DOT: Динамическое отслеживание объектов для визуального SLAM

Препринт - сентябрь 2020 г.

5 авторов, в том числе:

Ирен Баллестер
Университет Сарагосы
2 ПУБЛИКАЦИИ **1** ЦИТИРОВАНИЕ

ПОСМОТРЕТЬ ПРОФИЛЬ

Хавьер Сивера
Университет Сарагосы
90 ПУБЛИКАЦИЙ **4 111** ЦИТИРОВАНИЙ

ПОСМОТРЕТЬ ПРОФИЛЬ

Клаус Х. Штробль
Немецкий аэрокосмический центр (DLR)
31 ПУБЛИКАЦИЯ **627** ЦИТИРОВАНИЙ

ПОСМОТРЕТЬ ПРОФИЛЬ

Рудольф Трибель
Технический университет Мюнхена
178 ПУБЛИКАЦИЙ **3,562** ЦИТИРОВАНИЙ

ПОСМОТРЕТЬ ПРОФИЛЬ

DOT: Динамическое отслеживание объектов для визуального SLAM

Ирен Баллестер [1,2]
iballestercampos@gmail.com

Алехандро Фонтан [1,2]
alejandro.fontanvillacampa@dlr.de

Хавьер Сивера [1]
jcivera@unizar.es

Клаус Х. Штробль [2]
klaus.strobl@dlr.de

Рудольф Трибель [2,3]
rudolph.triebel@dlr.de

[1] Университет Сарагосы

[2] Немецкий аэрокосмический центр (DLR)

[3]Технический университет Мюнхена

Рисунок 1. **Верхний ряд**: Кадры соответствуют ORB-SLAM2 [17], оценивающему траекторию камеры из потока изображений в бенчмарке The KITTI [7]. **Средний ряд**: Модифицированный ORB-SLAM2, работающий с масками сегментации, сгенерированными DOT, которые различают движущиеся и статичные объекты. **Нижний ряд**: Модифицированный ORB-SLAM2 с использованием масок сегментации, предоставленных Detectron2 [23], которые кодируют все потенциальные динамические объекты. Обратите внимание, что от самой статичной сцены (левая колонка) до самой динамичной (правая колонка) DOT способен избегать движущихся объектов, сохраняя при этом статичные. DOT достигает компромисса между этими двумя противоположными сценариями, оценивая фактическое состояние движения объектов, чтобы получить более высокую надежность и точность отслеживания.

Аннотация

В этой статье мы представляем DOT (Dynamic Object Track- ing), фронт-энд, который добавляется к существующим системам SLAM и может значительно повысить их надежность и точность в высокодинамичных средах. DOT сочетает в себе сегментацию экземпляров и многоракурсную геометрию для создания масок для динамических объектов, что позволяет системам SLAM, основанным на жестких моделях сцены, избегать таких областей изображения в своей опти- мизации.

Чтобы определить, какие объекты действительно движутся, DOT сегментирует первые экземпляры потенциально динамичных объектов, а затем, с учетом оценки движения камеры, отслеживает такие объекты, минимизируя ошибку фотометрической репроекции. Такое кратковременное отслеживание повышает точность сегментации по сравнению с другими подходами. В итоге создаются только активные динамические маски. Мы оценили DOT с ORB-SLAM 2 [17] на трех публичных наборах данных. Наши результаты показывают, что наш подход значительно улучшает точность и устойчивость ORB-SLAM 2, особенно в высокодинамичных сценах.

1. Введение

Одновременная локализация и картирование, известная под аббревиатурой SLAM, является одной из фундаментальных возможностей для автономной навигации роботизированных платформ [4]. Ее целью является совместная оценка движения робота и карты его окружения на основе информации встроенных датчиков. Визуальный SLAM, для которого датчиками являются в основном или исключительно камеры, является одной из самых сложных и в то же время актуальных конфигураций.

Несмотря на значительные достижения в области SLAM за последние два десятилетия, большинство современных систем по-прежнему предполагают статичную среду, где относительное положение между точками сцены не меняется, а единственное движение осуществляется камерой. При таком допущении модели SLAM приписывают визуальные изменения исключительно относительному движению камеры. Обычный подход [16, 17] заключается в моделировании динамических областей как аут-лиеров, игнорируя их в процессе отслеживания позы и построения карты. Однако в течение нескольких кадров, пока такие динамические области не будут отброшены как выбросы, их данные используются в оптимизации SLAM, что вносит ошибки и несоответствия в оценку карты и положения камеры. Более того, для методов SLAM на основе признаков, которые

отслеживание небольшого числа точек изображения, ошибки, вызванные относительно небольшим числом совпадений в динамических областях, являются значимыми и могут привести к сбою системы.

Мир и реальные приложения, в которых должен работать робот или AR-система, далеко не статичны. В качестве показательных примеров можно привести автономную навигацию автомобилей или дронов, AR в многолюдных сценах или даже задачи по исследованию планет, где плохая текстура делает системы SLAM нестабильными в присутствии теней или других роботов. Разработка систем SLAM, достаточно надежных для работы в высокодинамичных средах, необходима для многих приложений.

Как показано на рисунке, целью данной работы является разработка стратегии обработки изображений, которая повышает устойчивость визуальной системы SLAM в динамических средах. Наш конкретный вклад заключается в разработке "Динамического отслеживания объектов" (DOT), фронт-энда, который сочетает в себе сегментацию экземпляров с многоракурсной геометрией для отслеживания движения камеры, а также движения динамических объектов, используя прямые методы [5]. Результатом такой предварительной обработки является маска, содержащая динамические части каждого изображения, которую система SLAM может использовать, чтобы избежать установления соответствий в таких областях.

Наши экспериментальные результаты на трех различных публичных наборах данных показывают, что наша комбинация семантической сегментации и геометрического слежения превосходит современные достижения в динамических сценах. Мы также считаем важным, что DOT реализована как независимый внешний модуль и, следовательно, легко подключается к существующим системам SLAM. Поскольку DOT включает кратковременное отслеживание маски, мы избегаем сегментации всех кадров в последовательности, что значительно экономит вычисления. Наконец, хотя мы настроили и оценили DOT для конкретной области автомобильной навигации, наша стратегия может быть использована и в других приложениях.

2. Связанная работа

SLAM в динамических средах является открытой исследовательской проблемой с большой научной библиографией. Мы разделим различные подходы на три основные категории.

Первая категория, наиболее общая, моделирует сцену как набор нежестких частей, включая, таким образом, де-формируемые и динамические объекты [18, 12, 13]. Хотя эта линия поиска является наиболее общей, она также является наиболее сложной. В данной работе мы будем предполагать внутриобъектную жесткость, что является предпосылкой для двух других категорий динамических визуальных SLAM.

Вторая категория направлена на повышение точности и надежности визуального SLAM путем реконструкции только статической части сцены. Динамические объекты сегментируются и учитываются для отслеживания позы камеры и оценки карты. В этом направлении DynaSLAM [2], построенная на базе ORB-SLAM2 [17], нацелена на оценку карты статической части сцены и повторное использование ее в долгосрочных приложениях. Динамические объекты являются

устраняется путем сочетания 1) семантической сегментации для потенциально движущихся объектов и 2) многоракурсной геометрии для выявления несоответствий в жесткой модели. Для семантической сегментации используется маска R-CNN [9], которая обнаруживает и классифицирует объекты в сцене по различным категориям, некоторые из которых были предварительно заданы как потенциально динамичные (*например*, автомобиль или человек). DynaSLAM был разработан для маскировки всех потенциально подвижных объектов в сцене, что приводит к более низкой точности по сравнению с оригинальным ORB-SLAM2 в сценах, содержащих потенциально подвижные объекты, которые на самом деле не движутся (*например,* сцены с большим количеством припаркованных автомобилей). Целью данной работы является именно преодоление этой проблемы, поскольку только те объекты, которые движутся в данный момент, будут помечены как динамические.

Другой работой, в которой используется аналогичный подход, является StaticFusion [20], плотная RGB-D визуальная SLAM система, в которой сегментация выполняется с использованием 3D реконструкции фона сцены как способа распространения временной ин- формации о статических частях сцены.

Наконец, третье направление работы в динамическом визуальном SLAM, которое выходит за рамки сегментации и подавления динамических объектов, включает такие работы, как MID-Fusion [24], MaskFusion [19], DynSLAM [1] и ClusterVO[11]. Их целью является одновременная оценка положения камеры и нескольких динамических объектов. Для этого в MID- Fusion [24] и MaskFusion [19] создаются подкарты каждого возможного движущегося объекта и выполняется совместная оценка положения объектов и камеры.

Большинство упомянутых систем [24, 1, 19, 11, 2] используют методы глубокого обучения, которые в некоторых случаях не могут быть реализованы в реальном времени из-за узкого места, вызванного ограниченной частотой сегментации сети. Разработанный в данной работе вклад устраняет требование сегментации всех кадров, что позволяет системе не зависеть от частоты сегментации сети, тем самым позволяя реализовать ее в реальном времени.

3. DOT

3.1. Обзор системы

На рисунке 2 показан обзор нашего предложения. На вход DOT подаются RGB-D или стереоизображения с определенной скоростью видео, а на выходе получается маска, кодирующая статические и динамические элементы сцены, которая может быть непосредственно использована системами SLAM или одометрии.

Первый блок (*Instance Segmentation*) соответствует CNN, который сегментирует по пикселям все потенциально динамичные объекты. В наших экспериментах, проведенных с использованием наборов данных автономного вождения, только автомобили были сегментированы как потенциально движущиеся. Как будет показано далее, поскольку DOT отслеживает маску от кадра к кадру, эту операцию не нужно выполнять в каждом кадре.

Блок *обработки изображения* извлекает и разделяет

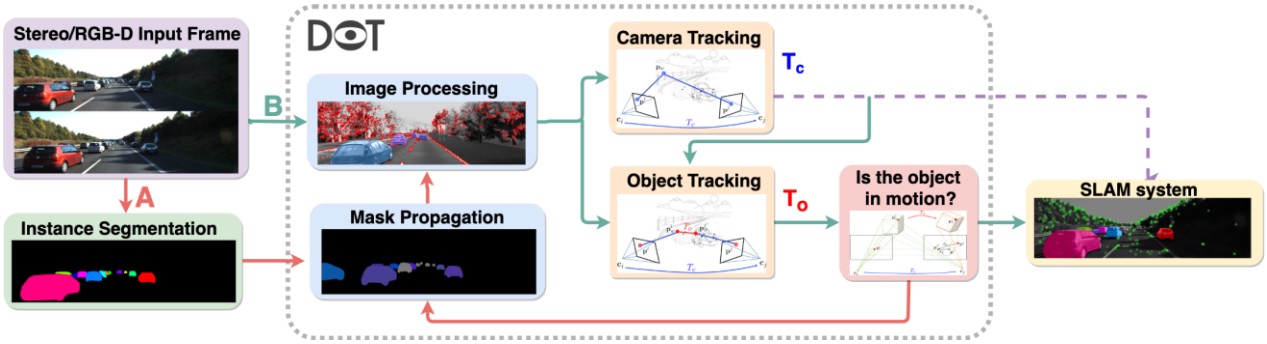

Рисунок 2. **Обзор DOT**. Путь A (красный) показывает обработку кадров, которые получают маску сегментации из сети. Путь B (зеленый) показывает обработку для кадров, которые получат маску сегментации, геометрически распространяемую DOT.

точки, относящиеся к статичным областям изображения, и точки, находящиеся в динамичных объектах. Поза камеры отслеживается с использованием только статической части сцены. На основе этого блока и с учетом позы камеры движение каждого из сегментированных объектов оценивается независимо (*отслеживание объектов*).

Следующий блок (*Движется ли объект?*) определяет, используя геометрические критерии, действительно ли объекты, помеченные сетью как потенциально динамичные, движутся. Эта информация используется для обновления масок, кодирующих статические и динамические области каждого кадра, а также для питания связанной одометрии / визуальной системы SLAM.

Наконец, DOT генерирует новые маски из оценок движения объектов (*Mask Propagation*), поэтому не каждый кадр должен быть сегментирован сетью (см. рис. 3). Учитывая значительную вычислительную нагрузку при сегментации экземпляров, это может быть существенным преимуществом DOT по сравнению с другими современными методами.

3.2. Сегментация экземпляров

Мы используем глубокую сеть Detectron2 [23] для сегментации всех потенциально подвижных объектов, присутствующих на изображении. Выходные данные сети были модифицированы для получения на одном изображении всех масок сегментации. Области изображения, не отнесенные к потенциально подвижным категориям, получают метку "фон" и в последующих блоках считаются статичными.

Мы используем базовую модель COCO Instance Segmentation с маской R-CNN R50-FPN 3x [14][15]. Классы были ограничены теми, которые считаются потенциально подвижными, исключая людей, поскольку отслеживание людей выходит за рамки данной работы. Если потребуются другие категории, сеть может быть доработана с использованием этих весов в качестве отправной точки или обучена с нуля на собственном наборе данных.

Для последовательного отслеживания объектов на нескольких кадрах мы включили этап согласования между масками, вычисленными DOT, и масками, предоставленными сетью.

Новые обнаружения, которые не могут быть сопряжены ни с одним из существующих объектов, используются для инициализации новых экземпляров.

3.3. Отслеживание камеры и объектов

На основе сегментации экземпляров, выполненной на предыдущем этапе, мы стремимся оценить движение камеры и динамических объектов. Поскольку движение камеры и движение объектов связаны на изображениях, мы проводим оценку в два этапа. Сначала мы находим позу камеры как относительное преобразование $\mathbf{T}_c \in \mathbf{SE}(3)$, а затем вычитаем его для оценки движения объекта $\mathbf{T}_o \in \mathbf{SE}(3)$.

Наша оптимизация связана с недавними подходами прямая визуальная одометрия и SLAM [5], целью которых является поиск движения, минимизирующего ошибку фотометрической репроекции. **Оптимизация.** Как для вычисления положения камеры, так и для последующей оценки движения объекта, мы выполняем Оптимизация по Гауссу-Ньютону

$$(\mathbf{J}^T \Sigma_r^{-1} \mathbf{J}) \mathbf{x} = -\mathbf{J}^T \Sigma_r^{-1} \mathbf{r}, \quad (1)$$

где $\mathbf{J} \in \mathbb{R}^{n \times 6}$ содержит производные остаточной функции (уравнения (3) и (5)), а $\Sigma_r \in \mathbb{R}^{n \times n}$ - ди-агональная матрица, содержащая ковариации фотометрических остатков $\mathbf{r} \in \mathbb{R}^n$. Позиционные инкременты алгебры Ли

^вектор в ма^тричном представлении касательного

пространства

преобразования с использованием умножения левой матрицы и оператора экс- поненциальной карты $\exp(\ )$. Обе оптимизации ини- циализируются с моделью постоянной скорости и многомасштабным изображением пирамиды для облегчения сходимости.

Отслеживание камеры. Движение камеры оценивается по статическим точкам сцены и многоракурсным ограничениям [8], предполагая, что калибровка камеры и глубина точек известны. Проекция статической точки $\mathbf{p}$ с ее пиксельных координат $\mathbf{p}^j$ в опорном кадре F_j на ее соответствующие координаты $\mathbf{p}'$ в кадре F_i осуществляется следующим образом:

$$\mathbf{p}' = \Pi(\mathbf{T}_c \Pi_c^{-1}(\mathbf{p}^j, z_j)), \quad (2)$$

$\in$

$\in$

$\in$

$\in$

$\mathbf{x} \in \mathfrak{se}(3)$, причем - является оператором

[21], выражаются в виде вектора $\mathbf{x} \in \mathbb{R}^6$. Мы обновим

$\mathbf{P}$

$\in \mathbf{P}$

где Π и Π^{-1} соответствуют моделям перспективной проекции и обратной проекции, соответственно, а z_j - глубина точки в системе отсчета F_j.

Поза камеры оптимизируется путем минимизации ошибки репроекции фото-метрии

$$\sum_{p \in P} \gamma\left(I_j(\mathbf{p}) - I_i\left(\Pi\left(\exp(\hat{\mathbf{x}}_{se(3)})\mathbf{T}_{cj}^{-1}\Pi_c^{-1}(\mathbf{p}, z_j)\right)\right)\right), \quad (3)$$

которая вычисляется как сумма всех разностей интенсивностей между точками в их опорном кадре и их проекцией на отслеживаемый кадр. Мы используем норму Хьюбера γ.

Отслеживание объекта. После оценки $\mathbf{T}_c$, позы каждого потенциально динамического объекта может быть оценена аналогичным образом с использованием точек изображения, принадлежащих такому объекту. Моделируя потенциально динамический объект как твердое тело с позой $\mathbf{T}_o$, проекция каждой точки $\mathbf{p}$ в кадре F_j к его координатам в кадре F_i является:

$$\mathbf{p}'_{ij} = \Pi(\mathbf{T}_c \mathbf{T}_o \Pi_o^{-1}(\mathbf{p}, z_j)). \quad (4)$$

Аналогично уравнению 3, мы оцениваем $\mathbf{T}_o$ путем минимизации следующей ошибки фотометрической репроекции

$$\sum_{p \in Q} \gamma\left(I_j(\mathbf{p}') - I_i(\Pi(\mathbf{T}_c \exp(\hat{\mathbf{x}}_{se(3)})\mathbf{T}_o \Pi_o^{-1}(\mathbf{p}, z_j)))\right). \quad (5)$$

3.4. Качество отслеживания, выбросы и окклюзии

Окклюзии, изменения условий освещения и ошибки сегментации оказывают значительное влияние на точность определения объектов и положения камеры. Как видно из алгоритма 1, мы разработали несколько стратегий, которые мы применяем после этапа отслеживания объектов, чтобы уменьшить их влияние.

Качество отслеживания. Внешний вид динамических объектов сильно меняется, что приводит к большим ошибкам отслеживания. Для моделирования внешнего сходства мы использовали коэффициент корреляции Пирсона $\varphi_o \in [-1, 1]$. Эта метрика отражает степень линейной корреляции между эталонной интенсивностью точек и их соответствующими оценками, следовательно, она инвариантна к изменениям усиления и смещения. Обратите внимание, что эта метрика также может быть применена к отслеживанию камеры φ_c, хотя изменения внешнего вида фона обычно менее выражены.

Алгоритм 1 Динамическое отслеживание объектов

```
функция OBJECT TRACKING(P, Q, O)
                                  d P = статические точки
                                  d Q = динамические точки
                                  d O = множество объектов
    mask ← ∅d Динамическая маска для вычисления

    {T_c , φ_c } ← отслеживание камеры (P) d
Отслеживание камеры
    end if
    для объекта O indod                Отслеживание объектов
        if is visible (object , T_c) then
            {T_o , φ_o } ← объект трека (T_c, Q_o, маска)

        if φ_o < th_φ then break
        end if
        объект ← отклонение выбросов (φ_o)
        маска ← обновить маску (объект)
                ←

                    maskis объект движется? (объект)
        end if

    end for

    маска возврата
конец функции
```

Окклюзии. Динамические объекты могут заслонять друг друга другое. Удаление окклюдированных частей как выбросов не было удовлетворительным. эффективным в наших экспериментах. Мы применили стратегию, состоящую из отслеживания объектов от ближайших до дальних.

Отбраковка выбросов. Общий подход для обнаружения аутлиеров заключается в определении абсолютного порога фотометрической ошибки (3) (5). Более сложные работы [5] адаптируют его в соответствии с медианным остатком, размытием движения или изменениями освещенности. Как показано на рисунке 4, мы предлагаем установить порог относительно линейной зависимости между интенсивностями, чтобы ошибки не зависели от фотометрических изменений в изображении.

thest, последовательно обновляя их соответствующие маски. Таким образом, на каждой итерации мы обновляем точки более удаленных объектов, которые были закрыты более близкими.

3.5. Находится ли объект в движении?

Этот блок получает на вход матрицы преобразования камеры $\mathbf{T}_c$ и объектов $\mathbf{T}_o$ и оценивает, движутся объекты или нет. Его выходной сигнал, который будет использоваться системами SLAM или одометрии, - это маски, хранящие области изображения, занятые динамическими объектами, и то, находятся ли они в движении или нет. Маски получаются путем проецирования пикселей каждого объекта на новый кадр с использованием $\mathbf{T}_c$ и $\mathbf{T}_o$, оцененных на предыдущем этапе.

Наблюдение движения объекта непосредственно в $\mathbf{T}_o$ создает, из-за распространяющегося шума изображения, трудности в установлении абсолютных порогов, которые определяют, находится ли объект в движении. В данной работе мы решили наблюдать за движением объектов с помощью двумерных измерений изображения. Мы обозначаем нашу метрику как *динамическую диспаратность*, которая представляет собой расстояние в пикселях между проекцией точки как если бы она была статичной $\mathbf{p}^i$ и ее фактической проекцией $\mathbf{p}^i$. Для каждого объекта мы вычисляем
медиана динамических неравенств его точек $\tilde{\mathbf{p}} \in Q$:

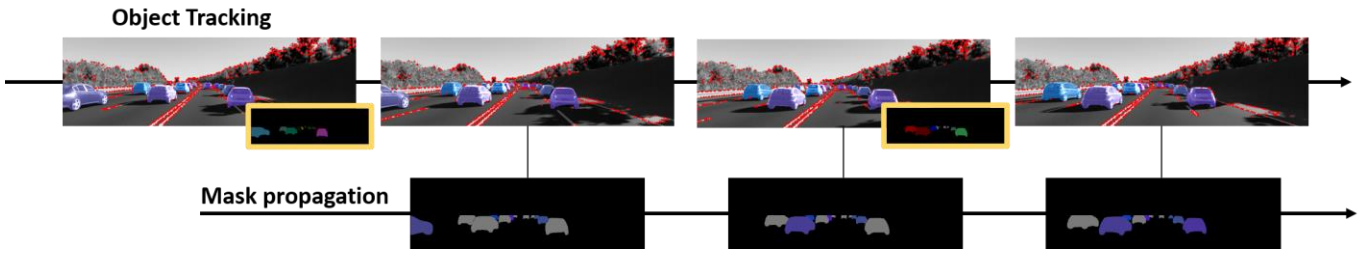

Рисунок 3. **Образец сегмента потока вычислений**. В верхнем ряду показано, как DOT оценивает слежение за камерой и объектами. Обратите внимание, что маски сегментации из сети (желтые кадры) нужны не во всех кадрах. В нижнем ряду показаны сегментационные маски, генерируемые DOT, которые кодируют классификацию движения: в движении (цвет), статичный (черный) и ненаблюдаемый (серый).

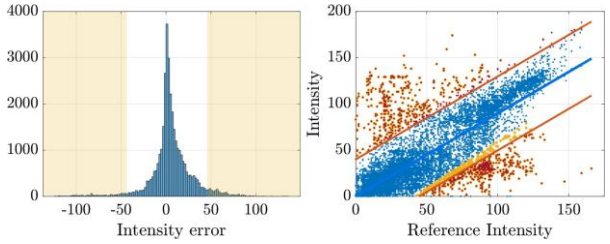

Рисунок 4. **Отбраковка выбросов**. **Слева:** гистограмма фотометрических ошибок для объекта. Затененная область соответствует точкам, повторно перемещенным с постоянным порогом. **Справа:** Линейная зависимость между интенсивностями. Обратите внимание на различные точки, помеченные как выбросы абсолютным (желтый) и относительным (красный) порогами из-за изменения фотометрии.

Рисунок 5. **Диспаратность в сравнении с энтропией**. Сравнение динамических диспропорций, создаваемых различными объектами в движении. Обратите внимание, что объекты с высокими значениями энтропии (более яркий красный цвет) производят большее смещение пикселей изображения.

$$d_d = \text{medp}\ \cdots^i, \mathbf{p}^i, \forall p \in \tilde{Q}. \quad (6)$$

Трехмерное движение точки создает различные изображения в зависимости от 1) координат изображения, 2) глубины и 3) относительного угла между направлениями движения объекта и камеры.

Из нелинейной оптимизации позы (см. уравнение. (1)) мы можем получить неопределенность в оценке движения объекта $\Sigma_x = (\mathbf{J}^T \Sigma r\text{-}1J)^{-1}$. Предполагая, что k-мерный
Гауссово распределение, его дифференциальная энтропия равна:

$$H(\mathbf{x}) = \frac{1}{}\log((2\pi e)^k|\Sigma|). \quad (7)$$

Дифференциальную энтропию можно рассматривать как неопределенность позы, полученную в результате минимизации фотометрических остатков. Другими словами, наблюдения трехмерных движений

с высокими значениями энтропии приведут к большим смещениям пикселей изображения (см. Рисунок 5). С другой стороны, наблюдения с низкой энтропией приведут к небольшим расхождениям изображения.

Исходя из этого, алгоритм классификации движения объектов работает следующим образом. Мы сравниваем динамические диспропорции (6) с переменным порогом $\Delta d = f(H(x))$, который плавно растет с увеличением энтропии. Мы помечаем как "в движении" все те объекты, чья динамическая диспаратность превышает этот порог ($d_d > \Delta d$). Для каждого значения ниже порога энтропии H_{min} мы предполагаем, что движение объекта не может быть обнаружено. Таким образом, для обозначения объекта как статичного необходимо, чтобы движение было наблюдаемым ($H(x) > H_{min}$) и чтобы медиан динамической диспаратности был меньше переменного порога ($d_d < \Delta d$).

Хотя выбор оптимальной функциональной формулировки требует дальнейшего изучения, это выражение удовлетворяет требованиям и показало хорошие результаты в данной работе (см. раздел 4.1). Рисунок 3 - пример маски, распространяемой DOT. Объекты, помеченные как "в движении", представлены цветом, в то время как объекты, помеченные как "статичные", исчезают в черном цвете. Автомобили, представленные серым цветом, - это те, которые нельзя отнести ни к статичным, ни к динамичным.

3.6. Распространение маски

DOT использует две маски сегментации, доступные в каждом кадре: одна создается нейронной сетью, а другая распространяется из предыдущего кадра. Искажение одной сегментации в другую позволяет надежно связать экземпляры, найденные в разных кадрах, в один и тот же 3D-объект.

Распространение состояния. Отнесение новых семантических экземпляров к уже существующим объектам позволяет нам предсказывать их движение (что очень важно для быстро движущихся объектов). Кроме того, можно сохранить классификацию движения в случае, если объект перемещается в позицию, где движение не наблюдается (см. раздел 3.3).

Независимая сегментация. Наше предложение позволяет распространять семантические маски сегментации из исходного семени во времени и пространстве, устраняя необходимость сегментации каждого кадра. Запуск нейронной сети на более низкой частоте облегчает отслеживание объектов в реальном времени на низкопроизводительных устройствах.

платформы. Дополнительным преимуществом является то, что DOT может заполнить пробелы, когда сеть временно теряет представление об объекте между последовательными изображениями.

4. Экспериментальные результаты

Хотя потенциальные применения DOT охватывают широкий спектр - от обнаружения объектов до дополненной реальности или автономного вождения, в данной работе мы проводим интенсивную оценку, чтобы продемонстрировать, в какой степени "знание о движении объектов" может повысить точность системы SLAM.

4.1. Оценка в сравнении с базовыми показателями

Базовые показатели. Наши эксперименты оценивают траекторию движения камеры.

тории с использованием современной системы SLAM в трех различных конфигурациях. В частности, мы используем ORB-SLAM2 [17], с ее RGB-D и стерео реализацией. Три конфигурации, разработанные для оценки DOT, следующие:

Без масок: ORB-SLAM2 запускается с помощью авторской реализации на немодифицированных изображениях. Предполагается жесткая сцена, поэтому все точки на изображениях (включая те, которые принадлежат движущимся объектам) могут быть выбраны ORB-SLAM2.

Маски динамических объектов: ORB-SLAM2 получает на вход, помимо изображений, маски динамических объектов, содержащие потенциально динамические объекты, находящиеся в движении. Мы модифицировали ORB-SLAM2 таким образом, чтобы он не извлекал точки из таких движущихся объектов.

Все маски: ORB-SLAM2 получает все маски, полученные сетью сегментации экземпляров. В этой конфигурации все потенциально динамические объекты удаляются без проверки того, движутся они на самом деле или нет.

Seq.	ATE [м]			ATE/ATE		
	маски		маски	маски	best	маски
01	**1.10**	DOT 8	1141.3	1.00	1.04	1.26
06	0.11	**0.07**	0.081	.67 1.	001.18	
18	4.77	**1.00**	1.50	4.79 1.		001.51
20	29.42	**9.12**	13.54	3.23 1.		001.49
enorm	192.6%	100.0%	137.8%			

Seq.	ATE NoDOTAllмаскимаскимаскимаскимаскимаскимаскимаскимаскимаскимаскимаски				[m]ATE/ATE best NoDOTAllмаскимаскимаскимаскимаскимаскимаскимаскимаскимаскимаскимаски		
0	**1.771**		.802.081			1.02	1.18
1		.00				1.21	1.33
2	**6.377**		.718.451			1.00	1.04
3		.00					
4	3.723	**.703**	.841	.01			
5				1.00		1.01	1.00
6		0.400.400.40		1.12		1.09	1.00
7		0.270.	**260.24**	1.03		1.00	1.14
8	0.	**400.**	390.45				
9	**0.**		630.680.671	.00	1.08	1.07	
10	0.		**520.510.511**	2.71		1.00	.01 1.
		001.00					
enorm	112.7%	100.0%	130.3%				

Таблица 1. DOT против базовых уровней (*без масок* и со *всеми масками*) в V- KITTI. **Слева:** ATE [м]. **Справа:** Превышение ATE над лучшим ATE для каждой последовательности.

Подмножества последовательностей. Мы оцениваем вышеуказанные конфигурации на трех подмножествах последовательностей из набора KITTI Vision Bench- mark Suite [7], содержащего стереопоследовательности городских и дорожных сцен, записанных с автомобиля и используемых для исследований в области автономного вождения. Мы используем Virtual KITTI [6] [3], син- тетический набор данных, состоящий из 5 последовательностей, виртуально клонированных из KITTI [7], KITTI *Odometry*, предопределенное подмножество KITTI *Odometry*.

Таблица 2. DOT по сравнению с базовыми показателями (*без масок* и со *всеми масками*) в *одометрии* KITTI. **Слева:** ATE [м]. **Справа:** ATE по сравнению с лучшим ATE для каждой последовательности.

Seq.	ATE			[m]ATE/ATEbest		
	NoDOT маски	DOTAll маски	NoDOTAll маски	NoDOT маски	DOTAll маски	NoDOTAll маски
0926-0009	**1.23**	1.24	1.44	1.00	1.01	1.17
0926-0013	**0.26**	**0.26**	0.27	1.00	1.00	1.03
0926-0014	0.86	0.82	**0.78**	1.11	1.06	1.00
0926-0051	0.37	**0.36**	0.37	1.02	1.00	1.02
0926-0101	**8.66**	10.26	12.37	1.00	1.18	1.43
0929-0004	0.32	**0.30**	**0.30**	1.08	1.03	1.00
1003-0047	13.81	**1.25**	2.23	11.01	1.00	1.78
εnorm242	3%	100,0%	115,9%			

Таблица 3. DOT по сравнению с базовыми показателями (*без масок* и со *всеми масками*) в KITTI
Сырой. **Слева:** ATE [м]. **Справа:** Превышение ATE над лучшим ATE для каждой последовательности.

последовательности, специально разработанные для разработки и оценки систем визуальной одометрии, и выборка последовательностей, отобранных из *необработанной* секции KITTI из-за большого количества движущихся объектов [10].

Мы запускаем RGB-D версию ORB-SLAM2 в виртуальном KITTI, поскольку предоставляются синтетические изображения глубины, в то время как для других подмножеств мы запускаем стерео версию ORB-SLAM2 над цветными стереопарами. В качестве исходной истины для реальных последовательностей используется точная система локализации GPS. **Метрики оценки.** Как это обычно бывает при оценке SLAM в реальном времени, чтобы учесть не детерминированные эффекты, мы запускаем каждую конфигурацию 10 раз для каждой последовательности и сообщаем медианные значения. Все эксперименты проводились на ноутбуке с процессором Intel Core i5 и 8 ГБ оперативной памяти.

Мы сообщаем об абсолютной ошибке траектории (ATE), как указано в [22], которая представляет собой среднеквадратичную ошибку (RMSE) оцененного положения всех кадров относительно GPS-истины после выравнивания обеих траекторий.

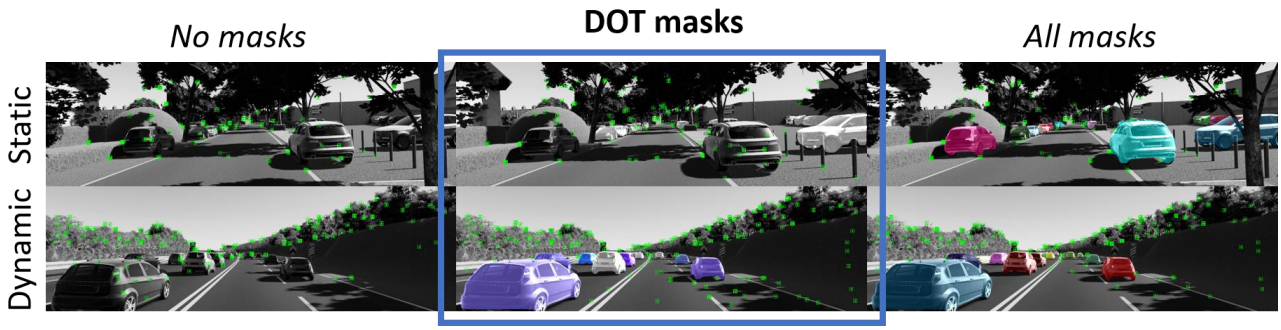

Рисунок 6. **Адаптация содержания сцены**. Примеры результатов для трех исследованных конфигураций. **Слева**: *без масок*. **Центр**: *DOT-маски*. **Справа**: *Все маски*. В верхнем ряду показана статичная сцена, в которой настройка "*Все маски"* отбрасывает все точки статичных объектов, которые могут способствовать точности отслеживания. Напротив, в нижнем ряду показано, как настройка "*Без масок"* позволяет извлечь точки движущихся объектов, которые могут привести к сбою системы. В обоих случаях недостаточное понимание сцены ухудшает работу SLAM. DOT успешно идентифицирует припаркованные автомобили как статичные, а движущиеся - как динамичные. Обратите внимание, как DOT достигает компромисса между этими двумя противоположными сценариями, оценивая фактическое состояние движения объектов, что приводит к лучшей оценке траектории.

Для более легкого сравнения между DOT и двумя другими конфигурациями, мы сообщаем среднее значение ошибок, нормированных на значение, полученное с помощью DOT для каждой последовательности.

$$\varepsilon norm = \frac{1}{n} \sum_{i=0}^{n} \frac{\varepsilon i}{\varepsilon_{DOT}}.$$

В правых колонках таблиц 1, 2, 3 показано значение ATE, нормированное на лучшее значение ATE в каждой последовательности из трех конфигураций. Таким образом, значение, равное 1, идентифицирует лучший результат, в то время как значения > 1 указывают на более низкую производительность. Цветовая шкала показывает соотношение ошибок между лучшим результатом (зеленый) и худшим (красный).

Точность отслеживания. Данные ATE в Таблице 1, соответствующие последовательностям V-KITTI, показывают улучшение точности нашей системы на 92,6% и 37,8% по сравнению с конфигурациями "*Без масок"* и "*Все маски",* соответственно. Кроме того, DOT показывает лучшие результаты для 3 из 5 оцениваемых последовательностей.

Таблица 2 содержит результаты ATE для 11 траекторий KITTI *Odometry,* оцененных с тремя различными конфигурациями. В этом случае DOT получает общую производительность, которая на 12,7% и 30,3% лучше, чем *Без масок* и *Все маски*, соответственно. По сравнению с V-KITTI, эта группа последовательностей содержит меньше динамических элементов, поэтому использование масок даже вредно. Согласно спецификации набора данных, точность определения положения камеры на местности, собранной GPS, составляет не более 10 см. Поэтому между тремя конфигурациями в последовательностях 3, 4, 5, 6, 7 и 10 нет существенных различий. Считается, что это связано с небольшим количеством движущихся объектов, а также с богатой текстурой изображений, которая обеспечивает большое количество статичных точек для оценки движения камеры.

Различия между последовательностями и методами более очевидны в последнем наборе последовательностей, представленном в таблице 3, характеризующемся обилием движущихся объектов. В целом, DOT достигает улучшения точности ATE на 142,3% по сравнению с методом "*Без масок"* и на 15,9% по сравнению с методом "*Все маски*". Еще раз отметим, что отбрасывание динамических объектов в последовательности 1003-0047 значительно снижает ошибки отслеживания. Последовательности 0926-0009, 0929-0004 и 1003-0047 были клонированы для создания синтетических последовательностей V-KITTI (1, 18 и 20).

Как и ожидалось, поскольку содержание сцен идентично, то и

качественный анализ результатов.

Цветовая шкала, используемая в таблицах 1, 2, 3, показывает, как DOT стремится приблизиться к наилучшему решению, когда оно не является самой точной траекторией (зеленый цвет). Это доказывает, что, хотя использование масок может быть удобным, точность значительно повышается, если удаляются только те объекты, которые были проверены как находящиеся в движении. Эти результаты показывают, что DOT достигает стабильно хорошей производительности как для статических, так и для динамических сцен.

Адаптация содержания сцены. На рисунке 6 показаны два сценария, которые влияют на точность SLAM в сцене с динамическими объектами. В нижнем ряду показана дорога, где все автомобили находятся в движении (секвенция 20 в таблице 1). Высокий динамизм всех транспортных средств в сцене нарушает предположение о жесткости в ORB-SLAM2 и приводит к отказу системы. Аналогично, движущиеся объекты в последовательности 18 (табл. 1) вызывают сбой в отслеживании ORB-SLAM2 в 6 из 10 испытаний (в этих случаях удалось оценить только 56% траектории).

В верхнем ряду показана городская сцена с несколькими автомобилями, припаркованными по обеим сторонам дороги (секв. 01 в табл. 1). В отличие от предыдущего случая, наихудшей конфигурацией является использование всех масок сегментации, поскольку большое количество точек с высоким информационным содержанием удаляется для отслеживания. Результаты ATE в таблице 1 для этой последовательности показывают, что извлечение точек из большей области приводит к более точной оценке траектории.

Подводя итог, заметим, что неиспользование динамических масок объектов увеличивает ошибку траектории из-за совпадения точек на движущихся объектах. Однако применение масок без проверки того, находится ли объект в движении, отбрасывает большое количество информации, особенно когда большая часть сцены является

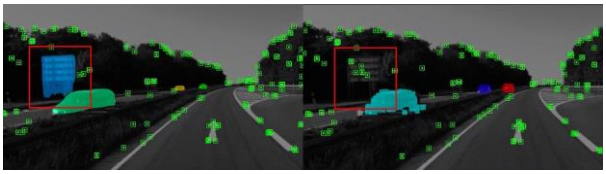

Рисунок 7. **Ошибка сегментации**. Сравнение между *всеми масками*
и *маски DOT*. Обратите внимание, что неправильный сегмент из Detectron2 (знаку в красном квадрате присвоена метка автомобиля) правильно классифицирован DOT как статический.

занятые транспортными средствами. DOT достигает компромисса между этими двумя противоположными сценариями, оценивая фактическое мо- тивационное состояние объектов, чтобы получить более высокую точность и надежность отслеживания.

Замыкание контура. Не все различия в точности траектории обусловлены плохой работой системы слежения. Модуль замыкания контура в ORB-SLAM2 уменьшает дрейф и, следовательно, неточности, вызванные динамическими объектами или удалением припаркованных автомобилей. Мы заметили, что ORB-SLAM2, работающий с *масками DOT,* способен закрыть петлю 6 из 10 запусков в последовательности 9 KITTI *Odometry* (см. таблицу 2), в то время как при использовании *масок All* не было выявлено ни одной. Это приводит к более широкой вариабельности ошибок.

Ошибки сегментации. По сравнению с другими подходами, DOT способен смягчить ошибки сегментации. Нейронные сети иногда неправильно маркируют статические объекты (*например,* дорожные знаки или здания) как динамические, DOT исправляет эту ошибку, повторно маркируя объект как статический (см. рис. 7). Другой пример: когда сеть не срабатывает в одном из кадров последовательности, DOT может заполнить пробел, распространяя маску объекта.

4.2. Распространение маски

Как объясняется в разделе 3.6, наш подход позволяет снизить частоту сегментации сети за счет распространения уже существующих масок в промежуточных кадрах. На рисунке 8 показано количество правильно помеченных пикселей за вычетом ошибочно помеченных (наземная истина черным цветом) на каждом кадре V- KITTI, когда DOT использует 100% сегментаций Detectron2 (красный), 50% (синий), 33% (желтый) и 25% (зеленый). Обратите внимание, что маски остаются точными при распространении, за исключением случаев, когда происходят сбои в отслеживании или движущийся объект входит в сцену между сегментациями (см. также intersection over union для V-KITTI в таблице 4). Мы

Тариф	Seq01	Seq02	Seq06	Seq18	Seq20
1.0	0.88	0.88	0.84	0.90	0.89
0.5	0.74	0.83	0.67	0.85	0.84
0.33	0.72	0.80	0.60	0.85	0.81
0.25	0.69	0.78	0.55	0.84	0.81

считаем, что этот результат может быть полезен для отслеживания объектов в реальном времени, особенно для высокочастотных потоков изображений.

5. Выводы

DOT - это новый алгоритм для систем SLAM, который надежно обнаруживает и отслеживает движущиеся объекты, комбинируя сегментацию объектов и уравнения геометрии в нескольких ракурсах. Наша оценка с ORB-SLAM2 в трех публичных системах

Таблица 4. **Пересечение над объединением** в наборе данных V-KITTI для различных скоростей сегментации.

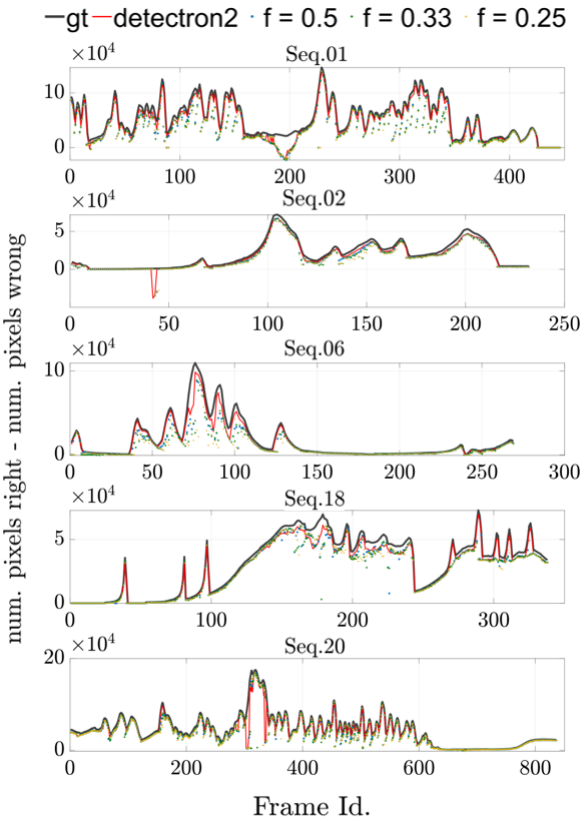

Рисунок 8. **Распространение масок**. Мы показываем для каждого кадра V-KITTI количество правильно помеченных пикселей за вычетом ошибочно помеченных пикселей в соответствии с базовой истиной (черный), когда DOT использует все маски из Detectron2 (красный), 50% (синий), 33% (желтый) и 25% (зеленый).

Наборы данных для исследования автономного вождения [7][6][3] демонстрируют, что информация о движении объектов, генерируемая DOT, позволяет нам сегментировать динамический контент, значительно повышая его надежность и точность.

Независимость DOT от SLAM делает ее универсальным фронт-эндом, который может быть адаптирован с минимальными интеграционными работами к любой современной системе визуальной одометрии или SLAM. В отличие от других систем, отслеживание маски в DOT снижает скорость сегментации (обычно требующей больших ком- пьютерных затрат), уменьшая вычислительные потребности по сравнению с современным уровнем техники.

Ссылки

[1] Йоан Андрей Барсан, Пейдонг Лю, Марк Поллефейс и Ан-дреас Гейгер. Надежное плотное картирование для крупномасштабных динамических сред. *2018 IEEE Международная конференция по робототехнике и автоматизации (ICRA)*, май 2018.

[2] Berta Besco´s, Jose´ M. Fa´cil, Javier Civera, and Jose´ Neira. DynSLAM: отслеживание, отображение и закрашивание в динамических сценах. *CoRR*, abs/1806.05620, 2018.

[3] Йоханн Кабон, Наиля Мюррей и Мартин Хуменбергер. Вир-туал КИТТИ 2, 2020.

[4] Сезар Кадена, Лука Карлоне, Генри Каррильо, Ясир Латиф, Давиде Скарамуцца, Хосе Нейра, Ян Рид и Джон Дж. Леонард. Прошлое, настоящее и будущее одновременной локализации и картирования: Навстречу эпохе надежного восприятия. *IEEE Transactions on robotics*, 32(6):1309-1332, 2016.

[5] Якоб Энгель, Владлен Колтун и Даниэль Кремерс. Прямая разреженная одометрия. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611-625, 2017.

[6] Адриен Гейдон, Цяо Ванг, Йоханн Кабон и Элеонора Виг. Виртуальные миры как прокси для анализа многообъектного слежения. *Труды конференции IEEE по компьютерному зрению и распознаванию образов*, страницы 4340-4349, 2016.

[7] Андреас Гайгер, Филипп Ленц и Ракель Уртасун. Готовы ли мы к автономному вождению? Комплекс для проверки зрения KITTI. *Конференция по компьютерному зрению и распознаванию образов (CVPR)*, 2012.

[8] Ричард Хартли и Эндрю Зиссерман. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[9] K. He, G. Gkioxari, P. Dolla´r, and R. Girshick. Mask R-CNN. In *2017 IEEE International Conference on ComputerVision (ICCV)*, pages 2980-2988, 2017.

[10] Jiahui Huang, Sheng Yang, Tai-Jiang Mu, and Shi-Min Hu. ClusterVO: кластеризация движущихся объектов и оценка ви- суальной одометрии для себя и окружения, 2020.

[11] Jiahui Huang, Sheng Yang, Zishuo Zhao, Yu-Kun Lai, and Shi-Min Hu. ClusterSLAM: бэкэнд SLAM для симультанной кластеризации жестких тел и оценки движения. 2019.

[12] Маттиас Иннманн, Михаэль Цолльхофер, Маттиас Нисснер, Кристиан Теобальт и Марк Штаммингер. VolumeDeform: Объемная нежесткая реконструкция в реальном времени. Октябрь 2016 года.

[13] Хосе Ламарка, Шайфали Парашар, Адриен Бартоли и Джей Эм Монтьель. Defslam: Отслеживание и отображение деформа- ингибирование сцен из монокулярных последовательностей. *препринт arXiv:1908.08918*, 2019.

[14] Цун-И Лин, Майкл Мейр, Серж Белонги, Любомир Бурдев, Росс Гиршик, Джеймс Хейс, Пьетро Перона, Дэва Раманан, К. Лоуренс Зитник и Петр Долльр. Microsoft COCO: Common Objects in Context, 2014.

[15] Инк. Matterport. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow, 2019. URL: https://github.com/matterport/Mask_RCNN [Online. Accedido el 03/12/2019].

[16] Рауль Мур-Арталь, Хосе Мария Мартинес Монтьель и Хуан Д Тардос. ORB-SLAM: универсальная и точная монокулярная система. Система SLAM. *IEEE transactions on robotics*, 31(5):1147-1163, 2015.

[17] Рауль Мур-Арталь и Хуан Д. Тардос. ORB-SLAM2: Система SLAM с открытым исходным кодом для монокулярных, стерео и RGB-D камер. *IEEE Transactions on Robotics*, 33(5):12551262, 2017.

[18] Ричард А. Ньюкомб, Дитер Фокс и Стивен М. Сейтц. Dynamicfusion: Реконструкция и отслеживание нежестких сцен в реальном времени. *Конференция IEEE по компьютерному зрению и распознаванию образов (CVPR)*, июнь 2015 года.

[19] Мартин Рнз, Мод Буффье и Лурдес Агапито. MaskFusion: Распознавание, отслеживание и реконструкциянескольких движущихся объектов в реальном времени, 2018.

[20] Raluca Scona, Mariano Jaimez, Yvan R. Petillot, Maurice Fallon, and Daniel Cremers. StaticFusion: Background Reconstruction for Dense RGB-D SLAM in Dynamic Environments. In *2018 ICRA*. IEEE.

[21] Хауке Страсдат. *Локальная точность и глобальная согласованность для эффективного визуального SLAM*. Докторская диссертация, факультет вычислительной техники, Имперский колледж Лондона, 2012.

[22] Юрген Штурм, Николас Энгельхард, Феликс Эндрес, Вольфрам Бургард и Даниэль Кремерс. Эталон для оценки систем rgb-d slam. В *2012 IEEE/RSJ Interna- tional Conference on Intelligent Robots and Systems*, pages 573-580. IEEE, 2012.

[23] Юксин Ву, Александр Кириллов, Франциско Масса, Ван-Йен Ло, и Росс Гиршик. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[24] Бинбин Сюй, Вэньбин Ли, Димос Тзуманикас, Майкл Блош, Эндрю Дэвисон и Стефан Лейтенеггер. MID-Fusion: основанная на восьмерке объектно-уровневая многоинстанционная динамика. SLAM, 2018.