Trudy MAI. 2025. No. 145. (In Russ.)

Научная статья
УДК 004.8
URL: https://trudymai.ru/published.php?ID=186898
EDN: https://www.elibrary.ru/ZSPFMU

ПРИМЕНЕНИЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧЕ ВИЗУАЛЬНОЙ ОДОМЕТРИИ

П.Г. Корыткин✉

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова», г. Москва, Россия

✉ korytkinpg@my.msu.ru

Цитирование:. Корыткин П.Г. Применение глубоких нейронных сетей в задаче визуальной одометрии // Труды МАИ. 2025. № 145. URL: https://trudymai.ru/published.php?ID=186898

Аннотация. Задача повышения точности ориентации робототехнических комплексов и беспилотных летательных аппаратов сохраняет свою актуальность. Существующие решения, реализующие визуальную одометрию на основе алгоритмов, требуют ручной настройки параметров, а также чувствительны к освещённости и цвету. Научной новизной работы является применение нейронных сетей на этапе обработки изображений визуальной одометрии для выделения признаков. В работе рассматривается изучение применения нейронной сети, построенной на основе комбинации модифицированных архитектур MobileNet V2 и U-Net для выделения особых признаков на изображении. Выполнена модификация архитектуры U-Net – заменена транспонированная свёртка на комбинацию из слоёв, выполняющих: растяжение изображения, нормализацию, свёртку с функцией активации, свёртку, нормализацию, функцию активации для повышения качества обучения по метрике F1-Score. Для обучения нейронной сети подготовлено два датасета на основе 4 видеозаписей, из которых 2 синтетические и 2 записанные на камеру

видеорегистратора. Первый датасет состоял из отмасштабированных цветных кадров исходного изображения до разрешения 224x224x3, второй из квадратов фиксированного разрешения 128x128x3 полученных из исходного изображения. Для получения карты признаков для датасетов использовался алгоритм SIFT применяемый к изображениям видеозаписей для формирования чёрно-белой карты, где чёрный – отсутствие признака, белый – наличие признака. На этих датасетах обучалось 4 нейронных сети из которых 2 нейронные сети обучались на датасете состоящем из сегментов разрешения 64x64x3 и 128x128x3 для которых входные изображения – цветные. Одна на сегментах, отмасштабированных до разрешения 64x64x3 и преобразованных в чёрно-белый формат. Одна с цветными изображениями разрешения 224x224x3. Лучший результат по метрике F1-Score у нейросетевого детектора, работавшего с чёрно-белыми изображениями разрешения 64x64x3. Для апробирования выбрана простая система визуальной одометрии, в которую был встроен детектор особых признаков на основе разработанной нейронной сети. Выполнено апробирование полученной системы визуальной одометрии на датасете KITTI и сравнение с исходной системой визуальной одометрии, использующей детектор SIFT. Полученное программное решение показало свою работоспособность. В результате тестирования выявлено, что нейросетевой детектор находит большее число признаков, чем детектор SIFT. В одном из трёх маршрутов KITTI нейросетевой детектор показал превосходство. В двух других маршрутах выявлен дрейф и накопление ошибок, связанное с встречным трафиком при отсутствии движения транспортного средства, на котором установлена камера.

Ключевые слова: визуальная одометрия, нейронные сети, особенности изображений, SIFT, MobileNet V2, U-Net.

APPLYING DEEP LEARNING NETWORKS IN VISUAL ODOMETRY PROBLEMS

P.G. Korytkin ✉

Federal State Budget Educational Institution of Higher Education M.V. Lomonosov Moscow State University, Moscow, Russia

✉ korytkinpg@my.msu.ru

Citation: Korytkin P.G. Applying deep learning networks in visual odometry problems // Trudy MAI. 2025. No. 145. (In Russ.). URL: https://trudymai.ru/published.php?ID=186898

Abstract. Improving the orientation accuracy of robotic systems remains an important task. Existing algorithm-based visual odometry solutions require manual parameter adjustment and are sensitive to illumination and color. A promising approach to improving existing algorithmic visual odometry systems is to integrate neural networks into the image processing stages of visual odometry. This paper examines the application of a neural network built using a combination of the MobileNet V2 and U-Net architectures for image feature extraction. Two datasets were prepared for training the neural network, consisting of four video recordings. First dataset consisted of color frames of the original image scaled to a resolution of 224x224x3; second one consisted of squares of a fixed resolution of 128x128x3 obtained from the original frame image. To obtain a feature map for the datasets, the SIFT algorithm was applied to video images to form a black-and-white map, where black indicates the absence of a feature and white indicates the presence of a feature. Four neural networks were trained on these datasets, two of which were trained on a dataset consisting of 64x64x3 and 128x128x3 resolution segments respectively for which the input images were color. One was trained on segments scaled to 64x64x3 resolution and converted to black-and-white format, and the other with color images of 224x224x3 resolution. The best result in terms of F1 Score was achieved by the neural network detector working with black-and-white 64x64x1 images. A simple visual odometry system was modified to use neural network-based feature detector for testing. A combined visual odometry system with neural network-based feature detector was tested on the KITTI dataset and compared with the original visual odometry system using the SIFT detector. The resulting software solution demonstrated its applicability. It was found that the neural network detector identifies a larger number of features than the SIFT detector. On one of the three KITTI routes used for testing, the neural network detector demonstrated superiority. On the other two routes, drift and error accumulation were detected due to oncoming traffic while the vehicle carrying the camera was stationary.

Keywords: visual odometry, neural networks, image features, SIFT, MobileNet V2, U-Net.

Введение

Повышение автономности робототехнических комплексов (РТК) и беспилотных летательных аппаратов связано с необходимостью повышения точности позиционирования в пространстве. В настоящее время для позиционирования применяются комплексирование информации от инерциальных измерительных модулей (IMU) и глобальной системы позиционирования (GPS). Но такой подход имеет недостатки. Недостатком IMU является накопление ошибок со временем [1], чувствительность к проскальзыванию колёс [2], низкая точность маловесных IMU. Существенный недостаток GPS – невозможность применения в замкнутых пространствах, а также местах, где сигнал GPS может отсутствовать, подавляться, подменяться, переотражаться [3]. Поэтому остаётся актуальным поиск дополнительных источников информации о навигационных параметрах для комплексирования вместе с существующими.

Перспективное направление для применения, которое активно развивается – визуальная одометрия (ВО), которая выполняет вычисление угловых и линейных координат на основе набора изображений. Основными этапами ВО являются: захват изображений, поиск признаков, сопоставление признаков, оценка поворота и перемещения [4]. Существующие традиционные методы ВО, для поиска особых признаков, применяют некоторый алгоритм, который необходимо предварительно настроить, учитывая параметры освещённости и шумы изображений, получаемых с камеры для оптимальной работы в заданных условиях. Необходимость настройки параметров для системы, работающей в динамических условиях с изменчивым освещением и погодными условиями, является недостатком, усложняющим применение ВО. Популярными алгоритмами поиска признаков, которые применяются в системах ВО являются: SIFT, ORB. Данная область с применением алгоритмов активно развивается с момента появления нашумевшей статьи [5] Дэвида Нистера, Олега Народицкого и Джеймса Бергана, которая ввела такое понятие, как “visual odometry”, а также популяризовала визуальную одометрию для исследования, где описывается несколько алгоритмов определения изменения позиции робота на основе

монокулярной и стерео-видеопоследовательности. Основная идея алгоритма авторов, для монокулярной схемы, состояла в поиске особенностей на последовательности кадров, отслеживании относительной позиции между тремя кадрами, используя специальный алгоритм и RANSAC, уточняемый итерационно, после чего треки на основе триангуляции, преобразовывались в точки 3D пространства на основе первого и последнего наблюдения, и если алгоритм выполнялся не в первый раз, то новая реконструкция перемещалась в координатную систему предыдущей. На основе известных точек, вычислялось положение камеры. В стерео-видеопоследовательности, устанавливалось соответствие между левым и правым изображением, после чего соответствия триангулировались в точки 3D пространства, на основе найденных особенностей на определённом количестве кадров, а затем вычислялась позиция стереосистемы, с использованием упреждающего RANSAC и итерационным уточнением. Также в статье указывается, что абсолютная ориентация в стерео-видеопоследовательности, использующая несколько кадров, приводит к разрушительным ошибкам в силу низкой точности полученной карты глубины. В перемещении, состоящем из трёх кругов и суммарным расстоянием в 184 метра и возвращении в ту же точку, погрешность в расстоянии между конечными точками составила 4.1 метра. Важным является то, что в силу необходимости настройки данные алгоритмы чувствительны к изменениям среды, освещения, шумам на изображениях, что ограничивает их применение в сложных и изменчивых условиях, в которых обычно работают беспилотные летательные аппараты и робототехнические комплексы. Помимо классических алгоритмических подходов, существуют подходы, построенные на основе нейронных сетей, которые используя некоторую архитектуру нейронной сети позволяют в результате её применения к исходному изображению в комбинации с некоторыми параметрами на выходе получить угловые и линейные координаты или перемещение. Данные нейросетевые подходы как правило имеют более нескольких миллионов параметров так как на входе требуется обрабатывать изображение большого разрешения 1280x720 или 640x480 пикселей. Помимо этого, данные системы могут включать в себя различные

механизмы памяти для нейронных сетей, сохраняющие некоторую информацию в процессе прохождения маршрута, что позволяет повысить точность, а также реализовать замыкание петель маршрута для снижения дрейфа. Известным примером в данной области является DROID-SLAM [6]. При этом главным недостатком такого подхода является то, что требуется большой объём видеопамяти, как для обучения, так и применения данной системы, которого как правило нет в компактных устройствах по типу NVIDIA Jetson. Помимо этого, большую роль играет то, что в данный момент дополнительная память RAM и VRAM обходится значительно дороже, чем раньше из-за имеющегося дефицита на рынке электроники [7], что требует дополнительных оптимизаций и экономии памяти для снижения денежных затрат. В случае DROID-SLAM – это минимум 11 Гб VRAM при работе, а при обучении – 24 Гб, чего как правило нет на мобильных устройствах. Поэтому на данном этапе развития систем визуальной одометрии, построенных целиком на основе нейронных сетей, затруднено применение, при наличии ограничений, соответствующих мобильным системам: вычислительных, энергетических, массогабаритных. При этом в последнее время во многих устройствах уже доступны нейросетевые ускорители небольшой производительности, а также видеоускорители, которые позволяют запустить небольшие нейросетевые модели, как напрямую, так и с использованием слоёв совместимости (например, Vulkan), решающие некоторую узкоспециализированную задачу. Одним из успешных примеров таких комбинированных подходов является статья авторов [8], где автором удалось применить на беспилотном летательном аппарате комбинацию из архитектур долгой-короткой памяти (LSTM) и свёрточных слоёв (CNN) для повышения точности навигационных параметров в случае отсутствия сигнала GPS. Поэтому перспективным является применение существующих алгоритмических систем визуальной одометрии с нейросетевым детектором особенностей на изображениях, как дополнительного источника навигационных параметров [9], при этом менее чувствительного к изменению среды и освещения, а также не требующего ручной настройки и подбора параметров для выделения особых точек на изображении.

Цель работы – разработка детектора особых признаков на основе глубоких нейронных сетей и, исследование его применения для поиска особых точек на изображении в существующей системе ВО [10] и сравнение, полученных результатов, с применением алгоритма SIFT, для поиска особых точек, реализованного в фреймворке OpenCV [11].

Постановка задачи

В настоящий момент существует множество архитектур нейронных сетей. Как правило, объём требуемой видеопамяти зависит от числа параметров. Число параметров и математических операций значительно влияет на скорость обучения и применения глубокой нейронной сети. Помимо этого, у архитектур нейронных сетей имеются гиперпараметры, которые возможно настроить при подготовке архитектуры, влияя на итоговое поведение нейронной сети, число обучаемых параметров, а также скорость обучения, валидации и выполнения. Настройка гиперпараметров значительно влияет на итоговый результат [12]. В результате изучения информации об архитектурах кодировщиков серии MobileNet [13, 16, 17] и DeepLab [14, 15] решено выбрать кодировщик MobileNet V2 [16] для кодирования признакового пространства. Главное преимущество архитектур серии MobileNet состоит в оптимизации для применения в мобильных устройствах. Для декодирования признакового пространства выбрана архитектура U-Net [18], которая была доработана.

Поэтому итоговая нейросетевая архитектура строилась на основе применения комбинации архитектур MobileNet V2 с уменьшенным числом остаточных блоков и модифицированной архитектуры U-Net. Для реализации данной архитектуры использовался фреймворк TensorFlow [19,20] и язык программирования Python. Применение фреймворка, как правило, позволяет более гибко вести разработку, а также управлять полученным решением для дальнейшего внедрения и применения. Обучение, валидация и тестирование работоспособности нейронных сетей выполнялось на видеоадаптере NVIDIA GeForce RTX 3050 с 8 Гб VRAM, если не указано иное. Данный видеоадаптер выбран не случайно, а по той причине, что старшие модели

современных встраиваемых процессорных модулей серии NVIDIA Jetson такие, как Jetson T4000 и T5000 позиционируемые, как применимые в беспилотных летательных аппаратах и робототехнических комплексах, имеют сопоставимую производительность по TFLOPS.

Используемые датасеты

В силу того, что для большего числа параметров у глубокой нейронной сети нужно больше видеопамяти и больше обучающих изображений, то решено ограничиться тремя наборами разрешений изображений: 64x64, 128x128 и 224x224, с которыми работала нейронная сеть.

Для обучения, тестов и валидации использовалось два обучающих набора данных (датасета), которые подготовлены на основе кадров четырёх видеозаписей: двух видеозаписей с видеорегистратора автомобиля и двух синтетических видеозаписей из игрового симулятора Euro Truck Simulator 2. Для данных изображений создавалась пара, состоящая из исходного изображения и изображения чёрно-белого, содержащего результаты применения алгоритма SIFT [21] к исходному изображению, где наличие признака закрашивалось белым цветом, а отсутствие чёрным. Первый датасет – состоит из фрагментов исходного кадра разрешения 128x128 – 874358 штук изображений, а второй из всего кадра, отмасштабированного до разрешения 224x224 – 20000 штук изображений. Для датасета состоящего из фрагментов кадра в силу перекоса в сторону отсутствия найденных признаков на изображении, для избежания переобучения в сторону отсутствия признаков, выполнена балансировка.

Для выполнения балансировки проведен сбор статистической информации по числу признаков на изображении и числу таковых изображений на фрагментах кадра. В результате выяснено, что для числа признаков от 0 до 9 меньшим числом является 15158 изображений. Поэтому верхней границей для количества изображений было установлено число в 15158 изображений, а остальные изображения отброшены из группы для данного числа признаков. В итоге из подготовленных 874358 пар изображений отбросили 187100 изображения. Итоговый датасет фрагментов кадра после балансировки был

разделён на две группы, первая – обучающая 90%, и вторая 10% – валидационная.

Информация о использованном датасете фрагментов кадра:

1. Обучающий: 670632 пары изображений.
2. Валидационный: 16626 пар изображений.

Информация о использованном датасете кадров:

1. Обучающий: 18000 пар изображений.
2. Валидационный: 2000 пар изображений.

Пример изображения из датасета фрагментов кадра на рисунке 1, пример изображения датасета из кадров, отмасштабированных до разрешения 224x224 на рисунке 2.

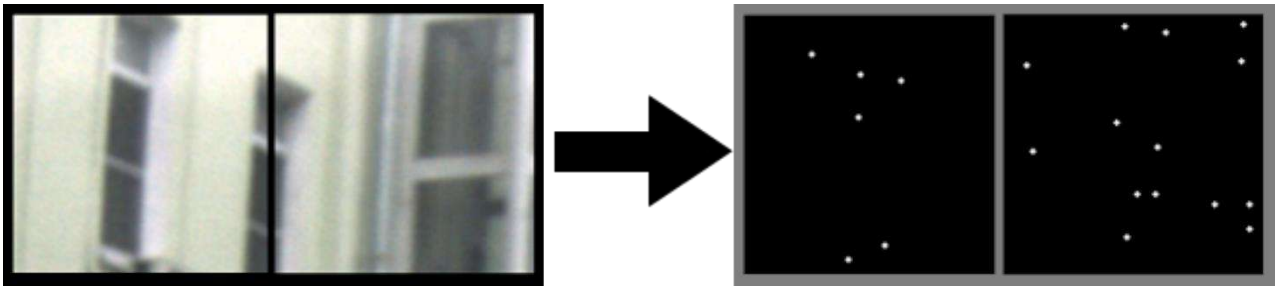

Рисунок 1 – Подготовленный набор частей исходных кадров 128x128x3 и признаков, содержащихся на них 128x128x1.

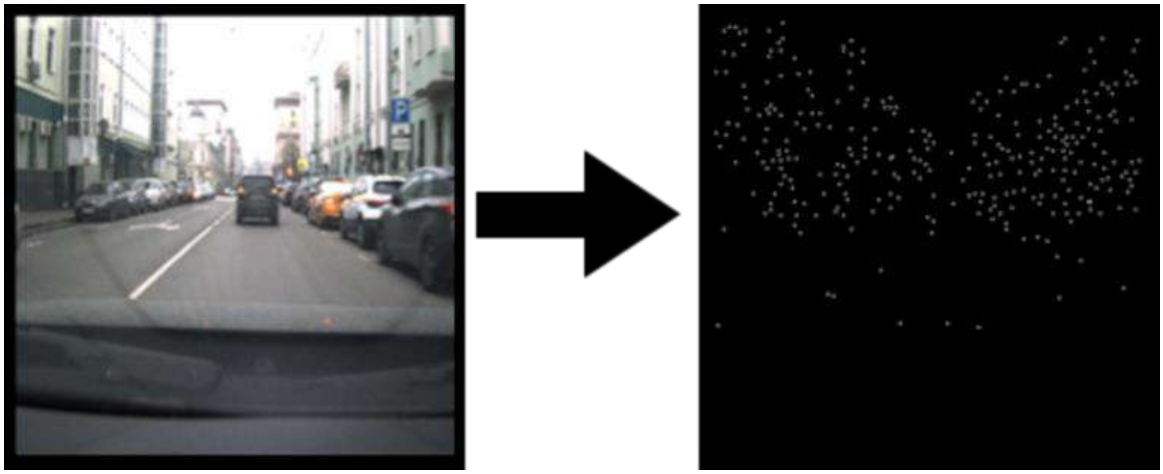

Рисунок 2 – Подготовленный набор изображений кадров видеозаписи 224x224x3 и их карт, содержащих признаки 224x224x1.

Для решения задачи поиска признаков возможны два подхода: первый подход – работа с сегментами исходного изображения фиксированного размера, второй – работа с исходным изображением, отмасштабированным до фиксированного разрешения. Первый подход требует дополнительного времени для предварительной подготовки сегментов изображений, но может обрабатываться нейронной сетью пакетно, а второй теряет детали исходного изображения, так как у изображения снижается разрешение и появляются артефакты, связанные с применением алгоритма, уменьшающего разрешение. Для исследования решено применить оба подхода и сравнить их для выбора лучшего.

Архитектуры нейронных сетей

Для решения задачи детектирования особых признаков необходим кодировщик и декодировщик некоторой архитектуры. В основе исследованной архитектуры лежит комбинация модифицированных архитектур MobileNet V2 и U-Net для выделения особых признаков на изображении. Соответственно MobileNet V2 выступает кодировщиком, а U-Net – декодировщиком.

MobileNet V2 – кодировщик, состоит из классической архитектуры MobileNet V2, но при этом содержит меньшее число остаточных блоков 13, а не 16, как в оригинальной архитектуре. Данное изменение вызвано тем, что наличие блоков 14-16 классической архитектуры приводит к избыточному снижению размерности, которое не требуется в данной задаче. Помимо того, для снижения числа фильтров в свёрточных слоях, в MobileNet V2 выставлен параметр Alpha=0.35. Изменение данного параметра позволяет уменьшить число параметров в итоговой архитектуре нейронной сети.

U-Net – декодировщик, необходим для получения карты признаков из выхода архитектуры MobileNet V2. При этом применение классической архитектуры U-Net, включающее транспонированные свёртки, показало значительные проблемы, влиявшие на качество обучения в силу того, как устроена транспонированная свёртка. На рисунке 3 виден результат при обучении на 20 эпохах, результаты не удовлетворительные в силу того, что

полученные признаки занимают не точечный размер, а некоторую область вокруг точки, что не позволило применить исходную архитектуру для задачи выявления точечных признаков.

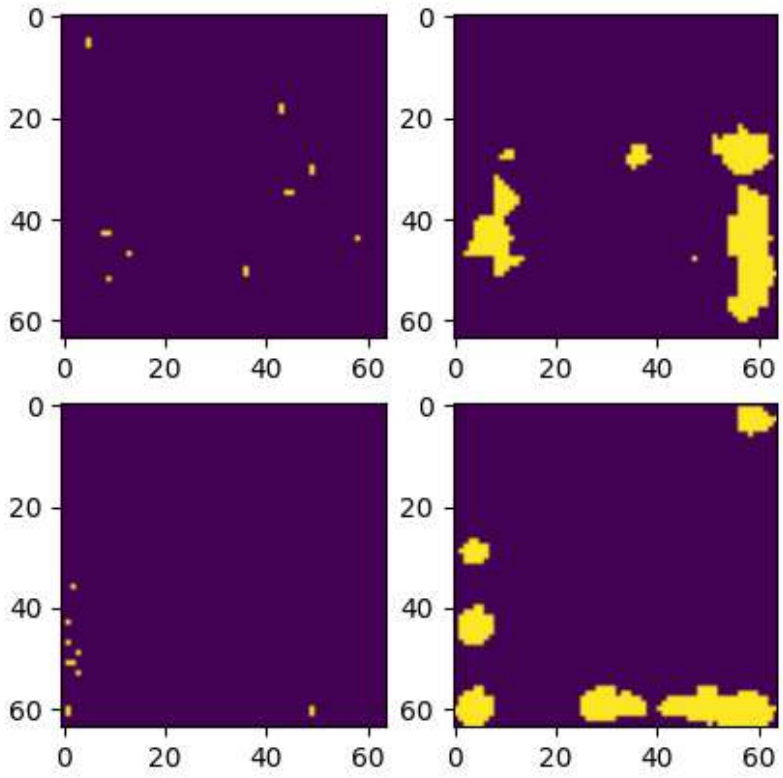

Рисунок 3 – По вертикали слева – ожидаемый результат, справа – полученный при обучении.

Для применения такой архитектуры потребуется дополнительная обработка результатов для поиска некоторого центра области, где предположительно находится признак. Поэтому потребовалось доработать декодировщик для улучшения качества получения точечных признаков. Для решения этой задачи изучалось применение четырёх подходов из которых выбран лучший. Классический подход U-Net включает в себя один слой транспонированной свёртки. В исследуемых подходах данный слой заменялся на комбинацию других слоёв.

В нашем случае из-за того, что итоговое изображение содержит два класса и имеется значительный перекос в сторону класса, где нет итогового признака, то классические метрики качества обучения по типу Accuracy не применимы. Поэтому качество обучения отслеживалось на основе метрики F1-Score [22],

которая не чувствительна к перекосу классов в обучающей выборке. При обнаружении плато или переобучения на валидационном наборе данных по данной метрике выполнялась автоматическая остановка обучения.

Вместо транспонированной свёртки использовался следующий набор слоёв:

1. Слой увеличения разрешения в 2 раза по горизонтали и 2 раза по вертикали.
2. Слой, выполняющий соединение прошлого слоя и информации от обходного соединения (skip-connection) для улучшения передачи градиентов.
3. Слой свёртки с числом фильтров в зависимости от расположения в архитектуре блока повышения разрешения: 96, 64, 32, 16, число фильтров снижается с повышением разрешения. Ядро свёртки 3x3, и отступы, для сохранения размерности выходного слоя.
4. Пакетная нормализация.
5. Функция активации ReLU.

При этом в части подходов слои 3-5 содержались два раза подряд, чтобы улучшить качество генерализации для точечных признаков.

Исследование качества применения данных слоёв проводилось при обучении на 20 эпохах, при применении во входных изображениях цветных изображений 64x64x3 и выходных 64x64x1. В качестве набора данных использовались изображения 128x128x3 из сегментов кадров масштабированные до разрешения 64x64x3.

Исследовались следующие наборы слоёв:

1. Набор слоёв, содержащий слои 1-5 и обычные свёртки.
2. Набор слоёв, содержащий слои 1-5 и дополнительный повтор слоёв 3-5 с обычными свёртками.
3. Набор слоёв, содержащий слои 1-5 и поканальные свёртки (depthwise convolution) с поточечными свёртками (pointwise convolution).
4. Набор слоёв, содержащий слои 1-5 и дополнительный повтор слоёв 3-5 с поканальными свёртками (depthwise convolution) с поточечными свёртками (pointwise convolution).

Для анализа полученных нейронных сетей из валидационного набора данных взяты несколько изображений, на рисунках 4-7 показано сравнение выходных результатов, которые ожидаются и которые предсказывает нейронная сеть.

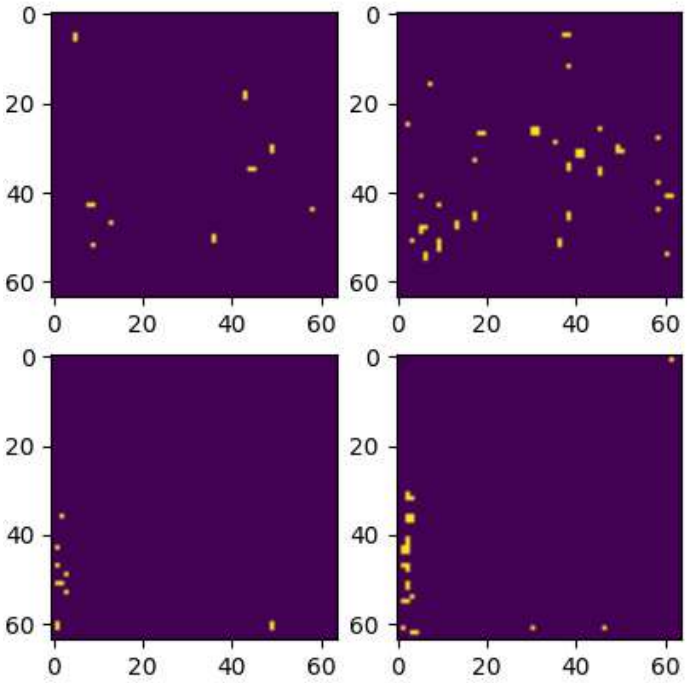

Рисунок 4 – Валидация результатов работы модификации 1 архитектуры U-Net.
По вертикали слева – ожидаемый результат, справа – полученный на валидации.

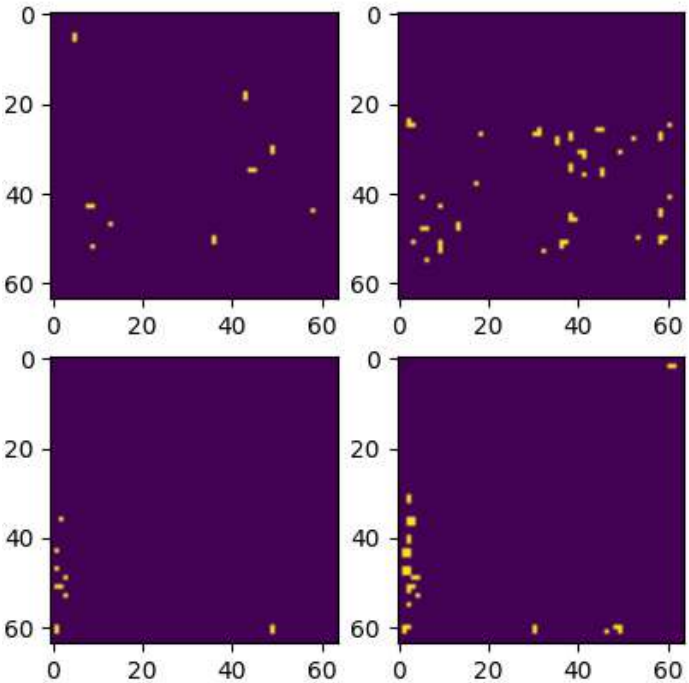

Рисунок 5 – Валидация результатов работы модификации 2 архитектуры U-Net.
По вертикали слева – ожидаемый результат, справа – полученный на валидации.

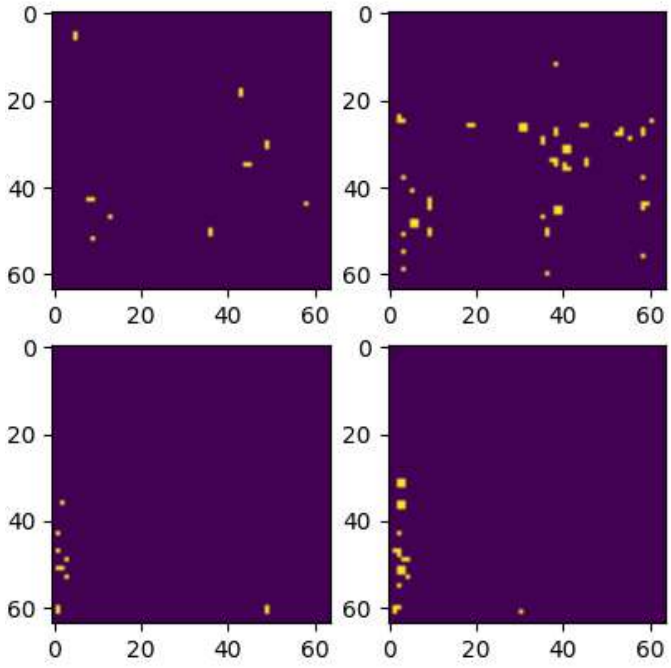

Рисунок 6 – Валидация результатов работы модификации 3 архитектуры U-Net.
По вертикали слева – ожидаемый результат, справа – полученный на валидации.

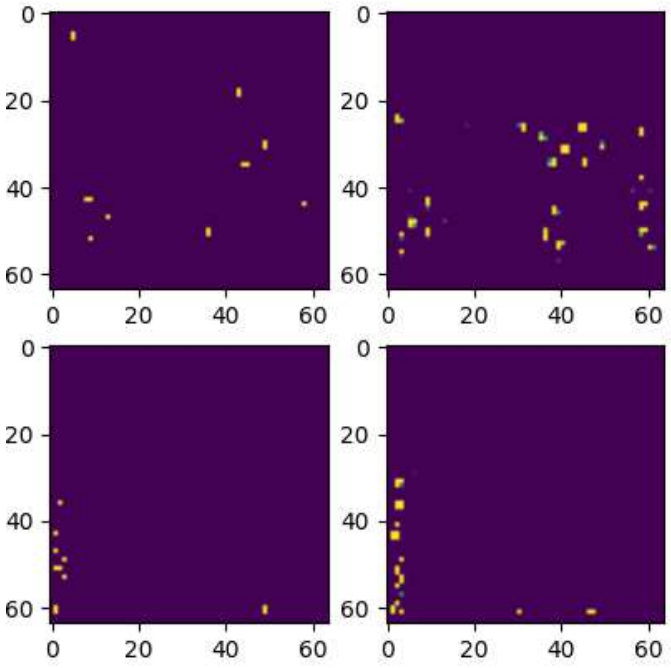

Рисунок 7 – Валидация результатов работы модификации 4 архитектуры U-Net.
По вертикали слева – ожидаемый результат, справа – полученный на валидации.

В процессе обучения собраны метрики, показанные в таблице 1. Помимо исследуемых модификаций U-Net в таблице приведён контрольный вариант, использующий транспонированные свёртки. По метрике F1-Score видно, что любой из подходов, использующих модификации U-Net, превосходит применение

транспонированных свёрток. На основе полученной информации по метрике F1-Score выбран подход 2, применяющий классические свёртки. Замедление скорости обучения по времени между подходами 1 и 2 в 17,6% не существенно. При этом исследовании модификаций архитектуры U-Net применялся видеоадаптер NVIDIA GeForce RTX 5070 Ti с 16 Гб VRAM.

Таблица 1

Метрики, собранные при исследовании модификаций архитектуры U-Net

Модификация архитектуры U-Net	F1-Score на валидации	Precision на валидации	Recall на валидации	Время обучения на NVIDIA GeForce RTX 5070 Ti	Число параметров всей архитектуры
1	0.2370	0.2437	0.3457	17 минут	1470225
2	0.2493	0.2120	0.3025	20 минут	1602593
3	0.2230	0.1774	0.3002	17 минут	722212
4	0.2367	0.3021	0.1946	18 минут	740412
Контроль	0.0670	0.0485	0.1702	13 минут	872481

Подробная схема полученной архитектуры представлена на рисунке 8.

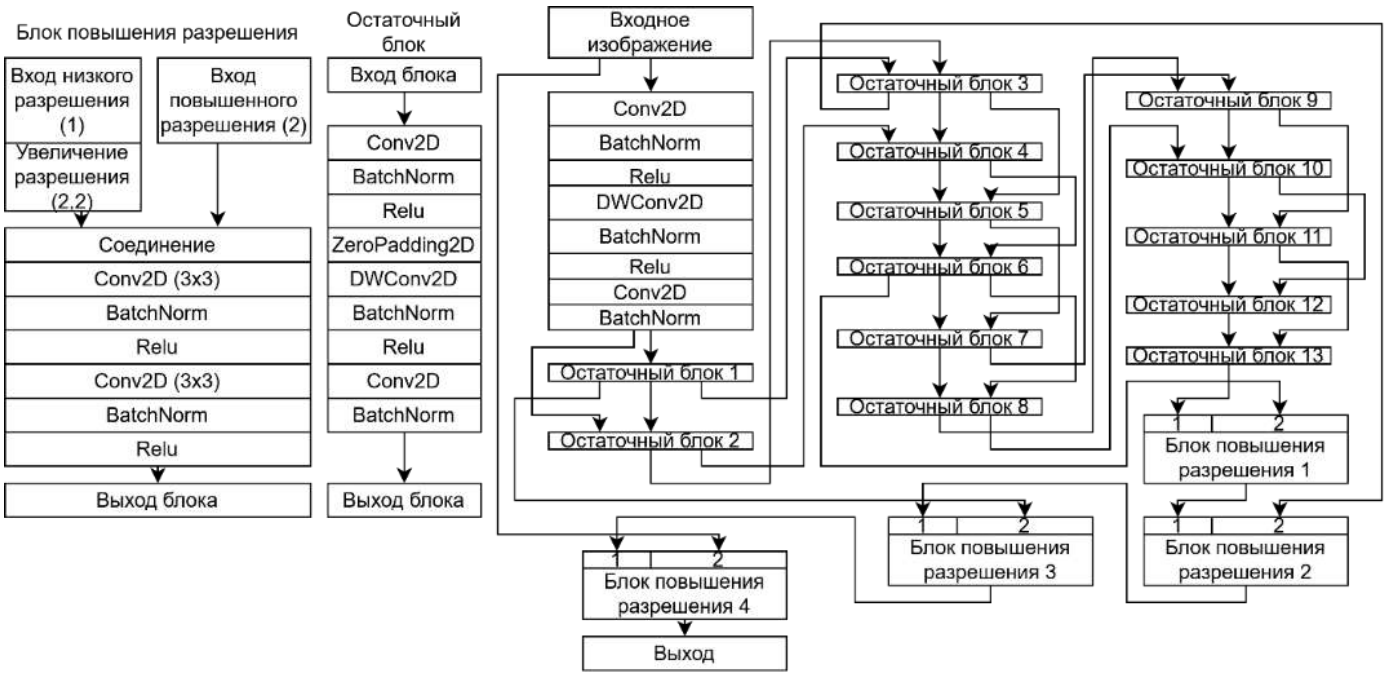

Рисунок 8 – Принципиальная схема разработанной архитектуры нейронной сети.

Слева на схеме представлен остаточный блок архитектуры MobileNet V2, а рядом с ним блок повышения разрешения. Декодировщик U-Net содержит вместо транспонированных свёрток комбинацию из слоёв: свёртки, соединения тензора, полученного из выхода свёртки и тензора от обходного маршрута, нормализации и функции активации, свёртки, нормализации и функции активации. Архитектура MobileNet V2 имеет меньшее число остаточных блоков 13 вместо 16 и Alpha=0.35 в отличии от классической архитектуры. Итоговое число слоёв 153 штуки. После каждой эпохи выполнялась валидация. При этом, как было сказано ранее, число фильтров в свёртках блока повышения разрешения уменьшается с повышением разрешения. В блоке повышения разрешения число фильтров такое: 1 блок – 96 фильтров, 2 блок – 64 фильтра, 3 блок – 32 фильтра, 4 блок – 16 фильтров. Данный подход позволяет задействовать меньшее число параметров.

При обучении, для повышения качества обучения, использовался случайный набор из аугментаций, где каждая аугментация имела шанс повлиять на обучающее изображение в 50%:

1. Изменения яркости.
2. Изменения контрастности. Независимо к каждому каналу.
3. Зеркалирования по горизонтали.
4. Зеркалирования по вертикали.
5. Случайного вращения вокруг центра изображения на произвольный угол.

Информация о разработанных нейронных сетях, их входном изображении и результатах обучения:

1. Фрагмент цветного изображения разрешения 64x64x3, число параметров нейронной сети 601009, F1-Score 0.2766. Обучение длилось 6 часов.
2. Фрагмент цветного изображения разрешения 128x128x3, число параметров нейронной сети 1602593, F1-Score 0.2335. Обучение длилось 3.33 часа.
3. Фрагмент чёрно-белого изображения 64x64x1, число параметров нейронной сети 600433, F1-Score 0.4498. Обучение длилось 2 часа.

4. Цветное изображение, отмасштабированное до разрешения 224x224x3, число параметров нейронной сети 558656, F1-Score 0.4128. Обучение длилось 4 часа.

В силу того, что нейронная сеть 3 показала лучший результат по метрике F1-Score, выполнено её тестирование для сравнения с алгоритмом SIFT. После чего выполнено апробирование данной нейронной сети для получения признаков в системе ВО и сравнение с применением алгоритма SIFT.

Тестирование нейронной сети

Для работы с изображениями большего разрешения, чем поддерживает на входе нейронная сеть, разработан алгоритм, который выполняет подготовку фрагментов изображения фиксированного размера на основе исходного кадра, соответствующих разрешению, которое ожидает нейронная сеть на вход.

Исследование работоспособности выявления признаков на фотографиях показало наличие мест, где нейронная сеть не нашла признаки в некоторых случаях. Это места близкие к краям фрагментов, на которые разрезалось исходное изображение рисунок 9. Вызвано это тем, что настройки алгоритма SIFT исключали детектирование признаков на краях изображения, что и повлияло на то, какие признаки обнаруживает итоговая нейронная сеть.

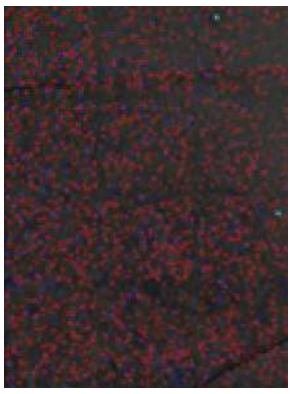

Рисунок 9 – Пример возникающих пустот без признаков на краях фрагментов, на которые разрезалось исходное изображение в части случаев.

Для тестирования нахождения признаков нейросетевым детектором взяты три изображения с различными погодными условиями, освещением, структурой, ориентацией. На рисунке 10 представлены тестовые изображения, на которых выполнено сравнение с алгоритмом SIFT.

Рисунок 10 – Тестовые изображения для проверки и анализа работоспособности нейросетевого детектора признаков. Слева направо: изображения 1, 2, 3.

Для данных тестовых изображений выполнено применение алгоритма SIFT и нейросетевого детектора, а также собрана статистика о том, сколько признаков найдено, сколько совпало и как долго выполнялся поиск. Изображения и их приближения для удобства изучения затемнены, на них расположены точки, в которых находятся признаки. Сами изображения показаны на рисунках 11-13. На изображениях видно, что нейросетевые признаки расположены не хаотично, а в местах, которые сохраняют прослеживаемость даже при поворотах камеры таких как углы и границы. Важно отметить, что в данных случаях не проявились места, где не находились признаки на краях разрезаемых фрагментов.

Рисунок 11 – Тестовое изображение 1 слева и в приближении в 2 и в 4 раза. Синим цветом – признаки SIFT, красным – признаки нейросетевого детектора.

Рисунок 12 – Тестовое изображение 2 слева и в приближении в 2 и в 4 раза.
Синим цветом – признаки SIFT, красным – признаки нейросетевого детектора.

Рисунок 13 – Тестовое изображение 3 слева и в приближении в 3 раза.
Синим цветом – признаки SIFT, красным – признаки нейросетевого детектора.

Статистика, собранная для данных изображений представлена в таблице 2. При данном тестировании применялся видеоадаптер NVIDIA GeForce RTX 5070 Ti с 16 Гб VRAM. Важно отметить, что небольшая часть признаков совпала по координатам пиксель в пиксель. В силу того, что данный видеоадаптер производительнее, чем применявшийся в обучении и апробировании то видно, что скорость обработки нейросетевых признаков при высоком разрешении изображения значительно выше и при этом проигрыш по времени исполнения по сравнению с алгоритмом SIFT от полутора до двух раз.

Таблица 2

Статистика по тестовым группам изображений при сравнении нейросетевого детектора признаков с детектором SIFT

Группа изображений	Число признаков SIFT	Число признаков нейронной сети	Время обработки SIFT	Время обработки нейронной сети	Число совпавших признаков
1	3999	908	197 мс	338 мс	68
2	5173	869	204 мс	323 мс	61
3	5225	1238	172 мс	368 мс	113

Апробирование нейронной сети

Важно отметить, что нейросетевой детектор признаков должен работать в составе существующей системы визуальной одометрии, что позволяет утверждать, что признаки имеют устойчивость и прослеживаемость, что играет важную роль для системы ВО. Поэтому апробирование имеет направленность в сторону интеграционного тестирования поведения строимого маршрута системой ВО. В случае наличия явных проблем с нейросетевым детектором или неприменимости данного нейросетевого детектора для данной задачи ожидаемо, что характер маршрута, строимого системой ВО, будет иметь вид некоторой случайной ломанной, не похожей на маршрут транспортного средства, или кривой находящейся в районе точки начала отсчёта на двумерной карте.

Апробирование выполнялось на наборе данных KITTI на маршрутах 0, 1, 2 для монокулярного варианта. Существующая система ВО является монокулярной, поэтому восстанавливает только относительное перемещение. Для восстановления информации об абсолютном перемещении и ошибке расстояния в метрах выполнялось умножение на модуль вектора перемещения из информации о точном перемещении, предоставленной авторами датасета KITTI [23]. Подробная статистическая информация, собранная при апробировании на датасете KITTI представлена в таблице 3.

Таблица 3

Основная статистическая информация о расстоянии в метрах
для визуальной одометрии датасета KITTI. Меньше – лучше.

Метрика	Способ получения признаков	KITTI 0 М.	KITTI 1 М.	KITTI 2 М.	KITTI 0 Сек.	KITTI 1 Сек.	KITTI 2 Сек.
Max	SIFT	581.13	228.41	459.4	0.35	0.48	0.47
	Нейронная сеть 3	827.22	1837.77	398.05	12.82	7.29	15.42
P_{50}	SIFT	193.65	76.15	180.36	0.04	0.09	0.04
	Нейронная сеть 3	465.04	815.86	78.63	0.05	0.05	0.06
P_{95}	SIFT	537.03	204.11	373.5	0.11	0.12	0.12
	Нейронная сеть 3	802.42	1823.43	304.52	0.22	0.16	0.31

На рисунке 14 изображены двумерные карты маршрута на которых: красный маршрут – данные, собранные авторами с GPS/OXTS, синий маршрут – результат ВО с использованием SIFT признаков, зелёный маршрут – результат ВО с использованием нейросетевых признаков. По изображениям маршрута видно, что нейросетевой детектор выполняет задачу по поиску признаков, что подтверждается строимым маршрутом.

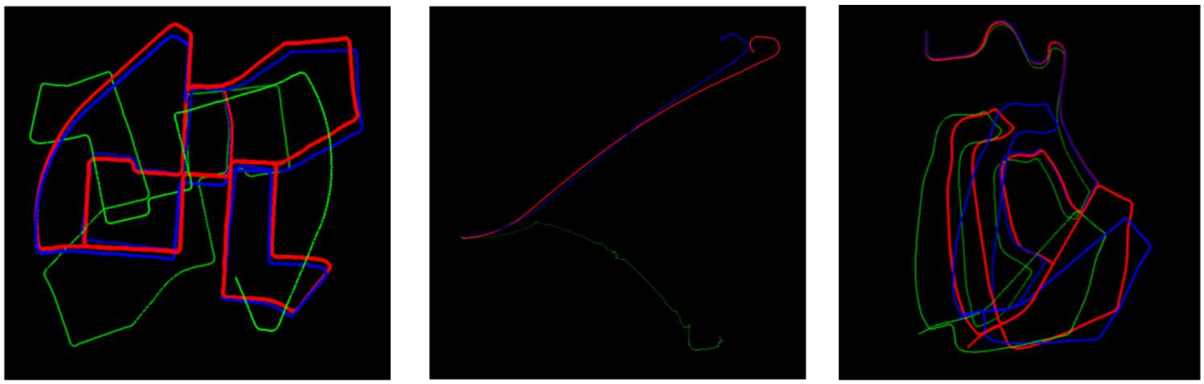

Рисунок 14 – Изображения двумерных карт маршрутов слева направо 0, 1, 2 из датасета KITTI.

Результаты апробирования:

1. В маршруте 2 применение нейронной сети 3 для детектирования признаков показало меньшее итоговое расстояние до точного положения, чем SIFT. 398.05 метров против 459.4 метров.

2. В маршрутах 0 и 1 нейросетевой детектор в системе ВО показал большую накопленную ошибку расстояния до точного положения. В данных маршрутах наблюдаются остановки транспортного средства и движение встречного трафика, что привело к тому, что часть найденных признаков оказывалось на движущихся транспортных средствах и данная система ВО при избыточном числе признаков не выполняла отсеивание таких признаков, что привело к тому, что накапливалась угловая ошибка. Поэтому траектория движения смещалась и возникал дрейф на таких участках остановок.

3. По временным затратам применение нейронной сети на NVIDIA GeForce RTX 3050 уступает алгоритму SIFT до 4 раз. Основными причинами является то, что по сравнению с SIFT появляется необходимость задействовать предварительную обработку изображений на CPU и далее шину PCI для копирования информации в VRAM GPU, что вносит задержки. Помимо этого,

после выполнения вычислений на GPU результаты копируются обратно в RAM для извлечения признаков с изображений с использованием CPU.

4. Нейросетевой детектор находил большее число признаков по сравнению с SIFT, что вызывало замедление работы данного алгоритма BO, так как большее число признаков требует больше времени на обработку. Важно, что вопрос оптимизации числа найденных признаков для увеличения прослеживаемости и устойчивости не изучался.

Выводы

1. Подтверждена возможность применения глубоких нейронных сетей для задачи выделения признаков на изображении в существующих системах визуальной одометрии. В отличие от традиционных алгоритмов (например, SIFT), нейросетевой подход не требует ручной настройки параметров и демонстрирует меньшую чувствительность к изменениям освещения и цвета, что делает его более устойчивым в динамичных условиях.

2. Разработана и оптимизирована архитектура нейронной сети, сочетающая MobileNet V2 (кодировщик), содержащий уменьшенное число блоков, и модифицированный U-Net (декодировщик), с заменёнными транспонированными свёртками на комбинацию из 8 слоёв: увеличения разрешения, соединения тензора увеличенного разрешения и тензора из обходного соединения, свёртки, нормализации, функции активации, свёртки, нормализации и функции активации. Такая модификация позволила получать необходимые точечные признаки на выходе нейронной сети, что критически важно для детектора признаков, применяемого в задаче визуальной одометрии.

3. Лучший результат по метрике F1-Score: 0.4498 был достигнут при использовании чёрно-белых изображений разрешения 64×64×1, что свидетельствует о эффективности предложенной архитектуры при работе с упрощёнными входными данными, содержащими меньшее число входных каналов. Также это подтверждает, что снижение размерности и использование бинарных карт признаков повышает качество обучения и точность детекции точечных признаков.

4. Нейросетевой детектор выявляет большее количество признаков, чем SIFT, что увеличивает информативность данных, используемых для оценки перемещения. Однако увеличение числа признаков приводит к росту вычислительных затрат и замедлению работы системы ВО. Поэтому для повышения эффективности требуется отбрасывать избыточные признаки для снижения вычислительных затрат.

5. При апробировании на датасете KITTI нейросетевой детектор показал превосходство в одном из трёх маршрутов (маршрут 2), где точность восстановления траектории была выше по сравнению с использованием признаков, находимых алгоритмом SIFT. Однако в маршрутах с остановками и встречным движением трафика (маршруты 0 и 1) наблюдался дрейф при остановке и накопление ошибок, возникающий из-за того, что система не отсеивала признаки на движущихся объектах. Это указывает на необходимость дальнейшей оптимизации фильтрации признаков.

6. Работоспособность системы подтверждена на видеоадаптере NVIDIA GeForce RTX 3050 с 8 ГБ VRAM, что позволяет предположить её возможность применения на встраиваемых платформах серии Nvidia Jetson T5000, имеющих сопоставимую производительность. Это делает предложенный подход перспективным для внедрения в мобильные робототехнические и беспилотные системы.

7. Предложенный комбинированный подход, сочетающий в себе нейросетевой детектор признаков и существующую алгоритмическую систему ВО, является перспективным направлением повышения точности и устойчивости навигационных параметров в системах, имеющих ограниченные ресурсы, особенно в условиях, где GPS-сигнал недоступен или ненадёжен.

В результате исследования подтверждена возможность эффективного применения нейронных сетей для этапа выделения признаков в системах визуальной одометрии, что позволяет снизить зависимость от внешних факторов таких, как освещение и погода, и улучшить устойчивость системы. Предложенный метод показал высокую эффективность при работе с реальными данными, а также реализуемость на мобильных вычислительных платформах.

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов.

Conflict of interest

The author declares no conflict of interest.

Список источников

1. Ермаков П. Г., Гоголев А. А. Сравнительный анализ схем комплексирования информации бесплатформенных инерциальных навигационных систем беспилотных летательных аппаратов // Труды МАИ, 2021, №117, http://mai.ru//upload/iblock/c31/xon4nnv6t4aum3wqzj4b7kfbsol369la/Ermakov_Gogolev_rus.pdf. DOI: 10.34759/trd-2021-117-11
2. Maimone M. Autonomous Navigation Results from the Mars Exploration Rover (MER) Mission / M. Maimone, A. Johnson, Y Cheng, R. Willson, L. Matthies // Experimental Robotics IX, 2006. P. 3-12. DOI:10.1007/11552246_1
3. Антонов Д. А., Жарков М. В., Кузнецов И. М., Чернодубов А. Ю. Методы повышения точности и помехозащищенности навигационного обеспечения транспортного средства // Труды МАИ, 2016, №90, http://mai.ru//upload/iblock/277/antonov_zharkov_kuznetsov_chernodubov_rus.pdf
4. D. Scaramuzza, F. Fraundorfer, Tutorial Visual Odometry // IEEE ROBOTICS & AUTOMATION MAGAZINE, 2011, Vol. 18, Iss. 4.
5. D. Nister, O. Naroditsky and J. Bergen, "Visual odometry," Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., Washington, DC, USA, 2004, pp. I-I, doi: 10.1109/CVPR.2004.1315094.
6. Teed, Zachary and Deng, Jia, “DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras”, Advances in neural information processing systems, 2021.
7. You Might Need to Say Goodbye to Affordable PCs; A Price Hike Storm Is Set to Hit in H2 2026 as Memory Shortages & Windows 10 EOL Collide: //

https://wccftech.com/you-might-need-to-say-goodbye-to-affordable-pcs/ (дата обращения: 20.12.2025).

8. Olivier Brochu Dufour and Abolfazl Mohebbi and Sofiane Achiche, An Attention-Based Deep Learning Architecture for Real-Time Monocular Visual Odometry: Applications to GPS-free Drone Navigation, 2024 https://arxiv.org/abs/2404.17745

9. L. Yu, E. Yang, B. Yang, Z. Fei and C. Niu, "A Robust Learned Feature-Based Visual Odometry System for UAV Pose Estimation in Challenging Indoor Environments," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-11, 2023, Art no. 5015411, doi: 10.1109/TIM.2023.3279458.

10. MonoVO-python // Github URL: https://github.com/uoip/monoVO-python (дата обращения: 08.02.2025).

11. Bradski, G. The OpenCV Library. Dr. Dobb's Journal of Software Tools. 15.01.2008

12. Дорошев А.С., Шеломанов Д.А. Методика подбора гиперпараметров нейросетевой модели в задачах оптической навигации // Труды МАИ. 2025. № 142. URL: https://trudymai.ru/published.php?ID=185106

13. Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications // arXiv.org, 2017, DOI: https://doi.org/10.48550/arXiv.1704.04861

14. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs // arXiv.org, 2017 https://doi.org/10.48550/arXiv.1606.00915

15. Liang-Chieh Chen, George Papandreou, Florian Schroff, Hartwig Adam, Rethinking Atrous Convolution for Semantic Image Segmentation // arXiv.org, 2017 https://arxiv.org/abs/1706.05587

16. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks // The IEEE Conference on Computer Vision

and Pattern Recognition (CVPR), 2018, pp. 4510-4520, DOI: https://doi.org/10.48550/arXiv.1801.04381

17. Andrew Howard, Mark Sandler and Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam, Searching for MobileNetV3 // arXiv.org, 2019 https://arxiv.org/abs/1905.02244

18. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

19. API Documentation // Tensorflow URL: https://www.tensorflow.org/api_docs/ (дата обращения: 08.02.2025).

20. T. Ganegedara, Tensorflow In Action // MANNING SHELTER ISLAND., 2022

21. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94

22. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015 Aug 12;15:29. doi: 10.1186/s12880-015-0068-x. PMID: 26263899; PMCID: PMC4533825.

23. Geiger, P. Lenz and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 3354-3361, DOI: 10.1109/CVPR.2012.6248074.

References

1. Ermakov P. G., Gogolev A. A. Comparative analysis of information integration architectures of strapdown inertial navigation systems for unmanned aerial vehicles. Trudy MAI, 2021, no 117, http://mai.ru//upload/iblock/c31/xon4nnv6t4aum3wqzj4b7kfbsol369la/Ermakov_Gogolev_rus.pdf. DOI: 10.34759/trd-2021-117-11

2. Maimone M. Autonomous Navigation Results from the Mars Exploration Rover (MER) Mission / M. Maimone, A. Johnson, Y Cheng, R. Willson, L. Matthies // Experimental Robotics IX, 2006. P. 3-12. DOI:10.1007/11552246_1

3. Antonov D. A., Zharkov M. V., Kuznetsov I. M., Tchernodoubov A. Y. Vehicle navigation system accuracy and noise immunity improvement techniques. Trudy MAI, 2016, no 90, http://mai.ru//upload/iblock/277/antonov_zharkov_kuznetsov_chernodubov_rus.pdf

4. D. Scaramuzza. Tutorial Visual Odometry / D. Scaramuzza, F. Fraundorfer // IEEE ROBOTICS & AUTOMATION MAGAZINE, 2011, Vol. 18, Iss. 4.

5. D. Nister, O. Naroditsky and J. Bergen, "Visual odometry," Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., Washington, DC, USA, 2004, pp. I-I, doi: 10.1109/CVPR.2004.1315094.

6. Teed, Zachary and Deng, Jia, “DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras”, Advances in neural information processing systems, 2021.

7. Olivier Brochu Dufour and Abolfazl Mohebbi and Sofiane Achiche, An Attention-Based Deep Learning Architecture for Real-Time Monocular Visual Odometry: Applications to GPS-free Drone Navigation, 2024 https://arxiv.org/abs/2404.17745

8. You Might Need to Say Goodbye to Affordable PCs; A Price Hike Storm Is Set to Hit in H2 2026 as Memory Shortages & Windows 10 EOL Collide: // https://wccftech.com/you-might-need-to-say-goodbye-to-affordable-pcs/ (дата обращения: 20.12.2025).

9. L. Yu, E. Yang, B. Yang, Z. Fei and C. Niu, "A Robust Learned Feature-Based Visual Odometry System for UAV Pose Estimation in Challenging Indoor Environments," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-11, 2023, Art no. 5015411, doi: 10.1109/TIM.2023.3279458.

10. MonoVO-python // Github URL: https://github.com/uoip/monoVO-python (дата обращения: 08.02.2025).

11. Bradski, G. / The OpenCV Library. Dr. Dobb's Journal of Software Tools. 15.01.2008

12. Doroshev A.S., Shelomanov D.A. Methodology for selecting hyperparameters of a neural network model in optical navigation tasks. Trudy MAI. 2025. No. 142. (In Russ.). URL: https://trudymai.ru/eng/published.php?ID=185106

13. Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications // arXiv.org, 2017, DOI: https://doi.org/10.48550/arXiv.1704.04861

14. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs // arXiv.org, 2017 https://doi.org/10.48550/arXiv.1606.00915

15. Liang-Chieh Chen, George Papandreou, Florian Schroff, Hartwig Adam, Rethinking Atrous Convolution for Semantic Image Segmentation // arXiv.org, 2017 https://arxiv.org/abs/1706.05587

16. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520, DOI: https://doi.org/10.48550/arXiv.1801.04381

17. Andrew Howard, Mark Sandler and Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam, Searching for MobileNetV3 // arXiv.org, 2019 https://arxiv.org/abs/1905.02244

18. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

19. API Documentation // Tensorflow URL: https://www.tensorflow.org/api_docs/ (дата обращения: 08.02.2025).

20. T. Ganegedara, Tensorflow In Action // MANNING SHELTER ISLAND., 2022

21. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94

22. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015 Aug 12;15:29. doi: 10.1186/s12880-015-0068-x. PMID: 26263899; PMCID: PMC4533825.

23. Geiger, P. Lenz and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 3354-3361, DOI: 10.1109/CVPR.2012.6248074.

Информация об авторах

Павел Георгиевич Корыткин, аспирант, Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова», г. Москва, Россия; e-mail: korytkinpg@my.msu.ru

Information about the authors

Pavel G. Korytkin, Postgraduate Student; Federal State Budget Educational Institution of Higher Education M.V. Lomonosov Moscow State University, Moscow, Russia; e-mail: korytkinpg@my.msu.ru

Получено 14 ноября 2025 ● Принято к публикации 25 декабря 2025 ● Опубликовано 30 декабря 2025
Received 14 November 2025 ● Accepted 25 December 2025 ● Published 30 December 2025