Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Научно-образовательный центр когнитивного моделирования

Направление подготовки / специальность: 09.04.01 Информатика и вычислительная техника
Направленность (профиль) подготовки: Методы и технологии искусственного интеллекта

РАЗРАБОТКА И ИССЛЕДОВАНИЕ НЕЙРОСЕТЕВЫХ МЕТОДОВ МУЛЬТИМОДАЛЬНОГО ТРЕХМЕРНОГО ОБНАРУЖЕНИЯ ОБЪЕКТОВ

(магистерская диссертация)

Студент:
Матыкина Ольга Вячеславовна

_____
(подпись студента)

Научный руководитель:
Юдин Дмитрий Александрович,
канд. техн. наук

_____
(подпись научного руководителя)

Консультант *(при наличии)*:
_____
_____
(подпись консультанта)

Москва 2025

Содержание

1 Введение **4**

2 Обзорно-постановочный раздел работы **6**

2.1 Обзор источников данных 6

2.2 Обзор существующих подходов к 3D-детекции объектов 12

2.3 Формальная постановка задачи 15

3 Гибридный метод трехмерного обнаружения объектов на основе изображений и карт глубин **19**

3.1 Общая схема метода 19

3.2 Модуль сегментации RGB-изображений 20

3.3 Алгоритм оценки расстояний до препятствий 22

3.4 Особенности обучения 23

4 Трансформерный метод комплексирования данных радара и камеры для трехмерного обнаружения объектов **25**

4.1 Архитектура предлагаемого подхода RCDINO 25

4.2 Визуальный энкодер 25

4.3 Адаптер DINOv2 26

4.4 Разреженный радарный энкодер 26

4.5 Плотный радарный энкодер 26

4.6 Последовательный декодер 27

5 Вычислительные эксперименты **29**

5.1 Сбор и подготовка наборов данных 29

5.1.1 Сбор данных 29

5.1.2 Объединение и доразметка наборов данных 31

5.1.3 Описание открытого набора данных nuScenes 32

5.2 Оценка качества работы гибридного метода 33

5.3 Оценка качества работы RCDINO 37

6 Заключение **43**

Аннотация

Трёхмерное восприятие сцены с использованием нескольких сенсоров является ключевым элементом в системах автономного вождения и робототехнике. В данной работе рассматриваются два подхода к задаче 3D-детекции объектов: гибридный метод, основанный на сегментации сцены и оценке расстояний до объектов, а также мультимодальная трансформерная модель RCDINO, которая улучшает визуальные признаки, объединяя их с семантически богатыми представлениями из предварительно обученной фундаментальной модели DINOv2. Эксперименты на датасете nuScenes показали, что метод RCDINO достигает передовых результатов среди моделей, использующих данные от камеры и радара: 56.4 NDS и 48.1 mAP. Также реализован собственный частично размеченный датасет с синхронизированными данными радара и стерео-камеры, позволяющий протестировать устойчивость разработанных решений в условиях реального мира.

1 Введение

Трёхмерное восприятие окружающей среды играет ключевую роль в задачах автономного вождения, робототехники и интеллектуального транспорта. Одной из центральных задач в этой области является обнаружние объектов в трёхмерном пространстве, обеспечивающее безопасное и точное взаимодействие агента с окружающим миром. Особенно актуальной становится задача мультимодальной 3D-детекции, где данные с различных сенсоров, таких как камеры, радары, лидары, комбинируются для точного и надежного обнаружения объектов.

Камеры предоставляют изображения высокого разрешения, богатые текстурной и цветовой информацией, однако не позволяют напрямую измерять глубину. Радар, напротив, обеспечивает надёжную информацию о расстоянии и скорости объектов даже в сложных погодных условиях [62], но страдает от низкого пространственного разрешения и отсутствия семантики. Эффективное объединение этих модальностей представляет собой важную и всё ещё открытую проблему, в том числе из-за зашумления радарных данных.

На протяжении последних лет было предложено множество подходов к задаче радарно-камерной 3D-детекции объектов. Среди них можно выделить два основных направления. Первое включает методы, которые напрямую решают задачу обнаружения объектов: они локализуют и классифицируют цели в трёхмерном пространстве, используя такие представления, как вид сверху (BEV, bird's-eye view) для объединения признаков или архитектуры на основе запросов, в которых обучаемые запросы позволяют постепенно уточнять параметры объектов. Второе направление включает двухэтапные подходы, где сначала выполняется сегментация сцены, а затем на основе маски объекта и карты глубины восстанавливаются его пространственные координаты и габариты. Такие методы часто оказываются эффективными в условиях ограниченного количества аннотированных 3D-данных.

Несмотря на достигнутые успехи, существующие решения сталкиваются с рядом трудностей. В частности, ограниченная семантическая насыщенность визуальных признаков затрудняет обнаружение объектов на дальних дистанциях, что критически важно для надёжного автономного восприятия. Кроме того, большая часть исследований сосредоточена на условиях городской среды, в то время как менее формализованные и разнообразные сценарии, встречающиеся за её пределами, остаются слабо охваченными.

В данной работе рассматриваются представители обоих направлений. С одной стороны, предлагается гибридный метод, при котором задача 3D-детекции сводится к сегментации сцены и оценке расстояний до обнаруженных объектов с использованием карты глубины. На первом этапе нейросетевая модель сегментирует объекты, потенциально являющиеся препятствиями (например, камни, автомобили). Затем с помощью карты глубины восстанавливаются ограничивающие боксы объектов в трёхмерном пространстве.

С другой стороны, был разработан метод RCDINO, представляющий собой end-to-end решение для 3D-детекции на основе трансформерной архитектуры. Предлагаемый подход сочетает визуальный бэкбон ResNet18 [6] с предобученной моделью DINOv2 [51], извлекающей универсальные семантические признаки изображений. Для эффективной интеграции этих признаков в архитектуру детектора используется специальный адаптер. Модель также учитывает пространственную информацию радара, что позволяет улучшить локализацию объектов и устойчивость к неблагоприятным условиям. Проведённые эксперименты на датасете nuScenes [18] показывают, что предложенное решение превосходит базовую версию модели по метрикам точности детекции и работает на уровне state-of-the-art.

Структура работы следующая. В разделе 2 представлен обзор существующих датасетов и подходов к трехмерному обнаружению объектов, а также описана формальная постановка задачи. В разделе 3 описан предложенный гибридный метод обнаружения объектов по изображениям и картам глубины. В разделе 4 описана архитектура предложенной модели RCDINO. В разделе 5 приводятся детали реализации, результаты экспериментов и сравнение с другими методами. Заключение подводит основные итоги и обсуждает направления для дальнейших исследований.

2 Обзорно-постановочный раздел работы

2.1 Обзор источников данных

Существует несколько типов сенсоров, используемых в задачах восприятия окружающей среды, каждый из которых обладает своими особенностями, преимуществами и ограничениями. Для правильного выбора сенсора в системах пространственного восприятия важно понимать, какие параметры способен измерять каждый тип, в каких условиях он наиболее эффективен для решения задачи трехмерного обнаружения объектов.

RGB-камера фиксирует визуальную информацию в трёх каналах: красном, зелёном и синем. Она обеспечивает высокое разрешение и богата семантическим содержанием, что делает её незаменимой для задач классификации, распознавания объектов и семантической сегментации. Однако такая камера не даёт прямой информации о глубине сцены и чувствительна к условиям освещения — в темноте, при ярком свете, направленном на камеру, или осадках её надёжность значительно снижается.

Несмотря на ограничения, RGB-камеры широко используются в полевых условиях благодаря простоте установки и богатству визуальных признаков. Ниже перечислены примеры датасетов, включающих RGB-изображения с аннотациями, релевантными для решения задачи обнаружения объектов в сложных условиях.

В RUGD [13] представлены разнообразные объекты неурбанизированной местности. Изображения состовляют видеопоследовательности, однако разметка проходимости подходит лишь для малогабаритных роботов. В отличие от него, YCOR [12] содержит аннотацию по уровням проходимости (гладкая тропа, неровная тропа и т.п.), что позволяет адаптировать классы под конкретные типы транспортных средств.

INSTSnowRoadDetection [75] представляет интерес для работы в условиях зимнего движения: он содержит маски дороги в заснеженной среде, но не аннотирует препятствия. Датасет Soil [78, 79] включает аннотации земляных насыпей, критически важных для безопасного движения в карьерах. RoadObstacle21 [25] фокусируется на аномалиях на асфальтированных дорогах и может быть полезен в будущих исследованиях.

RGBD-камера объединяет визуальные данные с картой глубины, обеспечивая

как текстурную, так и геометрическую информацию о сцене. Глубина может быть получена с помощью стереозрения или структурированного света. Такие камеры точны на малых и средних расстояниях, однако имеют ограниченную дальность действия и чувствительны к внешнему освещению.

Объединение RGB и глубины позволяет точно оценивать расстояние до объектов и тем самым решать задачу трехмерного обнаружения. Ниже приведены датасеты, использующие RGBD-камеры.

Датасет Freiburg Forest [5] предлагает разметку лесной среды, однако не включает проезжую часть и препятствия, что ограничивает его применимость для решения поставленной задачи. Набор NPO [56] содержит положительные (выше уровня дороги) и отрицательные (ниже уровня дороги) препятствия, но не аннотирует проезжую часть. Изображения содержат сложные сцены с различными погодными условиями и типами местности. В результате, модели, предобученные на общих сегментационных датасетах [10, 11], плохо справляются с выделением дороги в заснеженных сценах этого датасета. ORFD [35] охватывает различные погодные и световые условия, но не содержит разметки препятствий. Наконец, PotholeMix [76] сосредоточен на мелких дефектах дорожного покрытия, не представляющих угрозу для тяжёлых машин.

Лидар (LiDAR, Light Detection and Ranging) использует лазерные импульсы для измерения расстояний до объектов и формирует облако точек, описывающее трёхмерную структуру сцены. Лидары обеспечивают высокую точность измерения глубины, не зависят от освещённости и особенно эффективны для задач пространственной локализации, 3D-детекции и построения карт.

Для задач оценки проходимости лидар может давать богатую геометрию местности и объектов, однако точность выводов во многом зависит от структуры разметки. RELLIS-3D [24] является датасетом, включающим лидарные данные, но ограниченным в контексте движения на машине по пересечённой местности из-за специфики аннотаций, подготовленной для небольшого колесного робота.

Радар Существует несколько типов радаров, отличающихся своими характеристиками и назначением. Для правильного выбора сенсора в системах восприятия важно понимать, какие параметры способен измерять каждый тип радара и в каких сценариях его применение наиболее оправданно.

Например, помимо радаров, используемых в автомобильных системах, существует также радар, проникающий вглубь материала — ground-penetrating radar (GPR), применяемый для поиска объектов под землёй, таких как трубы, кабели,

пустоты или археологические артефакты. Он работает в более низком частотном диапазоне, способен проникать сквозь грунт и строительные материалы, но не предназначен для обнаружения или отслеживания надземных объектов в реальном времени.

В этой работе мы остановимся на двух других типах радаров, которые широко применяются в автомобильных системах восприятия и активно используются для решения задачи пространственного обнаружения объектов: 3D-радаре и 4D-радаре.

3D-радар измеряет три величины: дальность до объекта (range), азимутальный угол (azimuth) и допплеровскую скорость (radial velocity). Именно наличие измерения скорости даёт третье измерение в термине *3D*, а не вертикальное положение цели. Антенны таких радаров расположены в горизонтальной плоскости, и потому 3D-радар не способен определить угол возвышения (elevation angle), а значит, и высоту объекта.

В свою очередь, 4D-радар добавляет четвёртое измерение — угол возвышения. Для этого антенны в 4D-радаре размещаются не только в горизонтальной, но и в вертикальной плоскости, что позволяет измерить направление прихода сигнала в двух угловых координатах: азимуте (θ) и возвышении (φ). Таким образом, становится возможным точно определить положение объекта во всех трёх пространственных координатах (x, y, z), а допплеровская скорость выступает в роли четвёртой размерности. На рисунке 1 показано, как размещение антенн в 3D- и 4D-радарах влияет на способность измерять угол возвышения и, соответственно, на точность пространственного позиционирования объектов.

Отсутствие информации о возвышении у 3D-радаров приводит к так называемой неоднозначности по высоте (elevation ambiguity): отражения от объектов, находящихся на разных высотах, проецируются в одну и ту же точку в горизонтальной плоскости. Например, 3D-радар не может отличить отражение от крыши грузовика от отражения от его колеса — оба сигнала интерпретируются как находящиеся на одной высоте, будто они расположены на уровне земли.

В отличие от этого, 4D-радар способен определить вертикальное положение источника отражения и, тем самым, точно различить, где именно в пространстве находится объект. Это особенно важно для задач трёхмерной детекции и навигации в условиях бездорожья, где препятствия можно разделить на положительные и отрицательные в зависимости от вертикальной координаты их положения относительно дороги.

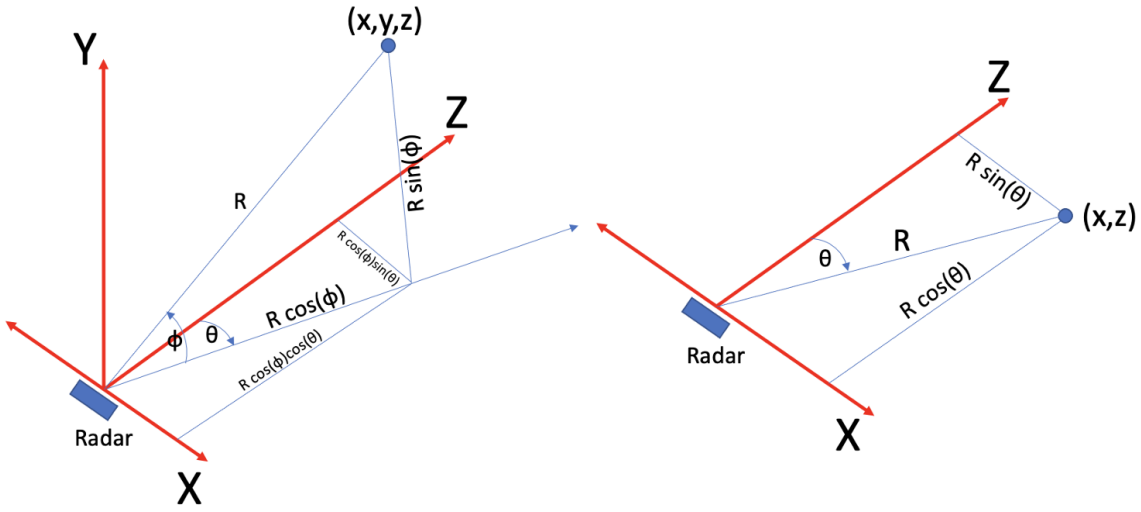

Рис. 1: Сравнение 3D- и 4D-радаров (источник: [50]). Слева - 4D-радар, способный измерять угол возвышения, справа - 3D-радар, не имеющий такой возможности.

В городской среде, где дорожные условия, как правило, более предсказуемы, 3D-радара может быть достаточно для надёжного восприятия окружающей сцены. Однако в более сложных сценариях — например, при движении по бездорожью — критически важной становится возможность оценивать высоту объектов, и в таких случаях предпочтение следует отдавать 4D-радару.

4D-радар Наиболее проработанными остаются городские датасеты с 3D-разметкой и синхронизированными данными от нескольких сенсоров. Astyx [14] представляет собой компактный, но тщательно размеченный датасет из 500 кадров с радара, лидара и камеры. Его преимущество — большое количество аннотаций (около 3000 объектов) с характеристиками в 3D: координаты, размеры, ориентация, степень окклюзии и неопределённости. Однако ограниченный объём и отсутствие погодного разнообразия снижают его пригодность для устойчивого обучения.

TJ4DRadSet [41] более масштабен — 7757 кадров в 44 последовательностях с 3D-аннотациями и ID треков, подходящий для задач отслеживания. Ещё более полным является K-Radar [36]: 35 тысяч размеченных кадров, снятых в различных погодных условиях (дождь, туман, снег), с использованием нескольких радаров, лидаров и стереокамер. Он предоставляет реалистичную и разнообразную мультимодальную информацию.

View-of-Delft [31], включающий 8693 размеченных кадра и 13 классов объектов, использует данные с радара, лидара и стереокамеры. Однако сцены исключительно

городские, что ограничивает его применимость в условиях бездорожья.

Ряд специализированных наборов данных ориентирован на задачи локализации и навигации в сложных условиях. NTU4DRadLM [59] включает 4D радар, лидар и тепловизор, собранные в off-road среде. Он радароцентричен и ценен для SLAM-исследований, особенно при плохой видимости, однако не содержит RGB-изображений и аннотаций объектов, что ограничивает его применение в задачах детекции.

Похожую направленность имеет Mine and Forest Radar Dataset [64], собранный в карьерах, шахтах и лесах. Он использует 4D радары и лидары, но объём данных невелик, а объектные аннотации отсутствуют. Один из самых насыщенных по количеству сенсорных данных — MSC-RAD4R [57]: более 90 тыс. кадров с 4D радаром, стереоизображениями, лидаром, GPS и IMU, снятых днём и ночью, в дыму и снегу. Однако, как и другие SLAM-наборы, он не содержит аннотаций объектов.

Некоторые датасеты ориентированы на исследование необработанных сигналов радара. Например, RADIal [38] предоставляет ADC-данные, а также аннотации для сегментации свободного пространства и рамки объектов в 2D/3D. Такие данные позволяют разрабатывать подходы, использующие признаки необработанных радарных данных, а не облаков точек.

Таким образом, несмотря на растущее число мультимодальных датасетов с 4D-радарами, большинство из них либо ограничены городской средой, либо предназначены для задач SLAM и локализации, а не детекции. Отдельные наборы предоставляют богатую сенсорную информацию, но не содержат объектной разметки. В то же время off-road датасеты либо не содержат RGB-изображений, либо так же не имеют разметки объектов. Это указывает на дефицит полноценных мультимодальных наборов данных, пригодных для обучения моделей 3D-обнаружения объектов в природной среде.

3D-радар Несмотря на то, что 4D-радары более применимы для решения задачи трёхмерного обнаружения объектов вне городской среды, датасеты с 3D-радарами значительно шире распространены и могут использоваться для обучения и оценки алгоритмов обнаружения.

nuScenes [18] является наиболее популярным мультимодальным датасетом, включающим RGB-изображения, облака точек лидара, радарные данные и подробную аннотацию 3D-объектов. Хотя он ориентирован на городское движение, структура и формат данных делают его удобной основой для предварительной тренировки

моделей, включая методы с использованием BEV-представления. Однако отсутствие off-road-сцен и аномалий снижает его применимость для задач в природной среде. Еще один городской датасет, RadarScenes [22] — крупный датасет с радарными облаками точек и аннотацией 11 категорий объектов. Включает более 150 последовательностей и метаинформацию о движении транспортного средства. Подходит для обучения и оценки радарных детекторов в дорожных условиях.

Однако наиболее актуальной задача мультимодального обнаружения объектов становится в сложных погодных условиях, где радарные данные особенно ценны. Для обучения и проверки методов, решающих эту задачу, существует несколько специализированных датасетов. Один из них — RADIATE [23] — датасет, ориентированный на восприятие в сложных погодных условиях (дождь, снег, туман). Включает 44000 кадров с радаром, лидаром и стереокамерой. Содержит аннотации для 8 категорий объектов. Особенность — отсутствие данных о скорости из радара, но хорошее разнообразие погодных условий. Ещё один ценный источник для обучения восприятия в сложных условиях — DENSE датасет [19]. Этот набор данных охватывает 12000 кадров в реальных дорожных сценах и 1500 кадров, снятых в контролируемой туманной камере. Датасет включает редкие погодные условия: сильный туман, дождь и снег. Он содержит синхронизированные данные от двух RGB-камер, двух лидаров, 3D радара, а также NIR и FIR датчиков. Разметка охватывает как 2D-, так и 3D-боксы по основным категориям: легковые автомобили, пешеходы, крупные транспортные средства. Богатство сенсорной информации и экстремальные погодные условия делают этот датасет особенно ценным для исследований по мультимодальному восприятию в неблагоприятных условиях. Однако, несмотря на наличие осадков и плохой видимости, сцены по-прежнему остаются в рамках дорожной инфраструктуры, что ограничивает применение набора в off-road-среде.

Boreas [46] охватывает все сезоны и время суток, записан на одном маршруте в течение года. Содержит данные с радаров, лидаров и камер. Из 44 последовательностей только часть размечена, зато есть снег, дождь и ночные сцены. Полезен для задач устойчивой детекции. CaRINA [4] делает акцент на сложных покрытиях: булыжник, выбоины, неасфальтированные участки. Этот датасет включает лидар, стереокамеру и радар, но является небольшим (900 кадров). Полезен для оценки устойчивости алгоритмов обнаружения неровностей на дороЫгах.

Среди датасетов с 3D-радарами также существуют наборы с нестандартным форматом радарных данных. Так например CARRADA [21] — небольшой

датасет с синхронизированными видео и радарными данными в формате дальность–угол–доплер, что позволяет анализировать движение объектов. Предоставлены аннотации в виде масок и ограничивающих рамок для пешеходов, велосипедистов и автомобилей. Датасет полезен для задач детекции и сегментации, но охватывает только простые дорожные сцены. Еще один датасет — Oxford Radar RobotCar [8, 20] — предоставляет обширный набор синхронизированных изображений, радарных сканов в полярных координатах, полученных в условиях города, с разнообразными погодными сценариями. Однако он лишён аннотаций объектов, что ограничивает его применение в задачах обнаружения.

Еще один датасет, который содержит данные радара, но не подходит для задач 3D-детекции, — MulRan [17]. MulRan содержит радарные и лидарные данные, записанные в нескольких городах в разное время и по разным маршрутам. Он предназначен для задач локализации и place recognition, включает 6D траектории, но не содержит семантической разметки объектов и не подходит для обучения детекторов.

Таким образом, среди существующих датасетов ни один не охватывает в полной мере задачу 3D-детекции и оценки проходимости в природной среде с использованием радара и камеры. Это мотивирует создание собственного набора данных, адаптированного под нужды обнаружения объектов в условиях бездорожья, включая аннотацию препятствий.

2.2 Обзор существующих подходов к 3D-детекции объектов

Методы сегментации сцены

Частью решения задачи трехмерного обнаружения двухэтапным подходом является использование 2D-сегментации для нахождения масок объектов, которые затем используются для построения трехмерных представлений. Существует множество методов сегментации, которые могут применяться в этой области.

Классические подходы включают сверточные нейронные сети [42, 27] и семейство YOLO [43, 16], особенно в задачах обнаружения дефектов дорожного покрытия. Более специализированные архитектуры, такие как InconSeg [56], используют информацию о глубине сцены для локализации препятствий.

TokenFusion [32] вводит трансформерную архитектуру для модального слияния изображений и карт глубины с заменой неинформативных токенов на агрегированные межмодальные характеристики. Современные универсальные модели, такие как

Mask2Former [30], решают задачи семантической, инстанс- и паноптической сегментации, а OneFormer [61] обеспечивает единое представление для разных подзадач. Эти архитектуры легли в основу комплексных решений, таких как OnePeace [60], BEIT-3 [55] и EVA [52].

Помимо рассмотренных методов 2D-сегментации, существуют методы трехмерной сегментации, один из которых следует парадигме «сверху вниз», распространенной среди методов 3D-детекции. Методы, следующие обратной парадигме «снизу вверх», выучивают эмбеддинги для каждой точки и используют их для кластеризации, тем самым формируют набор предположений об объектах. Напротив, парадигма «сверху вниз» позволяет напрямую предсказывать инстанс-предположения, а затем фильтровать их методом немаксимального подавления. Одним из методов, следующих этой парадигме, является TD3D [70], который представляет собой сеть, состоящую из сверточных разреженных слоев. Несмотря на главное предназначение метода - 3D-инстанс-сегментация, первая часть сети генерирует предположения об объектах в виде трехмерных ограничивающих боксов, что является решением задачи 3D-детекции.

Еще одним методом 3D-сегментации является OneFormer3D [66]. Как и модель 2D-сегментации OneFormer [61], этот метод позволяет обучать одну модель на общем паноптическом наборе данных для совместного решения всех трех задач сегментации.

Методы 3D-детекции

Классический end-to-end подход к 3D-детекции объектов распространен шире, чем двухстадийный, методы в этой области можно разделить на несколько категорий в зависимости от используемых представлений и архитектур.

BEV-представления Наиболее распространённый подход основан на переходе в представление «вид сверху» (Bird's Eye View, BEV), где данные с различных сенсоров проецируются в единую координатную систему. Основная трудность построения BEV-представления по изображениям от камеры заключается в неточной оценке глубины, и BEVDepth [44] решает эту проблему с помощью внедрения обучаемого модуля глубины. Чтобы точнее предсказывать скорость объектов и уменьшить количество пропущенных детекций, во многих работах используют временную информацию. BEVFormer [26] первым ввёл последовательное временное моделирование

в задаче мультикамерной 3D-детекции объектов. Авторы SOLOFusion [40] обнаруживают, что долгосрочное и краткосрочное временное слияние хорошо дополняют друг друга, и предлагают соответствующий подход. Другой подход, который использует BEV-представления и временное слияние карт признаков является FMFNet [34].

BEV также может выступать в качестве унифицированного представления для слияния признаков из разных модальностей. BEVFusion [45] стал одним из первых методов, использующих BEV как общее пространство для слияния. BEV позволяет явно сопоставить данные из разных источников, что делает его очень удобным для объединения модальностей. Еще одна работа BEVCar [63] предлагает новый подход для объединённой сегментации объектов и сегментации карты в пространстве BEV. Ключевая новизна заключается в первоначальном обучении кодирования необработанных радарных данных, которое затем используется для эффективной инициализации проекции признаков изображений в BEV-пространство. В качестве бэкбона используется фундаментальная модель DINOv2 [51] с адаптером, что и послужило источником вдохновения в нашей работе.

Несмотря на успехи использования единого пространства, ошибки выравнивания между различными BEV-представлениями могут снижать качество слияния. Некоторые подходы пытаются решить эту проблему с помощью модулей внимания. Так, RCBEVDet [68] преобразует радарные точки в BEV-признаки с помощью RadarBEVNet и объединяет их с признаками изображений через модуль внимания. Аналогично, CRN [49] применяет деформируемое кросс-модальное внимание для согласования изображений и радара в BEV-пространстве.

Методы вне BEV-пространства Существуют и альтернативные методы вне BEV-представлений. Одним из современных подходов к 3D-детекции, основанным только на изображениях от камеры, является ImVoxelNet [39]. Его особенностью является сверточная архитектура и возможность работать как с одним, так и с несколькими изображениями от различно расположенных камер. Этот метод хорошо работает как в indoor, так и в outdoor-сценах, отличаясь только головами детекции в зависимости от домена. MVFusion [58] вводит семантически согласованный радарный энкодер SARE и применяет трансформеры для построения устойчивых признаков. SpaRC [69] использует слияние через фруструм и локальное самовнимание для повышения точности локализации. Метод HyDRa [71] применяет Height Association Transformer для улучшенной оценки глубины с использованием радарных признаков. RVCDet [33] решает проблемы производительности и надежности 3D-детекции

на основе LiDAR в реальном времени, представляя быстрый динамический вокселизатор, совместимый с pillar-based моделями, и легкую дополнительную голову для фильтрации ложных срабатываний.

Методы на основе запросов Методы, основанные на запросах (query-based), представляют собой ещё одну важную парадигму в 3D-детекции объектов. В таких подходах запросы объектов взаимодействуют с токенами входных данных через модуль внимания для уточнения предсказаний. DETR3D [28] вводит набор точек в качестве опорных, каждая из которых соответствует одному объектному запросу. Впоследствии исследователи сосредоточились на эффективном извлечении признаков из мультикамерных изображений для этих объектных запросов. PETR [37] предлагает позиционное кодирование признаков изображения в координатах лидара. CAPE [47] строит позиционное кодирование в координатах камеры, чтобы уменьшить влияние изменений внешних параметров камер. StreamPETR [53] предлагает алгоритм долгосрочного временного объединения на основе запросов, который позволяет передавать историческую информацию об объектах в текущий кадр.

Современные методы всё чаще сочетают преимущества BEV-представления и query-based архитектур. Так, RaCFormer [67] предлагает извлекать признаки, релевантные объектным запросам, как из BEV-пространства, так и напрямую из изображений, избегая ошибок, связанных с неточной оценкой глубины путем обучения карт глубины по радару и признакам изображений. Кроме того, метод вводит адаптивную инициализацию запросов в полярных координатах и использует доплеровское смещение карт BEV-признаков для повышения точности временного моделирования. RCTrans [65] также реализует query-based схему, но делает упор на обогащение радарных признаков с помощью Radar Dense Encoder и объединяет их с изображениями на уровне токенов. Для предсказания объектов применяется последовательный декодер, который пошагово уточняет позиции объектов.

2.3 Формальная постановка задачи

Общая постановка задачи

Задача мультимодального трёхмерного обнаружения объектов формализуется следующим образом. Пусть I — множество изображений сцены, полученных с одной или нескольких калиброванных камер, а $M \in \{D, R\}$ — пространственная модальность, где D — карта глубины от стерео-камеры, а R — облако точек, полученное с

радара. Требуется по данным (I, M) определить множество ограничивающих трёхмерных параллелепипедов $B = \{B_1, \ldots, B_N\}$, соответствующих объектам на сцене. Задача решается с помощью модели детектора:

$$B = f_{\text{det}}(I, M), \tag{2.1}$$

где f_{det} — функция, реализуемая гибридной или нейросетевой моделью с параметрами Θ. Параметры Θ оптимизируются путём минимизации функции потерь, включающей классификационную и регрессионную составляющие. Конечная цель — обучение модели по следующему критерию:

$$\hat{\Theta} = \arg\min_{\Theta} \mathcal{L}(f_{\text{det}}(I, M; \Theta), B^{\star}),$$

где $B^{\star}$ — истинные аннотации объектов.

Каждый элемент $B_i \in B$ имеет следующий вид:

$$B_i = [x_c, y_c, z_c, l, w, h, v_x, v_y, \theta, \text{class}], \tag{2.2}$$

где $(x_c, y_c, z_c) \in \mathbb{R}^3$ — координаты центра объекта, $l, w, h \in \mathbb{R}_+$ — габаритные размеры объекта (длина, ширина, высота), $v_x, v_y \in \mathbb{R}$ — проекции скорости объекта на оси x и y, $\theta \in [-\pi, \pi)$ — угол ориентации объекта в горизонтальной плоскости, $\text{class} \in \mathcal{C}$ — категориальная метка, определяющая принадлежность объекта к одному из классов $\mathcal{C}$. Полный вектор параметров объекта участвует в задаче регрессии, за исключением категориальной метки, которая прогнозируется как задача классификации.

Функция потерь $\mathcal{L}$ включает две основные компоненты:

$$\mathcal{L}(o, \hat{o}) = \omega_1 \cdot \mathcal{L}_{\text{cls}}(c, \hat{c}) + \omega_2 \cdot \mathcal{L}_{\text{reg}}(b, \hat{b}),$$

где $o = (c, b)$ — истинная метка и параметры бокса, $\hat{o} = (\hat{c}, \hat{b})$ — предсказанные значения.

Классификационная составляющая основана на Focal Loss [9]:

$$\mathcal{L}_{\text{cls}}(c, \hat{c}) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t),$$

где $p_t = \begin{cases} \hat{c}, & \text{если } c = 1 \\ 1 - \hat{c}, & \text{если } c = 0 \end{cases}$, $\alpha_t = 0{,}25$ — параметр балансировки классов, $\gamma = 2$ — параметр фокусировки, уменьшающий вклад легко классифицируемых примеров.

Регрессионная составляющая — это средняя L1-норма между предсказанными и истинными параметрами бокса:

$$\mathcal{L}_{\text{reg}}(b, \hat{b}) = \frac{1}{K} \sum_{k=1}^{K} \left| b_k - \hat{b}_k \right|,$$

где K — количество регрессируемых параметров (координаты, размеры, скорости и угол).

Веса компонентов функции потерь устанавливаются как $\omega_1 = 2.0$ для классификационного лосса и $\omega_2 = 0.25$ для регрессионного лосса.

Частная постановка задачи

Пусть на изображении I с помощью модели сегментации получены маски объектов. По маскам получены двумерные ограничивающие прямоугольники (bounding boxes), а с помощью эвристического алгоритма из карты глубины D получены расстояния до объектов $\{r_i\}$. Центр 2D-бокса обозначим как $\mathbf{m}_i = [u_i, v_i, 1]^T \in \mathbb{R}^3$, а матрица внутренних параметров камеры — $K \in \mathbb{R}^{3\times3}$, где:

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.3}$$

Используя уравнение проекции для модели камеры-обскуры [1]:

$$\rho \cdot \mathbf{m}_i = K \cdot \mathbf{X}_i, \tag{2.4}$$

где $\rho = r_i$ — расстояние до объекта, $\mathbf{X}_i = [x_i, y_i, z_i]^T$ — координаты объекта в системе координат камеры, получим:

$$\mathbf{X}_i = r_i \cdot K^{-1} \cdot \mathbf{m}_i. \tag{2.5}$$

Таким образом, мы напрямую восстанавливаем координаты центра объекта в пространстве камеры, исходя из пиксельных координат и расстояния до объекта. На рисунке 2 представлена визуализация компонентов уравнения восстановления координат центра объекта в трехмерных координатах.

Получив трёхмерную точку $\mathbf{X}_i$, можно задать приближённый ограничивающий параллелепипед объекта B_i по аналогии с общей моделью:

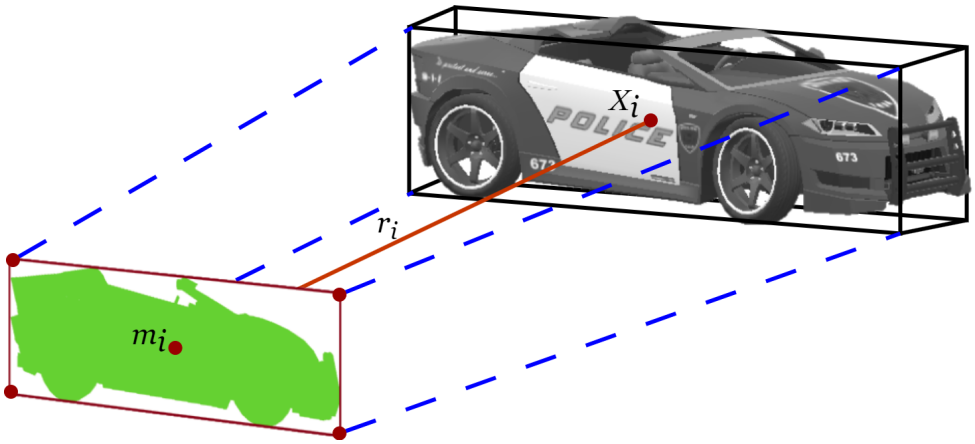

Рис. 2: Восстановление трёхмерной точки объекта по двумерной маске и расстоянию до него

$$B_i = [x_i, y_i, z_i, l_i, w_i, h_i, 0, 0, 0, \text{class}], \tag{2.6}$$

где $(x_c, y_c, z_c) = \mathbf{X}_i$, скорости и ориентация считаются равными нулю из-за отсутствия информации о движении, размеры l_i, w_i, h_i получаются из эмпирических соотношений или априорных предположений, class определяется моделью сегментации.

Таким образом, задача трёхмерного обнаружения объектов может быть сведена к следующей постановке:

$$B \approx f_{\text{det}}(I, D),$$

где f_{det} — алгоритм, извлекающий трёхмерные описания объектов на основе 2D-масок и карт глубины, полученных с помощью гибридного алгоритма, который описан далее.

3 Гибридный метод трехмерного обнаружения объектов на основе изображений и карт глубин

Описание предлагаемого гибридного метода ранее представлено в статье, принятой к рассмотрению в журнал «Труды МФТИ» [73].

3.1 Общая схема метода

В соответствии с формальной постановкой задачи (см. раздел 2.3), рассматривается построение отображения

$$B = f_{\text{det}}(I, D),$$

которое по входным данным — изображению I и соответствующей карте глубины D — возвращает множество трёхмерных ограничивающих параллелепипедов B, описывающих положение, габариты и класс обнаруженных объектов.

Гибридный метод реализует приближённое вычисление функции f_{det} как последовательность следующих этапов:

Сегментация объектов на изображении По изображению I выполняется сегментация объектов, что позволяет получить маски объектов на сцене:

$$M_{\text{obs}} = f_{\text{seg}}(I),$$

где f_{seg} — композиция двух нейросетевых моделей: предобученной YOLOv11-seg [80] и дообучаемой Mask2Former [30].

Оценка расстояний до объектов с использованием карты глубины Для каждой маски $M_i \in M_{\text{obs}}$ определяется расстояние до объекта по формуле:

$$r_i = \min(D \cap M_i),$$

где операция $\cap$ обозначает поэлементное маскирование карты глубины D маской объекта M_i. Данный этап реализован эвристически и не содержит обучаемых параметров.

Восстановление 3D-позиции объекта Точка $m_i = [u_i, v_i, 1]^T$, соответствующая центру маски M_i, преобразуется в трёхмерную координату с использованием матрицы внутренних параметров камеры K:

$$X_i = r_i \cdot K^{-1} m_i.$$

Формирование ограничивающего параллелепипеда Размеры параллелепипеда (l_i, w_i, h_i) задаются либо на основе априорной статистики по соответствующему классу, либо вручную. Компоненты скорости и ориентации полагаются равными нулю, поскольку отсутствуют соответствующие данные.

Итоговое представление объекта имеет вид:

$$B_i = [x_i, y_i, z_i, l_i, w_i, h_i, 0, 0, 0, \text{class}_i],$$

где (x_i, y_i, z_i) — центр объекта в координатах камеры, а class_i — его класс.

Таким образом, функция f_{det} аппроксимируется в виде композиции:

$$f_{\text{det}}(I, D) = f_{\text{bbox}}(f_{\text{seg}}(I), D),$$

где f_{bbox} — эвристическая процедура построения 3D-боксов на основе масок сегментации и карты глубины. Компоненты f_{bbox} фиксированы и не обучаются.

Схема на рисунке 3 иллюстрирует гибридный метод обнаружения объектов на дороге на основе комплексирования сенсорных данных. Изображения поступают в нейросетевой компонент метода, состоящий из двух моделей сегментации. Затем с помощью эвристических алгоритмов полученные маски сегментации обрабатываются и дополняются оценкой расстояния до объектов. В результате работы метода формируются маски объектов и расстояния до них. По ним, по описанному алгоритму, можно восстанавить 3D-координаты центров объектов и сформировать 3D-боксы.

3.2 Модуль сегментации RGB-изображений

Модуль сегментации использует изображения I и карты глубины D для получения масок объектов M_{obj}.

В качестве модели семантической сегментации используется Mask2Former [30], демонстрирующий высокую точность при работе с разнообразными типами поверхностей и мелкими объектами. Архитектура Mask2Former (рисунок 4) основана на трансформерах и содержит три ключевых компонента:

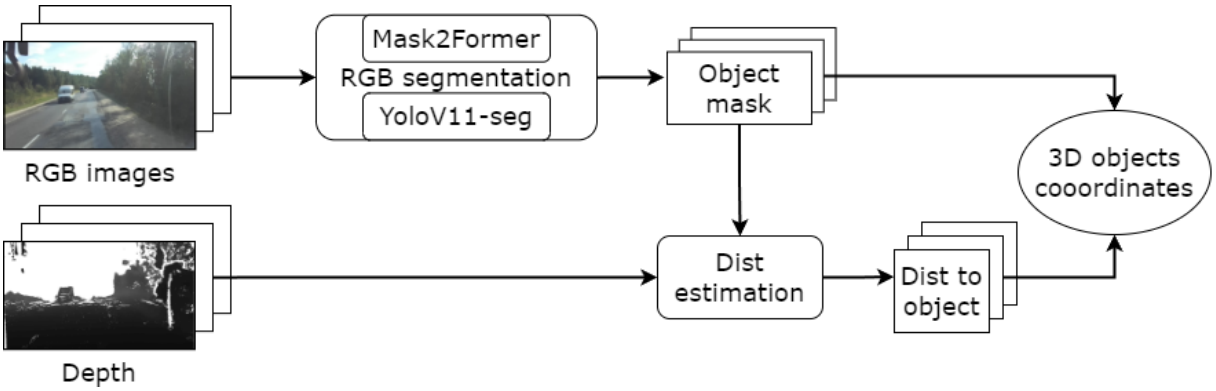

Рис. 3: Схема гибридного метода обнаружения объектов на основе комплексирования сенсорных данных

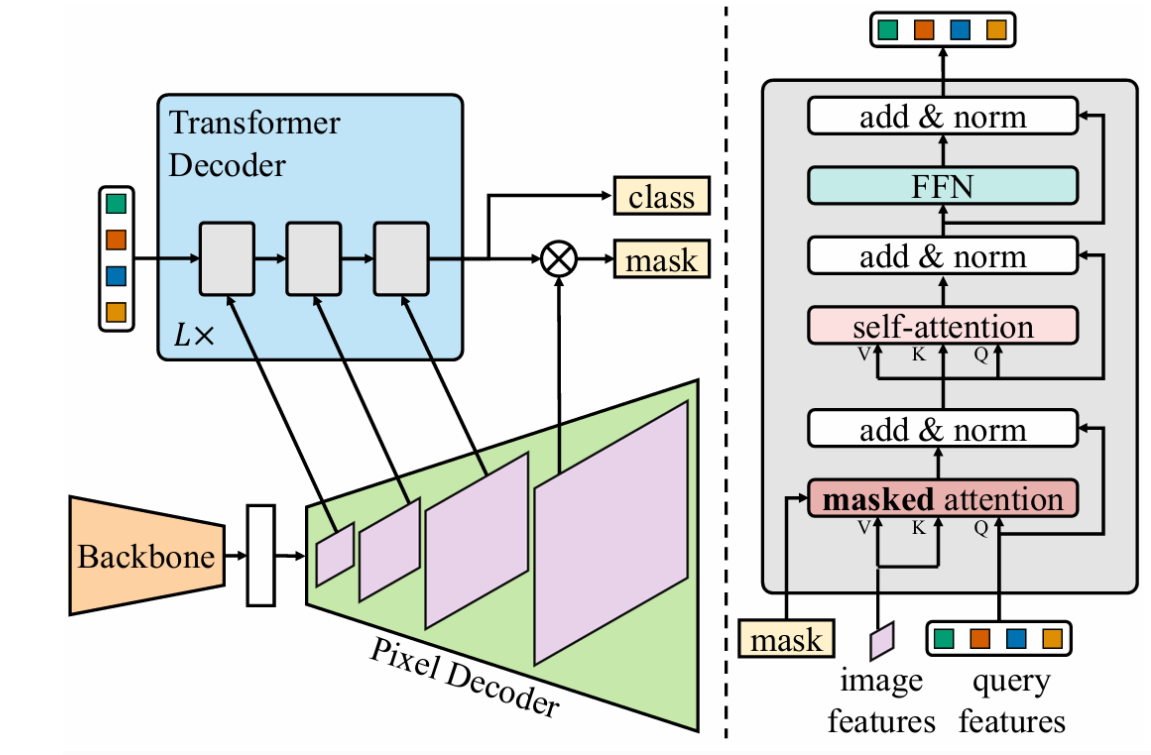

Рис. 4: Архитектура Mask2Former, часть модуля сегментации предложенного метода (источник: [30])

- Backbone (обычно Swin Transformer) для извлечения мультимасштабных признаков.
- Pixel Decoder для построения плотных карт признаков.
- Transformer Decoder с маскированным вниманием для предсказания масок.

Эта архитектура отличается универсальностью — она успешно применяется как

для семантической, так и для инстанс-сегментации и паноптической сегментации. Благодаря трансформерной архитектуре Mask2Former обеспечивает качественную сегментацию как крупных, так и мелких структур. Это особенно важно в задачах детектирования ландшафтных дефектов и мелких объектов (ям, трещин), а также точного выделения области дороги.

Мы используем Mask2Former в режиме семантической сегментации и выделяем следующие классы объектов:

- отрицательные препятствия (ямы, трещины);
- дорожное покрытие любых типов (грунт, асфальт);
- положительные препятствия (отвалы земли, заборы, здания и временные конструкции).

Для инстанс-сегментации подвижных объектов выбрана YOLOv11-seg [80]. Выбор обусловлен оптимальным балансом между скоростью и точностью для детектирования критически важных объектов: автомобилей, пешеходов, животных. Эта модель с высокой скоростью делает качественные предсказания с использованием предобученных весов. YOLOv11-seg эффективно решает задачу обнаружения подвижных и потенциально опасных объектов, дополняя Mask2Former, который ориентирован на структурные элементы сцены.

Таким образом, совмещение моделей Mask2Former и YOLOv11-seg позволяет объединить преимущества точной сегментации ландшафта и быстрых предсказаний для подвижных объектов, что повышает надежность метода в условиях реальной дорожной ситуации.

3.3 Алгоритм оценки расстояний до препятствий

В рамках этапа постобработки и оценки расстояний до объектов, маски объектов M_{obj}, полученные с помощью модуля сегментации, обрабатываются с использованием алгоритма *ObstacleDist*. Этот алгоритм позволяет сформировать геометрически непрерывные маски объектов и оценить расстояния до них по карте глубины D. Алгоритм включает следующие шаги:

1. Для каждого типа препятствий (положительные, отрицательные) из семантической маски от Mask2Former выделяются непрерывные полигоны, соответствующие отдельным препятствиям.

2. Маски, полученные от модели инстанс-сегментации YOLOv11-seg, уже являются непрерывными полигонами, поэтому они просто добавляются в массив масок без постобработки.

3. Полученные на предыдущих шагах полигоны сопоставляются с картой глубины D. Расстояние до препятствия оценивается по формуле:

$$Dist_{obj_i} = min(D \cap M_{obj_i}),$$

где $\cap$ понимается как поэлементное произведение матрицы глубины и бинарной маски препятствия, в которой нули заменены на большое число (например, 50000).

4. Результаты сохраняются в виде множества пар «маска препятствия — расстояние», которые далее используются для визуализации в системе помощи водителю или для решения задачи трехмерного обнаружения.

3.4 Особенности обучения

Из всей вычислительной цепочки f_{det} обучаемым компонентом является исключительно модель Mask2Former. Модель YOLOv11-seg используется в предобученном виде и не дообучается.

Mask2Former обучается с использованием составной функции потерь:

$$\mathcal{L} = \mathcal{L}_{\text{mask}} + \lambda_{\text{cls}}\,\mathcal{L}_{\text{cls}}, \tag{3.1}$$

где масочный лосс определяется как

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{ce}}\,\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\,\mathcal{L}_{\text{dice}}. \tag{3.2}$$

Классификационный лосс $\mathcal{L}_{\text{cls}}$ представляет собой взвешенную многоклассовую кросс-энтропию, применяемую ко всем предсказаниям (в том числе не сопоставленным с объектами аннотации):

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\lambda_i \sum_{c=1}^{C} y_{ic}\log(p_{ic}), \tag{3.3}$$

где N — общее количество предсказаний, C — число классов, $y_{ic} \in \{0, 1\}$ — индикатор истинного класса, p_{ic} — предсказанная вероятность класса c, а λ_i — вес: $\lambda_i = 2.0$ для сопоставленных предсказаний и $\lambda_i = 0.1$ для несопоставленных.

Масочный лосс $\mathcal{L}_{\text{ce}}$ представляет собой бинарную кросс-энтропию между предсказанными и истинными масками, рассчитываемую независимо для каждого пикселя:

$$\mathcal{L}_{\text{ce}}(p, y) = -\left[y \log(p) + (1 - y) \log(1 - p)\right], \tag{3.4}$$

где $y \in \{0, 1\}$ — истинное значение пикселя, $p \in (0, 1)$ — предсказанная вероятность принадлежности к положительному классу.

Функция потерь Dice [7], широко применяемая в задачах сегментации, определяется как:

$$\mathcal{L}_{\text{dice}}(p, y) = 1 - \frac{2 \sum_i p_i y_i + \epsilon}{\sum_i p_i + \sum_i y_i + \epsilon}, \tag{3.5}$$

где p_i и y_i — значения предсказаний и истинной маски для пикселя i, а ϵ — малая константа для численной устойчивости.

Весовые коэффициенты масочного лосса заданы как $\lambda_{\text{ce}} = 5$, $\lambda_{\text{dice}} = 5$.

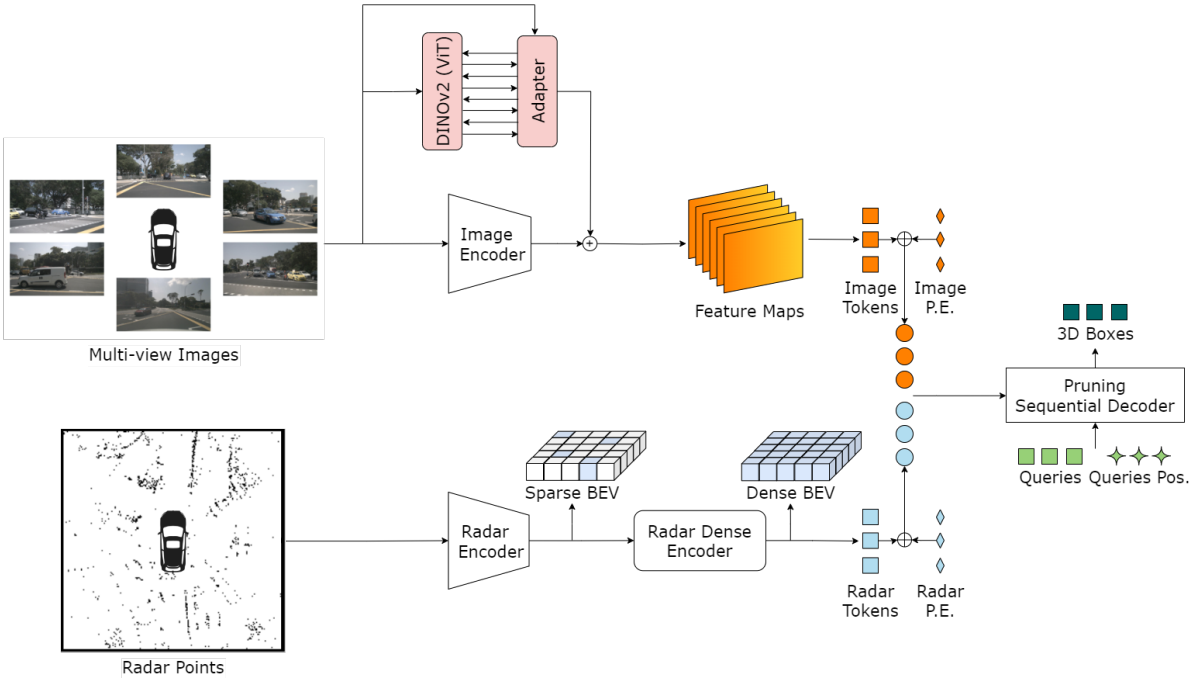

Рис. 5: Схема архитектуры модели RCDINO. Модель дополнительно использует предобученную модель DINOv2 для извлечения визуальных признаков. Радарная информация обрабатывается с помощью двух энкодеров. Трансформерный декодер итеративно уточняет запросы объектов на основе признаков двух модальностей.

4 Трансформерный метод комплексирования данных радара и камеры для трехмерного обнаружения объектов

4.1 Архитектура предлагаемого подхода RCDINO

Предлагаемая модель **RCDINO** (Radar-Camera DINO) представляет собой мультимодальный 3D-детектор с трансформерным декодером, который итеративно уточняет запросы к объектам сцены на основе признаков от камеры и радара. Архитектура модели включает пять основных компонентов: визуальный энкодер, адаптер DINOv2, разреженный радарный энкодер, плотный радарный энкодер и последовательный декодер. Схема архитектуры представлена на рисунке 5.

4.2 Визуальный энкодер

Для каждого изображения из набора входных камер извлекаются признаки при помощи стандартного сверточного или трансформерного бэкбона, например,

ResNet [6] или Vision Transformer (ViT) [15]. Эти признаки формируют компактное представление изображения, которое затем объединяется с признаками, извлечёнными из DINOv2, и используется для дальнейшей обработки.

4.3 Адаптер DINOv2

Признаки из модели DINOv2 извлекаются с нескольких внутренних слоёв с помощью адаптера, что позволяет использовать предобученную модель без дообучения или с ним. Адаптер состоит из:

- сверточной сети, извлекающей признаки изображения с разными разрешениями;
- двух ключевых модулей — **инжектора** и **экстрактора**, реализующих многоуровневое деформируемое внимание на основе предобученных весов DINOv2.

Признаки с разных уровней с помощью сверток и интерполяции приводятся к единой размерности, совместимой с визуальным бэкбоном. Далее к признакам бэкбона с обучаемым весом прибавляются полученные на этом этапе признаки DINOv2.

Интеграция признаков DINOv2 расширяет визуальное представление за счёт высокоуровневой семантики, полученной на масштабных разнородных данных, что повышает точность восприятия сцены в сложных условиях.

4.4 Разреженный радарный энкодер

Для обработки радарных точек используется подход, аналогичный Futr3D [54]. Входной массив $P \in \mathbb{R}^{N_r \times 5}$ включает 3D-координаты, компенсированную скорость и временное смещение. Сначала применяется MLP для извлечения признаков пилларов, которые затем распределяются по сетке BEV. Компонент используется в том же виде, что и в бейзлайне.

4.5 Плотный радарный энкодер

Для компенсации разреженности радарных признаков введён плотный радарный энкодер, вдохновлённый архитектурой U-Net [3]. При построении радарной BEV-карты $B \in \mathbb{R}^{H_r \times W_r \times C_r}$ извлекаются признаки пониженной размерности $B_d = \{B_i \in \mathbb{R}^{\frac{H_r}{2^i} \times \frac{W_r}{2^i} \times C_i}, i = 1, 2, 3\}$. В самом нижнем уровне (B_3) применяется self-attention,

формируя адаптивно заполненную карту B_f, куда добавляется двумерное позиционное кодирование. Затем B_f повышается до исходного разрешения, а через skip-соединения объединяется с соответствующими уровнями B_d. Архитектура повторяет реализацию в бейзлайне.

4.6 Последовательный декодер

Вместо стандартного декодера, который обрабатывает мультимодальные признаки одновременно в одном большом блоке, мы используем последовательную структуру: признаки от камеры и радара последовательно подаются в два меньших по размеру слоя. Такой подход упрощает процесс слияния модальностей и снижает вычислительную сложность. Метод аналогичен тому, что используется в бейзлайне.

На каждом шаге входной запрос сначала проходит через слой внимания с признаками BEV, затем — с признаками RV (range view, вид в пространстве камеры). Такой подход уменьшает размерность операций на каждом этапе и облегчает слияние модальностей.

На вход декодера подаются:

- запросы ($query$) и их позиционные признаки ($query_pos$);
- ключи (key), соответствующие признакам сцены (визуальным и радарным), и их позиции (key_pos);
- временные признаки из предыдущих кадров ($temp_memory$ и $temp_pos$);
- начальные опорные точки ($reference_points$);
- обучаемые слои регрессии и классификации.

На каждом шаге декодера:

1. Производится двухэтапная обработка запросов — сначала через BEV-ветку, затем через RV-ветку с соответствующими позиционными признаками и временным контекстом.
2. Результат нормализуется, преобразуется и используется для предсказания классов объектов и регрессионных координат.

3. Координаты объектов уточняются прибавлением предсказанных смещений ΔR к координатам опорных точек с предыдущего шага:

$$R^{n+1} = R^n + \Delta R.$$

4. Предсказания накапливаются по мере прохождения слоёв, а промежуточные представления сохраняются.

Финальный вывод декодера включает:

- список предсказаний классов объектов;
- список соответствующих координат;
- набор промежуточных выходов с каждого шага.

Такой подход позволяет эффективно использовать временной и модальный контекст для уточнения детекции, сохраняя при этом модульность архитектуры.



Рис. 6: Примеры изображений из собранного набора данных

5 Вычислительные эксперименты

5.1 Сбор и подготовка наборов данных

5.1.1 Сбор данных

Для обучения и валидации моделей машинного зрения был собран обширный набор данных в целевых средах. Сбор проводился в разных регионах России в рамках натурных испытаний в городах Долгопрудный, Чехов и Алдан. В результате было собрано около 2 ТБ данных, включающих изображения с разрешением 1280x720 пикселей, снятые в различных погодных и ландшафтных условиях. На рисунке 6 представлены примеры из различных частей собранного набора данных.

Таблица 1 содержит подробную информацию о собранных данных для трёх экспериментальных наборов: VegaDolgoprudny, VegaChekhov и VegaSeligdar. Указаны объемы данных, включая количество RGB- и depth-фреймов для двух камер (zed1 и zed2), количество сообщений радара и среднее количество точек в каждом сообщении. Во всех наборах данные камер представлены с одинаковым разрешением 1280x720 пикселей. Общий размер данных увеличивается от VegaDolgoprudny (213 ГБ) к VegaSeligdar (1107 ГБ), что связано с продолжительностью записей, которая

варьируется от 46 минут до почти трёх часов.

Таблица 1: Сводная таблица по объему собранных данных

Параметр	VegaDolgoprudny	VegaChekhov	VegaSeligdar	Итого
Кол-во RGB-фреймов камеры zed1	27271	81602	148096	256969
Кол-во depth-фреймов камеры zed1	27302	81769	148092	257163
Кол-во RGB-фреймов камеры zed2	27584	88365	140992	256941
Кол-во depth-фреймов камеры zed2	27604	88339	140982	256925
Кол-во сообщений радара	25572	32275	94264	152111
Среднее количество точек радара в сообщении	53	33	55	47
Размер файлов с собранными данными, ГБ	213	677	1107	1997
Длительность, мин	46	130	169	345

Первый поднабор, VegaDolgoprudny, был сформирован в ходе поездок по бездорожью. Целью этого этапа являлось фиксирование дефектов дорожного покрытия, таких как ямы и лужи, для последующего использования данных в задаче обнаружения отрицательных препятствий. Сбор данных велся в условиях пересеченной местности, что позволило зафиксировать большое разнообразие дорожных дефектов и окружающих объектов.

Второй поднабор, VegaChekhov, был собран на территории завода и прилегающих территорий в городе Чехов. Здесь внимание уделялось наличию специфичных

препятствий целевого домена, включая кучи земли, строительные конструкции, заборы и специализированную технику.

Третий поднабор, VegaSeligdar, был собран в зимний период в карьере в городе Алдан (Республика Саха). Основной задачей являлся сбор данных в заснеженной среде для имитации условий эксплуатации в северных регионах. Данные охватывают как дневные, так и ночные сцены, включая движение техники по карьерным дорогам, обнаружение людей в зоне работ и фиксацию различных типов препятствий.

Благодаря охвату различных типов местности и погодных условий собранный набор данных обеспечивает широкую вариативность сценариев и пригоден для обучения моделей в задачах обнаружения объектов в реальных условиях эксплуатации.

5.1.2 Объединение и доразметка наборов данных

Для обучения и количественной оценки качества алгоритмов распознавания был разработан набор данных VegaFullV1.0, подробное описание структуры которого приведено в таблице 2.

Таблица 2: Распределение данных между обучением и валидацией объединенного набора данных VegaFullV1.0

Набор данных	Обучение	Валидация	Общее количество
YCOR [12]	600	151	751
Soil [78, 79]	84	14	98
INSTSnowRoadDetection [75]	225	15	240
VegaDolgoprudny	51	19	70
VegaChekhov	97	20	77
VegaSeligdar	81	20	101
Total	1118	239	1357

Собранные во время натурных экспериментов изображения были размечены с помощью инструмента CVAT [74]. При разметке были выделены три категории:

1. negative: ямы, лужи;

2. positive: столбы, заборы, дома, отвалы земли, сугробы снега.

3. road: дорога.

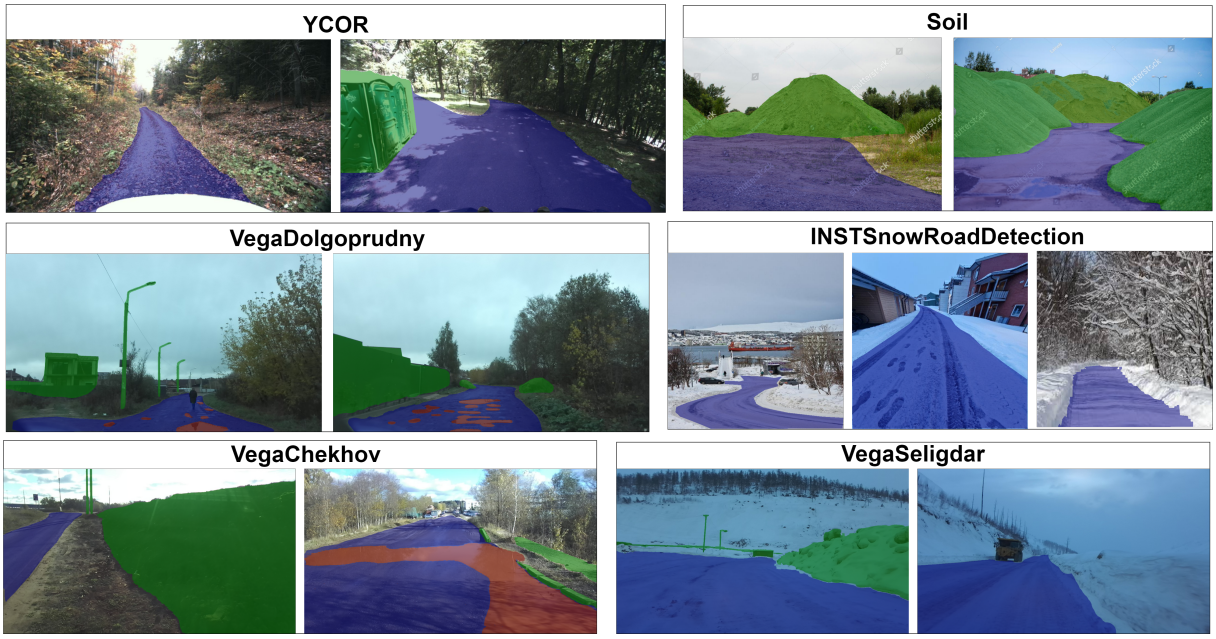

Рис. 7: Примеры размеченных изображений из открытых и собранных датасетов

На рисунке 7 приведены примеры размеченных изображений из открытых и собранных датасетов.

5.1.3 Описание открытого набора данных nuScenes

Как и в оригинальной работе [65], эксперименты проводились на популярном датасете для автономного вождения с аннотациями для трехмерного обнаружения объектов – nuScenes [18]. Этот датасет содержит 1000 сцен и разделён на три части: 700 для обучения, 150 для валидации и 150 для тестирования. Для каждого кадра в nuScenes присутствуют 6 изображений и 5 облаков точек радара, охватывающих 360°. В датасете примерно 1.4 миллиона размеченных 3д-ограничивающих рамок по десяти классам: автомобиль, грузовик, автобус, прицеп, строительная техника, пешеход, мотоцикл, велосипед, дорожный конус, барьер. Для некоторых классов заданы дополнительные атрибуты, описывающие состояние объекта, например, стоящий или движущийся.

В качестве основных метрик для оценки 3D детекции выбраны nuScenes detection score (NDS) и mean average precision (mAP). NDS — это взвешенная сумма mAP и других метрик, заданных авторами датасета, включая: Average Translation Error (ATE) (meters), Average Scale Error (ASE) (1 - 3D IoU), Average Orientation Error (AOE) (radians), Average Velocity Error (AVE) (m/s), Average Attribute Error (AAE) (1-accuracy).

- Среднюю ошибку смещения (Average Translation Error, ATE) в метрах,
- Среднюю ошибку масштаба (Average Scale Error, ASE) (1 - 3D IoU),
- Среднюю ошибку ориентации (Average Orientation Error, AOE) в радианах,
- Среднюю ошибку скорости (Average Velocity Error, AVE) в м/с,
- Среднюю ошибку атрибутов (Average Attribute Error, AAE) (1 - точность).

5.2 Оценка качества работы гибридного метода

Результаты экспериментов на собранных наборах данных

В таблице 3 приведены результаты оценки семантической сегментации для обучающей выборки, а также для отдельных частей обучающего датасета. Метрики по отрицательным категориям (IoU negative) для VegaSeligdar (ночь и день) не указаны, так как в разметке данного набора данных отсутствуют дефекты дороги, расположенные ниже уровня ее поверхности. Это обусловлено условиями среды в Якутии, где дороги в рассматриваемый период покрыты снегом и льдом, что исключает возможность появления таких дефектов. На валидационной выборке метрика качества mIoU достигает 73.47%, fwIoU достигает 92%. Ее улучшение возможно через итеративное расширение и доразметку обучающей выборки.

Таблица 3: Результаты оценки качества сегментации Mask2Former дефектов дорожного полотна, положительных препятствий и дороги в различных условиях

Датасет	mIoU	fwIoU	IoU road	IoU neg	IoU pos	Условия
VegaFullV1.0-train	94.30	96.82	95.92	88.59	95.17	осень/зима, день/ночь
VegaSeligdar (night)	93.14	95.14	93.06	-	89.38	зима, снег, ночь
VegaSeligdar (day)	96.43	97.90	97.88	-	92.87	зима, снег, день
VegaChekhov	95.85	97.69	97.45	94.39	92.80	осень, асфальт, день
VegaDolgoprudny	89.12	98.27	96.25	69.34	93.96	осень, бездорожье, день

На рисунке 8 приведены примеры работы обученной модели Mask2Former на валидационной выборке из собранных датасетов. Результаты показывают, что модель способна эффективно сегментировать объекты в различных условиях съёмки.

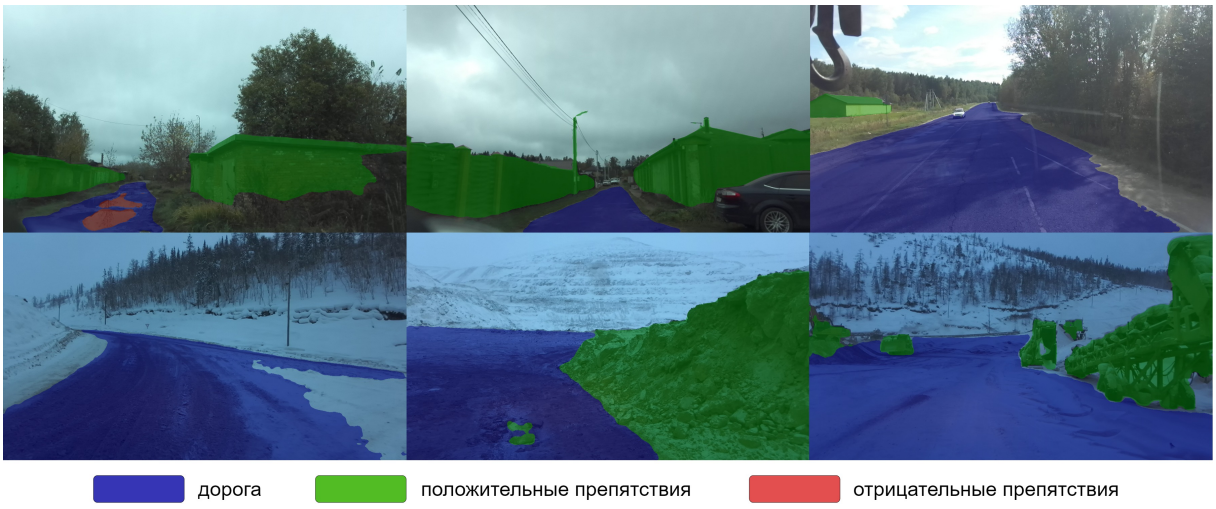

Рис. 8: Результат работы обученной модели Mask2Former на валидационной выборке



Рис. 9: Результат обнаружения препятствий моделями сегментации

На рисунке 9 приведены примеры обнаружения объектов с помощью обученной на собранном датасете модели Mask2Former и предобученной на датасете COCO [2] модели YOLOv11-seg. Визуализация также отображает оценку расстояний до объектов.

Результаты тестирования алгоритмов распознавания в бортовой системе

Разработанные алгоритмы были интегрированы в единую бортовую систему и протестированы в различных условиях эксплуатации. Программа написана на языке Python с использованием библиотек фреймворка ROS2 Foxy [77]. Данные от сенсоров камеры поступают синхронно, левое изображение стерео-камеры полностью согласованы с картой глубины.

На рисунке 10 представлена схема взаимодействия узлов и топиков системы. Синим цветом выделены основные узлы, реализующие сегментацию изображений и определение расстояний до препятствий. Красным цветом выделен топик, являющийся результатом работы модуля. Ниже подробно описаны выделенные узлы:

- **mask2former_node** получает сообщения из топика с изображениями и отправляет результаты сегментации в топик `segmentation`. Этот узел запускает

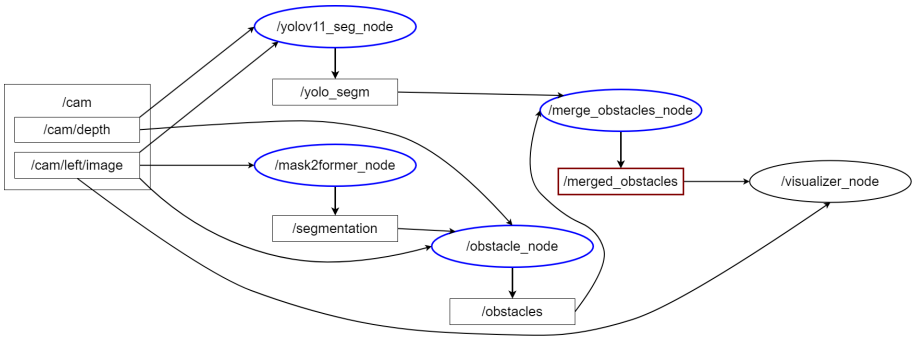

Рис. 10: Схема взаимодействия узлов и топиков разработанного модуля

нейросетевую модель Mask2Former [30], получает семантические маски и отправляет их в выходной топик.

- **yolov11_seg_node** получает сообщения из топиков с изображениями и глубинами, а затем отправляет результаты сегментации в топик `yolo_segm`. Этот узел запускает нейросетевую модель YOLOv11-seg [80], получает маски инстанс-сегментации. Затем из карт глубины вырезаются части, соответствующие маскам обнаруженных объектов, и вычисляются расстояния до них. Сообщения типа Objects отправляются в выходной топик.

- **obstacle_node** получает сообщения из топика `/segmentation` и от камеры глубины, а затем отправляет результаты в топик `/obstacles`. Этот узел получает семантическую маску дороги и препятствий, оставляет только категории препятствий. Затем семантическая маска каждого вида препятствий (отрицательные и положительные) разделяется на непрерывные полигоны, таким образом приводится к подобию масок инстанс-сегментации. Из карт глубины вырезаются части, соответствующие маскам обнаруженных объектов, и вычисляются расстояния до них. Сообщения типа Objects отправляются в выходной топик.

- **merge_obstacles_node** получает сообщения от узлов `yolov11_seg_node` и `obstacle_node` и отправляет их в топик `/merged_obstacles`. Этот узел объединяет сообщения о препятствиях, обнаруженных двумя нейросетевыми моделями.

Для быстрой передачи сообщений результирующего топика и удобной работы с ними реализован пользовательский тип сообщений `Objects`. Описание его полей приведено ниже:

- `std_msgs/Header header` — стандартный заголовок сообщения.

- `int32 num` — количество препятствий.
- `int32[] classes_ids` — массив категорий препятствий:
  - 1 — отрицательное препятствие,
  - 2 — положительное препятствие,
  - 4 — человек (категория 3 соответствует дороге).
- `Maskinroi[] masks` — массив масок препятствий, каждая содержит:
  - `int32 width` — ширина исходного изображения,
  - `int32 height` — высота исходного изображения,
  - `Roi roi` — минимальный прямоугольник, описывающий маску препятствия, с полями:
    - `int32 x` — координата левого верхнего угла по высоте,
    - `int32 y` — координата левого верхнего угла по ширине,
    - `int32 width` — ширина ограничивающего прямоугольника,
    - `int32 height` — высота ограничивающего прямоугольника.
  - `uint8[] mask_in_roi` — одномерный массив из 0 и 1, маска препятствия в ограничивающем прямоугольнике.
- `float32[] distances` — массив расстояний до препятствий.

В табл. 4 приведены замеры быстродействия компонентов системы.

Таблица 4: Замеры быстродействия компонентов системы

Алгоритм	**Оборудование**	**Время, мс**
mask2former_node	NVIDIA RTX 4070	48
yolov11_seg_node	NVIDIA RTX 4070	11
obstacle_node	Intel Core i7-13700HX	15

Общая производительность системы позволяет достигать частоты обработки выше 12 фреймов в секунду при обеспечении высокой точности. Основная вычислительная нагрузка ложится на модель Mask2Former, однако все компоненты работают с достаточной эффективностью для использования в реальных условиях.

5.3 Оценка качества работы RCDINO

Детали реализации

Как и в оригинальной процедуре обучения RCTrans [65], для предсказания на текущем кадре накапливается информация из 4 предыдущих кадров.

Количество декодерных слоёв трансформера установлено равным 6 при обучении и 3 при инференсе. Количество запросов, размер очереди памяти и количество распространяемых запросов установлены в 900, 512 и 128 соответственно.

Точки радара аккумулируются с 6 предыдущих кадров, как в CRAFT [48]. Размер BEV радара задан как 128×128.

Модель DINOv2 [51] использовалась вместе с адаптером. Были использованы веса предобученной модели DINOv2-small, размер патча 14×14. Чтобы использовать такую модель, размер входных изображений был изменен до 224×448 путем билинейной интерполяции. Признаки извлекались с 4 слоев DINOv2 и объединялись с помощью интерполяции и свертки. Затем полученные таким образом признаки интерполировались до размера 16×44 и складывались с призаками ResNet с обучаемым весовым коэффициентом.

Обучение сети проводилось в течение 16 эпох, начиная с предобученных весов оригинальной модели, с размером батча 12 и накоплением градиента в течение 3 шагов на 2 GPU NVIDIA A100. Оптимизация выполнена с помощью AdamW с коэффициентом регуляризации весов (weight decay) 10^{-2}, а скорость обучения регулируется фиксированным шагом обучения со значением 1×10^{-6} в течение 11 эпох и 1×10^{-7} в течение последних 5 эпох.

Результаты

Эксперимент показал, что интеграция предобученной модели DINOv2 с архитектурой RCTrans позволяет улучшить качество 3D-детекции объектов на наборе данных nuScenes [18]. Наблюдается рост метрик NDS на 0.4% и mAP на 0.7% по сравнению с бейзлайном и достигает 56.4 NDS и 48.1 mAP. В таблице 5 приведено сравнение результатов различных моделей трехмерного обнаружения на валидационной выборке датасета. На рисунке 11 представлены качественные результаты 3D-детекции объектов на валидационном наборе nuScenes. Для иллюстрации вынесены только изображения с одной камеры для каждой сцены, так как на них видны наибольшие различия. Видно, что обогащение визуальных признаков позволяет точ-



Рис. 11: Качественные результаты 3д-детекции объектов на валидационном наборе nuScenes. *Верхний ряд* демонстрирует, что обогащение визуальных признаков позволяет точно обнаруживать удаленные объекты, а *нижний ряд* показывает правильную классификацию плохо видимых объектов.

Таблица 5: Сравнение результатов 3д-детекции объектов на валидационном наборе nuScenes. C и R обозначают камеру и радар соответственно.

Метод	Вход	Бэкбон изображения	Размер изображения	NDS (%) ↑	mAP (%) ↑
BEVDepth [44]	C	ResNet50	256 × 704	47.5	35.1
BEVPoolv2 [29]	C	ResNet50	256 × 704	52.6	40.6
SOLOFusion [40]	C	ResNet50	256 × 704	53.4	42.7
StreamPETR [53]	C	ResNet50	256 × 704	54.0	43.2
CRN [49]	RC	ResNet18	256 × 704	54.3	44.8
RCBEVDet [68]	RC	ResNet18	256 × 704	54.8	42.9
RobuRCDet [72]	RC	ResNet18	256 × 704	55.0	45.5
RCTrans (baseline) [65]	RC	ResNet18	256 × 704	56.0	47.4
RCDINO	RC	ResNet18	256 × 704	**56.4**	**48.1**

но обнаруживать удаленные объекты, а также правильно классифицировать плохо видимые из-за погодных условий объекты.

Эксперименты с трансформерным декодером

Были проведены эксперименты с трансформерным декодером и архитектурой сети в целом для выявления влияния различных модальностей на качество 3D-обнаружения объектов. Авторами предложено необычное строение трансформерной

головы детекции, которое включает последовательное уточнение позиционных эмбеддингов запросов объектов. Заранее по референсным точкам запросов считаются 2D- и 3D-позиционные эмбеддинги запросов.

Сначала в трансформерный слой поступают 2D-позиционные эмбеддинги запросов, которые путём перекрёстного внимания с 2D-эмбеддингами радара (в координатах BEV-карты признаков) изменяют запросы. Затем в следующий трансформерный слой подаются 3D-позиционные эмбеддинги запросов, которые аналогично трансформируются через cross-attention с позиционными эмбеддингами изображений. Выходы этих двух слоёв декодера обновляют референсные точки запросов, и вся процедура повторяется. Таким образом, положение запросов уточняется итеративным способом.

Однако авторы [65] не провели исследование того, как отсутствие одного из компонентов процесса обновления позиционных эмбеддингов влияет на качество предсказаний. В данной части работы представлены такие эксперименты и сделаны соответствующие выводы.

В первую очередь были проведены эксперименты по отключению одной из модальностей из всей архитектуры сети. Удаление ветки камеры или радара, закономерно, привело к ухудшению качества предсказания. Обучение проводилось с инициализацией от предобученных весов полной модели, что позволило сократить время обучения. В случае удаления всей ветви, связанной с определённой модальностью, из декодера также исключался соответствующий позиционный эмбеддинг запросов — либо 2D-, либо 3D-. Для сохранения глубины сети слой с отсутствующим эмбеддингом заменялся на слой с оставшимся компонентом. Таким образом, сеть состояла из нескольких идентичных трансформерных слоёв.

Значительное ухудшение качества предсказаний при удалении одной из модальностей подчёркивает важность присутствия обоих компонентов позиционного кодирования запросов в трансформерной голове. Поэтому был проведён дополнительный эксперимент, в котором была предпринята попытка сохранить архитектуру токенов запросов в исходном виде, несмотря на отсутствие данных от одного из сенсоров.

Для удовлетворения этих условий признаки отсутствующего сенсора в голове детекции заменялись на тензор из единиц. Как и в предыдущих экспериментах, модель инициализировалась предобученными весами полной модели с бэкбоном ResNet18. Такая постановка имитирует поведение модели в ситуации, когда один из сенсоров перестаёт передавать данные.

В случае отключения радара BEV-карта его признаков заменялась на тензор,

заполненный единицами. При отключении камеры признаки изображений от бэкбона также заменялись на тензор из единиц, совпадающий по размерности с ожидаемым тензором признаков.

Было сделано интересное наблюдение: при таком отключении одной из модальностей качество модели, работающей на изображениях, ухудшилось на так сильно, как при полном удалении ветви радара, и, соответственно, без 2D-позиционных эмбеддингов запросов. Это может свидетельствовать о том, что сохранение структуры головы и позиционных эмбеддингов запросов даже без содержательных признаков даёт модели некоторое преимущество. Результаты описанных экспериментов представлены в таблице 6.

Таблица 6: Результаты экспериментов по отключению модальностей и компонентов трансформерной головы RCTrans. C и R обозначают камеру и радар соответственно.

Вход	Кол-во стадий декодера	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓	NDS ↑
R	1	0.0	111.5	63.4	71.1	92.8	58.3	11.5
R	2	1.3	98.5	61.2	66.0	80.2	55.7	14.5
C	1	23.0	89.2	28.7	74.0	99.1	26.2	29.8
C	2	30.1	79.7	28.4	63.7	51.0	25.7	40.2
RC	2	47.4	54.0	27.4	55.7	20.8	19.0	56.0

На рисунке 12 представлены качественные результаты 3D-детекции объектов на валидационном наборе nuScenes, полученные с различными способами удаления модальности радара. Можно видеть, что при отсутствии 2D-позиционных эмбеддингов запросов (т.е. при удалении ветви радара), модель с высокими вероятностями предсказывает несколько положений одного и того же объекта, отличающихся по глубине. Это особенно заметно на визуализации вида сверху. При этом сохранение 2D-позиционных эмбеддингов запросов (т.е. при замене признаков радара на тензор из единиц) позволяет модели более точно предсказывать положение объектов, хотя и с меньшей точностью, чем в случае полной модели. Это говорит о том, что обучение модели с сохранением структуры головы и позиционных эмбеддингов запросов даже без содержательных признаков даёт модели некоторое преимущество.

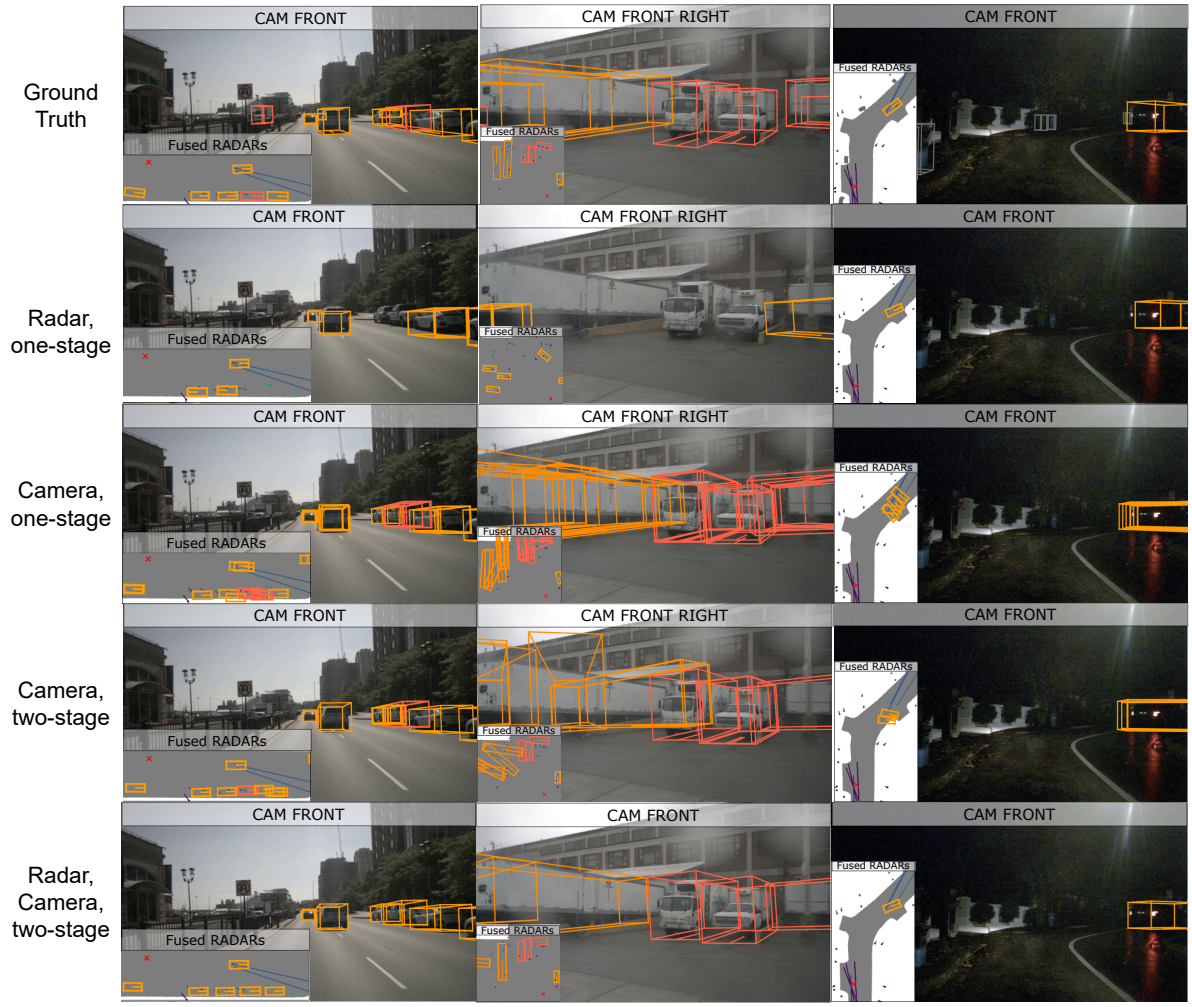

Рис. 12: Анализ влияния модальностей радара и камеры, а также компонентов декодера на качество 3D-детекции на валидационном наборе nuScenes. *Второй ряд* показывает, что удаление модальности камеры ухудшает качество обнаружения. *Третий* и *четвертый ряды* демонстрируют, что использование двухступенчатого декодера даже без информативных радарных признаков снижает неопределенность модели в локализации 3D bbox.

Инференс на собранном датасете

Для проверки обобщающей способности модели RCDINO на собранном датасете VegaFull была проведена процедура инференса на тестовой выборке. Из-за разницы в конфигурации сенсоров, а именно, из-за наличия только одной камеры и радара, модель была адаптирована к новым условиям. Произведено дообучение модели RCDINO на датасете nuScenes с использованием только одной камеры и радара, смотрящих вперед. Это позволило модели адаптироваться к новым условиям и использоваться для 3D-детекции объектов на собранном датасете без попытки предсказывать объекты с других сторон машины при отсутствии данных.

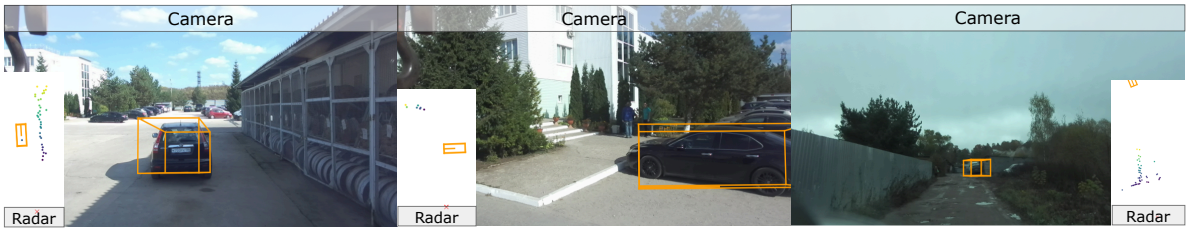

Рис. 13: Результаты 3д-детекции объектов моделью RCDINO на собранном наборе данных VegaFull

На рисунке 13 представлены результаты 3D-детекции объектов на собранном наборе данных. Модель обнаруживает объекты, что подтверждает её универсальность и способность к обобщению на новые данные. Однако при сборе датасета VegaFull не собирались данные о положении транспортного средства в пространстве, что критически важно для временного сопоставления предсказания и уплотнения радарных облаков точек. В условиях отсутствия такой информации, в каждый момент времени использовалось только одно радарное облако точек из текущего кадра. Временное сопоставление предсказаний применялось с искусственными данными об одометрии из датасета nuScenes, что не могло обеспечить точное позиционирование объектов в пространстве.

Для улучшения качества предсказаний в будущем планируется реализовать алгоритм визуальной одометрии, который будет использовать данные с камеры для оценки положения транспортного средства в пространстве. Это позволит , во-первых, уплотнять радарные облака точек данными с предыдущих кадров, а во-вторых, использовать информацию о движении транспортного средства для использования предсказаний с предыдущих шагов для улучшения точности предсказания на текущем кадре.

6 Заключение

Основные результаты работы Предложен новый гибридный подход к задаче трёхмерной детекции объектов, основанный на сегментации сцены и оценке расстояний до обнаруженных элементов. Метод был протестирован в бортовой системе в реальных условиях эксплуатации и показал вычислительную эффективность на уровне реального времени.

Также был разработан метод RCDINO — end-to-end архитектура 3D-детекции, использующая трансформеры и обогащённая признаками, извлечёнными из фундаментальной модели DINOv2. Эксперименты на датасете nuScenes показали, что интеграция визуальных признаков фундаментальной модели повышает точность обнаружения объектов по сравнению с базовой моделью и позволяет модели работать на уровне state-of-the-art.

Было исследовано влияние модальностей радара и камеры, а также компонентов декодера на качество 3D-детекции. Проведённые эксперименты показали, что использование двухстадийной структуры декодера и позиционных эмбеддингов запросов даже без содержательных признаков одной из модальностей позволяет модели лучше справляться с задачей 3D-детекции.

Кроме того, был разработан новый датасет VegaFull, содержащий изображения и радарные облака точек, собранные в сложных дорожных и погодных условиях. Датасет включает в себя сцены с различными типами дефектов дорожного полотна, положительными препятствиями и дорогой. Набор данных позволил провести обучение и тестирование алгоритмов сегментации для гибридного подхода, а также позволил оценить обобщающую способность модели RCDINO.

Ограничения и перспективы Гибридный метод, предложенный в работе, в основном опирается на визуальные данные, полученные с RGB-камеры. Это делает его чувствительным к неблагоприятным погодным условиям — таким как туман, снегопад или сильный дождь, — при которых качество изображений существенно снижается. Для повышения устойчивости метода к подобным условиям в будущем планируется использовать RGBD-сегментацию с компенсацией отсутствующих или искажённых данных глубины с помощью радара. Это позволит повысить надёжность детекции объектов в плохих погодных условиях.

Метод RCDINO также имеет определённое ограничение: в текущей реализации

признаки изображения не транслируются в BEV-пространство. Между тем, такие признаки при наличии точной глубинной информации могли бы существенно обогатить пространственное описание сцены. В дальнейшем планируется модифицировать архитектуру RCDINO, добавив обучаемый модуль оценки глубины, использующий радарные данные. Это обеспечит более точную трансформацию визуальных признаков в BEV-пространство и потенциально повысит точность локализации объектов.

Список литературы

1. *Moons T.*, *Gool L. V.*, *Vergauwen M.* 3D Reconstruction from Multiple Images: Part 1 - Principles // Found. Trends Comput. Graph. Vis. — 2009. — Т. 4. — С. 287—404. — URL: https://api.semanticscholar.org/CorpusID:54901473.
2. Microsoft coco: Common objects in context / T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick // Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. — Springer. 2014. — С. 740—755.
3. *Ronneberger O.*, *Fischer P.*, *Brox T.* U-net: Convolutional networks for biomedical image segmentation // Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. — Springer. 2015. — С. 234—241.
4. CaRINA dataset: An emerging-country urban scenario benchmark for road detection systems / P. Y. Shinzato, T. C. dos Santos, L. A. Rosero, D. A. Ridel, C. M. Massera, F. Alencar, M. P. Batista, A. Y. Hata, F. S. Osório, D. F. Wolf // 2016 IEEE 19th international conference on intelligent transportation systems (ITSC). — IEEE. 2016. — С. 41—46.
5. Deep multispectral semantic scene understanding of forested environments using multimodal fusion / A. Valada, G. L. Oliveira, T. Brox, W. Burgard // International symposium on experimental robotics. — Springer. 2016. — С. 465—477.
6. Deep residual learning for image recognition / K. He, X. Zhang, S. Ren, J. Sun // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — С. 770—778.
7. *Milletari F.*, *Navab N.*, *Ahmadi S.-A.* V-net: Fully convolutional neural networks for volumetric medical image segmentation // 2016 fourth international conference on 3D vision (3DV). — Ieee. 2016. — С. 565—571.
8. 1 Year, 1000km: The Oxford RobotCar Dataset / W. Maddern, G. Pascoe, C. Linegar, P. Newman // The International Journal of Robotics Research (IJRR). — 2017. — Т. 36, № 1. — С. 3—15. — DOI: 10.1177/0278364916679498.

9. Focal loss for dense object detection / T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár // Proceedings of the IEEE international conference on computer vision. — 2017. — С. 2980—2988.

10. Scene parsing through ade20k dataset / B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — С. 633—641.

11. The mapillary vistas dataset for semantic understanding of street scenes / G. Neuhold, T. Ollmann, S. Rota Bulo, P. Kontschieder // Proceedings of the IEEE international conference on computer vision. — 2017. — С. 4990—4999.

12. Real-time semantic mapping for autonomous off-road navigation / D. Maturana, P.-W. Chou, M. Uenoyama, S. Scherer // Field and Service Robotics: Results of the 11th International Conference. — Springer. 2018. — С. 335—350.

13. A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments / M. Wigness, S. Eum, J. G. Rogers, D. Han, H. Kwon // 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). — IEEE. 2019. — С. 5000—5007.

14. *Meyer M.*, *Kuschk G.* Automotive radar dataset for deep learning based 3d object detection // 2019 16th european radar conference (EuRAD). — IEEE. 2019. — С. 129—132.

15. An image is worth 16x16 words: Transformers for image recognition at scale / A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly [и др.] // arXiv preprint arXiv:2010.11929. — 2020.

16. *Bochkovskiy A.*, *Wang C.-Y.*, *Liao H.-Y. M.* Yolov4: Optimal speed and accuracy of object detection // arXiv preprint arXiv:2004.10934. — 2020.

17. Mulran: Multimodal range dataset for urban place recognition / G. Kim, Y. S. Park, Y. Cho, J. Jeong, A. Kim // 2020 IEEE international conference on robotics and automation (ICRA). — IEEE. 2020. — С. 6246—6253.

18. nuscenes: A multimodal dataset for autonomous driving / H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — С. 11621—11631.

19. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather / M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, F. Heide // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2020. — C. 11682—11692.

20. The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset / D. Barnes, M. Gadd, P. Murcutt, P. Newman, I. Posner // Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). — Paris, 2020. — URL: https://arxiv.org/abs/1909.01300.

21. Carrada dataset: Camera and automotive radar with range-angle-doppler annotations / A. Ouaknine, A. Newson, J. Rebut, F. Tupin, P. Pérez // 2020 25th International Conference on Pattern Recognition (ICPR). — IEEE. 2021. — C. 5068—5075.

22. RadarScenes: A real-world radar point cloud data set for automotive applications / O. Schumann, M. Hahn, N. Scheiner, F. Weishaupt, J. F. Tilly, J. Dickmann, C. Wöhler // 2021 IEEE 24th International Conference on Information Fusion (FUSION). — IEEE. 2021. — C. 1—8.

23. Radiate: A radar dataset for automotive perception in bad weather / M. Sheeny, E. De Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, A. Wallace // 2021 IEEE International Conference on Robotics and Automation (ICRA). — IEEE. 2021. — C. 1—7.

24. Rellis-3d dataset: Data, benchmarks and analysis / P. Jiang, P. Osteen, M. Wigness, S. Saripalli. — 2021.

25. Segmentmeifyoucan: A benchmark for anomaly segmentation / R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, M. Rottmann // arXiv preprint arXiv:2104.14812. — 2021.

26. BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers / Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, J. Dai // Computer Vision – ECCV 2022 / под ред. S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, T. Hassner. — Cham : Springer Nature Switzerland, 2022. — C. 1—18.

27. Convolution neural networks for pothole detection of critical road infrastructure / A. K. Pandey, R. Iqbal, T. Maniak, C. Karyotis, S. Akuma, V. Palade // Computers and Electrical Engineering. — 2022. — T. 99. — C. 107725.

28. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries / Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, J. Solomon // Conference on Robot Learning. — PMLR. 2022. — С. 180—191.

29. *Huang J.*, *Huang G.* Bevpoolv2: A cutting-edge implementation of bevdet toward deployment // arXiv preprint arXiv:2211.17111. — 2022.

30. Masked-attention mask transformer for universal image segmentation / B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, R. Girdhar. — 2022.

31. Multi-class road user detection with 3+ 1d radar in the view-of-delft dataset / A. Palffy, E. Pool, S. Baratam, J. F. Kooij, D. M. Gavrila // IEEE Robotics and Automation Letters. — 2022. — Т. 7, № 2. — С. 4961—4968.

32. Multimodal token fusion for vision transformers / Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, Y. Wang // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2022. — С. 12186—12195.

33. *Murhij Y.*, *Golodkov A.*, *Yudin D.* Rethinking voxelization and classification for 3d object detection // International Conference on Neural Information Processing. — Springer. 2022. — С. 470—481.

34. *Murhij Y.*, *Yudin D.* Fmfnet: improve the 3d object detection and tracking via feature map flow // 2022 International Joint Conference on Neural Networks (IJCNN). — IEEE. 2022. — С. 1—8.

35. Orfd: A dataset and benchmark for off-road freespace detection / C. Min, W. Jiang, D. Zhao, J. Xu, L. Xiao, Y. Nie, B. Dai // 2022 international conference on robotics and automation (ICRA). — IEEE. 2022. — С. 2532—2538.

36. *Paek D.-H.*, *Kong S.-H.*, *Wijaya K. T.* K-radar: 4d radar object detection for autonomous driving in various weather conditions // Advances in Neural Information Processing Systems. — 2022. — Т. 35. — С. 3819—3829.

37. Petr: Position embedding transformation for multi-view 3d object detection / Y. Liu, T. Wang, X. Zhang, J. Sun // European conference on computer vision. — Springer. 2022. — С. 531—548.

38. Raw high-definition radar for multi-task learning / J. Rebut, A. Ouaknine, W. Malik, P. Pérez // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2022. — С. 17021—17030.

39. *Rukhovich D.*, *Vorontsova A.*, *Konushin A.* Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. — 2022. — С. 2397—2406.

40. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection / J. Park, C. Xu, S. Yang, K. Keutzer, K. Kitani, M. Tomizuka, W. Zhan // arXiv preprint arXiv:2210.02443. — 2022.

41. TJ4DRadSet: A 4D radar dataset for autonomous driving / L. Zheng, Z. Ma, X. Zhu, B. Tan, S. Li, K. Long, W. Sun, S. Chen, L. Zhang, M. Wan [и др.] // 2022 IEEE 25th international conference on intelligent transportation systems (ITSC). — IEEE. 2022. — С. 493—498.

42. Use of parallel ResNet for high-performance pavement crack detection and measurement / Z. Fan, H. Lin, C. Li, J. Su, S. Bruno, G. Loprencipe // Sustainability. — 2022. — Т. 14, № 3. — С. 1825.

43. *Zhang J.*, *Qian S.*, *Tan C.* Automated bridge surface crack detection and segmentation using computer vision-based deep learning model // Engineering Applications of Artificial Intelligence. — 2022. — Т. 115. — С. 105225.

44. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection / Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, Z. Li // Proceedings of the AAAI conference on artificial intelligence. Т. 37. — 2023. — С. 1477—1485.

45. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation / Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, S. Han // 2023 IEEE international conference on robotics and automation (ICRA). — IEEE. 2023. — С. 2774—2781.

46. Boreas: A multi-season autonomous driving dataset / K. Burnett, D. J. Yoon, Y. Wu, A. Z. Li, H. Zhang, S. Lu, J. Qian, W.-K. Tseng, A. Lambert, K. Y. Leung [и др.] // The International Journal of Robotics Research. — 2023. — Т. 42, № 1/2. — С. 33—42.

47. Cape: Camera view position embedding for multi-view 3d object detection / K. Xiong, S. Gong, X. Ye, X. Tan, J. Wan, E. Ding, J. Wang, X. Bai // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2023. — С. 21570—21579.

48. Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer / Y. Kim, S. Kim, J. W. Choi, D. Kum // Proceedings of the AAAI Conference on Artificial Intelligence. T. 37. — 2023. — C. 1160—1168.

49. Crn: Camera radar net for accurate, robust, efficient 3d perception / Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, D. Kum // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2023. — C. 17615—17626.

50. Depth estimation from camera image and mmwave radar point cloud / A. D. Singh, Y. Ba, A. Sarker, H. Zhang, A. Kadambi, S. Soatto, M. Srivastava, A. Wong // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2023. — C. 9275—9285.

51. Dinov2: Learning robust visual features without supervision / M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby [и др.] // arXiv preprint arXiv:2304.07193. — 2023.

52. Eva: Exploring the limits of masked visual representation learning at scale / Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, Y. Cao // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2023. — C. 19358—19369.

53. Exploring object-centric temporal modeling for efficient multi-view 3d object detection / S. Wang, Y. Liu, T. Wang, Y. Li, X. Zhang // Proceedings of the IEEE/CVF international conference on computer vision. — 2023. — C. 3621—3631.

54. Futr3d: A unified sensor fusion framework for 3d detection / X. Chen, T. Zhang, Y. Wang, Y. Wang, H. Zhao // proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2023. — C. 172—181.

55. Image as a foreign language: Beit pretraining for vision and vision-language tasks / W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som [и др.] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2023. — C. 19175—19186.

56. Inconseg: Residual-guided fusion with inconsistent multi-modal data for negative and positive road obstacles segmentation / Z. Feng, Y. Guo, D. Navarro-Alarcon, Y. Lyu, Y. Sun // IEEE Robotics and Automation Letters. — 2023. — T. 8, № 8. — C. 4871—4878.

57. Msc-rad4r: Ros-based automotive dataset with 4d radar / M. Choi, S. Yang, S. Han, Y. Lee, M. Lee, K. H. Choi, K.-S. Kim // IEEE Robotics and Automation Letters. — 2023. — Т. 8, № 11. — С. 7194—7201.

58. Mvfusion: Multi-view 3d object detection with semantic-aligned radar and camera fusion / Z. Wu, G. Chen, Y. Gan, L. Wang, J. Pu // 2023 IEEE International Conference on Robotics and Automation (ICRA). — IEEE. 2023. — С. 2766—2773.

59. Ntu4dradlm: 4d radar-centric multi-modal dataset for localization and mapping / J. Zhang, H. Zhuge, Y. Liu, G. Peng, Z. Wu, H. Zhang, Q. Lyu, H. Li, C. Zhao, D. Kircali [и др.] // 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). — IEEE. 2023. — С. 4291—4296.

60. One-peace: Exploring one general representation model toward unlimited modalities / P. Wang, S. Wang, J. Lin, S. Bai, X. Zhou, J. Zhou, X. Wang, C. Zhou // arXiv preprint arXiv:2305.11172. — 2023.

61. Oneformer: One transformer to rule universal image segmentation / J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, H. Shi // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2023. — С. 2989—2998.

62. Perception and sensing for autonomous vehicles under adverse weather conditions: A survey / Y. Zhang, A. Carballo, H. Yang, K. Takeda // ISPRS Journal of Photogrammetry and Remote Sensing. — 2023. — Т. 196. — С. 146—177.

63. Bevcar: Camera-radar fusion for bev map and object segmentation / J. Schramm, N. Vödisch, K. Petek, B. R. Kiran, S. Yogamani, W. Burgard, A. Valada // 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). — IEEE. 2024. — С. 1435—1442.

64. *Kubelka V.*, *Fritz E.*, *Magnusson M.* Do we need scan-matching in radar odometry? // 2024 IEEE International Conference on Robotics and Automation (ICRA). — IEEE. 2024. — С. 13710—13716.

65. *Li Y.*, *Yang Y.*, *Lei Z.* RCTrans: Radar-Camera Transformer via Radar Densifier and Sequential Decoder for 3D Object Detection // arXiv preprint arXiv:2412.12799. — 2024.

66. Oneformer3d: One transformer for unified point cloud segmentation / M. Kolodiazhnyi, A. Vorontsova, A. Konushin, D. Rukhovich // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2024. — С. 20943—20953.

67. RaCFormer: Towards High-Quality 3D Object Detection via Query-based Radar-Camera Fusion / X. Chu, J. Deng, G. You, Y. Duan, H. Li, Y. Zhang // arXiv preprint arXiv:2412.12725. — 2024.

68. RCBEVDet: Radar-camera Fusion in Bird's Eye View for 3D Object Detection / Z. Lin, Z. Liu, Z. Xia, X. Wang, Y. Wang, S. Qi, Y. Dong, N. Dong, L. Zhang, C. Zhu // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2024. — С. 14928—14937.

69. SpaRC: Sparse Radar-Camera Fusion for 3D Object Detection / P. Wolters, J. Gilg, T. Teepe, F. Herzog, F. Fent, G. Rigoll // arXiv preprint arXiv:2411.19860. — 2024.

70. Top-down beats bottom-up in 3d instance segmentation / M. Kolodiazhnyi, A. Vorontsova, A. Konushin, D. Rukhovich // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. — 2024. — С. 3566—3574.

71. Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception / P. Wolters, J. Gilg, T. Teepe, F. Herzog, A. Laouichi, M. Hofmann, G. Rigoll // arXiv preprint arXiv:2403.07746. — 2024.

72. RobuRCDet: Enhancing Robustness of Radar-Camera Fusion in Bird's Eye View for 3D Object Detection / J. Yue, Z. Lin, X. Lin, X. Zhou, X. Li, L. Qi, Y. Wang, M.-H. Yang // arXiv preprint arXiv:2502.13071. — 2025.

73. *Матыкина О.*, *Ветошкин Л.*, *Юдин Д.* Гибридный метод обнаружения препятствий и краев дорожного полотна для мультимодальной бортовой системы технического зрения // Труды МФТИ. — 2025. — С. 1—13.

74. CVAT: Computer Vision Annotation Tool. — `https://www.cvat.ai/`.

75. InstSnowRoadDetection Dataset. — `https://universe.roboflow.com/debigp/instsnowroaddetection/dataset/2`.

76. PotholeMix Dataset. — `https://data.mendeley.com/datasets/kfth5g2xk3/2`.

77. ROS 2 Foxy — Installation Guide. — `https://docs.ros.org/en/foxy/Installation.html`.

78. Soil-2 Dataset. — https://universe.roboflow.com/pier-paolo-ceccaranelli-pnffz/soil-2-ozzkk/dataset/1.

79. Soil-3 Dataset. — https://universe.roboflow.com/pier-paolo-ceccaranelli-pnffz/soil-3/dataset/1.

80. YOLOv11: Next-gen object detection. — https://yolo11.com/.