A Survey on 3D Object Detection Methods for Autonomous Driving Applications

自动驾驶应用中的 3D 目标检测方法调查

论文引用：

Arnold E, Al-Jarrah O Y, Dianati M, et al. A survey on 3d object detection methods for autonomous driving applications[J]. IEEE Transactions on Intelligent Transportation Systems, 2019.

ABSTRACT

An autonomous vehicle (AV) requires an accurate perception of its surrounding environment to operate reliably. The perception system of an AV, which normally employs machine learning (e.g., deep learning), transforms sensory data into semantic information that enables autonomous driving. Object detection is a fundamental function of this perception system, which has been tackled by several works, most of them using 2D detection methods. However, the 2D methods do not provide depth information, which is required for driving tasks, such as path planning, collision avoidance, and so on. Alternatively, the 3D object detection methods introduce a third dimension that reveals more detailed object's size and location information. Nonetheless, the detection accuracy of such methods needs to be improved. To the best of our knowledge, this is the first survey on 3D object detection methods used for autonomous driving applications. This paper presents an overview of 3D object detection methods and prevalently used sensors and datasets in AVs. It then discusses and categorizes the recent works based on sensors modalities into monocular, point cloud-based, and fusion methods. We then summarize the results of the surveyed works and identify the research gaps and future research directions.

摘要

自动驾驶汽车（AV）需要准确感知其周围环境才能可靠地运行。通常采用机器学习（例如，深度学习）的 AV 的感知系统将感知数据转换成能够实现自动驾驶的语义信息。物体检测是这种感知系统的基本功能，已被多项工作所解决，其中大多数使用 2D 检测方法。但是，2D 方法不提供驱动任务所需的深度信息，例如路径规划，碰撞避免等。或者，3D 对象检测方法引入第三维，其显示更详细的对象的大小和位置信息。但是，需要提高这些方法的检测精度。据我们所知，这是第一个用于自动驾驶应用的 3D 物体检测方法的调查。本文概述了三维物体检测方法以及 AV 中普遍使用的传感器和数据集。然后，它讨论并将基于传感器模态的近期作品分类为单目，基于点云和融合方法。然后，我们总结了调查工作的结果，并确定了研究差距和未来的研究方向。

1. Introduction

Between the years 2016 and 2017, the number of road casualties in the U.K. was approximately 174,510, of which 27,010 were killed or severely injured casualties [1]. As reported by the U.S. Department of Transportation, more than 90% of car crashes in the U.S. are attributed to drivers' errors [2]. The adoption of connected and autonomous vehicles is expected to improve driving safety, traffic flow and efficiency [3]. However, for an autonomous vehicle to operate safely, an accurate environment perception and awareness is fundamental.

1. 引文

在 2016 年至 2017 年期间，英国的道路伤亡人数约为 174,510 人，其中 27,010 人死亡或重伤人员[1]。据美国运输部报道，美国 90％以上的车祸都归因于司机的错误[2]。联通和自动驾驶车辆的采用有望提高驾驶安全性，交通流量和效率[3]。然而，对于自主车辆安全运行，准确的环境感知和意识是基本的。

The perception system of an Autonomous Vehicle (AV) transforms sensory data into semantic information, such as identification and recognition of road agents (e.g., vehicles, pedestrians, cyclists, etc.) positions, velocity and class; lane marking; drivable areas and traffic signs information.

自主车辆（AV）的感知系统将感知数据转换为语义信息，例如识别和辨识道路智能体（例如，车辆，行人，骑车者等）的位置，速度和类别;车道标记;可行驶区域和交通标志信息。值得注意的是，对象检测任务具有根本重要性，因为未能识别和辨识道路智能体可能会导致与

Notably, the object detection task is of fundamental importance, as failing to identify and recognize road agents might lead to safety-related incidents. For instance, failing in detecting a leading vehicle can result in traffic accidents, threatening human lives [4].

安全相关的事件。例如，未能检测到前方车辆可能导致交通事故，威胁人类生命[4]。

One factor for failure in the perception system arises from sensors limitations and environment variations such as lighting and weather conditions. Other challenges include generalization across driving domains such as motorways, rural and urban areas. While motorways have well-structured lanes with vehicles following a standard orientation, urban areas exhibit vehicles parked at no particular orientation, more diverse classes such as pedestrians, cyclists, and background clutter such as bollards and bins. Another factor is occlusion, when one object blocks the view of another, resulting in partial or complete invisibility of the object. Not only objects' sizes can be verydissimilar, e.g.,comparinga truckwith a dog, but objects can be very close or far away from the subject AV. The object's scale dramatically affects the sensors' readings, resulting in very dissimilar representations for the objects of the same class.

感知系统失败的一个因素是传感器限制和环境变化，例如照明和天气条件。其他挑战包括高速公路，农村和城市等领域的推广。虽然高速公路拥有结构合理的车道，车辆遵循标准方向，但城市区域展示的车辆没有特定的方向，更多样化的类别，如行人，骑自行车者和背景杂物，如护柱和垃圾箱。另一个因素是遮挡，当一个对象阻挡另一个对象的视图时，导致对象的部分或完全不可见。不仅物体的尺寸可以非常相似，例如，比较卡车与狗，但是物体可以非常靠近或远离物体 AV。对象的比例会显着影响传感器的读数，从而导致同一类对象的表达形式非常不同。

Despite the aforementioned challenges, the performance of 2D object detection methods for autonomous driving has greatly improved, achieving an Average Precision (AP) of more than 90% on the well established "KITTI" object detection benchmark [5]. While 2D methods detect objects on the image plane, their 3D counterpart introduce a third dimension to the localization and size regression, revealing depth information in world coordinates. However, the performance gap between 2D and 3D methods in the context of AVs is still significant [6]. Further research should be conducted to fill the performance gap of 3D methods, as 3D scene understanding is crucial for driving tasks. A comparison between 2D and 3D detection methods is presented in Table I.

尽管存在上述挑战，但是用于自动驾驶的 2D 物体检测方法的性能已经大大提高，在完善的"KITTI"物体检测基准[5]上实现了超过 90％的平均精度（AP）。当 2D 方法检测图像平面上的对象时，它们的 3D 对应物为定位和大小回归引入第三维，在世界坐标中显示深度信息。然而，在 AV 背景下 2D 和 3D 方法之间的性能差距仍然很大[6]。应该进行进一步的研究以填补 3D 方法的性能差距，因为 3D 场景理解对于驾驶任务至关重要。表 I 中给出了 2D 和 3D 检测方法之间的比较。

TABLE I
2D VERSUS 3D OBJECT DETECTION

	Advantages	Disadvantages
2D Object Detection	Well established datasets and detection architectures. Usually RGB only input can achieve accurate results in the image plane.	Limited information: lack of object's pose, occlusion and 3D position information.
3D Object Detection	3D bounding box provides object size and position in world coordinates. These detailed information allows better environment understanding.	Requires depth estimation for precise localization. Extra dimension regression increases model complexity. Scarse 3D labelled datasets.

In previous work, Ranft and Stiller [7] reviewed machine vision methods for different tasks of intelligent vehicles, including localization and mapping, driving scene understanding and object classification. In [8], on-road object detection was briefly reviewed among other perception functions, however, authors predominantly considered 2D object detection. Mukhtar et al. [9] reviewed 2D vehicle detection methods for Driver Assistance Systems with focus on motion and appearance-based approaches using a traditional pipeline. A traditional pipeline consists of segmentation (e.g., graph-based segmentation [10] and voxel-based clustering methods [11]), hand-engineered feature extraction (e.g., voxel's probabilistic features [11]) and classification stages (e.g., a mixture of bag-of-words classifiers [12]).

在之前的工作中，Ranft 和 Stiller [7]回顾了智能车辆不同任务的机器视觉方法，包括定位和绘图，驾驶场景理解和对象分类。在[8]中，在其他感知功能中对路上物体检测进行了简要回顾，但作者主要考虑了二维物体检测。Mukhtar 等人[9]回顾了驾驶员辅助系统的 2D 车辆检测方法，重点是使用传统技术路线的基于运动和外观的方法。传统的技术路线包括分割（例如，基于图形的分割[10]和基于体素的聚类方法[11]），手工设计的特征提取（例如，体素的概率特征[11]）和分类阶段（例如，词袋分类器的混合[12]）。

Unlike traditional pipelines, which optimize each stage individually, end-to-end pipelines optimize the overall pipeline performance. An end-to-end detection method leverages learning algorithms to propose regions of interest and extract features from the data. The shift towards representation learning and end-to-end detection was possible by using deep learning methods, such as deep convolutional networks, which showed a significant performance gain in different applications [13], [14]. In this paper we focus on end-to-end pipelines and learning approaches, since these have become the state-of-the-art for 3D object detection and have rapidly progressed in recent years.

与传统技术路线可单独优化每个阶段不同，端到端技术路线可优化整体技术路线性能。端到端检测方法利用学习算法来提出感兴趣的区域并从数据中提取特征。通过使用深度卷积网络等深度学习方法，可以实现向表示学习和端到端检测的转变，这种方法在不同的应用中显示出显着的性能提升[13]，[14]。在本文中，我们关注端到端技术路线和学习方法，因为这些已经成为最先进的三维物体检测技术，并且近年来发展迅速。

This paper presents an overview of 3D object detection methods and prevalently used sensors and datasets in AVs. We discuss and categorize existing works based on sensor modality into: monocular-based methods, point cloud-based methods and fusion methods. Finally, we discuss current research challenges and future research directions. The contributions of this paper are as follows:

本文概述了三维物体检测方法以及 AV 中普遍使用的传感器和数据集。我们基于传感器模态讨论现有作品并将其分类为：基于单眼的方法，基于点云的方法和融合方法。最后，我们讨论当前的研究挑战和未来的研究方向。本文的贡献如下：

- ✓ summarizing datasets and simulation tools used to evaluate the performance of detection models
- ✓ providing a summary of 3D object detection advancements for autonomous driving vehicles
- ✓ comparing 3D object detection methods performances on a baseline benchmark
- ✓ identifying research gaps and future research directions.

- ✓ 总结用于评估检测模型性能的数据集和模拟工具；
- ✓ 提供自动驾驶车辆的 3D 物体检测发展的概述；
- ✓ 比较基线基准测试中的 3D 物体检测方法性能；
- ✓ 确定研究差距和未来研究方向。

This paper is structured as follows. Section II describes commonly used sensors for perception tasks in autonomous vehicles. Section III lists well-referenced datasets used for object detection in AVs. We review 3D object detection

此篇文章的结构如下：第 II 节描述了用于自动驾驶车辆中感知任务的常用传感器。第 III 节列出了用于 AV 中对象检测的充分引用的数据集。我们在第 IV 节中回顾了 3D 对象检测方法。第五部分比较了现有方法在基准

methods in Section IV. Section V compares the performance of existing methods on a benchmark dataset and highlights research challenges and potential research opportunities. Section VI provides a brief summary and concludes this work.

数据集上的表现，并突出了研究挑战和潜在的研究机会。第六节简要总结了这项工作。

2. SENSORS

Although humans primarily use their visual and auditory systems while driving, artificial perception methods rely on multiple modalities to overcome shortcomings of individual sensors. There are a wide range of sensors used by autonomous vehicles: passive ones, such as monocular and stereo cameras, and active ones, including lidar, radar and sonar. Since most research on perception for AVs focus on cameras and lidars, these two categories are described in higher detail. A more comprehensive report on current sensors for AV applications can be found in [15] and [16].

2. 传感器

尽管人类在驾驶时主要使用其视觉和听觉系统，但人工感知方法依赖于多种方式来克服各个传感器的缺点。自动驾驶车辆使用的传感器种类繁多：被动式传感器，如单目和立体摄像机，以及有源传感器，包括激光雷达，雷达和声纳。由于大多数关于 AV 感知的研究都集中在相机和激光雷达上，因此更详细地描述了这两个类别。有关 AV 应用的当前传感器的更全面的报告可以在[15]和[16]中找到。

2.1 Cameras

Monocular cameras provide detailed information in the form of pixel intensities, which at a bigger scale reveal shape and texture properties. The shape and texture information can be used to detect lane geometry, traffic signs [17] and the object class [7].

2.1 相机

单目相机以像素强度的形式提供详细信息，其以更大的尺度显示形状和纹理特性。形状和纹理信息可用于检测车道几何，交通标志[17]和对象类别[7]。

One disadvantage of monocular cameras is the lack of depth information, which is required for accurate object size and position estimation. A stereo camera setup can be used to recover depth channels. Such configuration uses matching algorithms to find correspondences in both images and calculate the depth of each point relative to the camera, demanding more processing power [18].

单目相机的一个缺点是缺少深度信息，这是精确的物体尺寸和位置估计所需的。立体摄像机装置可用于恢复深度通道。这种配置使用匹配算法来找到两个图像中的对应关系，并计算每个点相对于摄像机的深度，从而要求更高的处理能力[18]。

Other camera modalities that offer depth estimation are Time-of-Flight (ToF) cameras where depth is inferred by measuring the delay between emitting and receiving modulated infrared pulses [19]. This technology has been applied for vehicle safety applications [20], but despite the lower integration price and computational complexity has low resolution when compared to stereo cameras.

提供深度估计的其他相机模态是飞行时间（ToF）相机，其中深度通过测量延迟发射和接收调制的红外脉冲来推断[19]。该技术已应用于车辆安全应用[20]，但与立体相机相比，尽管集成度较低且计算复杂度较低，但分辨率较低。

Camera sensors are susceptible to light and weather conditions. Examples range from low luminosity at night-time to extreme brightness disparity when entering or leaving tunnels. The recent use of LEDs on traffic signs and vehicles brake lights creates a flickering problem. It happens as the camera sensor cannot reliably capture the emitted light due to the LEDs' switching behavior. Sony has

相机传感器易受光线和天气条件的影响。例子包括夜间的低亮度和进入或离开隧道时的极端亮度差异。最近在交通标志和车辆刹车灯上使用 LED 会产生晃动问题。这是因为由于 LED 的切换行为，相机传感器无法可靠地捕获发出的光。索尼最近宣布了一种新的相机技术，旨在减轻晃动效果并增强色彩动态范围[21]，如图 1 所示。此外，由于下雨或下雪天气，图像会降级。陈等人

recently announced a new camera technology designed to mitigate flickering effects and enhance colors dynamic range [21], as illustrated in Figure 1. Additionally, image degradation can occur due to rainy or snowy weather. Chen et al. [22] propose to mitigate this using a de-raining filter based on a multi-scale pyramid structure and conditional generative adversarial networks.

[22]建议使用基于多尺度金字塔结构和条件生成对抗网络的除雨过滤器来缓解这种情况。

Figure 1. IMX390 sensor sample image on a tunnel exit. The image on the left was taken with both LED flickering mitigation and High-Dynamic-Ranging (HDR) capability enabled. The top right image shows HDR functionality without LED flickering mitigation – note that the traffic sign velocity indicator does not appear. The bottom right image shows the image without any of the functionalities enabled, clearly showing the sensor capabilities. Image obtained from the Sony website [21].

图 1. 隧道出口处的 IMX390 传感器样本图像。左侧的图像是在启用 LED 闪烁缓解和高动态范围（HDR）功能的情况下拍摄的。右上角的图像显示 HDR 功能，没有 LED 闪烁缓解 - 请注意，交通标志速度指示器不会出现。右下角的图像显示的图像没有启用任何功能，清楚地显示了传感器功能。图片来自索尼网站[21]。

2.2 Lidar

Lidar sensors emit laser beams and measure the time between emitting and detecting the pulse back. The timing information determines the distance of obstacles in any given direction. The sensor readings result in a set of 3D points, also called Point Cloud (PCL), and corresponding reflectance values representing the strength of the received pulses. Unlike images, point clouds are sparse: the samples are not uniformly distributed in space. As active sensors, external illumination is not required and thus more reliable detection can be achieved considering adverse weather and extreme lighting conditions (e.g., night-time or sun glare scenarios).

2.2 激光雷达

激光雷达传感器发射激光束并测量发射和检测脉冲之间的时间。定时信息确定任何给定方向上的障碍物的距离。传感器读数产生一组 3D 点，也称为点云（PCL），以及表示接收脉冲强度的相应反射值。与图像不同，点云是稀疏的：样本在空间中不均匀分布。作为有源传感器，不需要外部照明，因此考虑到恶劣天气和极端光照条件（例如，夜间或太阳眩光场景），可以实现更可靠的检测。

Standard lidar models, such as the HDL-64L [23], use an array of rotating laser beams to obtain 3D point clouds in 360 degrees and up to 120m radius. This sensor can output

标准的激光雷达模型，例如 HDL-64L [23]，使用旋转激光束阵列来获得 360 度和 120 米半径的 3D 点云。该传感器每帧输出 12 万个点，在 10Hz 帧速率下达到每秒 12

120 thousand points per frame, which amounts to 1,200 million points per second on a 10 Hz frame rate. Velodyne recently announced the VLS-128 model [24] featuring 128 laser beams, higher angular resolution and 300m radius range. Figure 2 shows a comparison between the point densities of the two models. The announcement suggests that the increased point density might enhance the recall of methods using this modality but challenges real time processing performance. The primary challenge to the widespread use of lidar is its price: a single sensor can cost more than $70,000. Nevertheless, this price is expected to decrease in the following years with the introduction of solid state lidar technology [25] and large scale production.

亿个点。Velodyne 最近宣布推出 VLS-128 型号[24]，具有 128 个激光束，更高的角分辨率和 300 米半径范围。图 2 显示了两种模型的点密度之间的比较。该公告表明，增加的点密度可能会增强使用此模态的方法的召回，但会对实时处理性能提出挑战。广泛使用激光雷达的主要挑战是其价格：单个传感器的成本可能超过 70,000 美元。尽管如此，随着固态激光雷达技术[25]的引入和大规模生产，预计这一价格将在接下来的几年中下降。

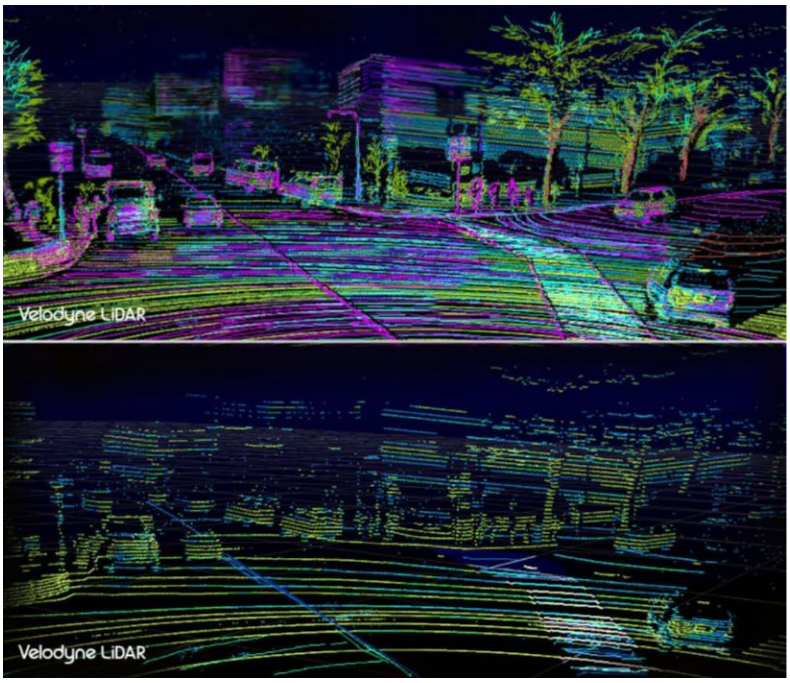

Figure 2. The two images show the point clouds obtained by two lidar sensors on the same scene. The top image was captured using the newer VLS-128 model while the bottom one used the standard HDL-64 model. Image obtained from [24].

图 2. 这两幅图像显示了同一场景中两个激光雷达传感器获得的点云。顶部图像使用较新的 VLS-128 模型捕获，而底部图像使用标准 HDL-64 模型。图像来自 [24]。

Some methods rely on both lidar and camera modalities. Before fusing these modalities it is required to calibrate the sensors to obtain a single spatial frame of reference. Park et al. [26] propose to use polygonal planar boards as targets that can be detected by both modalities to generate accurate 3D-2D correspondences and obtain a more accurate calibration. However, having spatial targets makes this method laborious for on-site calibration. As an alternative, Ishikawa et al. [27] devised a calibration method without spatial targets using odometry estimation of the sensors w.r.t. the environment to iteratively calibrate them.

一些方法依赖于激光雷达和相机模态。在融合这些模态之前，需要校准传感器以获得单个空间参照系。Park 等[26]建议使用多边形平面板作为目标，可以通过两种模态检测，以生成精确的 3D-2D 对应关系并获得更准确的校准。然而，具有空间目标使得该方法难以进行现场校准。作为替代方案，Ishikawa 等人[27]设计了一种没有空间目标的校准方法，使用传感器的测距估计 w.r.t.环境迭代校准它们。

2.3 Discussion

Monocular cameras are inexpensive sensors, but they lack depth information which is required for accurate 3D object detection. Depth cameras can be used for depth recovery, but fail in adverse lighting conditions and textureless scenes and ToF camera sensors have limited resolution. In contrast, lidar sensors can be used for accurate depth estimation during night-time, but is prone to noise during adverse weather, such as snow and fog, and cannot provide texture information. We summarize the advantages and disadvantages of each sensor modality in Table II.

2.3 讨论

单目相机是廉价的传感器，但它们缺少精确 3D 物体检测所需的深度信息。深度相机可用于深度恢复，但在不利的光照条件和无纹理场景中失败，并且 ToF 相机传感器的分辨率有限。相比之下，激光雷达传感器可用于在夜间进行精确的深度估计，但在恶劣天气（例如雪和雾）期间易于产生噪声，并且不能提供纹理信息。我们总结了表 II 中每种传感器模态的优缺点。

TABLE II

SENSORS COMPARISON

	Advantages	**Disadvantages**
Monocular Camera	Readily available and inexpensive. Multiple specifications available.	Prone to adverse light and weather conditions. No depth information provided.
Stereo Camera	Higher point density when compared to lidar. Provides dense depth map.	Depth estimation is computationally expensive. Poor performance with textureless regions or during night-time. Limited Field-of-View (FoV).
Lidar	360 degrees FoV, precise distance measurements. Not affected by light conditions.	Raw point cloud does not provide texture information. Expensive and large equipment.
Solid-State lidar	No moving mechanical parts, compact size. Large scale production should reduce final cost.	Limited FoV when compared to mechanical scanning lidar. Still under development.

3. DATASETS

As learning approaches become widely used the need of training data also increases. The availability of large scale image datasets such as ImageNet [28] allowed fast development and evolutionof image classification and object detection models. The same phenomena occurs in the driving scenario, where more data means broader scenario coverage. In particular, tasks such as object detection and semantic segmentation require finely labelled data. In this section we present common datasets for driving tasks, specifically to object detection.

3. 数据集

随着学习方法的广泛使用，训练数据的需求也在增加。大型图像数据集（如 ImageNet [28]）的可用性允许图像分类和对象检测模型的快速发展和演变。在驾驶场景中出现相同的现象，其中更多数据意味着更广泛的场景覆盖。特别是，诸如对象检测和语义分割之类的任务需要完全标记的数据。在本节中，我们将介绍用于驾驶任务的常用数据集，特别是对象检测。

One of the most used datasets in the driving context is KITTI [29], which provides stereo color images, lidar point clouds and GPS coordinates, all synchronized in time. Recorded scenes range from well-structured highways, complex urban areas and narrow countryside roads. The dataset can be used for multiple tasks: stereo matching, visual odometry, 3D

驾驶环境中最常用的数据集之一是 KITTI [29]，它提供立体彩色图像，激光雷达点云和 GPS 坐标，所有这些都在时间上同步。录制的场景包括结构良好的高速公路，复杂的城市区域和狭窄的乡村道路。数据集可用于多种任务：立体匹配，视觉测距，3D 跟踪和 3D 物体检测。特别地，特定对象检测数据集包含 7,481 个训练和 7,518

tracking and 3D object detection. In particular, the specific object detection dataset contains 7,481 training and 7,518 test frames, which are provided with sensor calibration information and annotated 3D boxes around objects of interest. The annotations are categorized in “easy, moderate and hard” cases, according to object size, occlusion and truncation levels.

个测试帧，其提供有传感器校准信息和围绕感兴趣对象的标注 3D 框。根据对象大小，遮挡和截断水平，标注被分类为“简单，中等和难”的情况。

Despite widely adopted, this dataset has several limitations. Notably, limited sensor configuration and lighting conditions: all the measurements were obtained by the same set of sensors during daytime and mostly under sunny conditions. In addition the classes frequency is highly unbalanced [30] – 75% car, 4% cyclist and 15% pedestrians. Furthermore, most scene objects follow a predominant orientation, facing the ego-vehicle. The lack of variety challenges the evaluation of current methods in more general scenarios, reducing their reliability for real-world applications.

尽管被广泛采用，但该数据集有一些局限性。值得注意的是，有限的传感器配置和照明条件：所有测量都是在白天通过同一组传感器获得的，并且主要是在阳光充足的条件下。此外，类别频率高度不平衡[30] - 75％的汽车，4％的骑车者和 15％的行人。此外，大多数场景物体遵循主要方向，面向自我车辆。缺乏多样性挑战了当前方法在更一般情况下的评估，降低了实际应用的可靠性。

Considering these limitations and the expensive process of obtaining and labeling a dataset, Gaidon et al. [31] proposed the Virtual KITTI dataset. The authors manually recreated the KITTI environment using a game-engine, 3D model assets and the original video sequences, see Figure 3. Different lighting and weather conditions, vehicles colors and models, etc., were adjusted to automatically generate labelled data. They provide approximately 17,000 frames consisting of the photo-realistic images, a depth frame, and pixel-level semantic segmentation ground-truth. Additionally, the authors assessed the transferability across real and virtual domains for a tracking application (which requires detection). They evaluated a tracker trained on real images and tested on virtual ones. The results revealed that the gap in performance is minimal, showing the equivalence of the datasets. They also concluded that the best performance was obtained when training on the virtual data and fine-tuning on real data.

考虑到这些限制以及获取和标记数据集的昂贵过程，Gaidon 等人[31]提出了虚拟 KITTI 数据集。作者使用游戏引擎，3D 模型资产和原始视频序列手动重建 KITTI 环境，参见图 3.调整不同的光照和天气条件，车辆颜色和模型等，以自动生成标记数据。它们提供大约 17,000 帧，包括照片般逼真的图像，深度帧和像素级语义分割地面实况。此外，作者还评估了跟踪应用程序（需要检测）的实际和虚拟域的可转移性。他们评估了一个在真实图像上训练并在虚拟图像上进行测试的跟踪器结果显示，性能差距很小，显示了数据集的等效性。他们还得出结论，在对虚拟数据进行训练和对实际数据进行微调时，可以获得最佳性能。

Simulation tools can be used to both generate training data on specific conditions or to train end-to-end driving systems [32], [33]. Using virtual data during training can enhance the performance of detection models on real environments. This data can be obtained through game-engines[34] or simulated environments [31]. CARLA [35] is an open-source simulation tool for autonomous driving that allows flexible environmental setup and sensor configuration. It provides several 3D models for pedestrians, cars and includes two virtual towns. Environmental conditions, such as weather

仿真工具既可用于生成特定条件下的训练数据，也可用于训练端到端驾驶系统[32]，[33]。在训练期间使用虚拟数据可以增强真实环境中检测模型的性能。这些数据可以通过游戏引擎[34]或模拟环境[31]获得。CARLA [35]是一种用于自动驾驶的开源仿真工具，允许灵活的环境设置和传感器配置。它为行人，汽车提供了几种 3D 模型，包括两个虚拟城镇。可以调整环境条件，例如天气和照明，以产生看不见的情景。虚拟传感器套件包括具有 ground-truth 分割帧的 RGB 和深度相机以及光线投射激光雷达模型。另一个模拟工具 Sim4CV [36]允许轻松

and lighting, can be adjusted to generate unseen scenarios. The virtual sensor suite includes RGB and depth cameras with ground-truth segmentation frames and a ray-casting lidar model. Another simulation tool, Sim4CV [36] allows easy environment customization and simultaneous multi-view rendering of the driving scenes, while providing ground-truth bounding boxes for object detection purposes.

的环境定制和驾驶场景的同时多视图渲染，同时提供用于物体检测目的的真值边界框。

Figure 3. Frames from 5 real KITTI videos (first column) and respective virtual clones on Virtual KITTI (second column). Image from [31].

图3. 来自5 个真实KITTI 视频（第一列）的帧和Virtual KITTI（第二列）上的相应虚拟克隆。图片来自[31]。

4. 3D OBJECT DETECTION METHODS

4. 3D 目标检测方法

We divide 3D object detection methods in three categories: monocular image, point cloud and fusion based methods. An overview of methodology, advantages and limitations for these methods is provided in Table III. The following subsections address each category individually.

我们将三维物体检测方法分为三类：单目图像，点云和基于融合的方法。表 III 提供了这些方法的方法，优点和局限性的概述。以下小节分别针对每个类别。

4.1 Monocular Image Based Methods

Although 2D object detection is a largely addressed task that has been successfully tackled in several datasets [37], [38], the KITTI dataset offers particular settings that pose challenges to object detection. These settings, common to most driving environments, include small, occluded or truncated objects and highly saturated areas or shadows. Furthermore, 2D detection on the image plane is not enough

4.1 基于单目图像方法

尽管 2D 物体检测是一项在很多数据集中成功解决的主要任务[37]，[38]，但 KITTI 数据集提供了针对物体检测的挑战。这些设置对于大多数驾驶环境来说是常见的，包括小的，遮挡的或截断的对象以及高度饱和的区域或阴影。此外，图像平面上的 2D 检测对于可靠的驾驶系统是不够的：对于这种应用，需要更精确的 3D 空间定位和尺寸估计。本节重点介绍能够仅基于单目图像

TABLE III
COMPARISON OF 3D OBJECT DETECTION METHODS BY CATEGORY

Category		Methodology/Advantages	Limitations/Drawbacks	Research Gaps
Monocular		Uses single RGB images to predict 3D object bounding boxes. Predicts 2D bounding boxes on the image plane then extrapolate them to 3D through re-projection constraints or bounding box regression.	The lack of explicit depth information on the input format limits the accuracy of localization performance.	CNNs that estimate depth channels could be investigated to increase localization accuracy.
Point-cloud	Projection	Projects point clouds into a 2D image and use established architectures for object detection on 2D images with extensions to regress 3D bounding boxes.	Projecting the point cloud data inevitably causes information loss. It also prevents the explicit encoding of spatial information as in raw point cloud data.	The encoding of the input image is performed with hand-engineered features (point density, etc.). Learned input representations could improve the detection results.
	Volumetric	Generates a 3 dimensional representation of the point cloud in a voxel structure and uses Fully Convolutional Networks (FCNs) to predict object detections. Shape information is encoded explicitly.	Expensive 3D convolutions increase models inference time. The volumetric representation is sparse and computationally inefficient.	Volumetric methods have not considered region proposals, which could improve both localization accuracy and processing time.
	PointNet	Uses feed-forward networks consuming raw 3D point clouds to generate predictions on class and estimated bounding boxes.	Considering the whole point cloud as input can increase run-time. Difficult establishing region proposals considering raw point inputs.	PointNet architectures rely on region proposals to limit the number of points. Proposal methods based uniquely on point-cloud data should be investigated.
Fusion		Fuses both front view images and point clouds to generate a robust detections. Architectures usually consider multiple branches, one per modality, and rely on region proposals. Allows modalities to interact and complement each other.	Requires calibration between sensors, and depending on the architecture can be computationally expensive.	These methods represent state-of-the-art detectors. However, they should be evaluated on more general scenarios including diverse lighting and weather conditions.

for reliable driving systems: more accurate 3D space localization and size estimation is required for such application. This section focuses on methods that are able to estimate 3D bounding boxes based only on monocular images. Since no depth information is available, most approaches first detect 2D candidates before predicting a 3D bounding box that contains the object using neural networks [39], geometrical constraints [40] or 3D model matching [41], [42].

Chen et al. [39] propose Mono3D, which leverages a simple region proposal algorithm using context, semantics, hand engineered shape features and location priors. For any given proposal, these features can be efficiently computed and scored by an energy model. Proposals are generated by exhaustive search on 3D space and filtered with Non-Maxima Suppression (NMS). The proposals are further scored by a Fast R-CNN [37] model that regresses 3D bounding boxes. The work builds upon the authors' previous work 3DOP [43], which considers depth images to generate proposals in a similar framework. Despite using only monocular images, the Mono3D model slightly improves the performance obtained by [43], which uses depth images. Pham and Jeon [44] extends the 3DOP proposal generation considering class-independent proposals, then re-ranks the proposals using both monocular images and depth maps. Their method outperforms both 3DOP and Mono3D

估计 3D 边界框的方法。由于没有可用的深度信息，大多数方法首先在使用神经网络[39]，几何约束[40]或 3D 模型匹配[41]，[42]预测包含对象的 3D 边界框之前检测 2D 候选。

陈等人[39]提出 Mono3D，它利用一个简单的区域提议算法，使用上下文，语义，手工设计的形状特征和位置先验。对于任何给定的提议，可以通过能量模型有效地计算和评分这些特征。通过对 3D 空间进行详尽搜索并使用非最大值抑制（NMS）进行过滤来生成提议。这些提议得到了 Fast R-CNN [37]模型的进一步评分，该模型回归了 3D 边界框。这项工作建立在作者之前的工作 3DOP [43]的基础上，该工作考虑了深度图像以在类似的框架中生成提案。尽管仅使用单目图像，Mono3D 模型略微提高了[43]所获得的性能，后者使用了深度图像。Pham 和 Jeon [44]扩展了 3DOP 提议生成，考虑了与类无关的提议，然后使用单眼图像和深度图对提案进行重新排序。尽管使用深度图像来改善提议，但他们的方法优于 3DOP 和 Mono3D 方法。

methods, despite using depth images to refine proposals.

An important characteristic of driving environments is severe occlusion present in crowded scenes where vehicles can block the view of other agents and themselves. Xiang et al. introduce visibility patterns into the model to mitigate occlusion effects through object reasoning. They propose the 3D Voxel Pattern (3DVP) [41] representation that models appearance through RGB intensities, 3D shape as a set of voxels and occlusion masks. This representation allows to recover which parts of the object are visible, occluded or truncated. They obtain a dictionary of 3DVPs by clustering the patterns observed on the data and training a classifier for each specific pattern given a 2D image segment of the vehicle. During the test phase the pattern obtained through classification is used for occlusion reasoning and 3D pose and localization estimation. They achieve 3D detection by minimizing the reprojection error between the projected 3D bounding box to the image plane and the 2D detection. Their pipeline is still dependent on the performance of Region Proposal Networks (RPNs).

驾驶环境的一个重要特征是在拥挤的场景中存在严重的遮挡，其中车辆可以阻挡其他智能体和他们自己的视野。Xiang 等人将可见性模式引入模型中以通过对象推理来减轻遮挡效应。他们提出了 3D 体素模式（3DVP）[41]表示，其通过 RGB 强度模拟外观，3D 形状作为一组体素和遮挡掩模。此表示允许恢复对象的哪些部分可见，被遮挡或截断。他们通过聚类在数据上观察到的模式并在给定车辆的 2D 图像片段的每个特定模式下训练分类器来获得 3DVP 的字典。在测试阶段期间，通过分类获得的模式用于遮挡推理和 3D 姿势和定位估计。它们通过最小化投影的 3D 边界框与图像平面之间的重投影误差和 2D 检测来实现 3D 检测。他们的流程仍然依赖于区域提案网络（RPN）的性能。

Although some RPNs were able to improve traditional proposal methods [37] they still fail to handle occlusion, truncation and different object scales. Extending the previous 3DVP framework, the same authors propose SubCNN [45], a CNN that explores class information for object detection at the RPN level. They use the concept of subcategory, which are classes of objects sharing similar attributes such as 3D pose or shape. Candidates are extracted using convolutional layers to predict heat maps for each subcategory at the RPN level. After Region of Interest (ROI) proposal the network outputs category classification along with refined 2D bounding box estimates. Using 3DVPs [41] as subcategories for pedestrian, cyclist and vehicle classes, the model recovers 3D shape, pose and occlusion patterns. An extrapolating layer is used to improve small object detection by introducing multi-scale image pyramids.

尽管一些 RPN 能够改进传统的提议方法[37]，但它们仍然无法处理遮挡，截断和不同的对象尺度。扩展了之前的 3DVP 框架，同一作者提出了 SubCNN [45]，这是一个探索 RPN 级别对象检测的类信息的 CNN。他们使用子类别的概念，子类别是共享类似属性（如 3D 姿势或形状）的对象类。使用卷积层提取候选者以预测 RPN 级别的每个子类别的热图。在感兴趣区域（ROI）提议之后，网络输出类别分类以及重新定义的 2D 边界框估计。使用 3DVP [41]作为行人，骑车人和车辆类的子类别，该模型恢复 3D 形状，姿势和遮挡模式。外推层用于通过引入多尺度图像金字塔来改善小物体检测。

Despite the previous 3DVP representations [41], [45] allow to model occlusion and parts appearance, they are obtained as a classification among an existing dictionary of visibility patterns common in the training set. Thus, may fail to generalize to an arbitrary vehicle pose that differs from the existing patterns. To overcome this, Deep MANTA [42] uses a many task network to estimate vehicle position, part localization and shape based only on monocular images. The vehicle shape consists of a set of key points that characterize the vehicle 3-dimensional boundaries, e.g. external vertices

尽管之前的 3DVP 表示[41]，[45]允许模拟遮挡和部件外观，但它们是作为训练集中常见的现有可见性模式字典中的分类而获得的。因此，可能无法概括为与现有模式不同的任意车辆姿态。为了解决这个问题，Deep MANTA [42]使用多任务网络来估计仅基于单眼图像的车辆位置，部件定位和形状。车辆形状由一组表征车辆三维边界的关键点组成，例如，车辆的外部顶点。他们首先通过两级改进区域 - 建议网络获得 2D 边界回归和零件定位。接下来，基于推断的形状，执行 3D 模型匹配以获得 3D 姿势。

of the vehicle. They first obtain 2D bounding regression and parts localization through a two-level refinement region-proposal network. Next, based on the inferred shape 3D model matching is performed to obtain the 3D pose.

Previous attempts performed either exhaustive search on the 3D bounding box space [39], estimated 3D pose through a cluster of appearance patterns [41] or 3D templates [42]. Mousavianet al. [40] first extend a standard 2D object detector with 3D orientation (yaw) and bounding box sizes regression. This is justified by the box dimensions having smaller variance and being invariant with respect to the orientation. Most models use L2 regression for orientation angle prediction. In contrast, the authors propose a Multi-bin method to regress orientation. The angle is considered to belong to one of n overlapping bins and a network estimates the confidence of the angle belonging to each bin along with a residual angle to be added to the bin center to recover the output angle. The 3D box dimensions and orientations are fixed as determined by the network prediction. Then 3D object pose is recovered solving for a translation matrix that minimizes the reprojection error of the 3D bounding box w.r.t. the 2D detection box on the image plane.

之前的尝试在 3D 边界框空间[39]上进行了穷举搜索，通过一组外观模式[41]或 3D 模板[42]进行了估计 3D 姿势。Mousavianet al[40]第一个扩展标准 2D 物体探测器，具有 3D 方向（偏航）和边界框尺寸回归。这通过具有较小方差并且相对于方向不变的盒子尺寸来证明。大多数模型使用 L2 回归进行方位角预测。相比之下，作者提出了一种回归方向的多仓方法。该角度被认为属于 n 个重叠区间中的一个，并且网络估计属于每个区间的角度的信度以及要添加到区间中心的残余角度以恢复输出角度。3D 框尺寸和方向由网络预测确定。然后恢复 3D 对象姿势，求解平移矩阵，该平移矩阵最小化 3D 边界框 w.r.t 图像平面上的 2D 检测框的重投影误差。

All previous monocular methods can only detect objects from the front-facing camera, ignoring objects on the sides and rear of the vehicle. While lidar methods can be used effectively for 360 degrees detection, [46] proposes the first 360 degrees panoramic image based method for 3D object detection. They estimate dense depth maps of panoramic images and adapt standard object detection methods for the equirectangular representation. Due to the lack of panoramic labelled datasets for driving, they adapt the KITTI dataset using style and projection transformations. They additionally provide benchmark detection results on a synthetic dataset.

所有以前的单目方法只能检测前置摄像头中的物体，忽略车辆侧面和后面的物体。虽然激光雷达方法可以有效地用于 360 度检测，[46]提出了基于第一个 360 度全景图像的 3D 物体检测方法。他们估计全景图像的密集深度图，并使标准物体检测方法适用于等距矩形表示。由于缺少用于驾驶的全景标记数据集，他们使用样式和投影变换来调整 KITTI 数据集。它们还在合成数据集上提供基准检测结果。

Monocular methods have been widely researched. Although previous works considered hand-engineered features for region proposals [39], most methods have shifted towards a learned paradigm for Region Proposals and second stage of 3D model matching and reprojection to obtain 3D bounding boxes. The main drawbacks of monocular based methods is the lack of depth cues, which limits detection and localization accuracy specially for far and occluded objects, and sensitivity to lighting and weather conditions, limiting the use of these methods for day time. Also, since most methods rely on a front facing camera (except for [46]), it is only possible to detect objects in front of the vehicle,

单目方法已被广泛研究。虽然以前的工作考虑了区域提案的手工设计特征[39]，但大多数方法已经转向学习区域提案的范例和第二阶段的 3D 模型匹配和重投影以获得 3D 边界框。基于单目视觉方法的主要缺点是缺乏深度线索，这限制了特别针对远端和遮挡物体的检测和定位精度，以及对照明和天气条件的敏感性，限制了这些方法在白天的使用。此外，由于大多数方法依赖于前置摄像头（[46]除外），因此只能检测车辆前方的物体，这与点云方法形成对比，原则上，这些方法在车辆四周具有覆盖范围。我们总结了表 IV 中单目方法的方法/贡献和局限性。

contrasting to point clouds methods that, in principle, have a coverage all around the vehicle. We summarize the methodology/contributions and limitations of monocular methods in Table IV.

TABLE IV
SUMMARY OF MONOCULAR BASED METHODS

Method	Methodology/Contributions	Limitations
Mono3D [39]	Improves detection performance over 3DOP that relied on the depth channel.	Poor localization accuracy given the lack of depth cues.
3DVP [41]	Novel 3DVP object representation includes appearance, 3D shape and occlusion information. Classification among an existing set of 3DVPs allows occlusion reasoning and recovering 3D pose and localization.	Fixed set of 3DVPs extracted during training limits generalisation to arbitrary object poses.
SubCNN [45]	Uses 3DVP representation to generate occlusion-aware region proposals. The proposals are refined and classified within the object representations (3DVP). Improves RPN model refinement network using CNNs.	Since the 3DVP representation is employed, this method has the same limitations as the previous one.
DeepManta [42]	CNN to predict parts localization and visibility, fine orientation, 3D localization and template, 3D template matching to recover 3D position.	Detection restricted to vehicles, ignoring other classes.
Deep3DBox [40]	Simplified network architecture by independently regressing bounding box size and angle. Then using image reprojection error minimization to obtain 3D localization.	The reprojection error is dependent on the BB size and angle regressed by the network. This dependence increases localization error.
360Panoramic [46]	Estimates depth for 360 degrees panoramic monocular images. Then adapt a CNN to predict 3D object detections on the recovered panoramic image. The only method capable of using images to detect objects at any angle around the vehicle.	Limited to vehicle detection and fails when the vehicle is too close to the camera. The resolution of the camera limits the range of detection.

4.2 Point Cloud Based Methods

Current 3D object detection methods based on point-clouds can be divided into three subcategories: projection based, volumetric representations and point-nets. Each category is explained and reviewed below, followed by a summary discussion.

4.2.1 Projection Methods

Image classification and object detection in 2D images is a well-researched topic in the computer vision community. The availability of datasets and benchmarked architectures for 2D images make using these methods even more attractive. For this reason, point cloud (PCL) projection methods first transform the 3D points into a 2D image via plane [47], cylindrical [48] or spherical [34] projections that can then be processed using standard 2D object detection models such as [49]. The 3D bounding box can then be recovered using position and dimensions regression.

Li et al. [48] uses a cylindrical projection mapping and a Fully Convolutional Network (FCN) to predict 3D bounding boxes around vehicles only. The input image resulting from the projection has channels encoding the points' height and distance from the sensor. This input is fed to a 2D FCN which down-samples the input for three consecutive layers and then uses transposed convolutional layers to up-sample these maps into point-wise "objectness" and bounding box (BB) prediction outputs. The first output defines if a given point is part of a vehicle or the background, effectively

4.2 基于点云方法

目前基于点云的三维物体检测方法可以分为三个子类别：基于投影，体积表示和点网。下面将解释和审查每个类别，然后进行总结性讨论。

4.2.1 投影方法

2D 图像中的图像分类和对象检测是计算机视觉社区中经过深入研究的主题。数据集和 2D 图像基准架构的可用性使得使用这些方法更具吸引力。出于这个原因，点云（PCL）投影方法首先通过平面[47]，圆柱[48]或球形[34]投影将 3D 点转换为 2D 图像，然后可以使用标准 2D 物体检测模型进行处理，例如[49]。然后可以使用位置和尺寸回归来恢复 3D 边界框。

Li 等人 [48]使用圆柱投影映射和完全卷积网络（FCN）来预测车辆周围的 3D 边界框。投影产生的输入图像具有编码点的高度和距传感器的距离的通道。该输入被馈送到 2D FCN，其对三个连续层的输入进行下采样，然后使用转置的卷积层将这些映射上采样为逐点"对象"和边界框（BB）预测输出。第一个输出定义给定点是车辆的一部分还是背景，有效地作为弱分类器工作。第二输出编码 3D 边界框的顶点，界定由第一输出调节的车辆。由于每辆车将有许多 BB 估计值，因此采用 NMS 策略来减少基于得分和距离的重叠预测。作者在 KITTI 数

working as a weak classifier. The second output encodes the vertices of the 3D bounding box delimiting the vehicle conditioned by the first output. Since there will be many BB estimates for each vehicle, an NMS strategy is employed to reduce overlapping predictions based on score and distance. The authors train this detection model in an end-to-end fashion on the KITTI dataset with loss balancing to avoid bias towards negative samples or near cars, which appear more frequently.

据集上以端到端的方式训练这种检测模型，并进行损失平衡，以避免偏向负样本或靠近汽车，这种情况看起来更频繁。

While previous methods used cylindrical and spherical projections, [30], [50], [51] use the bird-eye view projection to generate 3D proposals. They differ regarding the input representation: the first encodes the 2D input cells using the minimum, median and maximum height values of the points lying inside the cell as channels, while the last two use height, intensity and density channels. The first approach uses a Faster R-CNN [13] architecture as a base with an adjusted refinement network that outputs oriented 3D bounding boxes. Despite their reasonable bird-eye view results, their method performs poor orientation angle regression. Most lidar base methods use sensors with high point density, which limits the application of the resulting models on low-end lidar sensors. Beltrán et al. [51] propose a novel encoding that normalizes the density channel based on the parameters of the lidar being used. This normalization creates a uniform representation and allows to generalize the detection model to sensors with different specifications and number of beams.

虽然以前的方法使用圆柱形和球形投影，[30]，[50]，[51]使用鸟瞰视图投影来生成 3D 建议。它们在输入表示方面有所不同：第一个使用位于单元内的点的最小，中值和最大高度值作为通道对 2D 输入单元进行编码，而后两个使用高度，强度和密度通道。第一种方法使用更快的 R-CNN [13]架构作为基础，具有调整的改进网络，输出定向的 3D 边界框。尽管它们具有合理的鸟瞰视图结果，但它们的方法执行较差的方向角回归。大多数激光雷达基础方法使用具有高点密度的传感器，这限制了所得模型在低端激光雷达传感器上的应用。Beltrán 等[51]提出了一种新颖的编码，它基于所使用的激光雷达的参数来标准化密度通道。这种归一化产生了统一的表示，并允许将检测模型推广到具有不同规格和光束数量的传感器。

One fundamental requirement of safety-critical systems deployed on autonomous vehicles, including object detection, is real-time operation capability. These systems must meet strict response time deadlines to allow the vehicle to respond to the environment. Complex-YOLO [30] focus on efficiency using a YOLO [52] based architecture, with extensions to predict the extra dimension and yaw angle. While classical RPN approaches further process each region for finer predictions, this architecture is categorized as a single-shot detector, obtaining detections in a single forward step. This allows Complex-YOLO to achieve a runtime of 50 fps, up to five times more efficient than previous methods, despite inferior, but comparable detection performance.

部署在自动驾驶车辆上的安全关键系统的一个基本要求，包括物体检测，是实时操作能力。这些系统必须满足严格的响应时间期限，以允许车辆响应环境。Complex-YOLO [30]使用基于 YOLO [52]的架构专注于效率，并具有预测额外维度和偏航角的扩展。虽然经典 RPN 方法进一步处理每个区域以进行预测，但该体系结构被归类为单发检测器，在单个前向步骤中获得检测。这使得 Complex-YOLO 可以实现 50 fps 的运行时间，比以前的方法效率高出五倍，尽管性能较差，但具有可比性。

Quantifying the confidence of predictions made by an AV's object detection system is fundamental for the safe operation of such vehicle. As with human drivers, if the system has low confidence on its predictions, it should enter a safe state

量化 AV 物体检测系统对预测的置信度是这种车辆安全操作的基础。与人类驾驶员一样，如果系统对其预测的置信度较低，则应进入安全状态以避免风险。尽管大多数检测模型为每个预测提供分数，但它们倾向于使用

to avoid risks. Although most detection models offer a score for each prediction, they tend to use softmax normalization to obtain class distributions. Since this normalization forces the sum of probabilities to unity, it does not necessarily reflect the absolute confidence on the prediction. Feng et al. [53] uses a Bayesian Neural Network to predict the class and 3D bounding box after ROI pooling, which allows to quantify the network confidence for both outputs. The authors quantify epistemic and aleatoric uncertainties. While the former measures the model uncertainty to explain the observed object, the latter relates to observation noises in scenarios of occlusion and low point density. They observed an increase in detection performance when modeling aleatoric uncertainty by adding a constraint that penalizes noisy training samples.

softmax 归一化来获得类分布。由于这种归一化迫使概率之和为 1，因此它不一定反映预测的绝对置信度。冯等人。[53]使用贝叶斯神经网络预测 ROI 汇集后的类和 3D 边界框，这样可以量化两个输出的网络信任度。作者量化了认知和任意的不确定性。前者测量模型不确定性来解释观察对象，后者则涉及遮挡和低点密度情景下的观察噪声。他们通过添加惩罚嘈杂训练样本的约束来观察对任意不确定性进行建模时检测性能的提高。

4.2.2 Volumetric Convolutional Methods

Volumetric methods assume that the object or scene is represented in a 3D grid, or a voxel representation, where each unit has attributes, such as binary occupancy or a continuous point density. One advantage of such methods is that they encode shape information explicitly. However, as a consequence, most of the volume is empty, resulting in reduced efficiency while processing these empty cells. Additionally, since data is three dimensional by nature 3D convolutions are necessary, drastically increasing the computational cost of such models.

4.2.2 体积卷积方法

体积方法假设对象或场景在 3D 网格或体素表示中表示，其中每个单元具有属性，例如二进制占用或连续点密度。这种方法的一个优点是它们明确地编码形状信息。然而，因此，大部分体积是空的，导致在处理这些空单元时效率降低。另外，由于数据本质上是三维的，因此 3D 卷积是必要的，这大大增加了这种模型的计算成本。

To this effect [54], [55] address the problem of object detection on driving scenarios using one-stage FCN on the entire scene volumetric representation. This one-stage detection differs from two-stage where region proposals are first generated and then refined on a second processing stage. Instead, one-stage detectors infer detection predictions in a single forward pass. Li [54] uses a binary volumetric input and detects vehicles only. The model's output maps represent "objectness" and BB vertices predictions, similarly to the authors' previous work [48]. The first output predicts if the estimated region belongs to an object of interest, while the second predicts its coordinates. They use expensive 3D convolutions which limits temporal performance.

为此[54]，[55]解决了在整个场景体积表示上使用一级 FCN 对驾驶场景进行目标检测的问题。这种一阶段检测不同于两阶段，其中区域提议首先生成，然后在第二处理阶段重新建立。相反，一级检测器推断单个正向通过中的检测预测。Li [54]使用二进制体积输入并仅检测车辆。模型的输出图表示"对象性"和 BB 顶点预测，类似于作者以前的工作[48]。第一个输出预测估计的区域是否属于感兴趣的对象，而第二个输出预测其坐标。他们使用昂贵的 3D 卷积来限制时间性能。

Aiming at a more efficient implementation, [55] fixes BB sizes for each class but detects cars, pedestrians and cyclists. This assumption simplifies the architecture and together with a sparse convolution algorithm greatly reduces the model's complexity. L1 regularization and Rectified Linear Unit (ReLU) activation functions are used to maintain

为了更有效的实施，[55]确定每个班级的 BB 尺寸，但检测汽车，行人和骑自行车者。这种假设简化了体系结构，并与稀疏卷积算法一起大大降低了模型的复杂性。L1 正则化和修正线性单元（ReLU）激活函数用于维持卷积层的稀疏性。在推理期间，并行网络被独立地用于每个类。固定 BB 大小的假设允许直接在正样本的 3D

sparsity across convolutional layers. Parallel networks are used independently for each class during inference. The assumption of fixed BB sizes allows to train the network directly on the 3D crops of positive samples. During training they augment the data with rotation and translation transformation and employ hard negative mining to reduce false positives.

作物上训练网络。在培训期间，他们通过轮换和翻译转换来增加数据，并采用硬负挖掘来减少误报。

4.2.3 Point-Nets Methods

Point clouds consist of a variable number of 3D points sparsely distributed in space. Therefore, it is not obvious how to incorporate their structure to traditional feed-forward deep neural networks pipelines that assume fixed input data sizes. Previous methods attempted to either transform the point cloud raw points into images using projections or into volumetric structures using voxel representations. A third category of methods, called Pointnets, handle the irregularities by using the raw points as input in an attempt to reduce information loss caused by either projection or quantization in 3D space. We first review seminal work and then progress to driving specific applications.

4.2.3 点网方法

点云由在空间中稀疏分布的可变数量的 3D 点组成。因此，如何将其结构与传统的前馈深度神经网络流水线结合起来并不明显，这些流水线假设固定的输入数据大小。先前的方法试图使用投影将点云原始点转换为图像，或者使用体素表示将点云原始点转换为体积结构。第三类方法称为 Pointnets，通过使用原始点作为输入来处理不规则性，以试图减少由 3D 空间中的投影或量化引起的信息损失。我们首先审查开创性工作，然后进一步推动特定应用。

The seminal work in the category is introduced by PointNet [56]. Segmented 3D PCLs are used as input to perform object classification and part-segmentation. The network performs point-wise transformations using Fully Connected (FC) layers and aggregates a global feature through a max-pooling layer, ensuring independence on point order. Experimental results show that this approach outperforms volumetric methods [57], [58]. This model is further extended in PointNet++ [59], where each layer progressively encode more complex features in a hierarchical structure. The model generate overlapping sets of points and local attribute features are obtained by feeding each set to a local PointNet. Follow up work by Wang et al. [60] further generalize the PointNet architecture by considering points pair-wise relationships. More detailed information on convolutional neural networks for irregular domains is out of the scope of this paper but can be found in [61].

PointNet [56]介绍了该类别的开创性工作。分段 3D PCL 用作输入以执行对象分类和部分分割。网络使用 FullyConnected（FC）层执行逐点转换，并通过 max-pooling 层聚合全局特征，确保点顺序的独立性。实验结果表明，该方法优于体积法[57]，[58]。这个模型在 PointNet ++ [59]中进一步扩展，其中每个层在分层结构中逐步编码更复杂的特征。该模型生成重叠的点集，并通过将每个集合馈送到本地 PointNet 来获得局部属性特征。Wang 等人的后续工作。[60]通过考虑点对关系进一步概括了 PointNet 架构。关于不规则域的卷积神经网络的更多详细信息超出了本文的范围，但可以在[61]中找到。

The seminal methods assumed segmented PCLs that contain a single object, but the gap between object classification and detection is still an open question. VoxelNet [62] uses raw point subsets to generate voxel-wise features, creating a uniform representation of the point cloud, as obtained in volumetric methods. The first step randomly selects a fixed number of points from each voxel, reducing evaluation time and enhancing generalization. Each set of points is used by

开创性方法假设包含单个对象的分段 PCL，但对象分类和检测之间的差距仍然是一个悬而未决的问题。VoxelNet [62]使用原始点子集生成体素方面的特征，创建点云的统一表示，如在体积方法中获得的。第一步是从每个体素中随机选择固定数量的点，从而缩短评估时间并增强泛化能力。每个点集由体素特征编码（VFE）层使用以生成 4D 点云表示。该表示被馈送到 3D 卷积层，接着是 3D 区域提议网络以预测 BB 位置，大小和

a voxel-feature-encoding (VFE) layer to generate a 4D point cloud representation. This representation is fed to 3D convolutional layers, followed by a 3D region proposal network to predict BB location, size and class. The authors implement an efficient convolution operation considering the sparsity of the voxel representation. Different voxel sizes are used for cars and pedestrians/cyclists to avoid detail loss. Models are trained independently for each class, resulting in three models that must be used simultaneously during inference. In Frustum PointNet [63] detection is achieved by selecting sets of 3D points and using a PointNet to classify and predict bounding boxes for each set. The set selection criterion is based on 2D detections on the image plane, thus this method is classified as a Fusion method, reviewed in Section IV-C.

类别。考虑到体素表示的稀疏性，作者实现了一种有效的卷积运算。不同的体素尺寸用于汽车和行人/骑车者，以避免细节损失。模型是针对每个类独立训练的，因此在推理期间必须同时使用三个模型。在 Frustum 中，PointNet [63]通过选择 3D 点集并使用 PointNet 对每个集合的边界框进行分类和预测来实现检测。设定的选择标准基于图像平面上的 2D 检测，因此该方法被分类为融合方法，在第 IV-C 节中进行了回顾。

4.2.4 Discussion

Among point cloud based methods, the projection subcategory has gained most attention due to the proximity to standard image object detection. Particularly, it offers a good trade-off between time complexity and detection performance. However, most methods rely on hand engineered features when projecting the point cloud (density, height, etc.). In contrast, PointNet methods uses the raw 3D points to learn a representation in feature space. In this last category it is still necessary to investigate new forms of using a whole scene point cloud as input, as regular PointNet models assume segmented objects. Volumetric methods transform the point cloud into voxel representations where the space information in explicitly encoded. This approach causes a sparse representation which is inefficient given the need of 3D convolutions. We present a summary of point cloud-based methods in Table V.

4.2.4 讨论

在基于点云的方法中，投影子类别由于接近标准图像对象检测而获得最多关注。特别是，它在时间复杂度和检测性能之间提供了良好的折衷。然而，大多数方法在投射点云（密度，高度等）时依赖于手工设计的特征。相比之下，PointNet 方法使用原始 3D 点来学习特征空间中的表示。在最后一个类别中，仍然需要研究使用整个场景点云作为输入的新形式，因为常规 PointNet 模型假设分段对象。体积方法将点云转换为体素表示，其中空间信息被明确编码。这种方法导致稀疏表示，考虑到 3D 卷积的需要，这种表示是无效的。我们在表 V 中提供了基于点云的方法的摘要。

4.3 Fusion Based Methods

As mentioned previously, point clouds do not provide texture information, which is valuable for class discrimination in object detection and classification. In contrast, monocular images cannot capture depth values, which are necessary for accurate 3D localization and size estimation. Additionally, the density of point clouds tends to reduce quickly as the distance from the sensor increases, while images can still provide a means of detecting far vehicles and objects. In order to increase the overall performance, some methods try to use both modalities with different strategies and fusion schemes. Generally there are three types of fusion schemes [64]:

4.3 基于融合方法

如前所述，点云不提供纹理信息，这对于对象检测和分类中的类辨别是有价值的。相反，单目图像不能捕获深度值，这对于精确的 3D 定位和尺寸估计是必需的。另外，随着距传感器的距离增加，点云的密度趋于迅速减小，而图像仍然可以提供检测远处车辆和物体的手段。为了提高整体性能，一些方法尝试使用具有不同策略和融合方案的两种方式。通常有三种类型的融合方案[64]：

TABLE V
SUMMARY OF POINT CLOUD-BASED METHODS

SubCategory	Method	Methodology/Contributions	Limitations
Projection	VeloFCN [48]	Uses fully convolutional architecture with lidar point cloud bird-eye view projections. Output maps represent 3D bounding box regressions and "objectness" score, the likelihood of having an object at that position.	Detects vehicles only. Limited performance on small or occluded objects due to the loss of resolution across feature maps.
	C-YOLO [30]	Uses a YOLO based single-shot detector extended for 3D BB and orientation regression. The proposed architecture achieves 50 fps runtime, more than any previous method.	There is a tradeoff between inference time and detection accuracy. Single-shot networks underperform networks that use a second stage for refinement.
	TowardsSafe [53]	Uses variational dropout inference to quantify uncertainty in class and bounding box predictions. Aleatoric noise modelling allows the network to generalise better by reducing the impact of noisy samples in the training process.	The uncertaity estimation requires several forward passes of the network. This limits the temporal performance of this method, preventing real-time results.
	BirdNet [51]	Normalizes point cloud representation to allow detection generalisation across different lidar models and specifications.	Input image with only 3 channels encoding height, density and intensity information looses detailed information, which degrades performance.
Volumetric	3DFCN [54]	Extension of the FCN architecture to voxelized lidar points clouds. Single shot detection method.	Requires 3D convolutions, limiting temporal performance to 1 fps.
	Vote3Deep [55]	Proposes an efficient convolutional algorithm to exploit the sparsity of volumetric point cloud data. Uses L1 regularisation and Rectified Linear Unit (ReLU) to maintain sparsity.	Assumes fixed sizes for all detected objects, limiting the detection performance.
PointNet	VoxelNet [62]	Extends PointNet concept to point clouds in a scene scale. Uses raw 3D points to learn a volumetric representation through Voxel Feature Encoding layers. The volumetric features are used for 3D region proposal.	Expensive 3D convolutions limits time performance. Models are class specific, thus multiple models must be run in parallel at run time.

- ✓ **Early fusion:** Modalities are combined at the beginning of the process, creating a new representation that is dependent on all modalities.
- ✓ **Late fusion:** Modalities are processed separately and independently up to the last stage, where fusion occurs. This scheme does not require all modalities be available as it can rely on the predictions of a single modality.
- ✓ **Deep fusion:** Proposed in [64], it mixes the modalities hierarchically in neural network layers, allowing the features from different modalities to interact over layers, resulting in a more general fusion scheme.

- ✓ **早期融合：**在过程开始时将模态结合起来，创建一个依赖于所有模态的新表示。
- ✓ **后期融合：**模态分别独立处理，直到融合发生的最后阶段。该方案不需要所有模态都可用，因为它可以依赖于单一模态的预测。
- ✓ **深度融合：**在[64]中提出，它在神经网络层中分层次地混合模态，允许来自不同模态的特征在层上交互，从而产生更一般的融合方案。

Schlosser et al. [65] evaluate the fusion at different stages of a 3D pedestrian detection pipeline. Their model considered two inputs: monocular image and a depth frame. The authors conclude that late fusion yields the best performance, although early fusion can be used with minor performance drop.

Schlosser 等[65]评估 3D 行人检测流程不同阶段的融合。他们的模型考虑了两个输入：单目图像和深度框架。作者得出结论，后期融合产生了最佳性能，尽管早期融合可以用于较小的性能下降。

One fusion strategy consists of using the point cloud projection method, presented in Section IV-B.1, with extra RGB channels of front facing cameras along the projected PCL maps to obtain higher detection performance. Two of these methods [6], [64] use 3D region proposal networks (RPNs) to generate 3D Regions of Interest (ROI) which are then projected to the specific views and used to predict classes and 3D bounding boxes.

一种融合策略包括使用第 4.2.1 节中提供的点云投影方法，沿投影的 PCL 图具有前置摄像头的额外 RGB 通道，以获得更高的检测性能。这些方法中的两种[6]，[64]使用 3D 区域提议网络（RPN）来生成 3D 感兴趣区域（ROI），然后将其投影到特定视图并用于预测类和 3D 边界框。

The first method, MV3D [64], uses bird-eye and front view

第一种方法 MV3D [64]使用沿着前向摄像机的 RGB 通

projections of lidar points along the RGB channels of a forward facing camera. The network consists of three input branches, one for each view, with VGG [38] based feature extractors. The 3D proposals, generated based on the bird-eye view features only, are projected to each view's feature maps. A ROI pooling layer extracts the features corresponding to each view's branch. These proposal-specific features are aggregated in a deep fusion scheme, where feature maps can hierarchically interact with one another. The final layers output the classification result and the refined vertices of the regressed 3D bounding box. The authors investigate the performance of different fusion methods and conclude that the deep fusion approach obtains the best performance since it provides more flexible means of aggregating features from different modalities.

道的激光雷达点的鸟瞰和前视图投影。网络由三个输入分支组成，每个视图一个，基于 VGG [38]的特征提取器。仅基于鸟瞰视图功能生成的 3D 提议将投影到每个视图的要素图。ROI 池层提取与每个视图的分支对应的特征。这些特定于提议的特征在深度融合方案中聚合，其中特征映射可以彼此层次地交互。最终层输出分类结果和回归的 3D 边界框的重新定义的顶点。作者研究了不同融合方法的性能，并得出结论：深度融合方法获得了最佳性能，因为它提供了更灵活的方法来聚合来自不同模态的特征。

The second method, AVOD [6], is the first to introduce an early fusion approach where the bird-eye view and RGB channels are merged for region proposal. The input representations are similar to MV3D [64] except that only the bird-eye view and image input branches are used. Both modalities' feature maps are used by the RPN, achieving high proposal recall. The highest scoring region proposals are sampled and projected into the corresponding views' feature maps. Each modality proposal specific features are merged and a FC layer outputs class distribution and refined 3D boxes for each proposal. Commonly, loss of details after convolutional stages prevents detection of small objects. The authors circumvent this by upsampling the feature maps using Feature Pyramid Networks [66]. Qualitative results show robustness to snowy scenes and poor illumination conditions on private data.

第二种方法 AVOD [6]是第一种引入早期融合方法的方法，其中鸟眼视图和 RGB 通道合并用于区域提议。输入表示类似于 MV3D [64]，只是使用了 birdeye 视图和图像输入分支。RPN 使用两种模态的特征图，实现了高提案召回。最高得分区域提案被采样并投影到相应视图的要素图中。每个模态提议特定功能都已合并，FC 层为每个提案输出类别分布和重新定义的 3D 框。通常，卷积阶段后的细节丢失会阻止小物体的检测。作者通过使用特征金字塔网络[66]对特征图进行上采样来绕过这一点。定性结果显示对雪景的稳健性和私人数据的不良照明条件。

A second strategy consists of using the monocular image to obtain 2D candidates and extrapolate these detections to the 3D space where point cloud data is employed. In this category Frustum Point-Net [63] generates region proposals on the image plane with monocular images and use the point cloud to perform classification and bounding box regression. The 2D boxes obtained over the image plane are extrapolated to 3D using the camera calibration parameters, resulting in frustums region proposals. The points enclosed by each frustum are selected and segmented with a PointNet instance to remove the background clutter. This set is then fed to a second PointNet instance to perform classification and 3D BB regression. Similarly, Du et al. [67] first select the points that lie in the detection box when projected to the image plane, then use these points to perform model fitting, resulting in a preliminary 3D proposal. The proposal is

第二种策略包括使用单眼图像来获得 2D 候选并将这些检测外推到采用点云数据的 3D 空间。在这一类别中，Frustum Point-Net [63]使用单眼图像在图像平面上生成区域提议，并使用点云执行分类和边界框回归。使用相机校准参数将在图像平面上获得的 2D 盒外推至 3D，从而产生平截头体区域提议。选择每个平截头体所包围的点并使用 PointNet 实例进行分段以消除背景杂乱。然后将该集合馈送到第二个 PointNet 实例以执行分类和 3D BB 回归。同样，杜等人。[67]首先选择投影到图像平面时位于检测框中的点，然后使用这些点执行模型拟合，从而得到初步的 3D 建议。该提案由两阶段 CNN 处理，输出最终的 3D 框和信心分数。这两种方法中的检测都受到单眼图像提议的限制，这可能是由于照明条件等原因造成的限制因素。

TABLE VI
SUMMARY OF FUSION BASED METHODS

Method	Methodology/Contributions	Limitations
MV3D [64]	Uses bird-eye and front view lidar projections as well as monocular camera frames to detect vehicles. 3D proposal network based on the bird-eye-view. Introduces a deep fusion architecture to allow interactions between modalities.	Although far objects might be visible through the camera, the low lidar point density prevents detection of these objects. Specifically, the RPN based on the bird-eye view only limits these detections. Detects vehicles only.
AVOD [6]	Uses bird-eye lidar projection and monocular camera only. New RPN uses both modalities to generate proposals. A Feature Pyramid Network extension improves detection of small objects by up sampling feature maps. New vector representation removes ambiguities in the orientation regression. Can detect vehicles, pedestrians and cyclists.	Detection method only sensitive to objects in front of the vehicle due to the forward-facing camera used.
F-PointNet [63]	Extracts 2D detection from image plane, extrapolates detection to a 3D frustum, selecting lidar points. Uses a PointNet instance to segment background points and generate 3D detections. Can detect vehicles, pedestrians and cyclists.	Since proposals are obtained from the front view image, failing to detect objects in this view limits the detection performance. This limits the use of this method at night time, for example.

processed by a two-stage refinement CNN that outputs the final 3D box and confidence score. The detections in both these approaches are constrained by the proposal on monocular images, which can be a limiting factor due to lighting conditions, etc.

Fusion methods obtain state-of-the-art detection results by exploring complimentary information from multiple sensor modalities. While lidar point clouds provide accurate depth information with sparse and low point density at far locations, cameras can provide texture information which is valuable for class discrimination. Fusion of information at feature levels allow to use complimentary information to enhance performance. We provide a summary of fusion methods in Table VI.

融合方法通过探索来自多个传感器模态的补充信息来获得最先进的检测结果。虽然激光雷达点云在远处提供具有稀疏和低点密度的精确深度信息，但是相机可以提供对于类别辨别有价值的纹理信息。功能级别的信息融合允许使用免费信息来提高性能。我们在表 VI 中提供了融合方法的摘要。

5. References

[1] Beltrán J, Guindel C, Moreno F M, et al. Birdnet: a 3D object detection framework from LiDAR information[C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 3517-3523.

[1] A. Dhani, “Reported road casualties in Great Britain: Quarterly provisional estimates year ending September 2017,” U.K. Dept. Transport, London, U.K., Tech. Rep., Feb. 2018. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/ system/uploads/attachment_data/file/681593/quarterly-estimates-july-toseptember-2017.pdf

[2] S. Singh, “Traffic safety facts,” Nat. Highway Traffic Saf. Admin., U.S. Dept. Transp., Washington, DC, USA, Tech. Rep. DOT HS 812 115, Feb. 2015. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115

[3] Atkins Ltd. “Research on the impacts of connected and autonomous vehicles (CAVs) on traffic flow,” U.K. Dept. Transport, London, U.K., Tech. Rep. SO13994/3, May 2016. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/ system/uploads/attachment_data/file/530091/impacts-of-connected-andautonomous-vehicles-on-traffic-flow-summary-report.pdf

[4] K. Habib, “Technical report, Tesla crash,” Nat. Highway Traffic Saf. Admin., U.S. Dept. Transp., Washington, DC, USA, Tech. Rep. PE 16-007, Jan. 2017. [Online]. Available: https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF

[5] KITTI 3D Object Detection Online Benchmark. Accessed: Jun. 15, 2018. [Online]. Available: http://www.cvlibs.net/datasets/ kitti/eval_object.php?obj_benchmark=3d

[6] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, “Joint 3D proposal generation and object detection from

view aggregation," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2018, pp. 1–8.
[7] B. Ranft and C. Stiller, "The role of machine vision for intelligent vehicles," IEEE Trans. Intell. Vehicles, vol. 1, no. 1, pp. 8–19, Mar. 2016.
[8] S. D. Pendleton et al., "Perception, planning, control, and coordination for autonomous vehicles," Machines, vol. 5, no. 1, p. 6, Feb. 2017. [Online]. Available: http://www.mdpi.com/2075-1702/5/1/6
[9] A. Mukhtar, L. Xia, and T. B. Tang, "Vehicle detection techniques for collision avoidance systems: A review," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 5, pp. 2318–2338, May 2015.
[10] D. Z. Wang, I. Posner, and P. Newman, "What could move? Finding cars, pedestrians and bicyclists in 3D laser data," in Proc. IEEE Int. Conf. Robot. Automat. (ICRA), May 2012, pp. 4038–4044.
[11] A. Azim and O. Aycard, "Layer-based supervised classification of moving objects in outdoor dynamic environment using 3d laser scanner," in Proc. IEEE Intell. Vehicles Symp., Jun. 2014, pp. 1408–1414.
[12] J. Behley, V. Steinhage, and A. B. Cremers, "Laser-based segment classification using a mixture of bag-of-words," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Nov. 2013, pp. 4195–4200.
[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2015, pp. 91–99.
[14] Y. Zhang et al., "Towards end-to-end speech recognition with deep convolutional neural networks," in Proc. Interspeech, 2016, pp. 410–414, doi: 10.21437/Interspeech.2016-1446.
[15] J. Van Brummelen, M. O'Brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," Transp. Res. C, Emerg. Technol., vol. 89, pp. 384–406, Apr. 2018.
[16] S. Kuutti, S. Fallah, K. Katsaros, M. Dianati, F. Mccullough, and A. Mouzakitis, "A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications," IEEE Internet Things J., vol. 5, no. 2, pp. 829–846, Apr. 2018.
[17] M. Weber, P. Wolf, and J. M. Zöllner, "DeepTLR: A single deep convolutional network for detection and classification of traffic lights," in Proc. IEEE Intell. Vehicles Symp. (IV), Jun. 2016, pp. 342–348.
[18] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," IEEE Trans. Intell. Transp. Syst., vol. 14, no. 4, pp. 1773–1795, Dec. 2013.
[19] S. Hsu, S. Acharya, A. Rafii, and R. New, "Performance of a timeof-flight range camera for intelligent vehicle safety applications," in Advanced Microsystems for Automotive Applications. Berlin, Germany: Springer, 2006, pp. 205–219.
[20] O. Elkhalili et al., "A 64×8 pixel 3-D CMOS time of flight image sensor for car safety applications," in Proc. 32nd Eur. Solid-State Circuits Conf., 2006, pp. 568–571.
[21] Sony IMX390CQV CMOS Image Sensor for Automotive Cameras. [Online]. Available: https://www.sony.net/SonyInfo/News/Press/201704/ 17-034E/index.html
[22] Q. Chen, X. Yi, B. Ni, Z. Shen, and X. Yang, "Rain removal via residual generation cascading," in Proc. IEEE Vis. Commun. Image Process. (VCIP), Dec. 2017, pp. 1–4.
[23] Velodyne HDL-64E Lidar Specification. Accessed: Apr. 10, 2018. [Online]. Available: http://velodynelidar.com/hdl-64e.html
[24] Velodyne VLS-128 Announcement Article. Accessed: Apr. 10, 2018. [Online]. Available http://www.repairerdrivennews.com/2018/01/02/ velodyne-leading-lidar-price-halved-new-high-res-product-to-improveself-driving-cars/
[25] Leddar Solid-State Lidar Technology. Accessed: Apr. 10, 2018. [Online]. Available: https://leddartech.com/technology-fundamentals/
[26] Y. Park, S. Yun, C. S. Won, K. Cho, K. Um, and S. Sim, "Calibration between color camera and 3D LIDAR instruments with a polygonal planar board," Sensors, vol. 14, no. 3, pp. 5333–5353, 2014.
[27] R. Ishikawa, T. Oishi, and K. Ikeuchi. (Apr. 2018). "LiDAR and camera calibration using motion estimated by sensor fusion odometry." [Online]. Available: https://arxiv.org/abs/1804.05178
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2009, pp. 248–255.
[29] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2012, pp. 3354–3361.

[30] M. Simon, S. Milz, K. Amende, and H. Gross. (Mar. 2018). “ComplexYOLO: Real-time 3D object detection on point clouds.” [Online]. Available: https://arxiv.org/abs/1803.06199
[31] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, “Virtual worlds as proxy for multi-object tracking analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4340–4349.
[32] S. Chen, S. Zhang, J. Shang, B. Chen, and N. Zheng, “Braininspired cognitive model with attention for self-driving cars,” IEEE Trans. Cogn. Devel. Syst., to be published. [Online]. Available: https://ieeexplore.ieee.org/document/7954050, doi: 10.1109/TCDS.2017.2717451.
[33] H. Xu, Y. Gao, F. Yu, and T. Darrell, “End-to-end learning of driving models from large-scale video datasets,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3530–3538.
[34] B. Wu, A. Wan, X. Yue, and K. Keutzer. (Oct. 2017). “SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time roadobject segmentation from 3D LiDAR point cloud.” [Online]. Available: https://arxiv.org/abs/1710.07368
[35] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “CARLA: An open urban driving simulator,” in Proc. 1st Conf. Robot Learn. (CoRL), Nov. 2017, pp. 1–16.
[36] M. Müller, V. Casser, J. Lahoud, N. Smith, and B. Ghanem, “Sim4CV: A photo-realistic simulator for computer vision applications,” Int. J. Comput. Vis., vol. 126, no. 9, pp. 902–919, Sep. 2018.
[37] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Washington, DC, USA, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.
[38] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Represent., 2015.
[39] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, “Monocular 3D object detection for autonomous driving,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2147–2156.
[40] A. Mousavian, D. Anguelov, J. Flynn, and J. Košecká, “3D bounding box estimation using deep learning and geometry,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5632–5640.
[41] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, “Data-driven 3D voxel patterns for object category recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 1903–1911.
[42] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau, “Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1827–1836.
[43] X. Chen et al., “3D object proposals for accurate object class detection,” in Advances in Neural Information Processing Systems, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. New York, NY, USA: Curran Associates, Inc., 2015, pp. 424–432. [Online]. Available: http://papers.nips.cc/paper/5644-3d-objectproposals-for-accurate-object-class-detection.pdf
[44] C. C. Pham and J. W. Jeon, “Robust object proposals re-ranking for object detection in autonomous driving using convolutional neural networks,” Signal Process., Image Commun., vol. 53, pp. 110–122, Apr. 2017. [Online]. Available: http://www.sciencedirect.com/science/ article/pii/S0923596517300231
[45] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, “Subcategory-aware convolutional neural networks for object proposals and detection,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2017, pp. 924–933.
[46] G. Payen de La Garanderie, A. Atapour Abarghouei, and T. P. Breckon, “Eliminating the blind spot: Adapting 3D object detection and monocular depth estimation to 360° panoramic imagery,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 812–830.
[47] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, “Multi-view convolutional neural networks for 3D shape recognition,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 945–953.
[48] B. Li, T. Zhang, and T. Xia, “Vehicle detection from 3D Lidar using fully convolutional network,” in Proc. Robot., Sci. Syst. XII, AnnArbor, MI, USA, Jun. 2016. [Online]. Available: http://www.roboticsproceedings.org/rss12/
[49] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer. (Nov. 2016). “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size.” [Online]. Available: https://arxiv.org/abs/1602.07360
[50] S. L.Yu, T. Westfechtel, R. Hamada, K. Ohno, and S. Tadokoro, “Vehicle detection and localization on bird’s eye view elevation images using convolutional neural network,” in Proc. IEEE Int. Symp. Saf., Secur. Rescue Robot. (SSRR), Oct. 2017, pp. 102–109.

[51] J. Beltrán, C. Guindel, F. M. Moreno, D. Cruzado, F. García, and A. de la Escalera. (May 2018). “BirdNet: A 3D object detection framework from LiDAR information.” [Online]. Available: http://arxiv.org/abs/1805.01195
[52] J. Redmon and A. Farhadi, “YOLO9000: Better, faster, stronger,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 6517–6525.
[53] D. Feng, L. Rosenbaum, and K. Dietmayer. (2018). “Towards safe autonomous driving: Capture uncertainty in the deep neural network for Lidar 3D vehicle detection.” [Online]. Available: https://arxiv.org/abs/1804.05132
[54] B. Li, “3D fully convolutional network for vehicle detection in point cloud,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Sep. 2017, pp. 1513–1518.
[55] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, “Vote3Deep: Fast object detection in 3D point clouds using efficient convolutional neural networks,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2017, pp. 1355–1361.
[56] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, “PointNet: Deep learning on point sets for 3D classification and segmentation,” in Proc. Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2017, pp. 77–85.
[57] Z. Wu et al., “3D ShapeNets: A deep representation for volumetric shapes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1912–1920.
[58] D. Maturana and S. Scherer, “VoxNet: A 3D convolutional neural network for real-time object recognition,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2015, pp. 922–928.
[59] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep hierarchical feature learning on point sets in a metric space,” in Advances in Neural Information Processing Systems. New York, NY, USA: Curran Associates, Inc., 2017, pp. 5099–5108.
[60] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. (Jan. 2018). “Dynamic graph CNN for learning on point clouds.” [Online]. Available: https://arxiv.org/abs/1801.07829
[61] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, “Geometric deep learning: Going beyond Euclidean data,” IEEE Signal Process. Mag., vol. 34, no. 4, pp. 18–42, Jul. 2017.
[62] Y. Zhou and O. Tuzel. (Nov. 2017). “VoxelNet: End-to-end learning for point cloud based 3D object detection.” [Online]. Available: https://arxiv.org/abs/1711.06396
[63] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, “Frustum PointNets for 3D object detection from RGB-D data,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 918–927.
[64] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3D object detection network for autonomous driving,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6526–6534.
[65] J. Schlosser, C. K. Chow, and Z. Kira, “Fusing LIDAR and images for pedestrian detection using convolutional neural networks,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2016, pp. 2198–2205.
[66] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 2117–2125.
[67] X. Du, M. H. Ang, S. Karaman, and D. Rus, “A general pipeline for 3D detection of vehicles,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), Brisbane, QLD, Australia, May 2018, pp. 3194–3200, doi: 10.1109/ICRA.2018.8461232.
[68] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Sep. 2009.
[69] C. Redondo-Cabrera, R. J. López-Sastre, Y. Xiang, T. Tuytelaars, and S. Savarese, “Pose estimation errors, the ultimate diagnosis,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 118–134.
[70] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. (Aug. 2017). “On calibration of modern neural networks.” [Online]. Available: http://arxiv.org/abs/1706.04599