Нейросетевые методы сегментации изображений высокого разрешения

В.А. Офицеров, А.С. Конушин

***Аннотация*— В данной работе исследуется проблема интерактивной сегментации изображений, актуальная для современных приложений компьютерного зрения. Цель работы – повышение разрешения модели интерактивной сегментации в условиях ограниченных ресурсов. В работе проведен обзор существующих методов сегментации, предложен и усовершенствован базовый метод, что позволило улучшить показатели NoC N @ 90 bIoU с 16.97 до 12.25 на наборе данных HQSeg44k. Результаты демонстрируют, что новый метод улучшает разрешение сегментационных карт и повышает точность выделения объектов при ограниченных вычислительных мощностях, что подтверждает его потенциал для применения в различных областях, требующих точной сегментации изображений с минимальными ресурсами.**

***Ключевые слова*— Интерактивная сегментаций изображений, высокое разрешение, ограниченные ресурсы.**

I. Введение

В современном мире, где технологии компьютерного зрения играют всё более важную роль в различных сферах, интерактивная сегментация изображений обретает особую актуальность. Эта задача является важной для множества приложений, от исследований в области медицины до продвинутых систем искусственного интеллекта, используемых в автоматизированной обработке медиаконтента.

Интерактивная сегментация, позволяющая пользователям в реальном времени управлять процессом выделения объектов на изображении, становится ключевым элементом в улучшении качества и точности обработки данных. Задача сегментации встречается практически в любых проектах, связанных с компьютерным зрением. При этом стоимость и сложность разметки этой задачи очень высоки. На рис. 1 можно заметить разницу между разметкой без и с помощью предобученных методов интерактивной сегментации: при полностью ручной разметке без какой-либо помощи необходимо поставить большое количество точек по контуру изображения, и при этом маска сегментации может выглядеть неаккуратно. В то же время, используя систему интерактивной сегментации, предсказывающую маску сегментации, можно всего лишь за один или пару кликов добиться схожего по точности результата.

Рис. 1: Разметка изображений для задачи сегментации. Без(слева) и с(справа) системой интерактивной сегментации автоматически предсказывающей маску сегментации.

Актуальность исследования в данной области обусловлена быстрым развитием технологий и возрастающими требованиями к точности обработки визуальных данных. С каждым годом возрастает потребность в системах, способных эффективно и всё более точно обрабатывать большие объёмы изображений, при этом обеспечивая высокую степень взаимодействия с пользователем.

Несмотря на стремительное развитие, современные методы всё ещё не умеют идеально размечать любые изображения. Например, на рис. 2 можно увидеть проблему современных методов с пониманием мелких деталей и невозможностью предсказывать маски сегментации высокого разрешения. Стоит также отметить, что большинство лучших методов требуют использования огромного количества вычислительных ресурсов, доступных далеко не всем.

Особое внимание следует уделить методам, которые можно повышают разрешение сегментации в условиях ограниченных ресурсов, чтобы сделать их максимально доступными для любых подобных задач. Одним из основных применений интерактивной сегментации является редактирование фотографий пользователями и работа с их персональным контентом.

Статья получена 29 мая 2024.

Работа представляет собой результат магистерской диссертации.

Владислав Алексеевич Офицеров, Московский Государственный Университет им. М.В. Ломоносова (e-mail: ofitserovlad@yandex.ru).

Антон Сергеевич Конушин, Московский Государственный Университет им. М.В. Ломоносова (e-mail: ktosh@graphics.cs.msu.ru).

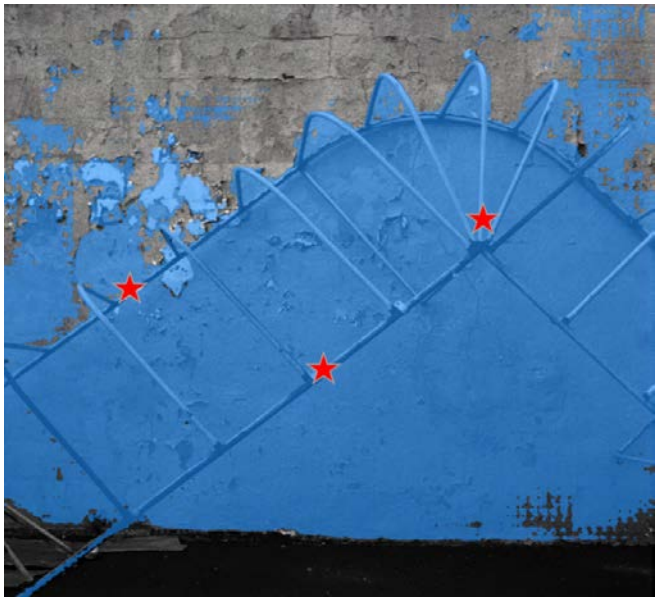

Рис. 2: Проблема алгоритмов интерактивной сегментации с выделением мелких объектов.

Важно, чтобы алгоритмы могли работать непосредственно на конечном устройстве, таком как телефон, планшет или ноутбук, без использования мощных GPU в облаке и без существенного расхода батареи. Поэтому прямой подход в наращивании нейросетевых моделей путём добавления тяжелых слоёв высокого разрешения не подходит.

Данное исследование направлено на изучение существующих подходов в области интерактивной сегментации и разработку нового метода, способного улучшить разрешение карты сегментации. Это будет способствовать не только повышению точности в обработке изображений, но и потенциально может открыть новые перспективы для их использования в самых разных областях деятельности.

II. Обзор литературы и существующих решений

С развитием нейронных сетей методы анализа изображений пережили значительный прогресс. Нейросетевые подходы предоставляют новый уровень точности и эффективности, что позволило существенно улучшить результаты сегментации изображений в различных областях, включая компьютерное зрение. Эти методы способны решать сложные задачи распознавания и обработки визуальной информации, превосходя традиционные методы по многим параметрам. В связи с этим, в данном обзоре основное внимание уделяется нейросетевым методам сегментации изображений, которые в настоящее время демонстрируют наиболее впечатляющие результаты и являются ведущими в этой области.

A. Методы сегментации

К настоящему времени методы сегментации изображений на основе глубокого обучения достигли значительных успехов. Современные лучшие методы (SOTA) обычно основаны на архитектурах свёрточных нейронных сетей (CNN) или на трансформерах, которые были адаптированы для задач компьютерного зрения.

Сеть U-Net [1], первоначально разработанная для медицинских изображений, имеет U-образную архитектуру. Сначала следует череда сверточных слоев, понижающих разрешение изображения, затем – соответствующие слои, повышающие разрешение. Дополнительно блоки одинакового размера соединены специальными пропущенными связями (англ. skip connections). Эта симметричная структура позволяет восстанавливать пространственную информацию, потерянную при понижении разрешения, и достигать высокой точности сегментации.

DeepLab (v3 и v3+) [2], [3] архитектуры используют артифициально атриумные свёрточные сети (ASPP) для улавливания контекста на разных масштабах. DeepLabv3 улучшает представление за счет использования различных размеров рецептивных полей, в то время как DeepLabv3+ добавляет модуль декодирования (англ. decoder) для более точного восстановления пространственной информации, что делает её одной из ведущих архитектур для задач сегментации.

Mask R-CNN - расширение [4] Faster R-CNN [5], которая, в свою очередь, является улучшенной версией R-CNN [6] и Fast R-CNN [7]. Faster R-CNN интегрирует региональное предложение (Region Proposal Network, RPN) непосредственно в процесс обучения, что позволяет существенно ускорить детекцию объектов. Mask R-CNN добавляет к этой архитектуре ветвь для сегментации объектов на уровне масок, что позволяет одновременно выполнять детекцию объектов и их сегментацию. Эта сеть эффективно решает задачу семантической сегментации, обеспечивая точное выделение объектов на изображении.

Transformer-базированные архитектуры являются продолжением работы Vision Transformer (ViT) [8], используют механизмы самовнимания для моделирования долгосрочных зависимостей в изображениях.

ViT представляет изображение в виде последовательности патчей (обычно 16x16 пикселей), которые затем линейно преобразуются в векторы. Эти векторы подаются на вход трансформера, где используется многоголовочное внимание для обработки информации. В отличие от CNN, ViT не использует свёртки, что позволяет ему захватывать более глобальные контексты. Однако, ViT требует большого количества данных для обучения и обладает высокой вычислительной сложностью.

Наиболее современным методом здесь является Mask2Former [9], который решает проблему универсальности и точности сегментации, используя единый архитектурный подход для различных задач сегментации. Mask2Former объединяет преимущества трансформеров и традиционных методов сегментации, применяя механизм внимания к объектам и областям интереса. Это позволяет модели адаптироваться к

различным уровням детализации и контекстам, улучшая результаты сегментации. Mask2Former также использует иерархическую структуру для обработки информации на разных масштабах, что способствует некоторому снижению вычислительной сложности при сохранении высокой точности.

HRNet архитектура [10] поддерживает высокое разрешение на протяжении всей сети, что позволяет лучше захватывать пространственные детали. HRNet соединяет многомасштабные представления параллельно, что способствует сохранению точности и позволяет более точно выполнять задачи сегментации, особенно для изображений с высоким разрешением и деталями. При этом HRNet требует значительно меньше вычислительных ресурсов по сравнению с другими высокоэффективными моделями, сохраняя при этом высокое качество сегментации, что делает её эффективной и производительной для различных приложений.

B. Методы интерактивной сегментации

Интерактивная сегментация изображений представляет собой динамично развивающуюся область, в которой сочетаются методы машинного обучения с возможностью пользовательского взаимодействия для уточнения результатов сегментации. Основные методы в этой области включают:

RITM (Reviving Iterative Training with Mask Guidance for Interactive Segmentation) [11] использует архитектуру HRNet, обеспечивающей высокое разрешение на всех уровнях сети, что критически важно для точной сегментации. Для обработки пользовательских взаимодействий, таких как клики, метод применяет круговые диски вокруг точек клика (см. рис. 3 (справа)), которые трансформируются в Conv1S блоки (см. рис. 3 (слева)) сверток для модификации карты признаков, тем самым направляя процесс сегментации более целенаправленно.

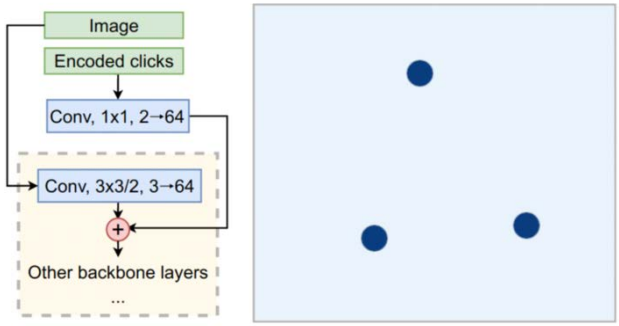

Рис. 3: Архитектура блока Conv1S и способы кодирования кликов пользователя дисками в RITM.

FocusCut [12] реализует усовершенствованный метод сегментации, основанный на механизмах внимания. Архитектура FocusCut включает в себя модуль внимания, который позволяет выделять и обрабатывать области интереса более детально. Это достигается с помощью Transformer-подобных механизмов, которые помогают сети фокусироваться на важных частях изображения, улучшая качество сегментации за счет увеличенной детализации в этих областях.

Архитектура SimpleClick [13] метода интерактивной сегментации состоит из трех частей:

Модуль извлечения признаков изображения ViT, обрабатывающий карты признаков одного размера.

Пирамида признаков с упрощенной архитектурой. Она содержит в себе 4 свёрточных слоя с разными шагами перемещения фильтра.

Легковесная подсеть с полносвязными слоями (MLP) для обработки карт признаков с каждого свёрточного слоя и увеличения их к одному размеру для последующей конкатенации. Далее полученный тензор переводится в формат одноканальной карты признаков, чтобы получить предсказанную сегментационную маску.

Данная модель обучалась с автоматической симуляцией кликов на основе текущих результатов сегментации и целевой маски-образца.

C. Методы адаптации

LoRA (Low-Rank Adaptation) [14] - метод, первоначально возникший для дообучения больших языковых моделей, позволяющий снизить объем ресурсов, необходимых для дообучения, путем замораживания основных параметров модели и добавления небольшого числа новых обучаемых параметров. Добавленные параметры для каждого слоя представляют собой две низкоранговые матрицы, произведение которых суммируется с основными параметрами слоя.

Преимущество данного метода в том, что ранг двух матриц является гиперпараметром и, таким образом, его можно выбрать так, чтобы данный адаптер практически не менял время работы метода, а также мог быть применим к любому типу архитектур, что является очень полезным моментом для данной работы.

В методе TOAST [15] предлагается “перенаправить” внимание нейронной сети, обученной под исходную задачу, на признаки, релевантные для новой задачи. Авторы берут исходную сеть, добавляют к ней в конец новый модуль, цель которого выявить, на что итоговой модели лучше “смотреть”, и используют эту информацию в слоях внимания исходной сети, прогоняя через нее изображение еще раз. Обучаемым при этом является только добавленный модуль, слои основной сети замораживаются.

На основе проведенного обзора методов сегментации изображений, можно сделать несколько ключевых выводов. Современные подходы к сегментации изображений демонстрируют значительные достижения благодаря применению глубоких нейронных сетей, в частности архитектур, основанных на сверточных нейронных сетях и трансформерах. Среди множества рассмотренных методов, архитектуры HRNet и RITM

выделяются как наиболее качественные и эффективные для задач сегментации изображений с высоким разрешением, особенно в условиях ограниченных вычислительных ресурсов.

HRNet, благодаря своей способности сохранять высокое разрешение на всех уровнях сети и обеспечивать параллельную обработку многомасштабных представлений, позволяет достигать высокой точности сегментации, сохраняя при этом эффективность использования ресурсов. Также хочется отметить, что для HRNet может быть применим адаптер LoRa, в отличии от TOAST, который применим только к трансформерным архитектурам.

Метод интерактивной сегментации RITM, который как раз и основан на архитектуре HRNet, также показывает высокие результаты благодаря использованию высокоразрешающей сети и эффективному взаимодействию с пользователем.

Исходя из вышеперечисленных преимуществ, методы HRNet и RITM являются наиболее подходящими для использования в дальнейшем в качестве базовых методов сегментации изображений с высоким разрешением. Они не только обеспечивают высокое качество сегментации, но и эффективны, особенно совместно с использованием с адаптерами, в условиях ограниченных вычислительных ресурсов, что делает их идеальными для широкого спектра приложений в области компьютерного зрения.

III. Исследование и решение задачи

Современные методы интерактивной сегментации изображений демонстрируют значительные успехи, однако в условиях ограниченных мощностей их применение или обучение зачастую становится проблематичным. В данной работе предлагается новый метод, основанный на RITM (Reviving Iterative Training with Mask Guidance for Interactive Segmentation) с использованием HRNet+OCR в качестве основного архитектурного блока, так как именно они по результатам обзора стали наиболее подходящими для поставленной задачи. Основное внимание уделяется разработке и внедрению легковесных адаптеров для улучшения качества сегментации без значительного увеличения вычислительных затрат. В числе предложенных улучшений — адаптеры на основе сверточных слоев или LoRa (которые как раз и позволяют не обучать/дообучать исходную сеть), а также использование специализированной функции потерь для улучшения сегментации мелких деталей и высокоразрешающих изображений.

Метод RITM с HRNet+OCR является эффективным решением для задач интерактивной сегментации изображений. HRNet (High-Resolution Network) поддерживает высокое разрешение на всех уровнях сети и обеспечивает параллельную обработку многомасштабных представлений, что способствует высокой точности сегментации. OCR (Object Contextual Representations) позволяет улучшить представление объектов в изображении за счет контекстуальной информации.

Важным аспектом метода RITM является использование функции потерь Normalized Focal Loss (NFL), разработанной для преодоления недостатков стандартной функции потерь binary cross entropy (BCE). BCE обрабатывает все примеры одинаково, что замедляет обучение на поздних стадиях, так как градиент от почти правильно сегментированных областей распространяется так же, как и от ошибочных регионов. Focal Loss (FL) [16] решает эту проблему, уменьшая вес правильно предсказанных областей. Normalized Focal Loss (NFL) нормализует градиент, что позволяет ускорить процесс обучения и улучшить точность.

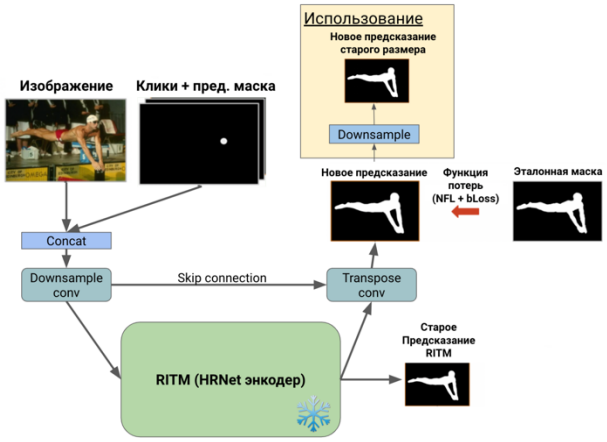

Рис. 4: Архитектура предложенного метода со сверточными слоями и дополнительной функцией потерь.

A. *Легковесные адаптеры*

Одним из усовершенствований предлагается внедрение LoRa (Low-Rank Adaptation), который представляет собой легковесный адаптер, применимый в том числе к архитектурам конволюционных нейронных сетей. Например, как это было сделано в ConvLoRa [17]. LoRa позволяет снизить количество параметров модели, сохраняя при этом высокую точность для конкретной задачи. Этот метод особенно полезен в условиях нашей задачи, так как адаптеры могут быть дообучены с минимальными затратами. В дополнение ко всему, с данным адаптером можно использовать произвольную функцию потерь, что позволяет выбирать её в зависимости от конкретной задачи.

Также предлагается внедрение адаптеров на основе сверточных слоев, чем-то похожим на используемые в архитектурах семейства UNet. Эти адаптеры включают конволюционные слои с шагом большим единицы (англ. Downsample Conv) в начале и обратные конволюционные слои (англ. Transposed Conv) в конце сети, что позволяет обрабатывать разрешения большего разрешения и также получить дополнительный выход с увеличенным разрешением, наряду с основным предсказанием. Данное улучшение мотивировано тем, что предыдущую сеть можно использовать практически

без изменений, добавляя лишь небольшой адаптер для улучшения качества сегментации.

Также дополнительно предлагается введение слоев пропусков (англ. skip connections), которые помогают сохранять важную информацию на всех уровнях сети и хорошо себя показывают в подобных архитектурах.

Помимо этого, добавление слоя конволюций в самом начале позволяет нам изменять входные данные. Предлагается добавление фильтра Собеля в качестве отдельного слоя к исходному изображению, что позволяет улучшить выделение краев объектов и, следовательно, потенциально повысить точность сегментации.

B. Функция потерь

Для улучшения сегментации мелких деталей и высокоразрешающих изображений предлагается использование Boundary Loss [18] в дополнение к NFL. Boundary Loss (BL) был специально разработан для улучшения предсказания границ объектов, что особенно важно для высокоразрешающих выходов сети.

Данную функцию потерь предлагается использовать как компоненту итоговой комбинированной функции потерь.

В данном разделе предложены способы усовершенствования интерактивной сегментации, включающие внедрение легковесных адаптеров и использование специализированной функции потерь. Эти улучшения направлены на повышение качества сегментации при минимальных вычислительных затратах, что делает их очень подходящими для приложения в реальном мире.

IV. Практическая часть, эксперимент

В этом разделе перейдем к описанию наборов данных и метрик для возможности тестирования предложенных усовершенствований из прошлой главы. Далее опишем сделанные эксперименты и их программную реализацию.

A. Наборы данных

Рассматривая возможность получения маски изображений более высокого разрешения очень важно правильно работать с данными. Большинство наборов данных, на которых тестируются все методы сегментации, имеют достаточно грубую оценку на многих изображениях. Тем не менее, использование этих датасетов (Таблица 1) также очень важно для понимания обобщенной способности методов, которые получатся в результаты экспериментов. Далее будем называть объединение всех этих наборов данных - "Регулярный тест".

Таблица 1 — Наборы данных, используемые для тестирования.

Название	Изображений	Объектов
GrabCut	50	50
Berkeley	96	100
DAVIS	345	345
SBD	2857	6671

В качестве более качественных данных хочется отметить набор данных для обучения COCO+LVIS*, который использовался в оригинальной статье RITM, и набор данных HQSeg44k [19], который является композицией 6 различных наборов данных с масками высокого разрешения: DIS [20] (обучение), ThinObject-5k [21] (обучение), FSS-1000 [22], ECSSD [23], MSRA-10K [24], DUT-OMRON [25].

Таблица 2 — Наборы более качественных данных, используемых для обучения и тестирования.

Название	Изображений	Объектов
COCO+LVIS*	99k	1.5M
HQSeg44k	44k	44k

B. Набор метрик

В данной работе использовались классическая метрика IoU и специально введенная метрика Boundary IoU [26], которая является более чувствительной к качеству границ сегментации, оценки качества моделей сегментации изображений. Также метрика NoC для интерактивной сегментации изображений. Метрика NoC используется для оценки эффективности интерактивной сегментации изображений. Она измеряет количество кликов, необходимых пользователю для достижения заданного уровня точности сегментации. Метрика NoC N @ 90 указывает на количество кликов, необходимых для достижения IoU = 90%.

C. Детали реализации

Все модели обучались на задачу бинарной сегментации с использованием двух функций потерь NFL и комбинированной функции потерь. Входные изображения подавали в виде кропов исходных изображений размером 320 × 480 в случае LoRa адаптеров и в два раза больших иначе. Везде использовали произвольное изменение размера изображений с коэффициентом масштабирования от 0,75 до 1,40 перед обрезкой. Во время тренировки используется горизонтальное смещение и случайное изменение значений яркости, контрастности и RGB в качестве аугментаций для уменьшения переобучения и лучшей сходимости. В качестве бэкбона использовался HRNet-18. Во время обучение LoRa исходная сеть была полностью неизменяема, а во время обучения конволюционных слоев в качестве адаптеров также дополнительно был разморожен первый слой сети. Коэффициент перед Boundery Loss равен 0.1. Во время тестирования использовали улучшенное временное тестирование (TTA) схожее с оригинальной статьей RITM. Во всех экспериментах использовался оптимизатор AdamW с weight decay = 1e-5, learning rate подбирался для каждой задачи отдельно, в экспериментах ниже представлены лучшие значение, с косинусной стратегией изменения значения на 55-100 эпохах. Во всех экспериментах использовалась ранняя остановка для сокращения времени, затраченного на обучение. Все эксперименты проходили на GPU NVIDIA GeForce GTX 1080 и Tesla V100.

D. Эксперименты

Для начала проведем эксперименты с LoRa адаптерами. Во всех таблицах будет использоваться метрика NoC @ 90 для IoU или bIoU. В таблице 3 представлены следующее эксперименты:

RITM: Повторение работы RiTM, обучение полностью всей сети с нуля в соответствии с оригинальной статьей.

RITM + LoRa: Дообучение LoRa адаптера сверх уже преодобученной ранее базовой модели, которая никак не дообучалась на текущем этапе.

RITM + LoRa + HQ-Seg44K: дообучение адаптеров будет происходить на обучающей части набора данных HQSeg-44k

RITM + LoRa + HQ-Seg44K + Boundary Loss: Аналогичное дообучение LoRa адаптера сверх уже преодобученной ранее базовой модели и измененная функция потерь.

Таблица 3 — Эксперименты с LoRa.

Метод	Регулярный тест		HQSeg-44k	
	IoU	bIoU	IoU	bIoU
RITM	5.83	6.17	8.82	16.97
RITM + LoRa	5.82	6.17	8.78	17.12
RITM + LoRa + HQ-Seg44K	5.94	6.25	8.23	13.67
RITM + LoRa + HQ-Seg44K + Boundary Loss	5.98	6.02	8.17	13.44

Теперь проведем эксперименты с адаптерами в виде конволюционных слоев. В таблице 4 представлены следующее эксперименты:

RITM: Аналогичное экспериментам с LoRa базовый метод повторения работы RITM.

RITM + Down/Transpose conv: Дообучение только добавленных адаптеров, состоящих из конволюционных слоев. Базовая сеть не дообучается.

RITM + Down/Transpose conv + HQ-Seg44K: Дополнительное дообучение только адаптеров с использованием обучающей части набора данных HQ-Seg44K. Базовая сеть не дообучается.

RITM + Down/Transpose conv + HQ-Seg44K + Skip: Добавление к прошлому пункту еще пропуск связи (англ. skip connection)

Таблица 4 — Эксперименты с конволюционными адаптерами.

Метод	Регулярный тест		HQSeg-44k	
	IoU	bIoU	IoU	bIoU
RITM	5.83	6.17	8.82	16.97
RITM + conv	5.857	5.98	9.41	15.81
RITM + conv + HQ-Seg44K	5.84	5.98	8.01	12.9
RITM + conv + HQ-Seg44K + Skip	5.83	5.97	7.92	12.73
RITM + conv + HQ-Seg44K + Skip + Boundary Loss + Собель	5.828	5.95	7.68	12.32
RITM + conv + HQ-Seg44K + Skip + Boundary Loss	5.825	5.9	7.65	12.25

RITM + Down/Transpose conv + HQ-Seg44K + Skip + Boundary Loss + фильтр Собеля: сверх изменений прошлого пункта изменяется функция потерь в соответствии с ранее описанной схемой и дополнение к архитектуре из прошлого пункта конкатенации к изображению на входе сети ее фильтра Собеля

RITM + Down/Transpose conv + HQ-Seg44K + Skip + Boundary Loss: Отличие от прошлого метода в отсутствие конкатенации к изображению на входе сети ее фильтра Собеля.

Эксперименты показали, что предложенный метод добавления легковесных конволюционных адаптеров позволяет повысить точность NoC N @ 90 bIoU с 16.97 до 12.25, при этом обгоняя методы на основе LoRa, которые достигают качества 13.44.

E. Программная реализация

Вся разработка велась на языке программирования Python 3.8 внутри Docker контейнера с cuda 10.1 и cudnn 7. Для разработки моделей использовались фреймворки PyTorch и PyTorch Lightning, аугментации в ходе обучения реализовывались с помощью библиотек Albumentations и с помощью Numpy. Для логирования результатов экспериментов был выбран инструмент Tensorboard и WandB. Для работы с данными и проведения аналитики использовались библиотеки NumPy, Pandas, Matplotlib, OpenCV и Scikit-Learn. Разработка велась в IDE PyCharm Community Edition в операционной системе Ubuntu 22.04. Объем кода: 3500 строк.

V. Заключение

В данной работе разработан метод для улучшения разрешения интерактивной сегментации изображений в условиях ограниченных ресурсов.

Проведен обзор существующих методов сегментации. В рамках данной работы наиболее подходящим оказался метод RITM, использующий HRNet архитектуру для сегментации изображений.

Разработан усовершенствованный метод: предложены модификации базового метода по добавлению легковесных адаптеров (конволюционных слоев и LoRA) и изменению функции потерь для них.

Проведена экспериментальная оценка, в результате которой, усовершенствованный метод с использованием конволюционных слоев в качестве адаптера и измененной функцией потерь оказался наилучшим по

результатам экспериментальной оценки и улучшил NoC N @ 90 bIoU базового метода до 12.25.

Библиография

[1] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer International Publishing.

[2] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

[3] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

[4] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[5] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[6] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[7] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[9] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1290-1299).

[10] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 43(10), 3349-3364.

[11] Sofiiuk, K., Petrov, I. A., & Konushin, A. (2022, October). Reviving iterative training with mask guidance for interactive segmentation. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 3141-3145). IEEE.

[12] Lin, Z., Duan, Z. P., Zhang, Z., Guo, C. L., & Cheng, M. M. (2022). Focuscut: Diving into a focus view in interactive segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2637-2646).

[13] Liu, Q., Xu, Z., Bertasius, G., & Niethammer, M. (2023). Simpleclick: Interactive image segmentation with simple vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22290-22300).

[14] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[15] Shi, B., Gai, S., Darrell, T., & Wang, X. (2023). Toast: Transfer learning via attention steering. arXiv preprint arXiv:2305.15542, 5(7), 13.

[16] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[17] Aleem, S., Dietlmeier, J., Arazo, E., & Little, S. (2024). ConvLoRA and AdaBN based Domain Adaptation via Self-Training. arXiv preprint arXiv:2402.04964.

[18] Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., & Ayed, I. B. (2019, May). Boundary loss for highly unbalanced segmentation. In International conference on medical imaging with deep learning (pp. 285-296). PMLR.

[19] Ke, L., Ye, M., Danelljan, M., Tai, Y. W., Tang, C. K., & Yu, F. (2024). Segment anything in high quality. Advances in Neural Information Processing Systems, 36.

[20] Qin, X., Dai, H., Hu, X., Fan, D. P., Shao, L., & Van Gool, L. (2022, October). Highly accurate dichotomous image segmentation. In European Conference on Computer Vision (pp. 38-56). Cham: Springer Nature Switzerland.

[21] Liew, J. H., Cohen, S., Price, B., Mai, L., & Feng, J. (2021). Deep interactive thin object selection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 305-314).

[22] Li, X., Wei, T., Chen, Y. P., Tai, Y. W., & Tang, C. K. (2020). Fss-1000: A 1000-class dataset for few-shot segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2869-2878).

[23] Shi, J., Yan, Q., Xu, L., & Jia, J. (2015). Hierarchical image saliency detection on extended CSSD. IEEE transactions on pattern analysis and machine intelligence, 38(4), 717-729.

[24] Cheng, M. M., Mitra, N. J., Huang, X., Torr, P. H., & Hu, S. M. (2014). Global contrast based salient region detection. IEEE transactions on pattern analysis and machine intelligence, 37(3), 569-582.

[25] Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M. H. (2013). Saliency detection via graph-based manifold ranking. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3166-3173).

[26] Cheng, B., Girshick, R., Dollár, P., Berg, A. C., & Kirillov, A. (2021). Boundary IoU: Improving object-centric image segmentation evaluation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 15334-15342).

High resolution image segmentation with deep learning models

Vladislav Ofitserov, Anton Konushin

Abstract— The work addresses the issue of interactive image segmentation, relevant to modern computer vision applications. The aim of the work is to improve the resolution of interactive segmentation models under limited resources. The work provides a review of existing segmentation methods and proposes an enhanced basic method, which improved the NoC N @ 90 bIoU metric from 16.97 to 12.25 on the HQSeg44k dataset. The results demonstrate that the new method enhances segmentation map resolution and improves object delineation accuracy with limited computational resources, confirming its potential for applications in various fields requiring precise image segmentation with minimal resources.

Keywords— Interactive image segmentation, high resolution, limited resources.

REFERENCES

[1] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer International Publishing.

[2] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

[3] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

[4] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[5] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[6] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[7] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[9] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1290-1299).

[10] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 43(10), 3349-3364.

[11] Sofiiuk, K., Petrov, I. A., & Konushin, A. (2022, October). Reviving iterative training with mask guidance for interactive segmentation. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 3141-3145). IEEE.

[12] Lin, Z., Duan, Z. P., Zhang, Z., Guo, C. L., & Cheng, M. M. (2022). Focuscut: Diving into a focus view in interactive segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2637-2646).

[13] Liu, Q., Xu, Z., Bertasius, G., & Niethammer, M. (2023). Simpleclick: Interactive image segmentation with simple vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22290-22300).

[14] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[15] Shi, B., Gai, S., Darrell, T., & Wang, X. (2023). Toast: Transfer learning via attention steering. arXiv preprint arXiv:2305.15542, 5(7), 13.

[16] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[17] Aleem, S., Dietlmeier, J., Arazo, E., & Little, S. (2024). ConvLoRA and AdaBN based Domain Adaptation via Self-Training. arXiv preprint arXiv:2402.04964.

[18] Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., & Ayed, I. B. (2019, May). Boundary loss for highly unbalanced segmentation. In International conference on medical imaging with deep learning (pp. 285-296). PMLR.

[19] Ke, L., Ye, M., Danelljan, M., Tai, Y. W., Tang, C. K., & Yu, F. (2024). Segment anything in high quality. Advances in Neural Information Processing Systems, 36.

[20] Qin, X., Dai, H., Hu, X., Fan, D. P., Shao, L., & Van Gool, L. (2022, October). Highly accurate dichotomous image segmentation. In European Conference on Computer Vision (pp. 38-56). Cham: Springer Nature Switzerland.

[21] Liew, J. H., Cohen, S., Price, B., Mai, L., & Feng, J. (2021). Deep interactive thin object selection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 305-314).

[22] Li, X., Wei, T., Chen, Y. P., Tai, Y. W., & Tang, C. K. (2020). Fss-1000: A 1000-class dataset for few-shot segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2869-2878).

[23] Shi, J., Yan, Q., Xu, L., & Jia, J. (2015). Hierarchical image saliency detection on extended CSSD. IEEE transactions on pattern analysis and machine intelligence, 38(4), 717-729.

[24] Cheng, M. M., Mitra, N. J., Huang, X., Torr, P. H., & Hu, S. M. (2014). Global contrast based salient region detection. IEEE transactions on pattern analysis and machine intelligence, 37(3), 569-582.

[25] Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M. H. (2013). Saliency detection via graph-based manifold ranking. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3166-3173).

[26] Cheng, B., Girshick, R., Dollár, P., Berg, A. C., & Kirillov, A. (2021). Boundary IoU: Improving object-centric image segmentation evaluation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 15334-15342).