

# Глава 1

## Введение

Введение.

# Глава 2

## Постановка задачи одновременного картирования и локализации

### 2.1 Одновременное картирование и локализация по видеопотоку (vSLAM)

Задача одновременного картирования и локализации по визуальным данным (visual-based simultaneous localization and mapping, vSLAM) возникает при навигации в неизвестной среде робота, не имеющего на борту никаких сенсоров, кроме единственной видеокамеры. Задача формулируется следующим образом: по изображениям с видеокамеры необходимо построить трёхмерную модель окружающего пространства, и определить траекторию перемещения камеры в этом пространстве.

Математически задачу можно сформулировать таким образом: существует набор точек в трёхмерном пространстве  $\{M_i\} = \mathbf{W}, M_i \in \mathbb{R}^3$ , называемый сценой. Данна последовательность кадров  $\{\mathcal{P}^t\}$ . Каждый кадр является проекцией точек сцены с ракурса  $\mathcal{R}_t = (x_t, y_t, z_t, p_t, r_t, w_t) \in \mathbb{R}^6$ . Числа  $x_t, y_t, z_t$  задают пространственное положение камеры в момент времени  $t$ , а  $p_t, r_t, w_t$  - углы направления главной оптической оси камеры в момент времени  $t$ .

Кадр представляется в виде трех матриц размер  $H \times W$ , содержащих числа от 0 до 1 - яркости соответствующих пикселей красной, синей и зеленой цветовых компонент:

$$\mathcal{P}^t = \{\mathcal{P}_{c,h,w}^t\}_{c \in [1 \dots 3], h \in [1 \dots H], w \in [1 \dots W]} \in [0, 1]^{3 \times H \times W}$$

Каждый элемент  $c$ -й матрицы кадра  $\mathcal{P}_{c,h,w}^t$  представляет собой яркость  $c$ -й цветовой компоненты точки, которая в момент времени  $t$  спроектировалась на позицию  $(h, w)$  в матрице камеры:

$$P(\mathcal{R}_t, M_i) = (h, w) \Rightarrow \mathcal{P}_{c,h,w}^t = I_c(\mathcal{R}_t, M_i),$$

где  $P(\mathcal{R}_t, M_i) : \mathbb{R}^6 \times \mathbb{R}^3 \rightarrow \mathbb{R}^2$  - функция проекции точки пространства на матрицу

камеры, принимающая на вход ракурс и положение точки в пространстве и возвращающая координаты проекции, а  $I_c(\mathcal{R}_t, M_i) : \mathbb{R}^6 \times \mathbb{R}^3 \rightarrow [0, 1]$  - функция яркости  $c$ -й цветовой компоненты точки  $M_i$ , рассматриваемой с ракурсом  $\mathcal{R}_t$ .

По имеющейся последовательности кадров  $\{\mathcal{P}_t\}$  необходимо найти  $(x_t, y_t, z_t)$  для всех моментов времени  $t$  - координаты ракурсов  $\mathcal{R}_t$ , а также как можно большее количество координат точек  $M_i$ .

## 2.2 Одновременное картирование и локализация по видеоданным с картами глубины (RGBD-SLAM)

Задача одновременного картирования и локализации по видеоданным и данным глубины (RGBD-SLAM) возникает при навигации в неизвестной среде робота, имеющего на борту видеокамеру и сенсор глубины. Задача формулируется следующим образом: по изображениям с видеокамеры и картам глубины этих изображений необходимо построить трёхмерную модель окружающего пространства, и определить траекторию перемещения камеры в этом пространстве.

Математически задачу можно сформулировать таким образом: существует набор точек в трёхмерном пространстве  $\{M_i\} = \mathbf{W}, M_i \in \mathbb{R}^3$ , называемый сценой. Данна последовательность кадров  $\{\mathcal{P}^t\}$  и карт глубины  $\{\mathcal{D}^t\}$ .

Кадр представляется в виде трех матриц размер  $H \times W$ , содержащих числа от 0 до 1 - яркости соответствующих пикселей красной, синей и зеленой цветовых компонент:

$$\mathcal{P}^t = \{\mathcal{P}_{c,h,w}^t\}_{c \in [1 \dots 3], h \in [1 \dots H], w \in [1 \dots W]} \in [0, 1]^{3 \times H \times W}$$

Каждый элемент  $c$ -й матрицы кадра  $\mathcal{P}_{c,h,w}^t$  представляет собой яркость  $c$ -й цветовой компоненты точки, которая в момент времени  $t$  спроектировалась на позицию  $(h, w)$  в матрице камеры:

$$P(\mathcal{R}_t, M_i) = (h, w) \Rightarrow \mathcal{P}_{c,h,w}^t = I_c(\mathcal{R}_t, M_i)$$

Карта глубины представляется в виде матрицы размера  $H \times W$ , содержащей положительные действительные числа - глубины соответствующих пикселей:

$$\mathcal{D}^t = \{\mathcal{D}_{h,w}^t\}_{h \in [0 \dots H], w \in [0 \dots W]}$$

Элемент матрицы карты глубины  $D_{h,w}^t$  представляет собой расстояние от положения камеры в момент времени  $t$  до точки, которая в момент времени  $t$  спроектировалась на позицию  $(h, w)$  в матрице камеры:

$$P(\mathcal{R}_t, M_i) = (h, w) \Rightarrow \mathcal{D}_{h,w}^t = \rho(M_i, (x_t, y_t, z_t))$$

Здесь  $P(\mathcal{R}_t, M_i) : \mathbb{R}^6 \times \mathbb{R}^3 \rightarrow \mathbb{R}^2$  - функция проекции точки пространства на матрицу камеры, принимающая на вход ракурс и положение точки в пространстве и возвращающая координаты проекции, а  $I_c(\mathcal{R}_t, M_i) : \mathbb{R}^6 \times \mathbb{R}^3 \rightarrow [0, 1]$  - функция яркости  $c$ -й цветовой компоненты точки  $M_i$ , рассматриваемой с ракурсом  $\mathcal{R}_t$ . Функция  $\rho : \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}_+$  задает евклидово расстояние между двумя точками в пространстве.

По имеющейся последовательности кадров  $\{\mathcal{P}_t\}$  и карт глубин  $\{\mathcal{D}_t\}$  необходимо найти  $(x_t, y_t, z_t)$  для всех моментов времени  $t$  - координаты ракурсов  $\mathcal{R}_t$ , а также как можно большее количество координат точек  $M_i$ .

## 2.3 Сведение RGBD-SLAM к vSLAM: восстановление карт глубин по видеопотоку

Задача восстановления карт глубин по видеопотоку возникает при навигации в неизвестной среде робота, не имеющего на борту никаких сенсоров, кроме видеокамеры. С помощью восстановления глубины по видеопотоку можно свести задачу vSLAM к задаче RGBD-SLAM, для которой разработаны более эффективные методы решения.

Задача восстановления карт глубин по видеопотоку формулируется следующим образом: по изображениям, поступающим с единственной видеокамеры, необходимо определить расстояния до всех объектов, изображенных на этих изображениях. Математически задачу можно сформулировать таким образом: существует набор точек в трёхмерном пространстве  $\{M_i\} = \mathbf{W}, M_i \in \mathbb{R}^3$ , называемый сценой. Данна последовательность кадров  $\{\mathcal{P}^t\}$ . Каждый кадр является проекцией точек сцены с ракурса  $\mathcal{R}_t = (x_t, y_t, z_t, p_t, r_t, w_t) \in \mathbb{R}^6$ . Числа  $x_t, y_t, z_t$  задают пространственное положение камеры в момент времени  $t$ , а  $p_t, r_t, w_t$  - углы направления главной оптической оси камеры в момент времени  $t$ .

Кадр представляется в виде трех матриц размер  $H \times W$ , содержащих числа от 0 до 1 - яркости соответствующих пикселей красной, синей и зеленой цветовых компонент:

$$\mathcal{P}^t = \{\mathcal{P}_{c,h,w}^t\}_{c \in [1 \dots 3], h \in [1 \dots H], w \in [1 \dots W]} \in [0, 1]^{3 \times H \times W}$$

Каждый элемент  $c$ -й матрицы кадра  $\mathcal{P}_{c,h,w}^t$  представляет собой яркость  $c$ -й цветовой компоненты точки, которая в момент времени  $t$  спроектировалась на позицию  $(h, w)$  в матрице камеры:

$$P(\mathcal{R}_t, M_i) = (h, w) \Rightarrow \mathcal{P}_{c,h,w}^t = I_c(\mathcal{R}_t, M_i),$$

где  $P(\mathcal{R}_t, M_i) : \mathbb{R}^6 \times \mathbb{R}^3 \rightarrow \mathbb{R}^2$  - функция проекции точки пространства на матрицу камеры, принимающая на вход ракурс и положение точки в пространстве и возвращающая координаты проекции, а  $I_c(\mathcal{R}_t, M_i) : \mathbb{R}^6 \times \mathbb{R}^3 \rightarrow [0, 1]$  - функция яркости  $c$ -й цветовой компоненты точки  $M_i$ , рассматриваемой с ракурсом  $\mathcal{R}_t$ .

По имеющейся последовательности кадров  $\{\mathcal{P}_t\}$  необходимо для всех  $h, w, t$  найти  $\mathcal{D}_{h,w}^t$  - расстояния от положения камеры в момент  $t$  до точек сцены, изображенных на кадре:

$$P(\mathcal{R}_t, M_i) = (h, w) \Rightarrow \mathcal{D}_{h,w}^t = \rho((x_t, y_t, z_t), M_i)$$

# Глава 3

## Оценка качества одновременного картирования и локализации

### 3.1 Метрики качества

Для оценки качества алгоритмов одновременного картирования и локализации (vSLAM) и их сравнения между собой необходимо выбрать метрику оценки качества. Алгоритмы vSLAM, как правило, дают в качестве выходных данных карту окружающей местности и траекторию перемещения камеры. Поэтому для оценки качества vSLAM существуют две группы метрик - метрики качества локализации и метрики качества картирования.

#### 3.1.1 Метрики качества локализации

Метрики качества локализации, как правило, сравнивают траекторию, вычисленную алгоритмом SLAM, с истинной траекторией. Траектория представляет собой набор поз, каждая из которых включает трехмерную позицию (положение камеры в пространстве) и трехмерную ориентацию (направление главной оптической оси камеры). Оценка качества локализации сводится к вычислению ошибки между набором истинных и предсказанных поз. В научной литературе используются абсолютные и относительные ошибки, а также ошибки смещения и поворота.

Обозначим истинную траекторию как  $\{(p_t, q_t)\}, t \in \{1, \dots, T\}$ , где  $p_t \in \mathbb{R}^3$  - трехмерная позиция камеры в момент времени  $t$ ,  $q_t \in \mathbb{R}^3$  - вектор направления главной оптической оси камеры в момент времени  $t$ . Предсказанную траекторию обозначим как  $\{(\hat{p}_t, \hat{q}_t)\}, t \in \{1, \dots, T\}$ .

Одной из наиболее распространенных метрик является абсолютная ошибка траектории (Absolute Trajectory Error, ATE). Она формулируется как среднеквадратичное отклонение точек предсказанной траектории от истинной:

$$ATE = \sqrt{\frac{1}{T} \sum_{t=1}^T \|p_t - \hat{p}_t\|_2^2} \quad (1)$$

Помимо абсолютной ошибки траектории, также широко применяется относительная ошибка позы (Relative Pose Error, RPE). Она формулируется как среднеквадратичное отклонение предсказанного смещения на каждом шаге от истинного:

$$\Delta p_t = M_{q_{t-1}}^{-1}(p_t - p_{t-1}); \Delta \hat{p}_t = M_{\widehat{q_{t-1}}}^{-1}(\hat{p}_t - \widehat{p_{t-1}})$$

$$RPE = \sqrt{\frac{1}{T} \sum_{t=1}^T \|\Delta p_t - \Delta \hat{p}_t\|_2^2}, \quad (2)$$

где  $M_{q_{t-1}}$ ,  $M_{\widehat{q_{t-1}}}$  - матрицы вращения, переводящие вектор  $(1, 0, 0)$  в векторы  $q_{t-1}$  и  $\widehat{q_{t-1}}$  соответственно.

**TODO:** Поясняющая картинка!

В работе [14] приводятся следующие метрики качества локализации: относительное смещение и относительная ошибка поворота. Данные метрики учитывают не только расхождение между траекториями, но и их длину:

$$E_{trans} = \frac{1}{T} \sum_{t=1}^T \frac{\|M_{q_{t-1}}(p_t - p_{t-1}) - M_{\widehat{q_{t-1}}}(\hat{p}_t - \widehat{p_{t-1}})\|_2}{\|p_t - p_{t-1}\|_2} \quad (3)$$

$$E_{rot} = \frac{1}{T} \frac{\angle(M_{q_{t-1}}(p_t - p_{t-1}), M_{\widehat{q_{t-1}}}(\hat{p}_t - \widehat{p_{t-1}}))}{\|p_t - p_{t-1}\|_2} \quad (4)$$

### 3.1.2 Метрики качества картирования

Оценка качества картирования является более сложной задачей, чем оценка качества локализации. Если при оценке качества локализации легко установить соответствие между точками истинной и предсказанной траекторий (по времени прохождения данных точек), то однозначного соответствия между точками истинной и предсказанной карты не существует. Как правило, соответствия устанавливаются методом ближайшего соседа - каждая точка предсказанной карты сопоставляется с ближайшей к ней точкой истинной карты. Такой подход применяется в программном пакете CloudCompare<sup>1</sup> и в работах [18][44]. Применяя среднеквадратичную ошибку (RMSE) в этом подходе, получаем абсолютную ошибку картирования (Absolute Mapping Error, AME).

По заданной истинной карте, представленной в виде трехмерного облака точек:

$$M = \{m_i \in \mathbb{R}^3; i \in [1; n]; n \in \mathbb{N}\} \quad (5)$$

---

<sup>1</sup><http://cloudcompare.org/>

и карте, построенной алгоритмом vSLAM:

$$M^* = \{m_j^* \in \mathbb{R}^3; j \in [1; N]; n \in \mathbb{N}\} \quad (6)$$

абсолютная ошибка картирования вычисляется следующим образом:

$$j' = \arg \min_i \|m_j^* - m_i\|_2$$

$$AME(M, M^*) = \sqrt{\frac{1}{N} \sum_{j=1}^N \|m_j^* - m_{j'}\|_2^2} \quad (7)$$

Однако существуют ситуации, когда абсолютная ошибка картирования, вычисленная по методу ближайшего соседа, не является репрезентативной. Например, при картировании помещений большой площади точки стен предсказанной карты могут быть сопоставлены с точками пола истинной карты. Таким образом абсолютная ошибка картирования может быть небольшой даже при неточном картировании. Так получилось в одном из экспериментов, описанных в разделе 6.3.1. При картировании помещений алгоритмом SLAM с глубинами, предсказанными нейросетью, обнаружилось, что карты визуально получались довольно неточные, при этом значение метрики AME на этих картах оказалось низким. При удалении точек пола из истинных и предсказанных карт значение метрики AME существенно выросло (см. рис. 2). Таким образом, в контексте оценки качества алгоритмов vSLAM метрика AME с сопоставлением по методу ближайшего соседа малоприменима.

В рамках данной работы был разработан новый способ сопоставления точек истинной и предсказанной карты, учитывающий контекст задачи визуального картирования и локализации и основанный на сопоставлении ракурсов, с которых видны точки истинной и предсказанной карт. Ниже приведено подробное описание разработанного метода.

Пусть  $M, M^*$  - истинная и построенная алгоритмом vSLAM карты (6, 5);  $m_i^*$  - точка карты  $M^*$ , попадающая в поле зрения камеры в момент времени  $t$ ;  $p_t^*, q_t^*$  - предсказанные методом vSLAM положение и ориентация камеры в момент времени  $t$ ;  $p_t, q_t$  - истинные положение и ориентация камеры в момент  $t$ . Нужно построить функцию  $f : M^* \rightarrow M$ , устанавливающую соответствие между точками построенной и истинной карты.

Обозначим матрицы вращения, заданные ориентациями  $q_t$  и  $q_t^*$ , как  $M_{q_t}$  и  $M_{q_t^*}$  соответственно. Вектор  $r_t = (M_{q_t^*})^{-1} M_{q_t} (m_i^* - p_t)$  соответствует направлению с позиции камеры на точку  $m_i^*$  в истинной карте (точка  $p_t + \alpha r_t$  в истинной карте будет видна в момент  $t$  под тем же ракурсом, что точка  $m_i^*$  в предсказанной карте, см. рис. 1). Точке  $m_i^*$  будет сопоставлена ближайшая точка истинной карты, которая видна под таким ракурсом:

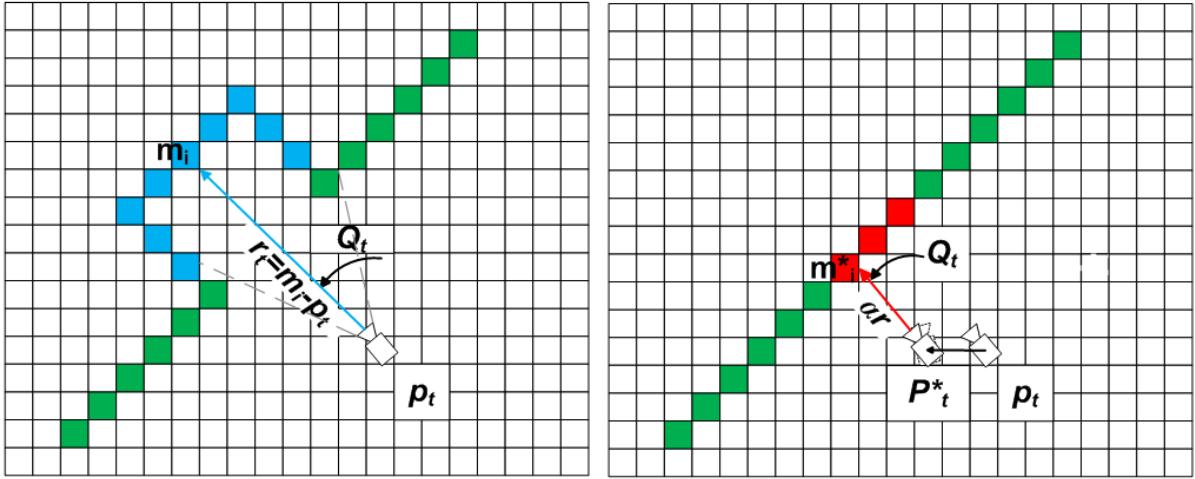


Рис. 1. Пример сопоставления точек истинной и предсказанной карты с использованием ракурса

$$f(m_i^*) = p_t + \alpha r_t; \quad \alpha = \arg \min_{\alpha'} : p_t + \alpha' r_t \in M \quad (8)$$

Абсолютная и относительная ошибки картирования с таким методом сопоставления будет выглядеть следующим образом:

$$AME(M, M^*) = \sqrt{\frac{1}{N} \sum_{i=1}^N \|m_i^* - f(m_i^*)\|_2^2} \quad (9)$$

$$RME(M, M^*) = \sqrt{\frac{1}{N} \sum_{i=1}^N \|M_{q_t}^{-1}(m_i^* - p_t^*) - M_{q_t}^{-1}(f(m_i^*) - p_t)\|_2^2} \quad (10)$$

При использовании данных метрик возникает следующая проблема. Одна точка карты может быть видна с разных позиций (т.е. для одной точки  $m_i^*$  есть несколько  $t$ , по которым можно построить разные соответствия). Поэтому нужно определиться со способом выбора  $t$ . В данной работе были рассмотрены следующие варианты:

- $t$  выбирается как момент, в который точка  $m_i^*$  попала в поле зрения камеры в первый раз;
- $t$  выбирается как момент, в который точка  $m_i^*$  попала в поле зрения камеры в последний раз;
- $t$  выбирается как момент видимости точки  $m_i^*$ , в который точка была наиболее близка к позиции камеры;
- $t$  рассматриваются все моменты  $t$ , в который точка попадала в поле зрения камеры; при вычислении метрики берется усредненное расстояние между точкой  $m_i^*$  и точками  $f(m_i^*, t)$  для всех  $t$ .

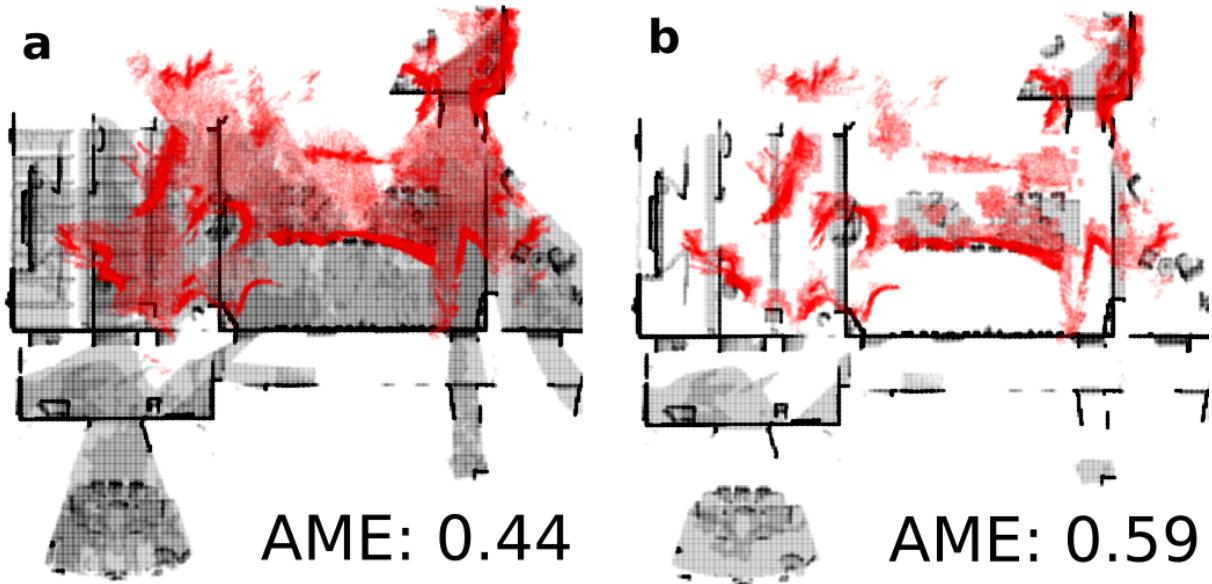


Рис. 2. Исходные карты (а) и карты с вырезанным полом (б). Чёрным отмечены точки истинной карты, красным - предсказанный методом vSLAM карты. Метрика AME на картах с полом значительно ниже, чем на тех же картах без пола.

Последний вариант выбора  $t$  приводит к большим вычислительным затратам, однако метрика с ним получается наиболее стабильной. Для оценки качества алгоритмов vSLAM в данной работе был выбран именно этот вариант.

Метрики AME и RME, вычисленные на основе функции сопоставления  $f$ , являются более подходящими для оценки качества алгоритмов vSLAM, поскольку они учитывают не только расстояния между точками, но и процесс построения карты алгоритмом vSLAM. Также данные метрики являются более устойчивыми к различным изменениям структуры карты (например, удаление точек пола), что подтверждено экспериментами, описанными в разделе 6.3.1.

### 3.2 Коллекции данных

Проведение натурных экспериментов на реальной робототехнической системе затратно, а повторяемость таких экспериментов затруднена, поэтому для тестирования алгоритмов vSLAM и их сравнения между собой обычно применяются предварительно собранные коллекции данных. Такие коллекции должны включать в себя запись видеопотока с камеры робота (в случае RGB-D камеры - запись изображений и данных о глубине), а также данные об истинных позициях робота и истинной карте окружающей местности для вычисления метрик качества. Подобные коллекции бывают двух видов: собранные в реальном мире и синтетические.

При сборке коллекций данных в реальном мире возникают сложности с определением истинной траектории робота и точной модели окружающей местности. Для вычисления точных данных о движении робота и его окружающей среде необходимо редкое и дорогостоящее оборудование, а также значительные временные затраты.

Как правило, в коллекциях данных из реального мира имеются данные лишь о траектории робота, собранные с помощью систем отслеживания движения (Motion-capture System [24]), а истинные карты окружающей среды отсутствуют. При тестировании методов vSLAM на таких коллекциях можно вычислить метрики ATE 1, RPE 2,  $E_{trans}$  3,  $E_{rot}$  4, но невозможно вычислить метрики AME и RME (7, 9, 10).

Синтетические коллекции данных, а также различные робототехнические симуляторы (например, [23] [35]) позволяют легко получить истинные карты и траектории, однако в большинстве обладают низкой фотореалистичностью, что критически важно для алгоритмов визуального картирования и локализации.

### 3.2.1 Коллекции данных из реального мира

Коллекции данных, собранные с помощью прогонов системы, оснащенной камерой, по некоторой траектории в реальном мире, на данный момент очень широко распространены. Имеется множество коллекций, в которых представлены прогоны робота в помещениях, а также несколько крупных коллекций, собранных с автомобиля, оснащенного большим количеством датчиков. В большинстве таких коллекций представлены истинные траектории движения робота, но отсутствуют истинные карты окружающей местности. Подобные датасеты хорошо подходят для оценки качества локализации методов vSLAM, однако оценка качества картирования на них затруднена или вовсе невозможна. Ниже представлен обзор наиболее известных наборов данных из реального мира для тестирования vSLAM.

Одним из самых известных датасетов для тестирования алгоритмов SLAM является KITTI [14]. Он содержит 22 сцены, записанные с автомобиля в городской среде. Общая длина проезда составляет 39 км. Датасет содержит порядка 41 тысячи кадров видеопотока, записанного со стереокамеры, а также данные трехмерного лазерного сканера (лидара) и инерциальной навигационной системы (ИНС) совместно с показаниями GPS. В данной коллекции также представлены истинные позиции, собранные с помощью агрегации данных GPS и ИНС, а также качественной пост-обработки. Истинная карта окружающих объектов не представлена, однако она может быть приблизительно воссоздана по истинным позициям автомобиля и облакам точек с лидара.

Помимо исходных данных с сенсоров, в KITTI также представлено программное обеспечение, необходимое для вычисления метрик качества локализации. Также в датасете предоставлены данные для тестирования некоторых других методов компьютерного зрения: семантической сегментации, детекции объектов, вычисления оптического потока. Данный датасет широко используется для оценки качества методов визуальной одометрии [15], а также для обучения различных нейросетевых моделей [27] [51] [34]. Однако оценка качества картирования на данном датасете затруднительна, поскольку истинных карт окружающей среды в датасете не предоставлено.

Одним из наиболее широко используемых датасетов для тестирования vSLAM в

помещениях является RGB-D SLAM dataset and benchmark [40], созданный учеными из Технического университета Мюнхена (TUM). Датасет содержит 39 последовательностей, записанных в различных помещениях с камеры Microsoft Kinect [49], установленной на малом колесном роботе. Последовательности были записаны в различных помещениях, длины траекторий не превосходят 40 метров. Помимо изображений и карт глубины с камеры Kinect, в датасете также имеются истинные траектории проезда робота, вычисленные с помощью высокоточной системы отслеживания движения (Motion Capture System).

Датасет от TUM также содержит набор программного обеспечения для тестирования алгоритмов SLAM и вычисления метрик ATE (1) и RPE (2). Благодаря набору ПО, а также наличию истинных траекторий, данный датасет хорошо подходит для оценки качества локализации методов vSLAM и RGB-D SLAM в помещениях. Однако оценка качества картирования на нем невозможна, так как не представлены истинные координаты точек окружающих объектов. Обучение нейронных сетей на данной коллекции затруднительно из-за ее небольшого объема, а также однообразности изображений.

Еще одним широко используемым для тестирования vSLAM датасетом является EuRoC [6]. В этом датасете представлены данные с видеокамеры, установленной на квадрокоптере, собранные с 11 прогонов по двум комнатам. Помимо видеоданных, также имеются данные ИНС и истинные траектории квадрокоптера, измеренные с помощью системы отслеживания движения (Motion Capture System). Карт глубины датасет не содержит, поэтому он не подходит для оценки качества алгоритмов RGB-D SLAM. Программного обеспечения для запуска алгоритмов vSLAM и вычисления метрик разработчики также не предоставили. Ввиду всего вышеперечисленного применимость данного датасета для оценки качества методов vSLAM весьма ограничена.

В настоящее время существуют также коллекции данных, в которых помимо траекторий предоставлены трехмерные модели окружающих объектов, что дает возможность оценить не только качество локализации, но и качество картирования. Одной из самых крупных таких коллекций является ScanNet [11]. В данной коллекции представлены 2.5 млн пар изображение-глубина, снятые RGB-D камерами в 1513 помещениях. Траектории камеры и трехмерные модели помещений были получены с помощью тщательной алгоритмической обработки данных RGB-D камеры и ИНС. Сенсоры, использованные при создании датасета, имеют достаточно большую погрешность, поэтому полученные таким образом траектории и карты сложно считать "истинными". При проведении экспериментов с методами vSLAM возможны искажения оценок их качества.

Датасет ScanNet изначально не предназначался для тестирования алгоритмов vSLAM, и в нем не предоставлено нужного для этого программного обеспечения и удобного формата данных. Поэтому тестирование методов vSLAM на данном датасете



Рис. 3. Пример симуляционной среды Gazebo (слева) и CoppeliaSim (справа). Однообразность текстур делает работу методов Visual SLAM затруднительной

те весьма затруднительно. Однако датасет может быть полезен для решения вспомогательных задач для visual SLAM - например, обучения нейросетей восстановления глубины и визуальной одометрии.

### 3.2.2 Симуляторы и коллекции синтетических данных

Благодаря бурному развитию вычислительной техники и компьютерных технологий, в последние годы были разработаны различные робототехнические симуляторы. Появление симуляторов дало широкие возможности для тестирования робототехнических алгоритмов, в том числе методов картирования и локализации. В симуляционной среде робот может перемещаться по произвольной траектории, для которой всегда известны истинные позиции, что дает возможность проведения неограниченного числа экспериментов по оценке качества методов SLAM.

Наиболее применяемыми на данный момент симуляторами являются Gazebo [23] и CoppeliaSim (V-REP) [35]. Эти симуляторы отличаются подробным моделированием различных физических процессов, благодаря чему в них можно имитировать работу различных сенсоров, таких, как ИНС и лазерные сканеры, а также видеокамер и RGB-D камер. Однако фотoreалистичность интерьеров в подобных симуляторах довольно низкая (см. рис. 3) - из-за повторяемости текстур и неточного моделирования распространения света могут возникнуть проблемы с тестированием алгоритмов Visual SLAM.

Проблемы фотoreалистичности частично решились с появлением симулятора Habitat [37]. В данном симуляторе не моделируются такие физические процессы, как инерция движения и распространение лазерных лучей, однако изображения сцен, используемые в нем, отличаются высокой фотoreалистичностью (см. рис. ??). Как правило, в симуляторе Habitat используются сцены из коллекции Gibson [46] или Matterport3D [8]. Их фотoreалистичность обусловлена тем, что сцены построены по реальным помещениям с помощью специальной камеры<sup>2</sup> и тщательной алгоритмической пост-обработки. Коллекция Matterport3D содержит около 60 сцен площадью в

<sup>2</sup><https://matterport.com/>



Рис. 4. Пример изображений с камеры робота в симуляционной среде Habitat на сцене из датасета Matterport3D

несколько сотен квадратных метров каждая. Коллекция Gibson содержит около 500 сцен площадью несколько десятков квадратных метров каждая. В обоих коллекциях представлены здания и помещения различного типа. Симуляционная среда дает возможность перемещаться в этих помещениях по произвольной траектории.

**TODO:** Ссылка на датасет MAOMaps!

В рамках данной работы на основе симулятора Habitat и коллекции сцен Matterport3D был собран новый датасет для оценки качества алгоритмов vSLAM. Датасет содержит 40 траекторий, которые разделены на 20 перекрывающихся пар, что дает возможность также тестировать алгоритмы объединения карт (Map Merging). Длины траекторий вариируются от 4 до 33 метров. В датасете представлены RGB-D данные (суммарно около 30000 пар картинка-глубина), а также истинные траектории перемещения камеры и истинные карты помещений, построенные с помощью метода обратной проекции по истинным позициям и точным картам глубин. Подробное описание собранного датасета доступно в работе [1]. Эксперименты по оценке качества методов vSLAM на нем описаны в разделе 6.3.1.

# Глава 4

## Методы решения задачи vSLAM

### 4.1 Классические методы

Как правило, для решения задачи vSLAM применяются методы, основанные на вычислении геометрических преобразований между позициями камеры по изображениям. Преобразования могут вычисляться как прямыми методами (например, минимизацией фотометрической ошибки), так и путем сопоставления извлеченных из изображений особых точек. По вычисленным преобразованиям восстанавливается траектория перемещения камеры. Также с использованием этих преобразований обычно строятся локальные карты - карты участка местности, попадающего в поле зрения камеры, которые затем соединяются в глобальную карту с помощью различных вероятностных методов. Ниже представлен обзор основных алгоритмов vSLAM, имеющих открытую программную реализацию.

#### 4.1.1 *ORB-SLAM* и *ORB-SLAM2*

Алгоритм ORB-SLAM [31] и его модификация ORB-SLAM2 [32] являются одними из наиболее популярных методов одновременного картирования и локализации. ORB-SLAM использует в качестве входных данных видеопоток с единственной камерой, а ORB-SLAM2 является расширением алгоритма для работы по данным стереокамеры или RGB-D камеры. В основе алгоритма лежит извлечение особых точек из изображений с помощью детектора ORB [36]. Высокая скорость детектора ORB позволяет методу работать в реальном времени в условиях ограниченных вычислительных ресурсов. Однако на карту наносятся лишь координаты особых точек, поэтому построенная алгоритмом карта получается разреженной (см. рис. 5) и является непригодной для планирования траектории (планировщик может проложить маршрут между двумя особыми точками, где может быть стена или другое препятствие).

Работа алгоритма ORB-SLAM разделена на три основных потока, выполняющихся параллельно:

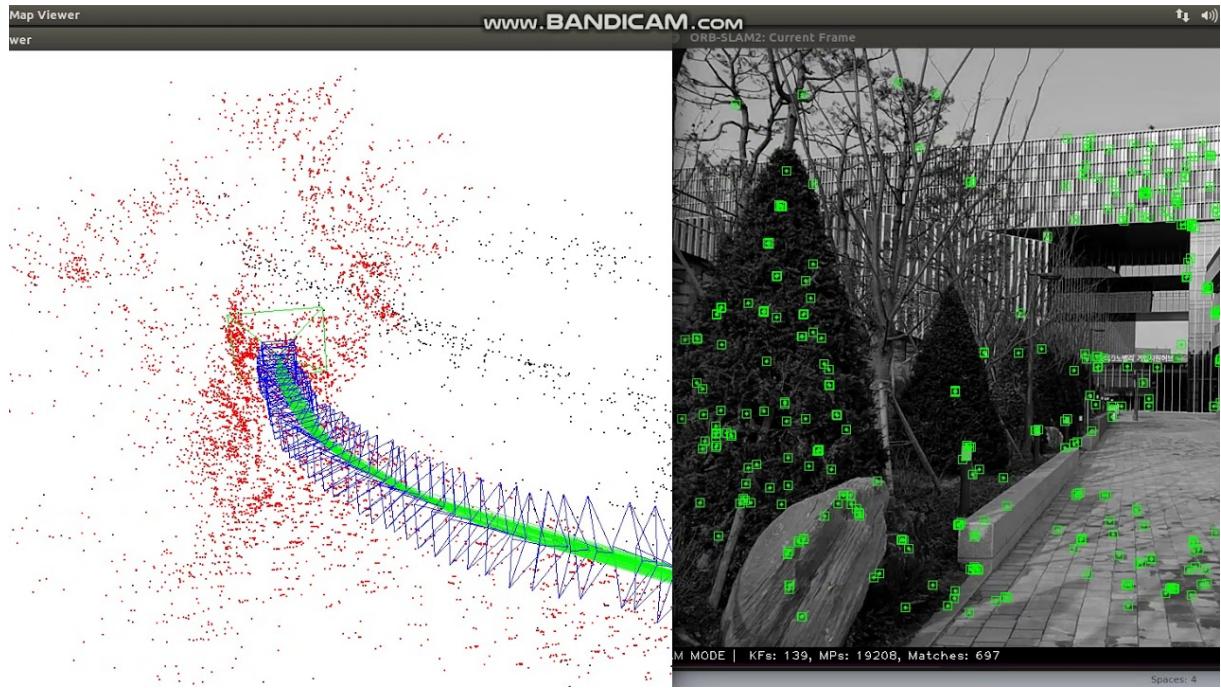


Рис. 5. Работа алгоритма ORB-SLAM: построенная карта (слева) и детекция особых точек на изображении (справа). Карта представляет собой сильно разреженный набор точек, что затрудняет навигацию в ней

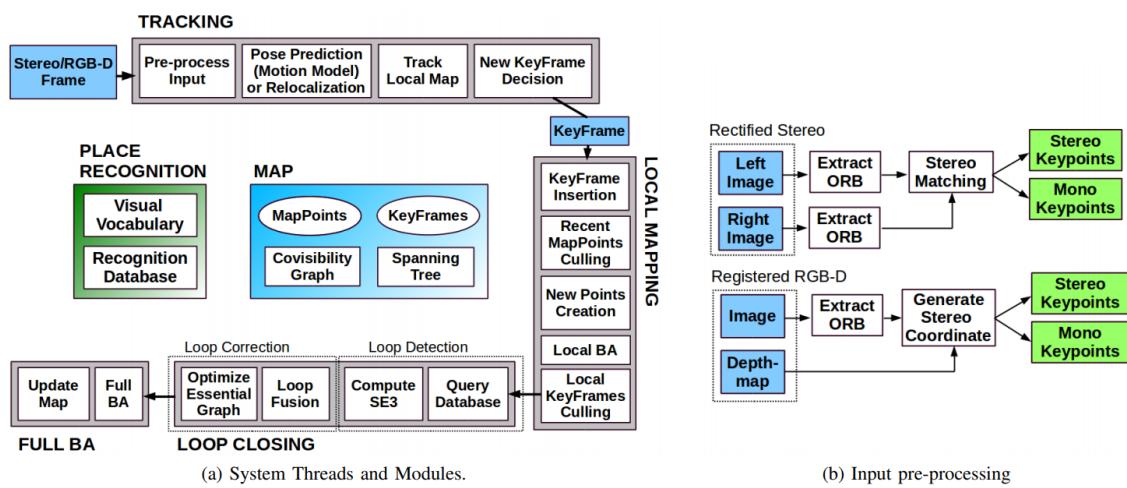


Рис. 6. Схема алгоритма ORB-SLAM

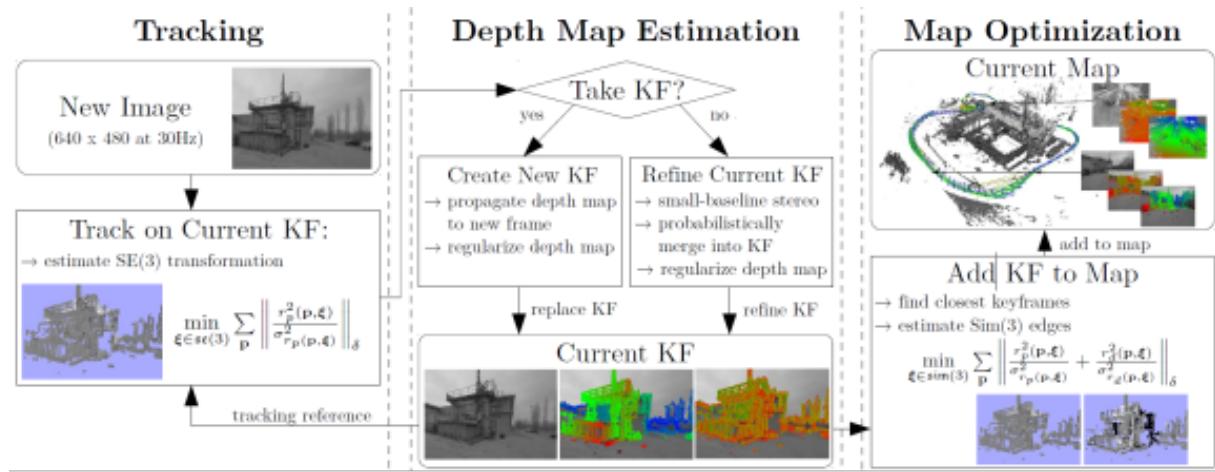


Рис. 7. Схема алгоритма LSD-SLAM

- **Tracking** - отслеживание кадров. Данный поток приблизительно определяет текущее положение камеры с помощью поиска похожего кадра и сопоставления особых точек на нем и текущем кадре.
- **Local Mapping** - построение и оптимизация локальной карты (вблизи текущего положения камеры).
- **Loop Closing** - поиск и замыкание циклов с помощью объединения похожих кадров.

Схема основных компонентов алгоритма изображена на рисунке 6.

#### 4.1.2 LSD-SLAM

Алгоритм LSD-SLAM [13] так же, как и ORB-SLAM, выполняет картирование и локализацию по данным с единственной камеры, однако он имеет другой принцип действия. Траектория перемещения камеры вычисляется не сопоставлением особых точек, а путем минимизации фотометрической ошибки по всему изображению. Благодаря использованию полного изображения вместо набора особых точек, LSD-SLAM строит более плотную карту, но также имеет более высокие требования к вычислительным ресурсам, чем ORB-SLAM. При работе алгоритма LSD-SLAM в реальном времени на маломощном вычислительном устройстве могут возникнуть проблемы. Недостатками данного алгоритма являются также высокие требования к калибровке камеры и неустойчивость к выбросам.

Алгоритм LSD-SLAM состоит из трех основных модулей: tracking, depth map estimation и map optimization. Схема алгоритма представлена на рисунке 7.

Модуль **tracking** определяет перемещение камеры, отслеживая входящие изображения и вычисляя преобразование подобия между ними. Для вычисления преобразования используется минимизация фотометрической ошибки между новым изображением и преобразованным текущим ключевым кадром.

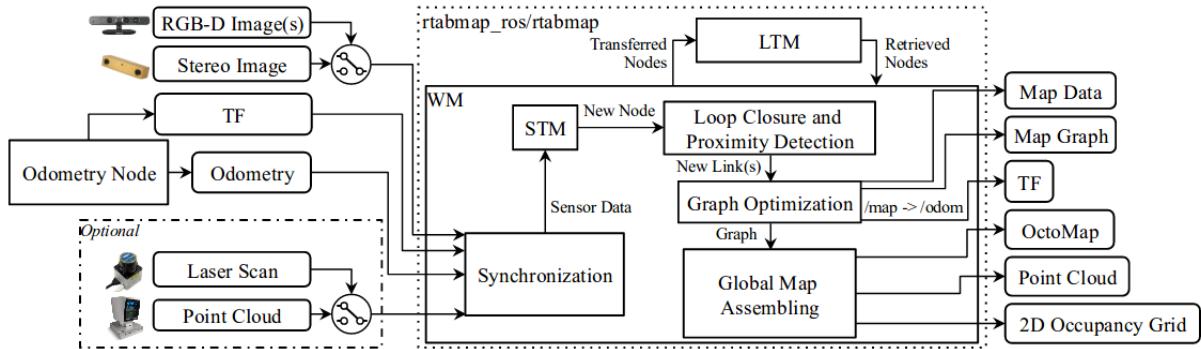


Рис. 8. Общая схема алгоритма RTAB-Мар

Модуль **depth map estimation** сравнивает кадр с текущим ключевым кадром, а затем уточняет или полностью заменяет его. Для сравнения используется взвешенная сумма расстояний и углов поворота между кадрами. Если она больше некоторого порога, ключевой кадр заменяется новым. По ключевому кадру вычисляется инвертированная карта глубины путем сопоставления точек с предыдущим ключевым кадром с использованием преобразования подобия, вычисленного модулем tracking. По вычисленному местоположению камеры и инвертированным картам глубины строится карта окружающей местности.

Модуль **map optimization** выполняет оптимизацию карты с использованием библиотеки g2o [17]. Оптимизация позволяет предотвратить накопление ошибок вычисления траектории и поддерживает точность построения карты.

Для хранения карты окружающей среды используется граф. Каждый узел этого графа хранит соответствующий ключевой кадр и инвертированную карту его глубины. Узлы соединяются ребрами, содержащими преобразование подобия между кадрами.

#### 4.1.3 RTAB-MAP

Алгоритм RTAB-MAP [25] предназначен для решения задачи SLAM с использованием информации о глубине изображений (по данным видеокамеры и лидара, или RGB-D камеры, или стереопары камер). Алгоритм использует три независимых процесса: вычисление движения камеры (одометрии), картирование и замыкание циклов. Схема алгоритма представлена на рисунке 8.

Для одометрии по кадрам вычисляются особые точки с помощью детектора BRIEF [7]. По сопоставлению особых точек на текущем и ключевом кадрах с помощью алгоритма PnP RANSAC [5] вычисляется перемещение камеры. Полученное положение камеры корректируется с помощью алгоритма Local Bundle Adjustment [50] и предсказаний на основе предыдущих движений камеры. Новый ключевой кадр добавляется, когда у текущего кадра и ключевого будет мало сопоставлений. Схема вычисления одометрии представлена на рисунке 9.

Картирование выполняется по локальным сеткам заполненности (occupancy grid),

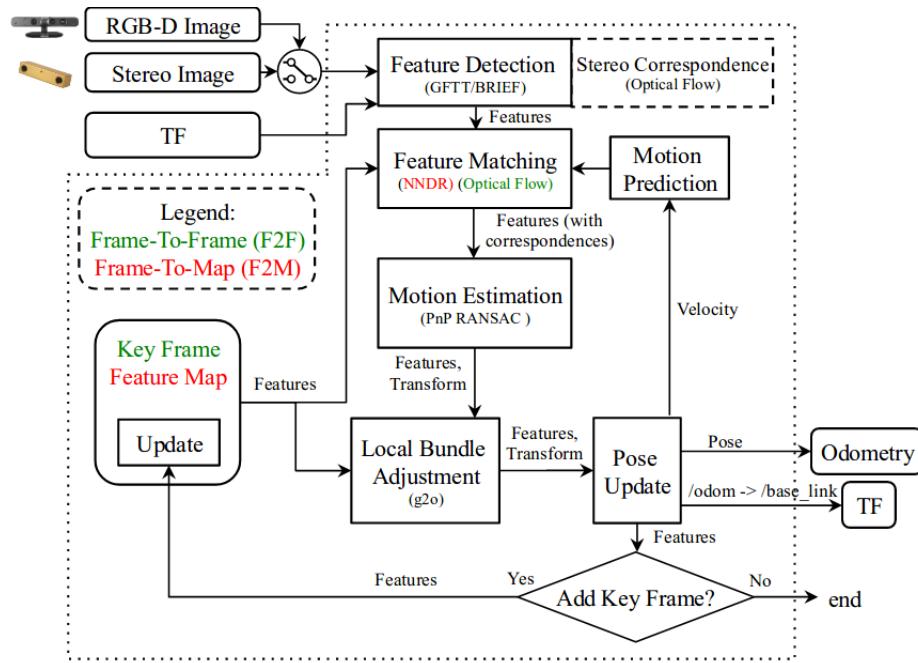


Рис. 9. Схема вычисления одометрии в методе RTAB-Мар

полученных из карт глубины. Локальные карты с помощью воксельного фильтра спиваются в глобальную карту. При замыкании цикла карта перестраивается. Схема процесса построения карты по локальным облакам точек представлена на рисунке 10.

Замыкание циклов основывается на сопоставлении особых точек на кадрах с видеопотока. Ключевая особенность данного метода - эффективное хранение изображений в памяти. Кадры хранятся в памяти как набор дескрипторов особых точек, организованный в kd-деревья. Дескрипторы извлекаются с помощью алгоритма SURF [1]. Алгоритм использует три вида памяти: WM (рабочая), в которой хранятся самые “полезные” кадры, STM (кратковременная), в которой хранятся последние кадры, и LTM (долгосрочная), в которой хранятся все кадры. Из STM в WM перемещаются те кадры, у которых больше всего похожих особых точек (похожесть мерится по дескрипторам). Для замыкания циклов используется кадр из рабочей памяти, который наиболее вероятно похож на текущий. Вероятности высчитываются байесовским фильтром. Схема процесса замыкания циклов представлена на рисунке 11.

Данный алгоритм имеет следующие преимущества в сравнении с другими методами SLAM:

1. Эффективная обработка данных с видеокамер и датчиков глубины в реальном времени
2. Эффективное замыкание циклов
3. Возможность работы в больших картах благодаря хранению долгосрочной памяти на жестком диске

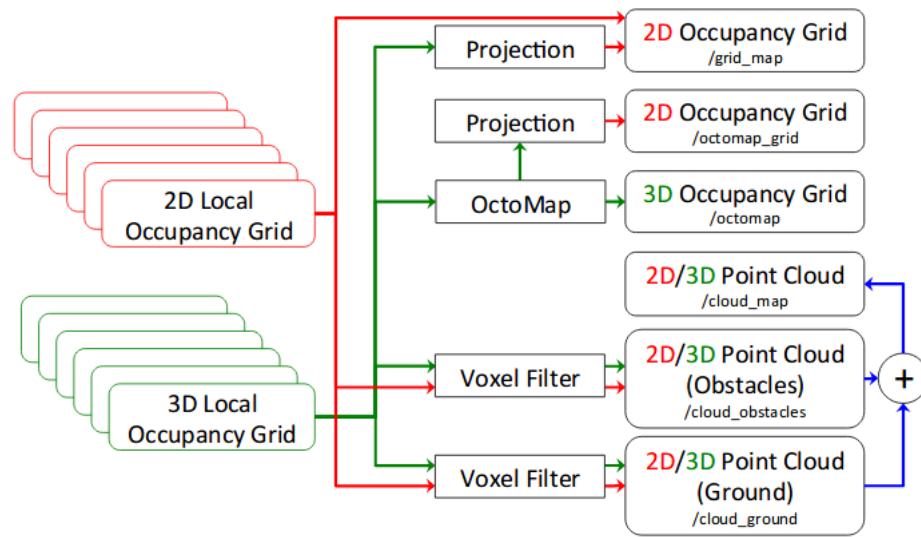


Рис. 10. Схема построения плотной глобальной карты по локальным сеткам

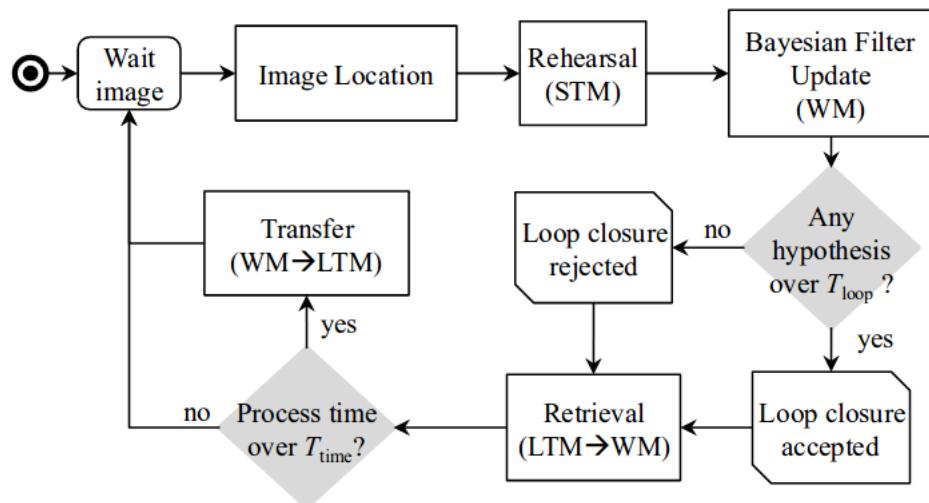


Рис. 11. Схема замыкания циклов в алгоритме RTAB-Мап

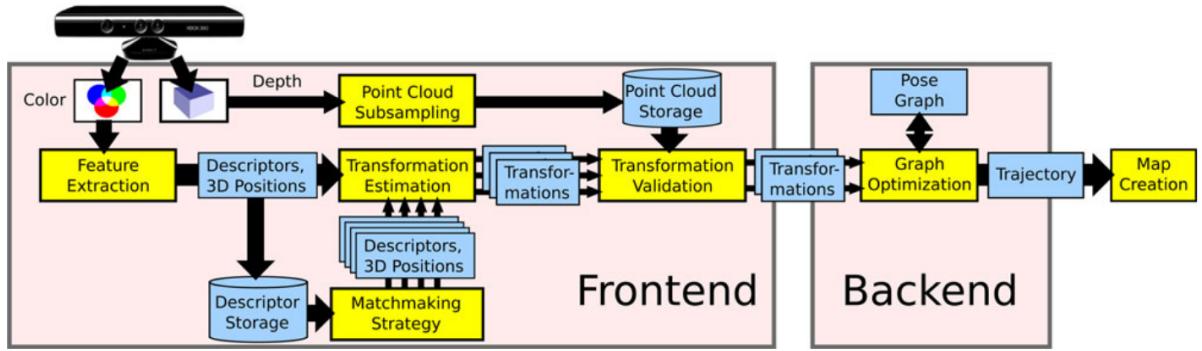


Рис. 12. Общая схема алгоритма RGBDSLAM\_v2

4. Высокая плотность построенной карты и возможность построения карты препятствий в формате Octomap
5. Легкость использования в различных приложениях, а также большое количество настраиваемых параметров

Помимо преимуществ, алгоритм RTAB-Мар обладает существенными недостатками:

1. Невозможность работы в монокулярном режиме
2. Потеря одометрии при отсутствии сопоставленных ориентиров
3. Высокая ресурсоемкость из-за необходимости обработки трехмерных облаков точек и построения плотной карты

#### 4.1.4 RGBDSLAM\_v2

Алгоритм RGBDSLAM\_v2 [12] также является популярным решением задачи одновременного картирования и локализации по данным с RGB-D камеры. Он появился примерно в то же время, что и описанный выше алгоритм RTAB-МАР, и основан на схожих принципах. Однако в деталях методов имеются существенные различия. Схема алгоритма RGBDSLAM\_v2 изображена на рисунке 12.

Вычисление перемещения камеры в алгоритме RGBDSLAM\_v2 осуществляется по тем же принципам, что и в алгоритме RTAB-МАР - путем сопоставления особых точек на ключевых кадрах. Изображение с камеры добавляется в множество ключевых кадров, когда у него не будет совпадающих особых точек с предыдущим кадром. Для извлечения особых точек используются детекторы SIFT [29], SURF [2] или ORB [36]. Сопоставления уточняются и фильтруются с помощью методов RANSAC [5] и ICP [10]. Преобразования, вычисленные по сопоставлениям особых точек, валидируются по картам глубин вероятностными методами.

По ключевым кадрам и найденным преобразованиям строится граф поз. Вершинами в этом графе являются ключевые кадры, ребрами - вычисленные алгоритмом

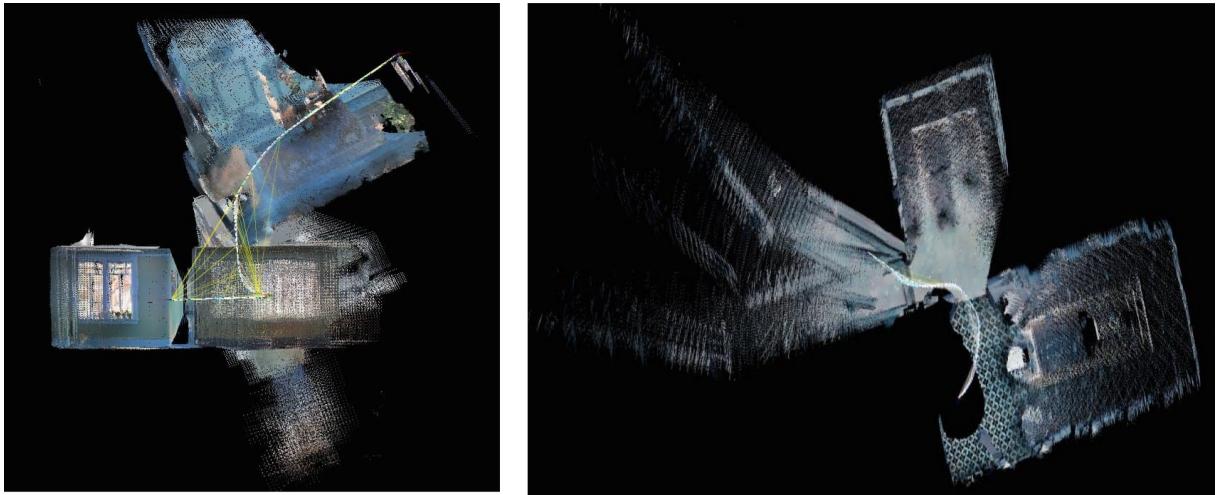


Рис. 13. Примеры некорректной работы алгоритма RGBDSLAM\_v2: ложное замыкание цикла (слева) и раздвоение коридора (справа)

геометрические преобразования между кадрами. По данному графу проводится глобальная оптимизация с помощью библиотеки g2o [17].

Метод замыкания циклов в алгоритме RGBDSLAM\_v2, так же, как и в RTAB-MAP, основан на вычислении преобразования по особым точкам между текущим кадром и похожими на него старыми ключевыми кадрами. Однако в алгоритме RGBDSLAM\_v2 используется более простой отбор кандидатов на "похожесть". По графу поз строится минимальное оствовное дерево, и для рассмотрения выбираются  $n$  предков текущего кадра в этом дереве, а также  $k$  случайно выбранных ключевых кадров в части дерева, оставшейся после удаления этих  $n$  предков, и еще  $l$  ключевых кадров, случайно выбранных по всему графу. Для эффективного замыкания циклов на разных длинах траекторий необходимо использовать разные  $n, k, l$ .

Алгоритм RGBDSLAM\_v2 обладает высокой вычислительной эффективностью, однако эксперименты, проведенные в работе [3], показали, что он обладает более низкой точностью по сравнению с алгоритмом RTAB-MAP. В частности, при резких поворотах робота могут найтись ложные замыкания или произойти раздвоение коридора на построенной алгоритмом карте (см. рис. 13).

## 4.2 Нейросетевые методы

В связи с бурным развитием вычислительной техники и нейронных сетей, в последние годы для решения задачи vSLAM также стали применяться методы, основанные на глубоком обучении. В подобных методах, как правило, вычисляются преобразования между позициями камеры с помощью сверточных нейронных сетей, также с помощью сверточных нейросетей вычисляются карты глубин по изображениям.

Одним из наиболее известных нейросетевых методов картирования и локализации является CNN-SLAM [42]. Как и в алгоритме LSD-SLAM [13], в CNN-SLAM из

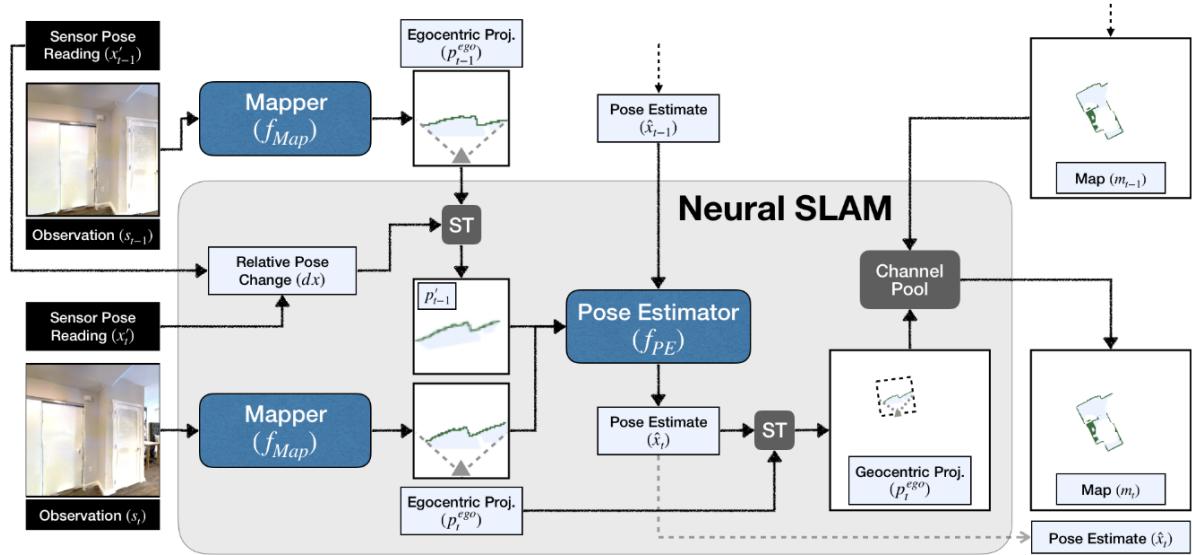


Рис. 14. Схема метода Active Neural SLAM

множества всех входящих кадров отбираются ключевые кадры. По ключевым кадрам строится граф позиций, который оптимизируется с помощью библиотеки g2o [17]. По каждому кадру вычисляется преобразование между этим кадром и текущим ключевым кадром путем минимизации фотометрической ошибки.

Метод CNN-SLAM принимает на вход видеопоток с камеры и по каждому изображению видеопотока вычисляет карту глубины с помощью полносверточной нейронной сети [26]. Также вместе с картой глубины вычисляется карта неопределенности, которая показывает, насколько глубина объектов, изображенных на текущем кадре, соотносится с глубинами этих объектов на других кадрах. Глубина каждого кадра оптимизируется исходя из карты неопределенности и глубины текущего ключевого кадра.

Помимо карты глубины, по каждому кадру входящего видеопотока с помощью нейросети выполняется семантическая сегментация. По картам глубин и картам сегментации с помощью глобальной сегментационной модели (GSM) [41] строится итоговая карта окружающей местности.

Эксперименты на датасетах TUM [40] и ICL-NUIM [18] показали, что по точности определения местоположения алгоритм CNN-SLAM более чем в два раза превосходит классический метод LSD-SLAM [13]. Однако качество метода CNN-SLAM, как и любого метода, использующего предобученные нейросети, сильно зависит от качества обучающей выборки.

Одним из наиболее известных недавних нейросетевых методов vSLAM является Active Neural SLAM [9]. Данный метод решает задачу vSLAM в контексте более глобальной задачи - исследование неизвестной местности (Exploration). Помимо картирования и локализации, алгоритм выполняет постановку цели для исследования и планирование траектории.

Метод Active Neural SLAM принимает на вход видеопоток с камеры и данные одо-

метрии с сенсоров робота. На выходе получается двумерная карта местности и траектория перемещения робота. В отличие от метода CNN-SLAM, где карта строится по предсказанным нейросетью глубинам с помощью различных методов оптимизации, в методе Active Neural SLAM локальные карты предсказываются непосредственно нейросетью по каждому изображению видеопотока. Перемещение робота также предсказывается нейросетью по локальным картам входящим данным одометрии (как правило, сильно зашумленным). Схема алгоритма Active Neural SLAM изображена на рисунке 14.

Эксперименты, проведенные на датасете Gibson [46], показали, что метод Active Neural SLAM способен исследовать в среднем 95% помещения, решая задачу vSLAM совместно с задачей планирования маршрутов и исследования неизвестной местности. Однако оценка качества построения карты и вычисления траектории в работе [9] не производилась.

### 4.3 Выводы

В данной главе были рассмотрены классические и нейросетевые методы решения задачи vSLAM. Среди классических методов были рассмотрены следующие:

1. ORB-SLAM - метод, основанный на сопоставлении особых точек на изображениях
2. LSD-SLAM - метод, основанный на вычислении геометрических преобразований путем минимизации фотометрической ошибки
3. RTAB-MAP - метод, принимающий на вход данные со стереокамеры или RGB-D камеры, обладающий эффективным замыканием циклов
4. RGBDSLAM - метод, принимающий на вход данные с RGB-D камеры, с более простым замыканием циклов по сравнению с RTAB-MAP

Алгоритм ORB-SLAM обладает высокой вычислительной эффективностью, однако строит разреженную карту местности, не пригодную для планирования маршрутов. Метод LSD-SLAM строит довольно плотную карту, однако он более требователен к вычислительным ресурсам и имеет в среднем более низкую точность. Методы RTAB-MAP и RGBDSLAM строят плотную карту и имеют довольно высокую точность, однако они требуют на вход данные со стереокамеры или RGBD-камеры. Для их работы по данным с единственной камеры может применяться восстановление глубин изображений с помощью нейронных сетей.

Помимо классических методов, были также рассмотрены нейросетевые методы картирования и локализации:

1. CNN-SLAM - построение карты по предсказанным нейросетью глубинам и картам сегментации с помощью глобальных моделей

2. Active Neural SLAM - построение карты и вычисление траектории с помощью нейросетей непосредственно

На некоторых коллекциях данных методы CNN-SLAM и Active Neural SLAM имеют более высокое качество по сравнению с классическими методами. Однако их применение на реальных роботах затруднено, поскольку данные методы требуют наличие графического ускорителя на борту, и качество их работы напрямую зависит от выборки, на которой производилось обучение нейросетей.

В данной работе был выбран алгоритм RTAB-MAP, поскольку он строит плотную трехмерную карту местности и не требователен к вычислительным ресурсам. Для применения алгоритма на данных с единственной видеокамеры используется восстановление карт глубин с помощью полносверточных нейронных сетей с легкой архитектурой, способных работать с высокой скоростью.

# Глава 5

## Вспомогательные задачи для vSLAM

### 5.1 Восстановление глубины по видеопотоку

Классические методы решения задачи vSLAM по данным с монокулярной камеры [31] [13], описанные в предыдущей главе, выдают разреженную карту, непригодную для планирования маршрутов. Построение плотной карты возможно при наличии информации о глубине изображений, например, при помощи методов [25], [12]. Таким образом, для построения плотной карты окружающей местности по данным с единственной видеокамеры необходимо решить задачу восстановления глубины изображений.

В настоящее время задача восстановления глубины изображений решается, как правило, с помощью глубоких нейронных сетей. Нейросети, предсказывающие глубину изображений по видеопотоку, делятся на два типа. Первые принимают на вход одиночное изображение, не используя информацию о контексте из видео. Вторые предсказывают глубину изображений с использованием информации об изменении кадров в видеопотоке. Ниже рассмотрены обе группы нейросетевых методов восстановления глубины.

#### 5.1.1 Восстановление глубины по одиночным изображениям

В настоящее время существует большое количество нейросетевых архитектур для восстановления глубины по одиночным изображениям. Как правило, такие нейросети состоят из сверточного энкодера и декодера, состоящего из нескольких слоев развертки и повышения дискретизации (Upsampling).

Одной из наиболее известный нейросетей восстановления глубины является FCRN [26]. Архитектура данной сети состоит из энкодера и декодера. В качестве энкодера была выбрана широко известная сеть ResNet [19], содержащая 50 сверточных слоев. В качестве декодера используется 5 блоков Up-Convolution или Up-Projection (см. рис. 15).

Обучение сети FCRN производилось на датасете NYU Depth v2 [39]. Исследова-

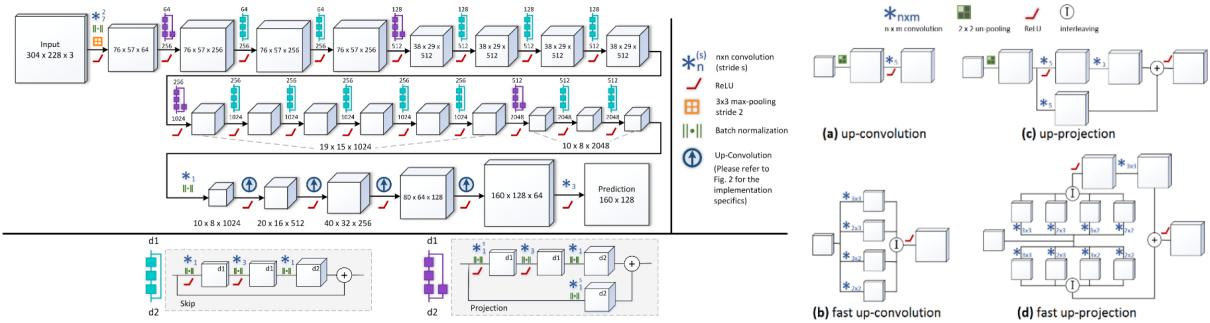


Рис. 15. Схема архитектуры FCRN (слева) и различных блоков декодера - Up-Projection и Up-Convolution (справа)

лись разные функции потерь - стандартная функция MSE, а также функции Huber и BerHu [33] [52], являющиеся комбинацией квадратичной и линейной ошибки. Наилучший результат показала архитектура с декодером Up-Projection, обученная с функцией потерь BerHu. Относительная ошибка этой архитектуры на валидационной выборке NYU Depth v2 составила 0.127. Была также измерена скорость работы данной архитектуры на видеокарте GTX Titan. Она составила 78 мс на одно изображение, или 13 кадров в секунду, что недостаточно для работы в реальном времени (для обработки всех кадров стандартного видеопотока необходима скорость не менее 30 кадров в секунду).

В 2020 году была предложена нейросетевая архитектура восстановления глубины с более высокой скоростью и качеством работы [20]. В этой архитектуре предсказание глубины уточняется с помощью механизма внимания, установленного между энкодером и декодером. Механизм внимания предсказывает, насколько глубина одного участка изображения может быть полезна для предсказания глубины другого участка, формируя карты внимания глубины (depth-attention maps). Для построения истинных карт внимания, необходимых для обучения сети, по изображенному на кадре участку сцены вычисляется несколько плоскостей. Значением внимания глубины одного пикселя для другого является максимальная сумма расстояний точек, спроецированных в эти пиксели, до плоскостей. Такой метод обосновывается тем, что у точек, расположенных в одной плоскости, легко вычислить глубину с помощью геометрических преобразований, если известна точная глубина хотя бы одной из этих точек.

С помощью вышеописанной архитектуры и механизма внимания была достигнута относительная ошибка 0.108 на датасете NYU Depth v2. Скорость работы данной архитектуры на видеокарте GTX 1080 составила 218 кадров в секунду. Однако на бортовых вычислителях робототехнических систем, не обладающих мощными видеокартами, скорость работы нейросети будет значительно меньше. Точно измерить скорость работы описанной нейросети на бортовых вычислителях не удалось, поскольку исходный код описанной архитектуры на момент написания данной работы предоставлен не был.

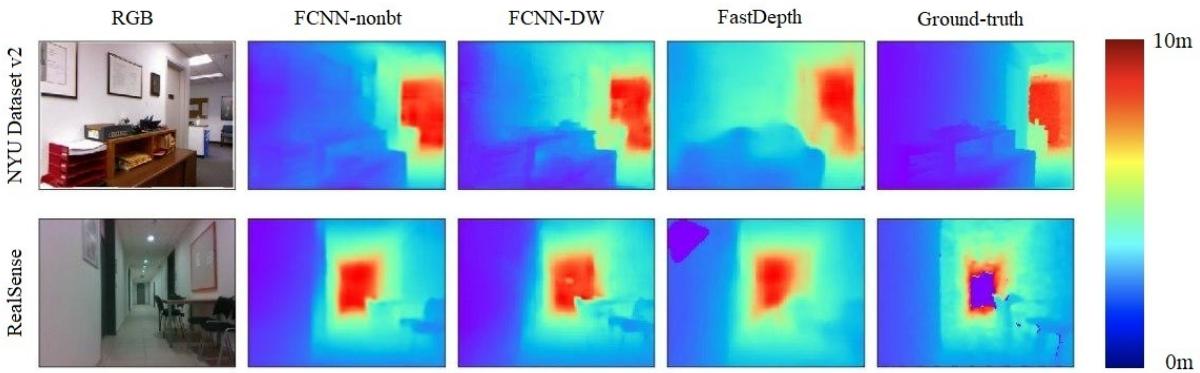


Рис. 16. Примеры глубин, предсказанных нейросетью FastDepth (справа), в сравнении с глубинами, предсказанными нейросетями из данной работы

В работе [45] описана архитектура FastDepth, пригодная для работы в реальном времени даже на маломощных вычислительных устройствах. С помощью различных технических и архитектурных оптимизаций авторам работы удалось добиться скорости восстановления глубины 175 кадров в секунду на встраиваемом компьютере NVidia Jetson TX2, который широко используется в различных робототехнических системах. Ошибка восстановления глубины на NYU Depth v2 при этом составила 0.158 - почти в полтора раза выше, чем у архитектуры из работы [20]. Контуры объектов на предсказанных картах глубины также получились размытыми (см. рис. 16), что может негативно сказаться на качестве карты при применении предсказанных глубин в задаче vSLAM.

В данной работе с учетом преимуществ и недостатков всех вышеописанных методов была разработана новая нейросетевая архитектура для достижения баланса между скоростью и качеством восстановления глубины. В основе разработанной архитектуры лежит классическая схема энкодер-декодер. В качестве энкодера выбрана сеть ResNet с 50 слоями, как и в архитектуре FCRN. В качестве декодера была выбрана серия сверток depthwise convolution, как и в архитектуре FastDepth. Также для повышения качества работы и четкости контуров были добавлены прямые связи (shortcuts) из энкодера в декодер. С помощью предложенной архитектуры удалось достичь приемлемого качества восстановления глубины (относительная ошибка 0.170 на NYU Depth v2) и вместе с тем приемлемой скорости работы. Примеры глубин, предсказанных разработанной нейросетью, показаны на рисунке 16 посередине.

### 5.1.2 Восстановление глубины с извлечением информации из видеопотока

При наличии видеопотока можно получить дополнительную информацию о глубине изображений из перемещения объектов на кадрах. Например, в методе LSD-SLAM [13] карты глубин изображений вычисляются с использованием преобразова-

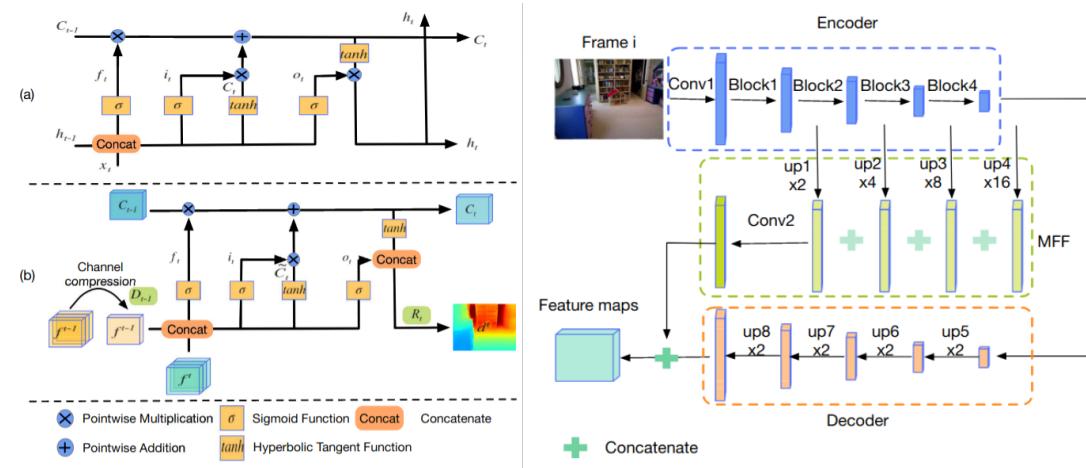


Рис. 17. Схема архитектуры ST-CLSTM: часть, кодирующая пространственные признаки (слева), и модуль Conv-LSTM (справа)

ний подобия, задающих перемещение камеры от кадра к кадру и вычисляемых путем минимизации фотометрической ошибки. С помощью глубоких нейронных сетей также можно извлекать временную информацию из видеопотока, тем самым уточняя карты глубин каждого изображения.

В работе [48] описана архитектура нейронной сети ST-CLSTM, в которой временная информация извлекается из видеопотока с помощью сверточно-рекуррентных блоков ConvLSTM. Модуль ConvLSTM представляет собой комбинацию слоев, используемую в традиционном модуле LSTM [16], в которой полносвязные слои заменены на сверточные.

Архитектура ST-CLSTM состоит из двух частей. Её схема изображена на рисунке 17. Первая часть имеет схему энкодер-декодер и по входящему изображению вычисляет набор пространственных признаков. В качестве энкодера используется сеть ResNet-18 [19]. Декодер представляет собой четыре блока, состоящих из последовательного повышения дискретизации и двух параллельных сверток. Для повышения информативности вычисляемых признаков в архитектуру первой части встроен также модуль слияния признаков с несколькими масштабами (Multi-scale feature fusion, MFF). Модуль MFF представляет собой конкатенацию выходов четырех блоков энкодера, приведенных к размеру исходного изображения с помощью слоев повышения дискретизации. Данная конкатенация преобразуется с помощью сверточного слоя и стыкуется с выходом декодера, в результате чего получается карта пространственных признаков.

Вторая часть архитектуры ST-CLSTM состоит из блока ConvLSTM и двух сверточных слоев. Слои блока LSTM используют информацию, извлеченную из карт пространственных признаков предыдущих изображений. Сверточные слои преобразуют данную информацию в карту глубины.

Для обучения сети ST-CLSTM использовалась комбинация пространственной и временной функции потерь. Пространственная функция потерь представляла со-

бой среднюю логарифмическую ошибку между предсказанной и истинной глубиной. Для эффективного обучения временной составляющей сети ST-CLSTM применялся генеративно-состязательный подход. В качестве генератора выступала непосредственно сеть ST-CLSTM, а в качестве дискриминатора - трехмерная сверточная нейронная сеть, принимающая на вход последовательность карт глубин и предсказывающая происхождение этих глубин (истинные или предсказанные нейросетью-генератором).

Эксперименты, проведенные на датасете NYU Depth v2 [39], показали, что с помощью архитектуры ST-CLSTM удалось добиться довольно высокого качества восстановления глубины. Относительная ошибка составила 0.132. В ходе экспериментов была также измерена скорость обработки одного изображения на видеокарте GTX 1080Ti. Скорость работы оказалась равной 33 кадрам в секунду, что достаточно для обработки видеопотока в реальном времени. Однако на менее мощной видеокарте, в частности, на бортовых вычислителях большинства робототехнических систем, данная архитектура будет работать значительно медленнее, что затруднит ее использование в реальном времени. Например, скорость работы на встраиваемом компьютере NVidia Jetson TX2 составила менее 10 кадров в секунду.

В работе [30] для предсказания глубины изображений используется традиционная полносверточная нейросеть, а предсказанные нейросетью глубины дополнительно уточняются с помощью оптического потока и метода COLMAP [38]. Оптический поток вычисляется нейронной сетью FlowNet2 [21].

С помощью вычисленного оптического потока нейросеть восстановления глубины может дообучаться на конкретном видео без данных об истинных глубинах изображений. По предсказанным нейросетью глубинам и перемещениях пикселей изображений (оптического потока) вычисляется временная консистентность, которая служит в данном случае функцией потерь при обучении.

Коррекция предсказанных глубин методом COLMAP позволила существенно повысить качество восстановления глубины в тех случаях, когда тестовая выборка значительно отличается от обучающей. Так, на датасете ScanNet [11] относительная ошибка предсказания глубины составила 0.073 (против 0.208 у полносверточной нейронной сети без коррекции). Однако данный подход малопригоден для применения в реальном времени на борту робота - нейросеть FlowNet, вычисляющая оптический поток, обрабатывает одну пару кадров более 100 мс на видеокарте компьютера NVidia Jetson TX2; а также метод COLMAP потребляет значительную часть ресурсов процессора, что затрудняет работу алгоритмов SLAM.

В работах [51], [47], [43] совместно с задачей восстановления глубины решается также задача визуальной одометрии - вычисления перемещения камеры по видеоданным с нее. Для вычисления карт глубин и перемещения камеры используются отдельные нейронные сети, которые обучаются совместно с использованием общей функции потерь, основанной на пространственно-временной консистентности. Таким

образом, возможно обучение нейросетей при отсутствии данных об истинной глубине изображений.

Совместное использование нейросетей глубины и позиции позволило достичь очень низкой ошибки восстановления глубины. Так, в работе [43] относительная ошибка на датасете NYU Depth v2 [39] составила 0.061 (по сравнению с ошибкой в 0.106 у классических полносверточных нейросетей). Однако параллельная работа двух нейронных сетей в реальном времени на бортовом вычислителе затруднительна. Так, в работе [43] время обработки одного кадра составляет 690 мс (при том, что в стандартном видеопотоке кадры приходят каждые 33 мс).

### 5.1.3 Выводы

Восстановление глубины изображений дает возможность строить плотную карту местности по данным с единственной видеокамеры с помощью методов vSLAM. В данной главе приведен обзор основных методов восстановления глубины по изображениям. Методы, использующие в качестве входных данных одиночное изображение ([26], [45]) обладают достаточно высокой скоростью работы для обработки стандартного видеопотока в реальном времени, однако имеют довольно низкое качество (относительная ошибка порядка 13-17%). Методы, принимающие во внимание информацию о перемещении между кадрами ([48] [30] [43]), имеют значительно более высокое качество (относительная ошибка восстановления глубины порядка 6-10%), однако обладают низкой скоростью работы, что затрудняет их применение на борту робототехнической системы в реальном времени.

В данной работе для решения задачи VSLAM была выбрана полносверточная нейронная сеть, принимающая на вход одиночные изображения. Архитектура нейронной сети была оптимизирована для работы в реальном времени на встраиваемом компьютере NVidia Jetson TX2, который может применяться в качестве бортового вычислителя на малых робототехнических системах. Подробное описание данной архитектуры приведено в главе 6.

# Глава 6

## Одновременное картирование и локализация с использованием вспомогательных методов

### 6.1 Описание метода

В данной работе представлен алгоритм одновременного картирования и локализации по видеопотоку с единственной камеры, основанный на восстановлении карт глубин изображений. По изображениям, поступающим с видеокамеры, вычисляются карты глубин с помощью полносверточной нейронной сети. По изображениям и предсказанным картам глубин осуществляется картирование и локализация с помощью метода RTAB-Мар, описанного в разделе 4.1.3. Схема алгоритма представлена на рисунке 18.

Нейросеть, используемая для восстановления глубины, имеет полносверточную архитектуру, состоящую из энкодера и декодера. Сеть принимает на вход трехканальное цветное изображение размера 320x240 и выдает карту глубины такого же размера. В качестве энкодера используется сеть ResNet-50 [19] без полносвязных слоев, предобученная на датасете MS Coco [28]. На выходе энкодера получается карта высокоуровневых признаков размерности 10x8x2048. Декодер состоит из пяти блоков, каждый из которых повышает пространственную размерность в два раза. Каждый блок состоит из слоя повышения дискретизации (Upsampling), свертки с ядром 1x1 (pointwise convolution) и поканальной свертки (depthwise convolution) с ядром 3x3. Схема архитектуры нейросети представлена на рисунке 19.

Обучение нейросети восстановления глубины проводилось на датасете NYU Depth v2 [39]. Датасет содержит 464 видеопоследовательности, снятые в разных помещениях с размеченными картами глубин. Суммарно во всех последовательностях содержится более 400000 пар изображение-глубина. Для обучения была отобрана выборка из 10% изображений датасета (порядка 40000). Из них около 30000 изображений составляла обучающая выборка и около 10000 - валидационная. В обучающей и

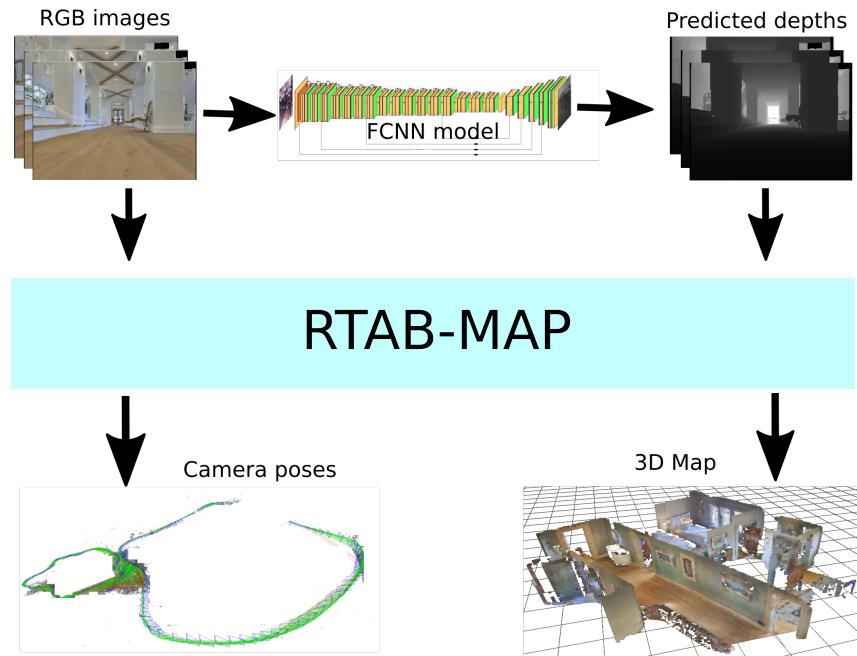


Рис. 18. Схема представленного в данной работе метода vSLAM

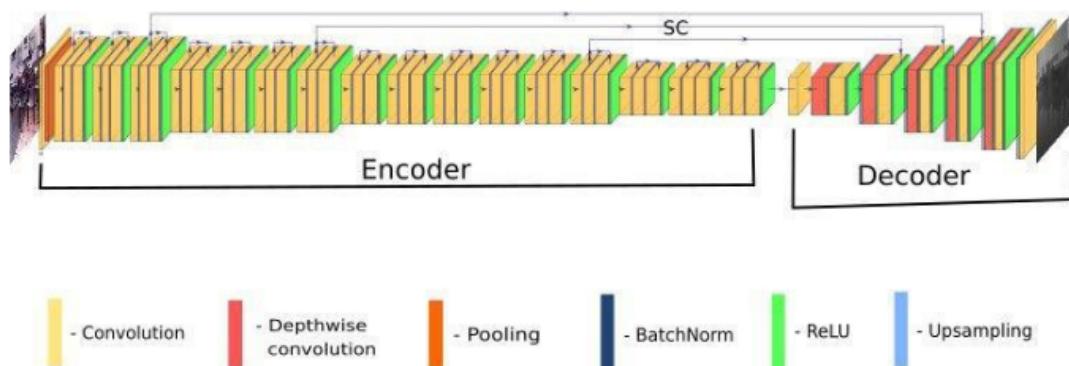


Рис. 19. Схема нейронной сети восстановления глубины

валидационной выборке использовались изображения из разных сцен.

Оптимизация параметров проводилась методом Adam [22] с параметрами  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  в течение 30 эпох. В качестве функции потерь использовалась комбинация среднеквадратичной и относительной ошибки между истинной и предсказанный глубиной:

$$L(D, \hat{D}) = \frac{1}{H \cdot W} \left( \sum_{i,j} (D_{i,j} - \hat{D}_{i,j})^2 + \alpha \sum_{i,j} \left( \frac{D_{i,j} - \hat{D}_{i,j}}{D_{i,j}} \right)^2 \right) \quad (11)$$

Таблица 1. Оценка качества восстановления глубины и сравнительный анализ различных нейросетей

Архитектура	NYU Dataset v2		RealSense			FPS
	RMSE	REL	RMSE	REL	$\delta^1$	
FastDepth [45]	<b>0.594</b>	0.170	1.315	0.292	0.467	<b>55</b>
FCNN-nonbt (ours) [4]	0.647	0.173	<b>0.936</b>	<b>0.202</b>	<b>0.608</b>	16
FCNN-DW (ours)	0.696	<b>0.158</b>	1.179	0.257	0.441	24

После обучения была проведена оценка качества нейросети на официальной тестовой выборке датасета NYU Depth v2, содержащей 1449 изображений с высокоточной глубиной с качественной пост-обработкой. Также для проверки устойчивости нейросети к изменениям выборки была проведена оценка качества на наборе данных, собранных с RGB-D камеры Intel Realsense. Набор содержал приблизительно 9500 пар изображение-глубина.

Для оценки качества вычислялись три основные метрики: среднеквадратичная ошибка (RMSE), относительная ошибка (REL), и также метрика  $\delta_1$  - доля пикселей, на которых относительная ошибка глубины не превысила 0.25. Помимо метрик качества, также была измерена скорость обработки изображений на бортовом компьютере NVidia Jetson TX2. Результаты сравнения представлены в таблице 1.

Эксперименты показали, что представленная в данной работе нейросеть имеет приемлемое качество восстановления глубины, сравнимое с другими современными архитектурами, и при этом обладает достаточной эффективностью для работы в реальном времени на встраиваемом компьютере (при скорости в 24 кадра в секунду сеть способна обрабатывать почти каждый кадр входящего видеопотока). Данная сеть имеет схожее качество по метрикам с архитектурой FastDepth [45], однако предсказанные глубины у нее получаются значительно более четкими (см. рис. 16). Также эксперименты показали, что данная нейросеть обладает более высокой обобщающей способностью, чем FastDepth (качество восстановления глубины на данных с RealSense оказалось лучше по всем метрикам).

Помимо оценки качества восстановления глубины, была также проведена оценка

качества работы метода vSLAM с разными нейросетевыми архитектурами восстановления глубины. Результаты оценки качества vSLAM описаны ниже.

## 6.2 Программная реализация

Тут про ROS, TensorRT, сетку и ноды запуска.

## 6.3 Эксперименты

### 6.3.1 Эксперименты в симуляторе

Описание экспериментов на датасете MAOMaps.

### 6.3.2 Эксперименты на реальном роботе

Описание экспериментов на живом роботе (МПРМ или Хаски).

## 6.4 Выводы

Выводы о работе нашего слама.

# Глава 7

## Применение в задаче исследования неизвестной местности

### 7.1 Описание задачи исследования неизвестной местности

#### 7.1.1 Постановка задачи

Постановка задачи эксплорейшена.

#### 7.1.2 Метрики качества

Метрики качества эксплорейшена - покрытая за n секунд площадь и т.д.

### 7.2 Описание метода

Описание нашего фронтъерного эксплорейшена.

### 7.3 Эксперименты

Описание экспериментов на Gibson.

### 7.4 Выводы

Выводы о качестве и стабильности работы эксплорейшена.

# Глава 8

## Заключение

Заключение.

# Список литературы

- [1] Herbert Bay, Tinne Tuytelaars, Luc Van Gool. “Surf: Speeded up robust features”. *European conference on computer vision*. Springer. 2006, c. 404—417.
- [2] Herbert Bay и др. “Speeded-up robust features (SURF)”. *Computer vision and image understanding* **110** 3 (2008), c. 346—359.
- [3] Andrey Bokovoy, Kirill Muraviev, Konstantin Yakovlev. “Map-merging algorithms for visual slam: Feasibility study and empirical evaluation”. *Russian Conference on Artificial Intelligence*. Springer. 2020, c. 46—60.
- [4] Andrey Bokovoy, Kirill Muravyev, Konstantin Yakovlev. “Real-time vision-based depth reconstruction with Nvidia Jetson”. *2019 European Conference on Mobile Robots (ECMR)*. IEEE. 2019, c. 1—6.
- [5] Eric Brachmann и др. “DSAC-differentiable RANSAC for camera localization”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, c. 6684—6692.
- [6] Michael Burri и др. “The EuRoC micro aerial vehicle datasets”. *The International Journal of Robotics Research* **35** 10 (2016), c. 1157—1163.
- [7] Michael Calonder и др. “Brief: Binary robust independent elementary features”. *European conference on computer vision*. Springer. 2010, c. 778—792.
- [8] Angel Chang и др. “Matterport3d: Learning from rgb-d data in indoor environments”. *arXiv preprint arXiv:1709.06158* (2017).
- [9] Devendra Singh Chaplot и др. “Learning to explore using active neural slam”. *arXiv preprint arXiv:2004.05155* (2020).
- [10] Dmitry Chetverikov, Dmitry Stepanov, Pavel Krsek. “Robust Euclidean alignment of 3D point sets: the trimmed iterative closest point algorithm”. *Image and vision computing* **23** 3 (2005), c. 299—309.
- [11] Angela Dai и др. “Scannet: Richly-annotated 3d reconstructions of indoor scenes”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, c. 5828—5839.
- [12] Felix Endres и др. “3-D mapping with an RGB-D camera”. *IEEE transactions on robotics* **30** 1 (2013), c. 177—187.

- [13] Jakob Engel, Thomas Schöps, Daniel Cremers. “LSD-SLAM: Large-scale direct monocular SLAM”. *European conference on computer vision*. Springer. 2014, c. 834—849.
- [14] Andreas Geiger, Philip Lenz, Raquel Urtasun. “Are we ready for autonomous driving? the kitti vision benchmark suite”. *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, c. 3354—3361.
- [15] Andreas Geiger и др. “The KITTI vision benchmark suite”. URL <http://www.cvlibs.net/datasets/kitti> 2 (2015).
- [16] Klaus Greff и др. “LSTM: A search space odyssey”. *IEEE transactions on neural networks and learning systems* **28** 10 (2016), c. 2222—2232.
- [17] Giorgio Grisetti и др. “g2o: A general framework for (hyper) graph optimization”. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China*. 2011, c. 9—13.
- [18] Ankur Handa и др. “A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM”. *2014 IEEE international conference on Robotics and automation (ICRA)*. IEEE. 2014, c. 1524—1531.
- [19] Kaiming He и др. “Deep residual learning for image recognition”. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, c. 770—778.
- [20] Lam Huynh и др. “Guiding monocular depth estimation using depth-attention volume”. *European Conference on Computer Vision*. Springer. 2020, c. 581—597.
- [21] Eddy Ilg и др. “Flownet 2.0: Evolution of optical flow estimation with deep networks”. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, c. 2462—2470.
- [22] Diederik P Kingma, Jimmy Ba. “Adam: A method for stochastic optimization”. *arXiv preprint arXiv:1412.6980* (2014).
- [23] Nathan Koenig, Andrew Howard. “Design and use paradigms for gazebo, an open-source multi-robot simulator”. *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*. T. 3. IEEE. 2004, c. 2149—2154.
- [24] Kazutaka Kurihara и др. “Optical motion capture system with pan-tilt camera tracking and real time data processing”. *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*. T. 2. IEEE. 2002, c. 1241—1248.
- [25] Mathieu Labbe, Francois Michaud. “Memory management for real-time appearance-based loop closure detection”. *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2011, c. 1271—1276.
- [26] Iro Laina и др. “Deeper depth prediction with fully convolutional residual networks”. *2016 Fourth international conference on 3D vision (3DV)*. IEEE. 2016, c. 239—248.

- [27] Zhengqi Li, Noah Snavely. “Megadepth: Learning single-view depth prediction from internet photos”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, c. 2041—2050.
- [28] Tsung-Yi Lin и др. “Microsoft coco: Common objects in context”. *European conference on computer vision*. Springer. 2014, c. 740—755.
- [29] David G Lowe. “Distinctive image features from scale-invariant keypoints”. *International journal of computer vision* **60** 2 (2004), c. 91—110.
- [30] Xuan Luo и др. “Consistent video depth estimation”. *ACM Transactions on Graphics (TOG)* **39** 4 (2020), c. 71—1.
- [31] Raul Mur-Artal, Jose Maria Martinez Montiel, Juan D Tardos. “ORB-SLAM: a versatile and accurate monocular SLAM system”. *IEEE transactions on robotics* **31** 5 (2015), c. 1147—1163.
- [32] Raul Mur-Artal, Juan D Tardós. “Orb-slam2: An open-source slam system for monocular, stereo, and rgbd cameras”. *IEEE Transactions on Robotics* **33** 5 (2017), c. 1255—1262.
- [33] Art B Owen. “A robust hybrid of lasso and ridge regression”. *Contemporary Mathematics* **443** 7 (2007), c. 59—72.
- [34] Vaishakh Patil и др. “Don’t forget the past: Recurrent depth estimation from monocular video”. *IEEE Robotics and Automation Letters* **5** 4 (2020), c. 6813—6820.
- [35] S Rooban и др. “CoppeliaSim: Adaptable modular robot and its different locomotions simulation framework”. *Materials Today: Proceedings* (2021).
- [36] Ethan Rublee и др. “ORB: An efficient alternative to SIFT or SURF”. *2011 International conference on computer vision*. Ieee. 2011, c. 2564—2571.
- [37] Manolis Savva и др. “Habitat: A platform for embodied ai research”. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, c. 9339—9347.
- [38] Johannes L Schonberger, Jan-Michael Frahm. “Structure-from-motion revisited”. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, c. 4104—4113.
- [39] Nathan Silberman и др. “Indoor segmentation and support inference from rgbd images”. *European conference on computer vision*. Springer. 2012, c. 746—760.
- [40] Jürgen Sturm и др. “A benchmark for the evaluation of RGB-D SLAM systems”. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2012, c. 573—580.
- [41] Keisuke Tateno, Federico Tombari, Nassir Navab. “Real-time and scalable incremental segmentation on dense slam”. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2015, c. 4465—4472.

- [42] Keisuke Tateno и др. “Cnn-slam: Real-time dense monocular slam with learned depth prediction”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, с. 6243–6252.
- [43] Zachary Teed, Jia Deng. “Deepv2d: Video to depth with differentiable structure from motion”. *arXiv preprint arXiv:1812.04605* (2018).
- [44] Oliver Wasenmüller, Marcel Meyer, Didier Stricker. “CoRBS: Comprehensive RGB-D benchmark for SLAM using Kinect v2”. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2016, с. 1–7.
- [45] Diana Wofk и др. “Fastdepth: Fast monocular depth estimation on embedded systems”. *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, с. 6101–6108.
- [46] Fei Xia и др. “Gibson env: Real-world perception for embodied agents”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, с. 9068–9079.
- [47] Nan Yang и др. “D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, с. 1281–1292.
- [48] Haokui Zhang и др. “Exploiting temporal consistency for real-time video depth estimation”. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, с. 1725–1734.
- [49] Zhengyou Zhang. “Microsoft kinect sensor and its effect”. *IEEE multimedia* **19** 2 (2012), с. 4–10.
- [50] Zhengyou Zhang, Ying Shan. *Incremental motion estimation through local bundle adjustment*. US Patent 6,996,254. Февр. 2006.
- [51] Tinghui Zhou и др. “Unsupervised learning of depth and ego-motion from video”. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, с. 1851–1858.
- [52] Laurent Zwald, Sophie Lambert-Lacroix. “The berhu penalty and the grouped effect”. *arXiv preprint arXiv:1207.6868* (2012).