# 3D 目标检测方法研究综述

黄哲[1]，王永才[1,2]，李德英[1]
（1. 中国人民大学信息学院，北京 100872；2. 警务物联网应用技术公安部重点实验室，北京 100048）

**摘　要：**3D 目标检测是自动驾驶、虚拟现实、机器人等应用领域的重要基础问题，其目的是从无序点云中框取出描述目标最准确的 3D 框，例如紧密包围行人或车辆点云的 3D 框，并给出目标 3D 框的位置、尺寸和朝向。如今，基于双目视觉、RGB-D 相机、激光雷达构建的纯点云的 3D 目标检测，融合图像和点云多模态信息的 3D 目标检测，是两类主要的方法。首先介绍了 3D 点云的不同表示形式和特征提取方法，然后从传统机器学习类算法、非融合深度学习类算法、基于多模态融合的深度学习类算法 3 个层面，逐层递进地介绍各类 3D 目标检测方法，对类别内部和各类之间的方法进行分析和对比，深入分析了各类方法之间的区别和联系，最后论述了 3D 目标检测仍存在的问题和可能的研究方向，并对 3D 目标检测研究的主流数据集和主要评价指标进行了总结。

**关键词：**深度学习；3D 目标检测；多模态融合；点云；自动驾驶
**中图分类号：**TP183
**文献标志码：**A
**doi:** 10.11959/j.issn.2096-6652.202312

# A survey of 3D object detection algorithms

HUANG Zhe[1], WANG Yongcai[1,2], LI Deying[1]
1. School of Information, Renmin University of China，Beijing 100872, China
2. Key Laboratory of Police Internet of Things Application Ministry of Public Security, Beijing 100048, China

**Abstract:** 3D object detection is a fundamental problem in autonomous driving，virtual reality，robotics，and other applications. Its goal is to extract the most accurate 3D box characterizing interested targets from the disordered point clouds，such as the closest 3D box surrounding the pedestrians or vehicles. The target 3D box's location，size，and orientation are also output. Currently，there are two primary approaches for 3D object detection: (1) pure point cloud based 3D object detection，in which the point clouds are created by binocular vision，RGB-D camera，and lidar; (2) fusion-based 3D object detection based on the fusion of image and point cloud. The various representations of 3D point clouds were introduced. Then representative methods were introduced from three aspects: traditional machine learning techniques; non-fusion deep learning based algorithms; and multimodal fusion-based deep learning algorithms in progressive relation. The algorithms within and across each category were examined and compared，and the differences and connections between the various methods were analyzed thoroughly. Finally，remaining challenges of 3D object detection were discussed and explored. And the primary datasets and metrics used in 3D object detection studies were summarized.

**Key words:** deep learning, 3D object detection, multimodal fusion, point cloud, autonomous driving

## 0　引言

在自动驾驶、机器人、无人机等应用领域中，常通过激光雷达、双目视觉、RGB-D 相机等构建 3D 点云（point cloud）以描述周边环境，但点云信息是无序且缺少语义的。为检测出点云中的移动目

**收稿日期：**2023–01–11；**修回日期：**2023–03–01
**通信作者：**王永才，ycw@ruc.edu.cn
**基金项目：**国家自然科学基金资助项目（No.61972404，No.12071478，No.61732006）
**Foundation Items:** The National Natural Science Foundation of China (No.61972404, No.12071478, No.61732006)

标，或实现基于点云的目标检测与人机交互，从无序点云中框取出最准确描述目标的 3D 点云框，并给出目标 3D 框的空间位置、尺寸和朝向，这个过程被称为 3D 目标检测[1-2]，是上述各类应用的重要基础。

具体地，3D 点云可以由视觉传感器（包括单目、双目、RGB-D）或雷达传感器（包括超声雷达、激光雷达、毫米波雷达），融合定位传感器、惯性测量单元等，通过处理拍摄的图像或点云数据由视觉里程计或雷达里程计[3]方法构建。在点云中，任何物体以其表面的无序离散点表示，从这些无序离散点中，准确检测出代表目标（如行人、车辆）的点云块并给出目标位姿并非易事。而在实际三维空间中，物体是具有朝向的三维形状，因此在 3D 目标检测中，检测的目标近似为框取出最紧密包围目标的 3D 框。目标属性主要由 7 个参数来描述，包含物体的位置($x$,$y$,$z$)、长宽高（height, width, length）以及朝向 $\theta$，由于一般假设目标是在地面上运动的[4]，因此朝向指物体运动正方向在俯视图中的朝向，也有部分文献中用 3 个姿态角描述物体的空间三维姿态。

近年来，众多针对点云的 3D 目标检测方法被提出。按照输入数据模态不同，现有主流的 3D 目标检测方法可划分为以下 3 种。

（1）在雷达构建点云[5]上的 3D 目标检测。当激光雷达点云作为输入时，雷达数据本身具有精确的空间三维信息，3D 目标检测方法主要分为两类。一类方法是直接在点云上进行采样和特征提取，例如将 3D 点云数据转换为体素（voxel）、柱体（pillar）或基于 PointNet 进行点特征聚合等[6]。另一类方法是将 3D 点云数据转换为鸟瞰图（bird′s eye view，BEV）、前视图（front view，FV）、距离图像（range view）等 3D 多视图形式，并基于图像方法进行特征提取和 3D 目标检测。

（2）在图像构建点云上的 3D 目标检测。由图像输入构建点云时，由于 2D 图像缺少深度信息，通常以视觉里程计方式建立环境 3D 点云，或者结合深度相机、双目相机等获得深度信息[7]，针对图像建立点云的 3D 目标检测，可以将原始图像、深度图直接输入深度学习模块进行特征提取和目标检测[8]，也可以将原始图像与深度图像结合，将其转换为伪激光雷达数据，之后再以处理稠密激光雷达点云的方式进行后续的特征提取和目标检测。

（3）融合图像和点云的多模态融合的 3D 目标检测。对于同时包含图像和激光雷达的输入，通常通过多模态融合的方法实现 3D 目标检测，将在方法层面进行详细介绍。

另外，按照 3D 目标检测方法不同，可以将现有针对点云的 3D 目标检测方法划分为以下 3 种。

（1）基于传统的机器学习类型的方法[9]。这类方法主要使用目标模板或手工特征提取的方法进行特征提取与模板目标检测[10]。

（2）基于深度学习的单模态的 3D 目标检测方法[11]。这类方法主要包括以图像或点云为主的单模态深度学习的 3D 目标检测方法。这类方法针对图像数据和点云数据[12]的特点，将图像和点云数据转化为不同的表现形式[13]，设计不同的特征提取方法和目标检测方法。在这些单模态的 3D 目标检测工作中，其数据转化方法、特征提取方法和目标检测方法是理解和灵活应用 3D 目标检测方法的关键。

（3）融合图像、点云的基于深度学习的多模态 3D 目标检测方法[14]。这类方法基于图像数据和点云数据各自的特点（例如图像数据可以准确地进行 2D 目标检测，而点云数据具有更为准确的深度信息），结合二者的优势，设计融合的 3D 目标检测方法[15]，提升 3D 目标检测的效果[16]。

虽然过去已有一些学者对基于深度学习的 3D 目标检测研究进行了总结[1,2,17-18]，但是本文和它们有一些明显的区别。例如 Guo Y L 等人[1]研究和总结了基于深度学习的 3D 点云方法，涉及领域范围广、种类繁多，涵盖了检测、分割、追踪、分类等领域，但没有对 3D 目标检测领域进行仔细研究和描述。Arnold E 等人[2]总结了面向自动驾驶应用的三维目标检测方法，概述了经典算法和普遍使用的传感器、数据集，然而并没有深入地研究多模态融合检测算法及其未来的发展和挑战。Feng D 等人[17]和 Wang Y 等人[18]详细地总结了物体检测和语义分割中的融合方法，讨论了其中存在的挑战和未解决的问题。这些综述虽然对融合方法研究得非常深入，但仅针对某一具体领域，没有阐述传统和单模态的深度学习检测算法，不利于读者了解 3D 目标检测领域发展的全貌。此外，近年来 3D 目标检测领域发展迅速，新方法体系层出不穷，以上文献已不能全面概括其发展现状。综上所述，笔者认为对 3D 目标检测研究现状进行全面总结具有客观必要性。因此，本文专注于 3D 目标检测任务，从数据

表示、原理、方法分类、方法之间的关系等方面，对现有代表性 3D 目标检测的基础理论、各类检测方法的模型设计、数据集、评价指标等进行全面的总结和综述。

本文的主要贡献总结如下。

（1）与现有综述不同，本文是首个系统性、综合性、全面性讲解 3D 目标检测的中文综述，首次以 3D 目标检测中输入数据的不同表现形式为角度进行切入，针对各种数据的表现形式介绍相关解决方案的原理。具体地，将基于深度学习的方法分组为二维数据升维方法、三维数据降维方法、直接学习点云的方法、最新提出的基于 Transformer 的方法，在这些类别中详细介绍各类方法的细节，使读者可以更清晰地理解 3D 目标检测主要原理和各个方法之间的内在联系。

（2）按照 3D 目标检测模型设计方法不同，将 3D 目标检测方法分为传统 3D 目标检测方法、基于深度学习单模态 3D 目标检测方法、基于深度学习多模态融合的 3D 目标检测方法，对各个方法中的主要设计进行了详细介绍，并对各类方法的演进发展过程进行了分析和梳理。

（3）重点介绍了基于深度学习的 3D 目标检测最新进展，总结概括了最新的单模态和多模态数据集和评价指标。提出了在 3D 目标检测领域一些被忽视的开放问题和可能的研究方向，这些问题对未来应用于自动驾驶技术、机器人技术、医学图像的现实部署有参考意义。

3D 目标检测整体分类结构如图 1 所示，在传统 3D 目标检测方法中，主要介绍基于模板匹配、基于区域评分和基于滑窗的 3 类方法；在基于深度学习单模态 3D 目标检测方法中，主要介绍二维数据升维、三维数据降维、基于点云的以及基于 Transformer 的 3D 目标检测方法；在基于深度学习多模态融合的 3D 目标检测方法中，主要介绍融合对象组合方式、融合粒度、融合方案等。还详细介绍了不同方法中数据的表示形式、3D 目标检测的主要数据集和主要评价指标等。

## 1　数据表现形式

为了实现准确的 3D 目标检测，关键基础是特征提取，为了实现不同的特征提取，算法将原始的雷达和图像数据转换为不同的表示形式。不同表示形式的点云输入模型后进行特征提取，最后进行分类和回归。因此讨论输入数据的表示形式极其重要，它与后续网络结构的设计、特征提取的方法紧密相关。笔者将现有方法中涉及的数据表示形式进行整理和分类。

### 1.1　雷达数据的表示形式

激光雷达（light detection and ranging，LiDAR）捕获的点云是无序的，每个点包含 3 个坐标$(x,y,z)$和反射强度 $\sigma$ 等信息，点云提供了精确的深度信息，可以显著减轻二维图像中常见的遮挡问题。由于采集到的点云具有不规则性和稀疏性，因此合适地表示点云以进行有效处理是重要的基础问题。在现有研究中，3D 点云数据的主要表示形式可以划分为点云、体素、鸟瞰图、前视图、柱体、距离图像等。

• 点云

随着 PointNet[19]及其改进版本 PointNet++[20]的出现，原始的三维点云可以直接作为输入，进而预测点的特征，相比于基于体素降采样的方法，它们保留了更多的原始信息。点云同时提供了点的深度信息和反射率信息，一个点的深度信息可以用其笛卡尔坐标$[x,y,z]$进行编码，距离 $\rho=\sqrt{x^2+y^2+z^2}$，反射率 $\sigma$ 由反射强度给出。然而，这种方法通常计算成本很高，特别是在大型点云场景下。为了降低计算成本，点云

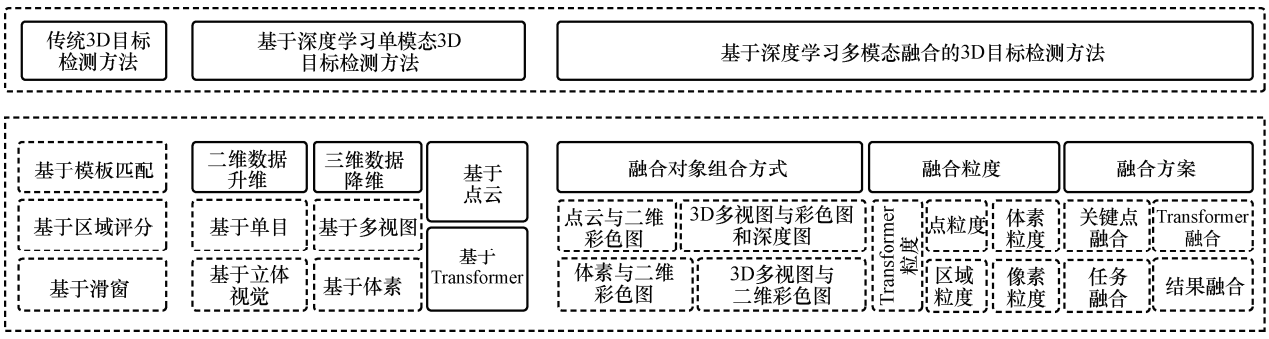



图 1　3D 目标检测整体分类结构

降采样（体素化）是一种必要而有用的方法。

- 体素

体素是用固定大小的立方块来表示不规则三维点云的一种数据结构，以体素为输入时，可以通过三维卷积神经网络有效地提取点特征并进行三维检测。体素可以看成粗粒度的点云。给定一个深度、高度和宽度分别为 $(D,H,W)$ 的大立方体表示输入点云，将其划分为等距离的网格[21]，三维空间中每个体素的深高宽为 $(v_d,v_h,v_w)$，则整个点云的三维体素化结果在各个坐标上生成的体素格（voxel grid）的个数如式（1）所示：

$$\frac{D}{v_d},\frac{H}{v_h},\frac{W}{v_w} \tag{1}$$

体素可以保存丰富的三维形状信息，但是这种表示也有一些缺点：首先在体素化过程中，LiDAR 采集的点云仅在物体表面存在，因此会产生大量的空体素，浪费大量的显存；其次每个体素内的点云都需要进行离散化操作，部分信息被丢失，降低了细粒度的定位精度。

- 鸟瞰图

3D 点云数据的另外一种表示方法是将它们投影到二维平面上，常用的两种表示方法是转换为鸟瞰图[22]和前视图[23]，以便它们可以通过二维卷积层进行处理，统称为 3D 多视图。激光雷达点投影到鸟瞰图的表示方式如图 2 所示，BEV[22]按高度、密度和强度编码点云。

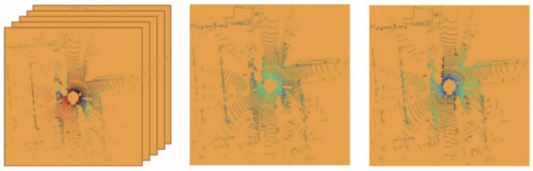



图 2 激光雷达点投影到鸟瞰图[22]

在 BEV 表示中，前后遮挡关系的物体在 BEV 投影上占据了不同的空间，可有效解决遮挡问题。BEV 表示同时保留了物体的长度和宽度，提供了物体在地平面上的位置，使定位任务更加容易。给定三维点 $p=(x, y, z, \sigma)$，$N$ 为单元格中的点数，top 为对应区域内点的最大属性值，则在 BEV 视图中的密度 $D_{xy}$、高度 $H_{xy}$、强度 $I_{xy}$ 分别如式（2）～式（4）所示：

$$D_{xy}=\min\left\{1,\frac{\log(N+1)}{\log 64}\right\} \tag{2}$$

$$H_{xy}=\text{top}_z^{(x,y)} \tag{3}$$

$$I_{xy}=\text{top}_\sigma^{(x,y)} \tag{4}$$

- 前视图

激光雷达点投影到前视图的表示方式如图3所示。

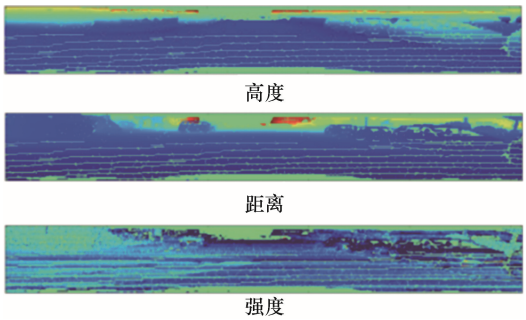



图 3 激光雷达点投影到前视图[23]

前视图[23]是将三维空间中的点投影到可以展开的圆柱表面上形成的图。给定三维点 $p=(x,y,z)$，它在前视图的坐标 $p_{\text{fv}}=(r,c)$ 可以利用式（5）～式（6）计算：

$$c=\lfloor a\tan 2(y,x)/\Delta\theta \rfloor \tag{5}$$

$$r=\left\lfloor a\tan 2(z,\sqrt{x^2+y^2})/\Delta\varphi \right\rfloor \tag{6}$$

其中，$\Delta\theta$ 和 $\Delta\varphi$ 分别为激光雷达的水平和垂直分辨率。前视图用三通道特征编码了高度、深度和反射强度 3 种信息。

BEV 相比 FV 有 3 个优点：首先，在 BEV 中物体会保持原有的物理大小，尺寸方差小；其次，BEV 中的物体占据了不同的空间，可以有效避免物体之间的遮挡；最后，在道路场景中，由于物体位于地面上，垂直位置的差异较小，因此 BEV 对于获得准确的三维边界框更为重要。由于它们都是通过使用一些统计特征来完成对长方体中点云的特征表达，如最大高度值、最大高度值对应点的强度值、点云个数、平均强度值等，这种手工统计（hand-crafted）特征的主要问题是丢弃了很多点云的点，存在物体尺度的变化，不可避免地会带来信息的损失。

- 柱体

柱体[24]与体素相似，只不过在划分格子时不对 $z$ 方向划分，将一个纵轴上的所有体素连在一起形成柱体，是一种新型点云表示方法，每个柱体是在笛卡尔坐标系下的平面上，以一定的步长对点云进行划分得到的一个三维的小单元格。

- 距离图像

距离图像[25]是将每圈激光线拉成直线再按行累

积而成，其中投影图的高为激光线数，宽为 LiDAR 扫描一圈的点数。距离图像可以很方便地获得更大范围的上下文信息，但是它丢失了原来的结构信息，出现尺度变化和遮挡问题[26]，因此这种投影方式很少直接做目标检测，而更适合分割任务[27]。

### 1.2　图像数据的表示形式

视觉传感器主要包括单目相机（monocular camera）和立体相机（stereo camera）。单目相机提供丰富的色彩和纹理信息，能更好地识别物体的特征，并且价格低廉、使用成本较低，但是单目相机缺乏深度信息。立体相机能够准确捕获图像中的每一个点到相机的距离，获取图像中每个点的三维空间坐标，提供密集的深度图。在 3D 目标检测中，图像数据的主要表示形式如下。

- 二维彩色图

以图像数据作为输入时，最常用的一种表示是直接使用二维彩色图，可以使用传统特征提取的方式进行操作，也可以用成熟的二维卷积神经网络[28]处理图像。

- 彩色图+深度图

二维特征图虽然具有丰富的纹理和边缘信息，但是缺乏三维感知中最重要的深度信息，因此常利用深度图[29]作为附加输入来帮助三维检测。

- 图像数据转换为伪激光雷达

给定立体图像或单目图像，将其反向投影到激光雷达坐标系中，转换成三维点云，通常被称为伪激光雷达（pseudo-Lidar）[30]，它的处理过程和真实激光雷达一样，不同点是密集程度要比真实激光雷达高，经常被用于深度补全等操作，虽然 pseudo-Lidar 可以提供重要的深度信息，但是原图像的深度信息在经过 2D 卷积处理后会发生剧烈的畸变扭曲，深度信息准确性低于真实雷达数据。

基于图像和点云这两种不同的数据，现有工作采用不同的方法进行特征提取。传统的特征提取方法主要利用特殊的特征描述子对图像或点云进行特征提取[31]，如 Harris、SIFT、ICP 等，这些方法计算简单，容易实现，但实时性不高，检测出的特征点较少，因此准确度不高。随着大数据时代的来临和深度学习技术的发展，通过神经网络提取特征的方法得到了广泛的应用，已经成为重要且有意义的研究方向。本文重点关注基于深度学习的 3D 目标检测算法。在此之前，首先简单地介绍传统 3D 目标检测算法。

## 2　传统 3D 目标检测算法

首先介绍使用传统机器学习方式进行 3D 目标检测的方法。将之分为基于模板匹配的检测方法、基于区域评分的检测方法以及基于滑窗的检测方法。

### 2.1　基于模板匹配的检测方法

基于模板匹配的检测方法[32]输入信息包括由激光扫描器、深度摄像头甚至 CAD 建模获得的 3D 模型，输入数据表现形式主要为 3D 点云。具体做法是，首先为目标建立特定的模板，通过随机选择或者传统特征提取算法获得关键特征，在模型数据库中对物体进行检索匹配[33]，寻找特征对应的关系，一旦确定了特征的对应关系就可以利用 ICP[34]或者绝对方向等识别出视觉相似的区域，进一步验证所选的对应关系并细化姿态估计，从而检测出目标。因此，基于模板匹配的检测算法主要用于识别特定的物体。

为了提升匹配的效率，Stein F 等人[35]提出了一种从场景数据中识别多个 3D 对象的方法：Structural indexing，该方法使用两种不同类型的结构索引进行匹配，但这种单一的索引只能匹配拥有简单线条的物体，无法匹配姿态复杂的物体。因此 Chua C S 等人[36]提出一种用于描述 3D 自由曲面的点表示形式：点签名（point signatures），相比于点坐标，点签名能够更完整地描述点的结构邻域，可以直接用于假设与具有相似签名的模型点的对应关系。由于遮挡问题，这种方法在杂乱无章的环境下的识别效率并不高。为了提升算法的适应性和鲁棒性，Frome A 等人[37]提出了一种新的霍夫投票机制，用于检测三维空间中的自由形状，在杂乱场景和有遮挡的情况下取得了不错的效果。基于模板匹配的检测方法可以识别特定的物体，针对性强，但是在实际的应用场景中，大量的物体需要一一建模匹配，因此计算量大，并且这种方法并不是端到端的，中间夹杂了如特征提取模块、匹配模块等模块，很难进行优化提升。

### 2.2　基于区域评分的检测方法

基于区域评分的检测方法[38]将 3D 模型分割成一些候选区域后，在每一个区域中利用人工设定特征对所有的区域进行评估、打分、排序，选取评分最高的区域作为最终结果，其输入数据的表现形式为 3D 点云。

Collet A 等人[39]开发了一个框架来执行场景分割，这个框架可以将场景分割成几个有意义的部分，同时丢弃背景杂乱的部分，作者结合图像和距

离图像数据，根据许多不同的形状和外观线索对单个 RGB-D 图像中的区域进行评分。为了减少识别相同的物体，提高效率，Shin J 等人[40]提出了在 3D 点云中无监督地发现重复对象的新方法，通过超像素分割场景为每个区域单独提取特征，最后通过 ICP 算法验证发现的对象并删除错误匹配。这种方法虽然提升了算法的效率，但是在大规模场景下，无法应对复杂的场景变化[41]，因此，Karpathy A 等人[42]提出了一种从室内环境的三维网格中发现对象模型的算法，首先将场景分解为一组候选网格段，然后对每个部分进行评分并排名。此外还提出了一种递归测量的方法，用来编码频繁出现的几何形状。基于区域评分的检测算法已经有区域生成网络（region proposal network, RPN）[43]的影子，但是在深度学习之前，这种检测算法大量依靠手工的方法做分割、分类、回归，因此，可扩展性和通用性比较差，并且这种方法存在过度分割的问题，容易导致漏检的情况出现。

### 2.3　基于滑窗的检测方法

为了克服 3D 目标检测在纹理、照明、形状、自遮挡、噪声和遮挡等方面的影响，基于滑窗的检测方法[44]将深度图加入二维图像中形成 RGB-D 图像，利用 RGB-D 图像进行目标检测，其输入数据的表现形式为彩色图+深度图。2014 年，Song S R 等人[44]提出了一种使用深度图像进行通用目标检测的算法，其主要思路是对于一个给定的对象类别（如椅子），使用计算机图形学（computer graphics，CG）模型进行建模。为了获得深度图，从数百个角度渲染每个 CG 模型，从深度图对应的三维点云中提取一个特征向量，用来训练分类器[45]，然后在三维空间中滑动一个三维检测窗口来匹配样本形状。最后，利用深度图分割进一步提高算法性能。这种方法实际上把检测问题划归为一种分类问题，训练的分类器只需要分辨窗口里是否有物体即可。但是基于滑窗的检测方法需要对所有的窗口分别计算一次分类结果，是一种暴力搜索算法，虽然训练分类器的过程很容易，但计算量大和内存消耗量大的缺点导致它无法高效地检测三维空间中的物体，此外在深度学习之前，分类器都比较简单，因此分类效果有待提升。

传统 3D 目标检测算法为基于深度学习的 3D 目标检测算法奠定了基础，笔者在后续的讲解中能够看到新兴的算法和前人的工作之间存在许多联系和相似之处。随着深度学习和自动驾驶技术的不断发展，3D 目标检测算法逐渐进入深度学习时代。

## 3　单模态的深度学习检测方法

从 2015 年开始，陆续出现很多尝试将基于深度学习的方法[46-47]用于 3D 目标检测，极大地提升了检测的精度和速度。目前一系列相关算法已经较为成熟，笔者将基于深度学习的 3D 目标检测方法分为两大类：单模态的检测方法和多模态融合的检测方法。多模态融合的深度学习检测方法越来越受到欢迎和重视，将在第 4 节进行详细讲解。首先介绍单模态的深度学习检测方法。

单模态的深度学习检测方法的输入数据主要包括图像或点云，不同的数据类型具有不同的检测框架。笔者按照数据表现形式将单模态的深度学习检测方法分为了 4 类：基于二维数据升维的方法、基于三维数据降维的方法、基于点云的方法以及基于 Transformer 的方法。单模态的深度学习检测方法时间轴如图 4 所示。

### 3.1　基于二维数据升维的方法

目前，除 PointNet 系列方法外，大部分的 3D 检测方法继承了 2D 检测方法的结构及设计思路，其改进的主要思想也来源于 2D 检测器。

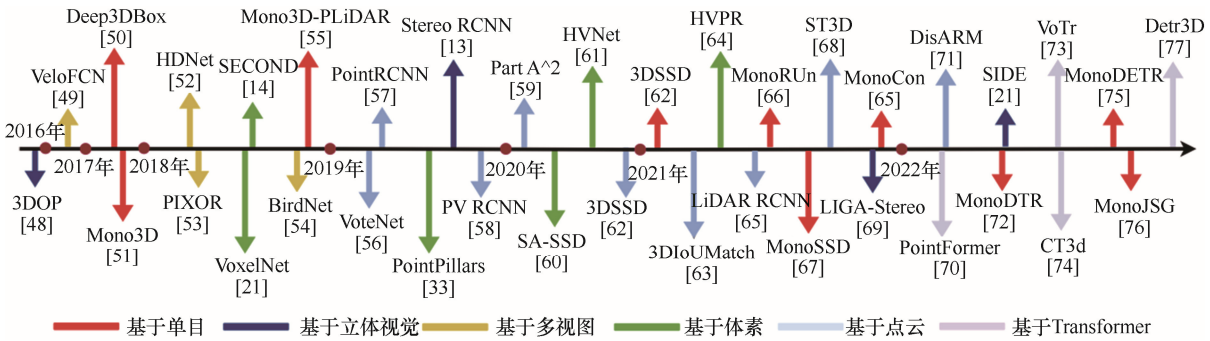



图 4　单模态的深度学习检测方法时间轴

基于深度学习的2D目标检测方法主要分为两大类[78-79]：基于候选区域的方法和基于回归的方法。前者先由算法生成一系列的目标候选区（region proposal），然后再通过卷积神经网络进行冗余候选区的排除和目标分类。这类经典的模型主要包括 Fast R-CNN[80]、Faster R-CNN[81]等。后者则直接将目标边界定位问题转换成端到端的回归问题。在这种方法中，不涉及生成候选框，图像会被缩放到同一尺寸，并以网格形式均等划分，如果目标中心落在某个网格中，该网格就负责预测目标。这类经典的模型主要包括 SSD[82]、YOLO[83]等。两种方法的区别导致了性能的不同，前者具有更高的检测准确率和定位准确率，而后者具有更快的运算速度。目前，2D 目标检测算法逐渐成熟，但是在自动驾驶领域，2D 目标检测无法提供物体的三维空间位置，因此一些 3D 目标检测方法尝试继承 2D 检测的结构和思路，形成了一种基于二维数据升维的 3D 检测方法。

基于二维数据升维的方法主要从图像出发，输入数据表现形式可以为二维彩色图，也可以为彩色图+深度图。单目图像主要利用给定的相机内参结合投射几何方程估算出物体粗略的三维尺寸，立体图像主要通过相机之间的相对位置，计算得到比单目视觉更强的空间约束关系。总而言之，这种方法利用高效的 2D 主干网络解决 3D 检测问题，统称为二维数据升维的方法。

（1）基于单目图像的方法

由于单目图像的方法缺少深度信息，从 2D 图像中直接获取对应的 3D 框的准确位置是件非常困难的事，因此常常把距离预测作为一个单值回归问题。

目前一部分单目检测算法利用几何关系估计物体的深度。He K M 等人[84]通过引入 10 个关键点分别计算 3 组 2D 投影中物体的高度，再结合几何关系得到物体的实际高度，最终结合 3 组估计的深度获得更好的深度预测结果。

另一部分方法利用透视投影原理，建立 3D 框在世界坐标系下和像素坐标系下的对应关系，然后通过最小化重投影误差计算目标距离。Chen Y J 等人[85]利用几何约束关系，结合相机内参及目标的物理尺寸和朝向信息，构造方程组，求解出目标的位置信息。除此之外，也可以在 2D 检测器之后加上深度预测分支，直接回归出物体的三维位置和方向[86]。

还有一种新型的单目检测方法利用现有比较成熟的深度估计，生成对应的稠密深度图，转换到 3D 空间得到伪激光雷达，然后再用基于点云的 3D 检测方法检测目标。Weng X S 等人[87]首先进行单目深度估计，并将输入图像提升到伪激光雷达点云表示，然后训练一个端到端的 3D 检测网络。为了处理大量噪声伪激光雷达，Weng X S 等人[87]使用包围框一致性约束，使其投影到图像上后与其对应的二维方案有较高的重叠度，使用实例掩码代替包围框作为二维方案的表示，以减少点云中不属于对象的点的数量，但是这种方法受限于单目深度估计的性能。因此 Lian Q 等人[88]提出了一种新颖的物体深度联合语义和几何误差测量方法，该方法利用像素级视觉线索改进边界框提案，然后提取相应的语义特征，构造了一个联合语义和几何的约束，从而提升检测精度。总体上，单目 3D 检测目前是单个物体独立的检测任务，深度估计、属性预测等都是相对独立的。由于缺少深度信息，这种方法存在定位不准确、精度不高等问题。

总体上，单目 3D 检测目前来说是单个物体独立的检测任务，深度估计、属性预测等都是相对独立的。

（2）基于立体视觉的方法

双目立体相机可以提供环境感知任务中最重要的深度信息，因此基于立体视觉的方法可以提供更加鲁棒的解法。但是在调查中发现，只有少量的工作利用立体视觉的方法进行 3D 目标检测。Chen X Z 等人[80]通过将物体大小先验、地平面先验和深度信息编码为能量函数生成三维提案，最后利用 Fast R-CNN[80]回归物体姿态和边框。Li P L 等人[90]充分利用立体图像中的稀疏、密集、语义和几何信息，提出了一种更快的检测方式。通过立体视觉中二维和三维之间的投影关系，将 3D 对象定位表示为学习辅助几何问题，之后利用 Roi Align 得到尺度一致的左右特征图并进行融合，最后采用类似于 Mask R-CNN[91]的结构进行关键点的预测。由于深度估计的误差会对其检测精度产生较大的影响，LIGA-stereo 算法[50]利用点云得到的检测结果引导基于立体视觉的方法，以学习到更多的几何特征。具体地，网络框架分为两个分支，一个是输入立体图像得到检测结果，另一个是输入点云数据得到检测结果，最后利用特征一致性约束，通过 BEV 特征的损失函数约束，强制性最小化立体图像特征图及其对应的点云特征图。为了使深度信息更加集中，Peng X D 等人[55]研究了深度信息中的稀疏性和

局部性，并将全局深度信息连接到结构感知中，使用结构感知注意力准确预测每个对象的中心深度，从而提升目标识别的精确率。

### 3.2 基于三维数据降维的方法

由于点云的稀疏性和无规则性，2D 检测器不能直接运用在点云中，因此找到合适的点云表示形式极其重要。一些方法将 3D 点云转换成 FV、BEV 等 2D 表现形式，再利用 CNN 进行处理。还有一些方法将点云转换为体素、柱体等规则的数据形式，再降维到 2D，最后采用 CNN 进行分类和回归，因此也可划归成 2D 检测问题。将这种 3D 点云转化成二维数据的方法统称为基于三维数据降维的方法。根据数据表现形式的不同，基于三维数据降维的方法可进一步划分为基于多视图的方法和基于体素的方法。

（1）基于多视图的方法

多视图方法的数据表现形式包括前视图、鸟瞰图及距离图像。首先，Li B 等人[76]将点云投影到前视图上，获得二维点图，然后在二维点图上应用一个卷积网络，并从卷积特征图中密集地预测三维边框来估计物体的位置和方向。为了加速网络的运算速度，在 VeloFCN 的基础上，LMNet[48]首次使用带有空洞卷积的全连接层结构进行物体检测。由于 FV 存在空间遮挡问题，因此虽然提升了速度，但损失了检测精度。

为了解决 FV 中的遮挡问题，PIXOR[92]将点云映射到鸟瞰图中，然后利用图像金字塔结构进行物体检测与定位，受益于 BEV 视角下遮挡率小的优势，PIXOR 取得了较好的目标检测性能。为了进一步提升基于多视图的检测性能，HDNET[93]继承了 PIXOR 的设计，通过车道线等区域缩小网络在 BEV 上的搜索范围，实现了三维物体的实时高精度检测。BirdNet[69]使用成熟的 2D 检测器在鸟瞰图上进行 2D 目标检测，提出了点云密度归一化来考虑不同激光雷达传感器之间差异，结果表明，该归一化显著提高了基于 BEV 方法的检测能力。虽然 BEV 方法可以保持空间的深度信息和物体的几何形状，同时能够轻易地解决遮挡问题，但对于行人、路标等身长头小的类别来说，高度采样后只有几个点，不利于后续特征的提取，另外，这种手工特征的方式存在信息缺失。

与前两种数据表现形式不同，LaserNet[94]将点云投影到 RV 上，通过全卷积网络生成一组预测，然后对 RV 中的每个激光雷达点预测一个类概率，并在 RV 中对边界框进行概率分布回归。但由于 RV 图像中存在尺度变化和遮挡问题，这种方法精度并不高。

（2）基于体素的方法

与基于多视图的方法不同的是，基于体素的方法直接利用深度学习对体素做特征提取，其数据表现形式为体素或柱体。VoxelNet[21]是基于体素方法的开山之作，将点云划分为一定数量的体素，作者提出了一种体素特征编码（voxel feature extraction，VFE）的方法，通过将点特征与局部聚集特征结合，实现了体素内部的点间交互，通过叠加多个 VFE 层可以学习复杂特征，来表征局部 3D 形状信息，使网络能够学习点云的形状信息。在 VFE 的基础之上，不少算法研究出更好的特征编码方式。HVNet[49]认为较小的体素尺度可以捕获更精细的几何特征，可以更好地定位对象，但需要较长的推理时间；较大的体素尺度适合于更大的特征图和较快的推理速度，但对于小物体来说性能较差，因此提出了混合体素特征编码结构（hybrid voxel feature extraction, HVFE）。HVNet 在 KITTI 数据集的 Cyclist 类别上获得了较好的成绩，但在 Car 类别上仍然低于纯点云的方法，因此可以得出这种利用多尺度混合体素的方式对提升小物体的精确度可以起到积极的作用。VoxelNet 针对点云规范化做出了巨大的贡献，并采用 3D CNN 的方式提取特征，但是 3D CNN 存在一个严重的问题就是空体素也会参加运算，这会导致显存的浪费。因此为了减少计算时间，SECOND[12]在其基础之上采用稀疏卷积代替传统的 3D CNN，减少内存的占用，提升运算效率。

尽管 SECOND 利用稀疏卷积消除了空体素带来的不必要的计算，但是昂贵的 3D 卷积依然存在，阻碍了速度的进一步提升。2019 年，PointPillars[24]的提出消除了这个瓶颈，设计了新的编码方式：Pillar。PointPillars 由 3 部分组成，首先利用柱体的方式将点云转化为稀疏伪图像，然后使用 2D 网络进行特征学习，最后使用 SSD[95]头进行包围框的回归，在速度上是 SECOND 的 3 倍。

体素和柱体这两种编码点云的方式为后续的工作提供了良好的基础，大量的工作开始围绕基于体素的方法不断地改进[53]，但是在编码点云的过程中，存在着一些问题：由于编码伴随着降采样，一些对检测有用的信息可能会被丢弃，另外无效的填充浪费了大量的计算。总之，基于体素的方法已被

广泛应用于三维目标检测，但其性能仍然受到体素化误差的限制。

### 3.3 基于点云的方法

以上两类方法都是通过某种方式处理输入的点云数据，得到 BEV、体素、柱体等表示后去做 3D 目标检测，而基于点云的方法则是直接处理点云，利用点云本身的稀疏性，捕获点云的特征而不做格外的处理，这类方法会尽可能多地保留原始点云的几何形状，其数据表现形式为 3D 点云。

PointNet 是第一个可以端到端地处理点云数据的神经网络架构，整体思想简单直观：输入为所有无序的点云坐标，经过一系列的多层感知机对所有点进行特征编码，再用一个 T-Net 对特征进行对齐，每个点用一个特征向量来表示，然后综合所有点生成一个全局特征向量，再将全局特征向量与每个点的特征向量进行融合，PointNet 虽然利用了每个点的特征及全局特征，但并没有利用局部特征。因此改进后的 PointNet++提出了分层特征提取（set abstraction，SA）模块，并使用多尺度分组（multi-scale grouping, MSG）和多分辨率分组（multi-resolution grouping, MRG）提取点的局部特征，充分利用了数据中的局部信息。

为了减小 3D 检测框的搜索范围，PointRCNN[52]利用语义分割技术得到有效的前景点回归检测框，通过点云池化等操作得到每个搜索框的特征，并结合语义分割得到的预测结果对搜索框进行修正和打分，但这种方式会造成大量的检测框冗余，计算量大，效率不高。之后作者又提出了 Part-A^2 Net[54]，充分利用原始数据集中的标签内部信息，以减少检测框的冗余，设计出 part aware 和 part-aggregation 网络。为了同时利用体素和点云的优点，PV- RCNN[96]提出了 voxel set abstraction（VSA）操作，将稀疏卷积中不同尺度的体素及其特征投影回原始的 3D 空间，以关键点为球中心，在每个尺度上聚合周围体素的特征。这个过程结合了基于点云的和基于体素的两种点云特征提取的方式，同时将整个场景的多尺度信息融合到少量的关键点中。为了进一步提供准确的定位信息，3DIoUMatch[97]认为如何合理标注数据集是当前 3D 应用的瓶颈问题，作者提出了一种两阶段的过滤策略为 3D 边框提供更精确的定位置信度，并利用改进后的非极大值抑制（non-maximum suppression，NMS）方法去除重复框。为了更好地利用 3D 上下文信息，DisARM[61]提出了一种双向网络框架来提取上下文的信息，计算不同候选框之间的权重，包括空间距离权重和特征距离权重，然后融合点云之间的特征，进而提高目标检测的性能。总之，由于 LiDAR 可以提供可靠的深度信息，因此基于点云的 3D 检测算法明显优于基于二维图像的算法。

### 3.4 基于 Transformer 的方法

Transformer[98]是一个完全依赖自注意力机制来计算的转换模型，具有捕获远程依赖关系的能力，Transformer 在自然语言处理和图像领域得到广泛应用后，基于 Transformer 的架构也被用于 3D 点云目标检测。Mao J G 等人[60]首次提出了一种将 Transformer 应用于稀疏体素的新思想。针对体素的稀疏特性，提出了一种特殊的注意机制和快速体素查询方案，极大提高了计算效率，并展示出优越的检测性能。PointFormer[57]则专门为点云设计出一个基于 Transformer 的骨干网络，引入注意力算子提取点云特征。为了进一步捕获多尺度表示之间的依赖关系，PointFormer 将局部特征与高分辨率的全局特征集成在一起，同时引入坐标优化模块，改善了候选框的生成。为了更好地提取 3D 候选框中的特征，Wang Y 等人[59]提出了一种基于通道层面的自注意力机制，为了同时利用局部信息和全局信息，作者在通道层面上为每个点赋予了不同的权重，一定程度上提高了模型的精度。

另一种方案是将 DETR[98]拓展到 3D 目标检测中，DETR3D[58]提出了一种多视角的检测框架，可以直接推理出 3D 检测框，利用提取出的 2D 特征和推理出的 3D 点，设计出一种独特的 object queries，极大地提高了推理速度。为了更好地在单目检测中利用深度图像信息，MonoDETR[63]首次提出了一个端到端的检测器，能够自适应地探索深度引导下的信息图像特征。与 MonoDETR 设计类似，MonoDTR[71]引入深度觉察特征增强模块，避免从预训练的深度估计中获取不准确的深度先验信息，显著减少了计算时间。

总之，Transformer 用于 3D 目标检测领域逐渐被重视，有些基于 Transformer 的方法已经超过了基于卷积神经网络的方法，在这两年迅速发展，使其在图像分类、检测任务中表现出优越的性能。对典型的单模态 3D 检测方法的平均精度（average precision，AP）进行综合比较，结果见表 1。所有方法都遵循了官方 KITTI 数据集的评估方案，旋转的 3D IoU 分别为 0.7、0.5 和 0.5，分别针对汽车类、

**表 1　　典型的单模态 3D 检测方法在 KITTI 数据集上平均精度的比较**

| 方法 | | 输入 | 速度/fps | 车辆 | | | 行人 | | | 自行车手 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | E | M | H | E | M | H | E | M | H |
| 基于立体视觉 | Deep3DBox[85] | mono. image | — | — | — | — | — | — | — | — | — | — |
| | MonoCon[99] | | — | 22.50% | 16.46% | 13.95% | 13.10% | 8.41% | 6.94% | 2.80% | 1.92% | 1.55% |
| | Mono3D[100] | | — | 2.53% | 2.31% | 2.31% | — | — | — | — | — | — |
| | Mono3D-PLiDAR[87] | | — | 1.76% | 7.5% | 6.1% | — | — | — | — | — | — |
| | M3DSSD[101] | | — | 17.51% | 11.46% | 8.98% | 5.16% | 3.87% | 3.08% | 2.10% | 1.51% | 1.58% |
| | MonoRUn[102] | | | 19.65% | 12.30% | 10.58% | 10.88% | 6.78% | 5.83% | 1.01% | 0.61% | 0.48% |
| | MonoDETR[63] | | — | 25.00% | 16.47% | 13.58% | 33.60% | 22.11% | 18.60% | 28.84% | 20.61% | 16.38% |
| | MonoDTR[71] | | | 21.99% | 15.39% | 12.73% | 15.33% | 10.18% | 8.61% | 5.05% | 3.27% | 3.19% |
| | MonoJSG[88] | | | 24.69% | 16.14% | 13.64% | 11.02% | 7.49% | 6.41% | 5.45% | 3.21% | 2.57% |
| 基于单目 | 3DOP[89] | stereo image | — | — | — | — | — | — | — | — | — | — |
| | LIGA-Stereo[50] | | — | 84.92% | 67.06% | 63.80% | 81.39% | 64.66% | 57.22% | 40.46% | 30.00% | 27.07% |
| | CG-Stereo[103] | | — | 76.17% | 57.82% | 54.63% | 74.39% | 53.58% | 46.50% | 47.40% | 30.89% | 27.73% |
| | Stereo R-CNN[90] | | — | 47.58% | 30.23% | 23.72% | — | — | — | — | — | — |
| 基于多视图 | VeloFCN[76] | FV | 1.0 | — | — | — | | — | — | | — | — |
| | BirdNet[69] | BEV | 9.1 | 13.53% | 9.47% | 8.49% | 12.25% | 8.99% | 8.06% | 16.63% | 10.46% | 9.53% |
| | PIXOR[92] | BEV | 28.6 | — | — | — | — | — | — | — | — | — |
| | HDNet[93] | BEV | 20.0 | 82.11% | 71.70% | 67.08% | 50.32% | 40.97% | 37.87% | 77.63% | 63.78% | 55.89% |
| | LaserNet[94] | RV | 26.3 | | — | — | — | — | — | — | — | — |
| 基于体素 | VoxelNet[21] | voxel | 2.0 | 77.82% | 64.17% | 57.51% | — | — | — | — | — | — |
| | HVNet[49] | voxel | 31 | — | — | — | — | — | — | — | — | — |
| | SECOND[12] | voxel | 26.3 | 84.65% | 75.96% | 68.71% | — | — | — | — | — | — |
| | PointPillars[24] | pillars | 62.0 | 82.58% | 74.31% | 68.99% | 51.45% | 41.92% | 38.89% | 77.10% | 58.65% | 51.92% |
| | HVPR[104] | voxel | 36.1 | 86.38% | 77.92% | 73.04% | 53.47% | 43.96% | 40.64% | | | |
| | SA-SSD[53] | voxel | 25.0 | 88.75% | 79.79% | 74.16% | — | — | — | — | — | |
| | Part-A^2[54] | point | 12.5 | 87.81% | 78.49% | 73.51% | 53.10% | 43.35% | 40.06% | 79.17% | 63.52% | 56.93% |
| 基于点云 | PointRCNN[52] | point | 10 | 86.96% | 75.64% | 70.70% | 47.98% | 39.37% | 36.01% | 74.96% | 58.82% | 52.53% |
| | VoteNet[105] | point | — | — | — | — | — | — | — | — | — | — |
| | PV RCNN[96] | point & voxel | 12.5 | 90.25% | 81.43% | 76.82% | 52.17% | 43.29% | 40.29% | 78.60% | 63.71% | 57.65% |
| | 3DSSD[106] | point | 25.0 | 88.36% | 79.57% | 74.55% | 54.64% | 44.27% | 40.23% | 82.48% | 64.10% | 56.90% |
| | LiDAR RCNN[65] | point | — | 85.97% | 74.21% | 69.18% | — | — | — | — | — | — |
| | 3DIoUMatch[97] | point | — | 76.0% | 31.7% | 36.4% | 78.7% | 48.2% | 56.2% | 84.8% | 60.2% | 74.9% |
| | ST3D[51] | point | — | 72.94% | — | — | — | — | — | — | — | — |
| 基于 Transformer | DisARM[61] | point | — | — | — | — | — | — | — | — | — | — |
| | VoTr-TSD[60] | voxel | — | 89.04% | 84.04% | 78.68% | — | — | — | — | — | — |
| | PointFormer[57] | point | — | 87.13% | 77.06% | 69.25% | 50.67% | 42.43% | 39.60% | 75.01% | 59.80% | 53.99% |
| | CT3D[58] | point | — | 87.83% | 81.77% | 77.16% | — | — | — | — | — | — |
| | DETR3D[59] | BEV | — | — | — | — | — | — | — | — | — | — |

行人类和自行车手类，表格中“-”表示结果不可用，E、M、H 分别表示简单（easy）、中等（moderate）、困难（hard）。

### 3.5　小结

图 4 展示了一些典型的单模态基于深度学习的 3D 目标检测方法。通过观察时间序列，可以发现 2019 年以前研究主要是基于单目[107]和多视图[108]的方法；在 2019—2021 年，基于体素和点云的方法成为研究的主流，从效果上看这两种方法在 3D 目标检测方面取得了不错的成绩。2021 年之后，基于点云的研究继续受到关注，效果进一步提升，有意思的是基于单目视觉的方法在 2021 年开始高频率出现。

表 1 展示了这些典型算法在 KITTI 数据集上的速度和精度。结合表 1 可以发现基于单目视觉的模型精确度较低、难度较大，速度上并没有给出具体的数值。这一类模型正处在研究阶段，模型鲁棒性和灵活性较低，相比其他方法有足够的提升空间，而且单目视觉方法从应用的角度出发成本最低，因此在 2021 年以后吸引了较多的研究关注。基于立体视觉的模型虽然可以借助几何关系获取更高的表征，但从精确度、速度、鲁棒性和灵活性上仍然逊色于 LiDAR 的方法，研究较少。基于体素和基于点云的模型由于速度快、性能高、鲁棒性强等特点获得广泛的关注。在最新的研究工作中，Transformer 开始逐渐被应用到 3D 目标检测中，并取得了良好的效果。

## 4　多模态融合的深度学习检测算法

与单模态检测算法不同，多模态融合方法是将不同传感器获取的信息组合在一起，获得互补信息，得到更稳健、更精确的预测结果。虽然各种传感器融合网络已经被提出，但它们并不容易优于仅用雷达检测的网络。因为不同传感器采集数据的时间周期相互独立，并且图像稠密且规则，而点云稀疏且无序，尽管二者存在理论上的互补，但是在特征层或输入层上维度不同导致融合操作困难，因此不同数据如何进行有效的融合对网络模型的设计提出了新的挑战。

早期的融合方法主要分为：早期融合、晚期融合、深度融合。早期融合是在特征提取之前进行融合，可以保留和利用原始数据信息，但不同传感器采集的数据存在校准误差，导致模型不稳定；晚期融合是在特征提取之后，对特征层进行分类和回归时进行融合，这种方法避免了数据维度的不一致性，但计算成本高；深度融合是在特征层面上进行融合，可以使网络学习具有不同特征的表示，但在给定的网络架构下，找到最优的特征层并不容易。随着不同融合变体方法的出现，越来越多的算法已不再简单地归属于上述类别。如 MMF[109]既包含深度融合，又包含了晚期融合；PointPainting[73]既不属于早期融合，也不属于晚期融合，而是一种串行融合的方式。因此早期的融合方法没有特别清晰的界限，本身定义较为困难。

本节重点介绍基于多模态融合的深度学习 3D 目标检测方法，分析以下 3 个部分：（1）多模态数据的组合方式，即多个传感器的数据在输入过程中有哪些组合方式；（2）融合粒度，即来自多个传感器的数据以何种粒度进行融合；（3）融合方法，即对近几年的典型多模态算法进行分类并分析。多模态融合的深度学习检测方法时间轴如图 5 所示。

### 4.1　多模态数据的组合方式

在调查中发现许多融合方法使用了两种以上的数据，首先讨论融合算法中存在哪些数据组合方式。

（1）3D 多视图与二维彩色图融合

这种组合方式首先将 3D 点云投射到 BEV、FV 或 RV 上形成二维多视图，然后与二维彩色图共同输入网络进行目标检测，如图 6（a）所示。MV3D[22]是这个领域开创性的工作，它将点云转换成 BEV

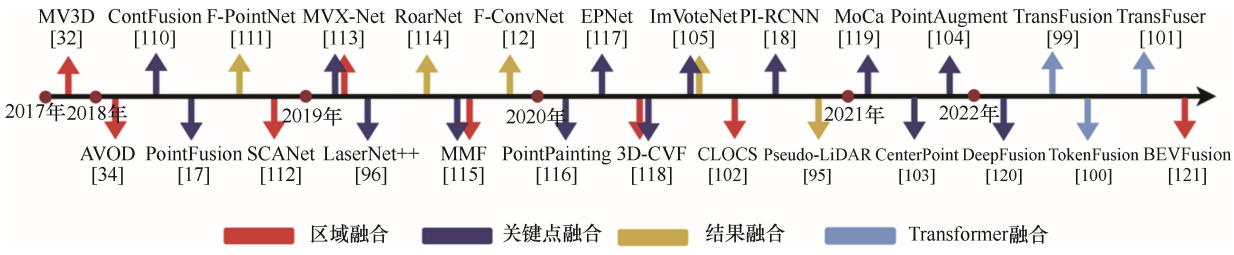
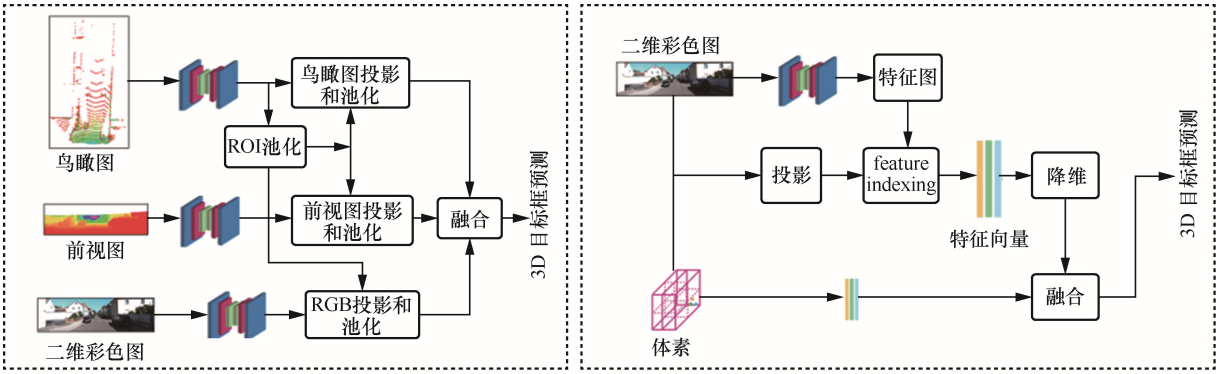



图 5　多模态融合的深度学习检测方法时间轴




(a) 3D多视图与二维彩色图　　(b) 体素与二维彩色图

图 6　两种组合方式的典型网络，从左到右依次是:（a）MV3D，（b）MVX-Net

和 FV 与图像相结合，在 BEV 和 FV 中利用 3D RPN 生成提案，图像则利用 2D CNN 生成提案，经过池化、特征整合，最后直接回归 3D 框的 8 个顶点得到最终的类别标签和 3D 边界框，相比于轴对齐的编码方式，减少了计算量。为了减少信息损失和计算成本，AVOD[23]舍弃了 FV，只利用 BEV 和图像作为输入，利用图像金字塔进行多尺度特征预测，提升了小目标的检测精度。此外，AVOD 重新设计了边界框的几何约束，只编码靠近地平面的 4 个角和 2 个高度值，高度值表示地面上顶部和底部角偏移量，地平面由传感器的高度决定，将 24 维的向量降为 10 维，很好地做到了编码降维的效果，减少了计算量，提升效率。

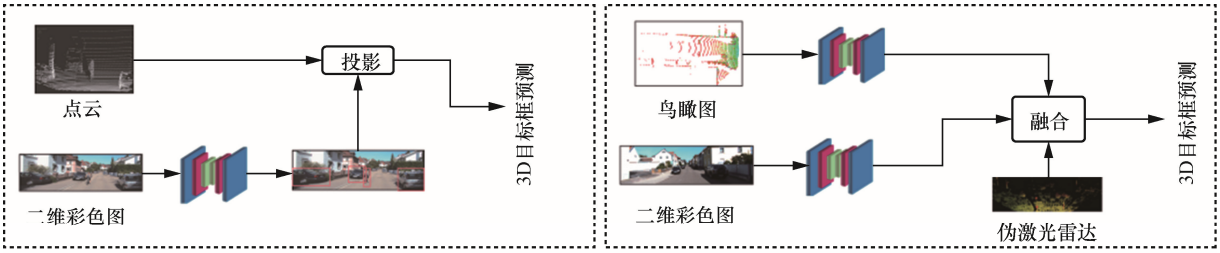
为了生成可靠的 3D 候选框，SCANet[70]同样舍弃了 FV 视图，同时通过引入注意力机制捕获多尺度的上下文信息，结合多尺度低维特征恢复带有丰富空间信息的高维特征。为了实现了高精度的三维空间物体检测定位，ContFuse[77]在多尺度、多传感器下对点云和图像进行深度连续融合，首先分别在 BEV 和点云上提取特征，然后将图像特征进行多尺度融合，并利用类似插值的过程将其投影到 BEV 视图上，由此 BEV 视图不仅具有了空间位置信息，还融合了图像特征信息。

（2）点云与二维彩色图融合

PointNet 和 PointNet++是可以直接处理点云的网络结构，因此可以直接将点云与二维特征图共同输入网络中。F-PointNets[74]是这个方向上开创性的工作，首先利用 2D 检测器生成二维预测框，利用标定矩阵将二维预测框投影到 3D 空间中，形成一个三维的锥体，然后在视锥中进行实例分割，最后利用 PointNet 对分割后的实例进行估计，如图 7（a）所示。除此之外，针对点云旋转不变性的特点，F-PointNets 加入了坐标变换网络（T-Net）使之具有更好的旋转不变性。T-Net 的作用是学习出一个仿射矩阵对输入的点云或特征进行平移、旋转等规范化操作，广泛应用在基于 LiDAR 的目标检测算法里。在后续的工作中，存在很多 F-PointNets 的变种，PointSIFT[75]将 SIFT 模块融合到网络中，利用捕获的方向信息和尺度不变性的特征提高 3D 点的分割性能。SIFRNet[72]将以上两种方法做了信息融合，一方面实现了尺度不变性，另一方面也在通道特征上增加了注意力机制。与原始的 F-PointNets 相比，这些方法在室内外数据集上均取得了显著的提升，但它们的精度通常受到二维检测器性能的限制。

（3）体素与二维彩色图融合

这种组合方式首先将 3D 点云转换成体素等表示形式，之后与二维彩色图共同输入网络中。MVX-Net[115]将每个非空体素投影到图像平面上，产生一个目标区域（region of interest，ROI），利用特征映射，将 ROI 中的特征附加到每个体素叠加的 VFE 层生成的特征向量中，如图 6（b）所示，MVX-Net 有效地融合了多模态信息，降低了假阳性率和阴性率。在 MVX-Net 的基础上，MoCa[116]重点分析了数据增强在多模态中的作用，并提出了多模态剪切的操作，进一步提升了检测性能。为了减少信息的丢失，3D-CVF[112]利用注意力机制权衡不同模态特征的重要性，生成联合特征图，利用联合特征图生成区域建议。由于空间信息较少，作者首先提取多尺度点云特征和图像特征，再由 PointNet 编码，最后利用融合的特征产生了最终的检测结果，从而进一步减少信息丢失。但是，三维体素的量化问题仍然不容忽视。




(a) 点云与二维彩色图融合　　(b) 3D多视图与彩色图和深度图融合

图 7　两种组合方式的典型网络，从左到右依次是：（a）F-PointNet，（b）MMF

（4）3D 多视图与彩色图和深度图融合

点云转换成 BEV 或 FV 视图后，存在深度信息丢失的问题，因此不少方法加入深度图像来补全缺失的深度信息，如图 7（b）所示。MMF[109]提出了第一个用于端到端的多传感器多任务的框架，它包含 2D 物体检测、地面估计以及深度补全。MMF 首先将点云投影成稀疏深度图与 RGB 连接成 RGB-D 送入 CNN，接着利用路面估计得到 BEV 视图，然后将这两部分进行融合，最后经过全连接、非极大值抑制等产生 2D、3D 边界框。这种多任务的方法使得网络可以更好地学习特征表示，更好地生成更精确的物体检测结果。

### 4.2　融合粒度

在第 4.1 节讨论了多个传感器的数据是如何组合的，第 4.2 节将介绍数据以怎样的粒度进行融合。数据经过处理后得到的特征图有粗细粒度之分，细粒度特征可以关注更多的细节，如纹理、颜色；粗粒度特征能够学习到物体的全局特征，如位置、方向。根据融合粒度的粗细，将融合粒度的方法分为以下 5 种。

（1）区域粒度（ROI wise）

ROI wise[22]是通过 ROI 池化等操作，将不同类别数据输出的特征图统一池化，在 ROI 级别上进行融合的操作。这种操作主要出现在 2018 年以前[23]，是多模态方法刚刚出现的时期。这种方法可以更好地利用成熟的特征提取网络，融合方式简单，可以端到端优化，但在融合过程中会丢失空间几何信息和深度信息，并且融合的 ROI[122]存在背景噪声，过于粗糙，因此它是一种粗粒度的融合方式，不适用于精细的目标检测任务。

（2）点粒度（point wise）

点粒度是将点云中的每一个点通过标定矩阵找到图像上的每一个像素点，寻找点云与图像像素之间的一一对应关系，然后融合，是一种更细粒度的融合方式[110]。与上述两种融合粒度级别相比，点粒度不会出现“特征模糊”的问题，适合用于精细的目标检测任务。虽然这种方式可以解决密集图像和稀疏点云之间分辨率不匹配的问题，但是与体素粒度相比，点粒度的内存消耗量更高。

（3）体素粒度（voxel wise）

体素粒度[109]是将点云划分为体素的形式，利用点云提取网络得到稀疏的体素特征向量，然后将这些体素向量进行融合的操作。与区域粒度相比，体素粒度是一种稍微细粒度的融合方式，是区域粒度的缩小版。当体素的初度足够大时，可以近似看作 ROI，变成区域融合；当体素尺度足够小时，可以近似看成点，变成点粒度融合[49]。在体素化点云投影到 BEV 视图上时，需要对 BEV 和 RGB 视图进行对齐操作，因此点云同一个点可能会与多个像素相关联，造成“特征模糊”[109]问题，需要通过插值辅助对齐。

（4）像素粒度（pixel wise）

像素粒度[111]主要作用于距离图像和 RGB 图像的融合，在二维平面上建立像素之间的特征对齐[123]，然后通过 2D CNN 提取特征。距离图像无损失地保留了激光雷达的原始信息，距离图像虽然比点云更密集，但分辨率较低，并且很少用于目标检测，因此这种像素融合的方法并不多见。

（5）Transformer（Transformer wise）

这种方法是将 Transformer 强大的表达能力运用到点云与图像融合中，通过 Transformer 的自注意力机制将关于 3D 场景的全局上下文推理直接集成到不同模态的特征提取层中[60]。在融合过程中，通过交叉注意力模块加权其他目标特征来增强当前目标特征，整合来自不同类型的输入数据。与前几种方法相比，这种方式具备更强大的学习能力，可以更方便、更容易地融合多模态数据，不需要标定矩阵等对齐操作，但存在严重的算力资源依赖和数据依赖。

### 4.3　融合方法

前面介绍了不同数据之间的组合方式、融合粒度，在这一部分将讨论具体的融合方法。笔者将融合方法分为以下 4 种方式：区域融合（region fusion）、关键点融合（point fusion）、结果融合（result fusion）、Transformer 融合（Transformer fusion）。

（1）区域融合

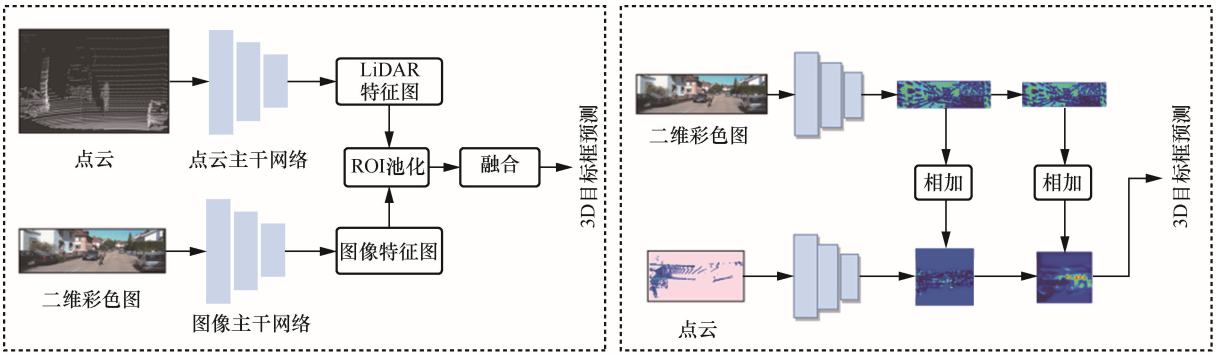
区域融合[70]是将许多粗粒度、大尺度特征图在区域粒度上进行融合，常见于两阶段检测框架。首先在第一阶段，不同数据通过不同的特征提取网络获得大小不等的特征图；在第二阶段，进行区域粒度融合并进行最终的分类和回归，如图 8（a）所示。除此之外，我们将多尺度特征融合中的一些大尺度体素也视为融合区域，当融合的粒度在点级以上时，统称为区域融合。在 2018 年前后，这种粗粒度的融合方法开始出现，经常用于 BEV 视图和图像视图的融合中。MMF[109]将 RGB 和 BEV 视图上的 ROI 特征拼接得到多尺度特征图，通过两层全连接网络，进一步优化目标框的回归，得到高质量的 2D 和 3D 检测结果。还有一些方法将点云划分成不同大小的体素，然后进行离散卷积操作，但在大规模场景下，划分的尺度会影响最终的检测结果，较小的体素能够捕获更精细的几何特征，但是需要较长的推理时间，较大的体素能够获得更快的推理速度，但是检测性能较低。为了平衡两者之间的关系，HVNet[49]提取了不同尺度体素下的点云特征用于后续的检测，认为一些大尺度的体素包含更多的点，可以视为大尺度特征图。区域融合可以端到端优化，但融合粒度粗糙，不适合执行精细的目标检测任务，其次，存在特征模糊的问题。

（2）关键点融合

关键点融合[110]通过标定矩阵找到点云与图像上各个点之间的对应关系，然后在点粒度上进行融合，如图 8（b）所示。这种方法直接使用原始点云提取空间几何特征，因此没有信息的损失。PointFusion[14]提取图像和点云特征后，通过相机外参计算出每一个像素到 3D 坐标的映射关系，对两类特征进行点粒度的融合。与 PointFusion 不同，ContFuse[77]计算不同视角之间的转换关系，将图像特征进行多尺度融合后投影到 BEV 视图上，不仅融合了图像特征及空间几何关系，而且提高了网络的感知能力。近几年来，许多算法尝试使用多种融合方法来提高算法的检测性能，如 MMF[109]在不同阶段使用了不同的融合方式。为了让融合过程更加稳健，融合的特征更具有表现力，PI-RCNN[16]充分利用图像中的语义信息，将分割模型输出的语义特征与 LiDAR 点的特征融合在一起，解决密集图像与稀疏点云之间分辨率不匹配的问题，除此之外，作者提出了一种新颖的融合模块并添加注意力机制使模型更具有表现力。EPNet[124]同样将图像的语义信息融合到点云数据之上，解决了原始相机图像带来的干扰问题。但这种方法只是简单地用语义特征装饰原始的激光雷达点云，没有考虑多模态中普遍存在的特征对齐和数据增广问题，因此，DeepFusion[110]提出了一个深度特征融合管道，解决了几何相关数据增加引起的对齐问题。虽然关键点融合可以充分利用原始数据信息，有效地提升检测性能，但是它在融合前需要计算并固定出标定矩阵，并且相比于区域融合，内存消耗量大。

（3）结果融合

结果融合[74]并没有融合图像和点云对应的特征，只是传递了来自图像的检测结果，其目的是减小三维空间的搜索范围，避免大规模遍历点云，这种不同数据之间传递信息的方式称为结果融合。这




(a) 区域融合　　(b) 关键点融合

图 8　区域融合、关键点融合方法示意

种方法最早应用于 F-PointNet[74]、F-ConvNet[11]，首先通过图像主干网络得到物体的位置信息，利用投影矩阵将 2D 信息映射到三维空间中，极大缩小了 3D 空间的搜索范围，然后利用点云主干网络得到物体的类别和坐标，如图 9（a）所示。然而受光照、遮挡等条件的影响，2D 目标检测会出现漏检现象。为了降低对 2D 检测器的依赖，RoarNet[113]不需要利用 2D 边界框筛选点云，而是利用建议区域的点云来预测是否含有物体，同时预测物体相对建议区域的位置，从而递归地使用相对预测位置作为下一次检测的建议区域。这种方式明显减少了可能的 3D 候选区域，减少了大范围内的目标搜索。结果融合与前两种融合方法相比，具有更简单的网络结构，它们不需要计算点云与图像之间的映射关系，也不需要处理点云对齐等问题。

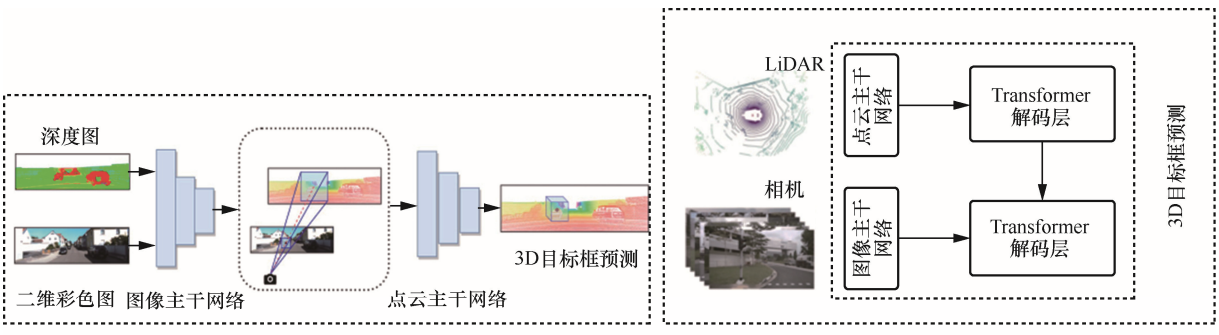
（4）Transformer 融合

上述 3 种融合方法在当前学术界的数据集上取得了很好的效果，但存在两个明显的问题。一方面，在融合过程中，标定矩阵是固定的，因此图像像素与激光点之间的关系也随之固定。另一方面，上述方法依赖高质量的传感器标定，这种依赖关系容易对整体性能产生影响，TransFusion[119]的出现解决了这种硬关联的问题，作者引入了一种有效且鲁棒的多模态检测框架，利用 Transformer 中的注意力机制，重新定位融合过程的焦点，实现从硬关联到软关联，从而提高对退化图像质量和传感器错位的鲁棒性，如图 9（b）所示。与 TransFusion 不同，TokenFusion[118]提出了一种有效的、通用的方法来组合多个单模态 Transformer，修剪多个单模态 Transformer，然后重新利用修剪后的单元进行多模态融合。为了捕获全局上下文信息，整合来自不同模态的数据，TransFuser[120]提出了一种新颖的多模态融合方法，将关于 3D 场景的全局上下文推理直接集成到不同模态的特征提取层中，在特征编码的多个阶段有效地整合来自不同模态的信息，改善了后期融合方法的局限性。总之，Transformer 融合是一个活跃的研究方向，为多模态数据特征学习提供了一种更新颖、有效的解决思路，具备更为强大的特征学习能力，但同时也存在严重的数据依赖和算力资源依赖等问题。

对典型的多模态 3D 检测方法平均精度进行综合比较，结果见表 2。所有方法都遵循了官方 KITTI 的评估方案，旋转的 3D IoU 分别为 0.7、0.5 和 0.5，分别针对汽车类、行人类和自行车手类。其中 Rg-F、Pt-F、Rs-F、Trs-F 分别代表区域融合、关键点融合、结果融合以及 Transformer 融合，(1)、(2)、(3)、(4) 分别代表第 4.1 节所描述的 4 种数据组合方式，“-”表示结果不可用。

### 4.4　小结

目前具体使用哪种融合方法使检测效果更好还没有统一的标准，从图 5 中可以观察到，越来越多的算法选择用鲁棒性强、性能高的 PointFusion 方法进行检测，除此之外，Transformer 技术开始在图像领域崭露头角，并取得了不错的效果，越来越多的人开始将 Transformer 迁移到计算机视觉领域。表 2 可以看到大多数融合方法都是基于 KITTI 测试的，但在 KITTI 排行榜上排名第一的方法主要是基于 LiDAR 的方法。相反在最新数据集，如 nuScenes 和 Waymo 上置顶的方法主要是基于多模态融合的，一方面的原因可能是不同数据集使用的激光雷达传感器具有不同的分辨率，多模态更适合点云稀疏的情况，另一方面的原因可能是大部分基于 LiDAR 的方法使用了更为激进的数据增强策略。总之，基于多模态融合的检测方法越来越成为研究的重点，也期待更多融合技术的出现。




(a) 结果融合　　(b) Transformer融合

图 9　区域融合、关键点融合方法示意图

表 2　典型的多模态 3D 检测方法在 KITTI 数据集上的平均精度比较

| 模型 | 融合方法 | 组合方式 | 速度/fps | 车辆 | | | 行人 | | | 自行车手 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | E | M | H | E | M | H | E | M | H |
| MV3D[22] | Rg-F | (1) | 2.8 | 74.97% | 63.63% | 54.00% | — | — | — | — | — | — |
| AVOD[23] | Rg-F | (1) | 12.5 | 76.39% | 66.47% | 60.23% | 36.10% | 27.86% | 25.76% | 57.19% | 42.08% | 38.29% |
| SCANet[58] | Rg-F | (1) | 11.1 | 79.22% | 67.13% | 60.65% | — | — | — | — | — | — |
| MVX-Ne[115] | Rg-F/Pt-F | (3) | 16.7 | 84.99% | 71.95% | 64.88% | — | — | — | — | — | — |
| MMF[109] | Rg-F/Pt-F | (4) | 12.5 | 88.40% | 77.43% | 70.22% | — | — | — | — | — | — |
| 3D-CVF[112] | Rg-F/Pt-F | (3) | — | 89.20% | 80.05% | 73.11% | — | — | — | — | — | — |
| CLOCS[125] | Rg-F | (2) | — | 86.38% | 78.45% | 72.45% | — | — | — | — | — | — |
| ContFuse[77] | Pt-F | (1) | 16.7 | 83.68% | 68.78% | 61.67% | — | — | — | — | — | — |
| PointFusion[14] | Pt-F | (2) | — | 77.92% | 63.00% | 53.27% | 33.36% | 28.04% | 23.38% | 49.34% | 29.42% | 26.98% |
| PointPainting[73] | Pt-F | (2) | 2.5 | 82.11% | 71.70% | 67.08% | 50.32% | 40.97% | 37.87% | 77.63% | 63.78% | 55.89% |
| EPNet[124] | Pt-F | (2) | — | 89.81% | 79.28% | 74.59% | — | — | — | — | — | — |
| PI-RCNN[16] | Pt-F | (2) | — | 88.27% | 78.53% | 77.75% | — | — | — | — | — | — |
| MoCa[116] | Pt-F | (3) | — | 50.9% | 43.7% | 40.0% | 76.1% | 61.0% | 53.4% | 86.0% | 75.9% | 70.7% |
| CenterPoint[126] | Pt-F | (1) | — | — | — | — | — | — | — | — | — | — |
| DeepFusiont[110] | Pt-F | (2) | — | — | — | — | — | — | — | — | — | — |
| PointAugmen[117] | Pt-F | (2) | — | — | — | — | — | — | — | — | — | — |
| RoarNet[113] | Rs-F | (2) | 10.0 | 83.71% | 73.04% | 59.16% | — | — | — | — | — | — |
| ImVoteNet[114] | Rs-F/Pt-F | (2) | — | — | — | — | — | — | — | — | — | — |
| Pseudo-LiDAR[48] | Rs-F | (4) | — | 54.53% | 34.05% | 38.25% | — | — | — | — | — | — |
| F-ConvNet[11] | Rs-F | (2) | 2.1 | 87.36% | 76.39% | 66.69% | 52.16% | 43.38% | 38.80% | 81.98% | 65.07% | 56.54% |
| F-PointNet[74] | Rs-F | (2) | 5.9 | 82.19% | 69.79% | 60.59% | 50.53% | 42.15% | 38.08% | 72.27% | 56.12% | 49.01% |
| TransFusion[119] | Tr-F | (2) | — | — | — | — | — | — | — | — | — | — |
| TokenFusion[118] | Tr-F | (2) | — | — | — | — | — | — | — | — | — | — |
| TransFuser[120] | Tr-F | (1) | — | — | — | — | — | — | — | — | — | — |

## 5　数据集和评价指标

### 5.1　数据集

如今，谷歌、Uber、特斯拉、Apollo 等互联网公司以及高校和科研院所开始将自动驾驶技术应用到现实中，例如 BMW 已经在慕尼黑附近的高速公路上进行了自动驾驶测试；谷歌已经在美国的 20 多个公路上行驶超过 800 万英里（约 1 287 万千米）来测试他们开发的无人驾驶汽车。大多数的 3D 目标检测方法是基于监督学习的，因此，需要大规模数据集[127]来训练这种深度神经网络。自 2013 年以来发布的一些真实世界的数据集见表 3，其中包括传感器配置、数据集大小、类别标签、采集地点以及公开地址等，其中 GNSS 为全球导航卫星系统（global navigation satellite system），第 2 列括号中的数字为相机的数量。

### 5.2　评价指标

（1）IoU（intersection over union）

IoU[4]是检测物体准确度的一个标准，也被称为 Jaccard 指数，已被广泛用于衡量有限样本集之间的相似性。对于目标检测来说，目标物体由二维图像中最小的矩形表示。基于这种表示，地面真实边界框 $B_g$ 与预测边界框 $B_d$ 之间的 IoU 计算定义如式（7）所示：

$$\mathrm{IoU}(B_g, B_d) = \frac{\text{Aera of overlap } B_g \text{ and } B_d}{\text{Aera of union } B_g \text{ and } B_d} \quad (7)$$

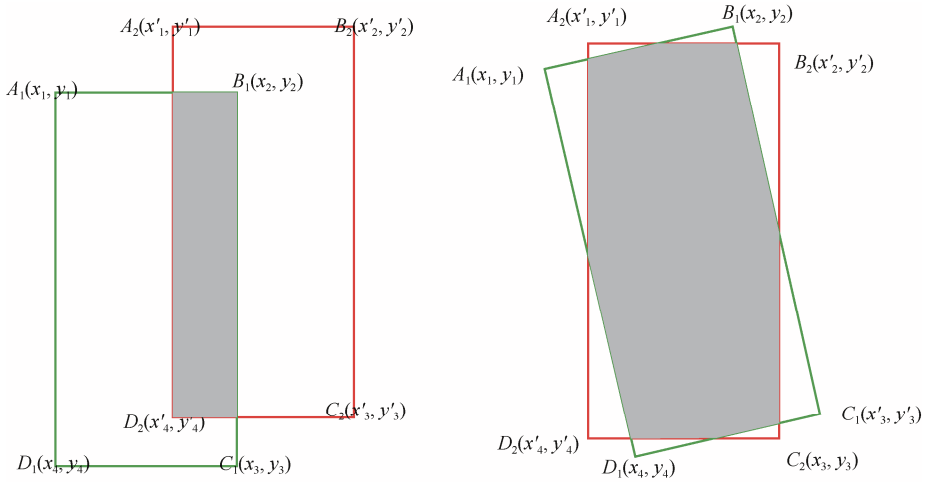
**表 3　　3D 目标检测数据集**

| 数据集 | 传感器 | 时间 | 大小 | 类别 | 采集地点 |
|---|---|---|---|---|---|
| KITTI[4] | 相机、LiDAR<br>GNSS、惯性传感器 | 2012 年 | 7481 帧<br>80 256 个对象 | 8 类 | 卡尔斯鲁厄（德国） |
| SUN3D[67] | 3D 相机 | 2013 年 | 254 个不同场景<br>捕获 415 序列 | 16 类 | 北美、欧洲、亚洲 |
| SUN RGB-D[66] | 3D 相机(4) | 2015 年 | 10 335 张室内场景<br>146 617 个 2D 边框<br>58 657 个 3D 框 | 10 类 | 普林斯顿大学（美国） |
| Multi-SpectralObject[128] | 视觉和热相机 | 2017 年 | 7 512 帧，5 833 个对象 | 3 类 | 日本 |
| MPO[62] | 相机、LiDAR、GNSS | 2017 年 | 1 569 帧 | 6 类 | — |
| ScanNet[64] | 3D 相机<br>深度传感器 | 2018 年 | 1 513 个室内场景 | 21 类 | — |
| S3DIS[56] | 3D 相机 | 2018 年 | 超过 70 000 张<br>RGB 图像 | 13 类 | 斯坦福大学（美国） |
| DBNet[129] | 相机、LiDAR、GNSS | 2018 年 | 超过 10k 帧 | 含 7 个数据集 | 中国 |
| KAIST[68] | 相机、LiDAR<br>GNSS、惯性传感器 | 2018 年 | 7512 帧<br>308 913 个对象 | 3 类 | 首尔（韩国） |
| A*3D[130] | 相机(2)<br>LiDAR | 2019 年 | 39 k 帧<br>230 k 个对象 | 7 类 | 新加坡 |
| Argoverse[131] | LiDAR(2)<br>相机(9) | 2019 年 | 113 个场景<br>300 k 轨迹 | 15 类 | 匹兹堡（美国）<br>宾西法尼亚州（美国）<br>佛罗里达州（美国） |
| PandaSet[132] | LiDAR(2)、相机(6)<br>GNSS、惯性传感器 | 2019 年 | 125 个场景 | 28 类 | 旧金山（美国） |
| ApolloScape[133] | 相机、LiDAR<br>GNSS、惯性传感器 | 2019 年 | 143 906 个图像帧<br>89 430 个物体 | 35 类 | 中国 |
| nuScence[134] | 相机(6)<br>LiDAR<br>Radars(6) | 2019 年 | 1 000 个场景;<br>1.4 万帧(照相机、雷达)<br>390 k 帧(3D 激光雷达) | 23 类 | 波士顿（美国）、新加坡 |
| BLVD[39] | 相机(5)<br>LiDAR(5) | 2019 年 | 120 k 帧<br>249 129 个对象 | 3 类 | 常熟（中国） |
| Waymo[135] | 相机<br>LiDAR | 2019 年 | 200 k 帧<br>12 M 对象(3D 激光雷达)<br>1.2 M 对象(2D 照相机) | 4 类 | — |
| H3D[136] | 相机(3)<br>LiDAR | 2019 年 | 27 721 帧<br>1 071 302 个对象 | 8 类 | 旧金山（美国） |
| A2D2[137] | 相机(6)<br>LiDAR(5) | 2020 年 | 40 k 帧(语义信息)<br>12 k 帧(3D 激光雷达)<br>390 k 帧未标记 | 37 类 | 凯默斯海姆（德国）<br>英戈尔施塔特（德国）<br>慕尼黑（德国） |

在大多数 2D 对象检测基准中，物体用轴对齐的方式进行标记。这种方法计算 IoU 简单，可以使用一些基本的数学函数来实现，如 max、min 等。图 10（a）说明了两个轴对齐的框之间的交集，其中阴影区域表示交集区域。然而，轴对齐不适用于 3D 目标检测，通常，3D 物体是具有 3 个旋转自由度的长方体，在自动驾驶场景下，假设所有物体都位于相对平坦的路面，三个旋转自由度减少为一个偏航角，如图 10（b）所示。这种表现形式广泛流行于 3D 目标检测基准，如 KITTI 和 nuScence。为了评估不同的方法，KITTI 提供了两种不同的 IoU 计算策略：BEV IoU 和 3D IoU。图 10（a）为轴对齐方式，常用于 2DIoU，图 10（b）为旋转计算方式[38]，常用于 3DIoU，交叉区域以灰色突出显示。

- BEV IoU[138]：将真值框和预测框投影到 BEV 上再计算 IoU。两个旋转矩形的 IoU 计算比轴对齐的方式复杂，因为它们有许多不同的方式相交，具体计算公式如式（8）所示：

$$\mathrm{IoU}_{\mathrm{BEV}}=\frac{\mathrm{Area}_{\mathrm{overlap}}}{\mathrm{Area}_{g}+\mathrm{Area}_{d}-\mathrm{Area}_{\mathrm{overlap}}} \tag{8}$$




(a) 轴对齐方式（2D-IoU）　　(b) 旋转计算方式（3D-IoU）

图 10　真值框 $B_g$（绿色）与预测框 $B_d$（红色）之间旋转重叠示意图

其中 $\text{Area}_g=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}\times\sqrt{(x_2-x_3)^2+(y_2-y_3)^2}$，$\text{Area}_d=\sqrt{(x'_2-x'_1)^2+(y'_2-y'_1)^2}\times\sqrt{(x'_2-x'_3)^2+(y'_2-y'_3)^2}$。

• 3D IoU[18]：直接在 3D 空间计算检测结果与真实值的 IoU。正如之前所提到的，自动驾驶中的 3D 物体通常由 7 个参数来描述，包含物体的位置 $(x,y,z)$、长宽高( height, width, length )以及朝向 $\theta$。在这种情况下，两个 3D 框的 IoU 可以由式（9）计算。

$$\text{IoU}_{3D}=\frac{\text{Area}_{\text{overlap}}\times h_{\text{overlap}}}{\text{Area}_g\times h_g+\text{Area}_d\times h_d-\text{Area}_{\text{overlap}}\times h_{\text{overlap}}} \tag{9}$$

其中，$h_{\text{overlap}}$ 表示在高度上的重合，$\text{Area}_{\text{overlap}}$ 表示交叉区域的重合。

（2）平均精度

AP 用来衡量算法在单个类别上的平均精度[139]。AP 值越高，表示对这个类别的检测精度越高[122]，计算式如式（10）所示，其中 $P$ 代表 Precision，$r$ 代表 Recall。由于 IoU 有 3 种计算方式，相应的 AP 也有 3 种计算方式：$\text{AP}_{2D}$、$\text{AP}_{3D}$、$\text{AP}_{BEV}$。在 KITTI 中，AP 是官方定义的评价指标，3D IoU 阈值 0.7、0.5、0.5 分别作为汽车、自行车和行人的类别。

$$\text{AP}=\int_0^1 P(r)\text{d}r \tag{10}$$

（3）平均精度均值（mean average precision，mAP）

通过对特定数据集中所有类的 AP 进行平均，可以很容易地获得 mAP，如式（11）所示：

$$\text{mAP}=\frac{1}{C}\sum_{i=1}^{|C|}\text{AP}_i \tag{11}$$

其中 $C$ 是感兴趣类别的一个子集，$\text{AP}_i$ 表示第 $i$ 类的 AP。

（4）平均方向相似度（average orientation similarity，AOS）

对于物体方向的预测，KITTI 提出了一种新颖的方法用于衡量检测结果与真实值的方向相似程度[4]，具体公式定义如式（12）所示：

$$\text{AOS}=\frac{1}{11}\sum_{r\in\{0,0.1,\cdots,1\}}\max_{\tilde{r}:\tilde{r}\geq r}s(\tilde{r}) \tag{12}$$

其中，$r=\dfrac{\text{TP}}{\text{TP}+\text{FN}}$，代表 PASCAL 物体检测的召回率，当检测到的 2D 预测框与真值框重叠至少 50%时，它们是正确的，方向相似性 $s\in[0,1]$ 被定义为所有预测样本与真实值的余弦距离归一化，如式（13）所示：

$$s(r)=\frac{1}{|D(r)|}\sum_{i\in D(r)}\frac{1+\cos\Delta_\theta^{(i)}}{2}\delta_i \tag{13}$$

其中，$D(r)$表示召回率 $r$ 下所有对象检测结果的集合，$\Delta_\theta^{(i)}$ 是检测物体 $i$ 的估计方向与真实方向的角度之差。为了惩罚多个检出结果匹配到同一个真值，当物体 $i$ 匹配到真实结果（重叠至少 50%时），$\delta_i=1$，否则 $\delta_i=0$。

## 6　未来研究方向

尽管基于点云的 3D 目标检测研究已经引起了

越来越多的关注，而且取得了一些成果，提出了许多相关的算法和理论。但从目前的研究和实用性来看，仍然存在诸多挑战，可能形成未来主要的趋势。

（1）近几年，基于图像和点云的多模态融合方法逐渐成为研究的热点。虽然各种传感器融合网络已经被提出，但笔者发现目前多模态融合的方法依然落后于基于点云的方法。事实上，由于不同传感器采集数据的周期相互独立，它们之间存在同步和校准误差，尽管二者存在理论上的互补，但是随着时间的漂移，传感器之间的标定存在误差，两种数据的同步和校准操作在工程上是一个巨大的挑战，如何将不同的数据更好地结合在一起仍然是一个棘手但值得研究的问题。最近一年，Transformer 技术开始利用其强大的特征学习能力，尝试用自注意力机制打破相机与点云之间的标定依赖，引入有效且鲁棒的多模态融合框架，未来基于 Transformer 的多模态融合技术将会是一个热点研究话题。

（2）从整体上看，由于缺乏深度信息，基于图像的方法与基于激光雷达的方法仍然存在巨大的差距，而单目和双目相机成本低廉，可以获取更充分的颜色和纹理信息，这为后续的研究提供了巨大的研究空间和提升空间，具有重要的意义。除此之外，与多模态的思想相反，为了避免意外造成雷达无法正常工作，降低单个传感器的过度依赖性，提高安全性能，利用单目或双目视觉的方法实现较高精度的 3D 目标检测也是值得关注的研究热点，例如在 2021—2022 年出现了多种基于单目视觉的 3D 目标检测方法。

（3）三维点云的几何形状信息在下游任务中起到很大的作用，但由于点云的无序性，点与点之间没有直接的对应关系，因此很难推断出这些不规则的点形成的潜在形状，也无法使用规则的卷积神经网络处理。除此之外，由于捕获的点云数量巨大，需要进一步转换和降采样，这导致三维点云中原本的几何形状信息丢失，损失大量的有用信息。如何建立搜索机制找到它们之间的潜在邻近信息，已有许多工作开始关注这个问题[140-141]，但由于点云的不规则性，不同场景、不同形状、不同大小的物体多种多样，提取的形状信息有很大的不确定性，亟待大量的实验和理论研究。

（4）数据增强对于提升模型的性能至关重要，与单模态目标检测相比，多模态目标检测使用的数据增强器类型相对较少，限制了精度进一步提升。此外，多模态数据集通常比图像数据集要小得多，如 KITTI 只有 80 256 个对象，而 ImageNet[139]则有上百万的对象，并且这些数据集中物体类别的分布非常不平衡，如车辆比行人、自行车手要多得多，夜晚物体类别远小于白天的物体类别，大物体类别远超过小物体类别，这会导致计算的 AP 值差距较大，尤其在室内的数据集（nuScences、ScanNet 等）中，因此需要解决数据集中类别分布不均衡的问题，已经有部分工作开始关注这个问题[116]。如何在多模态中更好地利用数据增强，并且在数据增强中维持多模态数据的均衡性也是一个需要探索的问题。

（5）大多数基于深度学习的目标检测方法如 PointNet、PointNet++、PointRCNN、VoteNet 等都适用于小点云场景。在实际应用中，由于雷达获得的点云数据通常是大规模的，因此，如何在大规模场景下有效地进行目标检测仍是一个值得关注的研究热点。

（6）目前大多数方法都是基于 CNN 架构的单帧感知，框架中的预测不依赖以前的帧，因为激光雷达每次只能捕捉场景的一个局部视图，导致单帧点云里物体点的分布总是不完整的。随着时间的推移和车辆的移动，传感器会不断生成包含同一物体的点云序列。因此，如何充分利用多帧点云序列得到更为完整的物体结构和位置信息，进而得到更为精准的检测结果，是一个亟待研究的方向[121, 142-143]。但据笔者调查，只有很少的工作包含了时间线索[144-145]，因此，基于时间序列的多模态感知算法仍等待进一步的研究。

## 7　结束语

3D 物体检测是自主驾驶、虚拟现实、机器人操作等领域的一个基本问题，其目的是从无序的三维点云中识别和定位物体。与图像相比，点云提供了丰富而准确的三维结构信息，这对于准确的物体定位至关重要。近年来，随着深度学习的发展，三维目标检测在近几年的人工智能顶刊、顶会上吸引了学者们的研究和关注。本文系统性地梳理了近 10 年来提出的主要方法，将现有的 3D 目标检测技术分为 3 类：（1）传统的机器学习算法，包括基于模板匹配的、基于区域评分的和基于滑窗的；（2）非融合深度学习算法，包括三维数据降维、二维数据升维、基于点云的和基于 Transformer 的算法；（3）基于

多模态融合的深度学习算法。

回顾和分析了各类方法的基本概念、各种目标检测算法、数据集和评价指标，描述了不同模型和方法之间的关系和差异，并对主流的算法进行对比。最后，展望了三维目标检测的未来挑战和可能的发展方向。

## 参考文献：


[1] GUO Y L, WANG H Y, HU Q Y, et al. Deep learning for 3D point clouds: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4338-4364.

[2] 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, 20(10): 3782-3795.

[3] QIAN R, LAI X, LI X. 3D object detection for autonomous driving: a survey[J]. arXiv preprint, 2021, arXiv: 2106.10823.

[4] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2012: 3354-3361.

[5] FRITZ M, SCHIELE B. Decomposition, discovery and detection of visual categories using topic models[C]//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2008: 1-8.

[6] CUI Y D, CHEN R, CHU W B, et al. Deep learning for image and point cloud fusion in autonomous driving: a review[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(2): 722-739.

[7] QIN Z Y, WANG J L, LU Y. MonoGRNet: a geometric reasoning network for monocular 3D object localization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8851-8858.

[8] FU H, GONG M M, WANG C H, et al. Deep ordinal regression network for monocular depth estimation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2002-2011.

[9] ALDOMA A, MARTON Z C, TOMBARI F, et al. Tutorial: point cloud library: three-dimensional object recognition and 6 DOF pose estimation[J]. IEEE Robotics & Automation Magazine, 2012, 19(3): 80-91.

[10] ARORA H, LOEFF N, FORSYTH D A, et al. Unsupervised segmentation of objects using efficient learning[C]//Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2007: 1-7.

[11] WANG Z X, JIA K. Frustum ConvNet: sliding Frustums to aggregate local point-wise features for amodal 3D object detection[C]//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2020: 1742-1749.

[12] YAN Y, MAO Y X, LI B. SECOND: sparsely embedded convolutional detection[J]. Sensors (Basel, Switzerland), 2018, 18(10): 3337.

[13] ZHOU Y, SUN P, ZHANG Y, et al. End-to-end multi-view fusion for 3D object detection in LiDAR point clouds[C]//Proceedings of Conference on Robot Learning. [S.l.:s.n.], 2020: 923-932.

[14] XU D F, ANGUELOV D, JAIN A. PointFusion: deep sensor fusion for 3D bounding box estimation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 244-253.

[15] XIE S N, LIU S N, CHEN Z Y, et al. Attentional ShapeContextNet for point cloud recognition[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4606-4615.

[16] XIE L, XIANG C, YU Z X, et al. PI-RCNN: an efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12460-12467.

[17] FENG D, HAASE-SCHÜTZ C, ROSENBAUM L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(3): 1341-1360.

[18] WANG Y, MAO Q, ZHU H, et al. Multi-modal 3D object detection in autonomous driving: a survey[J]. arXiv preprint, 2021, arXiv: 2106.12735.

[19] CHARLES R Q, HAO S, MO K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 77-85.

[20] QI C R, YI L, SU H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[J]. arXiv preprint, 2017, arXiv: 1706.02413.

[21] ZHOU Y, TUZEL O. VoxelNet: end-to-end learning for point cloud based 3D object detection[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4490-4499.

[22] CHEN X Z, MA H M, WAN J, et al. Multi-view 3D object detection network for autonomous driving[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6526-6534.

[23] KU J, MOZIFIAN M, LEE J, et al. Joint 3D proposal generation and object detection from view aggregation[C]//Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. New York: ACM Press, 2018: 1-8.

[24] LANG A H, VORA S, CAESAR H, et al. PointPillars: fast encoders for object detection from point clouds[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 12689-12697.

[25] FAN L, XIONG X, WANG F, et al. RangeDet: in defense of range view for LiDAR-based 3D object detection[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 2898-2907.

[26] SUN P, WANG W Y, CHAI Y N, et al. RSN: range sparse net for

efficient, accurate LiDAR 3D object detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 5721-5730.

[27] WU B C, WAN A, YUE X Y, et al. SqueezeSeg: convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud[C]//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2018: 1887-1893.

[28] REN M Y, POKROVSKY A, YANG B, et al. SBNet: sparse blocks network for fast inference[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8711-8720.

[29] LI P X, ZHAO H C, LIU P F, et al. RTM3D: real-time monocular 3D detection from object keypoints for autonomous driving[C]//Proceedings of 2020 16th European Conference on Computer Vision. Cham: Springer, 2020: 644-660.

[30] WANG Y, CHAO W L, GARG D, et al. Pseudo-LiDAR from visual depth estimation: bridging the gap in 3D object detection for autonomous driving[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 8437-8445.

[31] NG P C, HENIKOFF S. SIFT: predicting amino acid changes that affect protein function[J]. Nucleic Acids Research, 2003, 31(13): 3812-3814.

[32] JOHNSON A E, HEBERT M. Surface matching for object recognition in complex three-dimensional scenes[J]. Image and Vision Computing, 1998, 16(9/10): 635-651.

[33] CHEN H, BHANU B. 3D free-form object recognition in range images using local surface patches[J]. Pattern Recognition Letters, 2007, 28(10): 1252-1262.

[34] MIAN A, BENNAMOUN M, OWENS R. On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes[J]. International Journal of Computer Vision, 2010, 89(2): 348-361.

[35] STEIN F, MEDIONI G. Structural indexing: efficient 3D object recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(2): 125-145.

[36] CHUA C S, JARVIS R. Point signatures: a new representation for 3D object recognition[J]. International Journal of Computer Vision, 1997, 25(1): 63-85.

[37] FROME A, HUBER D, KOLLURI R, et al. Recognizing objects in range data using regional point descriptors[C]//Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2004: 224-237.

[38] ZHOU D F, FANG J, SONG X B, et al. IoU loss for 2D/3D object detection[C]//Proceedings of 2019 International Conference on 3D Vision. Piscataway: IEEE Press, 2019: 85-94.

[39] COLLET A, SRINIVASAY S S, HEBERT M. Structure discovery in multi-modal data: a region-based approach[C]//Proceedings of 2011 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2011: 5695-5702.

[40] SHIN J, TRIEBEL R, SIEGWART R. Unsupervised discovery of repetitive objects[C]//Proceedings of 2010 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2010: 5041-5046.

[41] HERBST E, HENRY P, REN X F, et al. Toward object discovery and modeling via 3D scene comparison[C]//Proceedings of 2011 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2011: 2623-2629.

[42] KARPATHY A, MILLER S, LI F F. Object discovery in 3D scenes via shape analysis[C]//Proceedings of 2013 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2013: 2088-2095.

[43] FELZENSZWALB P F, HUTTENLOCHER D P. Efficient graph-based image segmentation[J]. International Journal of Computer Vision, 2004, 59(2): 167-181.

[44] SONG S R, XIAO J X. Sliding shapes for 3D object detection in depth images[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2014: 634-651.

[45] MALISIEWICZ T, GUPTA A, EFROS A A. Ensemble of exemplar-SVMs for object detection and beyond[C]//Proceedings of 2011 International Conference on Computer Vision. Piscataway: IEEE Press, 2012: 89-96.

[46] SONG S R, XIAO J X. Deep sliding shapes for amodal 3D object detection in RGB-D images[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 808-816.

[47] 罗会兰, 陈鸿坤. 基于深度学习的目标检测研究综述[J]. 电子学报, 2020, 48(6): 1230-1239.
LUO H L, CHEN H K. Survey of object detection based on deep learning[J]. Acta Electronica Sinica, 2020, 48(6): 1230-1239.

[48] MINEMURA K, LIAU H, MONRROY A, et al. LMNet: real-time multiclass object detection on CPU using 3D LiDAR[C]//Proceedings of 2018 3rd Asia-Pacific Conference on Intelligent Robot Systems. Piscataway: IEEE Press, 2018: 28-34.

[49] YE M S, XU S J, CAO T Y. HVNet: hybrid voxel network for LiDAR based 3D object detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 1628-1637.

[50] GUO X Y, SHI S S, WANG X G, et al. LIGA-stereo: learning LiDAR geometry aware representations for stereo-based 3D detector[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 3133-3143.

[51] YANG J H, SHI S S, WANG Z, et al. ST3D: self-training for unsupervised domain adaptation on 3D object detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 10363-10373.

[52] SHI S S, WANG X G, LI H S. PointRCNN: 3D object proposal generation and detection from point cloud[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 770-779.

[53] HE C H, ZENG H, HUANG J Q, et al. Structure aware single-stage

3D object detection from point cloud[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11870-11879.

[54] SHI S S, WANG Z, WANG X G, et al. Part-A^2 Net: 3D part-aware and aggregation neural network for object detection from point cloud[J]. arXiv preprint, 2019, arXiv:1907.03670.

[55] PENG X D, ZHU X G, WANG T, et al. SIDE: center-based stereo 3D detector with structure-aware instance depth estimation[C]//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2022: 225-234.

[56] ARMENI I, SAX S, ZAMIR A R , et al. Joint 2D-3D-semantic data for indoor scene understanding[J]. arXiv preprint, 2017, arXiv: 1702.01105.

[57] PAN X R, XIA Z F, SONG S J, et al. 3D object detection with pointformer[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 7459-7468.

[58] SHENGA H L, CAI S J, LIU Y, et al. Improving 3D object detection with channel-wise transformer[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 2723-2732.

[59] WANG Y, GUIZILINI V, ZHANG T, et al. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries[C]//Proceedings of Conference on Robot Learning. [S.l.:s.n.], 2022: 180-191.

[60] MAO J G, XUE Y J, NIU M Z, et al. Voxel transformer for 3D object detection[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 3164-3173.

[61] DUAN Y, ZHU C, LAN Y, et al. DisARM: displacement aware relation module for 3D detection[J]. arXiv preprint, 2022, arXiv: 2203.01152.

[62] JUNG H, OTO Y, MOZOS O M, et al. Multi-modal panoramic 3D outdoor datasets for place categorization[C]//Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems. New York: ACM Press, 2016: 4545-4550.

[63] ZHANG R, QIU H, WANG T, et al. MonoDETR: depth-guided transformer for monocular 3D object detection[J]. arXiv preprint, 2022, arXiv: 2203.13310.

[64] DAI A, CHANG A X, SAVVA M, et al. ScanNet: richly-annotated 3D reconstructions of indoor scenes[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2432-2443.

[65] LI Z C, WANG F, WANG N Y. LiDAR R-CNN: an efficient and universal 3D object detector[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 7542-7551.

[66] SONG S R, LICHTENBERG S P, XIAO J X. SUN RGB-D: a RGB-D scene understanding benchmark suite[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 567-576.

[67] XIAO J X, OWENS A, TORRALBA A. SUN3D: a database of big spaces reconstructed using SfM and object labels[C]//Proceedings of 2013 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2014: 1625-1632.

[68] CHOI Y, KIM N, HWANG S, et al. KAIST multi-spectral day/night data set for autonomous and assisted driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(3): 934-948.

[69] BELTRÁN J, GUINDEL C, MORENO F M, et al. BirdNet: a 3D object detection framework from LiDAR information[C]//Proceedings of 2018 21st International Conference on Intelligent Transportation Systems. New York: ACM Press, 2018: 3517-3523.

[70] LU H H, CHEN X S, ZHANG G Y, et al. Scanet: spatial-channel attention network for 3D object detection[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 1992-1996.

[71] HUANG K C, WU T H, SU H T, et al. MonoDTR: monocular 3D object detection with depth-aware transformer[J]. arXiv preprint, 2022, arXiv: 2203.10981.

[72] ZHAO X, LIU Z, HU R L, et al. 3D object detection using scale invariant and feature reweighting networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 9267-9274.

[73] VORA S, LANG A H, HELOU B, et al. Point Painting: sequential fusion for 3D object detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4603-4611.

[74] QI C R, LIU W, WU C X, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 918-927.

[75] JIANG M, WU Y, ZHAO T, et al. PointSIFT: a SIFT-like network module for 3D point cloud semantic segmentation[J]. arXiv preprint, 2018, arXiv: 1807.00652.

[76] LI B, ZHANG T, XIA T. Vehicle detection from 3D lidar using fully convolutional network[J]. arXiv preprint, 2016, arXiv: 1608.07916.

[77] LIANG M, YANG B, WANG S L, et al. Deep continuous fusion for multi-sensor 3D object detection[C]//Proceedings of 2018 15th European Conference on Computer Vision. New York: ACM Press, 2018: 663-678.

[78] 尹宏鹏, 陈波, 柴毅, 等. 基于视觉的目标检测与跟踪综述[J]. 自动化学报, 2016, 42(10): 1466-1489.
YIN H P, CHEN B, CHAI Y, et al. Vision-based object detection and tracking: a review[J]. Acta Automatica Sinica, 2016, 42(10): 1466-1489.

[79] 王永森, 刘宏哲. 3D 目标检测技术的研究进展[C]//中国计算机用户协会网络应用分会 2019 年第二十三届网络新技术与应用年会论文集. 重庆:《计算机科学》编辑部, 2019: 177-182.
WANG Y S, LIU H Z. Study progress of advances in 3D object detection technology[C]//Proceedings of 2019 23rd Annual Conference on New Network Technologies and Applications of China Computer Users Association. Chongqing: Editorial Board of Computer Science, 2019: 177-182.

[80] GIRSHICK R. Fast R-CNN[C]//Proceedings of 2015 IEEE Interna-

tional Conference on Computer Vision. Piscataway: IEEE Press, 2016: 1440-1448.

[81] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[82] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//Proceedings of 14th European Conference on Computer Vision. Cham: Springer International Publishing, 2016: 21-37.

[83] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 779-788.

[84] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[85] CHEN Y J, TAI L, SUN K, et al. MonoPair: monocular 3D object detection using pairwise spatial relationships[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 12090-12099.

[86] MOUSAVIAN A, ANGUELOV D, FLYNN J, et al. 3D bounding box estimation using deep learning and geometry[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5632-5640.

[87] WENG X S, KITANI K. Monocular 3D object detection with pseudo-LiDAR point cloud[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Piscataway: IEEE Press, 2020: 857-866.

[88] LIAN Q, LI P, CHEN X. MonoJSG: joint semantic and geometric cost volume for monocular 3D object detection[J]. arXiv preprint, 2022, arXiv: 2203.08563.

[89] CHEN X Z, KUNDU K, ZHU Y K, et al. 3D object proposals using stereo imagery for accurate object class detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1259-1272.

[90] LI P L, CHEN X Z, SHEN S J. Stereo R-CNN based 3D object detection for autonomous driving[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 7636-7644.

[91] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2980-2988.

[92] YANG B, LUO W J, URTASUN R. PIXOR: real-time 3D object detection from point clouds[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7652-7660.

[93] YANG B, LIANG M, URTASUN R. HDNET: exploiting HD maps for 3D object detection[C]//Proceedings of Conference on Robot Learning. [S.l.:s.n.], 2018: 146-155.

[94] MEYER G P, LADDHA A, KEE E, et al. LaserNet: an efficient probabilistic 3D object detector for autonomous driving[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 12669-12678.

[95] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[96] SHI S S, GUO C X, JIANG L, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10526-10535.

[97] WANG H, CONG Y Z, LITANY O, et al. 3DIoUMatch: leveraging IoU prediction for semi-supervised 3D object detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 14610-14619.

[98] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 213-229.

[99] LIU X, XUE N, WU T. Learning auxiliary monocular contexts helps monocular 3D object detection[J]. arXiv preprint, 2021, arXiv: 2112.04628.

[100] HE T, SOATTO S. Mono3D++: monocular 3D vehicle detection with two-scale 3D hypotheses and task priors[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8409-8416.

[101] LUO S J, DAI H, SHAO L, et al. M3DSSD: monocular 3D single stage object detector[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 6141-6150.

[102] CHEN H S, HUANG Y Y, TIAN W, et al. MonoRUn: monocular 3D object detection by reconstruction and uncertainty propagation[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 10374-10383.

[103] LI C Y, KU J, WASLANDER S L. Confidence guided stereo 3D object detection with split depth estimation[C]//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2021: 5776-5783.

[104] NOH J, LEE S, HAM B. HVPR: hybrid voxel-point representation for single-stage 3D object detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 14600-14609.

[105] DING Z P, HAN X, NIETHAMMER M. VoteNet: a deep learning label fusion method for multi-atlas segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2019: 202-210.

[106] YANG Z T, SUN Y N, LIU S, et al. 3DSSD: point-based 3D single stage object detector[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11037-11045.

[107] LIU Z C, WU Z Z, TÓTH R. SMOKE: single-stage monocular 3D object detection via keypoint estimation[C]//Proceedings of 2020

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2020: 4289-4298.

[108] HERBST E, REN X F, FOX D. RGB-D object discovery via multi-scene analysis[C]//Proceedings of 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2011: 4850-4856.

[109] LIANG M, YANG B, CHEN Y, et al. Multi-task multi-sensor fusion for 3D object detection[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 7337-7345.

[110] LI Y, YU A W, MENG T, et al. DeepFusion: lidar-camera deep fusion for multi-modal 3D object detection[J]. arXiv preprint, 2022, arXiv: 2203.08195.

[111] GUPTA S, GIRSHICK R, ARBELÁEZ P, et al. Learning rich features from RGB-D images for object detection and segmentation[C]// Proceedings of European Conference on Computer Vision. Cham: Springer, 2014: 345-360.

[112] YOO J H, KIM Y, KIM J, et al. 3D-CVF: generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 720-736.

[113] SHIN K, KWON Y P, TOMIZUKA M. RoarNet: a robust 3D object detection based on RegiOn approximation refinement[C]// Proceedings of 2019 IEEE Intelligent Vehicles Symposium. Piscataway: IEEE Press, 2019: 2510-2515.

[114] QI C R, CHEN X L, LITANY O, et al. ImVoteNet: boosting 3D object detection in point clouds with image votes[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4403-4412.

[115] SINDAGI V A, ZHOU Y, TUZEL O. MVX-Net: multimodal VoxelNet for 3D object detection[C]//Proceedings of 2019 International Conference on Robotics and Automation. Piscataway: IEEE Press, 2019: 7276-7282.

[116] ZHANG W W, WANG Z, LOY C C. Multi-modality cut and paste for 3D object detection[J]. arXiv preprint, 2020, arXiv: 2012.12741.

[117] WANG C W, MA C, ZHU M, et al. PointAugmenting: cross-modal augmentation for 3D object detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 11789-11798.

[118] WANG Y, CHEN X, CAO L, et al. Multimodal token fusion for vision transformers[J]. arXiv preprint, 2022, arXiv: 2204.08721.

[119] BAI X, HU Z, ZHU X, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers[J]. arXiv preprint, 2022, arXiv: 2203.11496.

[120] PRAKASH A, CHITTA K, GEIGER A. Multi-modal fusion transformer for end-to-end autonomous driving[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 7073-7083.

[121] LIANG T, XIE H, YU K, et al. BEVFusion: a simple and robust LiDAR-camera fusion framework[J]. arXiv preprint, 2022, arXiv: 2205.13790.

[122] MANHARDT F, KEHL W, GAIDON A. ROI-10D: monocular lifting of 2D detection to 6D pose and metric shape[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2064-2073.

[123] GUPTA S, HOFFMAN J, MALIK J. Cross modal distillation for supervision transfer[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2827-2836.

[124] HUANG T T, LIU Z, CHEN X W, et al. EPNet: enhancing point features with image semantics for 3D object detection[C]// Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 35-52.

[125] PANG S, MORRIS D, RADHA H. CLOCs: camera-LiDAR object candidates fusion for 3D object detection[C]//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. New York: ACM Press, 2020: 10386-10393.

[126] YIN T W, ZHOU X Y, KRÄHENBÜHL P. Center-based 3D object detection and tracking[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 11779-11788.

[127] SIVIC J, RUSSELL B C, EFROS A A, et al. Discovering objects and their location in images[C]//Proceedings of 10th IEEE International Conference on Computer Vision Volume 1. Piscataway: IEEE Press, 2005: 370-377.

[128] SHI X P, CHEN Z X, KIM T K. Distance-normalized unified representation for monocular 3D object detection[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 91-107.

[129] CHEN Y P, WANG J K, LI J, et al. LiDAR-video driving dataset: learning driving policies effectively[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 5870-5878.

[130] PHAM Q H, SEVESTRE P, PAHWA R S, et al. A*3D dataset: towards autonomous driving in challenging environments[C]//Proceedings of 2020 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2020: 2267-2273.

[131] CHANG M F, LAMBERT J, SANGKLOY P, et al. Argoverse: 3D tracking and forecasting with rich maps[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 8740-8749.

[132] SCALE H. PandaSet: public large-scale dataset for autonomous driving[R]. 2019.

[133] HUANG X Y, CHENG X J, GENG Q C, et al. The ApolloScape dataset for autonomous driving[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2018: 1067-10676.

[134] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: a multimodal dataset for autonomous driving[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11618-11628.

[135] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability

in perception for autonomous driving: waymo open dataset[C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2443-2451.

[136] PATIL A, MALLA S, GANG H M, et al. The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes[C]//Proceedings of 2019 International Conference on Robotics and Automation. Piscataway: IEEE Press, 2019: 9552-9557.

[137] GEYER J. A2D2: AEV autonomous driving dataset[Z]. 2019.

[138] RAHMAN M M, TAN Y H, XUE J, et al. Notice of violation of IEEE publication principles: recent advances in 3D object detection in the era of deep neural networks: a survey[J]. IEEE Transactions on Image Processing, 2020, 29: 2947-2962.

[139] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[140] LIU Y C, FAN B, XIANG S M, et al. Relation-shape convolutional neural network for point cloud analysis[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 8887-8896.

[141] CASEY A D, SON S F, BILIONIS I, et al. Prediction of energetic material properties from electronic structure using 3D convolutional neural networks[J]. Journal of Chemical Information and Modeling, 2020, 60(10): 4457-4473.

[142] CASAS S, LUO W, URTASUN R. IntentNet: learning to predict intention from raw sensor data[J]. arXiv preprint, 2021, arXiv: 2101.07907.

[143] SUN Y X, ZUO W X, LIU M. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes[J]. IEEE Robotics and Automation Letters, 2019, 4(3): 2576-2583.

[144] CHEN X, SHI S, ZHU B, et al. MPPNet: multi-frame feature intertwining with proxy points for 3D temporal object detection[J]. arXiv preprint, 2022, arXiv: 2205.05979.

[145] XU J Y, MIAO Z W, ZHANG D, et al. INT: towards infiniteframes 3D detection with an efficient framework[J]. arXiv preprint, 2022, arXiv: 2209. 15215.

[作者简介]

**黄哲**（1996– ），女，中国人民大学信息学院博士生，主要研究领域为计算机视觉、3D 目标检测。

**王永才**（1978– ），男，博士，中国人民大学信息学院副教授、博士生导师，主要研究领域为物联网、智能感知、网络定位、视觉感知、惯导融合定位。

**李德英**（1965– ），女，中国人民大学信息学院教授、博士生导师，主要研究领域为物联网、智能网络算法与分析。