NEURAL INFORMATION PROCESSING SYSTEMS

RClicks: Realistic Click Simulation for Benchmarking Interactive Segmentation

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

AIRI

LOMONOSOV MOSCOW STATE UNIVERSITY

https://emb-ai.github.io/rclicks-project/

Interactive segmentation (IS)

Goal: to obtain high-quality pixel-level masks with limited user interaction

*Among various types of user input, **clicks** are the most common*

Interactive segmentation example using Segment Anything Model (SAM)

IS evaluation

1. Evaluation requires user inputs, gathering real-user data is impractical

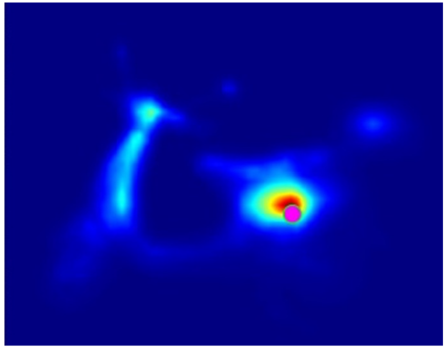
2. IS quality is assessed with a **baseline** clicking strategy: clicks are put in the center of the largest erroneous area

3. IS method might be **overfitted** for **baseline** clicks

4. To evaluate IS methods in a realistic way we propose a user **clickability model**

Real-users clicks

clicks simulated by **clickability model** and **baseline** click

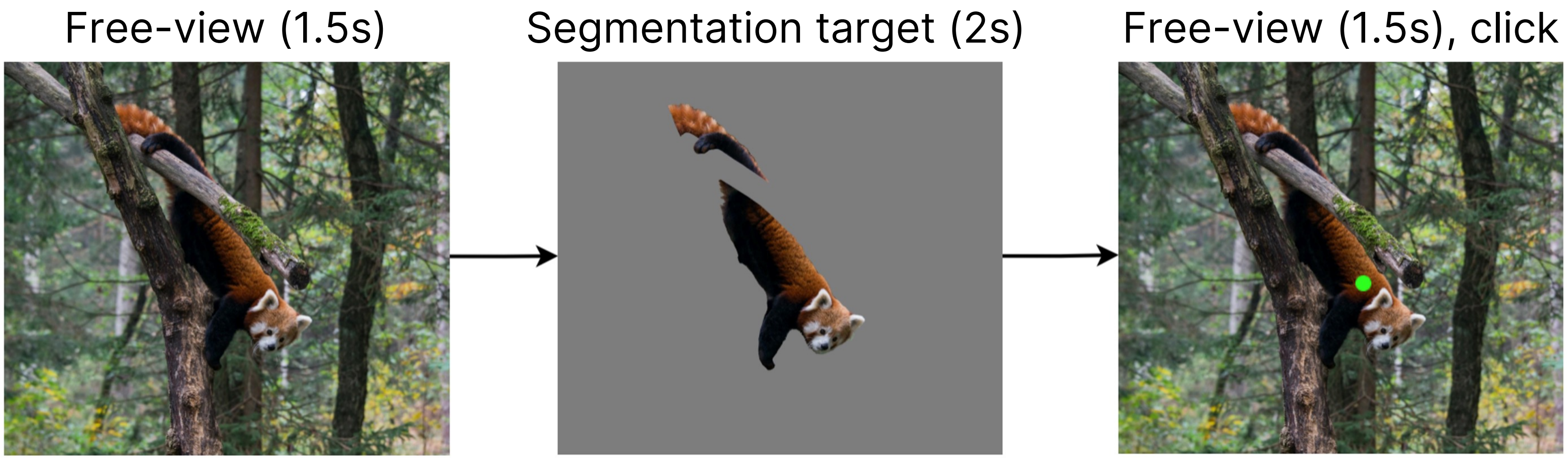


clicks distribution predicted by **clickability model**

Contributions

- Multi–round interaction dataset of **475 544 clicks**
- Novel **clickability model** for realistic click simulation
- RClicks — a **benchmark** for measurement of **real–world** annotation time and robustness of IS methods
- **Difficulty score** for IS instances on first real clicks

Data collection procedure



Data collection procedure

Users' click data

Our dataset is **based on** GrabCut, Berkeley, DAVIS, COCO-MVal, and TETRIS

To obtain error masks **for the subsequent rounds**, we applied SAM, SimpleClick, RITM, and synthetic distortions

We collected a dataset of clicks **both on PC** and **mobile devices**

Dataset	First #	Subseq. #	Sum #
GrabCut	2 395	3 427	5 822
Berkeley	4 859	6 937	11 796
DAVIS	16 975	23 687	40 662
COCO-MV	38 097	53 926	92 023
TETRIS	123 023	202 218	325 241
All	185 349	290 195	475 544

Number of collected clicks for each dataset in interaction rounds

Users' click data

Our dataset is **based on** GrabCut, Berkeley, DAVIS, COCO-MVal, and TETRIS

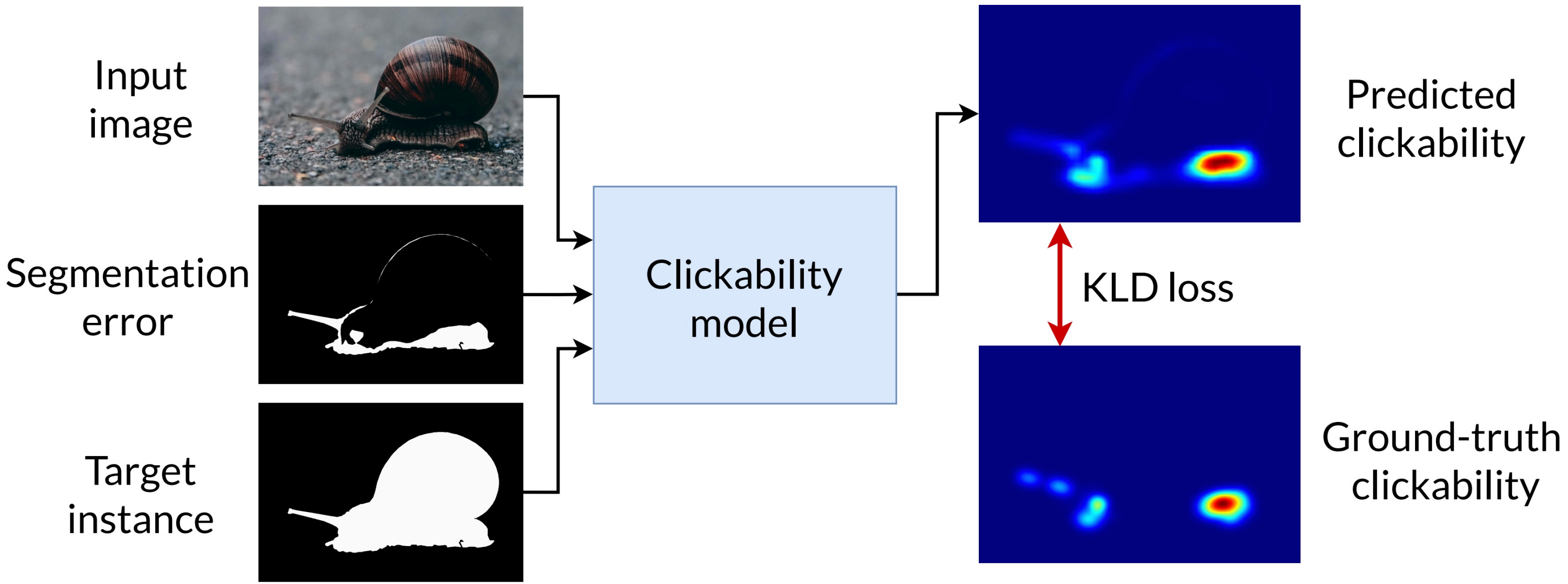
To obtain error masks **for the subsequent rounds**, we applied SAM, SimpleClick, RITM, and synthetic distortions

We collected a dataset of clicks **both on PC** and **mobile devices**

Dataset	First #	Subseq. #	Sum #
GrabCut	2 395	3 427	5 822
Berkeley	4 859	6 937	11 796
DAVIS	16 975	23 687	40 662
COCO-MV	38 097	53 926	92 023
TETRIS	123 023	202 218	325 241
All	185 349	290 195	$\boxed{475\,544}$

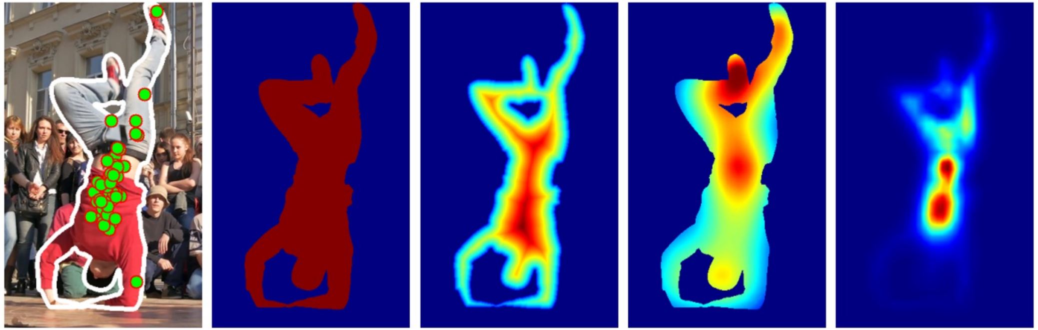
Number of collected clicks for each dataset in interaction rounds

Clickability model



Proposed clickability prediction pipeline

Clickability model



(a) Clicks (b) UD (c) DT (d) SM (e) Ours

Examples of **considered clickability models**:

(a) visualizes target object (white contour) and ground-truth clicks (green points)

(b) - (d) depict uniform distribution (UD), distance transform (DT), and saliency map (SM) respectively

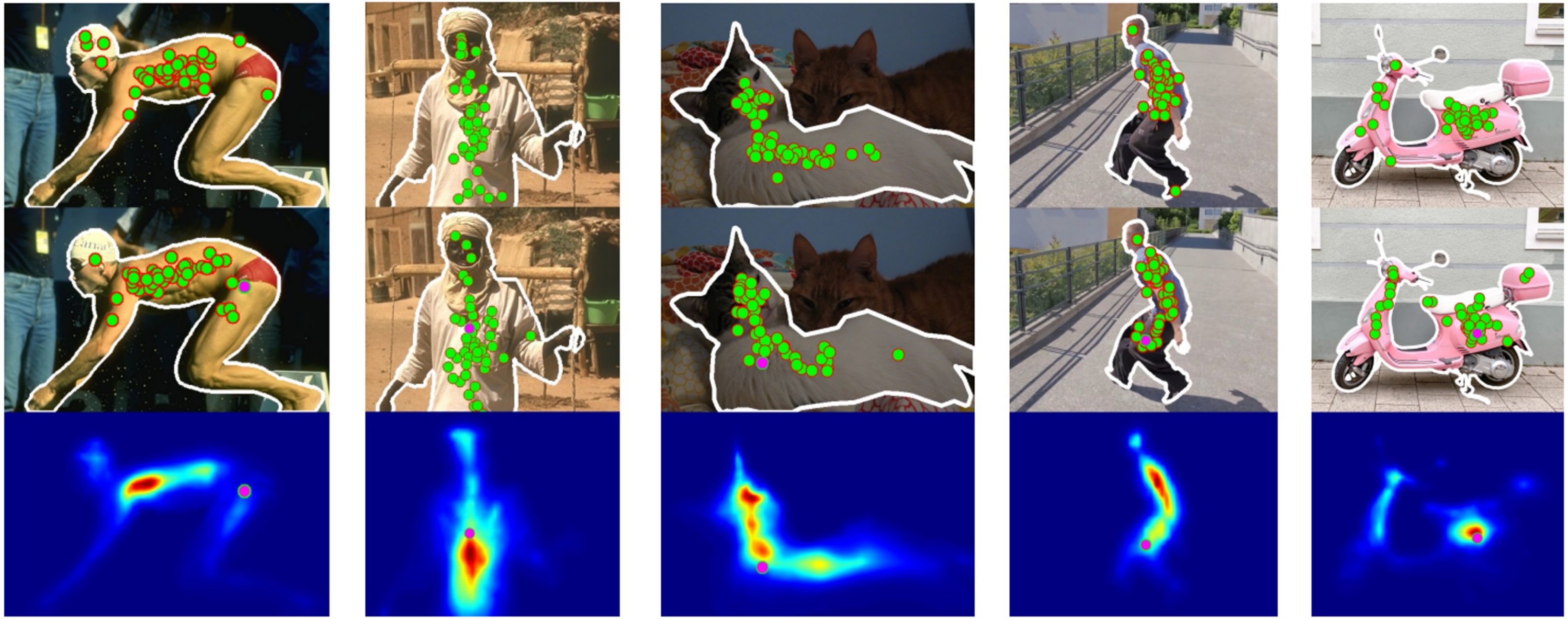
(e) - our predicted clickability map

Evaluation of various clickability models on real-user clicks of TETRIS validation part

Our approach **outperforms** existing clicking strategies in terms of the proximity of samples to real-user clicks

Model	KS $\uparrow$	$PL_1 \downarrow$	WD $\downarrow$	NSS $\uparrow$	PDE $\uparrow$
UD	0.10	0.57	0.17	3.99	1.36E-05
DT	0.14	0.52	0.16	6.45	2.76E-05
SM	0.13	0.51	0.15	4.79	1.83E-05
Ours	**0.55**	**0.40**	**0.08**	**9.11**	**4.69E-05**

Qualitative example



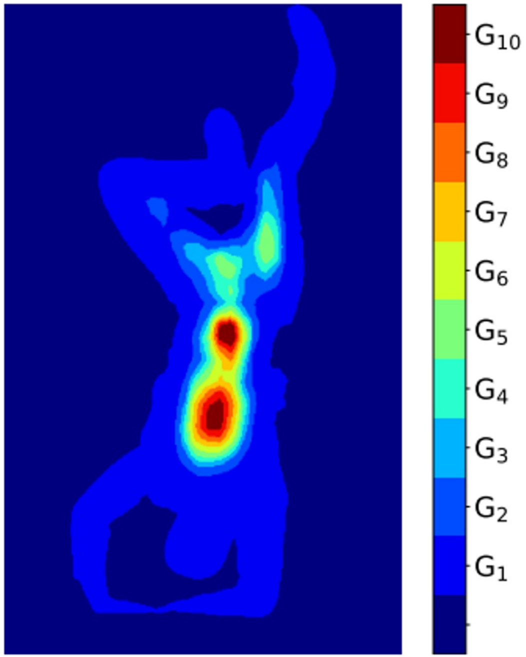
(a) GrabCut (b) Berkeley (c) COCO-MVal (d) DAVIS (e) TETRIS

Examples of real and predicted users' clicks and clickability map

RClicks benchmark

Using **Clicking Groups**, we propose the following objective IS robustness metrics:

- **Sample NoC** — mean and standard deviation of clicks (max 20) needed to achieve 90% IoU averaged across clicking groups (G_1 – G_{10})
- **ΔSB** — relative increase in Sample NoC compared to a baseline strategy
- **ΔGR** — relative increase in annotation time between G_1 and G_{10} clicking groups



Spatial distribution of **Clicking Groups** for the instance

Evaluation results

Method	Backbone	Data	DAVIS NoC$_{20}$@90 Sample (±std)	DAVIS ΔSB (+%)	DAVIS ΔGR (+%)	COCO–MVal NoC$_{20}$@90 Sample (±std)	COCO–MVal ΔSB (+%)	COCO–MVal ΔGR (+%)	TETRIS NoC$_{20}$@90 Sample (±std)	TETRIS ΔSB (+%)	TETRIS ΔGR (+%)
GPCIS	RN50	C+L	6.44±0.85	16.88	53.65	4.74±1.31	26.43	79.00	3.87±0.79	19.55	56.43
RITM	HR18	C+L	6.23±0.67	6.92	16.13	3.71±0.78	10.27	20.22	3.69±0.52	7.02	13.95
	HR18-IT	C+L	6.15±0.83	11.37	31.14	3.22±0.83	15.84	37.01	3.48±0.60	11.59	23.99
	HR32-IT	C+L	5.90±0.89	18.34	51.07	3.24±0.83	15.50	37.31	3.44±0.65	17.47	30.69
AdaptClick	ViT-B	C+L	4.97±0.40	8.60	15.14	2.93±0.58	9.44	19.75	2.62±0.37	6.99	12.94
SimpleClick	ViT-B	C+L	5.32±0.54	9.05	26.33	3.07±0.70	11.72	23.60	2.73±0.41	8.86	16.64
	ViT-L	C+L	5.03±0.42	8.71	16.67	2.67±0.56	8.05	20.88	*2.46±0.35*	7.11	10.01
	ViT-H	C+L	5.00±0.42	*7.06*	12.29	**2.57±0.54**	6.14	17.65	2.36±0.33	6.94	10.83
CFR-ICL	ViT-H	C+L	*4.53±0.46*	9.32	18.47	*2.70±0.63*	9.58	24.13	**2.12±0.34**	8.76	14.33
SAM	ViT-B	SA-1B	5.30±0.53	8.26	*11.27*	4.91±0.79	9.88	15.73	3.04±0.51	11.17	10.06
	ViT-L	SA-1B	5.21±0.41	8.82	11.59	4.81±0.63	8.89	14.97	2.60±0.40	8.11	7.08
	ViT-H	SA-1B	5.42±0.49	8.00	15.02	5.14±0.68	7.63	15.61	2.66±0.38	**5.95**	8.50
SAM-HQ	ViT-L	SA-1B	5.19±0.48	8.58	15.69	5.05±0.74	9.64	13.50	2.81±0.51	11.02	7.69
	ViT-H	SA-1B	5.16±0.44	8.15	18.36	4.97±0.68	7.71	*12.36*	2.75±0.41	*6.78*	7.95
SAM 2	Hiera-T	SA-V	4.65±0.28	**4.86**	**7.46**	3.86±0.64	7.79	13.14	3.11±0.50	9.45	3.57
	Hiera-B+	SA-V	4.67±0.33	8.49	15.86	3.75±0.61	*7.44*	12.67	3.02±0.47	9.51	*4.79*
	Hiera-L	SA-V	*4.61±0.29*	9.51	13.28	3.84±0.62	9.12	12.35	2.83±0.41	7.46	*4.10*
	Hiera-H	SA-V	**4.39±0.23**	7.55	10.03	3.42±0.51	**6.12**	**9.34**	2.74±0.38	6.51	4.87
SAM 2.1	Hiera-T	SA-V	4.67±0.32	7.08	8.99	3.91±0.68	8.45	*11.88*	3.11±0.50	9.75	**3.35**
	Hiera-B+	SA-V	4.63±0.32	9.72	14.30	3.76±0.62	8.16	12.35	3.04±0.49	9.59	4.70
	Hiera-L	SA-V	4.67±0.32	11.75	15.39	3.88±0.62	7.47	11.95	2.87±0.43	8.35	4.51
	Hiera-H	SA-V	4.44±0.25	10.35	*9.48*	3.51±0.52	*6.78*	9.91	2.81±0.39	7.41	4.50

Evaluation results of state-of-the-art interactive segmentation methods

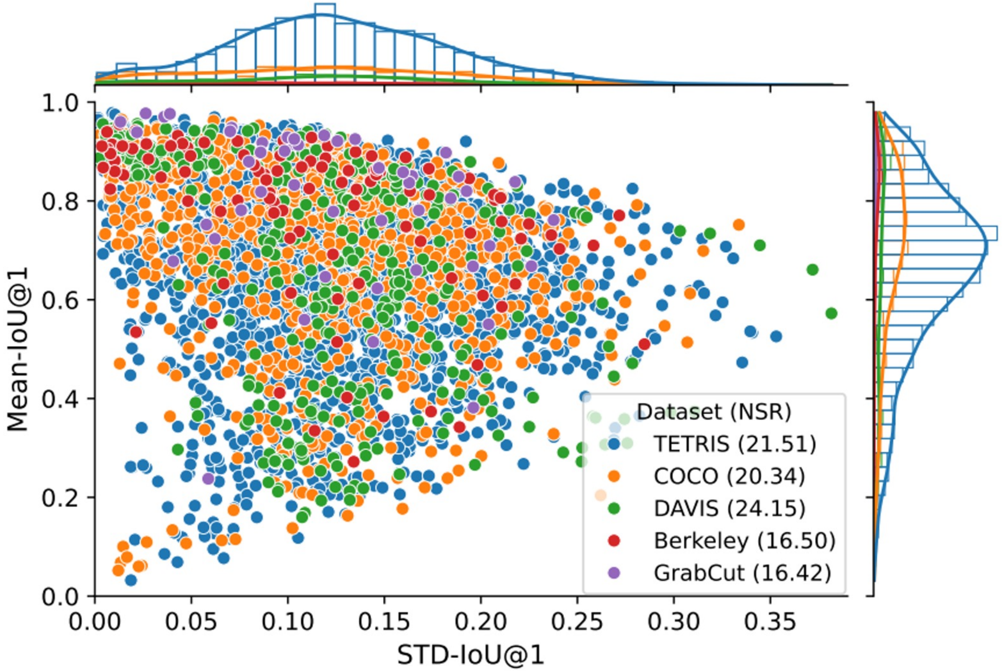
According to **Sample NoC** and **ΔGR** values, the best annotation time is achieved by SAM 2, CFR-ICL and SimpleClick, while the two latter methods are less robust compared to SAM-like methods

Datasets' difficulty

Scatter plot of the mean vs. standard deviation (STD) of IoU for the first real-users clicks

Difficulty score for every dataset — Noise to Signal Ratio (NSR). The higher score means the harder dataset for annotation:

$$\textbf{NSR} = \frac{\textbf{STD of IoU}}{\textbf{Mean of IoU}}$$



Main findings

Baseline strategy underestimates the real-world **annotation time** from **5% up to 29%**

Currently there is **NO segmentation method** that is optimal in terms of both performance and robustness on all datasets

DAVIS, with its 24.15 NSR, stands as the **hardest** dataset for annotation

Annotation time of users from different clicking groups varies from **3% up to 79%**