module is targeted at improving the segmentation in a specific local region determined by the maximum connected region in the difference map between M_1 and M_0. Concretely, it extracts patches from I, O_1, D, and the penultimate feature map of the coarse segmentation module, corresponding to the local region. Then, it regards theses patches as inputs, generating refined local segmentation logit $\hat{O}_1$ which is subsequently used for updating the coarse segmentation result.

To fortify the robustness against fluctuations in the initial mask, we introduce a regularization approach that ensures consistency between coarse segmentation results derived from different forms of initial masks. Additionally, we design a target-aware image zooming algorithm to retain the target content while downsampling the input image.

Learning with Mask Matching Regularization

In point-based IIS, a common approach is the iterative refinement of the current segmentation map with newly incorporated clicks (Sofiiuk, Petrov, and Konushin 2022a; Lin et al. 2022). While this iterative pipeline efficiently leverages previously generated segmentation results, it often lacks robustness against changes in the initial mask, as it relies on either a blank mask or the segmentation map of the last interaction step during training.

To mitigate this limitation, we introduce a novel regularization approach called *Mask Matching Regularization* (MaskMatch). Specifically, each training sample comprises four elements: the input image I, initial mask M_0, click map D, and ground-truth segmentation map G. Following (Chen et al. 2022), new positive (negative) clicks are synthesized from false negative (false positive) pixels in M_0. Next, two temporary masks $M'_{0,1}$ and $M'_{0,2}$ are generated according to two distinct strategies: 1) We input I, a blank mask, and a synthesized click point into the coarse segmentation module, yielding $M'_{0,1}$. 2) $M'_{0,2}$ is generated by perturbing G with boundary adjustment and region interference operations as in (Cheng et al. 2020) continuously, until it reaches the IoU value of $M'_{0,1}$.

With $M'_{0,1}$ and $M'_{0,2}$, the interactive segmentation process restarts for K additional steps (where K is randomly chosen from $\{0, 1, 2, 3\}$), producing two segmentation masks $M_{0,1}$ and $M_{0,2}$, respectively. Two click maps D_1 and D_2 can be acquired according to $M_{0,1}$ and $M_{0,2}$, respectively. These elements are then fed into the coarse segmentation module to generate two segmentation logits $O_{1,1} = \mathcal{F}_{coarse}(I, D_1, M_{0,1})$ and $O_{1,2} = \mathcal{F}_{coarse}(I, D_2, M_{0,2})$. A matching regularization term between $O_{1,1}$ and $O_{1,2}$ is then established:

$$L_{mr} = \Gamma(\mathbb{1}(\sigma(O_{1,1}) > 0.9) \circ \ell_{bce}(\sigma(O_{1,2}), \sigma(O_{1,1}))), \tag{1}$$

where $\circ$ denotes the element-wise product; $\mathbb{1}(\cdot)$ is the indicator function; $\sigma(\cdot)$ is the Sigmoid function; $\ell_{bce}(\cdot)$ represents the binary cross entropy function; $\Gamma(\cdot)$ is the element-wise average function. The optimization process is stabilized by back-propagating the gradient through $O_{1,2}$ but not $O_{1,1}$, and selecting only high-confidence pixels in $O_{1,1}$ for the computation of the regularization term.

The conventional supervised learning objective constrains predictions on $(I, D_1, M_{0,1})$:

$$\begin{aligned} L_{sup} =& \Gamma(\ell_{nf}(\sigma(O_{1,1}), G)) \\ &+ \Gamma(\ell_{nf}(\sigma(\hat{O}_{1,1}), \hat{G}) + \ell_{nf}(\hat{E}_{1,1}, \hat{G}_e)), \end{aligned} \tag{2}$$

where $\hat{O}_{1,1}$ is the segmentation logit of the refinement module; $\hat{G}$ is the GT mask of the local view; $\hat{E}_{1,1}$ is the edge map predicted by an auxiliary branch in the refinement module following (Chen et al. 2022), and $\hat{G}_e$ represents the GT of the edge map; $\ell_{nf}(\cdot)$ denotes the normalized focal loss function proposed in (Sofiiuk, Barinova, and Konushin 2019).

The overall loss function is formed by combining these two terms in Eq. (1) and (2):

$$L = L_{sup} + \mathbb{1}(\text{IoU}(M_{0,1}, G) > \alpha) \times L_{mr}, \tag{3}$$

where α $(= 0.8)$ is a constant; $\text{IoU}(M_{0,1}, G)$ calculates the intersection-over-union between $M_{0,1}$ and G. Here, if the quality of $M_{0,1}$ is not high, the mask matching regularization would be ignored, since $O_{1,1}$ may not be able to provide accurate supervision under such circumstance.

Target-Aware Image Zooming

Traditional methods often downsample images using standard interpolation algorithms to save computation but compromise essential visual details. Drawing inspiration from (Thavamani et al. 2023), we introduce *Target-Aware Image Zooming* (TAIZ), ensuring the preservation of crucial object information while reducing image resolution.

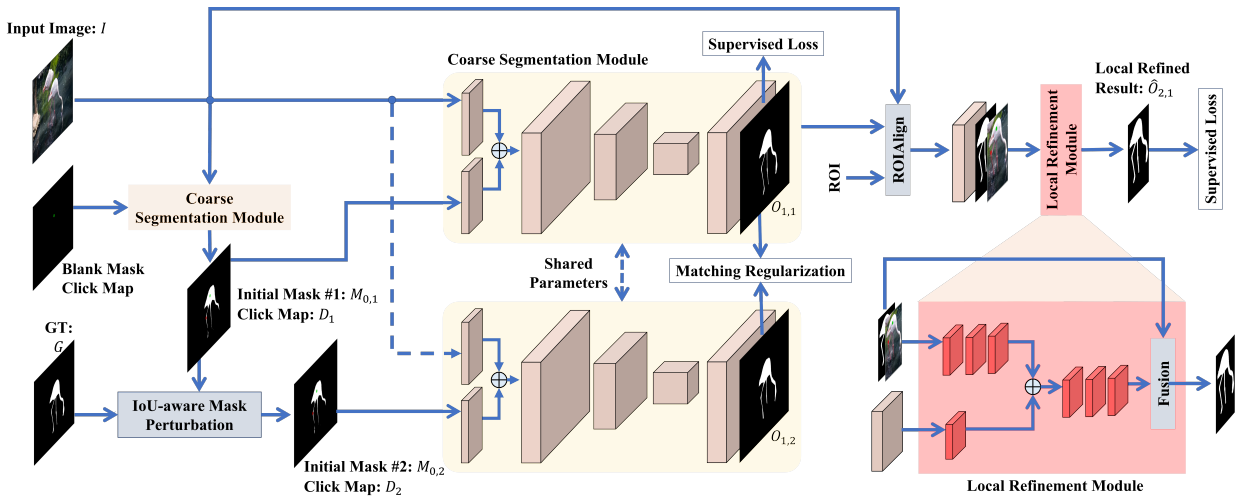

Figure 3: Workflow of our method during training. The network architecture is composed of a coarse segmentation module and a refinement module. Two types of initial masks are adopted to construct the mask matching regularization.

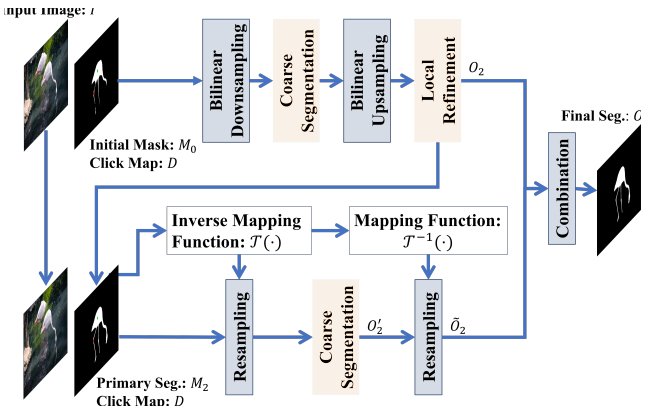

Figure 4: Inference process of our method.

The TAIZ operation counteracts the loss of visual details by employing denser pixel sampling in areas of interest. We initiate by defining a guidance map $S \in \mathbb{R}^{H \times W}$ which specifies pixel-wise locations for the target object. An inverse mapping function $\mathcal{T} : [0,1]^2 \rightarrow [0,1]^2$ is constructed to map a point (x, y) in the target image to $(\mathcal{T}_x(x), \mathcal{T}_y(y))$ in the source image. $\mathcal{T}_x(x)$ and $\mathcal{T}_y(y)$ determine the horizontal and vertical coordinates, respectively. Marginalizing S horizontally and vertically yields vectors $S_y \in \mathbb{R}^H$ and $S_x \in \mathbb{R}^W$ respectively, which are represented as $S_y = S \cdot \mathbf{1}^{W \times 1}$ and $S_x = (\mathbf{1}^{1 \times H} \cdot S)^{\mathrm{T}}$. Here, S_y and S_x reflect the importance levels of rows and columns, respectively. Then, $\mathcal{T}_x(x)$ and $\mathcal{T}_y(y)$ are calculated as below:

$$\mathcal{T}_x(x) = \frac{\int_{x'} x' S_x(x') \mathcal{K}(x, x')\, dx'}{\int_{x'} S_x(x') \mathcal{K}(x, x')\, dx'}, \tag{4}$$

$$\mathcal{T}_y(y) = \frac{\int_{y'} y' S_y(y') \mathcal{K}(y, y')\, dy'}{\int_{y'} S_y(y') \mathcal{K}(y, y')\, dy'}, \tag{5}$$

where $\mathcal{K}(x, x') = e^{-\frac{(x-x')^2}{2\sigma^2}}$ (with σ as the standard deviation) is the Gaussian kernel function. Owing to the weight modulation from S_x and S_y, this inverse mapping function concentrates insignificant points towards salient ones in the guidance map S. Hence, it can be used to downsample images without substantial information loss for the salient regions implied by S as visualized in Fig. 1 (b).

During training, the above TAIZ operation is used to distort half of images by regarding the GT mask as the guidance map, optimizing segmentation performance on TAIZ-processed images. During testing, the input image I, click map D, and initial mask M_0 are inputted into the interaction segmentation pipeline, producing a segmentation logit O_2. Bilinear interpolation is used to downsample inputs for efficiency. Applying a threshold of 0 to O_2 yields mask M_2. The union of M_2 and D serves as the guidance map for creating the inverse mapping function $\mathcal{T}(\cdot)$. Using this function, we obtain the low-resolution versions of I, M_2 and D as I', M_2', and D', respectively. These are then fed into the coarse segmentation module, producing O_2'. This is resampled to the original space using $\mathcal{T}^{-1}$, creating $\tilde{O}_2$. The final segmentation logit, O, is generated by combining O_2 and $\tilde{O}_2$ with the following formulation: $O = (1 - \lambda_t)O_2 + \lambda_t \tilde{O}_2$, where t is the interaction round. Considering the quality of guidance map is not high in early interaction rounds, we set λ_t to 0 if $t < T/2$; otherwise, $\lambda_t = \max(T/2, t)/T$. The inference process is depicted in Fig 4.

Experiments

Datasets and Evaluation Metrics

Datasets. The training images are collected from COCO (Lin et al. 2014) and LVIS (Gupta, Dollar, and Girshick 2019) datasets, containing 1.04×10^5 images and 1.6 million instance-level masks. Four publicly available datasets are used for evaluating IIS methods:

Method	GrabCut		Berkeley		SBD		DAVIS	
	NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90
RIS-Net	-	5.00	-	6.03	-	-	-	-
LD-vgg19	3.20	4.79	-	-	-	-	5.95	9.57
CAG-fcn8s	-	3.58	-	5.60	-	-	-	-
BRS-densenet	2.60	3.60	-	5.08	6.59	9.78	5.58	8.24
FCA-resnet101	-	2.24	-	4.23	-	-	-	7.90
FCA-res2net	-	2.08	-	3.92	-	-	-	7.57
f-BRS-resnet101	2.30	2.78	-	4.57	4.81	7.73	5.04	7.81
CDNet-resnet101	2.42	2.76	-	3.65	4.73	7.66	5.33	6.97
RITM-hrnet18s	1.54	1.68	-	2.6	4.04	6.48	4.7	5.98
RITM-hrnet18	1.42	1.54	-	2.26	3.80	6.06	4.36	5.74
RITM-hrnet32	1.46	1.56	-	2.10	3.59	5.71	4.11	5.34
EdgeFlow-hrnet18	1.60	1.72	-	2.40	-	-	4.54	5.77
FICI-hrnet18s	1.50	1.56	-	2.05	3.88	6.24	3.7	5.16
FICI-hrnet18	1.38	1.46	-	1.96	3.63	5.83	3.97	5.16
FocalClick-hrnet18s	1.48	1.62	1.60	2.23	4.43	6.79	3.90	5.23
FocalClick-segformerB0	1.40	1.66	1.59	2.27	4.56	6.86	5.04	5.49
FocalClick-segformerB3	1.44	1.50	1.55	1.92	**3.53**	5.59	3.61	4.90
Ours-segformerB0	1.42	1.54	1.64	2.18	4.43	6.75	3.81	5.39
Ours-segformerB3	**1.38**	**1.42**	**1.44**	**1.72**	3.55	**5.53**	**3.26**	**4.82**

Table 1: Performance of interactive image segmenation methods evaluated with NoC metrics on GrabCut, Berkeley, SBD, and DAVIS. Lower metric values indate better performance, and the best results are indicated by bold digits.

Method	Berkeley				DAVIS			
	NoF@90 ↓	IoU@5 ↑	BIoU@5 ↑	SPC ↓	NoF@90 ↓	IoU@5 ↑	BIoU@5 ↑	SPC ↓
f-BRS-B-resnet101	6	0.875	0.73	0.072	77	0.826	0.717	0.102
FCA-Net-resnet101	7	0.923	0.793	0.059	74	0.867	0.771	0.075
CDNet-resnet101	4	0.921	0.803	0.079	60	0.876	0.783	0.108
FICI-hrnet18s	0	0.958	0.883	0.044	51	0.907	0.830	0.069
FocalClick-segformerB0	2	0.957	0.798	**0.017**	54	0.903	0.724	**0.024**
FocalClick-segformerB3	**0**	0.962	0.896	0.037	**50**	0.912	0.841	0.048
Ours-segformerB0	**0**	0.961	0.810	0.027	51	0.912	0.748	0.040
Ours-segformerB3	**0**	**0.963**	**0.897**	0.054	**50**	**0.916**	**0.848**	0.067

Table 2: Performance of IIS methods evaluated with segmentation quality and efficiency metrics on Berkeley and DAVIS datasets. '↓' ('↑') means lower (higher) metric values indicate better performance.

- **GrabCut(Rother, Kolmogorov, and Blake 2004)** contains 50 images with single object masks.
- **Berkeley (Martin et al. 2001)** contains 96 images with 100 object masks.
- **SBD (Hariharan et al. 2011)** is comprised of 8,498 training images with 20,172 polygonal masks, and 2,857 validating images with 6,671 instance-level masks. Only the validating images are used for evaluation.
- **DAVIS (Perazzi et al. 2016)** contains 345 frames randomly sampled from 50 videos. Each frame is provided with high-quality masks.

Experimental Setting. We choose segformerB0 or segformerB3 (Xie et al. 2021) as the backbone of the segmentation model. During the training phase, in accordance with the approach described in (Chen et al. 2022), 30,000 images are randomly selected as the training dataset for each epoch. Initially, the images are downsampled using either bilinear interpolation or TAIZ. Data augmentation is subsequently applied, encompassing random flipping, resizing with a scale factor constrained within the interval $[0.75, 1.40]$, and randomized adjustments to brightness, contrast, and RGB coloration. σ is set to 11. The network parameters are optimized through the Adam algorithm, parameterized with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The model undergoes a training process of 230 epochs, with an initial learning rate of 5×10^{-4}. This learning rate is subsequently attenuated by a factor of 0.1 at the 190th and 220th epochs. Training is executed with a batch size of 24, using PyTorch as the implementation framework. All computational experiments are performed on a system equipped with two NVIDIA GeForce RTX 3090 GPUs, and the training duration for the proposed method is approximately 48 hours.

Evaluation Metrics. In assessing IIS methods, we adhere to the evaluation mechanism delineated in (Sofiiuk, Petrov, and Konushin 2022a; Chen et al. 2022). The maximum click number T is set to 20. The performance is quantitatively measured through five specific metrics:

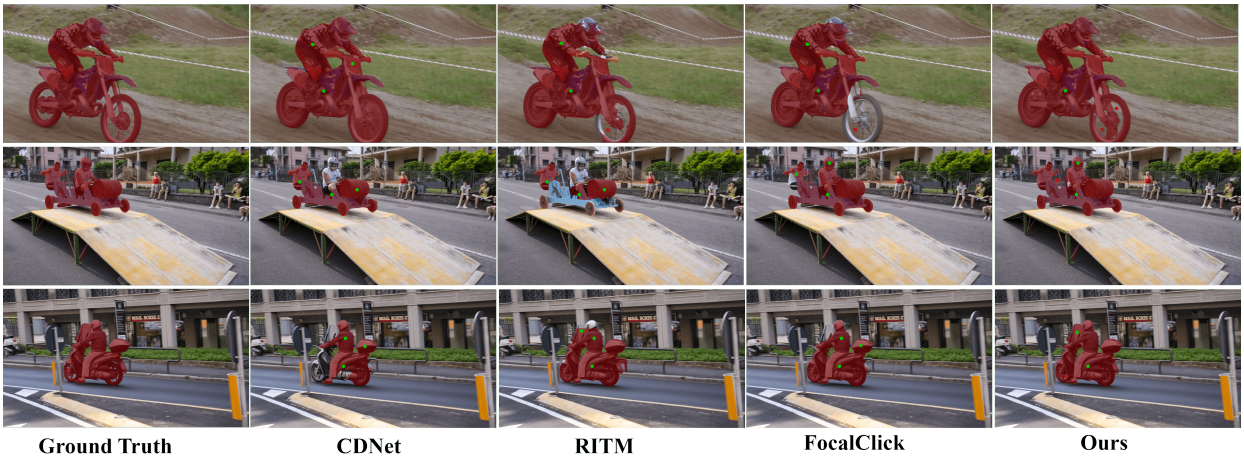

Figure 5: Qualitative comparisons of CDNet, RITM, FocalClick, and our method.

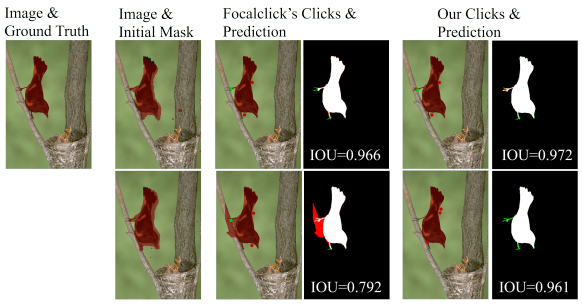

Figure 6: Comparison of the results obtained from the FocalClick and our method with different initial masks. In the 4th and 6th columns, white, red, and green indicate true positives, false positives, and false negatives, respectively.

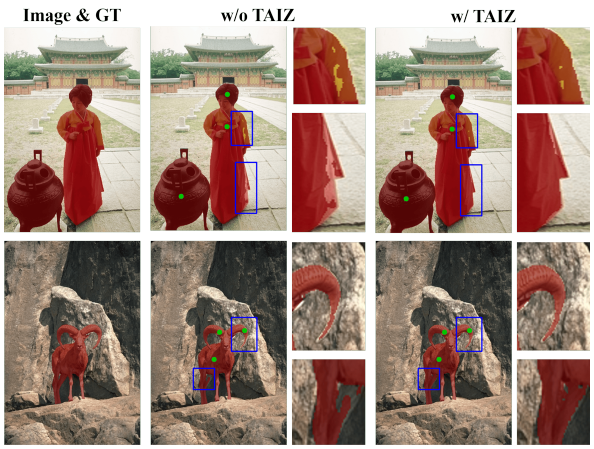

Figure 7: Qualitative comparison between method variations using or not using TAIZ algorithm.

1. NoC@IoU: Reflects the average number of clicks necessary to attain the specified IoU threshold.
2. NoF@IoU: Quantifies the number of instances where the model fails to reach the prescribed IoU threshold within the maximum allowable number of clicks.
3. IoU@N: Denotes the mean IoU achieved for testing images after N clicks.
4. BIoU@N: Signifies the mean boundary IoU of the testing images after N clicks.
5. SPC: Represents the mean computational time required for inference following each click.

These metrics collectively provide a comprehensive and robust evaluation of the model's efficacy and efficiency.

Comparison with Other Methods

In our comparative analysis presented in Table 1, our method is benchmarked against existing techniques, including RIS-Net (Liew et al. 2017), LD (Li, Chen, and Koltun 2018), CAG (Majumder and Yao 2019), BRS (Jang and Kim 2019), FCA (Lin et al. 2020), f-BRS (Sofiiuk et al. 2020), CD-Net (Chen et al. 2021), RITM (Sofiiuk, Petrov, and Konushin 2022a), EdgeFlow (Hao et al. 2021b), FICI (Wei, Zhang, and Yong 2023), and FocalClick, using the NoC metrics. Owing to the incorporation of the mask matching regularization and TAIZ operation, our proposed algorithm consistently necessitates fewer clicks to achieve IoU thresholds of 85% and 90%. This establishes its superiority over contemporary state-of-the-art methods. Notably, the Berkeley and DAVIS datasets manifest significant enhancements when processed by our method. For instance, based on the segformerB3, we achieve reductions of 0.11 and 0.35 in NoC@85 for Berkeley and DAVIS, respectively, relative to the prevailing state-of-the-art methods. Correspondingly, NoC@90 witnesses declines of 0.2 and 0.08 on these datasets.

Complementary metrics, encompassing NoF@90, IoU@5, BIoU@5, and SPC on the Berkeley and DAVIS datasets, are cataloged in Table 2. Our approach consistently

Method	Berkeley		DAVIS	
	NoC@85	NoC@90	NoC@85	NoC@90
FocalClick	1.98	2.50	3.97	5.46
Ours	1.86	2.38	3.82	5.31

Table 3: Average performance on correcting initial masks.

Method	Berkeley		DAVIS	
	NoC@85	NoC@90	NoC@85	NoC@90
baseline	1.55	1.92	3.61	4.90
TAIZ	1.53	1.90	3.54	4.90
MaskMatch	**1.43**	1.84	3.48	4.89
Replace L_{mr} with L_{sup}	1.51	1.78	3.49	4.88
Ours	1.44	**1.72**	**3.26**	**4.82**

Table 4: Ablation study for core components of our method.

Threshold	Berkeley		DAVIS	
	NoC@85	NoC@90	NoC@85	NoC@90
0.7	1.74	2.02	3.58	5.00
0.8	**1.44**	**1.72**	**3.26**	**4.82**
0.9	1.55	1.90	3.49	4.97

Table 5: Performance of our method using different thresholds to activate the mask matching regularization.

Variants	Berkeley		DAVIS	
	NoC@85	NoC@90	NoC@85	NoC@90
Prev. Mask	**1.38**	1.75	3.32	4.88
Points	1.46	1.81	3.34	4.91
Prev. Mask & Points	1.44	**1.72**	**3.26**	**4.82**

Table 6: Averaged NoC obtained with different choices of guidance map.

achieves superior values in segmentation quality metrics like IoU@5 and BIoU@5, without imposing significant time overheads. Specifically, compared to the second-ranking FocalClick, our approach enhances the BIoU@5 metric by 0.007 on DAVIS when using segformerB3 as the backbone. The SPC for our approach based on segformerB3 is 0.054s for Berkeley and 0.067s for DAVIS, indicating a computation time that is congruent with real-time applications. Compared to FocalClick, our method maintains the same model complexity and brings minimal additional computational or memory overhead during inference. Fig. 5 offers a visual depiction of the results. Relative to competitive techniques, our method yields more refined segmentations with enhanced boundary fidelity.

Robustness against Fluctuation in Initial Mask. To assess the resilience of our approach to fluctuation in the initial mask, we introduce perturbations to the GT mask of each test image ten times, guided by a threshold of 0.8 IoU. The capability of FocalClick and our method to rectify errors on these perturbed masks is detailed in Table 3. A visual representation comparing the outcomes of FocalClick and our approach for two initial masks is depicted in Fig. 6. Both quantitative and qualitative evaluations corroborate that our technique consistently delivers more stable and refined segmentation results irrespective of mask initialization.

Ablation Study

We conduct ablation studies on the Berkeley (Martin et al. 2001) and DAVIS (Perazzi et al. 2016) datasets, employing segformerB3 as the backbone for the segmentation model.

Core Component Analysis. Table 4 delineates the outcomes from diverse configurations of our approach. The baseline configuration excludes both the MaskMatch and TAIZ modules. Introducing either the MaskMatch or TAIZ modules distinctly enhances the NoC metrics on both datasets. Integrating both modules accentuates this improvement. Specifically, the integration of MaskMatch and TAIZ reduces the NoC@85 metric by 0.07 and 0.13, respectively, on the DAVIS dataset. When employed both modules, the decrease in NoC@85 is 0.22 on DAVIS, in comparison to employing only MaskMatch with standard bilinear interpolation for image downsampling. This substantiates the supplementary benefit of TAIZ over standard bilinear interpolation. Fig. 7 illustrates the segmentation outcomes from two method variations: one incorporating TAIZ (denoted as w/ TAIZ) and the other excluding it (denoted as w/o TAIZ). Visual assessments indicate that TAIZ brings a more comprehensive target extraction.

Moreover, we try to substitute the matching regularization term in Eq. (1) with the supervised learning term from Eq. (2), denoted by "Replace L_{mr} with L_{sup}". This alteration underperforms compared to our finalized model, underscoring the efficacy of MaskMatch.

Choice of Threshold Value for Activating MaskMatch. Table 5 analyzes the threshold, α, for MaskMatch activation. We find 0.8 as the optimal value. Lower thresholds may produce masks far from the GT, while higher ones may limit training samples in the MaskMatch process.

Comparison of Different Guidance Strategies in TAIZ. Table 6 assesses three guidance map strategies for the TAIZ module: 1) using the click map, 2) using the prior mask, and 3) combining both. The third strategy proves most effective, capturing the target's comprehensive representation and local click details.

Conclusion

In this paper, we introduce a cutting-edge approach termed *Variance-insensitive and Target-preserving Mask Refinement* for the point-based interactive image segmentation task. Our methodology encompasses a mask matching regularization, fortifying consistency in predictions arising from diverse initial masks. Such regularization substantially alleviates the prediction sensitivity to initial mask fluctuations. To alleviate the dilution of target information during input image downsampling, we deploy a target-aware image zooming mechanism, complementing traditional interpolation techniques. Comprehensive evaluations on datasets—GrabCut, Berkeley, SBD, and DAVIS—confirm our model's superiority over existing methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NO. 62376206, NO. 62003256, NO. 62322608), in part by the Shenzhen Science and Technology Program (NO. JCYJ20220530141211024), and in part by the Open Project Program of the Key Laboratory of Artificial Intelligence for Perception and Understanding, Liaoning Province (AIPU, No. 20230003).

References

Bertalmio, M.; Sapiro, G.; Caselles, V.; and Ballester, C. 2000. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 417–424.

Bonaccorso, G. 2017. *Machine learning algorithms*. Packt Publishing Ltd.

Chen, X.; Zhao, Z.; Yu, F.; Zhang, Y.; and Duan, M. 2021. Conditional Diffusion for Interactive Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7345–7354.

Chen, X.; Zhao, Z.; Zhang, Y.; Duan, M.; Qi, D.; and Zhao, H. 2022. FocalClick: Towards Practical Interactive Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1300–1309.

Cheng, H. K.; Chung, J.; Tai, Y.-W.; and Tang, C.-K. 2020. CascadePSP: Toward Class-Agnostic and Very High-Resolution Segmentation via Global and Local Refinement. In *CVPR*.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hao, Y.; Liu, Y.; Wu, Z.; Han, L.; Chen, Y.; Chen, G.; Chu, L.; Tang, S.; Yu, Z.; Chen, Z.; and Lai, B. 2021a. EdgeFlow: Achieving Practical Interactive Segmentation with Edge-Guided Flow. arXiv:2109.09406.

Hao, Y.; Liu, Y.; Wu, Z.; Han, L.; Chen, Y.; Chen, G.; Chu, L.; Tang, S.; Yu, Z.; Chen, Z.; and Lai, B. 2021b. EdgeFlow: Achieving Practical Interactive Segmentation With Edge-Guided Flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 1551–1560.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, 991–998.

Jang, W.-D.; and Kim, C.-S. 2019. Interactive Image Segmentation via Backpropagating Refinement Scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Z.; Chen, Q.; and Koltun, V. 2018. Interactive Image Segmentation With Latent Diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liew, J.; Wei, Y.; Xiong, W.; Ong, S.-H.; and Feng, J. 2017. Regional Interactive Image Segmentation Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2746–2754.

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3159–3167.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Lin, Z.; Duan, Z.-P.; Zhang, Z.; Guo, C.-L.; and Cheng, M.-M. 2022. FocusCut: Diving Into a Focus View in Interactive Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2637–2646.

Lin, Z.; Zhang, Z.; Chen, L.-Z.; Cheng, M.-M.; and Lu, S.-P. 2020. Interactive Image Segmentation With First Click Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mahadevan, S.; Voigtlaender, P.; and Leibe, B. 2018. Iteratively Trained Interactive Segmentation. arXiv:1805.04398.

Majumder, S.; and Yao, A. 2019. Content-Aware Multi-Level Guidance for Interactive Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, 416–423 vol.2.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Gool, L. V.; Gross, M.; and Sorkine-Hornung, A. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rother, C.; Kolmogorov, V.; and Blake, A. 2004. "GrabCut": Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Trans. Graph.*, 23(3): 309–314.

Sajjadi, M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29.

Sofiiuk, K.; Barinova, O.; and Konushin, A. 2019. AdaptIS: Adaptive Instance Selection Network. arXiv:1909.07829.

Sofiiuk, K.; Petrov, I.; Barinova, O.; and Konushin, A. 2020. F-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sofiiuk, K.; Petrov, I. A.; and Konushin, A. 2022a. Reviving Iterative Training with Mask Guidance for Interactive

Segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, 3141–3145.

Sofiiuk, K.; Petrov, I. A.; and Konushin, A. 2022b. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, 3141–3145. IEEE.

Thavamani, C.; Li, M.; Ferroni, F.; and Ramanan, D. 2023. Learning to Zoom and Unzoom. arXiv:2303.15390.

Wei, Q.; Zhang, H.; and Yong, J.-H. 2023. Focused and Collaborative Feedback Integration for Interactive Image Segmentation. arXiv:2303.11880.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33: 6256–6268.

Xu, N.; Price, B.; Cohen, S.; Yang, J.; and Huang, T. 2016. Deep Interactive Object Selection. arXiv:1603.04042.

Xu, N.; Price, B.; Cohen, S.; Yang, J.; and Huang, T. 2017. Deep GrabCut for Object Selection. arXiv:1707.00243.

Yu, H.; Zhou, Y.; Qian, H.; Xian, M.; and Wang, S. 2017. Loosecut: Interactive image segmentation with loosely bounded boxes. In *2017 IEEE International Conference on Image Processing (ICIP)*, 3335–3339. IEEE.

Zhou, M.; Wang, H.; Zhao, Q.; Li, Y.; Huang, Y.; Meng, D.; and Zheng, Y. 2023. Interactive Segmentation as Gaussian Process Classification. arXiv:2302.14578.