СЕКЦИЯ «СИСТЕМЫ И СЕТИ ИНФОКОММУНИКАЦИЙ»

УДК 004.932

ОЦЕНКА ЭФФЕКТИВНОСТИ СОВРЕМЕННЫХ ДЕТЕКТОРОВ ОБЪЕКТОВ В ЗАДАЧАХ ОДНОВРЕМЕННОЙ ЛОКАЛИЗАЦИИ, КАРТОГРАФИРОВАНИЯ И ТРЕКИНГА (SLAMOT)

Левоненко И.И., Робачевский А.Д., *магистранты гр. 567001*

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Цветков В.Ю. – д-р техн. наук, профессор

Аннотация. В работе проводится сравнительный анализ современных детекторов объектов, применяемых в задачах одновременной локализации, картографирования и трекинга объектов (SLAMOT). Цель статьи – определить, какие архитектуры детекции обеспечивают наилучший баланс между точностью, скоростью и устойчивостью в условиях динамических сцен, характерных для мобильной робототехники и автономных систем. В качестве сравниваемых моделей рассматриваются одностадийные и многостадийные детекторы: семейство YOLO, SSD, Faster R-CNN, а также OWL как представитель современных универсальных моделей обнаружения. Для каждой архитектуры оцениваются метрики качества детекции, вычислительная эффективность, способность корректно отслеживать объекты во времени и влияние на стабильность SLAMOT-пайплайна. Результаты экспериментов позволяют определить наиболее подходящие модели для интеграции в системы SLAMOT и сформулировать рекомендации по выбору детектора в зависимости от требований к производительности и точности.

Ключевые слова. SLAMOT, детекция объектов, компьютерное зрение, одностадийные и многостадийные детекторы, оценка эффективности, автономные системы.

Введение

В условиях стремительного развития автономных робототехнических систем методы искусственного интеллекта перестали быть исключительно теоретическими моделями и превратились в ключевой инструмент обеспечения восприятия и навигации в реальном мире. Главную роль в этом процессе играют модели компьютерного зрения на основе глубоких нейронных сетей, способные извлекать высокоуровневые признаки из визуальных данных и выявлять сложные нелинейные зависимости, недоступные традиционным алгоритмам обработки изображений. Однако для достижения полноценной автономности роботам недостаточно лишь распознавать объекты или строить карту окружающего пространства. Необходима интеграция этих процессов в рамках задач одновременной локализации, картографирования и трекинга объектов (SLAMOT), где детекция объектов становится не вспомогательной функцией, а критическим компонентом системы восприятия [1,2].

Использование нейронных сетей в SLAMOT-системах позволяет отказаться от жёстко заданных алгоритмических правил и опираться на обученные модели, способные адаптироваться к динамическим, частично наблюдаемым и слабоструктурированным условиям реального мира. Актуальность исследования определяется необходимостью поиска оптимального компромисса между точностью детектирования, устойчивостью трекинга и вычислительной эффективностью, особенно в условиях ограниченных ресурсов бортовых вычислительных платформ. Одностадийные детекторы, такие как YOLO [3] и SSD [4], демонстрируют высокую скорость работы, что делает их привлекательными для мобильных роботов. Однако их влияние на стабильность SLAMOT-пайплайна и качество построения карт требует систематического анализа. Многостадийные архитектуры, например Faster R-CNN [5], обеспечивают более высокую точность, но уступают в производительности, что также ставит вопрос об их применимости в реальном времени. Дополнительный интерес представляет использование современных универсальных моделей, таких как OWL-ViT [6], способных выполнять zero-shot детекцию и расширять возможности системы без дополнительного обучения. Рассмотрение подобных моделей в сравнительном анализе даёт возможность исследовать эффективность подходов, поддерживающих обнаружение объектов вне заранее заданного набора категорий, и оценить их влияние на адаптивность SLAMOT-систем к новым или ранее неизвестным объектам. Это особенно важно в контексте автономных платформ, работающих в непредсказуемых и быстро меняющихся условиях окружающей среды.

Целью настоящей работы является комплексная оценка эффективности современных детекторов объектов в контексте их интеграции в SLAMOT-системы. Для достижения этой цели анализируется влияние архитектурных особенностей детекторов на устойчивость локализации, проводится классификация подходов к включению семантической информации в карты окружения, исследуются метрики производительности в условиях ограниченных вычислительных ресурсов, а

также рассматриваются современные программные инструменты для развертывания нейросетевых моделей на бортовых вычислительных устройствах.

Теоретические аспекты семантического восприятия в задачах SLAMOT

Семантическое восприятие является критически важным компонентом современных автономных систем, поскольку позволяет робототехнической платформе переходить от геометрического представления пространства к его интерпретации на уровне объектов. В математической формулировке задача семантического восприятия заключается в построении отображения:

$$\psi: I \to S, \tag{1}$$

где I – пространство входных сенсорных данных (изображений); а S – пространство семантических меток.

В контексте SLAMOT (Simultaneous Localization, Mapping, and Object Tracking) данная интерпретация становится фундаментом для повышения устойчивости локализации за счёт фильтрации динамических объектов и улучшения ассоциации данных. Результатом работы детектора объектов является набор кортежей

$$D = \{(b_i, c_i, p_i)\}_{i=1}^{N}, \tag{2}$$

где $b_i = (x_i, y_i, w_i, h_i)$ – координаты ограничивающей рамки; c_i – метка класса; p_i – вероятность обнаружения.

Эффективность последующих процессов трекинга и построения семантической карты напрямую зависит от точности и частоты обновления множества D, что определяется архитектурой нейронной сети.

Современные детекторы объектов эволюционируют в направлении оптимизации компромисса между точностью и вычислительной сложностью, разделяясь на одностадийные и многостадийные архитектуры. Одностадийные модели, к классу которых относятся семейства YOLO и SSD, реализуют принцип плотного прогнозирования (dense prediction), выполняя регрессию координат и классификацию объектов за один прямой проход сети. Функция потерь в таких архитектурах обычно формируется как взвешенная сумма ошибок классификации и локализации:

$$L = \lambda_{cls} L_{cls} + \lambda_{loc} L_{loc}, \tag{3}$$

где L_{loc} часто базируется на метрике *IoU* (Intersection over Union):

$$IoU(b, \hat{b}) = \frac{|b \cap \hat{b}|}{|b \cup \hat{b}|}, \tag{4}$$

или её модификациях CIoU/DIoU, улучшающих сходимость регрессии границ. Преимуществом данного подхода является линейная вычислительная сложность относительно размера входного тензора и минимальная задержка вывода, что делает такие модели предпочтительными для мобильных роботов с ограниченным энергопотреблением. Однако отсутствие этапа верификации регионов может приводить к снижению точности на сложных сценах с высокой окклюзией, что потенциально влияет на устойчивость трекинга в динамической среде.

Многостадийные архитектуры, представленные семейством R-CNN (в частности, Faster R-CNN), используют двухэтапный процесс, включающий генерацию регионов интереса (Region Proposal Network – RPN) и последующую классификацию предложенных областей. Математически это разделяет задачу на минимизацию функции потерь предложений

$$L_{RPN} = L_{cls}^{RPN} + L_{loc}^{RPN}, \tag{5}$$

и функцию потерь детекции:

$$L_{det} = L_{cls}^{det} + L_{loc}^{det}, \tag{6}$$

что позволяет достигать более высокой точности за счёт анализа регионов с высоким разрешением. Такие модели демонстрируют высокую эффективность при детекции мелких объектов и в условиях сложного фона благодаря механизму ROI Align. Однако их вычислительная сложность существенно выше из-за последовательного характера обработки и операций выборки регионов. В системах SLAMOT реального времени использование таких архитектур часто ограничено частотой кадров, но

они служат важным ориентиром при сравнительном анализе, позволяя оценить предельное качество детектирования при наличии достаточных вычислительных ресурсов.

Отдельное направление развития представляют универсальные модели открытого словаря (open-vocabulary), такие как OWL-ViT, основанные на трансформерных архитектурах и механизме самовнимания (Self-Attention). Ключевым элементом данных моделей является механизм внимания, вычисляемый как

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (7)$$

где Q – матрица запросов (*query*), содержащая векторы, описывающие, какую информацию каждый элемент последовательности стремится получить от других элементов; K – матрица ключей (*key*), представляющая векторы признаков, по которым определяется релевантность элементов относительно запросов; V – матрица значений (*value*), содержащая информацию, которая агрегируется на выходе механизма внимания; QK^T – матрица скалярных произведений запросов и ключей, определяющая степень соответствия между элементами последовательности; $\sqrt{d_k}$ – коэффициент масштабирования, где d_k – размерность векторов ключей; используется для стабилизации значений softmax и предотвращения чрезмерного роста градиентов; softmax(·) – нормализующая функция, преобразующая оценки внимания в вероятностное распределение весов;

Это позволяет модели устанавливать глобальные зависимости между участками изображения и текстовыми эмбеддингами классов без фиксации конечного набора категорий. Способность работать с текстовыми запросами и обнаруживать новые категории через вычисление косинусного сходства в пространстве признаков делает такие модели перспективными для SLAMOT-систем, функционирующих в непредсказуемых условиях. Несмотря на высокие вычислительные затраты, возможность адаптации к новым объектам без дообучения (zero-shot detection) открывает новые возможности для расширения семантического картографирования в долгосрочной перспективе.

Экспериментальная оценка эффективности детекторов объектов в SLAMOT-системах

Экспериментальная оценка эффективности детекторов в конвейере SLAMOT проведена для моделей, представляющих различные архитектурные парадигмы: одностадийных детекторов YOLOv8n и SSD (lite320/300), многостадийной архитектуры Faster R-CNN ResNet50 FPN, а также трансформерной модели открытого словаря OWL-ViT. Выбор данного набора обусловлен необходимостью анализа компромисса между вычислительной сложностью и точностью семантического восприятия в условиях ограниченных ресурсов мобильных платформ. Сравнение архитектур осуществляется по совокупности измеряемых характеристик: время инференса (inference latency), характеризующее быстродействие модели при обработке единичного кадра; потребление памяти (memory footprint), отражающее объём оперативной и видеопамяти, необходимый для выполнения прямого прохода; и размер модели (model size / parameters count), определяющий требования к хранению и передаче весовых коэффициентов. Данные метрики позволяют количественно оценить вычислительную сложность каждой архитектуры и её пригодность для развёртывания на ресурсоограниченных платформах, используемых в системах SLAMOT.

Непосредственное тестирование проводилось в рамках стандартизированного протокола, обеспечивающего воспроизводимость результатов. Первоначальный порог уверенности (confidence threshold) устанавливался на уровне 0.5, что соответствовало минимальной вероятности 50% корректной классификации объекта согласно выходному распределению модели. В случае отсутствия детекций, удовлетворяющих данному критерию, применялась адаптивная стратегия понижения порога с дискретным шагом до достижения содержательно валидных результатов, при этом нижняя граница поиска фиксировалась на нулевом значении.

В качестве основных метрик временной эффективности регистрировались медианное время инференса, характеризующее центральную тенденцию задержек обработки единичного кадра, а также робастная оценка вариативности – медианное абсолютное отклонение (MAD, Median Absolute Deviation), дополненное экстремальными значениями минимального и максимального времени выполнения. Указанные статистики вычислялись после предварительной фазы стабилизации (warmup), включавшей 10 итераций прогона модели на репрезентативном входном тензоре, что позволяло минимизировать влияние кэширования, динамической компиляции и инициализации вычислительного графа на итоговые замеры.

Характеристики ресурсоёмкости моделей включали пиковое потребление оперативной памяти (RAM) и видеопамяти (VRAM), фиксируемое в процессе выполнения прямого прохода, общее количество обучаемых параметров, а также оценочный размер модели в мегабайтах. Для архитектуры OWL-ViT размер вычислялся теоретически на основе произведения количества параметров на размерность представления весов в формате float32 (4 байта), тогда как для остальных детекторов применялся прямой замер размера сериализованного файла весов. Все замеры памяти отражали максимальное потребление в течение цикла инференса, что обеспечивало

консервативную оценку требований к аппаратным ресурсам при развёртывании в условиях ограниченной памяти.

На рисунке 1 приведены изображения, использованные для тестирования нейросетевых моделей.

Рисунок 1 – Тестовые изображения

На рисунке 2 представлен результат обработки используя Faster R-CNN ResNet50 FPN.

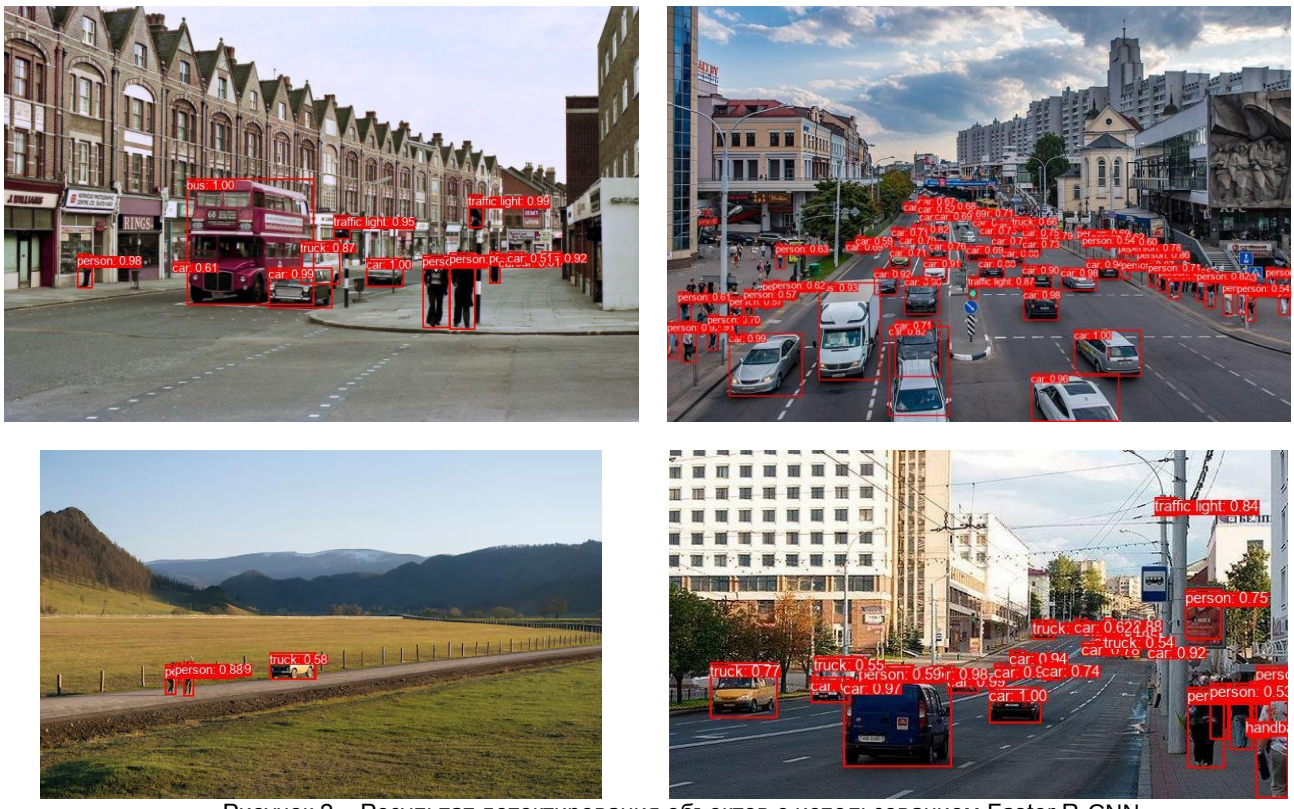

Рисунок 2 – Результат детектирования объектов с использованием Faster R-CNN.

На рисунке 3 представлен результат обработки изображений используя OWL-ViT.

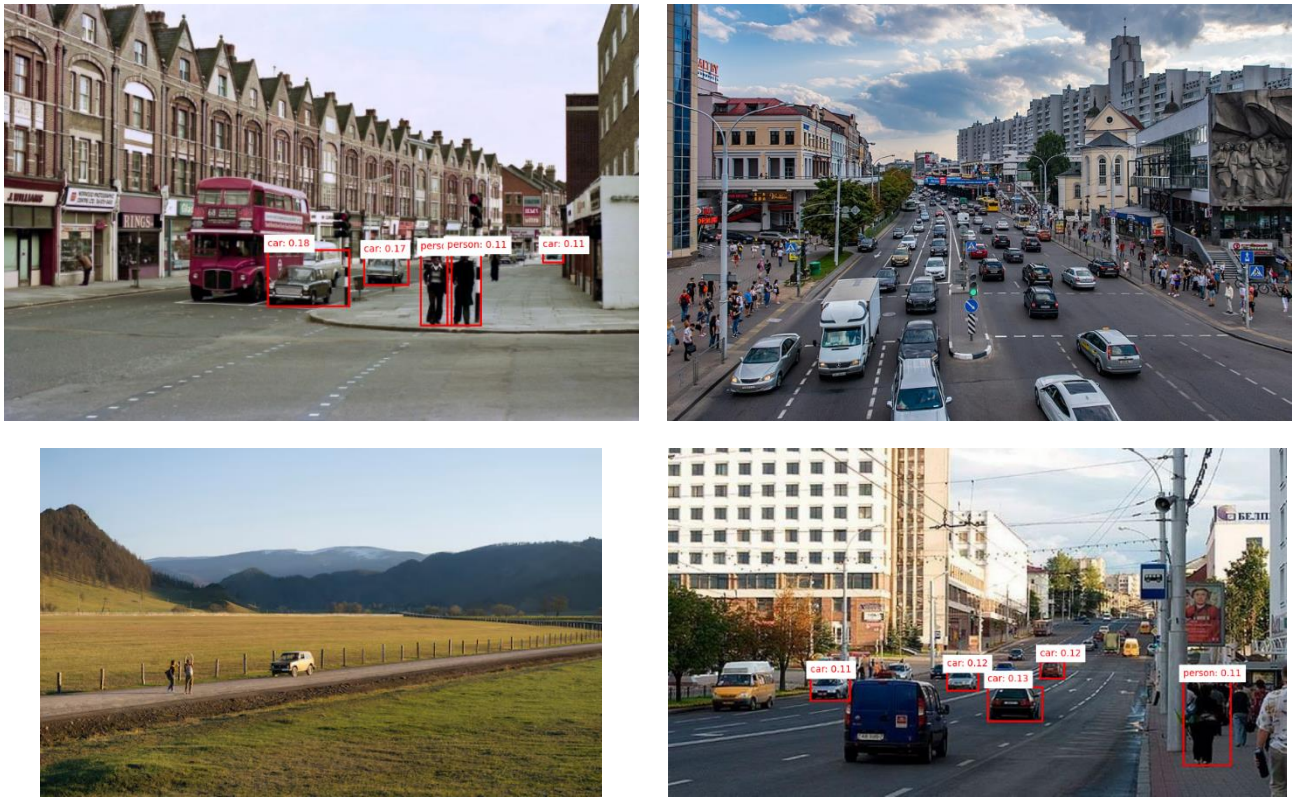

Рисунок 3 – Результат детектирования объектов с использованием OWL-ViT

На рисунке 4 представлен результат обработки изображений используя YOLOv8n.

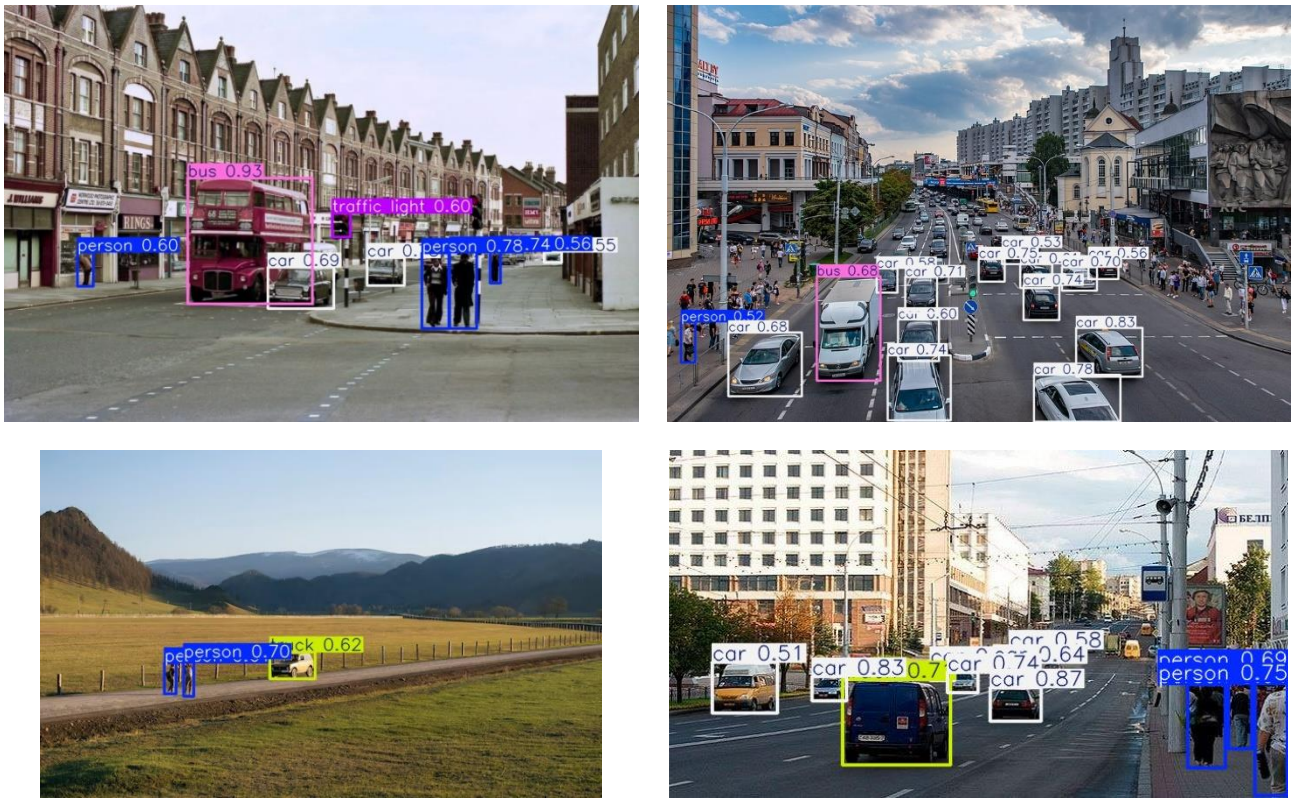

Рисунок 4 – Результат детектирования объектов с использованием YOLOv8n

На рисунке 5 представлен результат обработки изображений используя SSD (lite320).

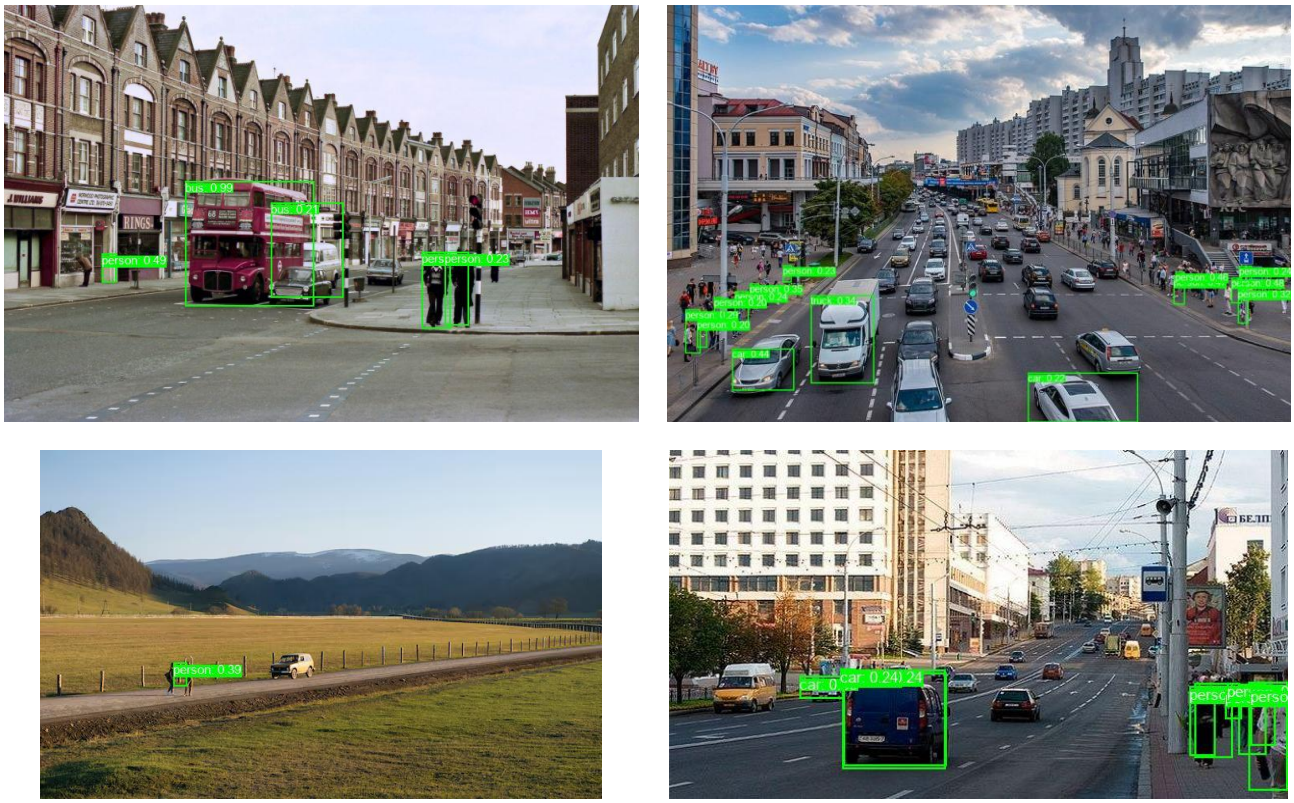

Рисунок 5 – Результат детектирования объектов с использованием SSD (lite320)

На рисунке 6 представлен результат обработки изображений используя SSD (300).

Рисунок 6 – Результат детектирования объектов с использованием SSD (300)

В таблице 1 обобщены ключевые метрики производительности исследуемых моделей детекции объектов.

Таблица 1 – Сводная таблица результатов детекторов объектов

Модель	Архитектура	Порог уверенности	Время инференса (медиана), с	Количество параметров, 10^6	Размер, МБ	Потребление RAM, МБ
YOLOv8n	One-stage	0,50	0,0198	3,16	6,25	672,30
SSDLite320	One-stage	0,20	0,0574	3,44	13,41	678,41
SSD300	One-stage	0,20	0,2080	35,64	135,99	891,25
OWL-ViT	Vision Transformer	0,00	0,3213	153,23	584,53	1212,10
Faster R-CNN	Two-stage	0,20	1,2483	41,76	159,79	1617,13

Экспериментальные замеры производительности пяти архитектур детекции объектов, выполненные на центральном процессоре при обработке четырех изображений схожего разрешения, продемонстрировали выраженную иерархию вычислительной эффективности, коррелирующую с архитектурной сложностью моделей.

Наиболее высокую скорость инференса показала модель YOLOv8n с медианным временем обработки кадра 19.8 мс и минимальным отклонением (MAD = 1.6 мс), что подтверждает эффективность одностадийных детекторов с оптимизированной архитектурой для задач реального времени. При этом модель характеризуется наименьшим объёмом параметров (3.16 млн) и размером весов (6.25 МБ), а также умеренным потреблением оперативной памяти (672 МБ). Качественный анализ результатов детекции выявил, что YOLOv8n обеспечивает приемлемый уровень локализации объектов, достаточный для большинства практических задач, несмотря на отдельные случаи пропуска малоразмерных или частично окклюзированных целей.

Модели семейства SSD продемонстрировали ожидаемый компромисс между точностью и быстродействием: облегчённая архитектура SSDLite320 MobileNetV3 обеспечила время инференса 57.4 мс при сохранении компактности (3.44 млн параметров), тогда как более ёмкая версия SSD300 VGG16 потребовала 208 мс на кадр, что объясняется увеличением глубины сети и количества параметров до 35.64 млн. В аспекте качества детекции модель SSD300 VGG16 превзошла облегчённую модификацию, демонстрируя более стабильное выделение объектов сложной формы, однако уступила YOLOv8n в точности локализации границ. Архитектура SSDLite320 заняла промежуточное положение, показав результаты, сопоставимые с базовыми требованиями к системам навигации в условиях ограниченных вычислительных ресурсов.

Трансформерная модель открытого словаря OWL-ViT, несмотря на значительное количество параметров (153.23 млн) и наибольший расчётный размер весов (584.53 МБ), показала время инференса 321.3 мс, что существенно быстрее двухстадийной архитектуры Faster R-CNN. Однако качественная оценка результатов детекции выявила существенные ограничения данной модели в рамках проведённого эксперимента: на тестовых изображениях наблюдалось критически низкое количество корректных детекций. Вероятной причиной данного явления выступает чувствительность архитектуры открытого словаря к формулировке текстовых запросов – семантическое несоответствие между промптами и визуальным содержанием кадров могло привести к подавлению большинства предсказаний механизмом пост-обработки. Таким образом, применение OWL-ViT требует тщательной инженерии запросов и, потенциально, адаптации пороговых стратегий под конкретную предметную область.

Наибольшее время обработки (1.248 с на кадр) и потребление ресурсов (1617 МБ RAM) зафиксировано для Faster R-CNN ResNet50 FPN, что согласуется с теоретическими ожиданиями для двухстадийных архитектур: необходимость последовательного выполнения региональной предложения и классификации существенно увеличивает вычислительную нагрузку. Тем не менее, именно данная модель продемонстрировала наивысшее качество детекции: точное выделение границ объектов, устойчивая работа при вариативном освещении и минимальное количество ложных срабатываний. Это подтверждает актуальность двухстадийных подходов для задач, где приоритетом является максимальная точность семантического восприятия, а временные задержки допустимы в рамках оффлайн-обработки или пост-анализа.

Все замеры производительности осуществлялись в условиях предварительной стабилизации вычислительного конвейера, для чего перед началом регистрации временных метрик выполнялась серия из десяти итераций прогона модели на репрезентативном входном тензоре (warmup-фаза). Данная процедура необходима для минимизации систематических погрешностей, вызванных «холодным» стартом: инициализацией вычислительного графа фреймворка, динамической компиляцией ядер (в случае использования бэкендов типа TorchScript или ONNX Runtime), заполнением кэшей процессора первого и второго уровня, а также установлением стационарного режима распределения оперативной памяти. Все замеры производительности осуществлялись в условиях предварительной стабилизации вычислительного конвейера, для чего перед началом

регистрации временных метрик выполнялась серия из десяти итераций прогона модели на репрезентативном входном тензоре (warmup-фаза). Данная процедура необходима для минимизации систематических погрешностей, вызванных «холодным» стартом: инициализацией вычислительного графа фреймворка, динамической компиляцией ядер (в случае использования бэкендов типа TorchScript или ONNX Runtime), заполнением кэшей процессора первого и второго уровня, а также установлением стационарного режима распределения оперативной памяти.

Заключение

В данной работе была произведена комплексная оценка эффективности современных архитектур детекции объектов в контексте их глубокой интеграции в сложные системы одновременной локализации, картографирования и трекинга (SLAMOT). Полученные экспериментальные данные демонстрируют выраженную прямую корреляцию между возрастающей архитектурной сложностью нейронных моделей и их снижающейся вычислительной эффективностью на стандартном оборудовании. Одностадийные детекторы, в частности легковесная версия YOLOv8n, обеспечивают наиболее оптимальный баланс высокой скорости обработки и приемлемой точности для критичных задач реального времени, требующих минимальной задержки отклика системы. В противоположность им, двухстадийные архитектуры, такие как Faster R-CNN ResNet50, демонстрируют наивысшее качество детекции и точность локализации границ, однако ценой существенного увеличения времени инференса и потребления оперативной памяти, что объективно ограничивает их применение исключительно сценариями оффлайн-анализа или постобработки данных на мощных серверах. Модели открытого словаря, представленные трансформером OWL-ViT, несмотря на значительную теоретическую гибкость в работе с произвольными текстовыми запросами, показали в текущих условиях низкую практическую эффективность без тщательной инженерии промптов и адаптации пороговых стратегий под конкретную предметную область. Кроме того, тестирование на центральном процессоре выявило критическую важность оптимизации памяти для всех моделей, так как избыточное потребление RAM может дестабилизировать работу основного конвейера SLAM. Таким образом, для развёртывания на ресурсоограниченных мобильных платформах и встраиваемых системах рекомендовано приоритетное использование лёгких одностадийных детекторов, в то время как более сложные архитектуры целесообразно применять для задач семантического обогащения карт, где приоритетом является максимальная точность восприятия окружающей среды, а не минимальная задержка обработки видеопотока.

Список использованных источников:

1. Cadena, C., Carlone, L., Carrillo, H. et al. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age // IEEE Transactions on Robotics. – 2016. – Vol. 32, No. 6. – P. 1309–1332.

2. Zhong, F., Wang, Y., Chen, X. et al. SLAMOT: A Survey on Simultaneous Localization, Mapping and Object Tracking // arXiv preprint arXiv:2303.09331. – 2023. – 42 p.

3. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 779–788.

4. Liu, W., Anguelov, D., Erhan, D. et al. SSD: Single Shot MultiBox Detector // European Conference on Computer Vision (ECCV). – 2016. – P. 21–37.

5. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks // Advances in Neural Information Processing Systems (NeurIPS). – 2015. – Vol. 28.

6. Minderer, M., Gritsenko, A., Stone, A. et al. Simple Open-Vocabulary Object Detection with Vision Transformers // European Conference on Computer Vision (ECCV). – 2022. – P. 728–755.

UDC 004.932

PERFORMANCE EVALUATION OF STATE-OF-THE-ART OBJECT DETECTORS FOR SIMULTANEOUS LOCALIZATION, MAPPING, AND OBJECT TRACKING (SLAMOT)

I.I. Levonenko, A.D. Robachevsky

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

V.Yu. Tsviatkou – Doctor of Engineering, professor

Annotation. This paper presents a comparative benchmark study of modern object detection architectures employed in Simultaneous Localization, Mapping, and Object Tracking (SLAMOT) pipelines. The primary objective is to identify which detection frameworks achieve the optimal trade-off between accuracy, inference latency, and robustness under dynamic scene conditions typical of mobile robotics and autonomous systems. The evaluated models encompass both single-stage detectors (YOLO family, SSD variants) and two-stage architectures (Faster R-CNN), alongside OWL-ViT as a representative of contemporary open-vocabulary, transformer-based detection paradigms. For each architecture, we assess detection quality metrics, computational efficiency (inference time, memory footprint, parameter count), temporal tracking consistency, and the impact on overall SLAMOT pipeline stability. Experimental results enable the identification of the most suitable models for integration into resource-constrained SLAMOT systems and provide actionable guidelines for detector selection based on application-specific requirements for real-time performance and semantic perception accuracy.

Keywords. SLAMOT, object detection, computer vision, single-stage and two-stage detectors, performance benchmarking, autonomous systems.