Раздел IV. Техническое зрение

УДК 004.853 DOI 10.18522/2311-3103-2024-1-247-257

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

КОМПЛЕКСИРОВАНИЕ МОДЕЛЕЙ СЕГМЕНТАЦИИ, СОПРОВОЖДЕНИЯ И КЛАССИФИКАЦИИ ДЛЯ РЕШЕНИЯ ЗАДАЧ ВИДЕОАНАЛИТИКИ

Комплексирование нескольких моделей в одну систему технического зрения позволит решать более сложные и комплексные задачи. В частности, для мобильной робототехники и беспилотных летательных аппаратов (БЛА) является актуальной проблемой отсутствие наборов данных для различных условий. В работе в качестве решения данной проблемы предлагается комплексирование нескольких моделей: сегментации, сопровождения и классификации. Это позволит значительно повысить качество решения сложных задач без дополнительного обучения. Модель сегментации позволяет выделять произвольные объекты из кадров, поэтому ее можно использовать в недетерминированных и динамических средах. Модель классификации позволяет определить необходимые для навигации объекты, которые затем сопровождаются с помощью третей модели. В работе подробно описан алгоритм комплексирования моделей. Ключевым элементом в алгоритме является коррекция предсказаний моделей, позволяющая достаточно надежно сегментировать и сопровождать различные объекты. Процедура коррекции предсказаний моделей решает следующие задачи: добавление новых объектов для сопровождения, валидация сегментированных масок объектов и уточнение сопровождаемых масок. Универсальность данного решения подтверждается работой в сложных условиях, на которых не обучали модели, например, подводная съемка или изображения с БЛА. Проведено экспериментальное исследование каждой из моделей в условиях открытой местности и в помещении. Наборы данных включали сцены актуальные для мобильной робототехники. В частности, в сценах присутствовали движущиеся объекты (человек, автомобиль) и возможные преграды на пути робота. Для большинства классов метрики качества сегментации превышали 80%. Основные ошибки связаны с размерами объектов. Проведенные эксперименты наглядно демонстрируют универсальность данного решения без дополнительного обучения моделей. Дополнительно проведено исследование быстродействия на персональном компьютере с различными входными параметрами и разрешением. Увеличение количества моделей значительно повышает вычислительную нагрузку и не достигает реального времени. Поэтому одним из направления дальнейших исследований является повышение быстродействия системы.

Нейронные сети; сегментация; сопровождения; классификация; видеоаналитика; системы технического зрения.

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

INTEGRATION OF SEGMENTATION, TRACKING AND CLASSIFICATION MODELS TO SOLVE VIDEO ANALYTICS PROBLEMS

The integration of several models into one technical vision system will allow solving more complex tasks. In particular, for mobile robotics and unmanned aerial vehicles (UAVs), the lack of data sets for various conditions is an urgent problem. In the work, the integration of several models is proposed as a solution to this problem: segmentation, maintenance and classification. The segmentation model allows you to select arbitrary objects from frames, which allows it to be used in non-deterministic and dynamic environments. The classification model allows you to determine the ob-

jects necessary for navigation or other use, which are then accompanied by a third model. The paper describes an algorithm for model aggregation. In addition to models, the key element is the correction of model predictions, which allows you to segment and accompany various objects reliably enough. The procedure for correcting model predictions solves the following tasks: adding new objects to accompany, validating segmented object masks and clarifying the associated masks. The versatility of this solution is confirmed by working in difficult conditions, for example, underwater photography or images from UAVs. An experimental study of each of the models was carried out in an open area and indoors. The data sets used make it possible to assess the applicability of models for mobile robotics tasks, that is, to identify possible obstacles in the robot's path, for example, a curb, as well as moving objects such as a person or a car. They demonstrated a sufficiently high quality of work. For most classes, the indicators exceeded 80% by various metrics. The main errors are related to the size of the objects. The conducted experiments clearly demonstrate the versatility of this solution without additional training of models. Additionally, a study of performance on a personal computer with various input parameters and resolution was conducted. Increasing the number of models significantly increases the computational load and does not reach real time. Therefore, one of the directions of further research is to increase the speed of the system.

Neural networks; segmentation; tracking; classification; video analytics; computer vision.

Введение. Основным препятствием на пути повсеместного использования нейронных сетей для обработки видеопотока является отсутствие или неполнота обучающих наборов данных. Эта проблема достаточно сильно ограничивает развитие автономности у мобильных роботов и беспилотных летательных аппаратов (БЛА). Для ее решения необходимо создавать новые принципы обработки видеопотока и обучения нейронных сетей. Повышения качества систем технического зрения (СТЗ) можно добиться за счет объединения нескольких моделей глубокого обучения, каждая из которых решает локальную задачу, но на достаточно высоком уровне. Это в свою очередь неизбежно увеличивает вычислительную нагрузку на обработку отдельного кадра видеопотока. При этом позволяет повысить эффективность СТЗ за счет обучения каждой отдельной модели на большом количестве примеров. При решении комплексных задач дорого и сложно формировать обучающие наборы данных, поэтому неизбежно будут возникать ошибки при распознавании кадров. Другое преимущество объединения нескольких моделей в одну систему заключается в возможности регулирования и настройки отдельных шагов обработки видеопотока. Комплексирование предсказаний позволит значительно повысить эффективность и интерпретируемость СТЗ.

В работах [1–4] уже предлагались механизмы комплексирования нескольких моделей. Чаще всего работы направлены на сопровождение одного конкретного объекта, либо же его части. В работе [5] предложен механизм комплексирования сегментации и сопровождения множества объектов, однако он имеет множество недостатков, связанных с некачественным выделением объектов и накоплением ошибки при сопровождении. На практике его сложно интегрировать для СТЗ мобильных роботов.

В настоящей работе основной упор при разработке механизма комплексирования ставился на повышение эффективности СТЗ за счет сопоставление предсказаний моделей, а также валидации предсказаний.

Универсальная модель сегментации. Сегментация изображений на отдельные объекты позволяет извлекать необходимую информацию для реализации систем управления мобильными роботами и БЛА. Наибольшую точность на сегодняшний день продемонстрировала модель универсальной сегментации SAM (Segment Anything Model) [6]. Ее ключевая особенность заключается в том, что при сегментации она не привязана к конкретным классам объектов. Ее обучали общему представлению, что такое объекты, поэтому ее можно применять для любых типов объектов и изображений без дополнительного обучения, что раньше не представлялось возможным. Авто-

ры данной модели собрали самый большой на сегодняшний день набор данных для сегментации, состоящий из 11 млн изображений и более 1 млрд масок. Преимущество рассматриваемой модели особенно актуально при работе в недетерминированной среде с заранее неопределенным набором объектов.

Принцип функционирования модели следующий. На изображение накладывается равномерная сетка из точек, для каждой из которой затем предсказывается маска объекта, на который попала данная точка. При этом обработка производится на уровне признаков изображений. Если задана одна точка – сеть вернет область в окрестности этой точки. Если две в различных частях кадра – на выходе будут две автономные области. При подаче двух и более точек, которые принадлежат одной области кадра, метод произведет операцию подавления немаксимумов, где оставляется одна точка, принадлежащая области, а остальные, принадлежащие той же области, но меньшие по показателю уверенности, подавляются, результатом будет одна область, которая назначается для всех переданных точек.

Еще одна особенность такого подхода заключается в возможном пересечении или наложении масок объектов. Например, при выделении на изображении здания и окон в случае семантической сегментации окна не будут являться частью здания, а будут выделены отдельной маской. В случае с SAM маска для окон будет выделена как отдельный объект «окно», и включена в маску «здания». Таким образом сеть лучше определяет семантику объектов во время обучения. Такой подход впервые использовался в работе [7], где наглядно продемонстрировали, что рассмотрение задачи семантической сегментации как попиксельной классификации не является эффективным способом обучения сети.

Архитектура SAM состоит из двух энкодеров и одного декодера. Энкодер в общем случае – это нейронная сеть, которая переводит изображение, текст или иные данные в сжатый вектор представлений, который обычно представляет из себя вектор чисел заданной длины. Первый энкодер выделяет вектор признаков из входного изображения. После обучения данный вектор признаков для каждого изображения несет некоторую высокоуровневую обобщенную информацию. Второй энкодер отвечает за извлечение признаков промпта (prompt – подсказка, запрос) – это запрос, подсказка, или инструкция для нейронной сети, которая обычно задается в текстовой форме. В сети SAM запрос может быть представлен координатами точки или прямоугольника, бинарной маской или текстовым описанием. Для каждой пары признаков промпта и изображения генерируется с помощью декодера бинарная маска, в которой 1 означает наличие объекта, соответствующего подсказке, а 0 его отсутствие. Для каждого промпта генерируется собственная бинарная маска. Во время обработки видеопотока с помощью SAM необходимо сопоставлять объекты на соседних кадрах последовательности для обеспечения согласованности между объектами.

Модель сопровождения. Задача сопровождения масок сегментации в видеопотоке появилась не так давно, что связано в первую очередь со сложностью формирования наборов данных для видео. Подразумевается, что на вход модели поступает кадр и соответствующая для него маска сегментации объектов. Модель выполняет сопровождение выделенных объектов на последующих кадрах, то есть для каждого изображения модель выдает маску сегментации. Такой подход может быть достаточно эффективен при сегментации первого кадра с помощью SAM с последующим сопровождением полученных масок объектов. Наибольшие показатели метрик подремонтировала модель для сопровождения множества объектов DeAOT (Decoupling Features in Association Objects with vision Transformers, разделение признаков, объединенное с визуальными трансформерами) [8, 9] ориентированная на иерархическое распространение результатов полуавтоматической разметки объектов для задачи сегментации объектов на видео. Модель имеет кратковременную и долговременную память. Отличие между ними состоит в количестве охватываемых кадров для запоминания. Основным ограничением при запоминании является память вычислительного устройства.

С практической точки зрения взаимодействие с методом выглядит следующим образом. На вход алгоритма подаются полученные тем или иным способом маски объектов (для каждого пикселя изображения задано значение класса объекта в диапазоне от 1 до 255, 0 – отсутствие объекта, фон). Далее нейронная сеть для последующих кадров видео самостоятельно выделит эти объекты за счет сопоставления признаков объектов на изображении с масками. Новая маска показывает, куда переместились на кадре соответствующие объекты. В силу правил кодирования, всего за один запуск на одном видео или на одной группе изображений может присутствовать не более 255 различных объектов, что накладывает некоторые ограничения, поэтому требуется предварительно обрабатывать данные сегментации так, чтобы не превысить это число.

Визуально-языковая модель. Для классификации отдельных объектов, выделенных с помощью универсальной модели сегментации может быть использована визуально-языковая модель. Особенность данного класса моделей состоит в мультимодальности представления данных, то есть они могут сопоставлять текстовые описания с изображениями. На сегодняшний день выложено в открытой доступ уже множество подобных моделей [10-12]. В работе используется модель CLIP [13], первая из класса визуально-языковых моделей. Принцип работы сети заключается в следующем. Она состоит из двух энкодеров, обученных на большом количестве примеров. Первый обрабатывает изображение и преобразовывает его в вектор признаков, размером 512 элементов. Вектора для похожих изображений ложатся в близкие области пространства признаков, вектора для разных изображений ложатся в далекие области пространства признаков. Второй энкодер принимает на вход короткий текст, одно или несколько слов, описывающих объект или понятие, а на выходе возвращает вектор той же размерности. Аналогично, вектора, соответствующие семантически близким понятиям, ложатся в пространстве признаков близко друг от друга, далекие понятия – далеко друг от друга. Пространства признаков для обеих сетей напрямую связаны между собой, то есть вектор для текстового описания объекта будет в пространстве признаков расположен близко к вектору для изображения этого объекта.

Исходя из этого предлагается следующий вариант реализации классификации областей с использованием модели CLIP. На вход поступает изображение с размеченным рассмотренными выше средствами набором областей. Каждая область извлекается из RGB-изображения стандартными средствами, и обрабатывается сетью, отвечающей за извлечение вектора признаков из изображений. Для всех понятий (объектов), которые должны распознаваться алгоритмом, а также ряда синонимов заранее построены вектора в пространстве признаков с помощью текстовой части сети. Далее для изображений каждого объекта извлекаются признаки и оценивается схожесть с признаками текстовых описаний.

Комплексирование нескольких моделей. Некоторые задачи, являются достаточно сложными, например, сопровождение масок сегментации в видеопоследовательности. Для таких задач дорого размечать большие наборы данных, поэтому сегодня они небольшие в сравнении с наборами, предназначенными для обработки отдельных изображений. Набор данных напрямую влияет возможность модели работать с новыми данными и новыми условиями съемки. Чаще всего при наличии малых наборов данных нейронных сети обладают низким качеством работы. Однако счет комплексирования моделей, обученных на больших наборах данных, можно решать сложные задачи без дополнительного обучения моделей. Разработанная система объединяет три ранее рассмотренные модели: SAM для выделения объектов, DeAOT для сопровождения, CLIP для анализа объектов. На рис. 1 приведена краткая схема комплексирования моделей. В момент инициализации SAM сегментирует объекты в первом кадре, затем с помощью модели CLIP определяются необходимые для сопровождения объекты. Они идентифицируются, то есть им присваивается уникальный ID для сопровождения. В рассмотренном примере выделены дорога, как зона для возможного перемещения и бордюр с газоном, как

потенциальная преграда на пути мобильного робота. Далее выделенные объекты сопровождаются с помощью DeAOT, который сегментирует объекты на последующих кадрах. Каждый N кадр (в текущей реализации N = 10) дополнительно обрабатывается с помощью SAM для коррекции предсказаний.

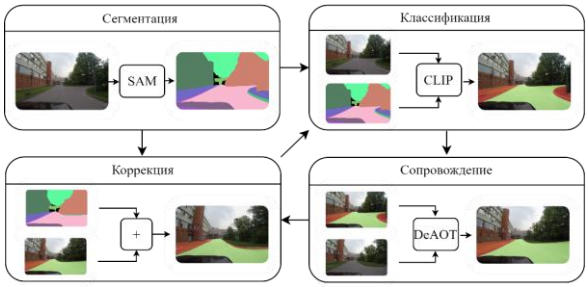

Рис. 1. Схема комплексирования нескольких моделей

Процедура коррекции решает следующие задачи: добавление новых объектов для сопровождения, валидация сегментированных масок объектов и уточнение сопровождаемых масок. Регулярно возникает ситуация, когда объекты на краях изображений некорректно сегментированы. Только при появлении большей части объекта на кадре удается понять границы его масок. Для решения этой проблемы реализован механизм валидации масок сегментации за счет сопровождения объектов в обратной последовательности. Если объект устойчиво сопровождается на нескольких кадрах, то, вероятно, он выделен корректно.

Выделение объектов из фона является достаточно простой операцией, поэтому детально не описывается. Если сегментированный SAM объект находится большей частью на фоне и не перекрывает другие объекты, то он идентифицируется и сопровождается на последующих кадрах как отдельный объект. Выделение новых объектов среди сопровождаемых является достаточно сложной задачей. Зачастую могут быть две крайности, когда несколько объектов выделены одной маской и когда один объект разделяется на несколько небольших областей. Реализованы определенные эвристические правила по выделению объектов на основе IoU (Intersection over Union) между масками сегментации и сопровождения. На рис. 2 приведен алгоритм сопоставления масок сегментации и сопровождения.

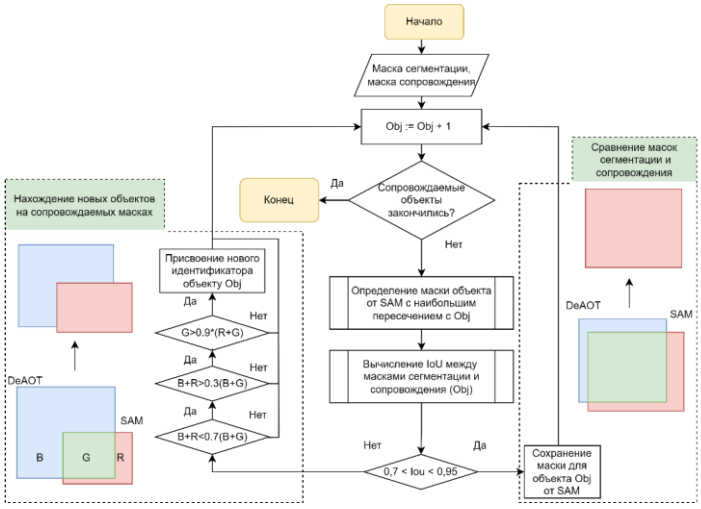

Рис. 2. Алгоритм сопоставления масок сегментации и сопровождения

Кроме того, SAM гораздо точнее определяет границы объектов в сравнении с DeAOT, поэтому при небольших различиях в предсказаниях масок одних и тех же объектов необходимо делать выбор в пользу масок SAM. Для этого реализован механизм сравнения масок при комплексировании объектов. Если значение IoU высоко, то маска корректируется в пользу сегментации SAM, который гораздо надежнее выделяет границы объектов.

Как упоминалось выше наиболее сложными задачами являются те, для которых нет репрезентативных наборов данных. Среди них можно выделить подводную съемку и изображение с камеры БЛА. На рис. 3 приведен пример работы разработанного алгоритма комплексирования моделей под водой для выделения протяженного объекта, а также с камеры БЛА, следующего за автомобилем [14]. Для БЛА с помощью CLIP среди всех масок определены дорога и автомобиль.

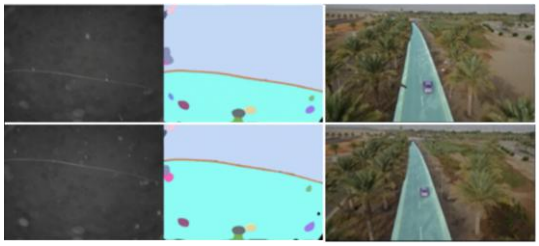

Рис. 3. Пример сегментации и сопровождения объектов

Проводить исследования на данных задачах достаточно сложно, так как отсутствует репрезентативное количество размеченных данных. Поэтому в настоящей работе эксперименты проводятся на видеороликах, сформированных для СТЗ мобильного робота.

Наборы данных. Для исследования эффективности и оценки качества работы моделей использовались наборы данных, полученные в условиях открытой местности и помещения. Для тестирования системы технического зрения в условиях открытой местности использовался обучающий набор данных для семантической сегментации. Разметка данных для семантической сегментации подразумевает установление каждому пикселю его класса. Как правило, для каждого класса объектов выделяется определенная интенсивность изображения. Все пикселы изображении приписываются к конкретному классу объектов. Разметка данных для семантической сегментации занимает большое количество времени. Для разметки использовался автоматизированный метод разметки для семантической сегментации [15, 16]. Он основан на нейросетевых моделях. Данный метод позволяет получить область, содержащую объект, используя лишь информацию о нажатиях по изображению. Пользователь предоставляет ввод в виде положительных и отрицательных нажатий. Обозначенные человеком точки с помощью предобученной модели возвращаются в виде размеченной маски, которую можно уточнить последующими нажатиями.

Участок, по которому мобильный робот может перемещаться, выделен в отдельный класс «дорога». Как правило, на данном участке в качестве перемещающихся преград выступают автомобили и люди, для которых также выделены отдельные классы «машина» и «человек» соответственно. Класс «инфраструктура» содержит небольшие объекты, например – клумбы, столбы. Для мобильного робота бордюр будет являться непреодолимой преградой, поэтому для него выделен отдельный класс «бордюр». Также существуют участки территории, которые похожи на дорогу, однако доступ к ним невозможен, например, если они расположены на уровне с бордюром, на который робот не сможет подняться. Для них также выделен отдельный класс под названием «пешеходная зона». Здания, деревья, га-

зон, небо выделены в соответствующие классы «здание», «дерево», «трава», «небо». Размеченный набор данных составил 172 изображения, которые получены в различное время года. Этот набор данных исследовался в задаче навигации мобильного робота в работе [17] с использованием семантической сегментации.

Для исследования качества моделей в помещении использованы 6 видеозаписей из набора данных VSPW [18], которые описаны в табл. 1.

Таблица 1

Описание набора данных в помещении

ID	Количество объектов	Описание сцены	Движение камеры
72	7	Картинная галерея с людьми	Вращение и перемещение
84	14	Гостиничный холл без людей	Вращение и перемещение
132	11	Аудитория в школе с детьми и учителем	Камера статична
1151	23	Перемещение с камерой по комнате	Вращение и перемещение
1153	20	Перемещение с камерой по комнате с человеком	Вращение и перемещение
1410	8	Человек в библиотеке	Камера статична

Для каждого кадра из видеопоследовательности имеется разметка для семантической сегментации. Стоит отметить, что в большинстве видеороликов производится перемещение и вращение камеры. Каждый видеоролик обладает своим набором классов, они варьируются от 7 до 23. Большинство сцен содержат перемещающихся людей, что усложняет работу модели сопровождения.

Экспериментальное исследование моделей. Особенность сегментации методом SAM заключается в достаточно большом количестве получаемых масок, которые могут разделять один объект на составные элементы. При ручной разметке объектов автомобиль выделяется одной маской. Для возможности сопоставления масок, полученных с помощью SAM и размеченных вручную, реализован алгоритм сопоставления масок, суть которого состоит в автоматическом слиянии масок, находящихся в пределах одного размеченного объекта. На рис. 4 приведен пример слияния масок, полученных с помощью SAM. Также для сравнения приведена ручная разметка.

После слияния масок можно рассчитать метрики качества сегментации. В качестве метрик использовались IoU и Dice. Для каждого класса значение вычисленных метрик для оценки качества сегментации в условиях открытой местности приведено в табл. 2. Из данной таблицы видно, что качество сегментации достаточно высокое.

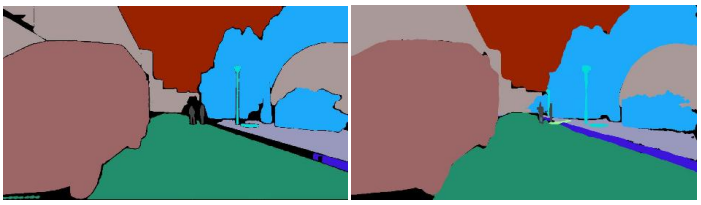

Рис. 4. Сравнение сегментации SAM (слева) с ручной разметкой (справа) после слияния масок

Наибольшее количество ошибок замечено у объектов малых размеров. В частности, на людях и объектах инфраструктуры, таких как фонарные столбы, клумбы или бордюр. При увеличении размеров объектов в кадре, например, при приближении человека к камере, качество сегментации резко возрастает. Это связано с тем фактом, что

разметка SAM (генерация областей сегментации) производится не по всему кадру, а от заданных исходных точек, которые располагаются в области кадра равномерной сеткой 20 на 20 точек. Если точка попала в малый объект, то он будет обнаружен.

Таблица 2

Результат сегментации и сопровождения на открытой местности

Наименование класса	Сегментация		Сопровождение	
	Dice, %	IoU, %	Dice, %	IoU, %
человек	35,02	23,01	31,69	21,08
дорога	97,9	95,95	96,36	94,6
трава	87,99	81,76	86,87	79,68
здание	94,59	90,2	89,46	83,17
небо	92,91	89,36	77,32	73,05
дерево	90,1	84,03	90,73	85,51
инфраструктура	72,45	62,32	57,5	47,12
машина	88,62	85,4	84,58	81,62
пешеходная зона	79,79	77,17	78,54	75,15
бордюр	77,43	67,34	77,43	67,34

Для видеопоследовательностей в помещении количество классов достаточно велико, поэтому в табл. 3 приведены усредненные по классам значения для каждой видеопоследовательности. Как и в случае с сегментацией на открытой местности, основную сложность представляют объекты небольшого размера. Общее качество сегментации с помощью SAM высокое, поэтому он может быть применен в рамках поставленной задачи.

Таблица 3

Результат сегментации и сопровождения в помещении

Идентификатор	Сегментация		Сопровождение	
	Dice, %	IoU, %	Dice, %	IoU, %
72	89,71	84,77	93,9	92,33
84	86,7	80,45	94,62	92,35
132	83,92	74,91	95,1	91,8
1151	84,8	82,09	88,79	87,3
1153	84,78	82,59	86,3	84,86
1410	81,47	73,98	97,65	95,48

Достаточно сложно полноценно оценить качество работы модели сопровождения на всей видеопоследовательности, так как набор данных сформирован не для видео, а для отдельных кадров. Чтобы решить данную проблему использовался следующий прием. Бралась пара размеченных кадров с близким расположением камеры. Затем один использовался в качестве размеченного, а на другом происходило сопровождение и подсчет метрики, затем процедура повторялась в обратном порядке. После этого полученный результат проверялся. Таким образом было выбрано и обработано 50 пар, то есть 100 изображений, для которых был выполнен подсчет метрик. В табл. 2 приведены рассчитанные метрики Dice и IoU. Для большинства классов были получены достаточно высокие значения, от 70% до 80% точности. Выделяются только два класса, «человек» и «инфраструктура». Первый содержит пиксели для людей, второй небольшие бетонные клумбы вдоль дороги, а также осветительные столбы. Их низкий результат связан с тем, что в силу правил подсчета метрик для каждого класса цена ошибки тем выше, чем его площадь на изображении ниже. Люди и инфраструктурные объекты занимают на изображениях настолько малую площадь, что очень небольшие ошибки приводят к существенному падению точности.

Видеозаписи в помещении имеют разметку для каждого кадра. Основная проблема состоит в том, что при перемещении камеры неизбежно появляются новые объекты, о которых модель ничего не знает, но которые будут вносить ошибку при пря-

мом сравнении с разметкой по соответствующим метрикам. Их необходимо как-то отслеживать. Чтобы модель могла их отслеживать, при их первом появлении в кадре (при изменении состава объектов в кадре) необходимо переинициализироваться этим кадром, и на его основе происходит сопровождение последующих кадров. Результаты расчета соответствующих метрик приведены в табл. 3. Для всех последовательностей выбранного набора данных DeAOT показал очень высокие показатели, что демонстрирует его применимость для решения задач как снаружи, так и внутри помещений.

Классификация объектов. Каждый отдельный объект вырезается по ограничивающему прямоугольнику и классифицируется с помощью CLIP. Однако, при этом на полученном изображении не гарантирована единственность объекта. На рис. 5 слева вверху приведен пример, на котором видно, что вместе со зданием в кадре также сохранились дорога, небо и трава, которые могут повлиять на результат классификации. Чтобы явным образом обеспечить наличие только одного объекта, в работе рассмотрено несколько способов подавления фона: размытые, обнуление, заливка белым цветом. На рис. 5 приведен пример обработки каждым из способов: слева вверху исходное изображение, справа вверху размытие, слева внизу обнуление, справа внизу заливка белым цветом. Сравнение результатов обработки приведено в табл. 4.

Рис. 5. Иллюстрация различных способов предобработки изображений

Из табл. 4 видно, что наибольшая точность классификации достигается при использовании размытия фона, поэтому при комплексировании используется данный метод предобработки данных.

Таблица 4

Результаты классификации объектов с различными способами предобработки изображений

Способ предобработки	Точность, %
Без предобработки	83,27
Размытие фона	88,61
Обнуление фона	75,98
Заливка белым цветом	76,33

Для классификации изображений при помощи CLIP необходимо составить словарь синонимов. Одного слова на класс может быть недостаточно, так как оно может не совсем точно описывать конкретный объект. Дополнительная проблема состоит в том, что заранее неизвестен набор данных, с помощью, которого обучался CLIP. Поэтому узнать на каких именно словах обучалась модель (и какие позволят получить наилучшие результаты) невозможно. Для каждого из набора данных составлен словарь синонимов для каждого класса, содержащий по 5-10 понятий. Качество классификации при этом составляет выше 90% для большинства объектов. Основные ошибки при классификации составляют по-прежнему объекты малого размера.

Быстродействие. Поскольку разработанное решение содержит в себе сразу 3 модели, то вопрос вычислительной нагрузки стоит остро. При этом на практике ее можно регулировать в зависимости от поставленной задачи. Например, можно менять количество и положение точек для сегментации в SAM или разрешение изображений. Измерение быстродействия производилось с помощью усреднения времени обработки 50 кадров, так как сегментация обрабатывала только каждый 10 кадр, а остальные сопровождались. Все вычисления производились на ПК с GPU GeForce RTX 2080Ti. Результаты измерений только для сопровождения и быстродействия приведены в табл. 5. Процедура классификации масок не приводится, так как зависит от изначальной постановки задачи и вычислителя.

Таблица 5

Быстродействие сегментации и сопровождения

Разрешение	Количество точек на сторону	Потребляемая память, ГБ	Время на обработку кадра, с
1280 x 720	20	7,4	5,36
	10	6,8	3,96
	5	5,3	2,46
640 x 360	20	5	0,86
	10	4	0,7
	5	2,3	0,62

Можно заметить, что разработанное решение не позволяет обрабатывать видео в реальном времени. В текущей реализации модели достаточно сложно развернуть на встраиваемом вычислителе. Однако сегодня существует уже несколько научных публикаций по оптимизации вычисления SAM [19, 20].

Заключение. По результатам работы можно сделать следующие выводы. Предложенный в работе алгоритм комплексирования позволяет объединять предсказания нескольких моделей для решения более сложных задач СТЗ. Основная особенность предложенного решения заключается в необязательности обучения и как следствие использование СТЗ в не детерминируемых условиях. Подсчитаны метрики при решении задач: сегментации, сопровождении, классификации. Результаты экспериментальных исследований подтверждают эффективность моделей. При их комплексировании удается получать эффективную СТЗ для решения различных более сложных задач. Основным направлением дальнейших исследований является ускорение быстродействий для достижения обработки в реальном времени.

Результаты получены в рамках выполнения государственного задания Минобрнауки России №075-00697-24-00 от 27.12.2023 «Исследование методов создания самообучающихся систем видеонаблюдения и видеоаналитики на базе комплексирования технологий пространственно-временной фильтрации видеопотока и нейронных сетей» (FNRG 2022 0015 1021060307687-9-1.2.1).

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Yang J., et al*. Track anything: Segment anything meets videos, *CoRR*, 2023, Vol. abs/2304.11968. Available at: http://arxiv.org/abs/2304.11968.
2. *Cheng H.K., et al.* Tracking anything with decoupled video segmentation, *IEEE/CVF International Conference on Computer Vision,* 2023, pp. 1316-1326.
3. *Zhu J., et al.* Tracking anything in high quality, *CoRR*, 2023, Vol. abs/2307.13974. Available at: http://arxiv.org/abs/2307.13974.
4. *Liu Y., et al.* MobileSAM-Track: Lightweight One-Shot Tracking and Segmentation of Small Objects on Edge Devices, *Remote Sensing,* 2023, Vol. 15, No. 24, pp. 5665.
5. *Cheng Y., et al.* Segment and track anything, *CoRR*, 2023, Vol. abs/2305.06558. Available at: http://arxiv.org/abs/2305.06558.
6. *Kirillov A., et al.* Segment anything, *CoRR*, 2023, Vol. abs/2304.02643. Available at: http://arxiv.org/abs/2304.02643.

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. *Yang Z., Yang Y.* Decoupling features in hierarchical propagation for video object segmentation, *Advances in Neural Information Processing Systems,* 2022, Vol. 35, pp. 36324-36336.
9. *Yang Z., Wei Y., Yang Y.* Associating objects with transformers for video object segmentation, *Advances in Neural Information Processing Systems*, 2021, Vol. 34, pp. 2491-2502.
10. *Cherti M., et al.* Reproducible scaling laws for contrastive language-image learning, *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2023, pp. 2818-2829.
11. *Awadalla A., et al.* Openflamingo: An open-source framework for training large autoregressive vision-language models, *CoRR*, 2023, Vol. abs/2308.01390. Available at: http://arxiv.org/abs/2308.01390.
12. *Li J., Li D., Xiong C., Hoi S.* Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, *International Conference on Machine Learning,* 2022, pp. 12888-12900.
13. *Radford A., et al.* Learning transferable visual models from natural language supervision, *International conference on machine learning,* 2021, pp. 8748-8763.
14. *Mueller M., Smith N., Ghanem B.* A benchmark and simulator for uav tracking //Computer Vision–ECCV 2016: 14th European Conference. 2016. – P. 445-461.
15. Github: fbrs_interactive_segmentation. Available at: https://github.com/SamsungLabs/fbrs_interactive_segmentation.
16. *Sofiiuk K., Petrov I., Barinova O., Konushin A*. F-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation, *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2020, pp. 8623-8632.
17. *Fomin I., Arhipov A.* Selection of Neural Network Algorithms for the Semantic Analysis of Local Industrial Area, *International Russian Automation Conference,* 2021, pp. 380-385.
18. *Miao J., et al.* VSPW: A Large-scale Dataset for Video Scene Parsing in the Wild, *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2021, pp. 4133-4143.
19. *Zhang C., et al.* Faster Segment Anything: Towards Lightweight SAM for Mobile Applications, *CoRR*, 2023, Vol. abs/2306.14289. Available at: http://arxiv.org/abs/2306.14289.
20. *Wang A., et al*. RepViT-SAM: Towards Real-Time Segmenting Anything, *CoRR*, 2023, Vol. abs/2312.05760 Available at: http://arxiv.org/abs/2312.05760.

Статью рекомендовал к опубликованию к.т.н. Л.А. Станкевич.

Архипов Андрей Евгеньевич – Государственный научный центр РФ – Федеральное государственное автономное научное учреждение «Центральный научно-исследовательский и опытно-конструкторский институт робототехники и технической кибернетики»; e-mail: a.arkhipov@rtc.ru; г. Санкт-Петербург, Россия; тел.: +78125523351; м.н.с.

Фомин Иван Сергеевич – e-mail: i.fomin@rtc.ru; м.н.с.

Матвеев Виктор Дмитриевич – e-mail: v.matveev@rtc.ru; инженер.

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

Fomin Ivan Sergeevich – e-mail: i.fomin@rtc.ru; junior researcher.

Matveev Victor Dmitrievich – e-mail: v.matveev@rtc.ru; engineer.

УДК 004.383 DOI 10.18522/2311-3103-2024-1-257-267

Н.А. Бочаров, И.Н. Бычков, П.В. Коренев, Н.Б. Парамонов

ЖИВУЧЕСТЬ БОРТОВЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ НАЗЕМНЫХ РОБОТОТЕХНИЧЕСКИХ КОМПЛЕКСОВ

Исследования в области создания специализированных вычислительных комплексов для робототехнических комплексов (РТК) ведутся во многих мировых научных центрах и в том числе в нашей стране. Развитие возможностей сенсорных систем, систем глобальной нави-