Нейросетевые методы генерации HD-карт местности

Иванов Л.А.
ivanov.la@phystech.edu

Центр когнитивного моделирования МФТИ

2025 – Научно-исследовательский семинар



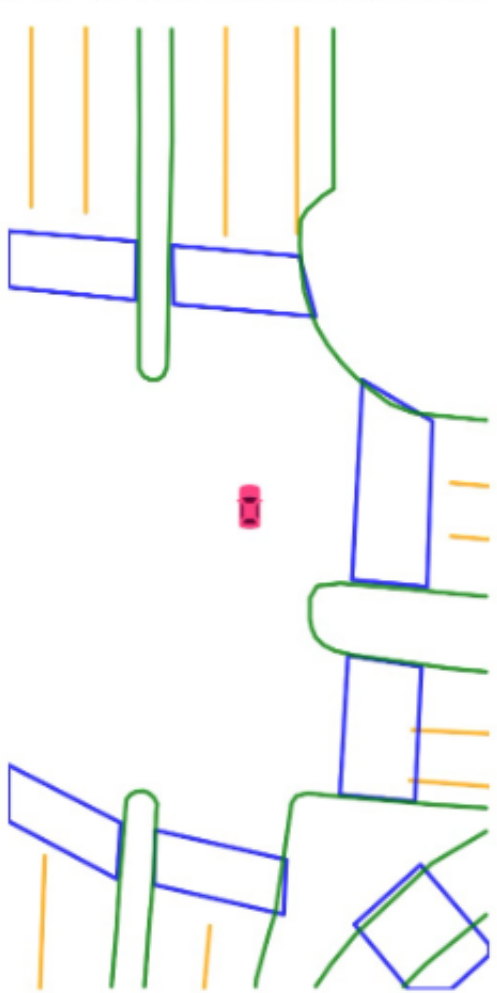


Мотивация

Карты высокой точности (HD Maps) являются одним из важнейших компонентов технологии автономного вождения при решении ряда задач:

- Навигации транспортного средства
- Прогнозировании поведения других участников движения
- Планировании собственного движения беспилотного транспорта

Цель работы Разработать эффективный алгоритм генерации HD-карты местности.



Метрики

- *Chamfer Distance* - метрика оценки схожести между двумя множествами точек, измеряющая среднее расстояние между точками одного множества и ближайшими точками другого множества.

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

$$\mathrm{CD}(S_1, S_2) = \mathrm{CD}_{\mathrm{Dir}}(S_1, S_2) + \mathrm{CD}_{\mathrm{Dir}}(S_2, S_1) \tag{2}$$

- *AP (Average Precision)* — это метрика измерения того, насколько точно модель определяет и классифицирует элементы карты.

$$\mathrm{AP} = \frac{1}{11} \sum_{r \in \{0.2, 0.5, \ldots, 1.5\}} \mathrm{AP}_r \tag{3}$$

Каждое предсказание классифицируется как истинно положительное (TP) или ложноположительное (FP) в зависимости от значения *Chamfer Distance*.

Постановка задачи

Входные данные: $D_t = (d_t^1, d_t^2, ..., d_t^N)$, где d_t^i — данные, полученные от i-го сенсора (лидара, камеры, GPS и др.) в момент времени t.
Выходные данные: карта $M_t = \{(P^1, c^1), (P^2, c^2), \ldots, (P^k, c^k)\}$, где каждая полилиния P^i задаётся как последовательность координат $P^i = \{p_1^i, p_2^i, \ldots, p_{n_i}^i\}$, $p_j^i \in \mathbb{R}^2$, а c^i — метка класса.
Нужно: решить следующую задачу оптимизации

$$\text{AP}(M_t, F_\theta(D_t)) \to \max,$$

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

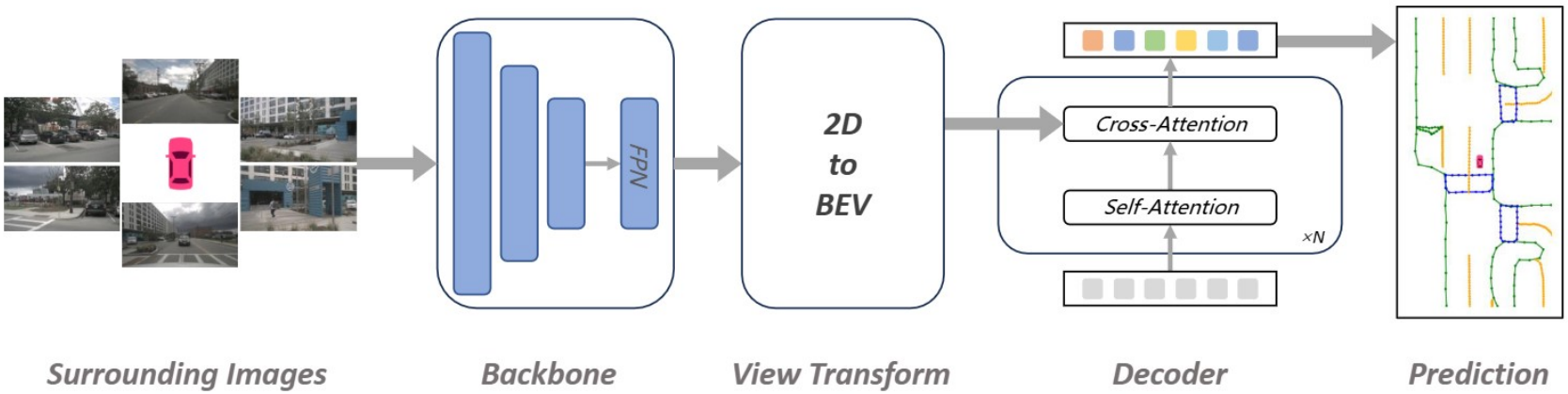
Существующие подходы

Нейросетевые методы построения HD-карт можно разбить на 3 группы:

- **сегментационные с пост обработкой** — HDMapNet, Lidar2Map;
- **трансформерные (end2end)** — VectorMapNet, MapTR, MapTRv2, PivotNet, GeMap, MapQR;
- **методы построения глобальной карты** — StreamMapNet, GlobalMapNet, MapTracker.

MapQR, 2024

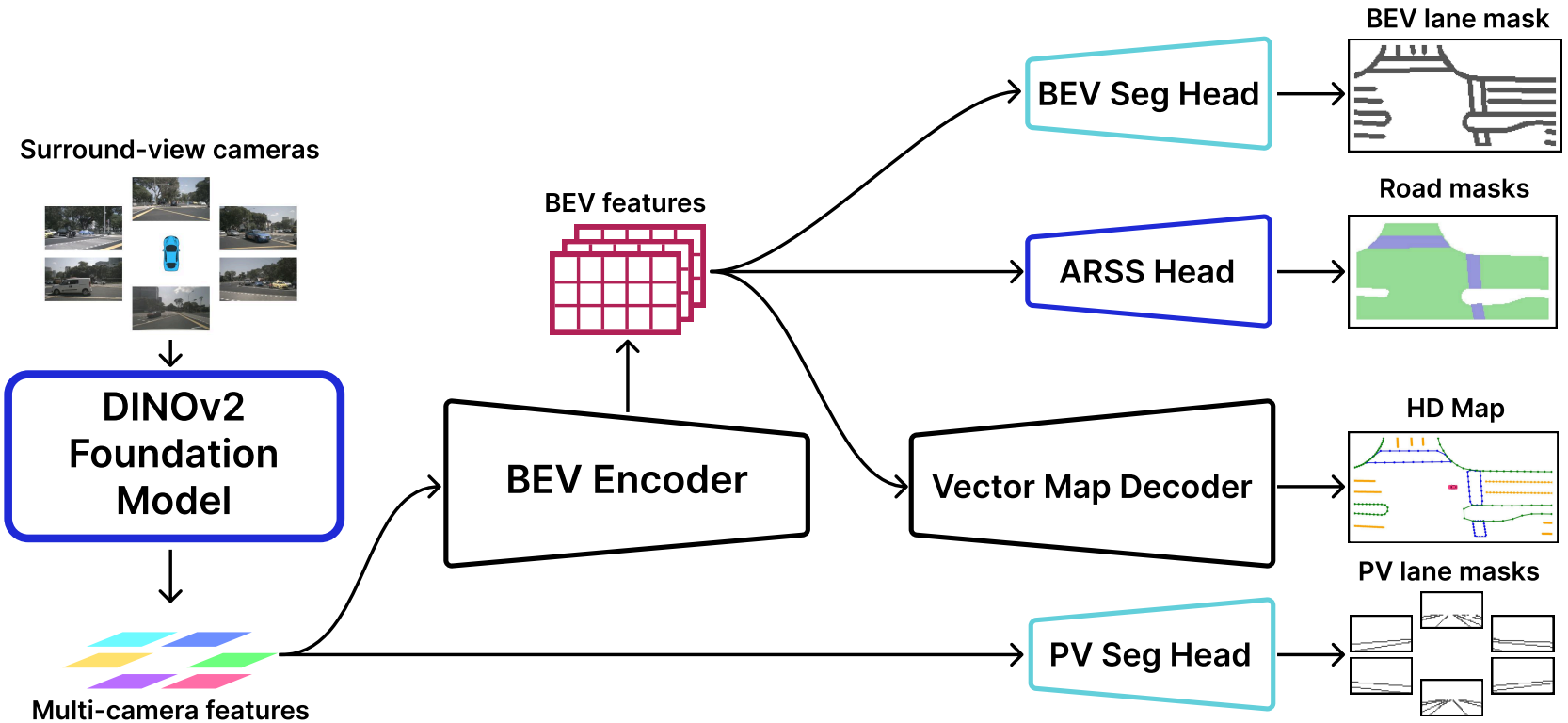
В MapQR[1] — это трансформерный метод генерации HD-карт, который решает проблему отсутствия учета контекстной информации при генерации точек. Данный подход устраняет проблему случайной инициализации точек, обеспечивая их привязку к конкретному местоположению.



[1] Zihao Liu и др. (2024). «Leveraging Enhanced Queries of Point Sets for Vectorized Map Construction». В: *European Conference on Computer Vision*.

MapFM

Была добавлена вспомогательная сегментационная голова ARSS для улучшения контекста BEV признаков, а также использована фундаментальная модель DINOv2 для излвченения признаков из изображений.



MapFM: Сравнение с SOTA методами

MapFM с DINOv2-small имеет прирост 1.5% mAP, с DINOv2-base 2.7% mAP при 16 эпохах обучения.

Method	**Backbone**	**Epochs**	AP_{div}	AP_{ped}	AP_{bound}	**mAP**
MapTR	ResNet50	24	51.5	46.3	53.1	50.3
MapTR	SwinT	24	45.2	52.7	52.3	50.1
BeMapNet	ResNet50	30	62.3	57.7	59.4	59.8
BeMapNet	SwinT	30	64.4	61.2	61.7	62.4
PivotNet	ResNet50	24	56.5	56.2	60.1	57.6
PivotNet	SwinT	24	60.6	59.2	62.2	60.6
MapTRv2	ResNet50	24	62.4	59.8	62.4	61.5
MGMapNet	ResNet50	24	66.1	64.7	69.4	66.8
MapQR (baseline)	ResNet50	24	68.0	63.4	67.3	66.3
MapQR (baseline)	SwinT	24	68.1	63.1	67.1	66.1
(ours) MapFM	DINOv2-small	24	**68.8**	**65.7**	**68.9**	**67.8**
(ours) MapFM	DINOv2-base	16	**68.8**	**67.8**	**71.3**	**69.0**

MapQR + Вспомогательная сегментационная голова

Дополнительная голова дала относительный прирост в 1% по отношению к **MapQR+SwinT** и 0,8% к **MapQR+ResNet50**.

Method	**Backbone**	AP_{div}	AP_{ped}	AP_{bound}	mAP
MapTR	ResNet50	51.5	46.3	53.1	50.3
MapTR	SwinT	45.2	52.7	52.3	50.1
PivotNet	ResNet50	56.5	56.2	60.1	57.6
PivotNet	SwinT	60.6	59.2	62.2	60.6
MapTRv2	ResNet50	62.4	59.8	62.4	61.5
MapQR	ResNet50	68.0	63.4	67.3	66.3
MapQR	SwinT	68.1	63.1	67.1	66.1
MapQR + ARSS	ResNet50	**68.8**	**64.9**	**67.6**	**66.7**
MapQR + ARSS	SwinT	68.1	**64.1**	**68.0**	**67.1**

MapFM: Сравнение backbone

Center for Cognitive Modeling

Резлультаты добавления фундаментальной модели DINOv2 и вспомогательной сегментационной головы.

Backbone	AP_{div}	AP_{ped}	AP_{bound}	mAP
ResNet50	68.8	64.9	67.6	66.7
SwinT	68.1	64.1	68.0	67.1
DINOv2-small	**68.8**	**65.7**	**68.9**	**67.8**
DINOv2-base	**68.8**	**67.8**	**71.0**	**69.3**

MapFM: агрегация DINO-признаков

Center for Cognitive Modeling

Также было проведено сравнение различных стратегий агрегации признаков с разных слоёв фундаментальной модели.

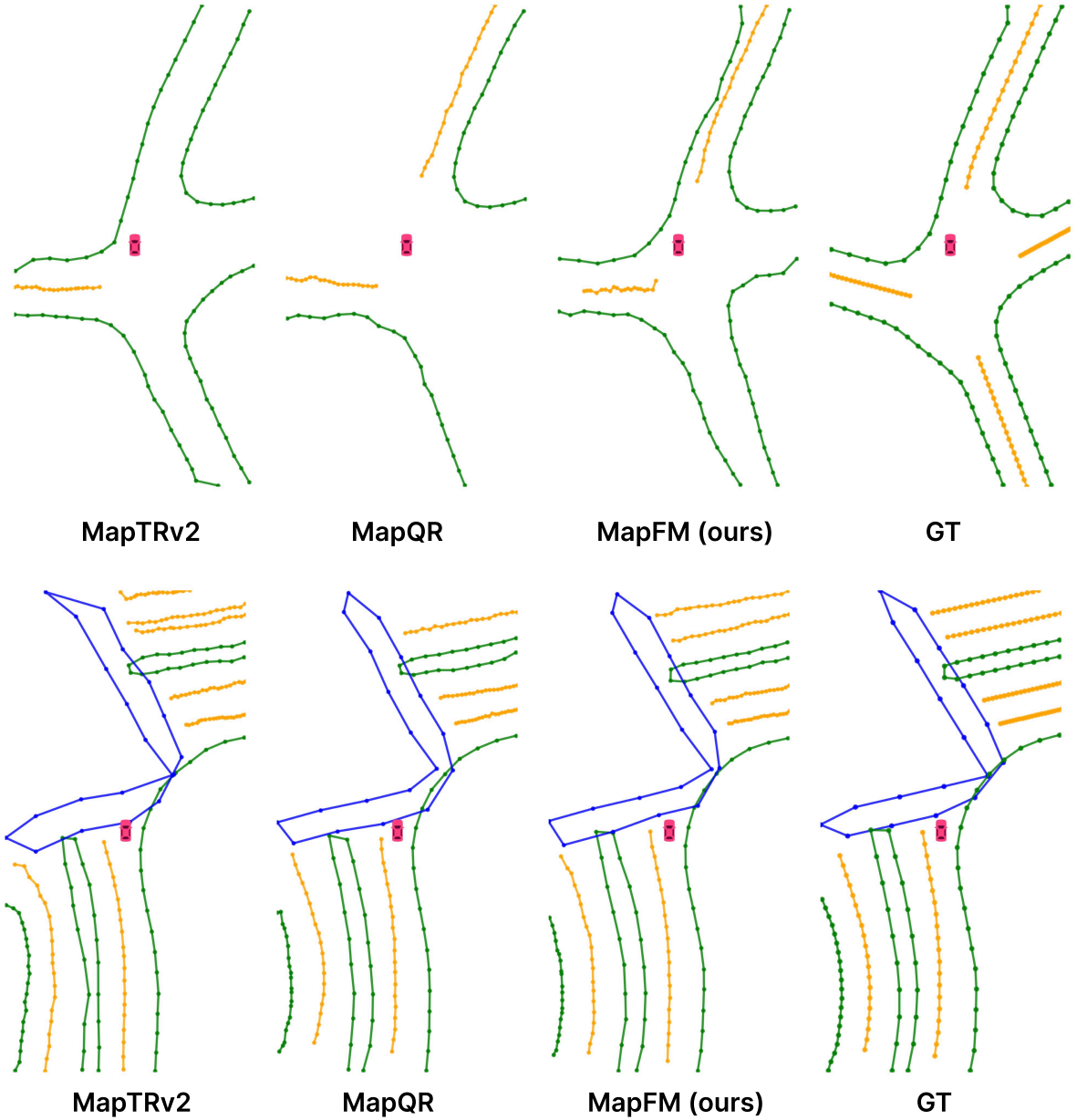
Feature Aggregation Strategy	**Backbone**	**Epochs**	**mAP**
Multi-layer CNN (4, 8, 12)	DINOv2-small	14	62.6
Feature concatenation (4, 8, 12)	DINOv2-small	14	63.9
Last Layer Features (12)	DINOv2-small	14	**64.1**

MapFM: анализ добавлениея ARSS головы

Совместное использование фундаментальной модели с вспомогательной сегментационной головы показывает явный прирост в качестве.

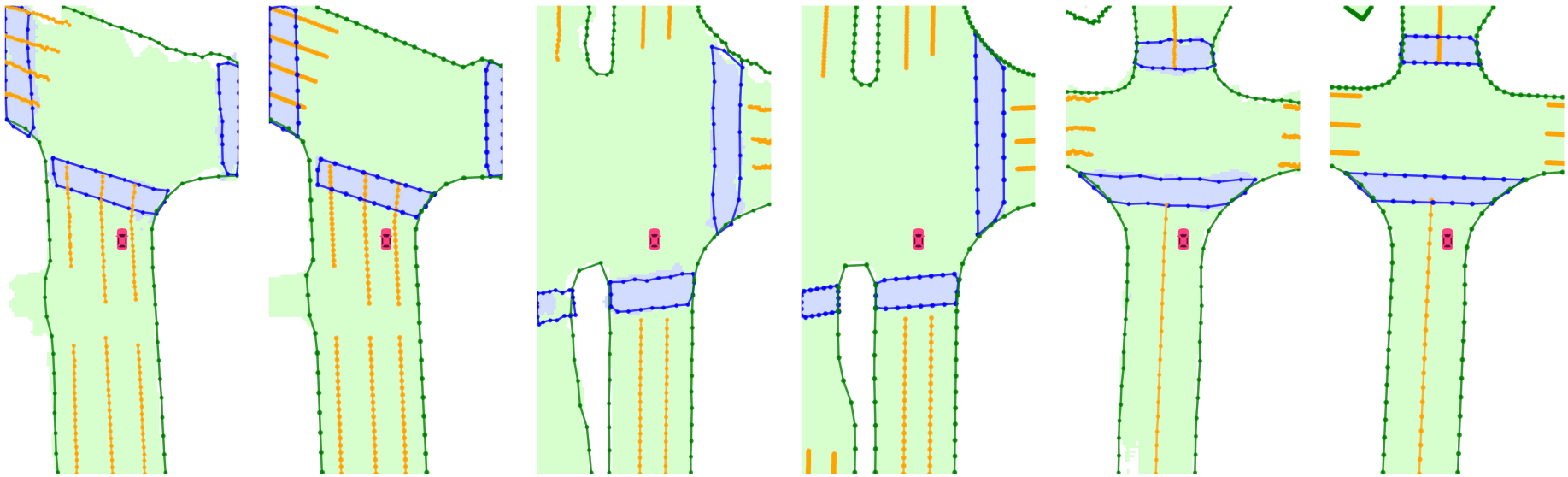
Config	**Epochs**	AP_{div}	AP_{ped}	AP_{bound}	**mAP**
MapFM w/o ARSS Head	14	62.2	60.8	66.1	63.0
MapFM w/ ARSS Head	14	**63.7**	**62.3**	**66.4**	**64.1**

Сравнение результатов генерации HD-карты



Результаты работы MapFM

Пример результата инференса метода MapFM с наложением HD-карты на сегментационные маски проезжей части.



Публикация результатов

Результаты исследования были оформлены и поданы на конференцию HAIS 2025.

MapFM: Foundation Model-Driven HD Mapping with Multi-Task Contextual Learning

Leonid Ivanov[1], Vasily Yuryev[1], and Dmitry Yudin[1,2]

[1] Moscow Institute of Physics and Technology, 141701, Dolgoprudny, Russia
[2] AIRI, 121170 Moscow, Russia

Abstract. In autonomous driving, high-definition (HD) maps and semantic maps in bird's-eye view (BEV) are essential for accurate localization, planning, and decision-making. This paper introduces an enhanced End-to-End model named MapFM for online vectorized HD map generation. We show significantly boost feature representation quality by incorporating powerful foundation model for encoding camera images. To further enrich the model's understanding of the environment and improve prediction quality, we integrate auxiliary prediction heads for semantic segmentation in the BEV representation. This multi-task learning approach provides richer contextual supervision, leading to a more comprehensive scene representation and ultimately resulting in higher accuracy and improved quality of the predicted vectorized HD maps. The source code is available at https://github.com/LIvanoff/MapFM.

Keywords: High-Definition Map · Foundation model · Neural network · Multi-Task learning.

Дальнейшние планы работы

Следующие эксперименты:

- Использование других фундаментальных моделей и применение мехназма внимания для агрегации их признаков;
- Добавление маскирования для BEV признаков.

Также планируется развитие работы в сторону построения **глобальной карты**.

Спасибо за внимание!

ivanov.la@phystech.edu