

© Я. Мурхиж, Дм. В. Учаев, Д. В. Учаев, Д. А. Гаврилов, Е. А. Татарина, 2025

ПРИМЕНЕНИЕ НЕЙРОННОЙ СЕТИ DDRNET ДЛЯ ПОВЫШЕНИЯ УСТОЙЧИВОСТИ VSLAM-РЕШЕНИЯ ORB-SLAM2 ПУТЕМ МАСКИРОВАНИЯ НА СНИМКАХ ПОДВИЖНЫХ ОБЪЕКТОВ

В работе проиллюстрирована возможность использования глубокой нейронной сети с двойным разрешением (DDRNet — Deep Dual-Resolution Networks) для решения в режиме реального времени задачи обнаружения на изображениях подвижных объектов (пешеходов, велосипедистов, автомобилей и др.). В работе представлено также программное решение DDR-SLAM в области VSLAM (Visual Simultaneous Localization and Mapping) в динамических средах. Данное VSLAM-решение представляет собой, по сути, улучшенную реализацию фреймворка ORB-SLAM2, в которой с помощью нейросети DDRNet производится маскирование подвижных объектов. Результаты экспериментальной апробации DDR-SLAM на последовательностях изображений из наборов данных KITTI и Cityscapes демонстрируют, что DDR-SLAM обеспечивает получение низких значений ошибок позиционирования в следующих сценариях: движение транспортного средства (ТС) в плотном потоке машин, длительное пребывание ТС в пробке, замедление движения ТС перед перекрестком, к которому приближаются также и иные ТС, и др. Показано также, что DDR-SLAM уверенно превосходит по точности позиционирования VSLAM-решение ORB-SLAM2 в тех сценариях, где присутствует большое число подвижных объектов. Исходный код DDR-SLAM доступен по ссылке: (<https://github.com/YznMur/DDRSLAM.git>).

Ключевые слова: глубокое обучение, одновременная локализация и построение карт (SLAM), семантическая сегментация, динамическая среда

ВВЕДЕНИЕ

В настоящее время во всем мире активно развиваются технологии автономного (беспилотного) вождения [1, 2]. Применение методов визуальной одновременной навигации и построения карты (VSLAM — Visual Simultaneous Localization and Mapping) позволяет получать информацию об окружающей среде, используя в качестве сенсоров только видеокамеры [3–5]. Использование VSLAM может быть затруднено в динамических средах, где присутствуют движущиеся объекты, такие как люди, автомобили и велосипеды, присутствие которых может приводить к ошибкам алгоритма и, как следствие, неправильным оценкам собственного положения и построению карты.

В последнее десятилетие активно развивается применение VSLAM для автономной робототехники благодаря низкой стоимости и компактным размерам. В работе [6] предложена система MonoSLAM [7], представляющая собой монохромную систему SLAM для работы в реальном времени. В работе [8] представлена система отслеживания камеры для дополненной реальности (PTAM — Parallel Tracking and Mapping), в которой предло-

жено разделить систему на два основных потока (отслеживание и построение карты), что позволило повысить надежность SLAM. В работе [9] предложен метод прямого отслеживания с вероятностным решением для учета влияния зашумленных значений глубины на процесс отслеживания. В работах [10, 11] представлены подходы, позволяющие решать без необходимости извлечения признаков и использования методов устойчивого сопоставления в оценке движения, комбинируя характерные точки с прямым отслеживанием оптического потока. Прямой метод стал популярным благодаря вычислительной эффективности [12]. Однако его нечувствительность к признакам является существенной проблемой. В SLAM точная оценка положения робота и построение карты окружающей среды являются критически важными, ограничения прямого метода могут препятствовать достижению этих целей. Ограничения прямого метода могут преодолеваются с помощью фреймворка визуальной инерциальной одометрии [13] ORB-SLAM (ORB — Oriented FAST and Rotated BRIEF) [14] и ORB-SLAM2 [15], которые представляют собой приложения SLAM, основанные на отслеживании признаков.

В большинстве систем SLAM динамические признаки (такие как пешеходы и транспортные средства) считаются ложными данными. В результате эти признаки не включаются в карту и не используются для отслеживания камеры. Общие подходы для обработки динамических признаков включают алгоритмы отбраковки выбросов, такие как RANSAC (RANDOM Sample Consensus) [16] в ORB-SLAM [17] и устойчивые функции стоимости в S-PTAM [18]. В последние годы для улучшения локализации и построения карт в динамических средах используются методы глубокого обучения, алгоритмы которых могут достаточно точно осуществлять обнаружение и сегментацию объектов в реальном времени [19, 20].

В работе [21] предложен метод SLAMIDE (Simultaneous Localization and Mapping in Dynamic Environments), который оценивает состояние 3D-признаков (стационарных или подвижных) с помощью обобщенного алгоритма максимизации ожиданий и использует обратимое связывание данных для включения подвижных объектов в единый фреймворк SLAM. В работе [17] представлена оффлайн-система визуального SLAM DynaSLAM, которая, основываясь на ORB-SLAM2, добавляет возможности обнаружения подвижных объектов и повышает устойчивость в динамических сценариях для моно-, стерео- и RGB-D конфигураций. В DynaSLAM изображения пропускаются через сверточную нейронную сеть (Mask R-CNN), которая выполняет семантическую сегментацию на уровне пикселей для всех априорно подвижных объектов (например, людей или транспортных средств). В работе [22] предложен фреймворк монохромного семантического SLAM под названием Dynamic-SLAM. Архитектура использует Single Shot Detector (SSD) для обнаружения объектов, который интегрирует предварительные знания для семантического определения движущихся объектов в новом потоке обнаружения. Для улучшения существующей сети SSD, склонной к низкому уровню обнаружения, в статье представлен алгоритм компенсации пропущенных обнаружений, основанный на инвариантности скорости в соседних кадрах, что значительно повышает уровень обнаружения. В [23] предложен Dynamic-DSO, который сочетает прямые методы с глубоким обучением для исключения подвижных кандидатов из ключевых кадров. Все вышеупомянутые исследования демонстрируют значительные усилия, направленные на решение проблем, связанных с динамическими средами в контексте VSLAM.

В настоящей работе предложена система DDR-SLAM, использующая сегментацию объектов на основе нейронной сети DDRNet [24] для достижения точной и надежной производительности

SLAM в динамических сценариях. DDR-SLAM представляет собой моно-, стерео- и RGB-D SLAM систему для динамических сред, которая одновременно оценивает положение камер, карту и траектории движущихся объектов через изображения с камеры. В качестве основы использована система ORB-SLAM2 [15], которая позволяет осуществлять локальное построение карты. Для обеспечения эффективной работы системы визуального SLAM в динамической среде используется DDRNet для идентификации и сегментации подвижных объектов, таких как пешеходы и автомобили, на уровне пикселей. DDR-SLAM успешно работает в реальном времени в сложных динамических условиях на открытой местности. Система решает задачи локализации Bundle adjustment, замыкания петли, разреженной реконструкции и другие проблемы в присутствии пешеходов и транспортных средств.

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАЗРАБОТАННОГО VSLAM-РЕШЕНИЯ

1.1. Используемые операционные средства

Для повышения точности определения положения и ориентации беспилотных транспортных средств (БТС) с помощью фреймворка ORB-SLAM2 целесообразным представляется усовершенствовать его за счет реализации блока обнаружения подвижных объектов и извлекать ключевые точки только для тех участков изображений, на которых не присутствуют подобного рода объекты. В этой связи в рамках проведенных исследований была разработана методика обнаружения с помощью нейросети DDRNet [25] подвижных объектов на изображениях с БТС-камер и реализовано VSLAM-решение DDR-SLAM, представляющее собой модификацию VSLAM-решения ORB-SLAM2, в которой при обнаружении ключевых точек изображений осуществляется маскирование подвижных объектов на изображениях.

В основу реализации методики обнаружения подвижных объектов положена модель DDRNet-23-slim, которая позволяет осуществлять семантическую сегментацию изображений в режиме реального времени и обеспечивает получение для изображений сегментационных карт, которые лучше соответствуют эталонным картам, чем сегментационные карты, полученные с помощью таких искусственных нейронных сетей (ИНС), как Fast-SCNN [26], DFANet [27], ERFNet [28], BiSENet [29], ICNet [30], DABNet [31], EDANet [32] и др. Стоит также отметить и то, что с целью выявления подвижных объектов на снимках, получаемых камерами БТС, модель DDRNet-23-slim была обучена на изображениях из наборов данных

KITTI (<https://www.cvlibs.net/datasets/kitti>) и Cityscapes (<https://www.cityscapes-dataset.com>).

Обнаружение с помощью сети DDRNet подвижных объектов на изображении предполагает, что данное изображение подается на вход сети DDRNet, а на выходе формируются бинарные маски, соответствующие выделенным классам подвижных объектов (пешеходам, велосипедистам, легковым автомобилям, грузовикам и др.). После объединения всех полученных масок создается результирующая маска для всех обнаруженных на изображении подвижных объектов.

Разработанное решение DDR-SLAM функционирует следующим образом. Первоначально получаемые камерами снимки пропускаются через сеть DDRNet, которая осуществляет обнаружение на снимках подвижных объектов. Вслед за этим для каждого из поступающих на вход DDRNet снимков формируется маска для всех обнаруженных на снимке подвижных объектов. На следующем шаге сформированные маски накладываются на соответствующие им снимки, и ключевые точки извлекаются только для тех участков снимков, которые не содержат обнаруженных с помощью сети DDRNet подвижных объектов. Далее по выделенным ключевым точкам снимков таким же образом, как и в ORB-SLAM2, осуществляется обнаружение и уточнение положения камер транспортного средства (ТС) с целью построения траектории движения ТС.

1.2. Сегментация потенциально динамического контента с использованием DDRNet

Для семантической сегментации изображений использована нейронная сеть DDRNet, характеризующаяся высокой производительностью и обеспечивающая решение задачи семантической сегментации изображений в режиме реального времени. Для выявления потенциально подвижных объектов на вход сети DDRNet подаются синхронизированные серые либо цветные изображения. На выходе сети DDRNet получаются маски, соответствующие классам потенциально подвижных объектов (пешеходы, велосипедисты, автомобили, грузовики и др.). Например, если на вход сети DDRNet подается RGB-изображение размерности $M \times N \times 3$, то на выходе получается массив размерности $M \times N \times L$, где L — количество объектов на изображении. Причем для каждого выходного канала получается своя бинарная маска. Путем объединения масок можно получить результирующую маску для всех потенциально подвижных объектов одного кадра — полную карту сегментации изображения. Карта сегментации показывает местоположение и контуры каждого подвижного объекта, обнаруженного DDRNet.

1.3. Низкозатратное отслеживание

После сегментации подвижных объектов на изображении положение камеры определяется с использованием статической части изображения. Как только потенциально движущиеся объекты изолированы, их границы обычно становятся областями с высоким контрастом и появляются характерные ключевые точки. Однако эти признаки в пределах контурных областей не используются в анализе. Для отслеживания позы камеры выполняются следующие шаги:

- извлечение признаков: создается карта признаков с помощью извлечения ключевых точек из статической части кадра изображения;
- поиск соответствий: отыскиваются соответствующие признаки в статических областях изображения;
- минимизация ошибки: оптимизируем позу камеры, минимизируя ошибку репроекции, которая представляет собой разницу между наблюдаемыми точками изображения и спроецированными точками из 3D-модели с использованием текущей позы камеры.

Представленный подход представляет собой более эффективный вычислительный вариант по сравнению с методом, используемым в ORB-SLAM2. В отличие от ORB-SLAM2, вход блока предсказания позы в разработанной системе состоит из RGB-кадров с масками сегментации. С помощью масок осуществляется извлечение ORB-признаков из сегментов изображения, которые классифицируются как статические. Кроме того, из-за высокой градиентной природы контуров сегментов удаляются ключевые точки, попадающие на контуры.

2. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

2.1. Условия экспериментов

Источники видеорядов

Для проверки надежности и точности системы DDR-SLAM в реальных динамических условиях была проведена серия экспериментов с различных точек зрения. Оценка DDR-SLAM осуществлялась с использованием KITTI Odometry Benchmark и набора данных Cityscapes. Выбор наборов данных обусловлен тем, что SLAM требует синхронизированных стереопоследовательностей с откорректированными изображениями.

Априорные знания нейронной сети DDRNet, используемой в эксперименте

Включают в основном классы подвижных объектов (человек, велосипед, автомобиль, мотоцикл,

автобус, поезд, грузовик и т.д.), которые могут быть распознаны сетью. Модель с априорными знаниями основана на сети DDRNet-23-slim, которая была обучена на наборах данных ImageNet, Cityscapes и KITTI360.

Экспериментальная среда

Экспериментальная среда выполнения представляет собой компьютер с конфигурацией: процессор Intel Core i7-9700K @ 3.60 ГГц, 8 ГБ оперативной памяти и видеокартой NVIDIA GeForce RTX 2060 с 6 ГБ видеопамяти.

2.2. Оценка эффективности на наборе данных KITTI

Набор данных KITTI является одним из крупнейших наборов данных для оценки алгоритмов компьютерного зрения в контексте автономного вождения, особенно для крупномасштабных сценариев. KITTI включает множество последовательностей, изображающих реальные уличные сцены, сопровождаемые точными данными о местоположении, собранными с помощью GPS и лазерного сканера Velodyne. Для проведения эксперимента отобрано 10 последовательностей для

сравнения траекторий DDR-SLAM, полученных для этих последовательностей с эталонными путями для проверки эффективности DDR-SLAM. Выбранные последовательности включают сцены на автомагистралях, городских дорогах, загородных дорогах и других. Эти сцены демонстрируют типичные объекты, встречающиеся в реальных уличных условиях, такие как пешеходы, велосипеды, мотоциклы, движущиеся транспортные средства, припаркованные автомобили на обочине и т.д.

Для оценки производительности ORB-SLAM2 и DDR-SLAM использовали оценки абсолютной ошибки позы (APE), включая среднеквадратичную ошибку (RMSE), среднюю ошибку (Mean), медианную ошибку (Median) и стандартное отклонение (STD). Для получения оценок APE и генерации траекторий движения использован пакет Python *evo*. Выравнивание траекторий с эталонным путем не использовалось при расчете APE, что позволило получить объективные оценки APE без занижения. В табл. 1 представлены статистические оценки ошибки APE, полученные для траекторий, восстановленных с помощью решений DDR-SLAM и ORB-SLAM2 для участвующих в апробации последовательностей из набора KITTI.

Табл. 1. Статистические оценки ошибки APE, полученные для траекторий, восстановленных с помощью решений DDR-SLAM и ORB-SLAM2 для участвующих в апробации последовательностей из набора KITTI

Последовательность	VSLAM-решение	RMSE	Mean	Median	STD
00	DDR-SLAM	9.83	8.80	8.54	4.38
	ORB-SLAM2	8.99	8.11	7.93	3.894
01	DDR-SLAM	13.41	12.50	13.83	4.85
	ORB-SLAM2	26.61	23.85	28.48	11.17
02	DDR-SLAM	13.63	13.07	13.37	3.87
	ORB-SLAM2	14.28	13.31	12.95	5.15
03	DDR-SLAM	9.10	7.80	7.51	4.69
	ORB-SLAM2	8.85	7.61	7.51	4.52
04	DDR-SLAM	2.60	2.23	1.99	1.34
	ORB-SLAM2	2.19	1.88	1.61	1.12

Табл. 1. Окончание

Последовательность	VSLAM-решение	RMSE	Mean	Median	STD
05	DDR-SLAM	2.81	2.62	2.64	1.02
	ORB-SLAM2	2.54	2.36	2.35	0.94
06	DDR-SLAM	2.98	2.633	2.50	1.38
	ORB-SLAM2	2.37	2.07	1.96	1.14
07	DDR-SLAM	13.70	13.53	13.47	2.18
	ORB-SLAM2	14.98	14.63	14.43	3.25
08	DDR-SLAM	8.51	7.11	8.51	4.67
	ORB-SLAM2	8.22	6.69	5.56	4.78
09	DDR-SLAM	6.48	5.35	4.04	3.66
	ORB-SLAM2	6.31	5.31	4.42	3.41

2.3. Оценка эффективности на наборе данных Cityscapes

Для оценки эффективности DDR-SLAM в практических условиях использован набор данных Cityscapes. Несмотря на то, что набор данных Cityscapes в основном предназначен для понимания сцены и сегментации изображений, он включает серию из более 100 000 кадров, записанных стереокамерой с высокой частотой кадров (17 кадров/с). Камеры в настройке Cityscapes имеют широкий динамический диапазон, что позволяет захватывать высококачественные изображения в сложных условиях освещения уличной среды.

К сожалению, набор данных Cityscapes не предоставляет эталонных данных для каждой последовательности изображений, а только неточную информацию GPS/одометрии транспортного средства, что затрудняет точную оценку производительности визуальной одометрии.

Тем не менее изображения из набора данных Cityscapes содержат множество движущихся объектов, включая автомобили, трамваи и грузовики. Для оценки точности разработанной системы отслеживания извлекаются меньшие сегменты, на-

зываемые S1, S2, S3 и S4, из длинной последовательности набора данных Cityscapes, а именно, стереопоследовательности Frankfurt. Извлеченные сегменты имеют различную длину от 400 до 5000 кадров. Для каждой небольшой последовательности вычислялись эталонные позы камеры с использованием координат GPS и метода проекции Меркатора.

Для устранения неточностей в эталонных данных, собранных с помощью GPS, использовалось преобразование SE(3) для выравнивания траекторий. В результате получено более точное соответствие оцененных траекторий эталонным траекториям. В табл. 2 представлены статистические оценки ошибки APE, полученные для траекторий, восстановленных с помощью решений DDR-SLAM и ORB-SLAM2 для участвующих в апробации фрагментов последовательности Frankfurt из набора Cityscapes.

Описание фрагментов последовательности Frankfurt из набора Cityscapes, используемых для апробации DDR-SLAM, представлено в табл. 3 и на рисунках.

Табл. 2. Статистические оценки ошибки APE, полученные для траекторий, восстановленных с помощью решений DDR-SLAM и ORB-SLAM2 для участвующих в апробации фрагментов последовательности Frankfurt из набора Cityscapes

Фрагмент последовательности Frankfurt	VSLAM-решение	RMSE	Mean	Median	STD
S1	DDR-SLAM	2.47	2.01	1.74	1.43
	ORB-SLAM2	8.06	4.20	1.53	6.88
S2	DDR-SLAM	3.29	2.77	2.38	1.77
	ORB-SLAM2	3.39	2.82	2.41	1.88
S3	DDR-SLAM	4.64	4.10	3.93	2.17
	ORB-SLAM2	4.63	4.09	3.95	2.16
S4	DDR-SLAM	0.31	0.27	0.25	0.14
	ORB-SLAM2	0.31	0.27	0.24	0.14

Табл. 3. Описание фрагментов последовательности Frankfurt из набора Cityscapes, используемых для апробации DDR-SLAM

Фрагмент последовательности Frankfurt *	Место съемки	Сценарии движения автомобиля
S1 (42 400–43 400)	город	Автомобиль приближается к регулируемому сигналами светофоров перекрестку и останавливается позади стоящего на перекрестке грузовика. По мере приближения автомобиля к грузовику последний становится доминирующим объектом сцены. Как только загорается зеленый сигнал светофора, грузовик начинает движение вперед и отрывается (рис., а).
S2 (55 050–55 400)	город	Автомобиль поворачивает направо перед группой пешеходов, которые становятся доминирующими объектами сцены и, как следствие, осложняют определение позы (рис., б).
S3 (13 110–13 400)	город	Доминирующим объектом сцены для автомобиля, стоящего перед пешеходным переходом, становятся переходящие улицу пешеходы и встречный автомобиль, поворачивающий налево (рис., в).
S4 (76 000–76 300)	город	Трамвай проезжает непосредственно перед движущимся автомобилем и тем самым существенно искажает оценку позы (рис., г).

Примечание: в скобках приводятся номера первого и последнего кадров последовательности Frankfurt, которые определяют последовательности S1–S4.



Рис. Сценарии движения автомобиля, реализуемые последовательностями S1 (а). S2 (б). S3 (в) и S4 (г)

ЗАКЛЮЧЕНИЕ

В настоящей работе представлена система DDR-SLAM как масштабируемая моно-/стереосистема визуального SLAM, основанная на семантической осведомленности и способная работать в реальном времени в динамических условиях. Система DDR-SLAM протестирована на реальных данных. Количественные оценки производительности DDR-SLAM подтверждают, что метод достаточно эффективно справляется со сложными ситуациями, в которых традиционные методы VSLAM испытывают затруднения, и сохраняет свою эффективность в статических средах, хотя может иметь ограничения, когда объекты, считающиеся статичными, на самом деле находятся

в движении или когда объекты, считающиеся динамическими, на самом деле неподвижны. В будущих работах планируется добавить отслеживание и проверку геометрической правдоподобности для решения указанных проблем и повышения точности предложенного метода DDR-SLAM.

Исследование выполнено за счет гранта Российского научного фонда № 25-27-00361, <https://rscf.ru/project/25-27-00361/>

СПИСОК ИСТОЧНИКОВ

1. Бутакова М.А., Щербань И.В., Мишин Н.А., Беляский Г.И. Два метода оценки оптического потока по видеоряду изображений // Искусственный интел-

- лект и принятие решений. 2025. № 1. С. 115–117. DOI: 10.14357/20718594250110
2. Гаврилов Д.А., Татаринова Е.А., Фортунатов А.А., Буздин В.Э., Учаев Д.В., Учаев Дм. В., Михайлов И.М., Терентьев М.В., Щелкунов Д.Н., Мурхиж Я., Поткин О.А. Разработка прототипа многофункциональной оптико-электронной системы кругового обзора для обеспечения визуального ориентирования беспилотных транспортных средств // Научное приборостроение. 2023. Т. 33. № 3. С. 37–48. URL: <http://iairas.ru/mag/2023/full3/Art3.pdf>
 3. Лапушкин А.Г., Гаврилов Д.А., Поткин О.А. Программное обеспечение для создания синтезированных данных и симулятор с обратной связью для тестирования алгоритмов машинного обучения // Научное приборостроение. 2023. Т. 33. № 1. С. 95–108. URL: <http://iairas.ru/mag/2023/full1/Art8.pdf>
 4. Лапушкин А.Г., Гаврилов Д.А., Щелкунов Н.Н., Бакеев Р.Н. Основные подходы к подготовке визуальных данных для обучения нейросетевых алгоритмов // Искусственный интеллект и принятие решений. 2021. № 4. С. 62–74. DOI: 10.14357/20718594210406
 5. Ващенко А.П., Гаврилов Д.А., Лапушкин А.Г. Разработка прототипа аппаратно-программного устройства панорамного обзора для систем безопасности и ситуационной осведомленности беспилотного транспортного средства // Вопросы кибербезопасности. 2021. № 5 (45). С. 87–95. URL: https://cyberrus.info/wp-content/uploads/2021/09/87-95-545-21_11.-Vashchenko.pdf
 6. Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. ImageNet: A large-scale hierarchical image database // 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, 2009. P. 248–255. DOI: 10.1109/CVPR.2009.5206848
 7. Davison A.J., Reid I.D., Molton N.D., Stasse O. Mono-SLAM: Real-Time Single Camera SLAM // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007. Vol. 29, no. 6. P. 1052–1067. DOI: 10.1109/TPAMI.2007.1049
 8. Engel J., Koltun V., Cremers D. Direct sparse odometry // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018. Vol. 40, no. 3. P. 611–625. DOI: 10.1109/TPAMI.2017.2658577
 9. Engel J., Schöps T., Cremers D. LSD-SLAM: Large-scale direct monocular SLAM // ECCV 2014. Springer, Cham, 2014. P. 834–849. DOI: 10.1007/978-3-319-10605-2_54
 10. Forster C., Pizzoli M., Scaramuzza D. SVO: Fast semi-direct monocular visual odometry // 2014 IEEE international conference on robotics and automation (ICRA). Hong Kong, 2014. P. 15–22. DOI: 10.1109/ICRA.2014.6906584
 11. Forster C. et al. SVO: semidirect visual odometry for monocular and multicamera systems // IEEE Transactions on Robotics. 2016. Vol. 33, no. 2. P. 249–265. DOI: 10.1109/TRO.2016.2623335
 12. Geiger A. et al. Vision meets robotics: The KITTI dataset // The international journal of robotics research. 2013. Vol. 32, no. 11. P. 1231–1237. DOI: 10.1177/0278364913491297
 13. Hong Y. et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes // arXiv preprint. arXiv:2101.06085. 2021. DOI: 10.48550/arXiv.2101.06085
 14. Klein G., Murray D. Parallel tracking and mapping for small AR workspaces // 2007 6th IEEE and ACM international symposium on mixed and augmented reality. Nara, 2007. P. 225–234. DOI: 10.1109/ISMAR.2007.4538852
 15. Bescos B. et al. DynaSLAM: tracking, mapping, and inpainting in dynamic scenes // IEEE robotics and automation letters. 2018. Vol. 3, no. 4. P. 4076–4083. DOI: 10.1109/LRA.2018.2860039
 16. Fischler M.A., Bolles R.C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography // Communications of the ACM. 1981. Vol. 24, no. 6. P. 381–395. DOI: 10.1145/358669.358692
 17. Mur-Artal R., Montiel J.M.M., Tardós J.D. ORB-SLAM: A versatile and accurate monocular SLAM system // IEEE transactions on robotics. 2015. Vol. 31, no. 5. P. 1147–1163. DOI: 10.1109/TRO.2015.2463671
 18. Leutenegger S. et al. Keyframe-based visual-inertial odometry using nonlinear optimization // The International Journal of Robotics Research. 2015. Vol. 34, no. 3. P. 314–334. DOI: 10.1177/0278364914554813
 19. Liao Y., Xie J., Geiger A. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023. Vol. 45, no. 3. P. 3292–3310. DOI: 10.1109/TPAMI.2022.3179507
 20. Lin T.Y. et al. Feature pyramid networks for object detection // Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, 2017. P. 2117–2125. DOI: 10.1109/CVPR.2017.106
 21. Mur-Artal R., Tardós J.D. ORB-SLAM2: an open-source slam system for monocular, stereo, and RGB-D cameras // IEEE transactions on robotics. 2017. Vol. 33, no. 5. P. 1255–1262. DOI: 10.1109/TRO.2017.2705103
 22. Pire T. et al. S-PTAM: Stereo parallel tracking and mapping // Robotics and Autonomous Systems. 2017. Vol. 93, P. 27–42. DOI: 10.1016/j.robot.2017.03.019
 23. Qin T., Li P., Shen S. VINS-mono: a robust and versatile monocular visual-inertial state estimator // IEEE transactions on robotics. 2018. Vol. 34, no. 4. P. 1004–1020. DOI: 10.1109/TRO.2018.2853729
 24. Bibby C., Reid I. Simultaneous localisation and mapping in dynamic environments (SLAMIDE) with reversible data association // Robotics: Science and Systems Proceedings III. Atlanta, 2007. 8 p. DOI: 10.15607/RSS.2007.III.014
 25. Pan H. et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes // IEEE Transactions on Intelligent Transportation Systems. 2022. Vol. 24, no. 3. P. 3448–3460. DOI: 10.1109/TITS.2022.3228042
 26. Poudel R.P.K., Liwicki S., Cipolla R. Fast-SCNN: Fast semantic segmentation network // arXiv preprint arXiv:1902.04502. 2019. DOI: 10.48550/arXiv.1902.04502
 27. Li H. et al. DFAnet: Deep feature aggregation for real-time semantic segmentation // Proceedings of the

- IEEE/CVF conference on computer vision and pattern recognition (CVPR). Long beach, 2019. P. 9522–9531. DOI: 10.1109/CVPR.2019.00975
28. *Romera E. et al.* ERFNet: Efficient residual factorized convNet for real-time semantic segmentation // IEEE Transactions on Intelligent Transportation Systems. 2017. Vol. 19, no. 1. P. 263–272. DOI: 10.1109/TITS.2017.2750080
 29. *Yu C. et al.* BiSeNet: Bilateral segmentation network for real-time semantic segmentation // Proceedings of the European conference on computer vision (ECCV). 2018. P. 325–341. URL: https://www.ecva.net/papers/eccv_2018/papers_ECCV/html/Changqian_Yu_BiSeNet_Bilateral_Segmentation_ECCV_2018_paper.php
 30. *Zhao H. et al.* ICNet for real-time semantic segmentation on high-resolution images // Proceedings of the European conference on computer vision (ECCV). 2018. 16 p. URL: https://openaccess.thecvf.com/content_ECCV_2018/papers/Hengshuang_Zhao_ICNet_for_Real-Time_ECCV_2018_paper.pdf
 31. *Li G. et al.* Depth-wise asymmetric bottleneck with point-wise aggregation decoder for real-time semantic segmentation in urban scenes // IEEE Access. 2020. Vol. 8. P. 27495–27506. DOI: 10.1109/ACCESS.2020.2971760
 32. *Lo S.Y. et al.* Efficient dense modules of asymmetric convolution for real-time semantic segmentation // Proceedings of the 1st ACM International Conference on Multimedia in Asia. 2019. P. 1–6. DOI: 10.1145/3338533.3366558

**Московский физико-технический институт
(государственный университет), г. Москва**
(Мурхиж Я., Гаврилов Д.А., Татаринова Е.А.)

**Московский государственный университет геодезии
и картографии, г. Москва** (Учаев Дм.В., Учаев Д.В.)

Контакты: Татаринова Елена Александровна
tatarinova.ea@mipt.ru

Материал поступил в редакцию 15.10.2025

USING DDRNET TO ENHANCE VSLAM STABILITY IN A MOVING SCENE

Ya. Murhij¹, Dm. Uchaev², D. Uchaev², D. Gavrilov¹, E. Tatarinova¹

¹Federal State Autonomous Educational Institution of Higher Education "Moscow Institute of Physics and Technology (National Research University)", Dolgoprudny, Moscow, Russia

²Federal State Budgetary Educational Institution of Higher Education "Moscow State University of Geodesy and Cartography", Moscow, Russia

The paper illustrates the possibility of using a deep dual-resolution neural network (DDRNet) for solving the problem of real-time detection of moving objects (pedestrians, cyclists, cars, etc.) in images. The paper also presents a DDR-SLAM software solution in the field of VSLAM (Visual Simultaneous Localization and Mapping) in dynamic environments. This VSLAM solution is, in fact, an improved implementation of the ORB-SLAM2 framework, in which the DDRNet neural network is used to mask moving objects. The results of experimental testing of DDR-SLAM on image sequences from the KITTI and Cityscapes datasets demonstrate that DDR-SLAM ensures low positioning error values in the following scenarios: vehicle movement in dense traffic; long-term stay of a vehicle in a traffic jam; slowing down the movement of a vehicle before an intersection, to which other vehicles are also approaching, etc. It is also shown that DDR-SLAM confidently outperforms the VSLAM solution ORB-SLAM2 in terms of positioning accuracy in scenarios where there is a large number of moving objects. The source code of DDR-SLAM is available at the link (<https://github.com/YznMur/DDRSLAM.git>).

Keywords: deep learning, simultaneous localization and mapping (SLAM), semantic segmentation, dynamic environment

REFERENCES

- Butakova M.A., Shcherban I.V., Mishin N.A., Belyavskiy G.I. [Two methods for estimating optical flow from a video sequence of images]. *Iskusstvennyi intellekt i prinyatie reshenii* [Artificial intelligence and decision making], 2025, no. 1, pp. 115–117. DOI: 10.14357/20718594250110 (In Russ.).
- Gavrilov D.A., Tatarinova E.A., Fortunatov A.A., Buzdin V.E., Uchaev D.V., Uchaev Dm.V., Mikhailov I.M., Terentiev M.V., Shchelkunov D.N., Murhizh Ya., Potkin O.A. [Development of a prototype of a multifunctional optical-electronic all-round viewing system to provide visual orientation of unmanned vehicles]. *Nauchnoe Priborostroenie* [Scientific Instrumentation], 2023, vol. 33, no. 3, pp. 37–48. (In Russ.). URL: <http://iairas.ru/mag/2023/full3/Art3.pdf>
- Lapushkin A.G., Gavrilov D.A., Potkin O.A. [Synthesized data creation software and feedback simulator for testing machine learning algorithms]. *Nauchnoe Priborostroenie* [Scientific Instrumentation], 2023, vol. 33, no. 1, pp. 95–108. (In Russ.). URL: <http://iairas.ru/mag/2023/full1/Art8.pdf>
- Lapushkin A.G., Gavrilov D.A., Shchelkunov D.N., Bakeev R.N. [The main approaches to the preparation of visual data for training neural network algorithms]. *Iskusstvennyi intellekt i prinyatie reshenii* [Artificial intelligence and decision making], 2021, no. 4, pp. 62–74. DOI: 10.14357/20718594210406 (In Russ.).
- Vashchenko A.P., Gavrilov D.A., Lapushkin A.G. [Development of a prototype of a hardware and software device for panoramic viewing for security systems and situational awareness of an unmanned vehicle]. *Voprosy kibernetiki* [Cybersecurity issues], 2021, no. 5 (45), pp. 87–95. (In Russ.). URL: https://cyberrus.info/wp-content/uploads/2021/09/87-95-545-21_11.-Vashchenko.pdf
- Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848
- Davison A.J., Reid I.D., Molton N.D., Stasse O. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, vol. 29, no. 6, pp. 1052–1067. DOI: 10.1109/TPAMI.2007.1049
- Engel J., Koltun V., Cremers D. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, vol. 40, no. 3, pp. 611–625. DOI: 10.1109/TPAMI.2017.2658577
- Engel J., Schöps T., Cremers D. LSD-SLAM: Large-scale direct monocular SLAM. *Computer Vision – ECCV 2014, ECCV 2014, Lecture Notes in Computer Science*, Springer, Cham, 2014, vol. 8690, pp. 834–849. DOI: 10.1007/978-3-319-10605-2_54
- Forster C., Pizzoli M., Scaramuzza D. SVO: Fast semi-direct monocular visual odometry. *2014 IEEE international conference on robotics and automation (ICRA)*, Hong Kong, 2014, pp. 15–22. DOI: 10.1109/ICRA.2014.6906584
- Forster C. et al. SVO: semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 2016, vol. 33, no. 2, pp. 249–265. DOI: 10.1109/TRO.2016.2623335

12. Geiger A. et al. Vision meets robotics: The KITTI dataset. *The international journal of robotics research*, 2013, vol. 32, no. 11, pp. 1231–1237. DOI: 10.1177/0278364913491297
13. Hong Y. et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv preprint*, arXiv:2101.06085, 2021. DOI: 10.48550/arXiv.2101.06085
14. Klein G., Murray D. Parallel tracking and mapping for small AR workspaces. *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, Nara, 2007, pp. 225–234. DOI: 10.1109/ISMAR.2007.4538852
15. Bescos B. et al. DynaSLAM: tracking, mapping, and inpainting in dynamic scenes. *IEEE robotics and automation letters*, 2018, vol. 3, no. 4, pp. 4076–4083. DOI: 10.1109/LRA.2018.2860039
16. Fischler M.A., Bolles R.C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981, vol. 24, no. 6, pp. 381–395. DOI: 10.1145/358669.358692
17. Mur-Artal R., Montiel J.M.M., Tardós J.D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, 2015, vol. 31, no. 5, pp. 1147–1163. DOI: 10.1109/TRO.2015.2463671
18. Leutenegger S. et al. Keyframe-based visual-inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 2015, vol. 34, no. 3, pp. 314–334. DOI: 10.1177/0278364914554813
19. Liao Y., Xie J., Geiger A. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, vol. 45, no. 3, pp. 3292–3310. DOI: 10.1109/TPAMI.2022.3179507
20. Lin T.-Y. et al. Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, 2017, pp. 2117–2125. DOI: 10.1109/CVPR.2017.106
21. Mur-Artal R., Tardós J.D. ORB-SLAM2: an open-source slam system for monocular, stereo, and RGB-D cameras. *IEEE transactions on robotics*, 2017, vol. 33, no. 5, pp. 1255–1262. DOI: 10.1109/TRO.2017.2705103
22. Pire T. et al. S-PTAM: Stereo parallel tracking and mapping. *Robotics and Autonomous Systems*, 2017, vol. 93, pp. 27–42. DOI: 10.1016/j.robot.2017.03.019
23. Qin T., Li P., Shen S. VINS-mono: a robust and versatile monocular visual-inertial state estimator. *IEEE transactions on robotics*, 2018, vol. 34, no. 4, pp. 1004–1020. DOI: 10.1109/TRO.2018.2853729
24. Bibby C., Reid I. Simultaneous localisation and mapping in dynamic environments (SLAMIDE) with reversible data association. *Robotics: Science and Systems Proceedings III*, Atlanta, 2007. 8 p. DOI: 10.15607/RSS.2007.III.014
25. Pan H. et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*, 2022, vol. 24, no. 3, pp. 3448–3460. DOI: 10.1109/TITS.2022.3228042
26. Poudel R.P.K., Liwicki S., Cipolla R. Fast-SCNN: Fast semantic segmentation network. *arXiv preprint*, arXiv:1902.04502, 2019. DOI: 10.48550/arXiv.1902.04502
27. Li H. et al. DFAnet: Deep feature aggregation for real-time semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Long beach, 2019, pp. 9522–9531. DOI: 10.1109/CVPR.2019.00975
28. Romera E. et al. ERFNet: Efficient residual factorized convNet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2017, vol. 19, no. 1, pp. 263–272. DOI: 10.1109/TITS.2017.2750080
29. Yu C. et al. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. *Computer Vision – ECCV 2018, Lecture Notes in Computer Science*, vol. 11217, Springer, Cham, 2018, pp. 334–349. DOI: 10.1007/978-3-030-01261-8_20
30. Zhao H. et al. ICnet for real-time semantic segmentation on high-resolution images. *Proceedings of the European conference on computer vision (ECCV)*, 2018. 16 p. DOI: 10.1007/978-3-030-01219-9_25
31. Li G. et al. Depth-wise asymmetric bottleneck with point-wise aggregation decoder for real-time semantic segmentation in urban scenes. *IEEE Access*, 2020, vol. 8, pp. 27495–27506. DOI: 10.1109/ACCESS.2020.2971760
32. Lo S.Y. et al. Efficient dense modules of asymmetric convolution for real-time semantic segmentation. *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, 2019, pp. 1–6. DOI: 10.1145/3338533.3366558

Contacts: Tatarinova Elena Alexandrovna,
tatarinova.ea@mipt.ru

Article received by the editorial office on 15.10.2025

INTRODUCTION

Autonomous (unmanned) driving technologies are currently being rapidly developed worldwide [1, 2]. The use of visual simultaneous navigation and mapping (VSLAM — Visual Simultaneous Localization and Mapping) methods allows one to obtain information about the environment using only video cameras as sensors [3–5]. Using VSLAM can be challenging in dynamic environments where moving objects such as people, cars, and bicycles are present. Their presence can lead to algorithm errors and incorrect position estimates and map construction.

In the past decade, the use of VSLAM in autonomous robotics has been rapidly developing due to its low cost and compact size. A Mono-SLAM system [7], a monochrome SLAM system for real-time operation, was proposed in [6]. [8] presents a camera tracking system for augmented reality (PTAM — Parallel Tracking and Mapping), which is proposed to divide the system into two main flows (tracking and map construction) to increase the robustness of SLAM. In [9], a direct tracking method with a probabilistic solution is proposed to take into account the influence of noisy depth values on the tracking process. In [10, 11], approaches are presented that allow, without the need for feature extraction and the use of robust matching methods in motion estimation, combining characteristic points with direct tracking of the optical flow. The direct method has become popular due to its computational efficiency [12]. However, its insensitivity to features is a significant problem. In SLAM, accurate estimation of the robot’s pose and plotting of an environment map are critical, and the limitations of the direct method may hinder the achievement of these goals. The limitations of the direct method can be overcome using the visual inertial odometry framework [13], ORB-SLAM (ORB — Oriented FAST and Rotated BRIEF) [14], and ORB-SLAM2 [15], which are feature tracking-based applications of SLAM.

In most SLAM systems, dynamic features (such as pedestrians and vehicles) are considered false data. As a result, these features are not included in the map and are not used for camera tracking. Common approaches for processing dynamic features include outlier rejection algorithms such as RANSAC (RANDOM SAMPLE CONSENSUS) [16] in ORB-SLAM [17] and robust cost functions in S-PTAM [18]. In recent years, deep learning methods have been used to improve localization and mapping in dynamic environments; the algorithms of these methods can perform object detection and segmentation in real time with sufficient accuracy [19, 20].

In [21], the SLAMIDE (Simultaneous Localization and Mapping in Dynamic Environments) method is proposed, which estimates the state of 3D features

(stationary or moving) using a generalized expectation maximization algorithm and uses reversible data linking to incorporate moving objects into a unified SLAM framework. In [17], an offline visual SLAM system, DynaSLAM, is presented, which, based on ORB-SLAM2, adds the ability to detect moving objects and improves robustness in dynamic scenarios for mono-, stereo-, and RGB-D configurations. In DynaSLAM, images are passed through a convolutional neural network (Mask R-CNN), which performs pixel-level semantic segmentation for all a priori moving objects (e.g., people or vehicles). In [22], a monochrome semantic SLAM framework called Dynamic-SLAM is proposed. The architecture uses a Single Shot Detector (SSD) which detects objects and integrates prior knowledge to semantically identify moving objects in a new detection stream. To improve the existing SSD network, which is prone to low detection rates, the paper presents a missed detection compensation algorithm based on velocity invariance in neighboring frames, which significantly improves the detection rate. In [23], Dynamic-DSO is proposed, which combines direct methods with deep learning to exclude moving candidates from key frames. All the above-mentioned studies demonstrate significant efforts aimed at addressing the challenges associated with dynamic environments in the context of VSLAM.

In this paper, we propose a DDR-SLAM system that utilizes DDRNet [24]-based object segmentation to achieve accurate and robust SLAM performance in dynamic scenarios. DDR-SLAM is a mono-, stereo-, and RGB-D SLAM system for dynamic environments that simultaneously estimates camera positions, maps, and trajectories of moving objects from camera images. As a basis, the ORB-SLAM2 system [15] was used: it allows for local construction of a map. To ensure the efficient operation of the visual SLAM system in a dynamic environment, DDRNet is used to identify and segment moving objects, such as pedestrians and vehicles, at the pixel level. DDR-SLAM successfully operates in real time under complex dynamic conditions in open areas. The system addresses localization, bundle adjustment, loop closure, sparse reconstruction, and other problems in the presence of pedestrians and vehicles.

1. GENERAL CHARACTERISTICS OF THE DEVELOPED VSLAM SOLUTION

1.1. *Applied operational tools*

To improve the accuracy of determining the position and orientation of unmanned vehicles (UVs) using the ORB-SLAM2 framework, it seems appropriate to enhance it by implementing a moving object detec-

tion unit and extracting keypoints only for those image areas that do not contain such objects. In this regard, as part of the research, a method for detecting moving objects in UV camera images using the DDRNet neural network [25] was developed. The DDR-SLAM VSLAM solution, a modification of the ORB-SLAM2 VSLAM solution, was implemented. In this solution, moving object masking is performed when keypoints are detected in images.

The implementation of the moving object detection method is based on the DDRNet-23-slim model, which enables semantic segmentation of images in real time and provides segmentation maps for images that better match reference maps than segmentation maps obtained using artificial neural networks (ANN) such as Fast-SCNN [26], DFANet [27], ERFNet [28], BiSENet [29], ICNet [30], DABNet [31], EDANet [32], etc. It is also worth noting that in order to detect moving objects in images obtained by UV-cameras, the DDRNet-23-slim model was trained on images from the KITTI (<https://www.cvlibs.net/datasets/kitti>) and Cityscapes datasets (<https://www.cityscapes-dataset.com>).

Detecting moving objects in an image using the DDRNet network involves feeding the image to the DDRNet network, which then generates binary masks corresponding to selected classes of moving objects (pedestrians, cyclists, cars, trucks, etc.). After combining all the resulting masks, a final mask is created for all moving objects detected in the image.

The developed DDR-SLAM solution operates as follows. Initially, images captured by the cameras are passed through the DDRNet network, which detects moving objects in the images. Subsequently, for each image input to the DDRNet network, a mask is generated for all moving objects detected in the image. In the next step, the generated masks are superimposed on the corresponding images, and keypoints are extracted only for those sections of the images that do not contain moving objects detected by the DDRNet network. Next, using the identified keypoints in the images, the position of the vehicle's cameras is detected and refined in the same manner as in ORB-SLAM2 to construct the vehicle's trajectory.

1.2. Segmentation of potentially dynamic content using DDRNet

For semantic image segmentation, the DDRNet neural network, characterized by high performance and enabling real-time semantic image segmentation, was used. To detect potentially moving objects, synchronized grayscale or color images are fed to the DDRNet input. The DDRNet outputs masks corresponding to classes of potentially moving objects (pedestrians, cyclists, cars, trucks, etc.). For example, if the DDRNet input is an RGB image of size $M \times N \times 3$, the output is an array of size $M \times N \times L$, where

L is the number of objects in the image. Each output channel produces its own binary mask. By combining these masks, a resulting mask can be obtained for all potentially moving objects in a single frame — a complete image segmentation map. The segmentation map shows the location and contours of each moving object detected by DDRNet.

1.3. Low-cost tracking

After segmenting moving objects in an image, the camera pose is determined using the static portion of the image. Once potentially moving objects are isolated, their boundaries typically become areas of high contrast, and characteristic keypoints appear. However, these features within the contour regions are not used in the analysis. To track the camera pose, the following steps are performed:

- feature extraction: a feature map is created by extracting keypoints from the static portion of the image frame;
- matching: matching features are found in static areas of the image;
- error minimization: we optimize the camera pose by minimizing the reprojection error, which is the difference between the observed image points and the projected points from the 3D model using the current camera pose.

The presented approach is a more computationally efficient alternative to the method used in ORB-SLAM2. Unlike ORB-SLAM2, the input to the pose prediction block in the developed system consists of RGB frames with segmentation masks. Using the masks, ORB features are extracted from image segments classified as static. Furthermore, due to the highly gradient nature of the segment contours, keypoints falling on the contours are removed.

2. EXPERIMENTAL STUDIES

2.1. Conditions of experiment

Sources of video sequences

To test the reliability and accuracy of the DDR-SLAM system in real-world dynamic conditions, a series of experiments were conducted from various perspectives. DDR-SLAM was evaluated using the KITTI Odometry Benchmark and the Cityscapes dataset. These datasets were chosen because SLAM requires synchronized stereo sequences with corrected images.

Prior knowledge of the DDRNet neural network used in the experiment

It primarily includes classes of moving objects (person, bicycle, car, motorcycle, bus, train, truck,

etc.) that can be recognized by the network. The model with prior knowledge is based on the DDRNet-23-slim network, which was trained on the ImageNet, Cityscapes, and KITTI360 datasets.

The experimental runtime

It is a computer with an Intel Core i7-9700K processor @ 3.60 GHz, 8 GB of RAM, and an NVIDIA GeForce RTX 2060 graphics card with 6 GB of video memory.

2.2. KITTI dataset-based performance evaluation

The KITTI dataset is one of the largest datasets for evaluating computer vision algorithms in the context of autonomous driving, especially for large-scale scenarios. KITTI includes multiple sequences depicting real-world street scenes, accompanied by precise location data collected using GPS and a Velodyne laser scanner. Ten sequences were selected for the experiment to compare the DDR-SLAM trajectories obtained for these sequences with reference paths to test the effectiveness of DDR-SLAM. The selected sequences include scenes on highways, urban roads, rural roads, and others. These scenes depict typical objects found in real street environments, such as pedestrians, bicycles, motorcycles, moving vehicles, parked cars on the side of the road, etc.

To evaluate the performance of ORB-SLAM2 and DDR-SLAM, absolute pose error (APE) estimates were used, including root mean square error (RMSE), mean error (Mean), median error (Median), and standard deviation (STD). The Python evo package was used to obtain APE estimates and generate motion trajectories. Trajectory alignment with the reference path was not used when calculating APE, which allowed us to obtain objective APE estimates without underestimating them. Tab. 1 presents statistical estimates of the APE error obtained for trajectories reconstructed using DDR-SLAM and ORB-SLAM2 solutions for the sequences from the KITTI set employed for the testing.

Tab. 1. Statistical estimates of the APE error obtained for trajectories reconstructed using DDR-SLAM and ORB-SLAM2 solutions for the sequences from the KITTI set taken for the testing

2.3. Performance evaluation using the Cityscapes dataset

The Cityscapes dataset was used to evaluate the performance of DDR-SLAM in a practical setting. Although the Cityscapes dataset is primarily designed

for scene understanding and image segmentation, it includes a series of over 100 000 frames recorded by a stereo camera at a high frame rate (17 fps). The cameras in the Cityscapes setup have a wide dynamic range, enabling the capture of high-quality images in challenging street lighting conditions.

Unfortunately, the Cityscapes dataset does not provide benchmark data for each image sequence, only coarse GPS/vehicle odometry information, making it difficult to accurately evaluate the performance of visual odometry.

However, the Cityscapes dataset images contain many moving objects, including cars, trams, and trucks. Smaller segments, called S1, S2, S3, and S4, are extracted from a long sequence of the Cityscapes dataset, namely the Frankfurt stereo sequence. The extracted segments have varying lengths from 400 to 5000 frames. For each small sequence, reference camera poses were calculated using GPS coordinates and the Mercator projection method.

To eliminate inaccuracies in the reference GPS data, an SE(3) transformation was used to align the trajectories. This resulted in a more accurate match between the estimated trajectories and the reference trajectories. Tab. 2 presents the statistical estimates of the APE obtained for trajectories reconstructed using the DDR-SLAM and ORB-SLAM2 solutions for the Frankfurt sequence fragments from the Cityscapes dataset.

Tab. 2. Statistical estimates of the APE error obtained for trajectories reconstructed using DDR-SLAM and ORB-SLAM2 solutions for the Frankfurt sequence fragments from the Cityscapes dataset used for testing.

A description of the Frankfurt sequence fragments from the Cityscapes dataset used for testing DDR-SLAM is presented in Tab. 3 and Fig.

Tab 3. Description of the Frankfurt sequence fragments from the Cityscapes dataset used to test DDR-SLAM

Note: The numbers in parentheses indicate the first and last frames of the Frankfurt sequence, which define sequences S1–S4.

Fig. Vehicle motion scenarios realized in sequences S1 (a), S2 (b), S3 (c), and S4 (d)

CONCLUSION

This paper presents DDR-SLAM as a scalable mono-/stereo visual SLAM system based on semantic awareness and capable of real-time operation in dynamic environments. DDR-SLAM is tested on real-world data. Quantitative performance evaluations of DDR-SLAM confirm that the method effectively handles challenging situations where traditional VSLAM

methods struggle and maintains its effectiveness in static environments, although it may have limitations when objects assumed to be static are actually moving, or when objects assumed to be dynamic are actually stationary. Future work plans to add tracking and geometric plausibility testing to address these issues and improve the accuracy of the proposed DDR-SLAM method.