УДК 629.331.5 DOI: 10.18324/2077-5415-2025-2-167-174

Тенденции развития методов трехмерного обнаружения объектов

Н.С. Филатов[1, 2]

[1]Санкт-Петербургский политехнический университет Петра Великого, ул. Политехническая, 29, Санкт-Петербург, Россия
[2]ООО «Медицинские Скрининг Системы», ул. Жуковского, 63, Санкт-Петербург, Россия
filatov.n@edu.spbstu.ru
orcid: 0000-0002-0657-1256
Статья поступила 14.04.2025, принята 12.05.2025

Трехмерное обнаружение объектов на основе только камер стало ключевым компонентом современных систем автономного вождения, предлагая экономически эффективную и универсальную альтернативу решениям, использующим лидары. В данной работе представлен систематический аналитический обзор последних достижений в области трехмерного обнаружения объектов по данным видеокамеры для робототехники, в частности для автономного вождения. С использованием инновационного подхода выполнен анализ публикационной активности по четырем основным эталонным наборам данных – KITTI, NuScenes, Argoverse и Waymo, поисковая системы Google Scholar применена в сочетании с большой языковой моделью для фильтрации нерелевантных результатов поиска. Анализ показал четкое смещение популярности от набора KITTI к NuScenes, что отражает растущие потребности в более разнообразных и сложных дорожных сценариях. Современные методы трехмерного обнаружения объектов по данным камеры классифицированы на пять групп: основанные на геометрии, основанные на предсказании глубины, многоэтапные, одноэтапные и использующие последовательность кадров. Показано, что показатели точности для набора данных NuScenes достигли качества наилучших методов на основе лидаров 2022 г. Однако отмечено, что современные визуальные модели как правило обладают значительно большим числом параметров по сравнению с методами, использующими только лидар. Кроме того, отмечено, что мультимодальные подходы, объединяющие данные камер и лидаров, достигают баланса по числу параметров и стабильно превосходят методы, использующие единственный тип данных.

Ключевые слова: монокулярное трехмерное обнаружение объектов; автономное вождение; большая языковая модель.

Trends in the development of 3D object detection methods

N.S. Filatov[1, 2]

[1]Peter the Great St. Petersburg Polytechnic University; 29, Polytechnicheskaya St., St. Petersburg, Russia
[2] "Medical Screening Systems", LLC; 63, Zhukovsky St., St. Petersburg, Russia
filatov.n@edu.spbstu.ru
Received 14.04.2025, accepted 12.05.2025

Camera-only 3D object detection has become a key component of modern autonomous driving systems, offering a cost-effective and versatile alternative to lidar-based solutions. The paper presents a systematic analytical review of recent advances in 3D object detection from video camera data for robotics, in particular for autonomous driving. Using an innovative approach, the publication activity on four major benchmark datasets - KITTI, NuScenes, Argoverse and Waymo - is analyzed, and the Google Scholar search engine combined with a large language model is applied to screen out irrelevant search results. The analysis shows a clear shift in popularity from the KITTI to NuScenes, reflecting the growing need for more diverse and complex motion scenarios. Current methods for 3D object detection from camera data are categorized into five groups: geometry-based, depth prediction-based, multi-stage, single-stage, and frame sequence-based. It is shown that the accuracy performance for the NuScenes dataset has reached the quality of the best 2022 lidar-based methods. However, it is noted that the state-of-the-art visual models have significantly more parameters compared to the lidar-only methods. Also, these state-of-the-art vision models typically possess larger parameter counts than lidar-only solutions. Furthermore, it is observed that multimodal approaches combining camera and lidar data achieve a balance in the number of parameters and consistently outperform methods using a single data type.

Keywords: monocular 3d object detection; autonomous driving; large language model.

Введение. Монокулярное трехмерное обнаружение объектов представляет собой задачу предсказания трехмерных ограничивающих рамок для объектов (например, транспортных средств или пешеходов) по единственному RGB-изображению. В контексте автономного вождения это означает, что по одному кадру камеры система должна определить пространственное положение, размеры и ориентацию каждого объекта в реальном мире. Эта задача по своей природе сложна, поскольку единственное двумерное изображение не содержит прямой информации о глубине, а проблема является некорректно поставленной ввиду неопределенностей масштаба и расстояния. Несмотря на указанные трудности, монокулярное трехмерное обнаружение объектов привлекает пристальное внимание научного сообщества, поскольку камеры являются не-

дорогими, легкими, обеспечивают богатую семантическую информацию (например, внешний вид, цвет и текстуру объектов) и широко распространены на современных транспортных средствах. По мере развития технологий трехмерного восприятия исследователи все чаще опираются на масштабные открытые наборы данных для обучения тестирования своих моделей. Анализ публикационной активности, выполненный в данной работе, показывает снижение доминирования набора данных KITTI в пользу более масштабных и разнообразных наборов данных, таких как nuScenes и Waymo, которые включают широкий спектр условий движения, погоды и классов объектов.

В настоящей работе представлен анализ публикаций, посвященных монокулярному трехмерному обнаружению объектов на основных наборах данных для автономного вождения. Анализ выполнен с использованием поисковой системы Google Scholar и фильтрации на основе большой языковой модели (БЯМ), что позволило эффективно исключить нерелевантные результаты поискового запроса. Далее систематически рассмотрены современные методы обнаружения, проведено разделение на пять категорий: основанные на геометрии, основанные на предсказании глубины, многоэтапные, одноэтапные и использующие последовательность кадров. Анализируя показатели качества на наборе данных NuScenes, отмечено, что методы, использующие только данные видеокамеры, достигли показателей качества наилучших систем, применяющих лидар 2022 г. Дополнительно проводится сравнение количества обучаемых параметров для методов, использующих различные датчики.

Эталонные наборы данных. Набор данных KITTI Vision Benchmark (2012/2013) [1] был первым, предоставившим аннотации трехмерных ограничивающих рамок для изображений, и в течение долгого времени являлся основным эталоном для оценки методов монокулярного трехмерного обнаружения. KITTI содержит 7481 обучающее изображение и 7518 тестовых изображений, снятых фронтальной камерой в г. Карлсруэ (Германия), и сопровождается соответствующими облаками точек, полученными с помощью лидара. Основное внимание уделено трем классам объектов (автомобили, пешеходы, велосипедисты) в относительно простых городских условиях при ясной погоде в дневное время. Будучи первым крупным открытым набором данных, он приобрел высокую популярность, однако KITTI имеет ограниченные масштабы и разнообразие условий (один город, одна камера, около 15 тыс. кадров), что стимулировало появление новых, более сложных наборов данных.

Набор данных NuScenes (2019) [2], разработанный компаниями Aptiv/Motional, предоставляет панорамный обзор вокруг автомобиля с помощью шести камер, а также данные лидара и радара. Он содержит 1000 сцен продолжительностью по 20 секунд каждая, снятых в Бостоне и Сингапуре, всего около 40 тыс. ключевых кадров с изображениями и аннотациями трехмерных объектов. Важным преимуществом NuScenes является разнообразие условий съемки: городские и загородные сцены, ночное время и дождливая погода, а также более широкий список классов объектов (автомобили, грузовики, автобусы, пешеходы, мотоциклы и другие). Аннотации объектов включают дополнительные атрибуты, такие как состояние объекта (например, припаркован или движется) и информация о скорости. Благодаря своей сложности и масштабу NuScenes быстро стал стандартным набором данных для оценки устойчивости методов монокулярного трехмерного обнаружения в сложных условиях.

Набор данных Waymo Open Dataset [3], выпущенный в 2019 г. компанией Waymo, отличается еще большим масштабом, чем предыдущие. В нем около 200 тыс. кадров с аннотациями, полученных с пяти камер и лидара, охватывающих панорамный обзор вокруг автомобиля. Набор включает три класса объектов (автомобили, пешеходы, велосипедисты), представленные в разнообразных городских и пригородных условиях и при различном освещении. Основное преимущество Waymo – большой объем данных, порядка 12 млн размеченных трехмерных рамок, что позволяет обучать модели, требующие большого количества данных. Несмотря на это, публикаций с оценкой на Waymo меньше, чем на KITTI или NuScenes, что объясняется высокой вычислительной сложностью набора данных.

Набор данных Argoverse (2019) [4] содержит сцены, записанные в Майами и Питтсбурге. Первая версия Argoverse включает более 30 тыс. изображений с трехмерными аннотациями и расширенную классификацию объектов (15 классов). Особое внимание уделяется дальнему обнаружению объектов (до 150 м).

Другие наборы данных также внесли вклад в развитие исследований. Среди них ApolloScape (ApolloCar3D), предоставляющий большое количество изображений с аннотациями только для автомобилей, Lyft Level 5 и Honda 3D (H3D), мультимодальные наборы с аннотациями объектов, а также менее распространенные A*3D, Cityscapes-3D и расширенный набор KITTI-360. Эти наборы данных, хотя и менее популярны, дают дополнительные возможности для исследований, ориентированных на специфические задачи (например, плотное городское движение или использование камер типа «рыбий глаз»). В заключение стоит отметить, что наборы данных KITTI, NuScenes и Waymo остаются наиболее распространенными и важными для оценки и развития методов монокулярного трехмерного обнаружения объектов.

Анализ публикационной активности. Анализ публикационной активности по теме позволяет оценить ее актуальность, определить тренд развития и выявить наиболее востребованные наборы данных среди исследователей. С этой целью был проведен систематический анализ публикаций с 2014 по 2024 гг., в анализ включены публикации, которые относятся к созданию новых методов трехмерного обнаружения объектов с использованием только видеокамер, которые были протестированы хотя бы на одном наборе данных из списка: Argoverse [4], NuScenes [2], KITTI [1], Waymo [3]. Статьи были найдены с использованием базы данных Google Scholar и далее для обеспечения точности проводимого анализа разделены на релевантные и не релевантные с помощью большой языковой модели. Общая схема процесса сбора данных приведена на рис. 1.

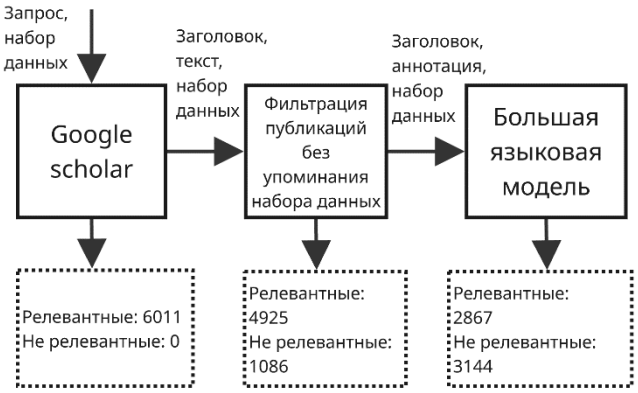

Рис. 1. Схема подготовки данных для анализа публикационной активности

Поиск по базе данных Google Scholar осуществлялся по запросам «monocular 3d object detection» и «multi-iview object detection», дополнительно к каждому запросу, согласно синтаксису Google Scholar, добавлялось требование включения каждого из рассматриваемых наборов данных, пример комбинированного запроса: «multi-view object detection AND Argoverse». Далее производилось удаление дубликатов и результатов поиска, которые не содержали упоминания целевого набора данных в тексте.

Однако далеко не все из полученных публикаций могли считаться относящимися к теме, поскольку в результаты поиска могли также попадать разнообразные обзоры, мультимодальные подходы, или вообще статьи, не связанные с обнаружением объектов. Также содержание названия целевого набора данных в тексте статьи не обязательно означает, что набор данных использовался в данной работе, а не был приведен в рамках литературного обзора. Таким образом, для заключения о релевантности публикации к теме монокулярного трехмерного обнаружения объектов в контексте определенного набора данных необходимо проводить логически суждения на основе названия публикации и ее аннотации, для автоматизации этого процесса была использована большая языковая модель, в частности использовалась модель Gemma 3 [5], содержащая 14 млрд параметров. Перед большой языковой моделью была поставлена следующая задача: по названию, аннотации статьи и целевому набору данных определить, является ли статья релевантной, описание задачи также содержало следующие критерии отбора:

– статья не должна быть обзорной;
– статья должная быть посвящена трехмерному обнаружению объектов, а не другим задачам, таким как навигация, реконструкция и т. п.;
– методы, предлагаемые в статье, должны использовать только видеокамеру, методы, использующие LiDAR или комбинацию датчиков, не допускаются.
– результаты работы должны быть протестированы на целевом наборе данных, в случае если указано, что тестирование проводилось на нескольких наборах данных.

После завершения обработки данных большой языковой моделью из 4925 исследований с прошлого этапа обработки 2867 исследований были отобраны как релевантные. Данные результаты прошли выборочный контроль: среди обработанных статей случайным образом были выбраны 200, релевантность которых независимо оценил человек, совпадение с результатами модели составило 96 %, что позволяет сделать заключение о высоком качестве обработки данных, пригодном для выполнения анализа (рис. 2).

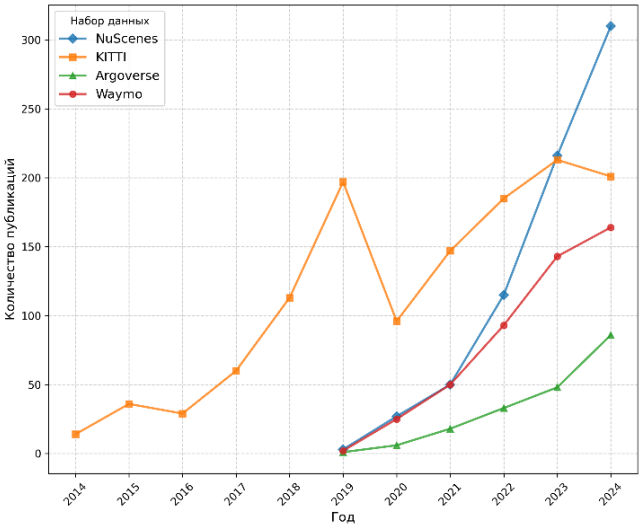

Рис. 2. Анализ публикационной активности для работ, связанных с монокулярным трехмерным обнаружением объектов

Результаты анализа позволяют заключить, что интерес исследователей к монокулярному трехмерному обнаружению объектов возрастает, и наиболее популярным набором данных в 2024 г. является NuScenes. Также видно, что набор данных KITTI, который появился значительно раньше других, уже не имеет выраженной тенденции к росту публикаций, связанных с ним.

Методы, основанные на геометрии. Методы монокулярного трехмерного обнаружения объектов на основе только камер подразделяются на геометрические, основанные на глубоком обучении, и гибридные подходы. Геометрические подходы используют проективную геометрию и априорное знание о форме объектов для извлечения информации о трехмерных объектах из изображений. Ранние методы, такие как Mono3D++ [6], опирались на геометрические предположения о плоской поверхности дороги и средних размерах объектов, что позволяло генерировать трехмерные рамки и оценивать их качество на основе двумерных признаков, например сегментации. В аналогичной работе [7] применяется гибридный подход, совмещающий глубокое обучение и геометрию: сначала сверточная нейронная сеть (CNN) предсказывает трехмерные размеры и ориентацию объекта, после чего его трехмерное положение рассчитывается путем минимизации ошибки перепроецирования. В таких подходах обеспечивается геометрически согласованная шестистепенная оценка позиции объекта без явного предсказания глубины.

Также предложены методы, использующие априорное знание об объектах в виде их трехмерных моделей. В методе Deep MANTA [8] двуступенчатая нейронная сеть предсказывает как положение объектов на двух-

мерном изображении, так и атрибуты и ориентацию трехмерных частей автомобилей, которые затем используются для размещения трехмерной модели в пространстве и получения итоговых предсказаний. Кроме того, исследовано применение трехмерных моделей для оценки ориентации объектов в трехмерном пространстве по двумерным изображениям [9].

Геометрические методы показали, что использование физических ограничений, таких как плоскость дороги и типичные размеры объектов, позволяют решать задачу трехмерного обнаружения объектов по единственному изображению.

Методы, основанные на предсказании глубины. Подходы, основанные на оценке глубины, предполагают явное вычисление промежуточного геометрического представления, такого как карта глубины, облако трехмерных точек. Основная идея заключается в том, что при точной оценке глубины изображение можно преобразовать в трехмерное пространство, в котором проще производить трехмерное обнаружение объектов.

Одним из классических примеров является подход Pseudo-LiDAR [10], в котором из одного изображения (или стереопары) сначала оценивается плотная карта глубины, после чего она проецируется обратно в трехмерное облако точек, имитирующее данные с лидара. Затем полученное облако («псевдо-лидар») анализируется методами, разработанными для работы с настоящими облаками точек. Несмотря на значительный прирост точности, который демонстрирует этот подход, качество итоговых результатов существенно зависит от точности оценки карты глубины.

Другим направлением стали методы, которые вместо прямого формирования облака точек преобразуют изображения в карту признаков, кодирующих глубины в формате BEV («вид сверху»). Ключевым примером является подход Lift, Splat, Shoot (LSS) [11], предполагается, что каждый пиксель изображения имеет распределение вероятностей по возможным значениям глубины. Затем признаки каждого пикселя «поднимаются» в трехмерную колонку, взвешенную в соответствии с этим распределением, и объединяются на единой плоскости BEV. Полученная таким образом карта признаков обрабатывается стандартными двумерными сверточными нейросетями для обнаружения объектов и сегментации. Аналогичных подходом является CaDDN [12], в котором производится предсказание категориального распределения глубины для каждого пикселя. Основная идея CaDDN состоит в том, что обучение сети на основе дискретных распределений глубины, полученных из редких данных лидара, обеспечивает четкое сосредоточение распределений на истинной глубине, одновременно сохраняя возможность моделировать неопределенность.

Таким образом, методы, основанные на глубине, представляют собой интуитивно понятные подходы решения задачи трехмерного обнаружения объектов, в которых инновации достигаются в способах моделирования и предсказания глубины по двухмерному изображению.

Многоэтапные методы. Многоэтапные методы трехмерного обнаружения объектов разбивают задачу на последовательные этапы, обычно состоящие из стадии двумерного обнаружения с последующей стадией трехмерной локализации. Такой подход объясняется тем, что задача двумерного обнаружения объектов уже достаточно хорошо изучена, и после локализации объекта на изображении пространство поиска для трехмерной оценки существенно сокращается.

Одним из примеров таких подходов является ROI-10D [13], который совмещает модели для оценки глубины, для двухмерного обнаружения объектов и регрессионную для оценки трехмерных параметров объекта в двухступенчатую архитектуру. На первом этапе формируются двумерные области интереса (ROI), а на втором уже для каждой области оцениваются размеры, ориентация и удаленность объекта. Важной особенностью ROI-10D является дифференцируемая проекция трехмерной рамки на изображение, что обеспечивает согласованность между результатами двумерного и трехмерного обнаружения и позволяет предварительно обучить сеть на больших наборах данных с двумерными аннотациями.

Другой пример данной категории методов – MonoPSR [14], в котором на основе готовых двумерных обнаружений и используя геометрию камеры, формируются грубые трехмерные предположения наличия объектов. Эти предположения существенно сокращают пространство поиска для последующего уточнения трехмерного положения и формы объекта. Отличительной чертой метода является одновременная оптимизация трехмерной ограничивающей рамки и восстановление облака точек, описывающее форму объекта в нормализованных координатах.

В общем случае многоступенчатые подходы позволяют эффективно использовать специализированную обработку на каждом этапе: на первом этапе решается хорошо изученная задача определения положения объектов в двумерном пространстве, а на втором – оценки расстояния и ориентации в трехмерном пространстве.

Одноэтапные методы. В одноступенчатых методах трехмерного обнаружения объектов производится попытка напрямую предсказывать трехмерные ограничивающие рамки из изображений за один проход нейросети, без этапов предсказания кандидатов объектов и их дополнительного уточнения. Среди них выделяются методы без заранее заданных шаблонов объектов, основанные на ключевых точках [15], в которых нейросеть одновременно выявляет центр объекта на изображении и производит регрессию его трехмерных параметров (размер, ориентацию). Другой пример – FCOS3D [16], расширяющий одноэтапный двумерный детектор FCOS на трехмерное обнаружение объектов. В работе [17] показано, что в переход от метода двухмерного обнаружения объектов к трехмерному обнаружению может быть достигнут добавлением всего одной функции потерь. Преимуществом таких методов является высокая скорость обработки и простота обучения благодаря отсутствию промежуточных этапов. Однако они должны неявно обучаться геометрическим зависимостям, что усложняет процесс обучения.

В области систем трехмерного обнаружения объектов по данным с камер по периметру автомобиля получили развитие одноступенчатые подходы, основанные на архитектуре трансформеров, такие как DETR3D [18]

и PETR [19]. DETR3D использует обучаемые токены, представляющие собой потенциальные трехмерные объекты, и с помощью механизма внимания агрегирует признаки с разных камер, избегая промежуточных этапов оценки глубины. PETR улучшает этот подход, кодируя позиционные эмбеддинги изображения с учетом их трехмерных координат на основе параметров камеры. Такой подход помогает нейросети лучше учитывать пространственный контекст, повышая точность обнаружения.

Важной проблемой одноступенчатых методов является неоднозначность глубины и появление ложных дубликатов объектов на разных расстояниях. Интересным решением этой проблемы является метод RayDN [20], который вводит в процесс обучения синтетические негативные примеры на разных расстояниях вдоль луча зрения камеры, заставляя модель лучше выучивать признаки оценки глубины и устранять дубликаты.

Таким образом, одноступенчатые методы обеспечивают простоту и высокую производительность, напрямую выучивая связи между изображением и трехмерным положением объектов.

Методы, использующие последовательность кадров. Методы с использованием временных признаков расширяют возможности монокулярного трехмерного обнаружения объектов за счет обработки последовательностей кадров, улучшая точность и устойчивость оценки объектов благодаря использованию информации из нескольких моментов времени.

В одном из таких подходов (StreamPETR [21]), используется временная передача признаков объектов между кадрами на основе объектных запросов. Это позволяет учитывать информацию о положении и движении объекта в предыдущих кадрах и стабилизировать его обнаружение в текущем кадре, снижая количество ложных обнаружений и пропусков. В результате метод значительно приближается по точности к системам на основе лидара.

Другой подход – VideoBEV [22] – реализует временную агрегацию признаков на основе рекуррентного объединения карт признаков вида сверху (BEV). Вместо обработки кадров по отдельности метод использует скрытое состояние, накапливающее информацию о сцене за несколько кадров, и обновляет его рекурсивно, что позволяет эффективно использовать временные признаки, ограничивая рост вычислительной сложности.

Более комплексный подход Sparse4D v3 [23] интегрирует задачи трехмерного обнаружения и сопровождения объектов в единой трансформерной архитектуре. Предлагается обработка последовательности кадров совместно, используя объектные запросы, которые агрегируют пространственно-временную информацию и представляют траектории объектов. Этот подход включает дополнительные обучающие задачи, такие как временное шумоподавление и оценка качества предсказаний, что делает обучение более устойчивым. В результате совместной оптимизации обнаружения и сопровождения объектов, улучшаются показатели качества для каждой из задач, в частности для удаленных объектов.

В результате совместной оптимизации обнаружения и сопровождения объектов улучшаются показатели качества для каждой из задач, в частности для удаленных объектов. Исследования также подтверждают, что регистрация траекторий движения транспортных средств на основе анализа последовательностей изображений и использования алгоритмов, таких как фильтр Калмана и оптический поток, способна значительно повысить устойчивость обнаружения и сопровождения объектов в условиях сложной дорожной обстановки и изменяющихся условий освещения [24]. Также исследуются методы для аналитического вычисления расстояния до объекта, при наличии нескольких изображений объекта в движении, что может быть применено для уточнения результатов работы прочих методов [25].

Все представленные временные подходы показывают, что эффективное использование информации о времени значительно повышает надежность и точность трехмерного обнаружения объектов на основе только видеокамер, существенно сокращая разрыв в производительности с системами, использующими лидары.

Анализ показателей качества. Показатели качества методов трехмерного обнаружения объектов, использующих только камеры, как правило, ниже показателей методов, применяющих лидар или и камеру и лидар. На рис. 3 представлена динамика развития методов, использующих различные датчики. Из данных следует, что показатели качества наилучших методов сильно возросли за интервал в несколько лет, при этом как правило методы, применяющие камеру, действительно в большинстве своем менее точны, чем другие, и, аналогично, мультимодальные методы, как правило, точнее методов, использующих единственную модальность. Однако также можно отметить, что методы, применяющие только видеокамеру, на данный момент имеют довольно высокую метрику mAP на уровне лучших методов, использующих лидар в 2022 г.

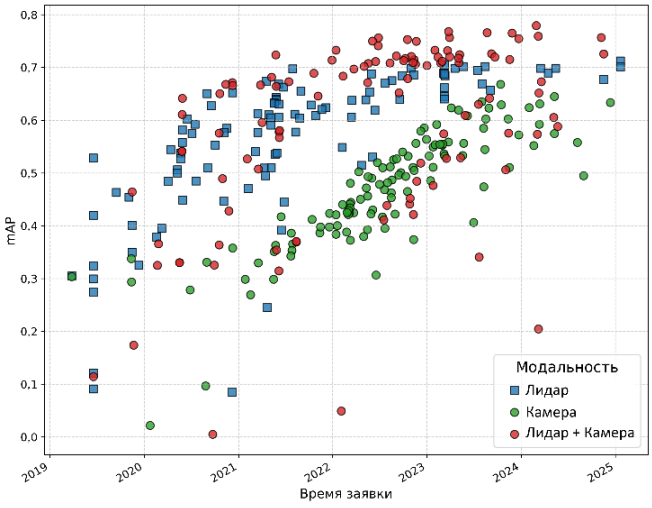

Рис. 3. График метрики mAP обнаружения объектов на наборе данных NuScenes для методов, использующих различные датчики

Аналогичное сравнение нескольких рассмотренных методов для обнаружения объектов по видеокамере в сравнении с лучшими методами из других категорий с открытым исходным кодом представлено в табл. 1.

Таблица 1. Показатели качества трехмерного обнаружения объектов на наборе данных NuScenes

Метод	Модальность	mAP	Количество параметров, млн	Год
FCOS3D [16]	C	0,358	55,06	2020
DETR3D [18]	C	0,412	78,67	2021
StreamPETR [22]	C	0,620	84,35	2024
RayDN [20]	C	0,631	334,27	2024
FocalFormer3D [26]	L	0,690	21,35	2023
IS-Fusion [27]	C+L	0,765	48,37	2024

Примечание. C – камера, L – лидар.

Используя открытый исходный код, для ряда методов было подсчитано количество обучаемых параметров. Методы обнаружения объектов только по видеокамере имеют довольно высокое количество обучаемых параметров из-за необходимости обрабатывать изображения с шести камер и решать неопределенность проецирования трехмерных объектов на двухмерные изображения. Напротив, минимальное количество обучаемых параметров в табл. 1 имеет метод, использующий лидар. Мультимодальный метод IS-Fusion [27] имеет наивысшую метрику качества и количество параметров больше, чем у методов использующих лидар, но меньше, чем у методов применяющих камеры. Это можно объяснить тем, что данные датчики являются взаимодополняющими, и в мультимодальной конфигурации можно использовать меньше обучаемых параметров на обработку изображений, поскольку уже много информации о трехмерных характеристиках объектов доступно из облака точек лидара.

Заключение. Показатели качества методов трехмерного обнаружения объектов, использующих только камеры, как правило ниже методов применяющих лидар или и камеру и лидар. На рис. 3 представлена динамика развития методов, использующих различные датчики. Из данных следует, что показатели качества наилучших методов сильно возросли за интервал в несколько лет, при этом, как правило, методы, использующие камеру действительно обычно менее точны чем другие методы, и, аналогично, мультимодальные методы в большинстве своем точнее методов, применяющих единственную модальность. Однако также можно отметить, что методы, использующие только видеокамеру, на данные момент, имеют довольно высокую метрику mAP на уровне лучших методов, применяющих лидар в 2022 г.

Таким образом, проведенный обзор подчеркивает стремительное развитие области монокулярного трехмерного обнаружения объектов для автономного вождения и робототехники. Это развитие во многом обусловлено переходом от крупным и более разнообразным наборам данных. Достигнутые успехи в точности позволяют говорить о возрастающей возможности использования визуальных систем в условиях экономически ограниченных ресурсов. В то же время значительно большее количество параметров в передовых методах, использующих только данные камеры, указывает на важность дальнейшего повышения их вычислительной эффективности.

Выполненный анализ публикационной активности с использованием Google Scholar и БЯМ подтвердил рост интереса научного сообщества к монокулярным методам трехмерного обнаружения объектов. Также была представлена систематическая классификация существующих методов на пять категорий, что позволяет лучше понять текущие достижения и направления развития. В перспективе можно предположить проведение ряда работ, направленных на повышение вычислительной эффективности и развития более продвинутых методов использования временной информации. Данный обзор призван помочь исследователям и разработчикам ориентироваться в динамично развивающейся области и способствовать созданию более надежных и доступных систем восприятия для автономных роботов.

Литература

1. Geiger A., Lenz P., Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite // 2012 IEEE conference on computer vision and pattern recognition. June 16-21 2012. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2012. P. 3354-3361.
2. Caesar H., Bankiti V., Lang A.H, Vora S., Liong V.E., Xu Q., Krishnan A., Pan Y., Baldan G., Beijbom O. nuscenes: A multimodal dataset for autonomous driving. [Electronic resource]. arXiv preprint arXiv:1903.11027 (date of address 19.02.2025).
3. Sun P., Kretzschmar H., Dotiwalla X., Chouard A., Patnaik V., Tsui P., Guo J., Zhou Y., Chai Y., Caine B., Vasudevan V., Han W., Ngiam J., Zhao H., Timofeev A., Ettinger S., Krivokon M., Gao A., Joshi A., Zhao S., Cheng S., Zhang Y., Shlens J., Chen Z., Anguelov D. Scalability in perception for autonomous driving: Waymo open dataset // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2020. P. 2446-2454.
4. Chang M.-F., Lambert J., Sangkloy P., Singh J., Bak S., Hartnett A., Wang D., Carr P., Lucey S., Ramanan D. Argoverse: 3d tracking and forecasting with rich maps // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2019. P. 8748-8757.
5. Team G., Kamath A., Ferret J., Pathak S., Vieillard N., Hussenot L. Gemma 3 Technical Report. [Electronic resource]. arXiv:2503.19786. arXiv, 2025 (date of address 03.03.2025).
6. He T., Soatto S. Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors // Proceedings of the AAAI Conference on Artificial Intelligence. 2019. Vol. 33, № 01. P. 8409-8416.
7. Mousavian A., Anguelov D., Flynn J., Kosecka J. 3d bounding box estimation using deep learning and geometry // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017. P. 7074-7082.
8. Chabot F., Chaouch M., Rabarisoa J., Teuliere C., Chateau T. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2017. P. 2040-2049.

9. Зубов И.Г. Оценка ракурса транспортного средства, полученного монокулярной камерой // Цифровая обработка сигналов и ее применение. 2019. № 31 P. 381-338.
10. Wang Y., Chao W.-L., Garg D., Hariharan B., Campbell M., Weinberger K.Q. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2019. P. 8445-8453.
11. Philion J., Fidler S. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D // Computer Vision – ECCV 2020. Vol. 12359. Cham: Springer International Publishing, 2020. P. 194-210.
12. Reading C., Harakeh A., Chae J., Waslander S.L. Categorical depth distribution network for monocular 3d object detection // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2021. P. 8555-8564.
13. Manhardt F., Kehl W., Gaidon A. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2019. P. 2069-2078.
14. Ku J., Pon A.D., Waslander S.L. Monocular 3d object detection leveraging accurate proposals and shape reconstruction // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2019. P. 11867-11876.
15. Liu Z., Wu Z., Tóth R. Smoke: Single-stage monocular 3d object detection via keypoint estimation // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2020. P. 996-997.
16. Wang T., Zhu X., Pang J., Lin D. Fcos3d: Fully convolutional one-stage monocular 3d object detection // Proceedings of the IEEE/CVF international conference on computer vision. 2021. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2021. P. 913-922.
17. Лупенко Н.Л., Богуш Р.П., Чен Х. Анализ методов определения абсолютного расстояния до объекта по изображению с одной видеокамеры с использованием нейронных сетей // Вестн. Полоцк. гос. Университета. Сер. С. Фундаментальные науки. 2024. Vol. 42, № 2. P. 24-33.
18. Wang Y., Guizilini V.C., Zhang T., Wang Y., Zhao H., Solomon J. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries // Conference on Robot Learning. Proceedings of Machine Learning Research, 2022. Auckland, New Zealand: Proceedings of Machine Learning Research, 2022. P. 180-191.
19. Liu Y., Wang T., Zhang X., Sun J. PETR: Position Embedding Transformation for Multi-view 3D Object Detection // Computer Vision – ECCV 2022. Vol. 13687. Cham: Springer Nature Switzerland, 2022. P. 531-548.
20. Liu F., Huang T., Zhang Q., Yao H., Zhang C., Wan F., Ye Q., Zhou Y. Ray Denoising: Depth-aware Hard Negative Sampling for Multi-view 3D Object Detection–Supplementary Material // European Conference on Computer Vision (ECCV). 2024. Vol 1. P. 190-209.
21. Wang S., Liu Y., Wang T., Li Y., Zhang X. Exploring object-centric temporal modeling for efficient multi-view 3d object detection // Proceedings of the IEEE/CVF international conference on computer vision. 2023. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2023. P. 3621-3631.
22. Han C., Yang J., Sun J., Ge Z., Dong R., Zhou H., Mao W., Peng Y., Zhang X. Exploring recurrent long-term temporal fusion for multi-view 3d perception // IEEE Robotics and Automation Letters. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2024. P. 6544-6551.
23. Lin X., Pei Z., Lin T., Huang L., Su Z. Sparse4D v3: Advancing End-to-End 3D Detection and Tracking. [Electronic resource]. arXiv:2311.11722 (date of address 10.03.2025).
24. Юдин Д.А., Горшкова Н.Г., Кныш А.С., Фролов С.В. Распознавание транспортных средств и регистрация их траектории движения на последовательности изображений // Вестн. Белгород. гос. технологич. ун-та. 2016. № 6. С. 139-148.
25. Маньшин И.М., Красноперов Н.С. Определение расстояния до объекта с помощью монокулярного зрения // Научный аспект. 2023. № 1. С. 538.
26. Chen Y., Yu Z., Chen Y., Lan S., Anandkumar A., Jia J., Alvarez J.M. Focalformer 3d: focusing on hard instance for 3d object detection // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2023. P. 8394-8405.
27. Yin J., Shen J., Chen R., Li W., Yang R., Frossard P., Wang W. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2024. P. 14905-14915.

References

1. Geiger A., Lenz P., Urtasun R. Are we ready for autonomous driving? The kitti vision benchmark suite // 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012. P. 3354–3361.
2. Caesar H., Bankiti V., Lang A.H, Vora S., Liong V.E., Xu Q., Krishnan A., Pan Y., Baldan G., Beijbom O. nuscenes: A multimodal dataset for autonomous driving. [Electronic resource]. arXiv preprint arXiv:1903.11027 (date of address 19.02.2025).
3. Sun P. et al. Scalability in perception for autonomous driving: Waymo open dataset // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. P. 2446-2454.
4. Chang M.-F., Lambert J., Sangkloy P., Singh J., Bak S., Hartnett A., Wang D., Carr P., Lucey S., Ramanan D. Argoverse: 3d tracking and forecasting with rich maps // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2019. P. 8748-8757.
5. Team G., Kamath A., Ferret J., Pathak S., Vieillard N., Hussenot L. Gemma 3 Technical Report. [Electronic resource]. arXiv:2503.19786. arXiv, 2025 (date of address 03.03.2025).
6. He T., Soatto S. Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors // Proceedings of the AAAI Conference on Artificial Intelligence. 2019. Vol. 33, № 1. P. 8409-8416.
7. Mousavian A., Anguelov D., Flynn J., Kosecka J. 3d bounding box estimation using deep learning and geometry // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017. P. 7074-7082.
8. Chabot F. et al. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. P. 2040-2049.
9. Zubov I.G. Estimation of a vehicle foreshortening, obtained by a monocular camera // Digital Signal Processing and its Applications. 2019. № 31 P. 381-338.
10. Wang Y. et al. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driv-

ing // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. P. 8445–8453.
11. Philion J., Fidler S. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D // Computer Vision – ECCV 2020 Cham: Springer International Publishing, 2020. Vol. 12359. P. 194-210.
12. Reading C., Harakeh A., Chae J., Waslander S.L. Categorical depth distribution network for monocular 3d object detection // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2021. P. 8555-8564.
13. Manhardt F., Kehl W., Gaidon A. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape // Proceedings of the IEEE/CVF conference on Computer vision and pattern Recognition. 2019. P. 2069-2078.
14. Ku J., Pon A.D., Waslander S.L. Monocular 3d object detection leveraging accurate proposals and shape reconstruction // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. P. 11867-11876.
15. Liu Z., Wu Z., Tóth R. Smoke: Single-stage monocular 3d object detection via keypoint estimation // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020. P. 996–997.
16. Wang T., Zhu X., Pang J., Lin D. Fcos3d: Fully convolutional one-stage monocular 3d object detection // Proceedings of the IEEE/CVF international conference on computer vision. 2021. P. 913-922.
17. Lupenko N.L., Bogush R.P., Chen X. Analysis of methods for determining the absolute distance to the object from the image from one video camera using neural networks // Bulletin of Polotsk State University. Ser. C. Fundamental Sciences. 2024. Vol. 42, № 2. P. 24-33.
18. Wang Y., Guizilini V.C., Zhang T., Wang Y., Zhao H., Solomon J. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries // Conference on Robot Learning. Proceedings of Machine Learning Research, 2022. Auckland, New Zealand: Proceedings of Machine Learning Research, 2022. P. 180-191.
19. Liu Y., Wang T., Zhang X., Sun J. PETR: Position Embedding Transformation for Multi-view 3D Object Detection // Computer Vision – ECCV 2022. Vol. 13687. Cham: Springer Nature Switzerland, 2022. P. 531-548.
20. Liu F., Huang T., Zhang Q., Yao H., Zhang C., Wan F., Ye Q., Zhou Y. Ray Denoising: Depth-aware Hard Negative Sampling for Multi-view 3D Object Detection–Supplementary Material // European Conference on Computer Vision (ECCV). 2024. Vol 1. P 190-209.
21. Wang S., Liu Y., Wang T., Li Y., Zhang X. Exploring object-centric temporal modeling for efficient multi-view 3d object detection // Proceedings of the IEEE/CVF international conference on computer vision. 2023. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2023. P. 3621-3631.
22. Han C., Yang J., Sun J., Ge Z., Dong R., Zhou H., Mao W., Peng Y., Zhang X. Exploring recurrent long-term temporal fusion for multi-view 3d perception // IEEE Robotics and Automation Letters. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2024. P. 6544 - 6551.
23. Lin X., Pei Z., Lin T., Huang L., Su Z. Sparse4D v3: Advancing End-to-End 3D Detection and Tracking. [Electronic resource]. arXiv:2311.11722 (date of address 10.03.2025).
24. Yudin D.A., Gorshkova N.G., Knysh A.S., Frolov S.V. Recognition of vehicles and registration of their trajectory on the sequence of images // Bulletin of Belgorod State Technological University. 2016. № 6. P. 139-148.
25. Manshin I.M., Krasnoperov N.S. Determination of the distance to the object using monocular vision // Scientific Aspect. 2023. № 1. P. 538.26.
26. Chen Y., Yu Z., Chen Y., Lan S., Anandkumar A., Jia J., Alvarez J.M. Focalformer 3d: focusing on hard instance for 3d object detection // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2023. P. 8394-8405.
27. Yin J., Shen J., Chen R., Li W., Yang R., Frossard P., Wang W. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024. New York: Publishing House of Institute of Electrical and Electronics Engineers, 2024. P. 14905-14915.