

TETRIS: Towards Exploring the Robustness of Interactive Segmentation

Andrey Moskalenko, Vlad Shakhuro, Anna Vorontsova, Anton Konushin,
Anton Antonov, Alexander Krapukhin, Denis Shepelev, Konstantin Soshin

Samsung Research

{andrey.mos, v.shakhuro, a.vorontsova, a.konushin, ant.antonov, a.krapukhin, d.shepelev, k.soshin}@samsung.com

Abstract

Interactive segmentation methods rely on user inputs to iteratively update the selection mask. A click specifying the object of interest is arguably the most simple and intuitive interaction type, and thereby the most common choice for interactive segmentation. However, user clicking patterns in the interactive segmentation context remain unexplored. Accordingly, interactive segmentation evaluation strategies rely more on intuition and common sense rather than empirical studies (e.g., assuming that users tend to click in the center of the area with the largest error). In this work, we conduct a real user study to investigate real user clicking patterns. This study reveals that the intuitive assumption made in the common evaluation strategy may not hold. As a result, interactive segmentation models may show high scores in the standard benchmarks, but it does not imply that they would perform well in a real world scenario. To assess the applicability of interactive segmentation methods, we propose a novel evaluation strategy providing a more comprehensive analysis of a model’s performance. To this end, we propose a methodology for finding extreme user inputs by a direct optimization in a white-box adversarial attack on the interactive segmentation model. Based on the performance with such adversarial user inputs, we assess the robustness of interactive segmentation models w.r.t click positions. Besides, we introduce a novel benchmark for measuring the robustness of interactive segmentation, and report the results of an extensive evaluation of dozens of models.

Introduction

Interactive segmentation methods are widely exploited for object removal, object selection, large dataset collection, medical image annotation and other tasks related to image labeling. Compared to conventional segmentation approaches, interactive methods provide higher quality masks that satisfy user requests better. Arguably the most well-explored, click-based interactive segmentation aims at selecting objects in an image according to multiple user input clicks (either positive or negative), comprising a *click trajectory*. However, the real user evaluation of each novel approach, that implies comparing it with an ever-growing number of predecessors, is completely unfeasible. Respectively,

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

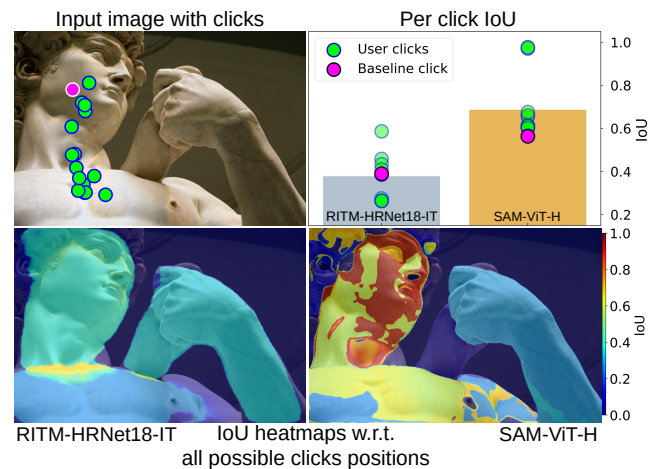


Figure 1: Single clicks made by different real users and the respective quality achieved. Top left: real users (green dots) do not click the way it is assumed in the standard testing procedures (magenta dot). Top right: the quality of two popular interactive segmentation models, a convolutional RITM (Sofiuk, Petrov, and Konushin 2022) and a transformer-based SAM (Kirillov et al. 2023), is widely spread around the average score (visualized with colored bars). Bottom: IoU heatmaps show that prediction quality fluctuates heavily depending on an actual click position.

in the common interactive segmentation benchmarks, clicks are not put by real users but automatically generated based on a history of interactions. Most existing methods 1) select a region with the largest error in the previous interaction round, and 2) click in the furthest point from the boundaries of this region. Hereinafter, we refer to this click generation scheme as the *baseline strategy*.

However, our study of real user clicks reveals this straightforward strategy does not emulate user behavior adequately. Besides, models tend to overfit to the baseline strategy, so that the accuracy might be high, but even a slight change of a click position causes a severe quality drop (Figure 1). Thus, the real-usage quality of the tested models remains unknown.

The evaluation protocols that assess quality for a single



Figure 2: Top row: images with overlapped ground-truth masks (white), real user clicks (green), and clicks generated with the baseline strategy (magenta). Two bottom rows: IoU scores of RITM and SAM, calculated on a grid for each possible integer click position; warmer colors correspond to higher scores. Apparently, IoU scores may vary dramatically within small regions of the same object: this shows that the state-of-the-art approaches are rather sensitive to the click position.

click trajectory cannot guarantee that tested methods are robust enough to perform well in various possible interaction scenarios. In this work, we formulate a multi-trajectory evaluation strategy. Particularly, we propose to generate click trajectories through a differentiable adversarial attack on the interactive segmentation model, and estimate the robustness based on a quality gap between trajectories.

Overall, our contributions are as follows:

- We conduct a pioneer study of real user clicking patterns in an interactive segmentation scenario. It reveals that users do not always click in the center of an area with the largest error, as assumed in the baseline methodology used in the most existing methods.
- To the best of our knowledge, we are the first to develop a procedure for generating user inputs via an adversarial attack for measuring robustness of interactive models. Relying on a differentiable rendering of user inputs, the proposed procedure remains fully differentiable and fast-convergent.
- We present a TETRIS benchmark with 2000 high-resolution images carefully selected and manually labeled with fine segmentation masks. The images depict common objects: 1000 images contain objects of various categories, and another 1000 portray people.
- We formulate an interactive segmentation robustness score, and evaluate the robustness of state-of-the-art methods, using TETRIS and the standard interactive segmentation benchmarks.

We believe that the methodology presented in this study will assist creating more robust and high-quality interactive models for real world applications.

Related Work

Benchmarking Interactive Segmentation. GrabCut (Rother, Kolmogorov, and Blake 2004) was the first interactive segmentation dataset. Then, the Berkeley (Martin et al. 2001) segmentation dataset was adapted for interactive

segmentation (McGuinness and O’connor 2010). The associated evaluation protocol implied assessing both object and boundary segmentation quality with IoU measure and required manual interaction with a method. (Xu et al. 2016) proposed an automatic procedure of benchmarking click-based interactive segmentation on PASCAL VOC 2012 (Everingham et al. 2012) and COCO (Lin et al. 2014) segmentation datasets; in this procedure, clicks were placed strictly in the center of the largest erroneous region, and the quality was assessed with IoU. The follow-up work (Li, Chen, and Koltun 2018) adapted DAVIS (Perazzi et al. 2016a) and SBD (Hariharan et al. 2011) (labeled with boundaries) datasets for interactive segmentation, using the same click generation strategy. We conduct our study on most commonly used datasets as well as TETRIS, which contains images of a significantly higher resolution.

Segmentation Metrics. The most common metric used to assess interactive segmentation is the Number of Clicks (NoC) (Jang and Kim 2019; Sofiiuk et al. 2020), required to achieve the predefined IoU score. NoC equally penalizes the cases where the desired score was achieved on the last interactions, and the cases where the threshold was not exceeded; we consider this to be a major drawback of this metric. Besides, it was noticed (Sofiiuk, Petrov, and Konushin 2022) that using the baseline strategy encourages the model to overfit to NoC, while the performance in a real scenario remains poor.

Accordingly, we do not use NoC, but consider an area under an IoU curve (Jang and Kim 2019) as a major metric. We consider 10 clicks and normalize the area to be within $[0, 1]$. The standard IoU score is edge-insensitive, so boundary metrics were additionally formulated for an ad-hoc assessment. The trimap IoU (Kohli, Ladický, and Torr 2009; Chen et al. 2018) is calculated within a distance from the ground truth mask boundary, ignoring distant erroneous pixels. The performance issue was addressed with approximations of F-measure (Csurka, Larlus, and Perronnin 2013; Perazzi et al. 2016b). McGuinness et al. (McGuinness and

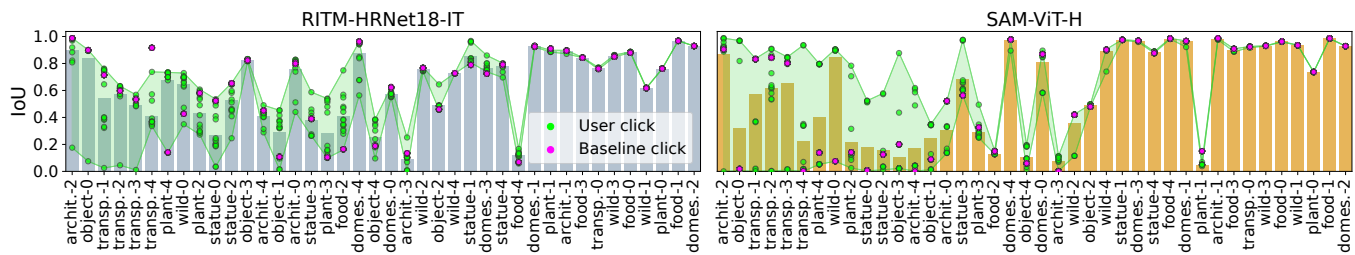


Figure 3: An IoU spread (a difference between a maximum and minimum IoU of user clicks) between predicted and ground truth masks in the first real user interaction round. Green points represent user clicks, magenta points depict the clicks generated with the baseline strategy. Columns are sorted by an average spread.

O’connor 2010) formulated a fuzzy boundary accuracy measure. (Cheng et al. 2020) proposed a mean Boundary Accuracy measure (mBA), further reworked into an MQ (Yang et al. 2020) score. In an image matting, boundary quality is evaluated with a trimap-based SAD, MSE (Xu et al. 2017), and perceptual Gradient and Connectivity errors (Xu et al. 2017). Since pronounced boundaries are especially important for high-resolution image processing, we also measure the boundary quality. To this end, we use a recently introduced Boundary IoU (Cheng et al. 2021a), which is intuitive and one of the most straightforward.

User Inputs. In this study, we consider only click-based approaches. However, numerous works were dedicated to other user input types. Bounding boxes were employed for selecting large image areas (Xu et al. 2016; Rother, Kolmogorov, and Blake 2004). In (Gueziri, McGuffin, and Laporte 2017), object selection was guided with manual strokes. In (Agustsson, Uijlings, and Ferrari 2019), an initial selection was made using bounding boxes obtained via extreme clicking (Papadopoulos et al. 2017), and then refined with strokes. (Cheng et al. 2021b) proposed a randomized uniform click and stroke generation strategy, where points were randomly sampled from the ground truth mask. Recently presented Segment Anything, or SAM (Kirillov et al. 2023), formulated a *promptable segmentation* task, where each prompt can be a point, a box, a mask, or a text.

Adversarial Attacks. Adversarial attack approaches are typically classified as either *black-box* or *white-box*, depending on whether the information about an attacked model is available. Black-box approaches (Wieland Brendel and Bethge 2018; Bhagoji et al. 2018; Su, Vargas, and Sakurai 2019) may compensate a lack of information with an extensive computation. Since we consider high-resolution images and numerous clicks, the amount of computations required in a black-box adversarial attack is unfeasible. Accordingly, we are restricted with white-box approaches.

The robustness of the conventional segmentation approaches was already explored (Kamann and Rother 2020); yet, as user inputs are not involved, the robustness could only be measured w.r.t image perturbations. A recent series of works (Guan et al. 2023; Zhang et al. 2023b; Qiao et al. 2023; Wang, Zhao, and Petzold 2023) measuring the robustness of SAM also focused on perturbing images rather than user prompts. In contrast, we fix input images and investi-

gate the robustness w.r.t user inputs. We propose a fully differentiable white-box attack for generating adversarial user inputs, and formulate robustness metrics accordingly.

TETRIS

In this section, we briefly describe our self-collected dataset serving as a basis of the robustness benchmark. By creating TETRIS, we focused on usability, which implies proper licensing, no privacy violation (all depicted people gave an articulated consent), and avoiding other issues that may limit or forbid using a dataset.

Object Classes

Object classes for TETRIS are chosen according to the task-specific requirements. Specifically, we select object classes seeming useful for image editing and labeling. Since we are unaware of any previous research dedicated to image editing scenarios, we cannot rely on a real-life object class distribution. So, we include classes present in PASCAL VOC 2012 (Everingham et al. 2012) and some other common classes from COCO (Lin et al. 2014). Overall, we consider 9 meta-classes: *transport*, *wild animal*, *object*, *domestic animal*, *food*, *architecture*, *plant*, *statue* are represented in TETRIS-THINGS, while TETRIS-PEOPLE contains only images of *people*.

Image Acquisition and Annotation

For TETRIS-THINGS, we manually selected 1000 images from Unsplash¹. For TETRIS-PEOPLE, we purchased 1000 photos directly from a crowdsourcing vendor. Age, gender, country, and race according to (Karkkainen and Joo 2021) were indicated by participants themselves. They also gave a mandatory consent to using their personal data and images; each user could donate from 1 to 5 photos. We restricted the photo resolution with at least of 2MP to ensure good quality of images.

The acquired images were segmented into polygonal regions using the CVAT (Sekachev et al. 2020) labeling tool, and each such region was marked either as a foreground, a background, or an uncertain region. Next, we apply matting (Park et al. 2022) to uncertain regions, so that they turned into either foreground or background. For images

¹<https://unsplash.com/license>

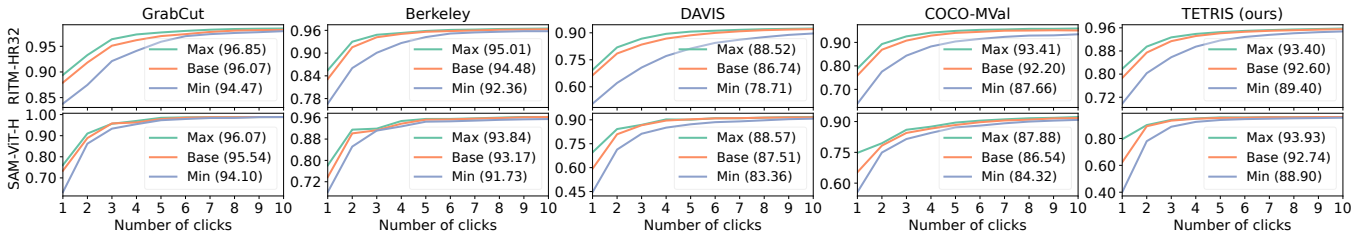


Figure 4: The minimizing, baseline, and maximizing trajectories of IoU for RITM and SAM models. Aggregated values (IoU-AuC) are given in brackets.

with several objects, we merged the corresponding alpha maps and thresholded them to obtain a binary segmentation mask. Finally, all masks were verified by expert annotators.

Exploring Robustness Issues

Below, we present the results of our real user study with two interaction rounds. Additionally, we analyze how the prediction quality varies in different possible click positions.

Real User Study

We selected five images per category from TETRIS-THINGS for a real user study. We used a crowdsourcing web annotation platform, and asked hundreds of users to label images with a simple annotation tool.

First interaction round. Each performer was exposed with 1) a source image, and 2) the same image, overlapped with a predicted semi-transparent mask (or a ground-truth mask in the first interaction round). In the first interaction round, we asked the annotators to put a single click on the target object. In the first round, only positive clicks are allowed. The total of 600 users participated, resulting in 15 interactions per each of 40 images.

Second interaction round. After completing the first round, the acquired user clicks were processed with RITM HRNet18 (Sofiiuk, Petrov, and Konushin 2022), and false positive and false negative per-pixel errors were calculated. For each type, we selected 40 samples with the largest error values. For clicks with the largest False Positive error, where the model predicted excessive masks, annotators were asked to make a negative click to exclude the redundant regions. Vice versa, for clicks with the largest False Negative error, users were requested to make the second positive click to cover missing areas. Other 1200 users were recruited in this round, providing 15 interactions per image.

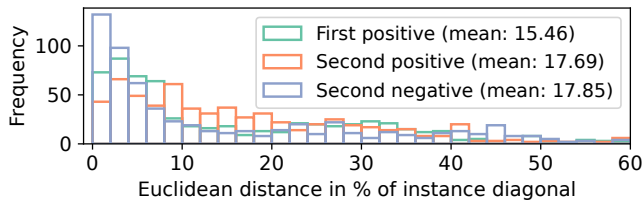


Figure 5: Distances between 1800 clicks made by real users and the ones generated using the baseline strategy.

Exhaustive Search

We also investigate the model quality in a full search over all integer input positions. Being arguably the simplest and the most intuitive way to measure the quality change w.r.t perturbed user inputs, this approach is resource-exhaustive: the number of forward passes grows linearly with both image height and width, making it impossible to evaluate on a full dataset in reasonable time.

We perform a brute-force on a pixel-wise grid and visualize the obtained results as heatmaps in Figure 2 (bottom rows). For each pixel, the color represents the IoU score obtained if clicking on this pixel; warmer hues mark higher IoU scores. Brute-force for a single 1024×1024 image takes over 4 hours using a single NVIDIA Tesla V100 to proceed, which encourages us to seek a faster approach for robustness evaluation, described below.

Analysis

We observe that the click position obtained with the baseline clicking strategy is consistent with the real user click only in case of convex objects of simple shapes (e.g., a pizza). For more complex geometries, users tend to click in different areas of density, or salience objects’ parts. Figure 5 shows distances between each user click and the click generated by the baseline strategy; all distances are normalized by an instance size (a diagonal length) for fair comparison between different objects. The error exceeds 15 percent on average, reaching a half of an instance diagonal size in some cases.

Nevertheless, user inputs are mostly gathered in a vicinity of the object’s “center” (being understood subjectively based on a common sense rather than formal criteria), which might not actually coincide with the point being the furthest from the boundaries. The divergence of real and generated user clicks is especially tangible in case of long, thin objects, i.e., a snake. Besides, we notice that the quality of the tested model may variate within a large range depending on a click position. As one can see in Figure 3, the click position significantly affects the quality in more than on a half test samples. It also shows that users might easily unintentionally place clicks in such adversarial points, providing an unexpectedly low segmentation quality.

Proposed Evaluation Protocol

Let us informally define a *robustness* of an interactive segmentation model as its ability to output the same mask for

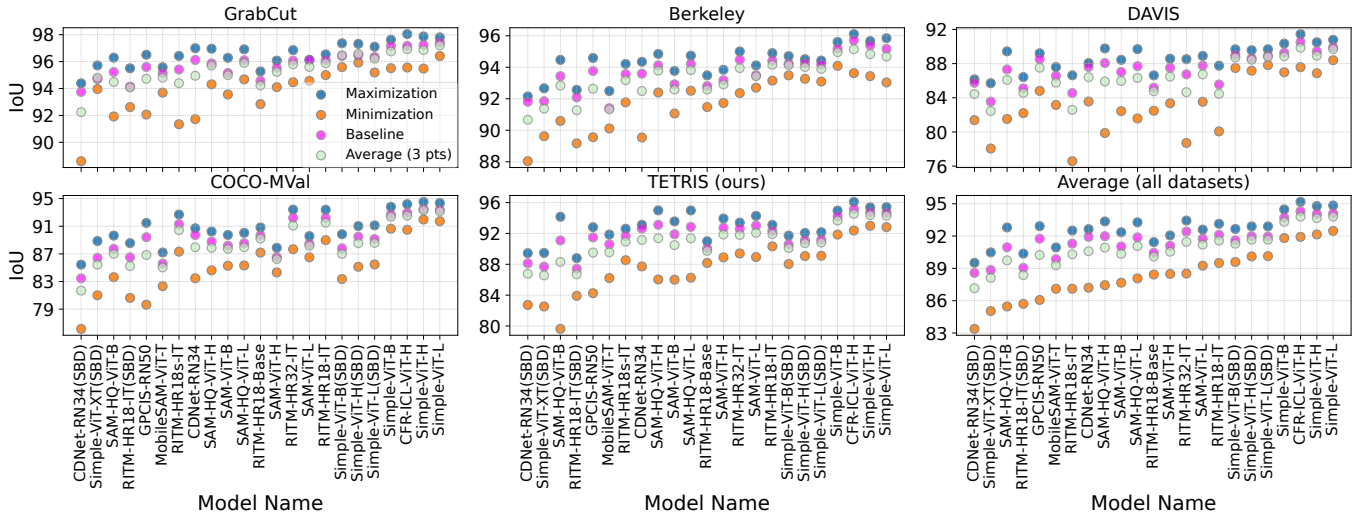


Figure 6: Visualization of IoU bounds of tested models. Columns are sorted by an average IoU-Min. As can be seen, there exists a dependency between the robustness (IoU-D, a delta between IoU-Max and IoU-Min) and the prediction quality.

any valid user input pointing to the same object. We consider only valid inputs: positive clicks must be placed within a yet unselected area of an object mask, while negative clicks should locate in some selected area outside the object mask.

Adversarial Inputs

To reduce the processing time from hours to seconds, compared to brute-force, we need to restrict the search space in a sensible way. We assume the center of the object to be a reasonable starting point, and then search for a local extreme in its vicinity.

We select click positions with a white-box targeted attack. The overall scheme of our method is shown in Figure 7. Using differentiable rendering (Ma et al. 2022), we encode clicks as maps with disks of a fixed radius marking click po-

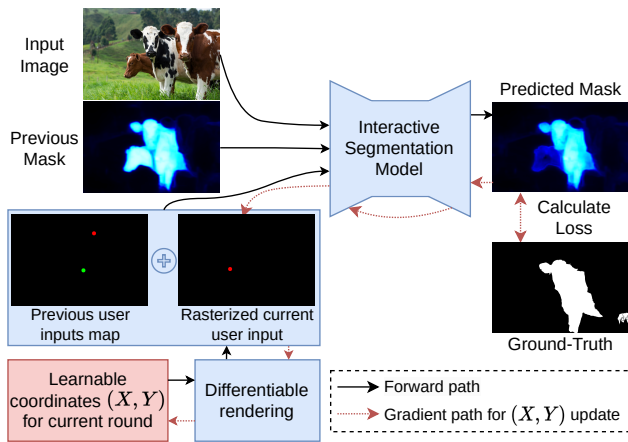


Figure 7: Overview of the proposed adversarial inputs generation. For models with raw (X, Y) coordinates, such as (Kirillov et al. 2023), we use a differentiable rendering step only for interaction location loss term.

sitions (since most models accept user inputs in this form). The radius is a hyperparameter, depending on the architecture of an attacked model. We calculate loss between the predicted and ground truth mask and run a gradient update to optimize click positions according to the chosen strategy. We use two strategies: one aims to minimize IoU, another targets at maximizing it. Surprisingly, even finding local extrema using the gradient descent method, a change of a click position has a great impact on the final quality (Figure 4).

Proposed Metrics

Sequentially optimizing each interaction, we obtain two click trajectories, referred to as the *minimizing trajectory* and *maximizing trajectory*. Similarly, we address the click trajectory obtained via the baseline clicking strategy, as the *baseline trajectory*. Based on the obtained trajectories, we propose a robustness metric specifically for the task of interactive segmentation (an intuitive explanation is given in Figure 8).

IoU/BIoU-Min/Max — the area under the minimizing/maximizing trajectory curve, the quality metric on a generated trajectory of the worst/best adversarial clicks.

IoU/BIoU-D — the difference between the area under

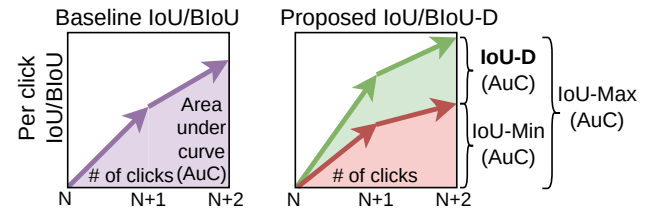


Figure 8: Left: the standard IoU/BIoU-AuC score. Right: the proposed IoU/BIoU-D robustness score. We consider 10 clicks and normalize the area to be within $[0, 1]$.

Method	Model	Data	GrabCut			Berkeley			DAVIS			COCO-MVal			TETRIS (ours)		
			IoU (AuC@10)	Min↑	Max↑	D↓	IoU (AuC@10)	Min↑	Max↑	D↓	IoU (AuC@10)	Min↑	Max↑	D↓	IoU (AuC@10)	Min↑	Max↑
MobileSAM	ViT-Tiny	SA-1B	93.69	95.58	1.89	90.11	92.49	2.38	83.16	87.60	4.44	82.32	87.20	4.88	86.22	91.85	5.64
SAM	ViT-B	SA-1B	93.56	96.28	2.72	91.06	93.77	2.71	82.44	88.44	6.00	85.28	89.73	4.45	86.01	93.56	7.55
	ViT-L		94.57	96.13	1.57	92.70	94.12	1.41	83.53	88.90	5.37	86.51	89.57	3.06	88.94	94.28	5.33
	ViT-H		94.10	96.07	1.97	91.73	93.84	2.11	83.36	88.57	5.21	84.32	87.88	3.56	88.90	93.93	5.04
SAM-HQ	ViT-B	SA-1B +44K	91.92	96.29	4.37	90.59	94.47	3.88	81.52	89.44	7.92	83.62	89.64	6.02	79.63	94.16	14.53
	ViT-L		94.68	96.91	2.23	92.51	94.74	2.24	81.57	89.69	8.12	85.32	90.04	4.72	86.26	94.98	8.72
	ViT-H		94.31	96.96	2.65	92.40	94.84	2.45	79.87	89.77	9.90	84.62	90.24	5.62	86.03	94.98	8.96
CDNet	RN34	C+L	91.72	96.99	5.27	89.54	94.35	4.81	83.56	88.06	4.50	83.46	90.70	7.24	87.72	93.12	5.39
	RN34	SBD	88.60	94.39	5.79	88.04	92.15	4.11	81.39	86.15	4.77	76.14	85.44	9.30	82.73	89.44	6.71
GPCIS	RN50	C+L	92.06	96.51	4.45	89.56	94.59	5.03	84.81	89.21	4.41	79.65	91.47	11.83	84.26	92.81	8.56
RITM	HR18s-IT	C+L	91.34	96.42	5.08	91.77	94.21	2.44	76.58	86.63	10.06	87.32	92.69	5.37	88.53	92.60	4.08
	HR18		92.84	95.27	2.43	91.47	93.49	2.01	82.48	86.62	4.14	87.18	90.82	3.64	88.17	90.99	2.82
	HR18-IT		95.00	96.54	1.54	93.15	94.90	1.75	80.06	87.76	7.69	88.98	93.39	4.40	90.32	93.11	2.79
	HR32-IT		94.47	96.85	2.39	92.36	95.01	2.65	78.71	88.52	9.81	87.66	93.41	5.74	89.40	93.40	4.00
	HR18-IT	SBD	92.62	95.50	2.88	89.17	92.57	3.40	82.20	86.43	4.22	80.62	88.56	7.93	83.91	88.80	4.89
SimpleClick	ViT-B	C+L	95.51	97.63	2.12	94.09	95.61	1.51	86.98	90.34	3.36	90.64	93.80	3.17	91.85	94.97	3.12
	ViT-L		96.41	97.80	1.39	93.04	95.85	2.81	88.40	90.79	2.39	91.73	94.37	2.64	<u>92.81</u>	95.42	<u>2.60</u>
	ViT-H		95.48	97.87	2.39	93.44	95.66	2.23	86.88	90.50	3.62	91.97	94.53	2.56	92.99	95.38	2.39
	ViT-XT	SBD	93.96	95.72	1.76	89.62	92.66	3.05	78.06	85.71	7.65	81.01	88.86	7.85	82.53	89.47	6.94
	ViT-B		95.59	97.36	1.77	93.49	94.71	1.22	87.49	89.67	<u>2.19</u>	83.36	89.84	6.48	88.04	91.72	3.68
	ViT-L		95.19	97.10	1.91	93.09	94.42	1.33	<u>87.82</u>	89.68	1.86	85.47	91.12	5.65	89.10	92.16	3.06
	ViT-H		<u>95.92</u>	97.32	<u>1.39</u>	93.27	94.52	<u>1.25</u>	87.17	89.57	2.40	85.13	91.02	5.89	89.07	92.06	2.99
CFR-ICL	ViT-H	C+L	95.56	98.04	2.48	<u>93.63</u>	96.11	2.48	87.58	91.45	3.87	90.49	94.19	3.71	92.38	96.09	3.72

Table 1: The quality and robustness scores of different models, measured on the standard datasets and our novel TETRIS dataset. The best results are bold, the second best are underlined. SimpleClick and CFR-ICL are more robust than other tested approaches. Still, even state-of-the-art models are extremely sensitive to the positions of user clicks, which may cause an unstable performance in a real-world scenario.

curves of maximizing and minimizing trajectories. As depicted in Figure 4, maximizing, minimizing, and baseline trajectories converge with an increasing number of clicks, and the accuracy gap decreases accordingly. Therefore, the difference between trajectories is the most divisible and hence informative for a few clicks; accordingly, we consider only 10 clicks in all our experiments.

Evaluation Setup

We narrow down our evaluation with recent methods having an open-source codebase: RITM (Sofiiuk, Petrov, and Konushin 2022), CDNet (Chen et al. 2021), SimpleClick (Liu et al. 2022), CFR-ICL (Sun et al. 2023), and GPCIS (Zhou et al. 2023). Besides, we experiment with promptable interactive segmentation methods from the SAM family: the original SAM (Kirillov et al. 2023), SAM-HQ (Ke et al. 2023), MobileSAM (Zhang et al. 2023a). Overall, we validate 23 checkpoints on the 5 interactive segmentation datasets: GrabCut, Berkeley, DAVIS, and COCO-MVal, as well as on our novel TETRIS dataset. The obtained quality scores are listed in Table 1 and presented in a visual form with the baseline strategy in Figure 6.

We run no more than 10 optimization iterations to restrict the number of calculations. The gradient updates are calculated with an Adam optimizer (Kingma and Ba 2014). To compare models with different input resolution fairly, we linearly scale the learning rate by an input size factor:

$LR = \frac{5\sqrt{H^2+W^2}}{400\sqrt{2}}$, where H, W denote image height and width in pixels, respectively. For minimizing and maximizing trajectories, the first iteration is selected with the baseline strategy and the consecutive clicks are placed greedily one by one.

Without any constraints, the maximization strategy yields points in between the object parts but outside the object mask. While providing the best quality, such click positions are unlikely to be made by a real user. Also, the minimizing optimization can easily converge outside an object of interest. Thus, to generate valid clicks in an adversarial optimization, we impose an additional constraint. We calculate a distance transform map for false positive and false negative areas. For a positive click, we sum up the distances in the false negative positions covered by a circle representing this click, for a negative click – in the false positive positions, respectively. This gives an *interaction location loss*. The total loss is a weighted sum of a Dice (Dice 1945) loss and an interaction location loss, scaled by 1000. During the optimization, we use the following update scheme:

1. Initialize an optimizable click position according to the baseline strategy;
2. Start the optimization by minimizing / maximizing a loss function;
3. Accept the new generated click, if IoU decreases / increases and the *interaction location loss* does not in-

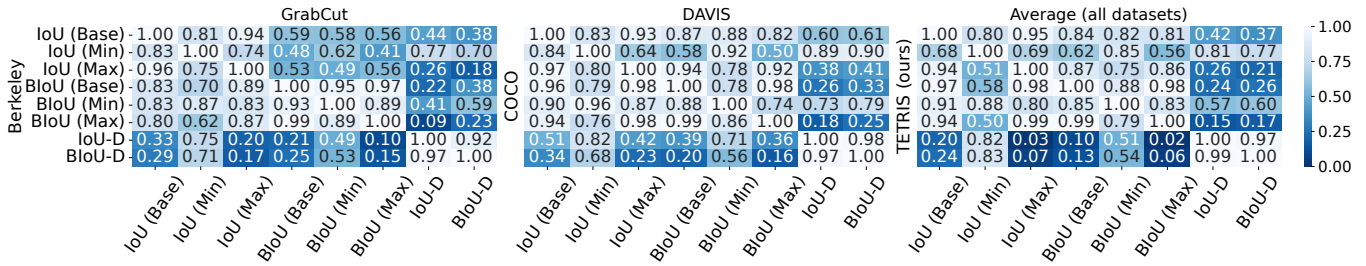


Figure 9: Cross-metric Spearman’s rank correlations measured for different datasets.

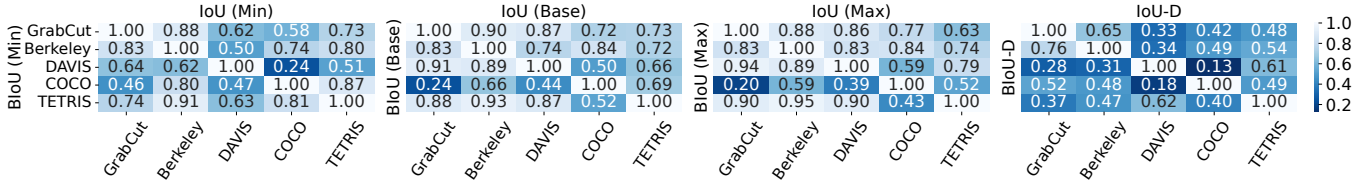


Figure 10: Cross-dataset Spearman’s rank correlations of IoU and BioU scores.

crease by more than 5%: (for objects thinner than the click radius, the loss inevitably gets worse even with a precise click, so we allow a small margin);

4. Save the predicted mask and click location from the best iteration, and use them in the next interaction round.

For models that accept raw click coordinates, we directly optimize the coordinates and use the differentiable rendering step only to compute *interaction location loss*. We follow the same evaluation procedure as is used in the original methods, applying ZoomIn (Sofiiuk et al. 2020), Cascade-Refinement (Sun et al. 2023), test-time augmentation flips (Sofiiuk, Petrov, and Konushin 2022), selecting a mask by a predicted score (Kirillov et al. 2023), etc.

Discussion

Based on the results of our user study (Figures 2, and 3) and robustness evaluation (Table 1; Figures 4, and 6), we can conclude that **state-of-the-art interactive segmentation models are extremely sensitive to the positions of user clicks**.

An exhaustive search on a pixel grid reveals that clicking on some coordinates within an object may unexpectedly cause a significant accuracy drop. We further show that for few points obtained through an adversarial attack, the quality may fluctuate significantly even within a small homogeneous area. We attribute such undesired behavior to the model selecting a part of an object (like a single slice of pepperoni) rather than the entire object (like a pepperoni pizza). Since the formulation of the interactive segmentation task is naturally fuzzy, such an ambiguity occurring is inevitable to a certain extent. Besides, the complexity of an object’s shape affects the model’s performance greatly. However, when developing an interactive segmentation model, one should aim to minimize those effects.

The minimizing, maximizing, and baseline trajectories do converge with an increasing number of clicks. The difference in quality is the largest during the first few interactions.

This actually means that **clicking in any sensible way (i.e. committing only valid clicks) will provide an acceptable result – but, possibly, after many interactions**.

Furthermore, we explore pairwise correlations between quality and robustness metrics (Figure 9) and compare model rankings on different datasets (Figure 10). It can be seen that:

- IoU/BioU-Base strongly correlates with IoU/BioU-Max; therefore, **the baseline evaluation protocol implicitly ranks models by the best possible quality, but does not reflect their performance in the worst case** (Figure 9);
- IoU/BioU-D strongly correlates with IoU/BioU-Min, while the correlations with IoU/BioU-Max is weaker (Figure 9). We attribute this to the fact that most of the quality spread is associated with a performance drop of the minimizing trajectory. According to Figures 4, and 6, optimizing adversarial inputs for the maximizing trajectory is much more difficult than searching for the worst clicks;
- The model ranking turns out to be dataset-specific. The ranking on TETRIS differs from the ranking on low-resolution datasets (Figure 10).

Conclusion

In this study, we showed that the prediction quality of click-based interactive segmentation models depends heavily on the click location. To this end, we conducted a real user study and analyzed 1800 participant responses. Guided by this empirical evidence, we proposed the adversarial input generation strategy, and formulated the robustness score, which is estimated based on multiple generated trajectories. We evaluated the robustness of dozens of open-sourced models on the well-known datasets; and also on the proposed TETRIS benchmark with 2000 high-resolution images manually labeled with fine segmentation masks.

References

- Agustsson, E.; Uijlings, J. R. R.; and Ferrari, V. 2019. Interactive Full Image Segmentation by Considering All Regions Jointly. In *CVPR*.
- Bhagoji, A. N.; He, W.; Li, B.; and Song, D. 2018. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proceedings of the European conference on computer vision (ECCV)*, 154–169.
- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *TPAMI*.
- 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*, 7345–7354.
- Cheng, B.; Girshick, R.; Dollár, P.; Berg, A. C.; and Kirillov, A. 2021a. Boundary IoU: Improving object-centric image segmentation evaluation. In *CVPR*.
- Cheng, H.; Xu, S.; Jiang, X.; and Wang, R. 2021b. Deep Image Matting with Flexible Guidance Input. In *BMVC*.
- 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*.
- Csurka, G.; Larlus, D.; and Perronnin, F. 2013. What is a good evaluation measure for semantic segmentation? In *BMVC*.
- Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302.
- Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.
- Guan, Z.; Hu, M.; Zhou, Z.; Zhang, J.; Li, S.; and Liu, N. 2023. Badsam: Exploring security vulnerabilities of sam via backdoor attacks. *arXiv preprint arXiv:2305.03289*.
- Gueziri, H.-E.; McGuffin, M.; and Laporte, C. 2017. Latency Management in Scribble-Based Interactive Segmentation of Medical Images. *IEEE Transactions on Biomedical Engineering*.
- Hariharan, B.; Arbelaez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic Contours from Inverse Detectors. In *ICCV*.
- 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*, 5297–5306.
- Kamann, C.; and Rother, C. 2020. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8828–8838.
- Karkkainen, K.; and Joo, J. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1548–1558.
- Ke, L.; Ye, M.; Danelljan, M.; Liu, Y.; Tai, Y.-W.; Tang, C.-K.; and Yu, F. 2023. Segment Anything in High Quality. *arXiv preprint arXiv:2306.01567*.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.
- Kohli, P.; Ladický, L.; and Torr, P. 2009. Robust Higher Order Potentials for Enforcing Label Consistency. *IJCV*.
- Li, Z.; Chen, Q.; and Koltun, V. 2018. Interactive image segmentation with latent diversity. In *CVPR*.
- 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 *ECCV*.
- Liu, Q.; Xu, Z.; Bertasius, G.; and Niethammer, M. 2022. SimpleClick: Interactive image segmentation with simple vision transformers. *arXiv preprint arXiv:2210.11006*.
- Ma, X.; Zhou, Y.; Xu, X.; Sun, B.; Filev, V.; Orlov, N.; Fu, Y.; and Shi, H. 2022. Towards Layer-wise Image Vectorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- 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 *ICCV*.
- McGuinness, K.; and O’connor, N. E. 2010. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2): 434–444.
- Papadopoulos, D. P.; Uijlings, J. R. R.; Keller, F.; and Ferrari, V. 2017. Extreme Clicking for Efficient Object Annotation. *ICCV*.
- Park, G.; Son, S.; Yoo, J.; Kim, S.; and Kwak, N. 2022. Matteformer: Transformer-based image matting via priortokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11696–11706.
- Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016a. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*.
- Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016b. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *CVPR*.
- Qiao, Y.; Zhang, C.; Kang, T.; Kim, D.; Tariq, S.; Zhang, C.; and Hong, C. S. 2023. Robustness of sam: Segment anything under corruptions and beyond. *arXiv preprint arXiv:2306.07713*.
- Rother, C.; Kolmogorov, V.; and Blake, A. 2004. GrabCut – Interactive Foreground Extraction using Iterated Graph Cuts. *ACM transactions on graphics (TOG)*, 23(3): 309–314.
- Sekachev, B.; Manovich, N.; Zhiltsov, M.; Zhavoronkov, A.; Kalinin, D.; Hoff, B.; TOSmanov; Kruchinin, D.; Zankevich,

A.; DmitriySidnev; Markelov, M.; Johannes222; Chenuet, M.; a andre; telenachos; Melnikov, A.; Kim, J.; Ilouz, L.; Glazov, N.; Priya4607; Tehrani, R.; Jeong, S.; Skubriev, V.; Yonekura, S.; vugia truong; zliang7; lizhming; and Truong, T. 2020. *opencv/cvat: v1.1.0*.

Sofiiuk, K.; Petrov, I.; Barinova, O.; and Konushin, A. 2020. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *CVPR*.

Sofiiuk, K.; Petrov, I. A.; and Konushin, A. 2022. Reviving Iterative Training with Mask Guidance for Interactive Segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, 3141–3145.

Su, J.; Vargas, D. V.; and Sakurai, K. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5): 828–841.

Sun, S.; Xian, M.; Xu, F.; Yao, T.; and Capriotti, L. 2023. CFR-ICL: Cascade-Forward Refinement with Iterative Click Loss for Interactive Image Segmentation. *arXiv preprint arXiv:2303.05620*.

Wang, Y.; Zhao, Y.; and Petzold, L. 2023. An empirical study on the robustness of the segment anything model (sam). *arXiv preprint arXiv:2305.06422*.

Wieland Brendel, J. R.; and Bethge, M. 2018. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In *International Conference on Learning Representations*.

Xu, N.; Price, B.; Cohen, S.; and Huang, T. 2017. Deep Image Matting. In *CVPR*.

Xu, N.; Price, B.; Cohen, S.; Yang, J.; and Huang, T. S. 2016. Deep interactive object selection. In *CVPR*.

Yang, C.; Wang, Y.; Zhang, J.; Zhang, H.; Lin, Z.; and Yuille, A. 2020. Meticulous Object Segmentation. *arXiv preprint arXiv:2012.07181*.

Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S.-H.; Lee, S.; and Hong, C. S. 2023a. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *arXiv preprint arXiv:2306.14289*.

Zhang, C.; Zhang, C.; Kang, T.; Kim, D.; Bae, S.-H.; and Kweon, I. S. 2023b. Attack-sam: Towards evaluating adversarial robustness of segment anything model. *arXiv preprint arXiv:2305.00866*.

Zhou, M.; Wang, H.; Zhao, Q.; Li, Y.; Huang, Y.; Meng, D.; and Zheng, Y. 2023. Interactive Segmentation As Gaussian Process Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19488–19497.