Министерство науки и высшего образования Российской Федерации
Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Магистерская программа
«Методы и технологии искусственного интеллекта»

Иванов Леонид Андреевич

Нейросетевые методы генерации HD-карт местности

09.04.01 Информатика и вычислительная техника

Литературный обзор для Магистерской диссертации

Научный руководитель:
к.т.н.
Юдин Дмитрий Александрович

Москва

2024 г.

Содержание

1 Введение **3**

2 Постановка задачи **4**

2.1 Метрики оценки качества генерации HD-карт 4

2.1.1 IoU 4

2.1.2 Chamfer distance 5

2.1.3 AP 5

2.2 Цель 6

3 Классификация методов построения HD-карт **6**

3.1 Построение HD-карт на основе семантической сегментации 7

3.1.1 HDMapNet, 2021 7

3.1.2 SuperFusion, 2022 9

3.1.3 Lidar2Map, 2023 10

3.1.4 Вывод 12

3.2 Генерация HD-карт с использованием трансформеров 12

3.2.1 ADMap, 2024 13

3.2.2 MapQR, 2024 15

3.2.3 Вывод 16

3.3 Генерация HD-карт с учетом временной информации 16

3.3.1 StreamMapNet, 2023 16

3.3.2 MapTracker, 2024 19

3.3.3 Вывод 21

4 Заключение **21**

1 Введение

Карты высокой точности (HD Maps) являются одним из важнейших компонентов технологии автономного вождения. Использование информации об окружающей среде, такой как границы полос движения, разделительные полосы и пешеходные переходы, играет решающую роль в навигации транспортного средства, прогнозировании движения и планировании. Традиционные методы построения HD-карт, такие как одновременная локализация и картографирование (SLAM) и ручное создание аннотаций к картам, требуют значительного времени и усилий человека. Эти ограничения приводят к переходу к обучаемым методам, использующим данные бортовых датчиков в реальном времени.

Методы, основанные на нейронных сетях, стали важным инструментом для автоматизации процесса создания HD-карт, позволив формировать карты, объединяя данные разной модальности.

Одним из первых подходов стало выполнение семантической сегментации сцены в представлении с высоты птичьего полёта (BEV). На выходе такого рода методы [1, 2, 3, 4] формируют семантическую маску с отсегминтированными краями дороги и разделительными полосами, которая впоследствии преобразуется в векторную карту посредством дополнительной постобработки. Такого рода ограничение привело к созданию монолитных методов [5, 6, 7, 8, 9, 10, 11], которые используют для генерации трансформерные декодеры, позволяющие получать структурированные карты в виде наборов точек напрямую, минуя этап постобработки. Таким образом получилось сильно повысить качество генерации HD-карты.

Для повышения точности и согласованности создаваемых HD-карт в современных подходах активно используется временная составляющая. Анализ временных последовательностей данных позволяет учитывать динамические изменения сцены и устранять артефакты, вызванные разовыми ошибками сенсоров или неточностью предсказания. Такие методы [12, 13, 14, 15] используют временные ряды изображений или последовательности данных различной модальности, что позволяет моделям учитывать взаимосвязь между кадрами. Это способствует более точному получению информации о местности в сложных и изменяющихся условиях.

Целью данного обзора является выбор наиболее перспективного подхода в задаче генерации карты высокой точности. Для достижения поставленной цели были сформулированы и решены следующие задачи:

1. Обзор работ на тему нейросетевых подходов генерации HD-карт местности.

2. Оценка преимуществ и недостатков рассматриваемых методов.

3. Сравнение и выбор перспективного подхода в качестве базового метода.

2 Постановка задачи

High-Definition Map Building (HD Map Building) — задача построения высокоточных карт местности. Основная цель построения HD-карт заключается в создании детализированного и актуального представление окружающей среды, включающее точные геометрические, топологические и семантические характеристики. Эти карты широко применяются в задачах автономного вождения и робототехники. Для обозначения карты местности в момент времени t используется M_t, где M_t представлена в виде множества полилиний $M_t = \{(P^1, c^1), (P^2, c^2), \ldots, (P^k, c^k)\}$, где каждая полилиния P^i задаётся как последовательность координат $P^i = \{p^i_1, p^i_2, \ldots, p^i_{n_i}\}$, $p^i_j \in \mathbb{R}^2$, а c^i — метка класса, характеризующая семантическую информацию (например, край дороги или пешеходный переход). Данные, используемые для построения карты, обозначаются как $D_t = (d^1_t, d^2_t, ..., d^N_t)$, где d^i_t данные, полученные от i-го сенсора (лидара, камеры, GPS и др.) в момент времени t.

2.1 Метрики оценки качества генерации HD-карт

Прежде чем перейти к цели задачи необходимо рассмотреть существующие метрики оценивания качества генерации HD-карт. Ниже представлен обзор основных метрик, используемых для оценки эффективности и точности генерации.

2.1.1 IoU

IoU (Intersection over Union) — это метрика, используемая для оценки качества моделей семантической сегментации в BEV (Bird's Eye View) представлении. Она измеряет степень пересечения между предсказанными и истинными масками для каждого элемента дороги.

$$\text{IoU}(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}, \tag{1}$$

где $D_1, D_2 \subseteq R^{H \times W \times D}$ представляют собой плотное представление фигур кривых, растеризованных на сетке; H и W - высота и ширина сетки, D - количество классов.

Данная метрика используется только для оцененивания методов основанных на семантической сегментации и имеет явный недостаток, поскольку в таком случае расчет метрики производится не для всего метода, а только для вывода получаемого с нейронной сети без последующей постобработки.

2.1.2 Chamfer distance

Chamfer distance — это метрика, используемая для оценки схожести между двумя множествами точек. Эта метрика измеряет среднее расстояние между точками одного множества и ближайшими точками другого множества.

$$\mathrm{CD}_{\mathrm{Dir}}(S_1, S_2) = \frac{1}{S_1} \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2 \tag{2}$$

$$\mathrm{CD}(S_1, S_2) = \mathrm{CD}_{\mathrm{Dir}}(S_1, S_2) + \mathrm{CD}_{\mathrm{Dir}}(S_2, S_1) \tag{3}$$

где CD_{Dir} - это расстояние между направленными скосами, а CD - расстояние между двунаправленными скосами; S_1 и S_2 - это два набора точек на кривых.

Оценка данной метрики используется в качестве порога для расчета средней точности (AP).

2.1.3 AP

AP (Average Precision) — это метрика, используемая для измерения того, насколько точно модель определяет и классифицирует элементы карты. AP рассчитывается как среднее значение точности при различных уровнях полноты:

$$\mathrm{AP} = \frac{1}{10} \sum_{r \in \{0.1, 0.2, \ldots, 1.0\}} \mathrm{AP}_r, \tag{4}$$

где AP_r — это точность при заданном уровне полноты r. Для вычисления AP все предсказания сортируются по убыванию их уверенности. Каждое предсказание классифицируется как истинно положительное (TP) или ложноположительное (FP) в зависимости от значения Chamfer Distance (CD). Если CD меньше порогового значения, предсказание считается истинно положительным, в противном случае — ложноположительным.

На основе всех предсказаний строятся пары точности и полноты, из которых вычисляются AP_r для фиксированных уровней полноты r. Итоговое значение AP представляет собой усреднённый показатель точности, учитывающий как пространственную, так и семантическую точность модели.

2.2 Цель

Цель задачи построения векторной HD-карты можно сформулировать через задачу оптимизации:

$$\mathrm{AP}(M_t, F_\theta(D_t)) \to \max, \tag{5}$$

где AP — метрика Average Precision, а F_θ — метод генерации HD-карты с настраиваемыми параметрами θ.

3 Классификация методов построения HD-карт

Нейросетевые методы построения HD-карт можно классифицировать по нескольким различным критериям:

- *по способу обработки данных* — **унимодальные** или **мультимодальные** методы. Унимодальные методы обрабатывают данные, поступающие только из одного источника (например, только изображения или лидарные данные). Мультимодальные методы напротив используют информацию, полученную сразу от нескольких источников;
- *по способу декодирования данных* — **сверточный декодер** или **трансформерный декодер**. **Сверточные декодеры** используются для семантической сегментации и получения маски сцены, которая в последствии используется для получения HD-карты, а **трансформерные декодеры** напрямую генерируют набор точек относящихся к различными элементам окружения сцены, выдывая тем самым конечный результат в виде HD-карты;
- *по использованию временной информации* — методы, использующие временную информацию, отличаются от статичных тем, что способны учитывать изменения в окружающей среде, например, движение транспортных средств или пешеходов.

3.1 Построение HD-карт на основе семантической сегментации

Подходы, описанные в данном разделе, выполняют семантическую сегментацию над BEV представлением сцены при помощи сверточных декодеров.

3.1.1 HDMapNet, 2021

Пионерской работой в области построения HD-карты является HDMapNet [1], который представляет собой мультимодальный метод семантической сегментации с постпроцессингом. Данный алгоритм использует модальность лидарного облака точек, а также изображения с камер, покрывающие обзор в 360 градусов для последующего получения полного представления о сцене с высоты птичьего полёта.

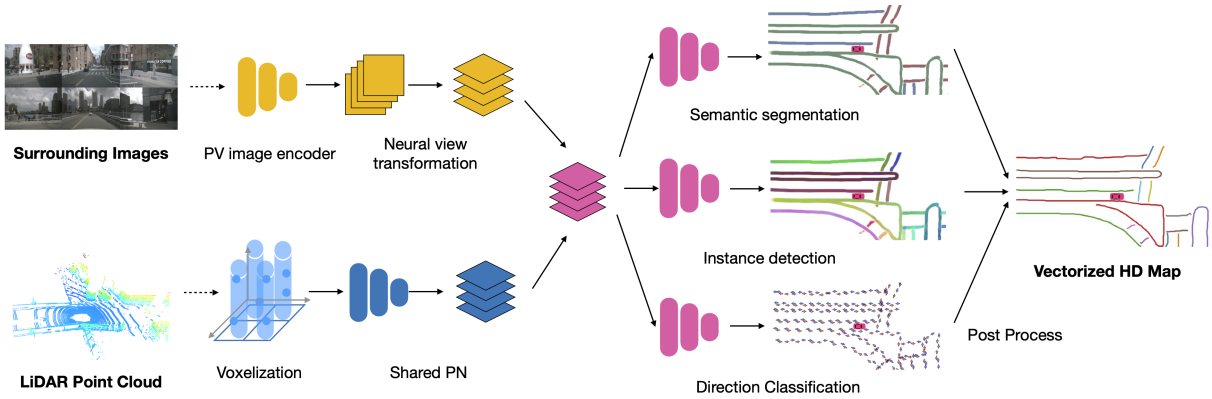

Рис. 1: Принцип работы алгоритма HDMapNet
(Источник: HDMapNet: An Online HD Map Construction and Evaluation Framework [1])

Как можно видеть на Рис.1 для кодирования модальности изображения используется две компоненты. Сначала при помощи кодировщика изображений получаются карты признаков с точки зрения перспективы, после чего происходит их трансформация в систему координат камеры, а затем в вид с высоты птичьего полета BEV_{img}.

$$\mathrm{BEV}_{\mathrm{img}} = \mathrm{I}_{\mathrm{enc}}(Image_{surr}) \tag{6}$$

Также параллельно происходит кодирования лидарного облака точек при помощи метода PointPillars [16], который преобразует трёхмерное облако точек в двухмерное представление. Это достигается за счет разбиения пространства на регулярную сетку (пиллары) и агрегации информации внутри каждого столбца. Затем данные

кодируются с использованием сверточной нейросети и на выходе также получаются BEV признаки BEV_{pc}.

$$\mathrm{BEV}_{\mathrm{pc}} = \mathrm{L}_{\mathrm{enc}}(PointCloud) \tag{7}$$

Полученные признаки BEV с двух кодировщиков объединяются путем конкатенации, после чего проходят процесс сегментации с использованием нескольких сверточных декодеров. На выходе формируется набор различных типов масок, которые с помощью алгоритма постобработки преобразуются в финальную векторизированную HD-карту.

$$\mathrm{Mask}_{\mathrm{sem}} = \mathrm{BEV}_{\mathrm{dec}_{\mathrm{sem}}}([\mathrm{BEV}_{\mathrm{pc}}, \mathrm{BEV}_{\mathrm{img}}]) \tag{8}$$

$$\mathrm{Mask}_{\mathrm{ins}} = \mathrm{BEV}_{\mathrm{dec}_{\mathrm{ins}}}([\mathrm{BEV}_{\mathrm{pc}}, \mathrm{BEV}_{\mathrm{img}}]) \tag{9}$$

$$\mathrm{Mask}_{\mathrm{dir}} = \mathrm{BEV}_{\mathrm{dec}_{\mathrm{dir}}}([\mathrm{BEV}_{\mathrm{pc}}, \mathrm{BEV}_{\mathrm{img}}]) \tag{10}$$

Преимущества

- Простой подход, позволяющий легко производить замену частей сети, например, кодировщика или декодера;
- Качественное отображение изображений из перспективы в BEV представление;
- Хорошее качество предсказания во время плохих погодных условий, а также в ночное время.

Недостатки

- Низкое качество построения HD-карты с нередко возникающими артефактами в виде исчезающих элементов дороги;
- Долгая постобработка сегментационных масок, которая может превышать время работы первичного получения масок с помощью обучаемого модуля;
- Отсутствие обучаемости на этапе постобработки, что делает невозможным автоматическую адаптацию или улучшение этого этапа в процессе обучения модели.

3.1.2 SuperFusion, 2022

Метод SuperFusion [2] объединяет данные с камеры и лидара на трех уровнях: слияние на уровне данных — объединяет информацию о глубине, полученную от лидара, для повышения точности оценки глубины изображения, слияние на уровне объектов — использует перекрестное внимание (Cross-Attention) для прогнозирование BEV признаков с помощью лидара на большом расстоянии с учетом характеристик изображения, и слияние на уровне BEV признаков — производит выравнивание BEV признаков, полученных с ветви лидарного кодировщика и ветви кодирования камеры, после чего объединяет их. Далее, аналогично HDMapNet [1], осуществляется предсказание трёх типов масок, которые используются для построения HD-карты.

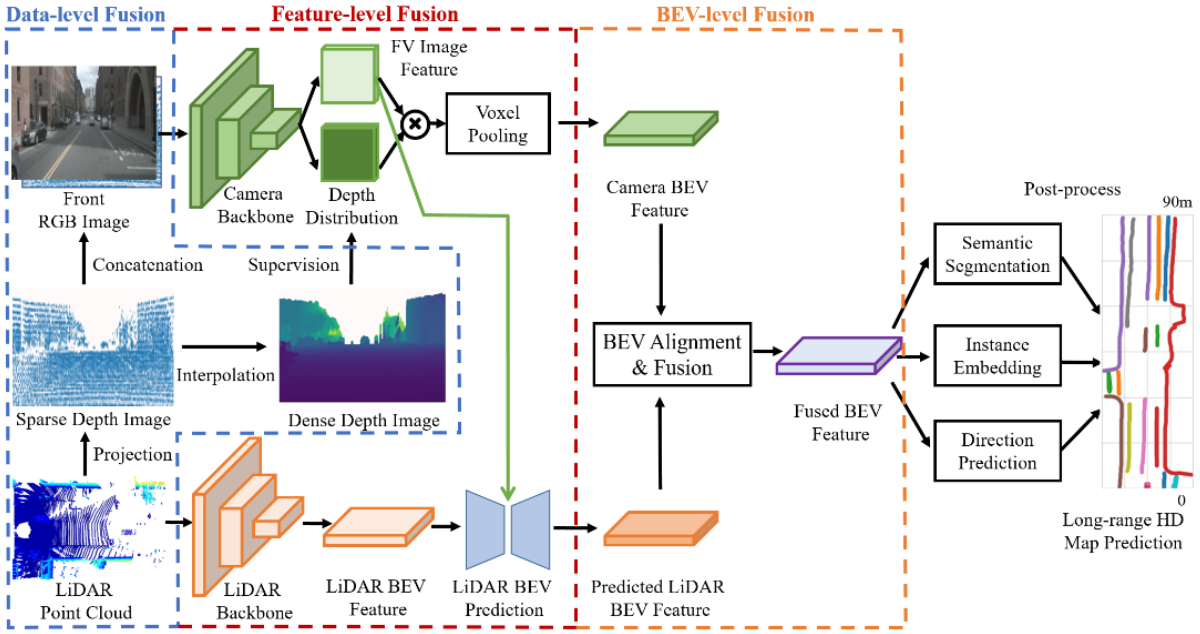

Рис. 2: Принцип работы алгоритма SuperFusion
(Источник: SuperFusion: Multilevel LiDAR-Camera Fusion for Long-Range HD Map Generation [2])

Как можно наблюдать на Рис.2 для получение более полного представления о сцене во время кодирования фронтального изображения применяется проецирование облака точек лидара в вид с камеры. Полученное облако точек затем применяется для создания изображения глубины, а также параллельно конкатенируется с исходным изображением для извлечения полезных признаков. Далее полученные признаки смешиваются с признаками, извлеченными из облака точек при помощи энкодера PointPillars [16], как в HDMapNet [1].

Преимущества

- Дальность предсказания вперед с фронтальной камеры до 90 метров, что может быть полезно для решения задач планирования пути.

Недостатки

- Все те же проблемы, что и у HDMapNet, не считая качества прогнозирования вдаль, но включая время отработки алгоритма постобработки;
- Ограниченный угол предсказания;
- Требуется каллибровка между лидаром и камерой.

3.1.3 Lidar2Map, 2023

Алгоритм Lidar2Map [3] также базируется на HDMapNet [1], но в отличие от предыдущих двух методов Lidar2Map использует модальность лидара и камеры вместе только во время обучения, во время логического вывода модели используется только лишь облако точек лидара, ветвь ответственная за кодирование изображений с камеры замораживается.

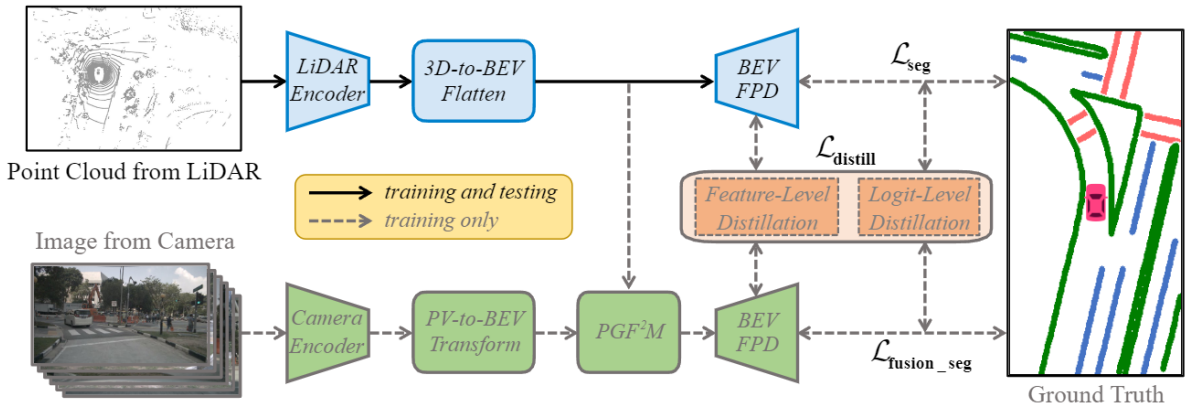

Рис. 3: Принцип работы алгоритма Lidar2Map
(Источник: LiDAR2Map: In Defense of LiDAR-Based Semantic Map Construction Using Online Camera Distillation [3])

Lidar2Map использует сеть на базе лидарного кодировщика в качестве основной ветви для извлечения признаков с помощью пирамидального декодера BEV признаков (BEV-FPD) для построения семантической карты. Во время обучения ветвь, основанная на камере, используется в качестве сети “Учителя” и адаптируется для

извлечения семантических признаков изображения. Выполняется дисстиляция знаний как на уровне признаков, так и на уровне вывода, что позволяет лидарной ветви извлекать выгоду из предоставляемых признаков изображения без дополнительных затрат при выводе.

В отличие от предыдущих методов Lidar2Map использует в качестве кодирования лидарного облака точек метод VoxelNet [17].

Общая функция потерь при обучении выглядит следующим образом:

$$\mathrm{L} = \mathrm{L}_{seg} + \mathrm{L}_{fusion} + \mathrm{L}_{distill}, \tag{11}$$

где $\mathrm{L}_{\mathrm{seg}}$ и $\mathrm{L}_{\mathrm{fusion}}$ это сегментационные функции качества ветви лидара и ветви объединения лидара и камеры соответственно. $\mathrm{L}_{\mathrm{distill}}$ включает в себя $\mathrm{L}_{\mathrm{feature}}$ и $\mathrm{L}_{\mathrm{logit}}$ для дистилляции знаний от ветви камеры к лидарной ветви.

Сегментационная ошибка для генерации семантической карты состоит из двух элементов L_{ce} и L_{ls} и выглядит следующим образом,

$$\mathrm{L}_{\mathrm{seg}} = \mathrm{L}_{\mathrm{ce}} + \mathrm{L}_{\mathrm{ls}}, \tag{12}$$

где L_{ce} — это кросс-энтропия. L_{ls} — это Лосс Ловазье, который используется для максимального увеличения количества пересечений по сравнению с объединением (IoU). Вычисление $\mathrm{L}_{\mathrm{fusion}}$ и $\mathrm{L}_{\mathrm{seg}}$ одинаково.

Преимущества

- Высокое качество семантической сегментации достигнутое при помощи декодера BEV-FPD;
- Высокая скорость работы нейросетевого модуля благодаря использованию лишь лидарной ветви для логического вывода.

Недостатки

- Всё еще большое время работы постобработки в случаях с большим количеством деталей в сцене.

3.1.4 Вывод

Методы, выполняющие семантическую сегментацию для получения HD-карты по средством постобработки, имеют ключевую проблему связанную со скоростью работы методов из-за чего не позволяют использовать их для работы онлайн в реальных условиях, поэтому возможность их применения сильно ограничена при создании системы картографирования, которая должна выстраивать карту местности с нуля или для обновления в реальном времени. Также несмотря на неплохие результаты сегементации полученной методом SuperFusion [2] и Lidar2Map [3] качество результатов построения HD-карты оставляет желать лучшего.

3.2 Генерация HD-карт с использованием трансформеров

Для устранения проблем, связанных с качеством и скоростью постобработки сегментационных методов, были разработаны полностью обучаемые монолитные подходы. Эти методы используют трансформер в качестве декодера, что позволяет значительно повысить эффективность и точность генерации HD-карт. VectorMapNet [18] первая трансформерная модель, которая для генерации элементов карты использует стратегию на основе авторегрессии, сначала при помощи Map Element Detector сеть генериует ключевые элементы и далее на их основе производит авторергессию остальных элементов для получения полилиний с помощью Polyline Generator. Такой подход оказался успешным и на момент июня 2022 года и на недолгий срок позволил VectorMapNet стать SOTA методом. Тем не менее, данный подход не получил широкой популярности, в частности из-за высокой вычислительной сложности алгоритма. Дальнейшее развитие трансформерных методов стало разиваться в сторону детекционного подхода, при котором производится одновременное предсказание всех объектов, как это делает DETR [19] и другие его модификации. Первым методом использующим данную парадигму стал MapTR [5], став SOTA методом на долгое время. Несмотря на это, у подобного подхода имеются свои проблемы связанные с тем, что MapTR не учитывает в полной мере взаимодействие между генерируемыми элементами, отсутствует моделирование сущностной взаимосвязи между экземплярами (элементами карты), поскольку моделирование происходит чисто на уровне отдельных точек. Также не учитываются многомасштабные объекты в BEV признаках и позиционная информация точек, что приводит к их дрожанию и смещению. Для решения данных проблем были разработаны несколько методов, такие как ADMap [8]

и MapQR [7]. Далее рассмотрим их подробнее.

3.2.1 ADMap, 2024

ADMap [8] - это трансформерный метод для создания векторизованных HD-карт, который устраняет проблему искажения топологии карты, вызванную дрожанием точек на карте. Данная проблема решается с помощью трех модулей: многомасштабное восприятие (Multiscale Perception Neck), интерактивное внимание (Instance Interactive Attention) и расчета ошибки при помощи Vector Direction Difference Loss (VDDL).

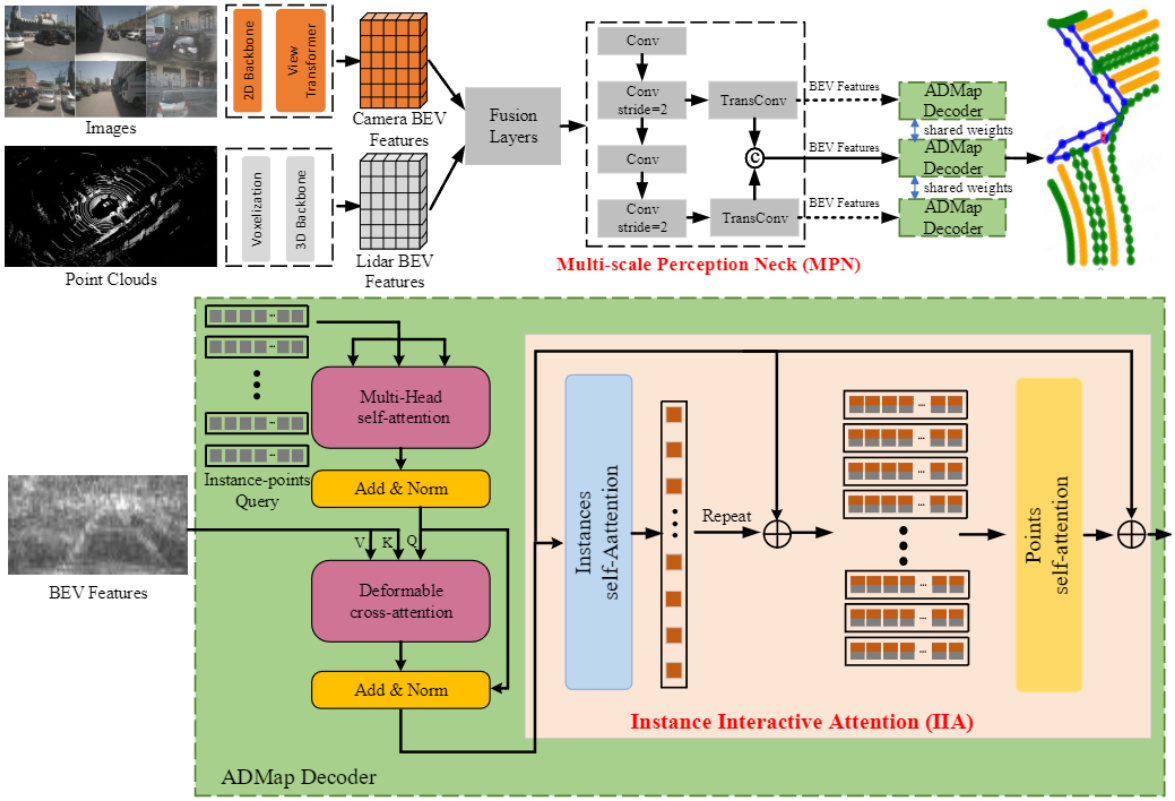

Рис. 4: Принцип работы алгоритма ADMap
(Источник: ADMap: Anti-disturbance Framework for Vectorized HD Map Construction [8])

Multiscale Perception Neck (MPN). Для отображения многомасштабных объектов на BEV признаках используется Multiscale Perception Neck. Во время обучения многомасштабные объекты BEV и карты признаков BEV на каждом уровне передаются в декодер ADMap, который позволяет сети прогнозировать информацию об экземплярах сцены в разных масштабах. В процессе логического вывода MPN вы-

водит только многомасштабные признаки BEV.

Instance Interactive Attention. Чтобы лучше изучить особенности каждого экземпляра на этапе декодирования, используется интерактивное внимание экземпляра (IIA), которое включает в себя Instance Self-Attention и Point Self-Attention. При использовании IIA для извлечения векторных представлений, соответствующих точкам, взаимодействия объектов между векторными представлениями экземпляров помогают сети изучать топологические отношения между точками экземпляра.

Vector Direction Difference Loss. При помощи VDDL анализируется направление векторов, соединяющих последовательные точки внутри объекта (экземпляра). Разница между предсказанным направлением вектора и реальным направлением используется как метрика, позволяющая точнее обучать сеть предсказывать позиции точек в последовательности. Если разница между реальным и предсказанным направлением вектора для какой-то точки оказывается значительной, это может указывать на резкие изменения в топологии объекта или сцены (например, на границы объектов, изгибы или другие сложные участки). Такие точки сложнее предсказать, поскольку они соответствуют менее предсказуемым участкам. Поскольку резкие изменения в топологии требуют большей точности, модель уделяет этим точкам повышенное внимание, чтобы точнее предсказать их положение и корректно отразить структуру объекта.

Преимущества

- SOTA качество генерации HD-карты на момент публикации (начало 2024 года);
- Решает проблему дрожащих точек;
- End-to-End подход;
- Мультимодальный подход, что нетипично для трансформерных методов в задаче генерации HD-карт;
- Высокая скорость работы.

Недостатки

- Не учитывается позиционная информация о точках.

3.2.2 MapQR, 2024

MapQR [7] еще один трансформерный метод для генерации HD-карт, который решает проблему отсутствия учета контекстной информации при генерации точек. Кроме того, данный подход устраняет проблему случайной инициализации точек, обеспечивая их привязку к конкретному местоположению.

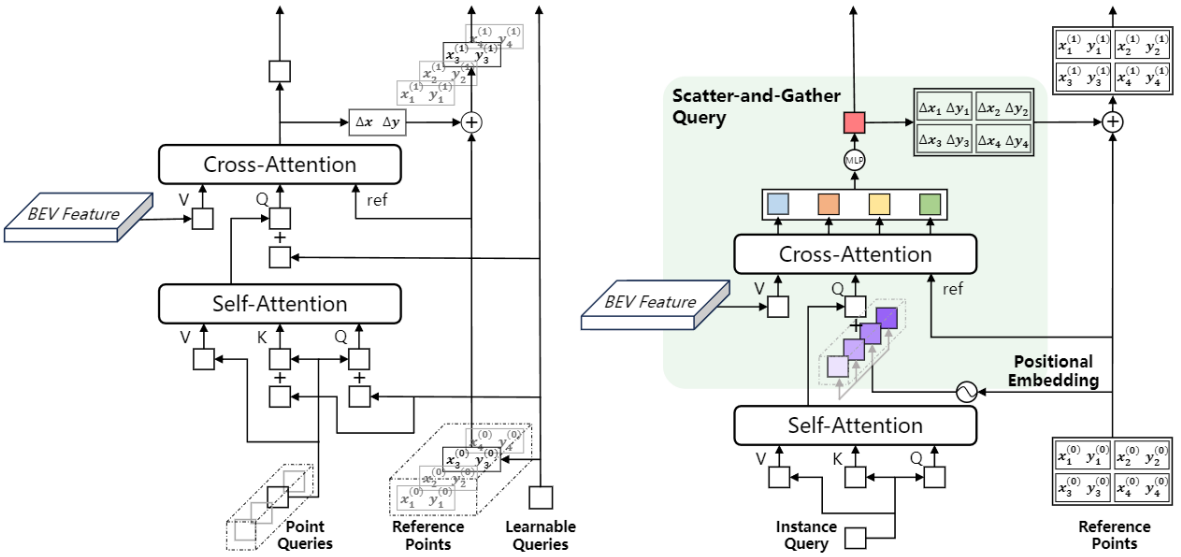

Рис. 5: Сравнение декодеров. Слева MapTR, справа MapQR
(Источник: Leveraging Enhanced Queries of Point Sets for Vectorized Map Construction [7])

Scatter-and-Gather Query. В MapQR используется только один тип запроса в качестве входных данных для декодера. В отличие от MapTR [5] в MapQR происходит моделирование на уровне сущностей, а не точек. Для чего используется Instance Query, который моделирует запрос к каждому элементу карты. После прохождения Self-Attention каждый запрос разделяется на подзапросы, которые используются для моделирования положения отдельных точек (scatter). Для этого применяется позиционное кодирование, основанное на референсных точках. Затем эти подзапросы передаются в Cross-Attention, где Attention Score вычисляется с использованием признаков в представлении Bird's Eye View, полученных от кодировщика. После этапа Cross-Attention подзапросы объединяются (gather) с помощью MLP, и итоговый вывод используется для уточнения положения референсных точек.

Преимущества

- Высокие результаты генерации HD-карты;

- Решает проблему дрожащих точек, а также используется информацию о местоположении точек при инициализации запросов;
- End-to-End подход;

Недостатки

- Отсутствие мультимодальности, данный метод использует только лишь изображения с камер.

3.2.3 Вывод

Генерация HD-карт с использованием трансформеров представляет собой значительный шаг вперед в создании более точных и эффективных карт для автономных транспортных средств. Благодаря полной обучаемости методов открываются новые горизонты для создания более точных и стабильных методово генерации HD-карт.

3.3 Генерация HD-карт с учетом временной информации

Описанные до этого моменты подходы ограничены использованием моментальной информации. Такой подход ограничивает стабильность рассмотренных методов и их производительность в сложных сценариях, когда участки сцены становятся частично или полностью невидимыми для сенсоров автономного транспортного средства, в основном из-за отсутствия временной информации. Более того, их производительность снижается при применении к более широким диапазонам восприятия. Включение временных данных, как в StreamMapNet [13], MapTracker [12] или GlobalMapNet [14], которые позволяют лучше справляться с заслонениями, так как они могут восстанавливать недостающие элементы на основе контекста из прошлых и будущих кадров.

3.3.1 StreamMapNet, 2023

StreamMapNet [13] обрабатывает последовательности синхронизированных изображений, снятых с нескольких камер автономных транспортных средств. Эти изображения используются для создания локальных HD-карт, представленных в виде набора векторных объектов (instances). Ключевым компонентом метода является буфер памяти, который хранит сгенерированные скрытые состояния (hidden states), включающие в себя BEV признаки и обновленные запросы объектов (object queries).

Таким образом буфер позволяет сохранять временную информацию и учитывать контекст предыдущих кадров.

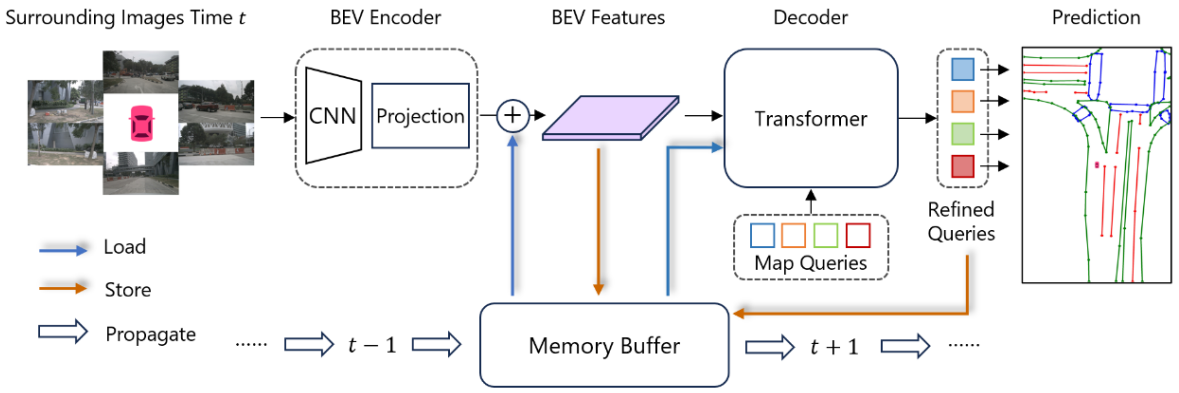

Рис. 6: Принцип работы алгоритма StreamMapNet
(Источник: StreamMapNet: Streaming Mapping Network for Vectorized Online HD Map Construction [13])

Streaming Strategy. Важной особенностью данного метода является подход к обработке временных данных. Модель использует потоковую стратегию (streaming strategy) для объединения информации во времени. В отличие от стратегии объединения кадров (stacking strategy), при которой несколько кадров обрабатываются одновременно как единый объем данных, что приводит к росту вычислительных затрат линейно с увеличением числа кадров, потоковая стратегия обрабатывает каждый кадр отдельно. Информация о предыдущих кадрах передается через скрытые состояния (hidden states), что позволяет создавать долгосрочные временные связи, поскольку скрытые состояния аккумулируют всю историческую информацию, а также снижать затраты на память и уменьшать задержку обработки по сравнению с объединением кадров. Таким образом, потоковая стратегия обеспечивает более эффективную и устойчивую обработку данных во времени. Для решения данной проблемы можно использовать Multi-Point Attention, у которого один запрос связан сразу с несколькими точками, которые совместно покрывают область объекта в BEV-пространстве, что позволяет модели извлекать признаки из более широкой области, учитывая форму и масштаб объекта.

Query Propagation. Query Propagation — это механизм, который сохраняет и переносит информацию о ключевых объектах между последовательными кадрами. В процессе Query Propagation модель оценивает объекты на текущем кадре и выби-

рает те, которые имеют наивысшие значения confidence scores (уровня уверенности). Эти запросы содержат информацию о положении, форме и классе объекта. Отобранные запросы передаются в следующий кадр в виде позиционных приоритетов (positional priors), сохраняя временные признаки (temporal features), накопленные за предыдущие кадры. Поскольку модель работает в системе координат, привязанной к собственному положению машины, необходимо корректировать положения объектов при передаче. Для этого используется MLP, который выполняет трансформацию запросов в скрытом пространстве признаков. Для повышения стабильности эта трансформация дополняется остаточной связью (residual connection), что позволяет сохранить информацию из предыдущих шагов.

Multi-Point Attention. Существующие методы генерации HD-карт в своей основе используют Deformable DETR [20], который назначает каждой точке запроса (query) так называемую референсную точку, как мы это увидели например MapTR [5] и MapQR [7]. Референсная точка играет роль якоря (anchor), вокруг которого извлекаются признаки из BEV-представления. В каждом слое трансформера референсная точка уточняется. Такой подход с одной референсной точкой может быть недостаточно гибким для сложных объектов или ситуаций, где полилинии или границы объекта занимают обширные области.

Преимущества

- Использование временного буфера для предсказания i-го шага;
- Сниженные вычислительные затраты благодаря использованию Streaming Strategy;
- Хорошая скорость работы алгоритма, при учете использования буферной информации.

Недостатки

- Отсутствие мультимодальности, данный метод использует только лишь изображения с камер;
- Несмотря на преимущества, передача информации через скрытые состояния может быть чувствительна к ошибкам в предыдущих кадрах, если скрытые состояния накапливают ошибки, это может повлиять на точность предсказаний в будущих кадрах;

- Проблема точности локализации автомобиля и собственного положения машины. Когда модель обрабатывает кадры по очереди, она сохраняет информацию о расположении объектов в скрытых состояниях. Чтобы информация, полученная из предыдущих кадров, была актуальной в текущем кадре, требуется точно преобразовать координаты объектов из предыдущих кадров в текущие, с учетом перемещения машины. Если глобальная локализация неверна, например, из-за ошибки в GPS или другой системе локализации, то система координат автомобиля будет определена неправильно. Это приведет к тому, что объекты будут локализованы в неправильных местах, и трансформация запросов из предыдущих кадров, которая основывается на этих данных, будет ошибочной.

3.3.2 MapTracker, 2024

Алгоритм MapTracker [12] представляет собой подход для онлайн-генерации HD-карт с использованием двух типов памяти: BEV-памяти и VEC-памяти (Vector), которые аккумулируют информацию из сенсорного потока. Эти памяти сохраняют важные данные о сцена в разных представлениях, что помогает обеспечивать стабильность и точность картирования в условиях движущегося автомобиля.

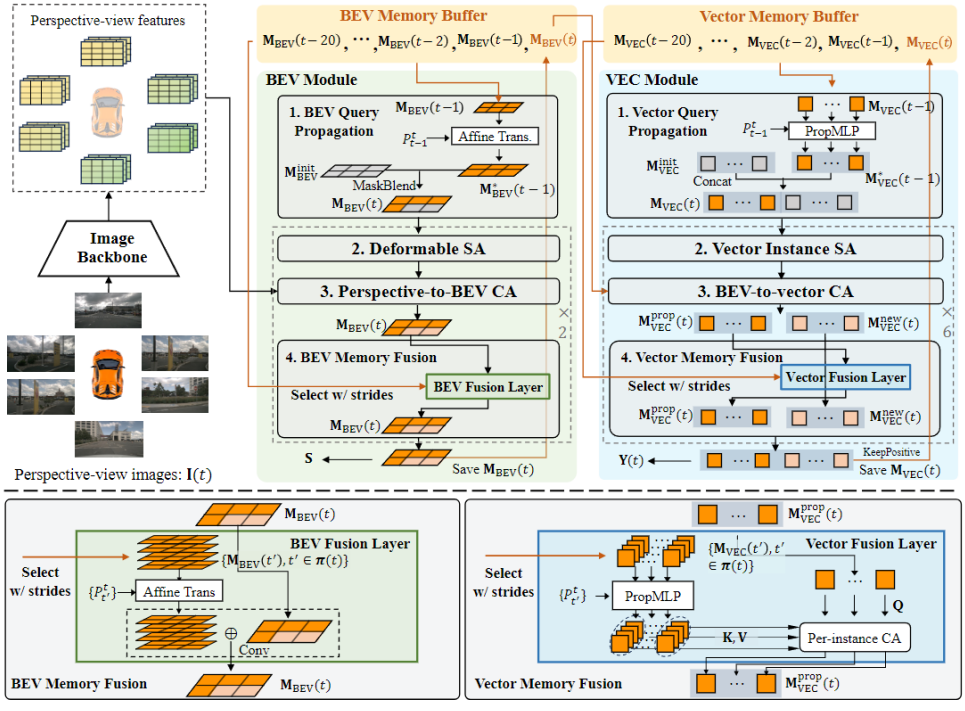

Рис. 7: Принцип работы алгоритма MapTracker
(Источник: MapTracker: Tracking with Strided Memory Fusion for Consistent Vector HD Mapping [12])

Bird's-eye-view memory. BEV-память - это представление сцены в виде латентного изображения с точки зрения сверху, которое отображает область вокруг автомобиля. BEV-память сохраняет контекст на несколько кадров, что позволяет следить за изменяющимися объектами на дорогах в реальном времени.

Vector memory. VEC-память - это представление объектов, таких как пешеходные переходы, разделители полос и дорожные границы, в виде латентных векторов. Каждый элемент дороги (например, разделительная полоса) сохраняется как векторное представление, что позволяет легко отслеживать изменения в дорожной инфраструктуре.

MapTracker сохраняет память из нескольких предыдущих кадров. Это предотвращает потерю данных, которая может происходить при использовании только текущей информации, особенно в сложных, загруженных окружениях (например, в местах с большим количеством машин, что затрудняет видимость дорожных объектов).

В MapTracker, аналогично StreamMapNet [13], используется механизм Query Propagation для ассоциации отслеживаемых элементов дороги между кадрами. То есть, запросы, полученные в одном кадре, передаются в следующий, чтобы продолжить отслеживание тех же объектов. Это помогает поддерживать согласованность дорожных элементов на протяжении всей последовательности кадров, таким образом для качественного склеивания кадров важно получать качественную локализацию автомобиля, иначе могут возникнуть большие проблемы с качеством предсказания карты местности.

Преимущества

- MapTracker сохраняет память из нескольких предыдущих кадров (до 1000 кадров) для более точного предсказания HD-карты в данный момент;

Недостатки

- Отсутствие мультимодальности, данный метод использует только лишь изображения с камер;
- Очень низкая скорость работы алгоритма;
- Высокие затраты на потребление памяти из-за очень большого размера памяти в 1000 кадров, а также хранения как BEV, так и латентных представлений.

3.3.3 Вывод

Описание методов StreamMapNet и MapTracker показывает, как интеграция временных данных позволяют улучшить качество карт, обеспечивая стабильность дорожных элементов на протяжении нескольких кадров. Основным преимуществом этих методов является использование механизмов передачи запросов (Query Propagation) между кадрами, что улучшает согласованность и точность картирования.

Однако методы также имеют свои ограничения, такие как отсутствие мультимодальности (работают только с изображениями), высокие вычислительные затраты, низкая скорость работы и зависимость от точности локализации автомобиля. В частности, ошибки в глобальной локализации могут существенно повлиять на качество картирования, поскольку трансформация запросов из предыдущих кадров будет ошибочной. Эти проблемы требуют дальнейших исследований и улучшений, чтобы сделать алгоритмы более эффективными и подходящими для реальных условий эксплуатации автономных транспортных средств.

4 Заключение

Method	Modality	Temporal info	Loc. Agnostic	Real-time	End-to-End	Year
HDMapNet [1]	L+C	-	+	-	-	2021
SuperFusion [2]	L+C	-	+	-	-	2022
VectorMapNet [18]	L+C	-	+	-	+	2023
Lidar2Map [3]	L	-	+	-	-	2023
MapTR [5]	L+C	-	+	+	+	2023
StreamMapNet [13]	C	+	-	+	+	2023
ADMap [8]	L+C	-	+	+	+	2024
MapQR [7]	C	-	+	+	+	2024
MapTracker [12]	C	+	-	-	+	2024

Таблица 1: Сравнительная таблица алгоритмов построения HD-карт

Современные методы генерации HD-карт активно развиваются, стремясь к созданию более точных, полных и надежных карт для автономных систем. Изначальные подходы, основанные на семантической сегментации с последующей постобра-

боткой, оказались ограничены из-за проблем со скоростью и качеством, что привело к разработке более эффективных трансформерных моделей, способных генерировать карты End-to-End. Такие модели как ADMap и MapQR, значительно улучшили качество и точность карт за счет механизмов внимания и учета позиционной информации. Дальнейшим шагом в развитии стали методы, интегрирующие временную информацию, такие как StreamMapNet и MapTracker, которые используют буферы памяти для отслеживания изменений во времени, что способствует повышению стабильности, но не несмотря на это, существуют и проблемы, такие как отсутствие мультимодальности, высокая вычислительная сложность и зависимость от точности локализации. Наилучшим вариантом в качестве базового метода является MapQR [7], несмотря на отсуствие мултимодальности данный трансформерный метод имеет высокое качество предсказания карты местности, а также высокий потенциал улучшения при помощи добавления дополнительной модальности лидара и использования временной информации.

Список литературы

1. HDMapNet: An Online HD Map Construction and Evaluation Framework / Q. Li, Y. Wang, Y. Wang, H. Zhao. — 2021. — arXiv: 2107.06307 [cs.CV].
2. SuperFusion: Multilevel LiDAR-Camera Fusion for Long-Range HD Map Generation / H. Dong, X. Zhang, J. Xu, R. Ai, W. Gu, H. Lu, J. Kannala, X. Chen // arXiv preprint arXiv:2211.15656. — 2022.
3. LiDAR2Map: In Defense of LiDAR-Based Semantic Map Construction Using Online Camera Distillation / S. Wang, W. Li, W. Liu, X. Liu, J. Zhu // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2023. — С. 5186—5195.
4. *Zhou B.*, *Krähenbühl P.* Cross-view Transformers for Real-time Map-view Semantic Segmentation //. — 2022. — С. 13760—13769.
5. MapTR: Structured Modeling and Learning for Online Vectorized HD Map Construction / B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, C. Huang //. — 2022.
6. MapTRv2: An End-to-End Framework for Online Vectorized HD Map Construction / B. Liao, S. Chen, Y. Zhang, B. Jiang, Q. Zhang, W. Liu, C. Huang, X. Wang // arXiv preprint arXiv:2308.05736. — 2023.
7. Leveraging Enhanced Queries of Point Sets for Vectorized Map Construction / Z. Liu, X. Zhang, G. Liu, J. Zhao, N. Xu // European Conference on Computer Vision. — 2024.
8. ADMap: Anti-disturbance framework for reconstructing online vectorized HD map / H. Hu, F. Wang, Y. Wang, L. Hu, J. Xu, Z. Zhang. — 2024. — arXiv: 2401.13172 [cs.CV]. — URL: https://arxiv.org/abs/2401.13172.
9. MGMap: Mask-Guided Learning for Online Vectorized HD Map Construction / X. Liu, S. Wang, W. Li, R. Yang, J. Chen, J. Zhu //. — 2024. — С. 14812—14821.
10. PivotNet: Vectorized Pivot Learning for End-to-end HD Map Construction / W. Ding, L. Qiao, X. Qiu, C. Zhang //. — 2023. — С. 3672—3682.
11. VectorMapNet: End-to-end Vectorized HD Map Learning / Y. Liu, T. Yuan, Y. Wang, Y. Wang, H. Zhao // ICML. — 2023. — С. 22352—22369.

12. MapTracker: Tracking with Strided Memory Fusion for Consistent Vector HD Mapping / J. Chen, Y. Wu, J. Tan, H. Ma, Y. Furukawa. — 2024. — arXiv: 2403.15951 [cs.CV]. — URL: https://arxiv.org/abs/2403.15951.

13. StreamMapNet: Streaming Mapping Network for Vectorized Online HD Map Construction / T. Yuan, Y. Liu, Y. Wang, Y. Wang, H. Zhao. — 2023. — arXiv: 2308.12570 [cs.CV]. — URL: https://arxiv.org/abs/2308.12570.

14. GlobalMapNet: An Online Framework for Vectorized Global HD Map Construction / A. Shi, Y. Cai, X. Chen, J. Pu, Z. Fu, H. Lu. — 2024. — arXiv: 2409.10063 [cs.CV]. — URL: https://arxiv.org/abs/2409.10063.

15. Neural Map Prior for Autonomous Driving / X. Xiong, Y. Liu, T. Yuan, Y. Wang, Y. Wang, H. Zhao. — 2023. — arXiv: 2304.08481 [cs.CV]. — URL: https://arxiv.org/abs/2304.08481.

16. PointPillars: Fast Encoders for Object Detection from Point Clouds / A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom. — 2019. — arXiv: 1812.05784 [cs.LG]. — URL: https://arxiv.org/abs/1812.05784.

17. *Zhou Y.*, *Tuzel O.* Voxelnet: End-to-end learning for point cloud based 3d object detection // CVPR. — 2018. — C. 4490—4499.

18. VectorMapNet: End-to-end Vectorized HD Map Learning / Y. Liu, T. Yuan, Y. Wang, Y. Wang, H. Zhao // International conference on machine learning. — PMLR. 2023.

19. End-to-End Object Detection with Transformers / N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko. — 2020. — arXiv: 2005.12872 [cs.CV]. — URL: https://arxiv.org/abs/2005.12872.

20. Deformable DETR: Deformable Transformers for End-to-End Object Detection / X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai. — 2021. — arXiv: 2010.04159 [cs.CV]. — URL: https://arxiv.org/abs/2010.04159.