УДК 004.853 DOI 10.18522/2311-3103-2025-2-212-220

В.Д. Матвеев, А.Е. Архипов, И.С. Фомин

РАЗРАБОТКА МОДЕЛИ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ RTC-SAM ДЛЯ ОПРЕДЕЛЕНИЯ ПРЕПЯТСТВИЙ НА ПУТИ МОБИЛЬНОГО РОБОТА

Задача определения препятствий перед мобильным роботом успешно и давно решена с применением лазерных и ультразвуковых датчиков. Однако, препятствия, не обнаруживаемые такими видами датчиков, могут угрожать безопасности робота. Для их обнаружения в работе предлагается использовать систему технического зрения (СТЗ), информацию с которой обрабатывает нейронная сеть семантической сегментации, возвращающая маску препятствия на кадре и его класс. Основой для такой сети стала сеть универсальной сегментации SAM, требующая доработки для применения к задаче семантической сегментации. Особенность данной сети состоит в ее универсальной применимости, то есть возможности выделения любых объектов в произвольных условиях съемки. При этом SAM не предсказывает семантику объекта. В данной работе предложен дополнительный модуль, позволяющий реализовать семантическую сегментацию за счет классификации признаков выделяемых объектов. Обоснована возможность использования такого модуля для решения задачи дополнения выхода сети новой информацией. Результат классификации далее поступает в тот же алгоритм фильтрации, что и маски, чтобы гарантировать соответствие между полученным результатом универсальной сети и дополняющего модуля. После интеграции модуля с моделью получена новая модель семантической сегментации, названная в работе RTC-SAM. С ее помощью проведена семантическая сегментация общедоступного набора данных с изображениями открытой местности. Полученный результат в 45 % по метрике IoU превосходит результат существующих методов на 13 %. Показанные в работе изображения результатов применения новой сети позволяют убедиться в ее работоспособности. Также описано тестирование разработанного решения с проведением исследования быстродействия разработанной модели на ПК и мобильном вычислителе. Алгоритм на мобильном вычислителе показывает недостаточную скорость для выхода в режим реального времени – больше 3,5 секунд на обработку одного кадра. В связи с этим, одно из направлений дальнейших исследований в области повышения быстродействия системы.

Нейронная сеть; сегментация; классификация векторов; вектор представлений; система технического зрения; препятствия.

V.D. Matveev, A.E. Arkhipov, I.S. Fomin

COMBINING SEGMENTATION, TRACKING, AND CLASSIFICATION MODELS TO SOLVE VIDEO ANALYTICS PROBLEMS

The task of detecting obstacles in front of a mobile robot has been successfully solved long ago using laser and ultrasonic sensors. However, obstacles that are not detected by these types of sensors may endanger the safety of the robot. To detect them in the work, it is proposed to use a technical vision system (STZ), the information from which is processed by a semantic segmentation neural network, which returns the mask of the obstacle on the frame and its class. The basis for such a network was the SAM universal segmentation network, which requires further development to be applied to the semantic segmentation task. The peculiarity of this network is its universal applicability, that is, the ability to select any objects in any filming situation. At the same time, SAM does not predict the semantics of the object. In this paper, an additional module is proposed that makes it possible to implement semantic segmentation by classifying the features of the selected objects. The possibility of using such a module to solve the problem of supplementing the network output with new information is substantiated. The classification result is then fed into the same filtering algorithm as the masks to ensure consistency between the result of the universal network and the complementary module. After integrating the module with the model, a new semantic segmentation model was obtained, called RTC-SAM in the work. It was used to perform semantic segmentation of a publicly available dataset with images of an open area. The 45% result obtained by the IoU metric exceeds the result of existing methods by 13%. The images of the results of using the new network shown in the work make it possible to verify its performance. It also describes the testing of the developed solution with a study of the performance of the developed model on a PC and a mobile computer. The algorithm on the mobile computer shows insufficient speed to enter real-time mode – more than 3.5 seconds to process one frame. In this regard, one of the directions of further research in the field of improving system performance.

Neural network; segmentation, classification of vectors; representation vector; vision system; obstacles.

Введение. Использование мобильных роботов в открытой среде сопряжено с определенными рисками, которые могут оказать сильное влияние на выполнимость миссии робота. Одним из таких рисков является потеря устройства из-за попадания в неопределенное препятствие, которое не было распознано стандартными датчиками мобильных роботов – Lidar или ультразвуковыми датчиками [1, 2]. К такому виду опасных препятствий могут относиться, например, водные преграды – которые не обнаруживаются с помощью 2D-Lidar. Такого рода преграды могут быть обнаружены с помощью системы технического зрения (СТЗ), поскольку при телеуправлении оператор всегда обходит данные препятствия. Для управления роботом без оператора необходимо реализовать алгоритм обработки изображения на мобильном вычислителе с целью семантической сегментации объектов. Получаемые маски объектов можно использовать для построения пути мобильного робота.

Сегодня наилучше качество роботы для семантической обработки изображений демонстрируют нейронные сети. Существует множество моделей семантической сегментации, которые обучены выделять различные объекты на кадрах. Например, известная модель PSPNet [3] представила пирамидальную структуру обработки изображений, на основе которой работают многие последующие модели сегментации. Или одна из первых сегментационных моделей на основе трансформерных архитектур обработки изображений – SegFormer [4]. Однако, большинство моделей не обладают важным для мобильных систем свойством – универсальностью. Дело в том, что существующие на сегодняшний день наборы данных для сегментации изображений пересеченной местности (некоторые из них будут описаны далее в работе) имеют примерно в 2-3 раза меньше объем, чем наборы данных для обучения автовождению – в связи с чем качество обучения моделей на таких наборах сильно ниже того же обучения сегментации дороги. И поэтому, для задач с малым количеством обучающих данных стоит использовать более универсальные подходы. Довольно важным шагом в преодолении этой проблемы стало появление универсальной модели сегментации SAM (Segment Anything Model) [5], которая была обучена на огромном наборе данных и приблизилась к уровню фундаментальных моделей по возможностям сегментации. Единственный серьезный недостаток данной сети в том, что она возвращает маски объектов без классов – то есть не работает как семантический сегментатор.

В связи с этим, данная работа посвящена разработке новой модели сегментации на базе SAM модели, которая позволит применить качество формирования масок этой модели и ее универсальность к задачам семантической сегментации. В первую очередь, к задачам семантической сегментации препятствий перед мобильным роботом. Она является продолжением работы [6], которая посвящена комплексированию нескольких нейронных сетей для получения алгоритма семантической сегментации и сопровождения объектов.

Универсальная модель сегментации. Сегментация изображений способна формировать необходимые области, с которыми может взаимодействовать мобильный робот, для построения его дальнейшей системы управления. Наиболее качественно моделью на сегодняшний день является модель универсальной сегментации SAM (Segment Anything Model) [5]. Благодаря особому подходу к сегментации – сегментации по подсказкам, данная сеть не привязывается к определенным классам объектов и способна сегментировать любые объекты. Кроме того, данная модель была подготовлена на самом большом из существующих наборов данных, включающем в себя 11 миллионов изображений и более 1 миллиарда масок – что также способствует формированию общего представления объектов.

Модель формирует по несколько масок для каждой подсказки, для каждой точки из заранее заданной сетки, после чего в процессе работы алгоритма выбирает наиболее подходящую для отрисовки. Такой подход был продемонстрирован ранее в работе [7], показав, что задача семантической сегментации через попиксельную классификацию изображений не эффективна в плане обучения сети.

Однако, из-за такого принципа работы, сети универсальной сегментации SAM не обладает семантикой объектов, которые она выделяет. Для нее это просто объекты, и на соседних кадрах даже при совпадении масок они будут выделены по-разному. Что полностью не соответствует задаче выделения препятствий перед мобильным роботом. В связи с чем было принято решение доработать рассматриваемую модель дополнитель-

ным модулем. Такой модуль должен работать без вмешательства в основную модель – чтобы не потерять накопленный полученные из самого крупного сегментационного датасета знания, и при этом реализовывать определение классов сегментированных объектов.

Классификация векторов представления сети. Вектор представления является частью скрытого состояния сети. Данные вектора формируются в ходе работы любых сетей и отражают входную и выходную информацию в виде некоторого массива чисел, понятного для конкретного шага сети. На работе с вектора представления сети, в частности – классификации, строится множество нейронных сетей. Так, многие языковые трансформеры, вроде GPT-3/4 или Claude [8, 9] переводят информацию в векторное представление для ее дальнейшего распознавания и присвоения текстам различных классов.

Кроме того, специальный тип обучения – контрастное обучение [10], полностью строится на работе с векторами представления. Суть этого типа обучения заключается в том, что векторы представления похожих объектов сближаются, а отличных друг от друга – намерено удаляются. Таким образом удается сформировать некоторое пространство, попадая в которое новые вектора признаков сразу же оказываются возле нужного класса. Как, например, это сделано в универсальной модели классификации CLIP [11] – модель обучалась сближать векторные представления изображений и векторные представления их текстовых описаний. И на этапе тестирования, на вход подается изображение – а на выходе его наиболее близкое текстовое описание.

Из приведенных выше примеров можно сделать вывод, что классификация сегментированных областей по векторам представлений может быть работоспособна. В качестве классификатора векторов представлений в данной работе будет использован многослойный персептрон, обученный на собранных векторах признаков и затем добавленный к оригинальной модели. Данный вид сети выбран потому, что небольшим размером (возможно уложиться в 100 тысяч параметров), в отличии от довольно известных сетей классификации, вроде VGG16 [12] – что напрямую сказывается на скорости обучения и нагрузке вычислителя, на котором работает алгоритм.

Классификация векторов представления. Следующим этапом необходимо выяснить, где в модели формируются наиболее подходящие для классификации вектора представлений. Архитектура SAM состоит из двух энкодеров (для изображений и подсказок) и одного декодера (для формирования маски). Энкодер изображений позволяет извлекать семантическое представление всего изображения, без привязки к какому-либо объекту, в то время как энкодер подсказок наоборот – выделяет пространственное представление без привязки к конкретному изображению. Объединение признаков из этих двух источников информации происходит в декодере, позволяя выделить в общем семантическом представлении нужный объект. На рис. 1 редставлена архитектура декодера с нанесенными на нее точками, в которых проводятся первые тесты возможности классификации.

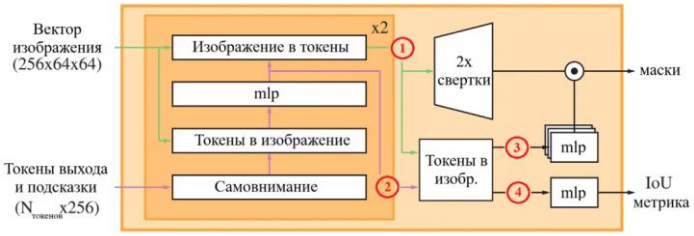

Рис. 1. Архитектура декодера и точки сбора векторов представления

Представленные точки представляют наибольший интерес, поскольку в каждой из них гарантировано произошло объединение признаков от изображений и подсказок, с дальнейшим проведением различных операций. Полученные результаты сравниваются по доле верно распознанных изображений, а также по размеру вектора признаков, который напрямую влияет на размер обучаемого многослойного персептрона.

Набор данных. Для тестирования алгоритма определения препятствий перед мобильным роботом был проведен поиск подходящих наборов данных. Существует множество открытых наборов данных, снятых с камер мобильных роботов или в сходной с ними перспективе [13–15]. Однако, в этих датасетах отсутствует разметка семантической сегментации, с которой можно было бы затем сравнить свое разработанное решение. Для тестирования рассматриваемой задачи разработки семантического сегментатора определения препятствий мобильным роботом было найдено всего два подходящих набора данных – RUGD [16] и Rellis-3d [17]. Оба этих датасета содержат примеры проездов мобильных роботов в открытой среде, а также разметку окружающих его объектов. На рис. 2 представлены примеры из этих наборов данных – в левой части из RUGD, в правой из Rellis-3d.

Рис. 2. Пример изображений из используемых наборов данных

В дальнейшей работе набор данных RUGD был использован для обучения и тестирования многослойного персептрона, в то время как на Rellis-3d проводилось только тестирование. Это связано с тем, что в оригинальной работе RUGD есть расчет метрик сегментации Intersection over Union (IoU), о которой будет сказано ниже, и таким образом обучившись на том же, на чем обучаются авторы работы и протестировав на том же возможно проведение сравнения качества семантической сегментации объектов с помощью новой модели нейронной сети.

Классы для многослойного персептрона и последующей семантической сегментации были выбраны следующие: «опорная поверхность» – часть пространства, по которому робот может проехать без каких-либо затруднений, «препятствие» – небольшие препятствия на пути робота, требующие внимание, однако возможно преодолимые, как например камни или невысокий кустарник, «деревья» – высокие препятствия, непреодолимые для робота, собственно деревья, фонарные столбы, «вода» – водные преграды, которые встречаются на некоторых проездах, хотя и редко; «небо» – класс, не имеющий отношения к обходу препятствий, однако присутствующий на любых кадрах вне зданий – и введенный для избежания ошибок классификации.

Разработка классификатора векторов. Обучение многослойного персептрона проходило на изображениях из обучающей выборки RUGD набора данных. Изображения подавались в сеть вместе с равномерной сеткой подсказок. Для каждой пары изображение – точка формировалось по три вектора представления внутри сети SAM (согласно алгоритму работы самой SAM). Затем полученные представления и соответствующие им маски сохранялись под классом, который формировался следующим образом. Каждая маска сравнивалась с областями, размеченными людьми в RUGD. Класс той области, с которой была наибольшая метрика IoU у полученной маски, и присваивался вектору представлений для дальнейшего обучения. Тестирование возможности использования классификатора векторов представления было проведено на одном небольшом проезде из RUGD с около 7500 сформированными векторами для обучения. В наилучшей точке приложения классификатора были обработаны все обучающий проезды изначального набора и получено более 130 тысяч векторов.

Затем проходило обучение самого многослойного персептрона с использованием хорошо изученного стохастического градиентного спуска (Stochastic Gradient Decent, SGD) [18] в течение 30 эпох. Далее этот классификатор тестировался на векторных представлениях тестовых проездов, сформированных SAM моделью.

Семантическая сегментация. После выбора точки классификации и подготовки многослойного персептрона он встраивается в изначальный алгоритм сети SAM, дополняя и не изменяя его. Таким образом получается новая модель сети семантической сегментации, RTC-SAM, чья обобщенная архитектура представлена на рис. 3. На ней желтым выделено то, что было в оригинальном SAM, а зеленым показаны доработанные части.

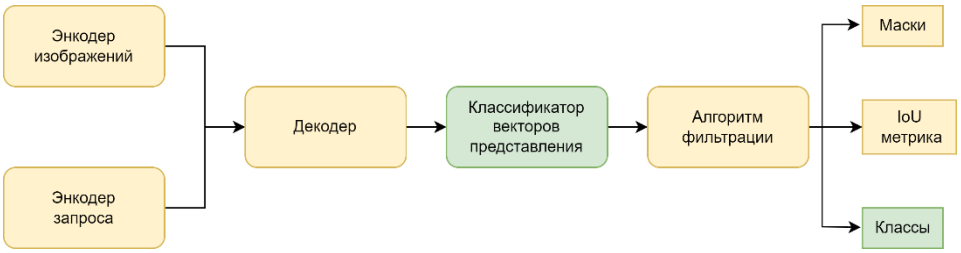

Рис. 3. Обобщенная архитектура разработанной сети

Затем, с помощью указанной модели была проведена семантическая сегментация тестовых проездов RUGD для возможности сравнения полученных результатов с разработанными ранее методами. Сравнение проводилось по метрике Intersection Over Union, которая представлена в формуле 1.

$$IoU = \frac{Overlap}{Union}, \quad (1)$$

где $Overlap$ – площадь пересечения разметки объекта и созданной алгоритмом маски,

$Union$ – объединенная площадь разметки объекта и созданной алгоритмом маски.

Далее проходило усреднение результатов по классам, поскольку в оригинальной работе RUGD уже дано mean IoU – т.е. среднее значение метрики; а также потому, что количество классов в настоящей работе меньше, чем в рассматриваемом наборе данных.

Выбор точки установки классификатора. После обучения классификатора в четырех точках, показанных выше на рис. 1, и проведения тестирования были получены результаты, сформировавшие таблицу табл. 1.

Таблица 1

Результат обучения классификатора в различных частях декодера

№ точки	Количество элементов вектора признаков	Верно распознанные изображения из теста, %
1	1 048 576	95
2	1 792	87,5
3	1 024	92,5
4	256	82,5

Из таблицы видно, что наибольшую точность показывает первая точка. Однако, она также имеет и наибольший размер вектора признаков. Персептрон, являясь полносвязной моделью машинного обучения, с таким размером вектора признаков будет иметь больше 200 млн параметров – слишком большое значение, не подходящее для применения на мобильном вычислителе (упомянутая VGG16, принятая не удовлетворительной, имеет 138 млн). С другой стороны, третья точка, обладая близкой точностью, имеет в разы меньше параметров – около 200 тысяч. Данная точка больше подходит для создания классификатора векторов признаков, поскольку гораздо меньше увеличивает нагрузку на вычислитель потенциального робота. Она и выбрана для дальнейшей разработки алгоритма.

Результат семантической сегментации. После выбора точки применения классификатора он был обучен на проездах из обучающего набора RUGD, добавлен к изначальной модели сегментации по подсказкам, в соответствии с рис. 3, и затем общий полученный алгоритм испытан на тестовых проездах RUGD. В табл. 2 представлены результаты

сравнения полученной модели семантической сегментации с моделями, использованными при разработке самого датасета. В работе авторы использовали ResNet50 [19] в качестве энкодера для перевода изображений в пространство представлений, а в качестве декодера: простые последовательности блоков, упомянутую PSPNet или UperNet [20].

Таблица 2

Результат сравнения работы семантической сегментации разработанной моделью

Модель семантической сегментации	Среднее IoU, %
ResNet50 + восстанавливающая свертка, апсемплинг и проверка	32,24
ResNet50 + PSPNet	32,07
ResNet50 + PSPNet с проверкой	31,71
ResNet50 + Upernet	31,95
RTC-SAM (разработанная модель)	45,03

Из таблицы видно, что разработанная модель превосходит базовые методы, с которыми проходило сравнение, на довольно значительное количество процентов. Отсюда можно сделать вывод, что точность и аккуратность масок, формируемых SAM моделью – действительно довольно высока. А доработка основной модели небольшим (при размере SAM модели в 90 млн параметров) классификатором позволила применить этот высококачественный результат сегментации к новой задаче. На рис. показаны достигнутые результаты семантической сегментации на наборах RUGD и Rellis-3d. Зеленым цветом выделены опорные поверхности для робота; красным – препятствия; фиолетовым – деревья; голубым – небо.

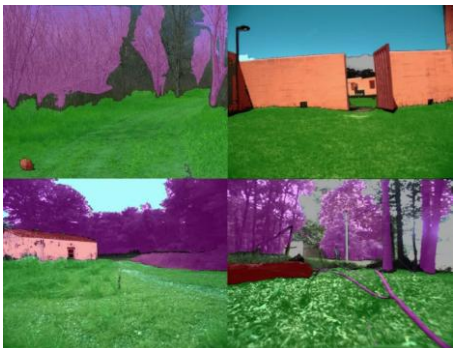

Рис. 4. Результат семантической сегментации с помощью разработанной модели

Исследование быстродействия модели на мобильном вычислителе. После подтверждения работоспособности разработанного алгоритма семантической сегментации на основе универсальной модели и обученного классификатора, было проведено тестирование быстродействия метода. Проведены замеры времени работы алгоритма как на мобильном вычислителе для предполагаемого робота – Nvidia Jetson AGX Xavier, так и на лабораторном ПК, чья работоспособность оценивается примерно в 3 раза выше по количеству операций с плавающей точкой (Flops), чем работоспособность самого современного встраиваемого Jetson AGX Orin. В табл. 3 представлены характеристики обоих устройств, на которых проходили замеры.

Быстродействие измерялось при различном количестве точек, которые выступают в качестве подсказок в алгоритме. Плотность этих точек напрямую влияет на количество выделяемых объектов и необходимых к обработке областей. В табл. 4 представлены результаты замеров времени, затрачиваемого на обработку одного изображения с различным количеством точек.

Таблица 3

Характеристики используемых устройств

Характеристика	Лабораторный ПК	Nvidia Jetson Xavier
Процессор	Intel Core i7-6700 @3.4ГГц x 8	Tegra Xavier 8xNvidia Carmel @2.26ГГц
Видеоускоритель	Nvidia GeForce RTX 2080 Ti 12 Гб	Nvidia Volta GV10B 32 Гб
Объем оперативной памяти	16 Гб	32 Гб, используется вместе с GPU

Таблица 4

Измерение быстродействия алгоритма, сек

Количество точек (подсказок) на 1 изображение	5x5	10x10	20x20
Лабораторный ПК	0,379	0,539	1,089
Nvidia Jetson Xavier	3,82	4,61	7,83

Из данной таблицы можно сделать несколько выводов. Во-первых, с ростом количества точек компьютер затрачивает больше времени на их обработку – двухкратный рост времени при переходе с 10x10 к 20x20 точек – в то время как у Jetson рост времени составляет 1,7 раз. Это показывает, что используемое на мобильном вычислителе программное обеспечение более оптимизировано и легче справляется со своей задачей – обработки вычислений с нейронными сетями, чем ПО на компьютере, где оно кроме таких вычислений предполагается для отрисовки графики и 3D моделей.

Второй важный вывод состоит в том, что мобильный вычислитель в целом работает в 10 раз медленнее лабораторного ПК и не выдает даже 1 кадра в секунду производительности. Такой результат не позволяет говорить о режиме реального времени для алгоритма на мобильном роботе и ставит задачу дальнейшего исследования в сторону увеличения производительности. Однако, доказательство возможности использования последних алгоритмов в области обработки изображений нейронными сетями на предыдущих поколениях вычислителей позволяет использовать эти вычислители в различных прикладных задачах – вместе с разрабатываемыми алгоритмами.

Заключение. В настоящей работе рассмотрена разработка сети семантической сегментации RTC-SAM для сегментации препятствий на пути мобильного робота. Сеть строится на основе сети универсальной сегментации SAM, которая формирует маски объектов по подсказкам – но без понимания классов. Для классификации сформированных масок предложено использование классификатора векторов представления признаков и обоснована возможность такого подхода. Далее выбраны точки, в которых проведение классификации признаков кажется наиболее возможным, а также обучены классификаторы на небольшом поднаборе для каждой их этих точек. Сделан вывод о том, что классификация векторов представления в рассматриваемой задаче возможна с достаточно высокими точностями до 95 % по долям верного распознавания, и выбрана наиболее подходящая для работы на мобильном устройстве точка.

Затем классификатор в этой точке переобучен на большем количестве обучающих данных выбранного датасета – и интегрирован в универсальную сеть сегментации SAM. Полученная таким образом сеть семантической сегментации RTC-SAM применена к тестовым данным выбранного датасета для получения метрики семантической сегментации. Данная метрика позволила сравнить имеющиеся подходы с разработанным, который показал свое превосходство на 13 % по метрике усредненного IoU.

Портирование разработанной сети на мобильный вычислитель Nvidia Jetson Xavier оказалось возможным, однако проведение замеров быстродействия на данном оборудовании показало, что на данном этап достичь режима реального времени на мобильном роботе не получится – алгоритм показал скорость работы меньше 1 кадра в секунду, в то время как лабораторный ПК показывает 2,64 кадра/сек.

В связи с этим, определены следующие направления дальнейших исследований. Во-первых, необходимо повысить скорость работы алгоритма. Для этого стоит проверить возможность оптимизации универсальной сети сегментации, так как именно она тратит наибольшее время в рассматриваемом методе. Во-вторых, провести применения разработанного подхода – семантической сегментации за счет классификации векторов представления – на новой версии универсальной сети SAM2.

Результаты получены в рамках выполнения государственного задания Минобрнауки России «Исследование методов анализа слабоструктурированных данных, обработки знаний и создания когнитивных агентов на базе комбинированных глубоких нейронных сетей» (FNRG-2025-0008 1024050200009-5-1.2.1;2.2.2)

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Peng Y. et al.* The obstacle detection and obstacle avoidance algorithm based on 2-d lidar // 2015 IEEE international conference on information and automation. – IEEE, 2015. – P. 1648-1653.
2. *Gibbs G., Jia H., Madani I.* Obstacle detection with ultrasonic sensors and signal analysis metrics // Transportation Research Procedia. – 2017. – Vol. 28. – P. 173-182.
3. *Zhao H. et al*. Pyramid scene parsing network // Proceedings of the IEEE confer-ence on computer vision and pattern recognition. – 2017. – P. 2881-2890.
4. *Xie E. et al.* SegFormer: Simple and efficient design for semantic segmentation with transformers // Advances in neural information processing systems. – 2021. – Vol. 34. – P. 12077-12090.
5. *Kirillov A. et al*. Segment anything // Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2023. – P. 4015-4026.
6. *Архипов А.Е., Фомин И.С., Матвеев В.Д.* Комплексирование моделей сегментации, сопровождения и классификации для решения задач видеоаналитики // Известия ЮФУ. Технические науки. – 2024. – № 1.
7. *Cheng B., Schwing A., Kirillov A.* Per-pixel classification is not all you need for semantic segmentation // Advances in Neural Information Processing Systems. – 2021. – Vol. 34. – P. 17864-17875.
8. *Anthropic A.I.* Claude 3.5 sonnet model card addendum // Claude-3.5 Model Card. – 2024. – Vol. 3.
9. *Kalyan K.S.* A survey of GPT-3 family large language models including ChatGPT and GPT-4 // Natural Language Processing Journal. – 2024. – Vol. 6. – P. 100048.
10. *Chen T. et al.* A simple framework for contrastive learning of visual representations // International conference on machine learning. – PMLR, 2020. – P. 1597-1607.
11. *Radford A. et al.* Learning transferable visual models from natural language super-vision // International conference on machine learning. – PMLR, 2021. – P. 8748-8763.
12. *Simonyan K.* Very deep convolutional networks for large-scale image recognition // arXiv preprint arXiv:1409.1556. – 2014.
13. *Jeong S., Kim H., Cho Y.* Diter: Diverse terrain and multi-modal dataset for field robot navigation in outdoor environments // IEEE Sensors Letters. – 2024.
14. *Shah D. et al.* Rapid exploration for open-world navigation with latent goal models // arXiv preprint arXiv:2104.05859. – 2021.
15. *Shah D. et al.* Ving: Learning open-world navigation with visual goals //2021 IEEE International Conference on Robotics and Automation (ICRA). – IEEE, 2021. – P. 13215-13222.
16. *Wigness M. et al.* A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments // 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). – IEEE, 2019. – P. 5000-5007.
17. *Jiang P. et al.* Rellis-3d dataset: Data, benchmarks and analysis // 2021 IEEE international conference on robotics and automation (ICRA). – IEEE, 2021. – P. 1110-1116.
18. *Amari S.* Backpropagation and stochastic gradient descent method // Neurocomputing. – 1993. – Vol. 5, No. 4-5. – P. 185-196.
19. *He K. et al.* Deep residual learning for image recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – P. 770-778.
20. *Xiao T. et al.* Unified perceptual parsing for scene understanding // Proceedings of the European conference on computer vision (ECCV). – 2018. – P. 418-434.

REFERENCES

1. *Peng Y. et al.* The obstacle detection and obstacle avoidance algorithm based on 2-d lidar, *2015 IEEE international conference on information and automation*. IEEE, 2015, pp. 1648-1653.
2. *Gibbs G., Jia H., Madani I.* Obstacle detection with ultrasonic sensors and signal analysis metrics, *Transportation Research Procedia*, 2017, Vol. 28, pp. 173-182.

3. *Zhao H. et al*. Pyramid scene parsing network, *Proceedings of the IEEE confer-ence on computer vision and pattern recognition*, 2017, pp. 2881-2890.
4. *Xie E. et al.* SegFormer: Simple and efficient design for semantic segmentation with transformers, *Advances in neural information processing systems*, 2021, Vol. 34, pp. 12077-12090.
5. *Kirillov A. et al*. Segment anything, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015-4026.
6. *Arkhipov A.E., Fomin I.S., Matveev V.D.* Kompleksirovanie modeley segmentatsii, soprovozhdeniya i klassifikatsii dlya resheniya zadach videoanalitiki [Integration of segmentation, tracking and classification models for solving video analytics problems], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2024, No. 1.
7. *Cheng B., Schwing A., Kirillov A.* Per-pixel classification is not all you need for semantic segmentation, *Advances in Neural Information Processing Systems*, 2021, Vol. 34, pp. 17864-17875.
8. *Anthropic A.I.* Claude 3.5 sonnet model card addendum, *Claude-3.5 Model Card*, 2024, Vol. 3.
9. *Kalyan K.S.* A survey of GPT-3 family large language models including ChatGPT and GPT-4, *Natural Language Processing Journal*, 2024, Vol. 6, pp. 100048.
10. *Chen T. et al.* A simple framework for contrastive learning of visual representations, *International conference on machine learning*. PMLR, 2020, pp. 1597-1607.
11. *Radford A. et al.* Learning transferable visual models from natural language super-vision, *International conference on machine learning*. PMLR, 2021, pp. 8748-8763.
12. *Simonyan K.* Very deep convolutional networks for large-scale image recognition // arXiv preprint arXiv:1409.1556, 2014.
13. *Jeong S., Kim H., Cho Y.* Diter: Diverse terrain and multi-modal dataset for field robot navigation in outdoor environments, *IEEE Sensors Letters*, 2024.
14. *Shah D. et al.* Rapid exploration for open-world navigation with latent goal models, *arXiv preprint arXiv:2104.05859*, 2021.
15. *Shah D. et al.* Ving: Learning open-world navigation with visual goals, *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13215-13222.
16. *Wigness M. et al.* A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments, *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5000-5007.
17. *Jiang P. et al.* Rellis-3d dataset: Data, benchmarks and analysis, *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 1110-1116.
18. *Amari S.* Backpropagation and stochastic gradient descent method, *Neurocomputing*, 1993, Vol. 5, No. 4-5, pp. 185-196.
19. *He K. et al.* Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
20. *Xiao T. et al.* Unified perceptual parsing for scene understanding, *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418-434.

Матвеев Виктор Дмитриевич – Центральный научно-исследовательский и опытно-конструкторский институт робототехники и технической кибернетики; e-mail: v.matveev@rtc.ru; г. Санкт-Петербург, Россия; тел.: +78125523351; инженер.

Архипов Андрей Евгениевич – Центральный научно-исследовательский и опытно-конструкторский институт робототехники и технической кибернетики; e-mail: a.arkhipov@rtc.ru; г. Санкт-Петербург, Россия; тел.: +78125523351; младший научный сотрудник.

Фомин Иван Сергеевич – Центральный научно-исследовательский и опытно-конструкторский институт робототехники и технической кибернетики; e-mail: i.fomin@rtc.ru; г. Санкт-Петербург, Россия; тел.: +78125523351; научный сотрудник.

Matveev Victor Dmitrievich – Russian State Scientific Center for Robotics and Technical Cybernetics; e-mail: v.matveev@rtc.ru; Saint-Petersburg, Russia; phone: +78125523351; engineer.

Arkhipov Andrey Evgenievich – Russian State Scientific Center for Robotics and Technical Cybernetics; e-mail: a.arkhipov@rtc.ru; Saint-Petersburg, Russia; phone: +78125523351; junior researcher.

Fomin Ivan Sergeevich – Russian State Scientific Center for Robotics and Technical Cybernetics; e-mail: i.fomin@rtc.ru; Saint-Petersburg, Russia; phone: +78125523351; researcher.