2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

doi 10.26089/NumMet.v27r444 УДК 004.932

Исследование устойчивости моделей интерактивной сегментации к состязательным кликовым подсказкам

А. В. Москаленко
Московский государственный университет имени М. В. Ломоносова,
факультет вычислительной математики и кибернетики, Москва, Российская Федерация
Институт искусственного интеллекта (AIRI), Москва, Российская Федерация
ORCID: 0000-0003-4965-0867, e-mail: and.v.moskalenko@gmail.com

В. И. Шахуро
Московский государственный университет имени М. В. Ломоносова,
факультет вычислительной математики и кибернетики, Москва, Российская Федерация
Институт искусственного интеллекта (AIRI), Москва, Российская Федерация
ORCID: 0000-0002-1586-9257, e-mail: shakhuro@airi.net

Д. С. Ватолин
Московский государственный университет имени М. В. Ломоносова,
факультет вычислительной математики и кибернетики, Москва, Российская Федерация
Институт искусственного интеллекта МГУ имени М. В. Ломоносова, Москва, Российская Федерация
ORCID: 0000-0002-8893-9340, e-mail: dmitriy@graphics.cs.msu.ru

Аннотация: Модели интерактивной сегментации широко используются для выделения объектов и ускорения разметки данных с помощью пользовательских подсказок. Наиболее простым, интуитивным и распространенным в большинстве систем разметки типом таких подсказок являются клики по интересующим объектам. Однако большинство существующих подходов к обучению и тестированию таких моделей опирается на нереалистичные искусственно сгенерированные кликовые подсказки, тогда как влияние реальных пользовательских вводов остается недостаточно изученным. На основе краудсорсингового исследования с участием 1800 пользователей в этой работе мы показали, что стандартные стратегии генерации кликов не всегда соответствуют реальному пользовательскому поведению, а высокое качество на традиционных тестах не гарантирует надежной работы в практических условиях. Полный перебор возможных положений пользовательских вводов мог бы решить проблему оценки устойчивости моделей, однако он требует колоссальных вычислительных ресурсов и неосуществим за разумное время. Поэтому мы переформулировали задачу как эффективную состязательную атаку белого ящика на модель интерактивной сегментации. Мы предложили метрики оценки устойчивости и с их помощью провели экспериментальное сравнение моделей на популярных наборах данных.

Ключевые слова: компьютерное зрение, интерактивная сегментация, сегментация с подсказками, кликовые подсказки, состязательные атаки, оценка устойчивости.

Благодарности: Работа выполнялась с использованием суперкомпьютера “МГУ-270” МГУ имени М. В. Ломоносова.

Для цитирования: Москаленко А.В., Шахуро В.И., Ватолин Д.С. Исследование устойчивости моделей интерактивной сегментации к состязательным кликовым подсказкам // Вычислительные методы и программирование. 2026. **27**, № 4. 723–740. doi 10.26089/NumMet.v27r444.

© А. В. Москаленко, В. И. Шахуро, Д. С. Ватолин

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

Assessing the resilience of guided segmentation models under adversarially perturbed point cues

Andrey V. Moskalenko
Lomonosov Moscow State University,
Faculty of Computational Mathematics and Cybernetics, Moscow, Russia
Artificial Intelligence Research Institute (AIRI), Moscow, Russia
ORCID: 0000-0003-4965-0867, e-mail: and.v.moskalenko@gmail.com

Vladislav I. Shakhuro
Lomonosov Moscow State University,
Faculty of Computational Mathematics and Cybernetics, Moscow, Russia
Artificial Intelligence Research Institute (AIRI), Moscow, Russia
ORCID: 0000-0002-1586-9257, e-mail: shakhuro@airi.net

Dmitriy S. Vatolin
Lomonosov Moscow State University,
Faculty of Computational Mathematics and Cybernetics, Moscow, Russia
Institute of Artificial Intelligence of Lomonosov Moscow State University, Moscow, Russia
ORCID: 0000-0002-8893-9340, e-mail: dmitriy@graphics.cs.msu.ru

Abstract: Guided segmentation models are widely used to extract objects and accelerate data annotation through user-provided prompts. The simplest, most intuitive, and most common type of such prompts in most annotation systems is clicks on objects of interest. However, most existing approaches to training and evaluating these models rely on unrealistic synthetically generated click prompts, while the impact of real user inputs remains insufficiently studied. Based on a crowdsourcing study involving 1800 users, we show that standard click generation strategies do not always reflect real user behavior, and that strong performance on conventional benchmarks does not guarantee reliable operation in practical settings. An exhaustive search over all possible user input locations could address the problem of assessing model resilience, but it would require enormous computational resources and is infeasible within a reasonable time. We therefore reformulate the problem as an efficient white-box adversarial attack on an interactive segmentation model. We propose resilience evaluation metrics and use them to experimentally compare models on popular datasets.

Keywords: computer vision, interactive segmentation, promptable segmentation, user clicks, adversarial attack, resilience assessment.

Acknowledgements: The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

For citation: A. V. Moskalenko, V. I. Shakhuro and D. S. Vatolin, "Assessing the resilience of guided segmentation models under adversarially perturbed point cues," Numerical Methods and Programming. **27** (4), 723–740 (2026). doi 10.26089/NumMet.v27r444.

---

1. Введение. Методы интерактивной сегментации широко применяются в интерактивных пользовательских интерфейсах для быстрого удаления и выделения объектов, сегментации больших наборов данных для обучения моделей, аннотирования медицинских изображений и решения других задач разметки изображений. По сравнению с автоматическими подходами к семантической сегментации интерактивные методы позволяют получать более качественные маски, лучше соответствующие запросам пользователя, опираясь лишь на простые пользовательские подсказки, такие как клики мышью, ограничивающие рамки, текстовые описания или грубые маски сегментации. В настоящей работе термином "клик" обозначается щелчок мышью, задающий точечную подсказку на изображении. Этот краткий термин используется далее для согласования с принятой в литературе терминологией интерактивной сегментации. Наибольшее число моделей интерактивной сегментации поддерживают кликовые подсказки и предполагают выделение объектов на изображении по нескольким положительным или отрицательным кликам пользователя,

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

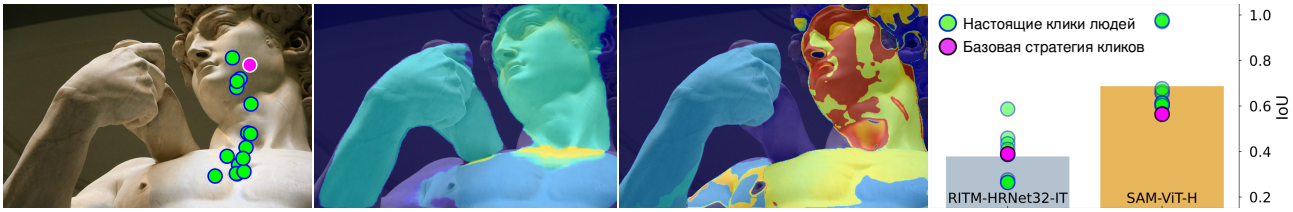


Рис. 1. Слева: кликовые подсказки настоящих пользователей (зеленые точки) и синтетический клик базовой стратегии (пурпурный). Две картинки в центре: тепловые карты качества IoU популярных моделей, демонстрирующие существенную зависимость качества от положения клика. Справа: значения IoU для пользовательских подсказок (зеленые точки) и базовой стратегии (пурпурные точки). Группы точек соответствуют методу RITM с сетью HRNet32 и итеративным обучением и методу Segment Anything с визуальным трансформером ViT-Huge

Fig. 1. Left: click prompts from real users (green points) and the synthetic click from the baseline strategy (purple). Two central pictures: IoU quality heatmaps for popular models demonstrating a significant dependence of quality on the click location. Right: IoU values for user prompts (green points) and the baseline strategy (purple points). The groups correspond to RITM with HRNet32 and iterative training and Segment Anything with a ViT-Huge image encoder

образующим траекторию кликов. Интерактивность таких моделей усложняет их тестирование, поскольку оценивать каждый новый подход с участием реальных пользователей и сравнивать его со все большим числом предшествующих методов практически невозможно. Поэтому в распространенных бенчмарках кликовые подсказки ставят не реальные пользователи, их автоматически генерируют с учетом предыдущих кликов. В большинстве существующих подходов стратегия генерации состоит в выборе области с наибольшей ошибкой на предыдущем раунде взаимодействия и в установке клика в точке, наиболее удаленной от границ этой области. Далее эту схему генерации кликов мы называем базовой стратегией. Такой подход не учитывает особенности того, как реальные пользователи взаимодействуют с моделями. Между тем именно пользовательский ввод в реальных условиях определяет качество результата, и, как показало наше исследование, этот результат может существенно различаться у разных пользователей. Цель данной работы — разработать и исследовать протокол оценки устойчивости моделей интерактивной сегментации к изменению положения пользовательских кликовых подсказок.

Для анализа поведения моделей в реальных сценариях использования и оценки их устойчивости к пользовательским подсказкам нами было проведено масштабное исследование с участием 1800 пользователей. Было установлено, что качество получаемых масок (метрики IoU (Intersection-over-Union) и Boundary-IoU) для одной и той же модели и одного и того же объекта заметно различается между пользователями, несмотря на то, что пользователи ставят клики близко друг к другу. Это указывает на высокую чувствительность моделей к естественному шуму во входных подсказках. Такая нестабильность (рис. 1) может быть связана с обучением на синтетических кликах базовой стратегии, которая недостаточно точно воспроизводит поведение пользователей. Поэтому качество исследуемых моделей в реальных условиях остается неизвестным.

Ввиду высокой стоимости сбора данных реальных пользователей мы провели масштабный перебор допустимых кликовых подсказок для наиболее популярных моделей интерактивной сегментации. Было установлено, что даже сдвиг клика на один пиксель может приводить к значительным изменениям качества (см. рис. 1–3). Однако полный перебор всех возможных подсказок вычислительно неосуществим за разумное время. Поэтому мы предложили переформулировку задачи в виде состязательной атаки белого ящика (white-box) в пространстве кликовых подсказок с регуляризатором на валидность этих подсказок. Используя эту методологию, мы предложили новую метрику для оценки устойчивости моделей сегментации к изменению кликовых подсказок и провели масштабный анализ популярных моделей.

Основной вклад работы заключается в следующем:

- Мы провели контролируемое краудсорсинговое исследование с участием 1800 пользователей и показали, что реальные кликовые подсказки систематически отклоняются от базовой стратегии выбора центра области с максимальной ошибкой, на которой основано большинство существующих методов.

- Предложена эффективная состязательная атака белого ящика для генерации состязательных кликовых подсказок, значительно ускоряющая стратегию полного перебора за счет дифференцируемой

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

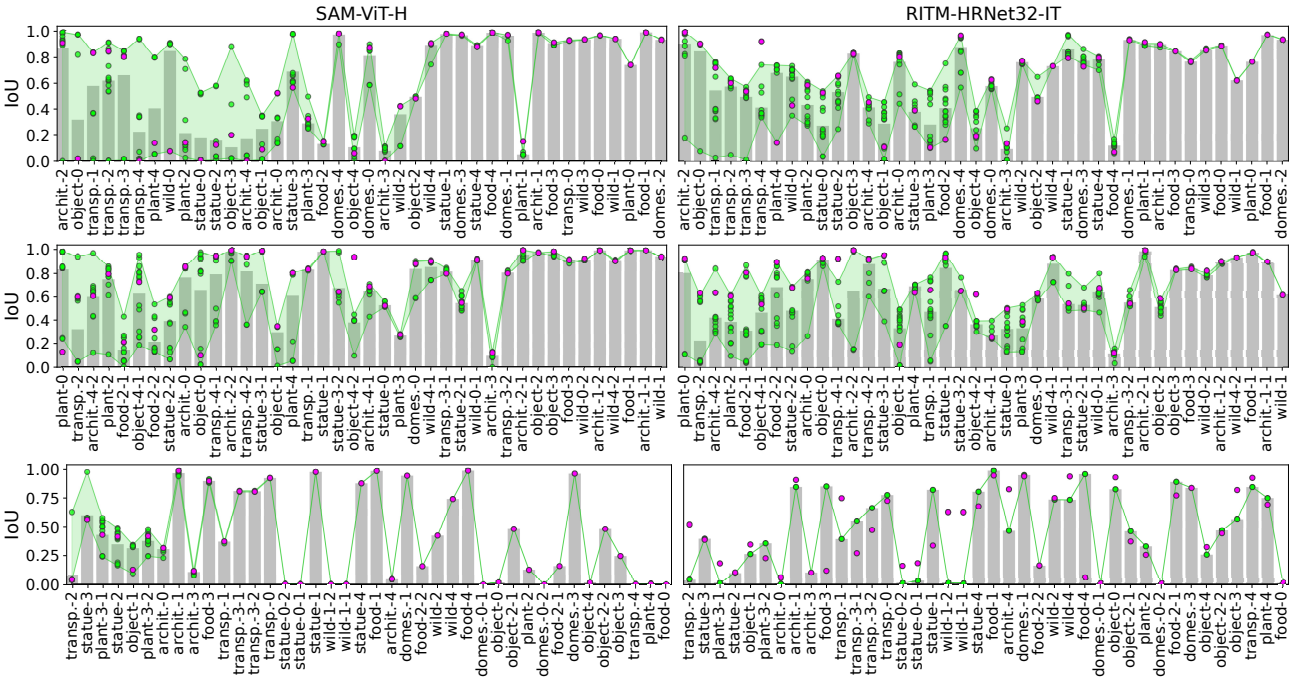


Рис. 2. Разброс IoU на 40 тестовых изображениях для двух моделей после первого, второго положительного и второго отрицательного раундов. Зеленые точки — клики пользователей, пурпурные — клики базовой стратегии. Столбцы отсортированы по разбросу

Fig. 2. IoU spread on 40 test images for two models after the first, second positive and second negative rounds. Green points — user clicks, purple — basic strategy clicks. Columns are sorted by spread

отрисовки и прямой градиентной оптимизации пользовательских вводов. Исходный код оценки и материалы экспериментов доступны по адресу https://github.com/image-seg/adv-clicks.

- Предложены метрики устойчивости моделей интерактивной сегментации и проведена масштабная оценка 23 современных моделей интерактивной сегментации на 4 наборах данных.

Предложенная методология открывает новые возможности для разработки более качественных и надежных моделей интерактивной сегментации, подходящих для применения в реальных условиях.

2. Обзор литературы.

2.1. Тестирование интерактивной сегментации. GrabCut [1] стал первым набором данных для исследования задачи интерактивной сегментации. Затем был адаптирован набор данных Berkeley [2, 3]. Протокол оценки требовал ручного взаимодействия пользователей с методом, а в качестве объективной метрики качества использовалась IoU, рассчитываемая как $\mathrm{IoU}\,(M, M^*) = \dfrac{|M \cap M^*|}{|M \cup M^*|}$, где M и M^* — предсказанная и эталонная маски сегментации соответственно. В работе [4] предложена автоматическая процедура оценки интерактивной сегментации с кликовыми подсказками на наборах данных PASCAL VOC 2012 [5] и COCO [6]; в ней клики ставились строго в центре наибольшей области ошибки, а качество оценивалось по IoU. В последующей работе [7] для интерактивной сегментации были адаптированы наборы DAVIS [8] и SBD [9] с той же стратегией синтетической генерации кликов.

Наиболее распространенная метрика интерактивной сегментации — число кликов (Number of Clicks, NoC), необходимых для достижения заданного IoU [10, 11]. Однако NoC одинаково штрафует случаи, когда порог достигается лишь на последних взаимодействиях, и случаи, когда он не достигается вовсе. Кроме того, в [12] отмечено, что базовая стратегия генерации кликов может приводить к переобучению модели под NoC без соответствующего улучшения качества в реальном применении. Поэтому в качестве основной метрики мы используем площадь под кривой IoU-AUC [10]: рассматриваем 10 кликов и нормируем значение к диапазону $[0, 1]$. Поскольку стандартная IoU слабо отражает качество границ объектов, мы дополнительно используем площадь под кривой Boundary-IoU (BIoU-AUC) [13].

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

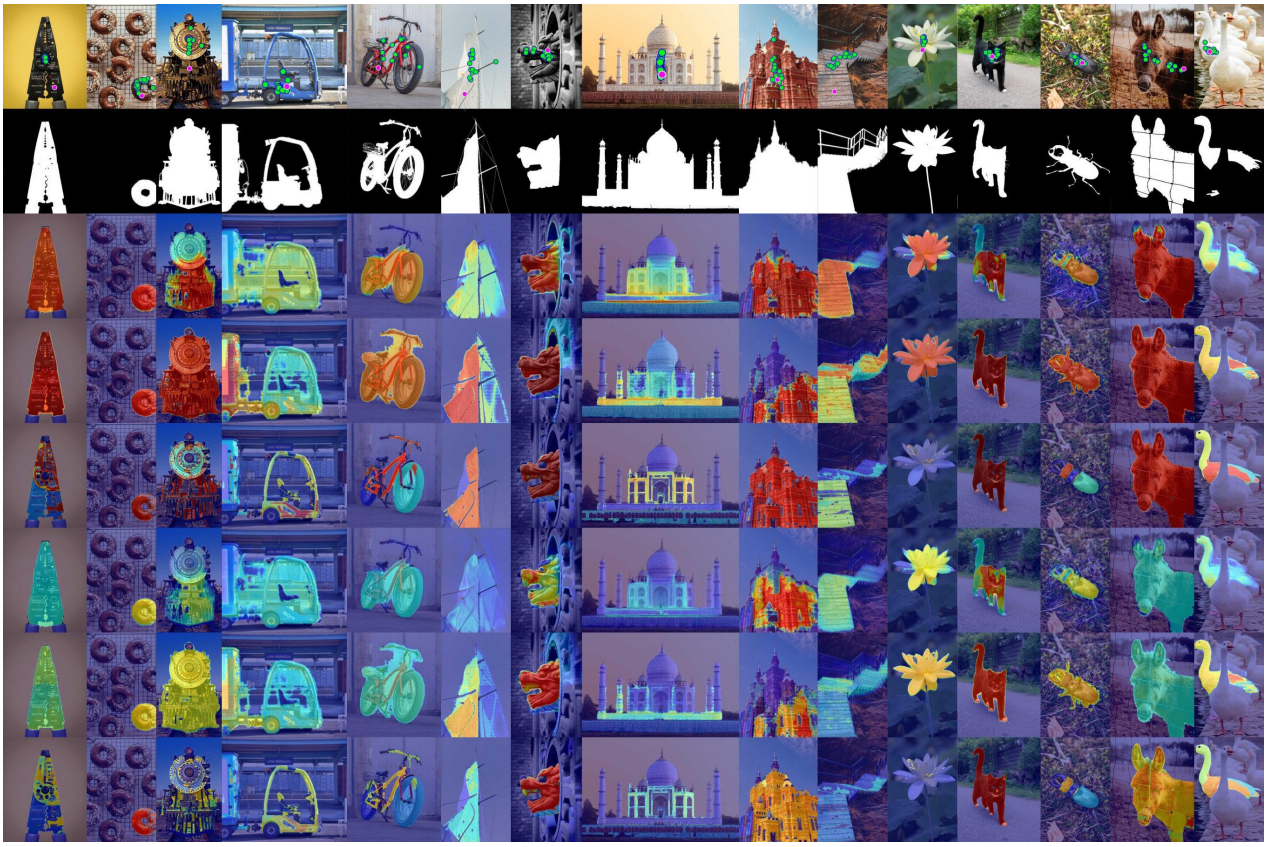


Рис. 3. Верхний ряд: изображения с кликами пользователей (зеленые) и базовой стратегии (пурпурные). Ниже: эталонная маска. Еще ниже: показаны 3 ряда тепловых карт IoU и 3 ряда BIoU для RITM, SimpleClick и SAM соответственно, рассчитанные полным перебором целочисленных положений клика. Более теплые цвета соответствуют более высоким значениям. Резкие локальные изменения IoU/BIoU показывают высокую чувствительность моделей к положению кликовой подсказки

Fig. 3. Top row: visualizations of user clicks (green) and the baseline strategy (purple). Below: the ground-truth mask. Further below: three rows of IoU heatmaps and three rows of BIoU heatmaps for RITM, SimpleClick, and SAM, respectively, calculated via an exhaustive search over integer click locations. Warmer colors correspond to higher values. Sharp local variations in IoU/BIoU indicate the models' high sensitivity to the location of the click prompt

2.2. Пользовательские подсказки. Существуют различные типы пользовательских подсказок. Для выделения крупных областей изображения применялись ограничивающие рамки [4, 1]. В работе [14] выделением объектов управляли с помощью нарисованных вручную штрихов. В [15] исходное выделение выполнялось с помощью ограничивающих рамок, полученных по экстремальным кликам [16], а затем уточнялось штрихами. В [17] предложена случайная стратегия генерации кликов и штрихов, при которой точки случайно выбирались из эталонной маски. В семействе моделей Segment Anything (SAM) [18] используются подсказки в виде кликов, ограничивающих рамок, масок или текста. Большинство существующих моделей поддерживают интерактивные подсказки в виде кликов. С пользовательской точки зрения такой формат является одним из наиболее быстрых и удобных способов взаимодействия: положительный клик используется для добавления области к маске, а отрицательный — для ее исключения.

2.3. Устойчивость моделей сегментации. Устойчивость традиционных методов сегментации уже исследовалась в других работах [19–21]. Однако при отсутствии пользовательских вводов ее можно было оценивать только относительно возмущений изображения. В ряде работ [22–26], посвященных устойчивости SAM, также рассматривались возмущения изображений, а не пользовательских подсказок.

Подходы к состязательным атакам обычно относят к методам черного или белого ящика в зависимости от доступности информации об атакуемой модели. Методы черного ящика [27–29] могут компенсировать недостаток информации большим объемом вычислений. Поскольку мы рассматриваем изображения высокого разрешения и множество последовательных кликовых подсказок, объем вычислений, требуемый

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

для атаки черного ящика и полного перебора, оказывается неосуществимым.

В отличие от предыдущих работ, мы фиксируем входные изображения и исследуем устойчивость относительно пользовательских подсказок. Мы предложили ресурсозатратную стратегию полного перебора, а также эффективную состязательную атаку белого ящика для генерации состязательных кликовых подсказок, которая позволила сформулировать метрики устойчивости.

2.4. Методы интерактивной сегментации. Рассматриваемые методы различаются способом использования подсказок и уточнения маски. RITM [12] обучается на последовательности взаимодействий и использует маску предыдущего раунда как дополнительный вход. CDNet [30] распространяет информацию о кликах по изображению посредством условной диффузии признаков и вероятностей принадлежности пикселей объекту. SimpleClick [31] использует неиерархический визуальный трансформер и совместное кодирование изображения и карт подсказок. CFR-ICL [32] дополняет маску повторными прямыми проходами с неизменными кликами и учитывает последовательные взаимодействия в функции потерь при обучении. GPCIS [33] формулирует сегментацию как попиксельную бинарную классификацию на основе гауссовского процесса с обучаемыми признаками изображения. SAM [18] объединяет кодировщики изображения и подсказок с декодировщиком маски; признаки изображения вычисляются один раз и затем используются при последующих взаимодействиях. SAM-HQ [34] уточняет детали маски дополнительным обучаемым выходным токеном и объединением признаков разных уровней. MobileSAM [35] заменяет тяжелый кодировщик SAM компактным TinyViT, сохраняя декодировщик маски.

3. Исследование устойчивости. Ниже представлены результаты нашего исследования устойчивости моделей интерактивной сегментации с участием реальных пользователей, состоящего из двух раундов взаимодействия кликовыми подсказками. Дополнительно мы анализируем изменение качества моделей при полном переборе возможных положений кликовых подсказок.

3.1. Исследование с участием реальных пользователей. Для исследования поведения реальных пользователей мы отобрали 40 тестовых изображений с платформы Unsplash, на которой фотографы публикуют изображения для свободного использования. Все отобранные изображения не входят в другие используемые наборы данных, что позволяет избежать пересечения с обучающими и тестовыми выборками. Для сбора данных интерактивных взаимодействий мы наняли 1800 пользователей через краудсорсинговую платформу Яндекс.Задания. Каждому участнику предлагалось разметить изображения в простом инструменте аннотирования за два раунда взаимодействия с использованием кликовых подсказок.

Каждому пользователю показывали исходное изображение и то же изображение с наложенной полупрозрачной предсказанной в предыдущем раунде маской (или эталонной маской в первом раунде взаимодействий). В первом раунде разметчикам предлагалось поставить одну кликовую подсказку на целевом объекте, т.е. разрешались только положительные клики. Всего было собрано по 15 взаимодействий уникальных пользователей для каждого из 40 изображений.

После первого раунда кликовые подсказки пользователей были обработаны моделью RITM HRNet18-IT-CL [12], после чего для предсказанных масок рассчитывались ложноположительные и ложноотрицательные ошибки. Затем было отобрано по 40 примеров с максимальными ошибками каждого типа. Для ложноположительных ошибок разметчики добавляли отрицательную подсказку, исключающую лишние области, а для ложноотрицательных — положительную, включающую пропущенные области. Во втором раунде участвовали 1200 пользователей, выполнивших по 15 взаимодействий для каждого из 80 изображений (рис. 4).

3.2. Анализ собранного набора данных. Мы наблюдаем, что положение кликовой подсказки, полученной базовой стратегией, редко совпадает с кликами реальных пользователей. Для объектов сложной геометрии пользователи часто выбирают разные плотные области маски или визуально значимые части объекта, обычно вблизи субъективного центра, который не всегда совпадает с точкой, наиболее удаленной от границ. В среднем расстояние от базового клика до пользовательских превышает 15% размера объекта, а в отдельных случаях достигает более половины. Также разброс качества после второго отрицательного клика оказался меньше, чем после второго положительного, что мы связываем с локальностью изменений, вызванных отрицательной подсказкой.

Кроме того, мы измерили качество интерактивных моделей на собранных пользовательских кликовых подсказках. Результаты показывают, что качество тестируемых моделей может существенно различаться между пользователями. Как видно на рис. 2, более чем для половины тестовых примеров положение клика заметно влияет на качество сегментации на обоих раундах взаимодействия. Это также

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

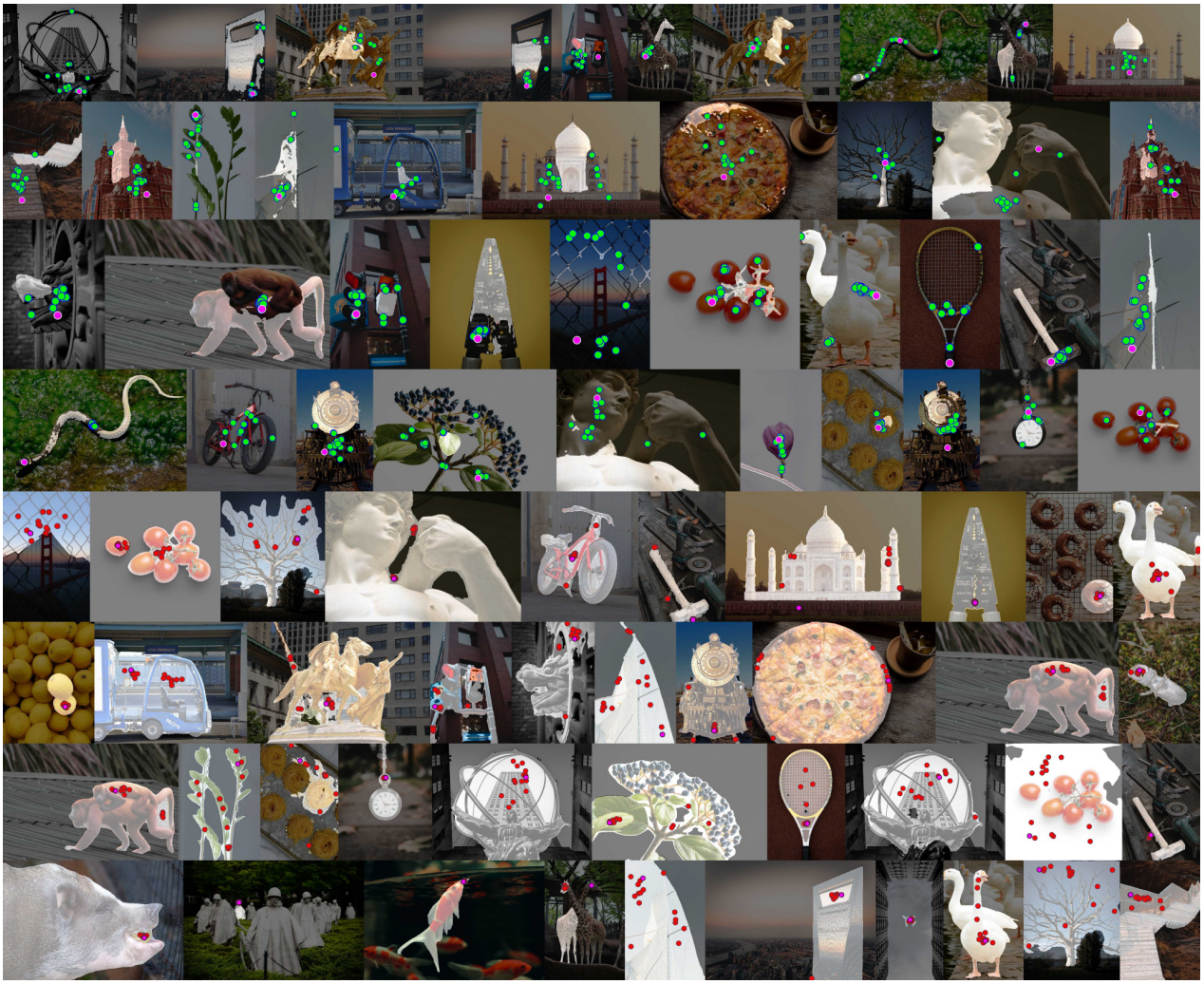


Рис. 4. Визуализация всех 80 тестовых изображений второго раунда взаимодействий с масками, ошибки которых пользователи исправляли положительными (зеленые) и отрицательными (красные) кликовыми подсказками. Дополнительно показаны корректирующие клики базовой стратегии (пурпурные)

Fig. 4. Visualization of all 80 test images from the second round of mask interactions, where users corrected errors using positive (green) and negative (red) click-based prompts. Also shown clicks from the baseline strategy (purple)

указывает на то, что пользователи могут непреднамеренно ставить клики в сложные для модели положения и получать неожиданно низкое качество сегментации.

3.3. Полный перебор. Ввиду высокой стоимости сбора реальных пользовательских подсказок мы дополнительно исследовали качество модели интерактивной сегментации с помощью полного перебора всех целочисленных положений кликовых подсказок. Это, по-видимому, наиболее прямой способ оценить изменение качества при возмущении пользовательского ввода. Однако такой подход требует значительных вычислительных ресурсов: число прямых проходов модели растет квадратично с размером изображения, поэтому провести такую оценку на полных наборах данных за разумное время невозможно.

Мы выполнили полный перебор по пиксельной сетке на небольшом числе тестовых примеров и визуализировали результаты в виде тепловых карт на рис. 3. Цвет каждого пикселя соответствует значению IoU, полученному при постановке кликовой подсказки в этот пиксель; более теплые оттенки обозначают более высокие значения IoU. Полный перебор для одного изображения размером 1024×1024 занимает порядка четырех часов на одной NVIDIA Tesla A100, что мотивировало нас разработать более эффективный подход к оценке устойчивости, описанный ниже.

4. Предлагаемый протокол оценки устойчивости. Неформально определим устойчивость модели интерактивной сегментации как ее способность выдавать одну и ту же маску для любой допустимой

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

пользовательской подсказки, указывающей на один и тот же объект. В предлагаемом протоколе допустимыми считаются подсказки следующего вида: положительные клики долж....ны располагаться в еще не выделенной области маски объекта, а отрицательные — в ранее выделенной области, но за пределами эталонной маски объекта.

4.1. Состязательные кликовые подсказки. Чтобы сократить время обработки с нескольких часов до нескольких секунд по сравнению с полным перебором, необходимо разумно ограничить пространство поиска. В качестве начальной инициализации координат пользовательской подсказки мы выбираем центр области с наибольшей ошибкой, а затем ищем локальный экстремум в ее окрестности.

Мы предлагаем находить положения кликовых подсказок с помощью состязательной атаки белого ящика. Общая схема метода приведена на рис. 5. С помощью дифференцируемого рендеринга [36] кликовые подсказки кодируются картами, на которых их положения отмечены дисками фиксированного радиуса, поскольку большинство моделей принимает пользовательские подсказки именно в таком виде. Радиус является гиперпараметром, зависящим от архитектуры исследуемой модели. Затем мы вычисляем функцию потерь между предсказанной и эталонной масками и выполняем обратное распространение ошибки для оптимизации координат кликовых подсказок в соответствии с выбранной стратегией оптимизации. Используются две стратегии: одна минимизирует IoU, другая максимизирует IoU. Несмотря на ограничения метода градиентного спуска в виде его возможности нахождения лишь локальных экстремумов, получаемых изменений положений кликовых подсказок оказывается достаточно, чтобы существенно повлиять на итоговое качество сегментации.

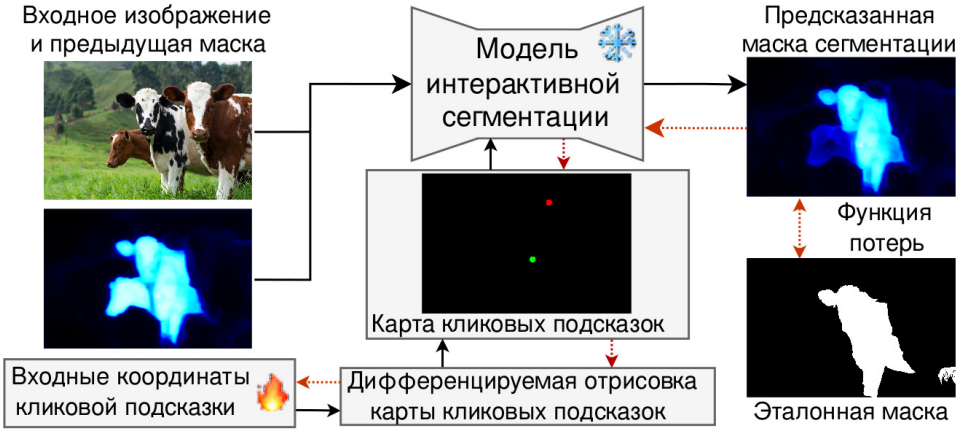


Рис. 5. Предложенная эффективная генерация состязательных кликовых подсказок для оценки устойчивости моделей интерактивной сегментации. Черные стрелки обозначают прямой проход, оранжевые и красные пунктирные — обратное распространение ошибки. Веса модели заморожены, оптимизируются только координаты клика. На карте кликовых подсказок зеленая точка обозначает положительный клик, красная — отрицательный, зеленый цвет обозначает пользовательские подсказки.

Fig. 5. Proposed method for efficient generation of adversarial click prompts to evaluate the resilience of interactive segmentation models. Black arrows indicate the forward pass, while orange and red dashed arrows indicate backpropagation. Model weights are frozen; only the click coordinates are optimized. In the click map, green point denotes a positive click and red — a negative click, green identifies user prompts.

4.2. Предлагаемые метрики. Последовательная оптимизация на каждом раунде взаимодействия задает две траектории кликовых подсказок: минимизирующую и максимизирующую. Траекторию, полученную базовой стратегией, далее называем базовой. По кривым зависимости IoU и BIoU от числа кликов, полученным для минимизирующей и максимизирующей траекторий, мы определяем метрики качества Min и Max и метрику устойчивости D (рис. 6) как разность между Max и Min. Базовая траектория используется как ориентир для сравнения.

Локальный поиск около базовой подсказки позволяет оценить, насколько качество меняется при изменении положения исходной подсказки. Для пользовательского сценария Min характеризует уязвимость к неудачному положению подсказки, Max — достижимое при таком поиске качество, а D — чувствительность к выбору точки. Малое D следует рассматривать совместно с высоким Min: постоянно низкое качество также может давать малый разброс. Базовая траектория сохраняется как привычный ориентир

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

для сравнения. Эти показатели предназначены для сравнительной оценки при одинаковых настройках поиска; они не являются гарантированными глобальными границами качества или моделью вероятности пользовательского выбора. Прямое сопоставление с реальными подсказками приведено в разделе 4.6.

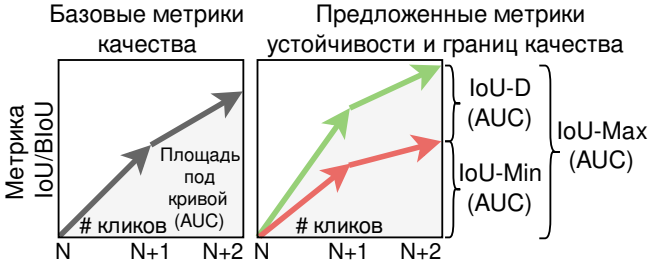


Рис. 6. Метрики качества и устойчивости. Слева: метрики качества IoU/BIoU-AUC. Справа: предлагаемые метрики устойчивости и границ качества. Площадь под кривой нормирована к диапазону $[0, 1]$ для 10 кликов

Fig. 6. Quality and stability metrics. Left: IoU/BIoU-AUC quality metrics. Right: the proposed stability and quality boundary metrics. The area under the curve is normalized to the $[0, 1]$ range for 10 clicks

Метрики IoU/BIoU-Min и IoU/BIoU-Max определяются как площади под кривыми минимизирующей и максимизирующей траекторий, то есть характеризуют качество для худших и лучших подсказок, найденных в рамках выбранного протокола оптимизации. Метрики устойчивости IoU-D/BIoU-D задаются как разности этих площадей. Поскольку наибольшие различия между траекториями обычно наблюдаются в первых раундах взаимодействия, при расчете AUC мы ограничиваемся 10 раундами.

4.3. Детали процесса оптимизации. Чтобы выбрать число шагов оптимизации, мы выполнили 101 шаг для модели RITM-HRNet18-IT [12] и измерили попарные разности координат кликовых подсказок на соседних шагах. Наибольшее влияние оказывают первые итерации, после чего эффект каждого шага постепенно уменьшается. Поэтому для сокращения вычислений число итераций ограничено 10. Оптимизация одного ввода для изображения 400×400 занимает около 10 с на NVIDIA Tesla A100.

Градиентные обновления вычисляются оптимизатором Adam [37]. Для сравнения моделей с разным входным разрешением скорость обучения линейно масштабируется: $LR = \dfrac{5\sqrt{H^2 + W^2}}{400\sqrt{2}}$, где H и W — высота и ширина изображения. Для минимизирующей и максимизирующей траекторий кликовые подсказки инициализируются базовой стратегией и оптимизируются последовательно.

Без ограничений максимизирующая стратегия может порождать клики между частями объекта, но вне его маски. Хотя такие точки дают наилучшее качество, реальный пользователь вряд ли их выберет. Минимизирующая траектория, в свою очередь, может сходиться к точке вне целевого объекта, зануляя предсказание маски, но не соответствуя корректной пользовательской подсказке. Поэтому при оптимизации мы вводим регуляризацию допустимых подсказок. Для положительного клика вычисляется карта расстояний до допустимой ложноотрицательной области, для отрицательного — до ложноположительной; внутри соответствующей области расстояние равно нулю. Регуляризатор R равен среднему по пикселям произведению этой карты на дифференцируемую карту диска нового клика. При поиске максимизирующей траектории минимизируется $(R + L_{\text{Dice}}/1000)$, а при поиске минимизирующей — $(R - L_{\text{Dice}}/1000)$, где L_{Dice} — функция потерь на основе коэффициента Dice [38]. Относительный вес регуляризатора 1000 выбран на основе сравнения норм градиентов слагаемых. Таким образом, регуляризатор минимизируется при обеих стратегиях и обеспечивает корректность получаемых в ходе атаки подсказок.

При оптимизации используется следующая схема обновления координат:

1. Инициализировать оптимизируемые координаты кликовой подсказки согласно базовой стратегии.

2. Запустить градиентную оптимизацию соответствующей функции потерь.

3. Принять новые координаты, если IoU уменьшается/увеличивается, а слагаемое регуляризатора возрастает не более чем на 5% относительно лучшей сохраненной итерации (для объектов тоньше радиуса клика регуляризатор неизбежно ухудшается даже при точном клике, поэтому допускается небольшой запас).

4. Сохранить предсказанную маску сегментации и координаты кликовой подсказки с лучшей итерации оптимизации и использовать их в следующем раунде взаимодействия.

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

4.4. Сравниваемые модели. В табл. 1 и 2 столбец “Методы” обозначает семейство алгоритмов, а столбец “Модели” — конкретную архитектуру и вариант обученных весов. RN34 и RN50 обозначают сверточные сети ResNet с 34 и 50 слоями; HR18 и HR32 — сети HRNet с шириной высокоразрешающей ветви 18 и 32 канала, HR18s — компактный вариант HRNet18. Обозначение ViT относится к визуальному трансформеру (Vision Transformer); B, L и H соответствуют вариантам Base, Large и Huge, XT — компактному xTiny в реализации SimpleClick. Используемое для MobileSAM обозначение ViT-Tiny соответствует архитектуре TinyViT. Суффикс IT у RITM указывает на итеративное обучение с передачей маски предыдущего раунда; CL — на обучение для интерактивной сегментации на объединении COCO и LVIS, SBD — на наборе Semantic Boundaries Dataset. Таким образом, сравниваются как разные семейства методов, так и варианты архитектуры и обучающей выборки внутри одного семейства.

Таблица 1. Показатели качества IoU и устойчивости различных моделей, измеренные на четырех стандартных наборах данных. Лучшие результаты выделены полужирным шрифтом, вторые по качеству подчеркнуты
Table 1. IoU quality and robustness metrics for various models, measured on four standard datasets. The best results are highlighted in bold, and the second-best results are underlined

Методы Methods	Модели Models	GrabCut			Berkeley			DAVIS			COCO-MVal		
		IoU-AUC@10			IoU-AUC@10			IoU-AUC@10			IoU-AUC@10		
		Min↑	Max↑	D↓	Min↑	Max↑	D↓	Min↑	Max↑	D↓	Min↑	Max↑	D↓
MobileSAM	ViT-Tiny	93.69	95.58	1.89	90.11	92.49	2.38	83.16	87.60	4.44	82.32	87.20	4.88
SAM	ViT-B	93.56	96.28	2.72	91.06	93.77	2.71	82.44	88.44	6.00	85.28	89.73	4.45
	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
	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
SAM-HQ	ViT-B	91.92	96.29	4.37	90.59	94.47	3.88	81.52	89.44	7.92	83.62	89.64	6.02
	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
	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
CDNet	RN34-CL	91.72	96.99	5.27	89.54	94.35	4.81	83.56	88.06	4.50	83.46	90.70	7.24
	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
GPCIS	RN50-CL	92.06	96.51	4.45	89.56	94.59	5.03	84.81	89.21	4.41	79.65	91.47	11.83
RITM	HR18s-IT-CL	91.34	96.42	5.08	91.77	94.21	2.44	76.58	86.63	10.06	87.32	92.69	5.37
	HR18-CL	92.84	95.27	2.43	91.47	93.49	2.01	82.48	86.62	4.14	87.18	90.82	3.64
	HR18-IT-CL	95.00	96.54	1.54	93.15	94.90	1.75	80.06	87.76	7.69	88.98	93.39	4.40
	HR32-IT-CL	94.47	96.85	2.39	92.36	95.01	2.65	78.71	88.52	9.81	87.66	93.41	5.74
	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
SimpleClick	ViT-B-CL	95.51	97.63	2.12	**94.09**	95.61	1.51	86.98	90.34	3.36	90.64	93.80	3.17
	ViT-L-CL	**96.41**	97.80	**1.39**	93.04	95.85	2.81	**88.40**	90.79	2.39	91.73	94.37	2.64
	ViT-H-CL	95.48	97.87	2.39	93.44	95.66	2.23	86.88	90.50	3.62	**91.97**	**94.53**	**2.56**
	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
	ViT-B-SBD	95.59	97.36	1.77	93.49	94.71	**1.22**	87.49	89.67	2.19	83.36	89.84	6.48
	ViT-L-SBD	95.19	97.10	1.91	93.09	94.42	1.33	87.82	89.68	**1.86**	85.47	91.12	5.65
	ViT-H-SBD	95.92	97.32	1.39	93.27	94.52	1.25	87.17	89.57	2.40	85.13	91.02	5.89
CFR-ICL	ViT-H-CL	95.56	**98.04**	2.48	93.63	**96.11**	2.48	87.58	**91.45**	3.87	90.49	94.19	3.71

4.5. Результаты. Мы провели оценку устойчивости современных методов с открытым исходным кодом (RITM [12], CDNet [30], SimpleClick [31], CFR-ICL [32] и GPCIS [33]), а также моделей из семейства Segment Anything (SAM [18], SAM-HQ [34] и MobileSAM [35]). Для SAM-подобных моделей растеризация кликовых подсказок выполняется лишь в регуляризующем слагаемом, а непосредственно в модель передаются только координаты. Мы придерживаемся той же процедуры оценки, что и в исходных методах, применяя ZoomIn [11], CFR [32], отражения для аугментации во время тестирования [12], выбор маски по предсказанному показателю [18] и другие приемы авторов исходных работ. Всего мы проверили 23 модели на 4 популярных наборах данных: GrabCut [1], Berkeley [2], DAVIS [8], COCO-MVal [6]. Визуализации состязательных кликовых подсказок, полученных нашей эффективной атакой, приведены на рис. 7.

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

Таблица 2. Показатели качества BIoU и устойчивости различных моделей, измеренные на четырех стандартных наборах данных. Лучшие результаты выделены полужирным шрифтом, вторые по качеству подчеркнуты

Table 2. BIoU quality and robustness metrics for various models, measured on four standard datasets. The best results are highlighted in bold, and the second-best results are underlined

Методы Methods	Модели Models	GrabCut			Berkeley			DAVIS			COCO-MVal		
		BIoU-AUC@10			BIoU-AUC@10			BIoU-AUC@10			BIoU-AUC@10		
		Min↑	Max↑	D↓	Min↑	Max↑	D↓	Min↑	Max↑	D↓	Min↑	Max↑	D↓
MobileSAM	ViT-Tiny	84.67	87.54	2.87	80.41	83.42	3.01	75.73	80.11	4.38	68.34	73.49	5.15
SAM	ViT-B	85.03	88.63	3.60	81.84	85.31	3.47	75.95	82.05	6.09	70.93	75.79	4.86
	ViT-L	86.82	88.72	**1.90**	84.39	86.27	**1.88**	77.25	82.69	5.44	72.76	76.02	**3.26**
	ViT-H	86.01	88.53	2.53	83.30	85.99	2.70	77.20	82.32	5.12	70.84	74.30	3.46
SAM-HQ	ViT-B	83.59	89.30	5.71	82.36	87.31	4.95	75.08	83.37	8.29	70.52	77.13	6.61
	ViT-L	**87.39**	90.22	2.83	85.38	88.16	2.78	75.55	83.93	8.37	72.87	78.13	5.26
	ViT-H	86.60	90.50	3.91	85.52	88.34	2.83	74.14	84.03	9.88	72.23	78.42	6.19
CDNet	RN34-CL	80.74	88.26	7.52	79.87	86.05	6.18	75.68	80.63	4.96	71.22	79.82	8.61
	RN34-SBD	72.76	81.52	8.75	74.71	80.55	5.84	71.77	76.98	5.22	60.76	71.88	11.12
GPCIS	RN50-CL	79.67	86.32	6.65	77.65	85.22	7.57	75.30	80.58	5.28	66.92	81.75	14.83
RITM	HR18s-IT-CL	79.80	86.11	6.31	81.63	84.87	3.24	68.58	78.58	10.00	75.95	82.73	6.78
	HR18-CL	80.07	83.93	3.86	80.60	83.48	2.88	74.21	78.63	4.42	75.23	80.07	4.84
	HR18-IT-CL	83.54	85.81	2.27	83.76	86.36	2.60	72.16	79.99	7.83	78.35	84.25	5.90
	HR32-IT-CL	82.81	86.72	3.91	83.31	87.03	3.71	71.31	81.10	9.79	77.23	84.54	7.31
	HR18-IT-SBD	77.77	82.00	4.23	76.74	80.92	4.18	72.32	77.02	4.71	67.73	77.06	9.32
SimpleClick	ViT-B-CL	84.99	88.49	3.50	86.25	88.42	2.17	79.59	83.39	3.80	81.14	85.57	4.43
	ViT-L-CL	86.54	89.26	2.73	85.68	89.31	3.63	81.54	84.29	2.75	82.85	86.80	3.95
	ViT-H-CL	85.77	89.38	3.61	86.17	89.26	3.09	80.16	84.11	3.95	**83.28**	**87.03**	3.75
	ViT-XT-SBD	79.80	82.79	2.99	75.66	80.17	4.51	66.81	75.46	8.65	66.50	76.97	10.47
	ViT-B-SBD	83.62	86.50	2.88	83.05	85.06	2.01	78.86	81.62	2.76	72.08	79.97	7.90
	ViT-L-SBD	83.33	86.38	3.05	82.98	85.15	2.17	79.41	81.67	**2.27**	74.66	81.75	7.09
	ViT-H-SBD	84.57	86.74	2.17	83.41	85.35	1.95	78.88	81.72	2.84	74.54	81.95	7.41
CFR-ICL	ViT-H-CL	87.14	**90.97**	3.83	**87.53**	**91.05**	3.51	**81.61**	**85.85**	4.24	81.62	86.57	4.94

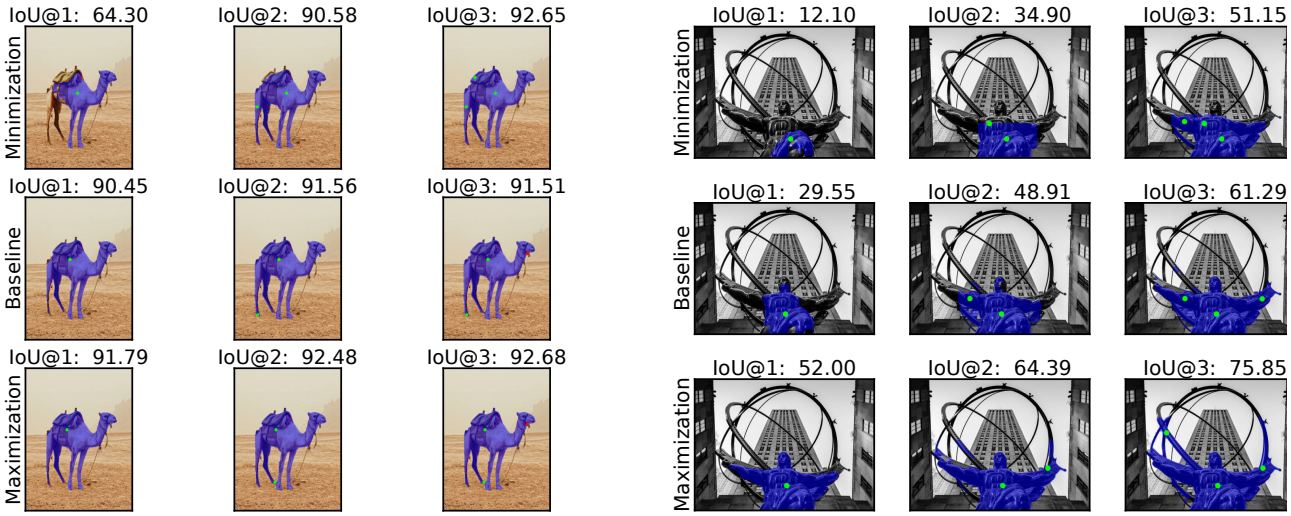


Рис. 7. Траектории трех последовательных кликов для модели RITM-HRNet18-IT: минимизирующая, базовая и максимизирующая. Синим показаны маски, зеленым — положительные клики, красным — отрицательные

Fig. 7. Trajectories of three successive clicks for the RITM-HRNet18-IT model: minimizing, baseline, and maximizing. Predicted masks are shown in blue, positive clicks — in green, and negative clicks — in red

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

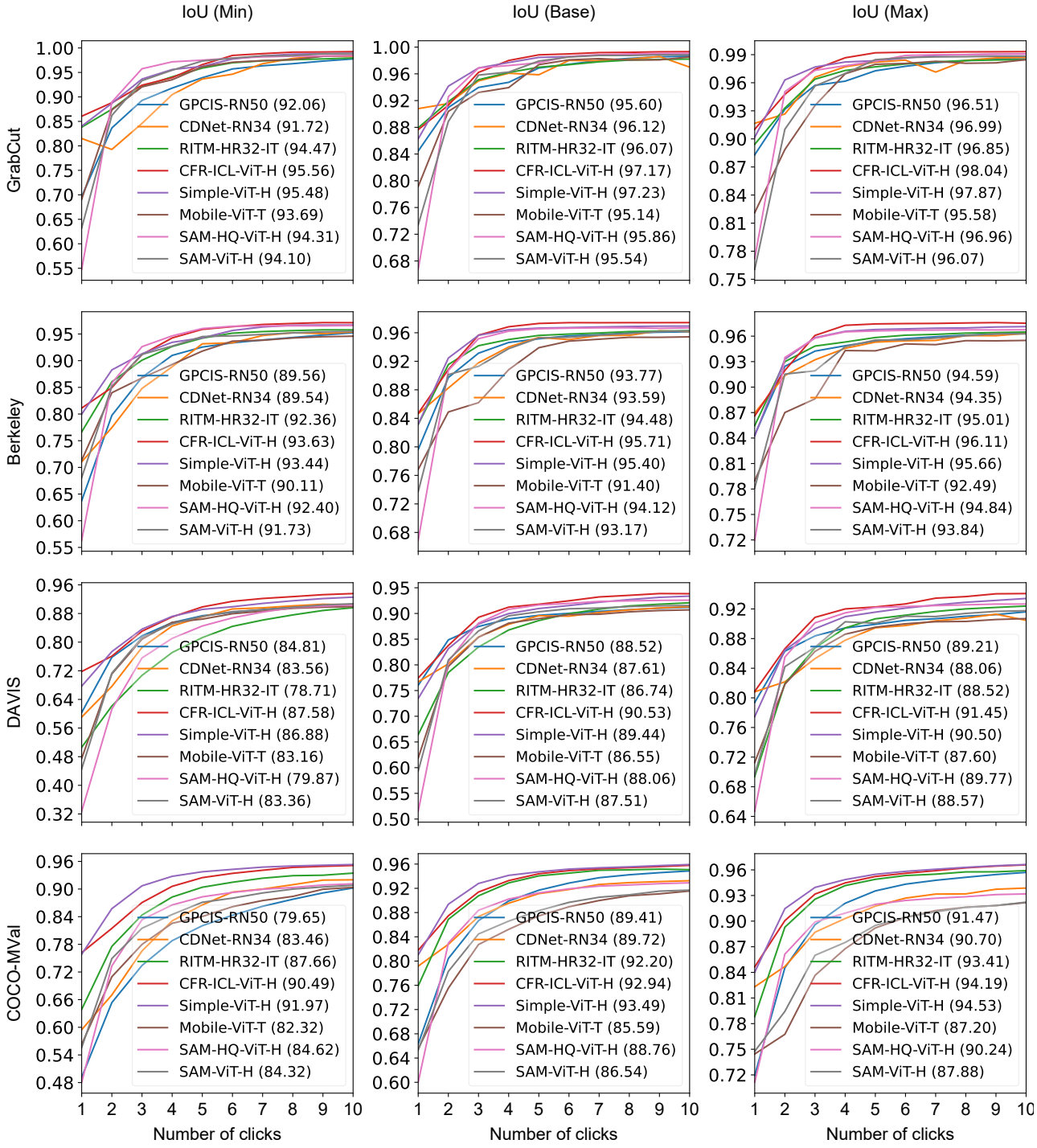


Рис. 8. Минимизирующая, базовая и максимизирующая траектории IoU для 8 семейств методов на 4 наборах данных. Агрегированное значение IoU-AUC указано в скобках. Каждая строка соответствует одному набору данных; столбцы показывают минимизирующую, базовую и максимизирующую стратегии соответственно

Fig. 8. Minimizing, baseline, and maximizing IoU trajectories for 8 families of methods across 4 datasets. The aggregated IoU-AUC value is indicated in parentheses. Each row corresponds to one dataset; the columns show the minimizing, baseline, and maximizing strategies, respectively

Полученные метрики приведены в табл. 1 и 2. Визуализация минимизирующей, базовой и максимизирующей траекторий представлена на рис. 8. Для наглядности показаны только самые крупные модели каждого семейства. Модели SimpleClick [31] и CFR-ICL [32] показали высокое качество минимизирующих и максимизирующих траекторий.

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

На основании результатов нашего краудсорсингового исследования с участием реальных пользователей (рис. 2 и 3) и оценки состязательной устойчивости (табл. 1, 2 и рис. 8) можно заключить, что современные модели интерактивной сегментации чрезвычайно чувствительны к положениям пользовательских кликов, что приводит к нестабильной работе в реальных условиях. В таблицах Min и Max приведены как нормированные площади под кривыми IoU/BIoU для 10 кликов, умноженные на 100; D выражена в процентных пунктах. Стрелки вверх и вниз указывают предпочтительное направление изменения показателя. Устойчивость оценивается по D совместно с качеством Min.

Полный перебор по пиксельной сетке показывает, что кликовые подсказки в некоторых координатах могут неожиданно вызывать значительное падение качества сегментации. Кроме того, для кликов, полученных с помощью нашей состязательной атаки, качество может существенно изменяться даже в небольшой семантически однородной области. Мы связываем такое нежелательное поведение с тем, что модель выделяет часть объекта, а не объект целиком. Задача интерактивной сегментации является некорректно поставленной — для одного изображения и пользовательской подсказки возможны несколько вариантов масок сегментации, поэтому подобная неоднозначность в определенной степени неизбежна. Кроме того, сложность формы объекта существенно влияет на качество модели. Однако при разработке моделей интерактивной сегментации следует стремиться свести эти эффекты к минимуму, чтобы гарантировать стабильное качество предсказания у разных пользователей.

Дополнительно мы исследовали корреляции между метриками качества и устойчивости. Из рис. 9 видно, что IoU/BIoU-Base сильно коррелирует с IoU/BIoU-Max, т.е. ранжирование по базовому протоколу близко к ранжированию по найденной максимизирующей траектории и не описывает возможные падения качества. IoU/BIoU-D сильно коррелирует с IoU/BIoU-Min, тогда как корреляция с IoU/BIoU-Max слабее. Мы связываем это с тем, что большая часть разброса качества обусловлена снижением показателей минимизирующей траектории. Отметим, что оптимизировать кликовые подсказки для максимизирующей траектории значительно сложнее, чем для минимизирующей, что также заметно по абсолютным значениям прироста и падения качества относительно базовой стратегии (см. табл. 1, 2 и рис. 8).

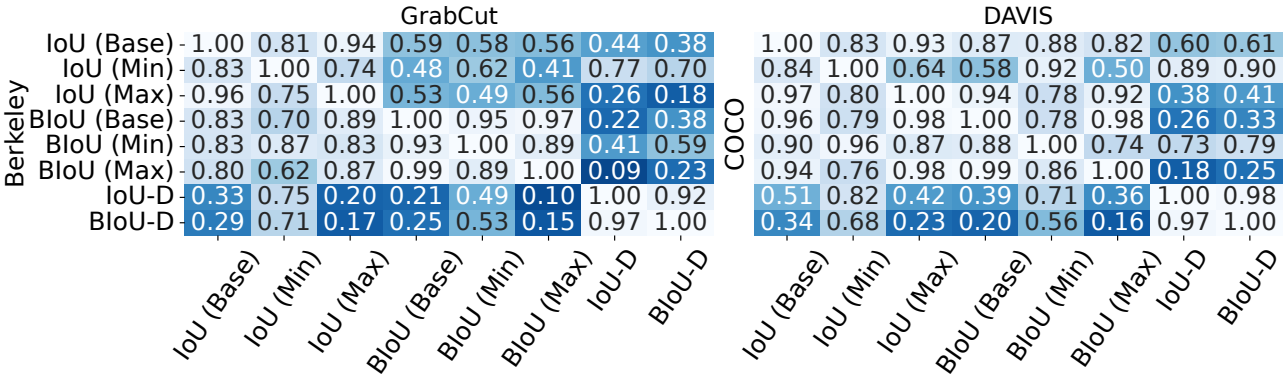


Рис. 9. Модули коэффициентов ранговой корреляции Спирмена между метриками качества и устойчивости. Слева: для наборов данных GrabCut (выше диагонали) и Berkeley (ниже диагонали). Справа: аналогично для наборов данных DAVIS и COCO

Fig. 9. Absolute values of Spearman rank correlation coefficients between quality and stability metrics. Left: for the GrabCut (above the diagonal) and Berkeley (below diagonal) datasets. Right: similarly for the DAVIS and COCO datasets

4.6. Сопоставление результатов атаки с пользовательскими данными. Для прямой проверки связи состязательной оценки с поведением пользователей мы выполнили дополнительный эксперимент на тех же 40 изображениях краудсорсингового исследования. Для каждой из 23 моделей были использованы 600 первых пользовательских кликов (по 15 на изображение) и результаты минимизирующей и максимизирующей стратегий на этих же изображениях. Для каждой пары “модель — изображение” были определены границы L и H метрики IoU после первого клика соответствующей стратегии. В интервал $[L, H]$ попали 58.51% пользовательских значений IoU, 6 моделей имеют покрытие более 70%. Для рангового сравнения сначала усреднялись по 40 изображениям величины L, $H - L$, минимум и диапазон IoU по 15 пользовательским кликам, затем вычислялись корреляции Спирмена по 23 моделям. Корреляция средней нижней границы с пользовательским минимумом составила $\rho = 0.9486$, а для средней ширины

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

интервала с пользовательским диапазоном $\rho = 0.9239$. Таким образом, эксперимент подтверждает целесообразность использования предложенной состязательной оценки для сравнительного ранжирования моделей и одновременно демонстрирует ограниченность покрытия отдельных пользовательских результатов, обусловленную ограничениями градиентного спуска.

4.7. Ограничения. Предложенный протокол требует доступа к градиентам модели и эталонной маске. Поиск локален, зависит от инициализации, числа шагов и регуляризации и может пропускать некоторые уязвимые положения клика. Регуляризация является мягким ограничением допустимости и не гарантирует ее строгого выполнения. Последовательная оптимизация также не дает глобального экстремума по всем траекториям взаимодействия. Краудсорсинговое исследование ограничено 40 изображениями и только первыми раундами взаимодействий ввиду экспоненциального роста возможных комбинаций последующих раундов взаимодействий.

5. Заключение. Краудсорсинговое исследование показало расхождение между широко используемой базовой стратегией и реальными подсказками пользователей. Предложенные метрики Min, Max и D позволяют оценивать качество найденных неблагоприятных и благоприятных траекторий и чувствительность к положению кликовой подсказки. На 40 изображениях для первого клика ранжирование 23 моделей по нижней границе согласуется с пользовательским минимумом ($\rho = 0.9486$), однако среднее покрытие пользовательских значений IoU составляет 58.51%, что указывает на ограниченность найденного интервала. Модели SimpleClick и CFR-ICL показали высокое качество минимизирующих траекторий, а преимущество по устойчивости зависит от специфики тестового набора данных. Например, на COCO-MVal модель SimpleClick ViT-H-CL одновременно достигает наибольшего IoU-Min (91.97%) и наименьшего IoU-D (2.56%). Поэтому при выборе модели для практического применения следует учитывать как базовое качество, так и предложенные метрики: качество минимизирующей траектории и разброс результатов при изменении положения подсказки.

Список литературы

1. *Rother C., Kolmogorov V., Blake A.* GrabCut — interactive foreground extraction using iterated graph cuts // ACM Transactions on Graphics (TOG). 2004. **23**, N 3. 309–314. doi 10.1145/1186562.1015720.
2. *Martin D., Fowlkes C., Tal D., Malik J.* A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics // Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vancouver, BC, Canada, July 07–14, 2001. New York: IEEE Press, 2001. pp. 416–423. doi 10.1109/ICCV.2001.937655.
3. *McGuinness K., O'Connor N.E.* A comparative evaluation of interactive segmentation algorithms // Pattern Recognition. 2010. **43**, N 2. 434–444. doi 10.1016/j.patcog.2009.03.008.
4. *Xu N., Price B., Cohen S., et al.* Deep interactive object selection // Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, Nevada, USA, June 26–July 1, 2016. New York: IEEE Press, 2016. pp. 373–381. doi 10.1109/CVPR.2016.47.
5. *Everingham M., Van Gool L., Williams C.K.I., et al.* The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. 2012.
6. *Lin T.-Y., Maire M., Belongie S., et al.* Microsoft COCO: Common objects in context // Proceedings 13th European Conference. Computer Vision — ECCV 2014, Zurich, Switzerland, September 6–12, 2014. ECCV 2014. Lecture Notes in Computer Science, Vol. 8693. Cham: Springer, 2014. pp. 740–755. doi 10.1007/978-3-319-10602-1_48.
7. *Li Z., Chen Q., Koltun V.* Interactive image segmentation with latent diversity // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18–23, 2018. IEEE Press, 2018. pp. 577–585. doi 10.1109/CVPR.2018.00067.
8. *Perazzi F., Pont-Tuset J., McWilliams B., et al.* A benchmark dataset and evaluation methodology for video object segmentation // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 27–30, 2016. IEEE Press, 2016. pp. 724–732. doi 10.1109/CVPR.2016.85.
9. *Hariharan B., Arbelaez P., Bourdev L., et al.* Semantic contours from inverse detectors // 2011 International Conference on Computer Vision, Barcelona, Spain, November 06–13, 2011. IEEE Press, 2011. pp. 991–998. doi 10.1109/ICCV.2011.6126343.
10. *Jang W.D., Kim C.S.* Interactive image segmentation via backpropagating refinement scheme // In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, June 15–20, 2019. IEEE Press, 2019. pp. 5292-5301. doi 10.1109/CVPR.2019.00544.

2026, **27** (4), 723–740. doi 10.26089/NumMet.v27r444

11. *Sofiiuk K., Petrov I., Barinova O., Konushin A.* f-BRS: Rethinking backpropagating refinement for interactive segmentation // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, June 13–19, 2020. IEEE Press, 2020. pp. 8620–8629. doi 10.1109/CVPR42600.2020.00865.
12. *Sofiiuk K., Petrov I.A., Konushin A.* Reviving iterative training with mask guidance for interactive segmentation // 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, October 16–19, 2022. IEEE Press, 2022. pp. 3141–3145. doi 10.1109/ICIP46576.2022.9897365.
13. *Cheng B., Girshick R., Dollár P., et al.* Boundary IoU: Improving object-centric image segmentation evaluation // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, June 20–25, 2021. IEEE Press, 2021. pp. 15329–15337. doi 10.1109/CVPR46437.2021.01508.
14. *Gueziri H.-E., McGuffin M., Laporte C.* Latency management in scribble-based interactive segmentation of medical images // IEEE Transactions on Biomedical Engineering. 2018. **65**, N 5. 1140–1150. doi 10.1109/TBME.2017.2777742.
15. *Agustsson E., Uijlings J.R.R., Ferrari V.* Interactive full image segmentation by considering all regions jointly // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, June 15–20, 2019. IEEE Press, 2019. pp. 11614–11623. doi 10.1109/CVPR.2019.01189.
16. *Papadopoulos D.P., Uijlings J.R.R., Keller F., Ferrari V.* Extreme clicking for efficient object annotation // 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 22–29, 2017. IEEE Press, 2017. pp. 4940–4949. doi 10.1109/ICCV.2017.528.
17. *Cheng H., Xu S., Jiang X., Wang R.* Deep image matting with flexible guidance input // Paper presented at 32nd British Machine Vision Conference, BMVC 2021, Virtual, Online. doi 10.5244/C.35.198.
18. *Kirillov A., Mintun E., Ravi N., et al.* Segment anything // 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, October 01–06, 2023. IEEE Press, 2023. pp. 3992–4003. doi 10.1109/ICCV51070.2023.00371.
19. *Kamann C., Rother C.* Benchmarking the robustness of semantic segmentation models // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, June 13–19, 2020. IEEE Press, 2020. pp. 8825–8835. doi 10.1109/CVPR42600.2020.00885.
20. *Воробьев Е.А., Намиот Д.Е.* Состязательное тестирование моделей сегментации изображений // Программная инженерия. 2025. **16**, № 4. 190–198. doi 10.17587/prin.16.190-198.
21. *Lokshin N.D., Khvostikov A.V., Krylov A.S.* Augmenting the training set of histological images with adversarial examples // Programming and Computer Software. 2023. **49**, N 3. 187–191. doi 10.1134/S0361768823030027.
22. *Guan Z., Hu M., Zhou Z., et al.* BadSAM: Exploring security vulnerabilities of SAM via backdoor attacks (student abstract) // Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI-24), Vancouver, Canada, February 20–27, 2024. Vol. 38, Iss. 21. Washington: AAAI Press, 2024. pp. 23506–23507. doi 10.1609/aaai.v38i21.30448.
23. *Zhang C., Zhang C., Kang T., et al.* Attack-SAM: Towards attacking segment anything model with adversarial examples // arXiv preprint arXiv:2305.00866. 2023. doi 10.48550/arXiv.2305.00866.
24. *Zhou Z., Song Y., Li M., et al.* DarkSAM: Fooling segment anything model to segment nothing // Advances in Neural Information Processing Systems. 2024. Vol. 37. pp. 49859–49880. doi 10.52202/079017-1579.
25. *Qiao Y., Hong C.S., Kang T., et al.* Robustness of SAM: Segment anything under corruptions and beyond // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2026. **48**, N 10. 12082-12099. doi 10.1109/TPAMI.2026.3698837.
26. *Wang Y., Zhao Y., Petzold L.* An empirical study on the robustness of the Segment Anything Model (SAM) // Pattern Recognition. 2024. **155**, N 12. Article Number 110685. doi 10.1016/j.patcog.2024.110685.
27. *Brendel W., Rauber J., Bethge M.* Decision-based adversarial attacks: Reliable attacks against black-box machine learning models // 6th International Conference on Learning Representations, Vancouver, Canada, April 30–May 3, 2018. https://openreview.net/group?id=ICLR.cc/2018/Conference.
28. *Bhagoji A.N., He W., Li B., Song D.* Practical black-box attacks on deep neural networks using efficient query mechanisms // Proceedings of the European Conference on Computer Vision (ECCV 2018). Lecture Notes in Computer Science, Vol. 11216. Cham: Springer, 2018. pp. 158–174. doi 10.1007/978-3-030-01258-8_10.
29. *Su J., Vargas D.V., Sakurai K.* One pixel attack for fooling deep neural networks // IEEE Transactions on Evolutionary Computation. 2019. **23**, N 5. 828–841. doi 10.1109/TEVC.2019.2890858.
30. *Chen X., Zhao Z., Yu F., et al.* Conditional diffusion for interactive segmentation // 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, October 10–17, 2021. IEEE Press, 2021. pp. 7325–7334. doi 10.1109/ICCV48922.2021.00725.
31. *Liu Q., Xu Z., Bertasius G., Niethammer M.* SimpleClick: Interactive image segmentation with simple vision transformers // Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, October 01–06, 2023. IEEE Press, 2023. pp. 22233–22243. doi 10.1109/ICCV51070.2023.02037.

32. *Sun S., Xian M., Xu F., et al.* CFR-ICL: Cascade-forward refinement with iterative click loss for interactive image segmentation // Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI-24), Vancouver, Canada, February 20–27, 2024. Vol. 38, Iss. 5. Washington: AAAI Press, 2024. pp. 5017–5024. doi 10.1609/aaai.v38i5.28306.
33. *Zhou M., Wang H., Zhao Q., et al.* Interactive segmentation as Gaussian process classification // 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, June 17–24, 2023. IEEE Press, 2023. pp. 19488–19497. doi 10.1109/CVPR52729.2023.01867.
34. *Ke L., Ye M., Danelljan M., et al.* Segment anything in high quality // Advances in Neural Information Processing Systems. 2023. **36**. 29914–29934. doi 10.52202/075280-1303.
35. *Zhang C., Han D., Qiao Y., et al.* Faster Segment Anything: Towards lightweight SAM for mobile applications // arXiv preprint arXiv:2306.14289. 2023. doi 10.48550/arXiv.2306.14289.
36. *Ma X., Zhou Y., Xu X., et al.* Towards layer-wise image vectorization // 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, June 18–24, 2022. IEEE Press, 2022. pp. 16293–16302. doi 10.1109/CVPR52688.2022.01583.
37. *Kingma D.P., Ba J.* Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. 2014. doi 10.48550/arXiv.1412.6980.
38. *Dice L.R.* Measures of the amount of ecologic association between species // Ecology. 1945. **26**, N 3. 297–302. doi 10.2307/1932409.

Получена *2 июля 2026 г.*
Принята *28 сентября 2026 г.*
Опубликована *17 ноября 2026 г.*

Информация об авторах

Андрей Викторович Москаленко — аспирант, науч. сотр.; 1) Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, 1, стр. 52, 119991, Москва, Российская Федерация; 2) Институт искусственного интеллекта (AIRI), Пресненская набережная, д. 6, стр. 2, 123112, Москва, Российская Федерация.

Владислав Игоревич Шахуро — к.ф.-м.н., ст. науч. сотр.; 1) Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, 1, стр. 52, 119991, Москва, Российская Федерация; 2) Институт искусственного интеллекта (AIRI), Пресненская набережная, д. 6, стр. 2, 123112, Москва, Российская Федерация.

Дмитрий Сергеевич Ватолин — к.ф.-м.н., заведующий лабораторией; 1) Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, 1, стр. 52, 119991, Москва, Российская Федерация; 2) Институт искусственного интеллекта МГУ имени М. В. Ломоносова, Ломоносовский проспект, д. 27, корп. 1, 119192, Москва, Российская Федерация.

References

1. C. Rother, V. Kolmogorov, and A. Blake, “GrabCut — Interactive Foreground Extraction Using Iterated Graph Cuts,” ACM Transactions on Graphics (TOG) **23** (3), 309–314 (2004). doi 10.1145/1186562.1015720.
2. D. Martin, C. Fowlkes, D. Tal, and J. Malik, “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, Vancouver, BC, Canada, July 07–14, 2001* (IEEE Press, New York, 2001), pp. 416–423. doi 10.1109/ICCV.2001.937655.
3. K. McGuinness and N. E. O’Connor, “A Comparative Evaluation of Interactive Segmentation Algorithms,” Pattern Recognition **43** (2), 434–444 (2010). doi 10.1016/j.patcog.2009.03.008.
4. N. Xu, B. Price, S. Cohen, et al., “Deep Interactive Object Selection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, Nevada, USA, June 26–July 1, 2016* (IEEE Press, New York, 2016), pp. 373–381. doi 10.1109/CVPR.2016.47.
5. M. Everingham, L. Van Gool, C. K. I. Williams, et al., *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results* (2012).

6. T.-Y. Lin, M. Maire, S. Belongie, et al., "Microsoft COCO: Common Objects in Context," in *Proceedings 13th European Conference. Computer Vision — ECCV 2014, Zurich, Switzerland, September 6–12, 2014*. ECCV 2014. Lecture Notes in Computer Science, Vol. 8693. (Springer, Cham, 2014), pp. 740–755. doi 10.1007/978-3-319-10602-1_48.
7. Z. Li, Q. Chen, and V. Koltun, "Interactive Image Segmentation with Latent Diversity," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18–23, 2018* (IEEE Press, 2018), pp. 577–585. doi 10.1109/CVPR.2018.00067.
8. F. Perazzi, J. Pont-Tuset, B. McWilliams, et al., "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 27–30, 2016* (IEEE Press, 2016), pp. 724–732. doi 10.1109/CVPR.2016.85.
9. B. Hariharan, P. Arbelaez, L. Bourdev, et al., "Semantic Contours from Inverse Detectors," in *2011 International Conference on Computer Vision, Barcelona, Spain, November 06–13, 2011* (IEEE Press, 2011), pp. 991–998. doi 10.1109/ICCV.2011.6126343.
10. W.-D. Jang and C.-S. Kim, "Interactive Image Segmentation via Backpropagating Refinement Scheme," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, June 15–20, 2019* (IEEE Press, 2019), pp. 5292-5301. doi 10.1109/CVPR.2019.00544.
11. K. Sofiiuk, I. Petrov, O. Barinova, and A. Konushin, "f-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, June 13–19, 2020* (IEEE Press, 2020), pp. 8620–8629. doi 10.1109/CVPR42600.2020.00865.
12. K. Sofiiuk, I. A. Petrov, and A. Konushin, "Reviving Iterative Training with Mask Guidance for Interactive Segmentation," in *2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, October 16–19, 2022* (IEEE Press, 2022), pp. 3141–3145. doi 10.1109/ICIP46576.2022.9897365.
13. B. Cheng, R. Girshick, P. Dollár, et al., "Boundary IoU: Improving Object-Centric Image Segmentation Evaluation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, June 20–25, 2021* (IEEE Press, 2021), pp. 15329–15337. doi 10.1109/CVPR46437.2021.01508.
14. H.-E. Gueziri, M. McGuffin, and C. Laporte, "Latency Management in Scribble-Based Interactive Segmentation of Medical Images," IEEE Transactions on Biomedical Engineering **65** (5), 1140–1150 (2018). doi 10.1109/TBME.2017.2777742.
15. E. Agustsson, J. R. R. Uijlings, and V. Ferrari, "Interactive Full Image Segmentation by Considering All Regions Jointly," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, June 15–20, 2019* (IEEE Press, 2019), pp. 11614–11623. doi 10.1109/CVPR.2019.01189.
16. D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari, "Extreme Clicking for Efficient Object Annotation," in *2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 22–29, 2017* (IEEE Press, 2017), pp. 4940–4949. doi 10.1109/ICCV.2017.528.
17. H. Cheng, S. Xu, X. Jiang, and R. Wang, "Deep Image Matting with Flexible Guidance Input," in *Paper presented at 32nd British Machine Vision Conference, BMVC 2021, Virtual, Online* doi 10.5244/C.35.198.
18. A. Kirillov, E. Mintun, N. Ravi, et al., "Segment Anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, October 01–06, 2023* (IEEE Press, 2023), pp. 3992–4003. doi 10.1109/ICCV51070.2023.00371.
19. C. Kamann and C. Rother, "Benchmarking the Robustness of Semantic Segmentation Models," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, June 13–19, 2020* (IEEE Press, 2020), pp. 8825–8835. doi 10.1109/CVPR42600.2020.00885.
20. E. A. Vorob'ev and D. E. Namiot, "Adversarial Testing of Image Segmentation Models," Programmnaya Inzheneriya **16** (4), 190–198 (2025). [in Russian]. doi 10.17587/prin.16.190-198.
21. N. D. Lokshin, A. V. Khvostikov, and A. S. Krylov, "Augmenting the Training Set of Histological Images with Adversarial Examples," Programming and Computer Software **49** (3), 187–191 (2023). doi 10.1134/S0361768823030027.
22. Z. Guan, M. Hu, Z. Zhou, et al., "BadSAM: Exploring Security Vulnerabilities of SAM via Backdoor Attacks (Student Abstract)," in *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI-24), Vancouver, Canada, February 20–27, 2024* Vol. 38, Iss. 21. (AAAI Press, Washington, DC, USA, 2024), pp. 23506–23507. doi 10.1609/aaai.v38i21.30448.
23. C. Zhang, C. Zhang, T. Kang, et al., "Attack-SAM: Towards Attacking Segment Anything Model with Adversarial Examples," arXiv preprint arXiv:2305.00866 (2023). doi 10.48550/arXiv.2305.00866.
24. Z. Zhou, Y. Song, M. Li, et al., "DarkSAM: Fooling Segment Anything Model to Segment Nothing," Advances in Neural Information Processing Systems **37**, 49859–49880 (2024). doi 10.52202/079017-1579.
25. Y. Qiao, C. S. Hong, T. Kang, et al., "Robustness of SAM: Segment Anything under Corruptions and Beyond," IEEE Transactions on Pattern Analysis and Machine Intelligence **48** (10), 12082–12099 (2026). doi 10.1109/TPAMI.2026.3698837.

26. Y. Wang, Y. Zhao, and L. Petzold, "An Empirical Study on the Robustness of the Segment Anything Model (SAM)," Pattern Recognition **155** (12), Article Number 110685 (2024). doi 10.1016/j.patcog.2024.110685.
27. W. Brendel, J. Rauber, and M. Bethge, "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models," in *6th International Conference on Learning Representations, Vancouver, Canada, April 30–May 3, 2018* https://openreview.net/group?id=ICLR.cc/2018/Conference.
28. A. N. Bhagoji, W. He, B. Li, and D. Song, "Practical Black-Box Attacks on Deep Neural Networks Using Efficient Query Mechanisms," in *Proceedings of the European Conference on Computer Vision (ECCV 2018)* Lecture Notes in Computer Science, Vol. 11216. (Springer, Cham, 2018), pp. 158–174. doi 10.1007/978-3-030-01258-8_10.
29. J. Su, D. V. Vargas, and K. Sakurai, "One Pixel Attack for Fooling Deep Neural Networks," IEEE Transactions on Evolutionary Computation **23** (5), 828–841 (2019). doi 10.1109/TEVC.2019.2890858.
30. X. Chen, Z. Zhao, F. Yu, et al., "Conditional Diffusion for Interactive Segmentation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, October 10–17, 2021* (IEEE Press, 2021), pp. 7325–7334. doi 10.1109/ICCV48922.2021.00725.
31. Q. Liu, Z. Xu, G. Bertasius, and M. Niethammer, "SimpleClick: Interactive Image Segmentation with Simple Vision Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, October 01–06, 2023* (IEEE Press, 2023), pp. 22233–22243. doi 10.1109/ICCV51070.2023.02037.
32. S. Sun, M. Xian, F. Xu, et al., "CFR-ICL: Cascade-Forward Refinement with Iterative Click Loss for Interactive Image Segmentation," in *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI-24), Vancouver, Canada, February 20–27, 2024* Vol. 38, Iss. 5. (AAAI Press, Washington, DC, USA, 2024), pp. 5017–5024. doi 10.1609/aaai.v38i5.28306.
33. M. Zhou, H. Wang, Q. Zhao, et al., "Interactive Segmentation as Gaussian Process Classification," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, June 17–24, 2023* (IEEE Press, 2023), pp. 19488–19497. doi 10.1109/CVPR52729.2023.01867.
34. L. Ke, M. Ye, M. Danelljan, et al., "Segment Anything in High Quality," Advances in Neural Information Processing Systems **36**. 29914–29934 (2023). doi 10.52202/075280-1303.
35. C. Zhang, D. Han, Y. Qiao, et al., "Faster Segment Anything: Towards Lightweight SAM for Mobile Applications," arXiv preprint arXiv:2306.14289 (2023). doi 10.48550/arXiv.2306.14289.
36. X. Ma, Y. Zhou, X. Xu, et al., "Towards Layer-Wise Image Vectorization," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, June 18–24, 2022* (IEEE Press, 2022), pp. 16293–16302. doi 10.1109/CVPR52688.2022.01583.
37. D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980 (2014). doi 10.48550/arXiv.1412.6980.
38. L. R. Dice, "Measures of the Amount of Ecologic Association between Species," Ecology **26** (3), 297–302 (1945). doi 10.2307/1932409.

Received *July 2, 2026*
Accepted *September 28, 2026*
Published *November 17, 2026*

Information about the authors

Andrey V. Moskalenko — PhD Student, Research Scientist; 1) Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Leninskie Gory, 1, building 52, 119991, Moscow, Russia; 2) Artificial Intelligence Research Institute (AIRI), Presnenskaya Naberezhnaya, 6, building 2, 123112, Moscow, Russia.

Vladislav I. Shakhuro — PhD., Senior Research Scientist; 1) Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Leninskie Gory, 1, building 52, 119991, Moscow, Russia; 2) Artificial Intelligence Research Institute (AIRI), Presnenskaya Naberezhnaya, 6, building 2, 123112, Moscow, Russia.

Dmitriy S. Vatolin — PhD., Head of Laboratory; 1) Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Leninskie Gory, 1, building 52, 119991, Moscow, Russia; 2) Institute of Artificial Intelligence of Lomonosov Moscow State University, Lomonosovsky Prospekt, 27, building 1, 119192, Moscow, Russia.