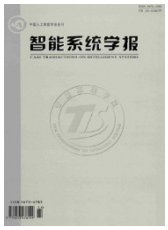

智能系统学报
CAAI Transactions on Intelligent Systems
ISSN 1673-4785,CN 23-1538/TP

《智能系统学报》网络首发论文

题目：面向自动驾驶目标检测的深度多模态融合技术
作者：张新钰，邹镇洪，李志伟，刘华平，李骏
收稿日期：2020-02-14
网络首发日期：2020-08-28
引用格式：张新钰，邹镇洪，李志伟，刘华平，李骏．面向自动驾驶目标检测的深度多模态融合技术[J/OL]．智能系统学报.
https://kns.cnki.net/kcms/detail/23.1538.TP.20200827.1334.016.html

DOI: 10.11992/tis.202002010
网络出版地址:

面向自动驾驶目标检测的深度多模态融合技术

张新钰[1,2], 邹镇洪[1,2], 李志伟[1,2], 刘华平[3], 李骏[1,2]
（1. 清华大学 汽车安全与节能国家重点实验室, 北京 100084; 2. 清华大学 车辆与运载学院, 北京 100084; 3. 清华大学 计算机科学与技术系, 北京 100084）

摘　要: 研究者关注利用多个传感器来提升自动驾驶中目标检测模型的准确率, 因此对目标检测中的数据融合方法进行研究具有重要的学术和应用价值。为此, 本文总结了近年来自动驾驶中深度目标检测模型中的数据融合方法。首先介绍了自动驾驶中深度目标检测技术和数据融合技术的发展, 以及已有的研究综述; 接着从多模态目标检测、数据融合的层次、数据融合的计算方法 3 个方面展开阐述, 全面展现了该领域的前沿进展; 此外, 本文提出了数据融合的合理性分析, 从方法、鲁棒性、冗余性 3 个角度对数据融合方法进行了讨论; 最后讨论了融合方法的一些公开问题, 并从挑战、策略和前景等方面作了总结。

关键词: 数据融合; 目标检测; 自动驾驶; 深度学习; 多模态; 感知; 计算机视觉; 传感器; 综述
中图分类号: TP274; TP212　　**文献标志码:** A　　**文章编号:** 1673−4785(2020)04−0001−14

中文引用格式: 张新钰, 邹镇洪, 李志伟, 等. 面向自动驾驶目标检测的深度多模态融合技术 [J]. 智能系统学报, 2020, 15(4): 1–14.
英文引用格式: ZHANG Xinyu, ZOU Zhenhong, LI Zhiwei, et al. Deep multi-modal fusion in object detection for autonomous driving[J]. CAAI transactions on intelligent systems, 2020, 15(4): 1–14.

Deep multi-modal fusion in object detection for autonomous driving

ZHANG Xinyu[1,2], ZOU Zhenhong[1,2], LI Zhiwei[1,2], LIU Huaping[3], LI Jun[1,2]
(1. State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 100084, China; 2. School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China; 3. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: In autonomous driving, there has been an increasing interest in utilizing multiple sensors to improve the accuracy of object detection models. Accordingly, the research on data fusion has important academic and application value. This paper summarizes the data fusion methods in deep object detection models of autonomous driving in recent years. The paper first introduces the development of deep object detection and data fusion in autonomous driving, as well as existing researches and reviews, then expounds from three aspects of multi-modal object detection, fusion levels and calculation methods, comprehensively showing the cutting-edge progress in this field. In addition, this paper proposes a rationality analysis of data fusion from another three perspectives: methods, robustness and redundancy. Finally, open issues are discussed, and the challenges, strategy and prospects are summarized.

Keywords: data fusion; object detection; autonomous driving; deep learning; multimodal; perception; computer vision; sensor; survey

作为自动驾驶技术的重要组成部分, 基于深度学习的目标检测技术持续受到研究人员的关注。尽管随着深度学习和计算机视觉领域的发展, 目标检测技术已经取得了显著的进步, 特别是将 DARPA[1]、PASCAL VOC2007[2] 等基于图像的目标检测任务的基准提升到了较高的水平。然而, 自动驾驶要求模型在复杂多变的场景下

收稿日期: 2020−02−14.
基金项目: 国家重点研发计划项目 (2018YFE0204300); 北京市科技计划项目 (Z191100007419008); 国强研究院项目 (2019GQG1010).
通信作者: 刘华平. E-mail: hpliu@tsinghua.edu.cn.

保持较高的准确率，基于单一传感器的算法即使在车道线检测这样的基础任务上也很难保持鲁棒性[3]。此外，不同于目标检测技术的其他应用场景，自动驾驶汽车上的多种传感器可以提供环境和车辆自身的多模态信息，并且它们在一定程度上存在互补关系[4]。因此，人们期待通过融合多模态数据来充分挖掘信息，并最终提高目标检测和其他自动驾驶模型的性能。本文先回顾了近年来目标检测技术和数据融合技术的发展，接着对面向自动驾驶的、基于多模态数据融合的目标检测技术进行了全面的概述，并比较和讨论了具体的融合理论和方法。与之前的研究[4-5]不同的是，本文只针对自动驾驶场景下的目标检测中的数据融合方法，从层次、计算和合理性等多个角度对其进行了全面深入的比较和分析，且进一步地总结了现有模型设计的策略，并给出了分析和建议。

1 背景

1.1 目标检测

作为计算机视觉领域的任务之一，目标检测是许多其他视觉任务的基础，如实例分割[6]与目标追踪[7]，旨在检测出不同类别物体的每个实例。考虑到数据的易得性和数据特征的丰富程度，其一般指以 RGB 图像为主要数据的目标检测，且通常在图像上使用边界框 (bounding box) 来定位物体并给出物体类别属性的概率[8]。根据定义，通常要求目标检测算法先在图像上搜索出可能包含目标的区域，再在此区域上进行分类，这种模型是多阶段模型。随着深度学习的发展，单步检测模型 (one-stage detection) 被提出，其可以在检测的同时进行分类，从而提高了检测速度。此外，针对不同模态的数据，检测的方法也不同。由于自动驾驶汽车上应用了激光雷达、雷达、深度相机 (RGB-Depth camera) 等多种传感器，因此自动驾驶中同样需要关注基于点云、深度图像或其他模态数据的方法，其中点云由激光雷达或雷达提供。对于图像上的目标检测，传统方法的准确率往往不如深度学习方法[8]，而后者则往往需要大型数据集和长时间的训练来学习特征。对于点云上的目标检测，优点是可以利用三维空间信息进行检测，缺点是空间维数的增加导致点云数据往往过于稀疏，造成模型拟合的效果不佳[9]。而深度图像，即深度相机记录的带有距离信息的 RGB 图像[10-11]，结合了图像和点云的特点，但因相机的性能不足尚未成为主流。综合来看，尽管针对多种模态数据的目标检测方法被不断提出，但大部分模型依然是基于图像。当前模型主要通过边界框的重合度来评价效果[12]，通过设置一个阈值 (intersection over union, IoU) 来决定是否正确预测。IoU 的计算方法一直在变化[13]，但并不影响本文对融合方法的讨论。

尽管在特定的数据集上，现有的基于计算机视觉和深度学习的目标检测方法取得了优异成绩，然而面对特定场景，特别是自动驾驶这类对鲁棒性、检测速度和准确率要求都很高的场景，现有模型的性能依然存在不足之处[4,8]。例如 2014 年针对基于车道线检测的综述[3]提到曝光对于视觉任务的影响，即使现有的网络针对小目标[14]、曝光不足[15]、分辨率低[16]的情形有所改善，然而极端过曝或欠曝的场景会导致图像数据对环境信息的记录严重损失，对此现有的基于单一图像网络依旧没有，也很难出现适用的解决方案。当自动驾驶汽车在路面上行驶的时候，很容易遇上光照变化幅度较大的区域，导致相机的记录失真，这将严重影响自动驾驶汽车的决策。为此，研究者考虑在视频流中利用连续帧的信息进行目标检测[7, 17,-18]，而另一主流研究方向则是利用多种传感器提供的多模态数据进行信息融合，再进行目标检测[4, 5]。2012 年德国的研究者提出了 KITTI 数据集[19]，其包含了多种车载传感器的数据和检测、分割等多个自动驾驶环境感知任务的标注。此后，Waymo[20]、Uber[21]和 Baidu[22]等公司先后推出针对自动驾驶的多模态数据集，为研究自动驾驶中的数据融合方法提供了极大便利。

1.2 多模态数据融合

考虑到上述在基于图像的目标检测模型中的问题，研究者考虑利用多模态数据的信息的互补来提升模型的鲁棒性。由于信息记录方式的不同，不同的传感器之间往往存在互补性[4]。比如 RGB 相机往往在光照条件不佳时难以记录有效信息，然而主动感知的传感器，如激光雷达、雷达和深度相机等，则不易受到外部环境条件的影响。对于激光雷达和雷达，它们记录的点云过于稀疏，获得的低分辨率数据难以用于高精度的检测，而 RGB 图像则可以提供稠密的数据。因此，如何理解与利用多种模态数据之间的关联与互补之处，成了多模态数据融合在应用中的重要问题。

具体到自动驾驶场景，不同的车载传感器既可以提供对同一环境的感知信息，如对前方道路

的 RGB 图像、热成像、深度图像、激光雷达点云和雷达点云等, 也可以提供对汽车自身的感知信息, 如车辆的行驶速度、路径等, 为连续地感知环境提供重要的估计参数。为此, 自动驾驶中的数据融合可以在多个任务中发挥作用, 比如目标检测[23]、目标跟踪和即时定位与建图 (simultaneous localization and mapping, SLAM)[24]。特别地, 不同于其他应用场景, 针对自动驾驶汽车的目标检测可以利用多种车载传感器。由于这些传感器被安装以记录前向场景的信息, 因此它们包含了对同一环境的多模态信息, 这使得它们既能很容易地配对, 又能被发掘出互补的信息[23]。对此最常见的融合模式是激光雷达与 RGB 相机的融合, 激光雷达点云可以主动感知较大范围内的物体, 因此不受光照条件的影响, 而 RGB 图像所提供的色彩、纹理等视觉信息则可以被用于更高精度的视觉任务[25-27]。然而现有的融合方法同样具有问题和挑战, 如多个传感器之间的配准[28]、部分传感器失灵的情形[26], 以及对更多样的融合方法的探究[4]。

1.3　已有研究

尽管已经有许多基于融合方法的目标检测模型被提出, 然而依然没有论文对目标检测模型中的融合机制进行完整且深入的研究。表 1 将本文与相关的文献综述进行了对比。Eduardo Arnold 等[9] 总结了自动驾驶任务中的 3D 目标检测方法, 其涉及到部分融合模型, 然而仅局限于 3D 检测的场景, 且没有深入分析融合方法的合理性。Feng 等[4] 总结了自动驾驶中适用于数据融合的数据集、感知任务、融合方法和现有问题等, 特别是对目标检测和语义分割中的融合方法进行了全面的归纳和分类, 然而没有对数据融合的冗余性进行分析, 也没有对基于数据融合的目标检测模型中的其他部分进行比较归纳; 罗俊海等[5] 对基于数据融合的目标检测方法进行了综述, 且采用了前面两个研究不一样的归纳方法, 然而其研究并非针对特定场景, 并且同样缺乏对数据融合的冗余性分析或合理性分析。特别地, 近年来, 目标检测中的融合方法缺乏统一、明确的定义, 不同的论文中模糊地遵循了“前融合, 中间融合, 后融合”的分类方法[23, 29-30], 然而各自在具体的实现细节上仍存在差异, 且现有的按融合阶段划分的方法, 在集成模型中不能很好地反映出融合步骤对于模型的作用。为此, 提出了新的根据融合结果作用的划分方法, 并给出了具体的定义, 相关细节将在下文中描述。

表 1　近年目标检测综述论文对比表

Table 1　Comparison of object detection review papers

文献	检测目标	划分方法	面向自动驾驶
[4]	2D&3D	基于融合层次	是
[5]	2D&3D	基于融合层次	否
[8]	2D	不涉及融合	否
[9]	3D	不涉及融合	是
本文	2D&3D	基于融合层次	是

2　融合方法

2.1　多模态目标检测

2.1.1　基于 RGB 图像

在应用深度学习技术前, 曾出现了 VJ Det、HOG Det 等方法[31-33]。2008 年 Felzenszwalb 等[34] 在所提出的 DPM 方法中首次涉及了边界框的概念, 此后对边界框的回归成为目标检测模型的经典思想。2012 年, AlexNet[35] 的出现成为了深度学习和计算机视觉发展的开端。从此, 基于卷积神经网络 (convolutional neural network, CNN) 的方法逐渐成为计算机视觉任务的主流[8]。从 2014 年起, Girshick 等[36-38] 先后提出了 R-CNN、Fast R-CNN, 基于 R-CNN 发展出来的神经网络成为两步目标检测 (two-stage detection) 的主要基准模型。R-CNN 先通过选择搜索 (selective search) 生成候选边界框, 再通过 CNN 提取特征, 最后进行分类。此外, 为了提高计算性能, Ren 等[15] 于 2015 年提出了 Faster R-CNN, 所使用的 Region Proposal Network(RPN) 成为经典的目标检测模型的设计思想。此后出现的特征金字塔网络 (feature pyramid networks, FPN)[39] 通过融合不同层次的语义来充分挖掘图像信息, 许多模型采用了这种结合浅层和深层语义的设计。

与两步检测相对的是单步目标检测 (one-stage detection), 其经典模型是 YOLO(you only look once)[40] 和 SSD(single shot detector)[41]。由名字可见, 单步检测不同于两步检测的提出候选边界框再进行分类的模式, 而是同时在图像的某一部分上预测边界框和分类。Redmon 等先后提出了 3 个版本的 YOLO 模型: YOLO[40]、YOLO9000[42]、YOLOv3[14], 其中 YOLO9000 和 YOLOv3 先后在预测边界框时使用锚点 (anchor) 的思想来替代全连接层、采用了多尺度模型、对边界框进行逻辑回归, 最终实现在提高准确率的同时保持了较高的检测速度, 且 YOLOv3 对于小目标的检测性能有显著提升。除了 YOLO 系列网络, 2015 年提出

的 SSD[41] 和 2017 年的 Retina-Net[43] 也受到了广泛的研究和使用。此后，仍然不断有各种模型在基准数据集上取得突破，然而这些模型依然遵循了单步或两步模型的经典设计。

2.1.2　基于激光雷达点云

激光雷达由于其主动感知的特性，以及相对雷达提供了更密集的点云，因此在自动驾驶的传感器中受到较多的应用[44-46]。图 1 中给出了 KITTI 数据集中 RGB 图像和激光雷达点云的示例，其中点云反射强度图被投影到相机成像平面，且所显示的均为灰度图。激光雷达不仅可以提供较大的感知范围，扫描半径可达 50~70 m，甚至更远，而且激光雷达可以提供环境的深度信息和反射率，且不受环境光照的影响。激光雷达数据为三维点云格式，因此点云的坐标自然地提供了物体相对激光雷达的三维空间坐标，可以用于三维空间的目标检测；点云的反射强度值反映了物体表面的材质，因此不同的物体可以根据反射强度被很容易地区分开。

图 1　KITTI 数据集示例[19]

Fig. 1　KITTI dataset examples

对于点云上的目标检测，现有多种方法，它们和点云的数据表现方式有关。由于点云的空间特性，它既可以在三维空间中执行三维检测[47]，也可以投影到二维平面，利用基于图像的二维目标检测模型进行计算[48]。具体地，投影主要包括前视图投影和鸟瞰图投影 (bird's eye view, BEV)，其中前视图投影一方面可以得到较为稠密的投影图像，另一方面可以投影到车载相机的像素平面上，从而可以和相机图像融合来执行目标检测。然而在模型中点云的 BEV 视图受到了更多的应用，主要有 3 个原因：1）物体在投影到 BEV 视图时会保留物理尺寸，而其他投影视图会产生透视效果；2）BEV 视图中的物体很少出现遮挡问题，且在空间上分布离散；3）在道路场景中，由于对象通常位于地面上并且垂直位置的变化很小，因此鸟瞰图有利于获取准确的 3D 边界框。此外，激光雷达还可以在投影图像上提供丰富的语义信息，比如深度信息[48]、反射率信息[48] 和物体高度信息[49] 等，根据多重语义信息可以更有效地在点云上进行视觉计算。

2.1.3　基于 RGB-D 图像和多光谱图像

深度相机，又称 RGB-D 相机，在 RGB 图像的对应像素上提供了深度信息，从而让所提供的 RGB-D 图像在一定程度上结合了 RGB 图像和激光雷达的优点。且由于 RGB-D 所提供的点云和图像自然配准，因此在计算上较单独的激光雷达点云和 RGB 图像更加方便。近年来研究者探索了 RGB-D 图像的信息挖掘方法，主要是将 RGB-D 图像中的两种信息分拆处理，包括对图像和点云投影得到的深度图像的联合处理[10]、对图像和点云的联合处理[50]。与深度图像类似的是多光谱图像。多光谱图像可以由多光谱相机或航空相机获得，也可以通过 RGB 相机和热成像相机 (红外光相机) 配准后获得。多光谱图像，特别是基于物体表面温度的热成像，可以避免环境可见光源对成像的影响，从而可以在包括夜间的多种环境下获得具有区分度的图像。Rutgers 大学[29] 和 Bonn 大学[30] 的研究者分别在 2016 年发表了基于多光谱成像的目标检测技术，是最早的一批利用多光谱图像进行目标检测的研究。此后，浙江大学的研究人员[51] 探究了多光谱成像在全天候目标检测下的应用前景，并获得了良好的实验效果。

2.1.4　基于其他数据来源

表 2 中列出了常见的自动驾驶场景中的传感器 (或数据来源) 以及对应的数据模态、数据提供的信息和使用的目标检测任务。

表 2　多种模态数据的比较

Table 2　Comparison of multiple modal data

传感器	模态	包含信息	检测目标
RGB相机	图像	RGB信息	2D
全景相机	图像	全景RGB信息	2D
深度相机	图像	RGB信息、深度	2D&3D
多光谱相机	图像	多光谱图像	2D
激光雷达	点云	深度、反射强度	2D&3D
雷达	点云	深度、径向速度	2D&3D
毫米波雷达	点云	深度、径向速度	2D
高精地图	地图	地图先验信息	2D&3D

除了上述 4 种常见的数据，高精地图 (HD map)、雷达 (radar) 和毫米波雷达 (millimeter wave radar) 同样被应用于自动驾驶的目标检测中。HDNet[52] 提供了一种融合激光雷达点云与高精地图的方法，且点云可以用于高精地图的构建，从

而建立起点云和高精地图之间的联系。通过往高精地图上添加交通语义信息, 如信号灯、道路指示标志、车辆信息等, 可以充分利用道路上的先验信息, 从而提高点云上目标检测模型的性能。雷达和毫米波雷达均可以提供点云, 但所提供的信息只包含深度 (三维空间) 信息, 且其点云较激光雷达点云更为稀疏, 因此应用范围不及激光雷达。然而雷达, 特别是毫米波雷达具有更大的射程, 可以提供更大距离的障碍物信息, 近年来有基于雷达[25]和毫米波雷达的研究[53], 同样值得关注。

2.1.5　基于多模态数据的目标检测

近年来, 自动驾驶领域对基于多模态数据融合的目标检测技术的研究兴起, 一方面是出于上述对单一模态数据的缺陷的考虑, 另一方面是出于对车载传感器稳定性的考虑。表 3 总结了近年来自动驾驶场景下基于多模态数据融合的深度目标检测方法。

表 3　深度目标检测数据融合方法统计

Table 3　Statistics of deep target detection data fusion methods

文献	传感器	点云表示方式	数据融合	特征融合	结果融合	辅助估计
[53]	毫米波雷达、RGB相机	前视图	√		√	√
[56]	激光雷达、RGB相机	前视图、鸟瞰图、体素化	√	√	√	√
[25]	雷达、长焦相机、短焦相机	前视图	√	√		
[57]	激光雷达、RGB相机	3D点云、体素化	√	√		
[26]	激光雷达、RGB相机	鸟瞰图		√		
[55]	激光雷达、RGB相机	前视图	√	√		
[27]	激光雷达、RGB相机	体素化的前视图、鸟瞰图			√	
[58]	激光雷达、RGB相机	鸟瞰图		√		√
[50]	深度相机	3D点云		√	√	
[51]	RGB相机、热成像相机	/	√	√		
[59]	激光雷达、RGB相机	鸟瞰图		√		
[60]	激光雷达、RGB相机	前视图		√		
[52]	激光雷达、高精地图	高精地图、栅格化鸟瞰图	√			√
[61]	激光雷达、道路先验信息	鸟瞰图		√		
[62]	激光雷达、RGB相机	6通道鸟瞰特征图		√	√	
[28]	激光雷达、RGB相机	前视图			√	
[63]	激光雷达、RGB相机	稀疏深度图、稠密深度图	√	√		
[64]	激光雷达、RGB相机	3D点云		√	√	
[65]	激光雷达、RGB相机	按深度生成3个前视图		√	√	
[66]	激光雷达、RGB相机	鸟瞰图	√		√	
[67]	深度相机	前视图	√	√	√	
[23]	激光雷达、RGB相机	鸟瞰图、前视图	√			
[68]	激光雷达、RGB相机	3D点云		√	√	
[48]	激光雷达、RGB相机	稠密深度图、稠密强度图		√		
[10]	深度相机、专家先验信息	前视图		√		
[49]	激光雷达、RGB相机	前视图	√	√		
[29]	RGB相机、热成像相机	/	√	√	√	
[30]	RGB相机、热成像相机	/	√	√		
[69]	激光雷达、RGB相机	稠密强度图			√	

针对第一方面, 近年来不断有针对多模态数据的目标检测方法被提出, 如 Cho 等[54]提出的对雷达点云、激光雷达点云和相机图像的专家融合方法, 用于车辆检测与跟踪; Schlosser 等[49]提出的基于 HHA 采样的激光雷达点云和相机图像的融合方法, 两种数据组成了 6 通道的扩展图像, 用于行人检测; DOU 等[55]提出了融合点云前向投影的深度图和相机图像的融合方法, 是一项

通过图像辅助点云上的 3D 检测的方法；2017 年清华大学提出 MV3D 模型[23]，采用了复杂的网络架构来融合图像与两个视角的点云投影，取得了当时的最佳性能。除此外还有众多的数据融合方法被提出，下文将从融合的层次、融合的计算方法两个角度来划分不同的融合方法，后一节将对现有模型中数据融合的冗余性进行分析。

2.2　数据融合的层次

由于激光雷达点云在融合模型中的表示方式影响到融合方法的选择，因此表 3 中也列出了相应的信息。根据 Feng 等的文献[4]，按数据融合步骤出现在目标检测模型的不同阶段，融合方法可以分为前融合 (early fusion)、中间融合 (middle fusion or deep fusion) 和后融合 (late fusion)3 种。其中，前融合针对原始数据或仅经过预处理的数据的融合，后融合将模型的多个分支的运算结果融合得到最终结果，中间融合则结合了前融合和后融合的特点，将多个模态的数据或其对应的特征图进行融合，再继续运算。中间融合可以仅存在于模型的中间步骤，也可以贯穿整个模型。在部分论文中，3 种融合也称为数据级融合、特征级融合和决策级融合[5]。然而，在多层次、多模型结合的深度学习模型中，模型分支的结果也是整个模型的中间特征[70]，大部分融合方法被归为中间融合，导致按阶段的划分方法缺乏区分度。为了更加直观地区分模型融合的程度，本文从融合的层次角度进一步对融合方法作了区分，分别是数据融合、特征融合、结果融合和辅助估计。根据本文提出的划分方法，融合方法划分的依据不再是融合步骤在模型中的先后位置，而是融合结果在模型中充当的作用，并且融合的划分不会受到融合数据的形式的限制，从而更加直观地反映了融合操作对于模型的作用。不同融合层次的定义和分析如下，且在图 2 中给出了示例。

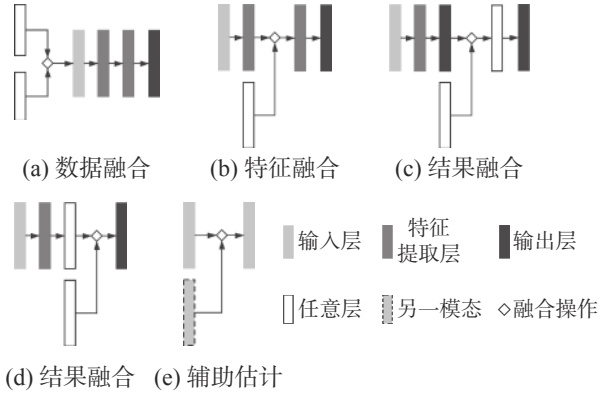

图 2　不同融合层次示例

Fig. 2　Examples of different levels of fusion

2.2.1　数据融合

数据融合，指输出数据将作为后续模型输入的融合操作。一方面，模型可以是完整的模型，也可以是作为集成模型某一分支或某一部分的子模型，；另一方面，输入的数据，也即被融合的数据，既可以是来自传感器的原始数据或仅经过预处理的数据[30, 63, 67]，也可以是集成模型的子模型输出的特征向量、特征图或计算结果[55- 56]。文献 [67] 对由深度相机提供的 RGB 图像和深度图像分别进行处理后融合，文中提供了多个可能的融合位置，当模型提取特征前就通过向量连接时，融合结果作为后续完整模型的输入，属于数据融合，也就是文献 [5] 中所提到的数据级的融合，或文献 [4] 中提到的前融合。在 DOU 等[55] 的研究中，点云被投影到相机像素平面，与语义分割后的图像和原始图像进行融合，所得融合图像作为后续目标检测模型的输入，且后续模型为完整的检测模型，因此是数据融合。在 Taohua 等[53] 的论文中，激光雷达点云被投影到 RGB 图像的像素平面，并被添加到图像作为扩展的第 4 通道，展示了一种常用的与 RGB 图像融合的方法。

2.2.2　特征融合

特征融合，指输出数据将作为模型的特征向量或特征图 (后续统称特征图) 的融合操作。具体地，特征融合通常指模型输入在模型计算过程中的特征图之间的融合操作，且融合的结果继续作为当前模型的特征图参与计算[25, 26]，但在某些模型中，原始数据经过预处理后也可能与特征图进行特征融合[56, 57]。DeepAI 在 2019 年的论文中提及了一种鲁棒性的融合方法，其中点云和图像分别由单独的 SSD 分支处理，其特征图进行交叉融合，同样的方法在另一篇车道线检测的论文中同样出现[71]。Liang 等[56] 提出了一种复杂的融合模型，其设计的 dense fusion 模块将 4 个分支的数据进行融合，再作为点云和图像两个处理网络的特征图继续处理，展现了特征层次的融合方法对数据来源的兼容性。UC Berkley 的研究者展现了一种新颖的特征融合方法[59]，通过设计一个稀疏矩阵来实现图像到图像的鸟瞰图、点云鸟瞰图到点云前视图的转化，并通过向量的连接来融合两种模态的特征。两种典型的融合方法在文献 [30] 中被展现：采用叠加点云到图像的扩展通道的方法来实现数据融合，通过连接特征图的方法来实现特征融合。

2.2.3　结果融合

结果融合，指输出数据作为模型的最终目标

输出, 或者输入数据本身是目标输出且融合的结果用于结果修正的融合操作。按定义, 结果融合主要包括并行的模型分支的计算结果的融合操作, 以及串行模型中, 后续子模型在已有结果上进行的针对同一目标输出的优化计算[56]。在文献 [69] 中, 模型将基于点云和基于图像两个分支的提取的区域建议 (region proposal) 进行融合, 从而得到所有的候选检测框, 由于目标检测模型的候选框由区域建议回归得到, 因此可以视为针对候选框的结果融合。类似的融合在文献 [29] 中也有体现。文献 [48] 可以视为一项只包含结果融合的研究。其所提及的融合模型对 3 个模型分支的边界框融合, 虽然此后仍有特征提取的操作, 但由于融合后的计算仅针对边界框 (网络输入) 的提升, 因此将后续对边界框处理的整个子网络视为结果融合模块。类似的设计见于文献 [68], 所提出的模型同样专注于对已有的检测框融合、调整, 这部分的融合操作仅限于对已有结果的提升, 因此属于结果融合。除此之外, 结果融合还广泛地被应用于文献 [51, 62-66] 中。

2.2.4　辅助估计

辅助估计, 指基于某些模态的数据, 对另一模型的数据进行估计, 以使其获得更佳的表示。不同于上述 3 种融合方式, 辅助估计没有减少数据的模态, 且辅助估计模块的存在与否不影响融合模型的完整性[28, 53, 56, 58]。从表 3 可见, 辅助估计并非常用的融合方法, 因为其要求融合的数据模态之间同时具有较强的相关性与较多的信息互补关系, 相关性弱的数据之间难以计算先验估计, 信息互补较少的数据之间难以获得数据质量的明显提升。在自动驾驶的感知方面, 较为常见的是激光雷达和 RGB 图像之间的辅助估计[28, 56], 如 KITTI 的 depth completion 任务[19], 它要求基于图像和稀疏点云估计得到稠密点云。文献 [56] 中的 Depth Completion 模块, 其网络基于由稀疏点云深度图和图像拼接得到的 4 通道扩展图像, 估计出稠密的点云深度图, 从而为后续的检测提供更佳丰富的空间深度信息。BMW 的研究者[28] 关注了多模态数据融合的另一个问题: 传感器或数据的配准。它们将点云强度图投影到图像平面进行校准, 估计出了从点云到图像平面的投影矩阵。最新的研究[53] 中也涉及了辅助估计, 模型通过雷达点云的深度信息来学习图像中物体的缩放与距离的关系, 从而为图像上的目标检测选择合适的候选框。

表 3 总结了近年来自动驾驶场景下基于多模态数据融合的深度目标检测方法, 不难发现, 特征融合是最常见的融合方法之一, 研究者认为这可能是因为特征融合可以有效地利用不同模态的深层语义信息。文献 [29-30, 49, 63-67] 中对不同位置的融合进行了测试, 实验结果表明, 较为靠后的特征融合更有利于挖掘信息, 提升模型性能, 然而过于靠后的融合只能提取高层语义信息, 反而降低了融合对模型性能的提升。现有的模型更多地考虑了多层次、多步、带有自适应权重的融合方式, 如文献 [10] 和 [60] 中均考虑了自适应权重, 研究者基于 LSTM[72] 设计了门控单元, 通过计算不同模态对融合模型的贡献来分配权重, 从而使得模型可以自适应地选择最有力的数据模态。Freiburg 大学的研究者[10] 简单地融合了基于深度图和 RGB 图像训练的模型, 实验表明, 带有权重的融合能够有效地反馈各模态数据在融合模型中的实际贡献。然而, 现有的实验场景仍然过于简单, 当应用场景发生明显、复杂的变化时, 如何设计一个鲁棒的自适应融合权重计算模型, 依然是一个开放问题。

2.3　数据融合的计算方法

在目标检测的融合方法中, 出现过多种多样的融合计算方法, 主要可以归为连接法、合并法、相加法、子网络法和辅助估计法 5 种。与文献 [4] 不同的是, 本文的分类方法添加了子网络法, 并提供了专家混合法 (mixture of experts) 的几种可能形式。

2.3.1　连接法

作为最直接的融合方法之一, 连接法直接将不同模态的数据或其特征图相连接, 特征的增加直接体现在数据或特征图的维数增加上[30, 57, 67]。常见的连接方式有两种: 1)第二模态的特征图 (数据) 拼接到第一模态的特征图 (数据) 上, 特征图的层数不变, 但每层的尺寸变大, 这种方法又称 concatenate; 2)将第 2 模态的特征图 (数据) 作为第一模态的特征图 (数据) 的扩展通道, 融合后的特征图通道变多, 每个通道的尺寸不变。前一种融合方式多见于特征融合阶段, 第 2 种融合方式多见于数据融合阶段, 但为了将第 2 种融合方式用于特征融合, 可以在融合后使用 1×1 卷积来调整通道数。第 2 种方法可以理解为 extended channel。PointFusion 模型[64] 将点云和 RGB 图像分别提取的特征图使用 concatenate 连接, 文献 [51] 同样也使用了 concatenate, 而文献 [49, 53] 则采用了扩展通道的形式。

2.3.2　合并法

合并法适用于多个同类候选元素, 如多个分支的同类目标输出的融合和候选框的融合[25, 51]。

文献 [51] 对模型分支输出的边界框和分类结果采用了加权求和的形式，文献 [25] 采用了长短两个焦距的相机，因此具有两个视野的图像，其中长焦距相机的图像是短焦距相机图像的子集，同时也具有更高的分辨率。因此模型将长焦相机图像中的边界框叠加到短焦相机图像的对应位置上，以在保持视野的同时提高模型对远处小目标的检测性能。

2.3.3　相加法

相加法通常应用于特征图的融合[60, 62]和感兴趣区域 (region of interest, ROI) 的处理，后者包括 ROI 的合并[23]以及 ROI 叠加到特征图上作为检测的约束条件[66]。有些模型考虑通过特征图的加权求和来融合特征[60]，对于 ROI 区域，既可以当作普通的特征图来直接合并，也可以作为模型的多层次约束。如文献 [66] 提出的方法，不断根据点云滤波减少搜索范围，将点云投影到图像上，并在点云范围内进行检测和分类。

2.3.4　子网络法

为了提升结果融合的效果，部分模型采用子网络来提取已有的目标输出的结果的特征，从而生成更精细的检测结果，这里的网络又称为网中网 (network in network, NiN)。文献 [68] 中设计了一个结果融合网络，对两个单独分支的检测框和分类结果分别进行了融合；类似地，文献 [48] 对 3 个分支输出的结果重新提取特征，经过多层感知机 (MLP) 和非极大抑制算法 (NMS) 后回归得到新的边界框和分类结果。此外，使用 NiN 来融合输出和高层特征的方法十分相似，文献[51, 56, 65]中将多个分支的结果用 NiN 融合后作为特征继续计算，而文献 [23] 则融合了多个分支的特征图作为特征继续计算，两者在结构上的相似性与它们性能的表现之间的关系还有待探究。

2.3.5　专家混合法

不同于上述可以泛化的方法，对于特定组合的多模态数据，需要使用特定的融合方法。文献 [54] 为了融合 RGB 图像、激光雷达点云和雷达点云，提出了一种基于运动模型的目标跟踪算法，以针对性将 3 种模态中的信息充分用于目标检测和追踪。文献 [65] 采用了三阶段的目标检测方法，首先在分段点云中生成检测种子，再根据种子计算锚点 (anchor)，最终生成并调整检测框。一种常见的任务是对稠密点云的估计，文献 [56, 63] 中均对点云的深度图进行了补全估计，以获得更稠密的空间信息。值得注意的是，专家融合法多见于非常用模态数据和先验信息的融合，比如高精地图[52, 75]、地图语义信息[61]和专家先验信息[10]，其中地图语义信息和专家先验信息的融合在深度学习模型中较为少见，对其的研究还有待开拓。

3　数据融合的合理性分析

尽管上述的数据融合模型都在实验中获得了良好的表现，很少有人去探究其中的原理，现有的融合模型中也很少有控制融合模型变量的对比实验，表 4 中展示了近年论文中的分析实验。

表 4　数据融合的合理性分析

Table 4　Analysis of the rationality of data fusion

文献	传感器	合理性分析	方法
[51]	RGB相机、热成像相机	融合方法分析	比较不同层次融合对性能的提升
[63]	激光雷达、RGB相机	融合方法分析	比较不同层次融合对性能的提升
[67]	深度相机	融合方法分析	比较不同层次融合对性能的提升
[49]	激光雷达、RGB相机	融合方法分析	比较不同层次融合对性能的提升
[29]	RGB相机、热成像相机	融合方法分析	比较不同层次融合对性能的提升
[30]	RGB相机、热成像相机	融合方法分析	比较不同层次融合对性能的提升
[25]	雷达、长焦相机、短焦相机	融合方法分析、数据冗余性分析	对比融合方法、融合与否的模型性能
[56]	激光雷达、RGB相机	融合方法分析、数据冗余性分析	比较不同层次和模态融合对性能的提升
[26]	激光雷达、RGB相`机	模型鲁棒性分析	使用随机模态丢失来提升模型鲁棒性
[52]	激光雷达、高精地图	模型鲁棒性分析	使用随机模态丢失来提升模型鲁棒性
[60]	激光雷达、RGB相机	模型鲁棒性分析	对多模态数据加噪声来验证模型的鲁棒性
[10]	深度相机、专家先验信息	模型鲁棒性分析	在连续变化的场景中测试融合的鲁棒性

文献 [4] 中提及, 目前仍缺乏足够的关于数据融合的冗余性的分析。不仅仅是冗余性分析, 关于融合模型的鲁棒性的分析也很少。但现有的研究中较多地涉及了融合方法的研究, 包括对融合层次和融合操作方法的对比实验, 为后续的研究提供了一定的实证支持。本文将融合方法的分析、模型的鲁棒性分析和数据冗余性分析并称为数据融合模型的合理性分析。如表 4 所示, 本文综述了自动驾驶场景下目标检测模型中的数据融合方法的合理性研究进展, 后续的研究将能在此基础上开展。3 个方面的分析如下:

3.1　融合方法分析

如上文所述, 目标检测中包含了多样的可能的融合方法, 不仅体现在融合层次的多样性上, 而且也体现在融合的具体计算方法的多样性上。通常的做法是根据控制变量原则, 分别在模型的不同阶段进行融合并比较结果[29, 63, 67, 73], 而文献 [25] 则比较了融合与否 (RGB 图像是否融合雷达点云) 以及融合方式 (连接法或相加法) 对模型的影响。现有的实验结果表明, 中间融合 (对应本文所述特征融合) 得到了最佳结果, 并且较晚的融合比较早的融合和过于晚的融合更佳, 而对于最佳融合阶段, 目前还没有研究给出相关估计。而对于融合的操作方法, 尽管文献 [4] 给出了一个全面的综述, 但目前还没有量化的比较。

3.2　模型鲁棒性分析

融合模型可以有效地结合多种模态数据的优点, 挖掘场景更深层次的语义信息。然而在实际应用中, 可能遇到极端的场景和数据, 或者遇上传感器异常的情况, 两种情形都将导致模型缺乏部分模态的数据或被输入部分模态的异常数据。与单一模态的模型不同, 由于具有多个模态数据的输入, 多模态模型可能在部分模态数据失效的同时持续运行而不被察觉, 但模型的性能已经受到影响; 另一种情况则与之相反, 研究人员希望多模态模型可以在部分模态失效的时候, 融合模型可以尽可能地减少部分模态损失导致的模型性能的下降[26, 60]。综上所述, 模型的鲁棒性应体现在两个方面: 对异常数据的检测; 自适应调节[10]。相对于自适应模型, 融合模型中对异常数据的检测还不多, 且现有的模型鲁棒性的研究主要在实验验证阶段: 文献 [52] 采用随机模态丢失的方法来训练模型, 一方面可以提升模型的鲁棒性, 另一方面验证了多模态模型在模态缺失的情形下依旧可以具有一定的准确率。类似地, 文献 [26] 采用了同样的方法, 模拟了雪雾天气下传感器的工作状态并通过数据融合获得了一个鲁棒的模型。关于自适应模型, 文献 [10] 和文献 [60] 中均考虑了自适应权重, 从而使得模型可以自适应地选择最有力的数据模态。

3.3　数据冗余性分析

尽管不断有新的融合模型被提出, 但却很少有研究定量或定性地分析, 不同模态的数据融合对模型性能的可能提升程度。为此存在两个问题: 1) 两个不同模态的数据融合造成的信息增益; 2) 不同模态数据的融合对模型不确定性的减少的影响有多大。根据目前的了解, 只有文献 [56] 量化比较了不同模态数据对于融合效果的贡献 (激光雷达点云和图像、建图模块 mapping、深度图、修正模块 refine 的融合)。此外, 而文献 [25] 中对融合信息 (雷达点云) 对模型影响的探究也体现了数据的冗余性分析。现有的实验结果表明, 在大部分任务上, 数据融合对模型的性能有提升, 但不是所有任务都有提升[25], 并且融合模型的效率也需要加入考虑。

3.4　融合实践的挑战与策略

3.4.1　挑战

尽管不断有基于数据融合的自动驾驶的目标检测模型出现, 然而目前依然缺乏对此的系统的描述和深入的探究。一方面, 缺乏针对目标检测中的数据融合方法的理论研究; 另一方面, 目前使用的融合方法仍显得过于简单[4]。从融合的合理性分析, 即使只考虑现有的方法也存在很多工作有待完成, 不仅缺乏不同方法之间的对比实验, 而且融合模型的可解释性也是一个重要问题。本文建议从模型的鲁棒性和数据冗余性两个角度来考虑融合的可解释性。

根据现有的论文方法, 在融合方面主要的挑战包括但不限于: 可解释的融合计算、对最佳融合阶段的估计、对融合数据的异常检测、自适应的融合机制、跨模态数据融合的信息增益估计以及数据融合对模型不确定性变化程度的估计。当前已经出现了一些验证实验和对比实验, 但还需要更深层次的研究。

3.4.2　策略

尽管现有的模型可能会在融合计算上偏向某一方法, 但目前仍然没有定量的分析或者推导来证明某一方法的优越性。然而, 在表 4 所记录的论文中, 对融合层次比较的实验结果基本支持中间融合优于前融合和后融合这一观点, 对应到本文的分类方法中则为特征融合优于数据融合和结果融合。然而, 按传统的前中后阶段划分对于集

成模型中的融合方法区分度不够高，因此简单地进行“中间融合”并不能很好地指导模型的设计。相反地，应当考虑融合的对象对于原始数据语义表征的深浅程度。

近年来，出现了在一定程度上超越“特定阶段的融合”的融合模式，例如自适应阶段的融合、同模态数据的融合等。在文献[71]中通过在神经网络的多层对不同模态数据的处理分支进行耦合，让神经网络学习融合参数，进而达到学习最优融合位置的目的。然而，由于会显著增加计算量，因此在深层网络中应当预先筛选融合的可能位置，缩小搜索范围。特别地，由于多模态数据的特征表示可能不同，现有方法通常考虑独立的前处理流程，并采取中间融合 (特征融合) 为主的策略。然而文献 [74] 中给出了新的思路，其采用图像序列数据结合 LSTM 网络结构进行车道线检测，挖掘了同一模态数据的潜在关联。同时，将非时序数据组合为时序数据，则可以和时序数据，例如音频，建立时序上的联系。除了独立的预处理分支外，多模态数据可以依据各自的特点重组数据形式，以实现更有效的融合。

进一步地，考虑仿生的角度，认知信息加工理论把人脑活动理解成一个信息加工系统，外界刺激导致多种感受器产生瞬时记忆，随后转换为工作记忆，并结合长时记忆进行对外界信息的理解判断。针对基于多源传感器的车辆环境感知系统基础，本文相应地提出一种面向自动驾驶的多模态认知融合框架的建议：通过特异性处理、多阶段多层次的融合等方式，提高道路场景下目标检测、语义分割等任务的准确率和鲁棒性；提升目标跟踪的精度，进而增强车辆环境感知系统的准确性和可靠性，为自动驾驶车辆迈向现实提供有力的理论依据和技术支持。

图 3 给出了一种面向自动驾驶的多模态认知融合框架。考虑到多模态数据的特征空间不一致，对多模态数据进行针对性的特征提取，然后在多模态整合模块中融合多通道特征;融合过程将结合来自连续帧的先验知识，能够增强整个系统对环境的认知能力，大幅提高目标检测、语义分割、目标跟踪等算法在曝光异常、雨雾天气等异常场景下的准确率和鲁棒性。

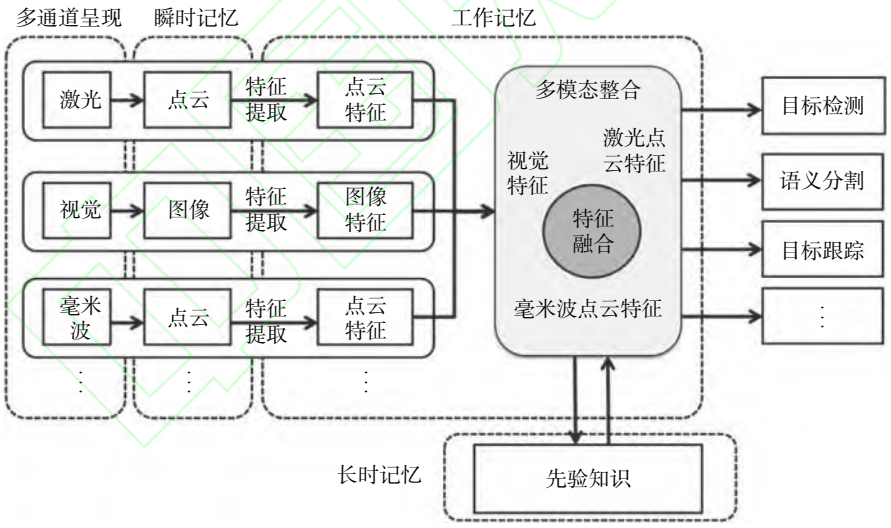

图 3　多模态认知融合框架

Fig. 3　Multimodal cognitive fusion framework

此外，根据最近的研究趋势，针对多任务学习的集成模型取得较好的效果[23, 56, 62]，这表明不同任务所挖掘的数据信息在一定程度上适用于其他任务。然而两个问题可能成为集成模型设计的限制，一是自动驾驶对模型复杂度和实时性的要求，使得模型不应当过于复杂；二是模型鲁棒性的限制，根据上述融合的合理性分析，复杂的模型需要考虑部分模态数据异常的情形，因此复杂的模型需要保证更高的鲁棒性。

4　结束语

数据融合是自动驾驶的感知任务的重要趋势，为此本文提供了对自动驾驶场景下目标检测中的多模态数据融合方法的综述。本文介绍了目标检测和数据融合的背景，并按融合层次、融合计算方法两个方面总结了当前的研究。我们认为，传统的“前中后”的划分方法对集成模型中的融合方法区分度不足，为此本文采用了新的层次定义方法并给出了明确的定义，从而能够帮助研

究人员更好地理解融合方法设计的动机和作用。此外，本文提出了数据融合的合理性分析并总结了现有的研究，还讨论了当前融合方法的挑战与策略，并提出了若干公开问题。

尽管新的融合方法不断被提出，然而现在依旧缺乏对此的理论分析和深入的对比实验。自动驾驶技术的发展依赖于高效、鲁棒的环境感知，相应的数据融合方法应当尽快被开发以保证车辆感知技术的性能。我们后续的工作包括对数据表现形式的研究和跨模态数据融合的信息增益度量，同时期待新的研究工作可以解决本文所讨论的融合方法中的挑战。

参考文献：

[1] URMSON C, ANHALT J, BAGNELL D, et al. Autonomous driving in urban environments: boss and the urban challenge[J]. Journal of field robotics, 2008, 25(8): 425–466.

[2] EVERINGHAM M, VANGOOL L, WILLIAMS C K I, et al. The PASCAL visual object classes challenge 2007 Results[EB/OL]. http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html

[3] HILLEL A B, LERNER R, LEVI D, et al. Recent progress in road and lane detection: a survey[J]. Machine vision applications, 2014, 25(3): 727–745.

[4] FENG D, HAASE-SCHUETZ C, ROSENBAUM L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges[J]. arXiv preprint arXiv: 1902.07830, 2019.

[5] 罗俊海, 杨阳. 基于数据融合的目标检测方法综述 [J]. 控制与决策, 2020, 35(1): 1–15.
LUO Junhai, YANG Yang. An overview of target detection methods based on data fusion[J]. Control and decision, 2020, 35(1): 1–15.

[6] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, et al. Simultaneous detection and segmentation[C]//European Conference on Computer vision. Zurich, Switzerland, 2014: 297−312.

[7] KANG K, LI H, YAN J, et al. T-CNN: tubelets with convolutional neural networks for object detection from videos[J]. IEEE transactions on circuits and systems for video technology, 2018, 28(10): 2896–2907.

[8] ZOU Z, SHI Z, GUO Y, et al. Object detection in 20 years: a survey[J]. arXiv preprint arXiv: 1905.05055, 2019.

[9] 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.

[10] MEES O, EITEL A, BURGARD W. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Daejeon, South Korea, 2016: 151−156.

[11] EITEL A, SPRINGENBERG J T, SPINELLO L, et al. Multimodal deep learning for robust RGB-D object recognition[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Hamburg, Germany, 2015: 681−687.

[12] YU J, JIANG Y, WANG Z, et al. UnitBox: an advanced object detection network[C]//ACM International Conference on Multimedia. Amsterdam, Netherlands, 2016: 516−520.

[13] REZATOFIGHI H, TSOI N, GWAK J, at al. Generalized intersection over union: a metric and a loss for bounding box regression[C]//IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 658−666.

[14] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J]. arXiv preprint arXiv: 1804.02767, 2018.

[15] REN S, HE K, GIRSHICK R, et al. Faster RCNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 91−99.

[16] YANG W, ZHANG X, TIAN Y, at al. Deep learning for single image super-resolution: a brief review[J]. IEEE transactions on multimedia, 2019, 21(12): 3106–3121.

[17] LU Y, LU C, TANG C K. Online video object detection using association LSTM[C]//IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2363−2371.

[18] WANG S, ZHOU Y, YAN J, at al. Fully motion-aware network for video object detection[C]//The European Conference on Computer Vision. Munich, Germany, 2018: 557−573.

[19] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//IEEE Conference on Computer Vision and Pattern Recognition. RI, USA, 2012: 3354−3361.

[20] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]//IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 2446-2454.

[21] CAESAR H, BANKITI V, LANG A H, et al. NuScenes: A multimodal dataset for autonomous driving[C]. //IEEE Conference on Computer Vision and Pattern Recognition.

Virtual, 2020: 2446-2454.

[22] HUANG X, CHENG X, GENG Q, et al. The Apolloscape dataset for autonomous driving[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 954−960.

[23] CHEN X, MA H, WAN J, et al. Multi-view 3D object detection network for autonomous driving[C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1907−1915.

[24] MUR-ARTAL R, TARDOS J D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. IEEE transactions on robotics, 2017, 33(5): 1255–1262.

[25] CHADWICK S, MADDERN W, NEWMAN P. Distant vehicle detection using radar and vision[C]//International Conference on Robotics and Automation. Montreal, Canada, 2019: 8311−8317.

[26] BIJELIC M, GRUBER T. Seeing through fog without seeing fog: deep sensor fusion in the absence of labeled training data[C]//IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 11621-11631.

[27] DU X, ANG M H, KARAMAN S, et al. A general pipeline for 3D detection of vehicles[C]//IEEE international Conference on Robotics and Automation. Brisbane, Australia, 2018: 3194−3200.

[28] BANERJEE K, NOTZ D, WINDELEN J, et al. Online camera lidar fusion and object detection on hybrid data for autonomous driving[C]//IEEE Intelligent Vehicles Symposium. Changshu, China, 2018: 1632−1638.

[29] LIU J, ZHANG S, WANG S, et al. Multispectral deep neural networks for pedestrian detection[C]//British Machine Vision Conference. York, UK, 2016: 1−13.

[30] FISCHER V, HERMAN M, BEHNKE S. Multispectral pedestrian detection using deep fusion convolutional neural networks[C]//European Symposium on artificial Neural Networks. Bruges, Belgium, 2016: 27−29.

[31] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]//IEEE Conference on Computer Vision and Pattern Recognition. Kauai, USA, 2001: 511−518.

[32] VIOLA P, JONES M. Robust real-time face detection[J]. International journal of computer vision, 2004, 57: 137–154.

[33] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 886−893.

[34] FELZENSZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model[C]//IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1−8.

[35] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Advances in neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097−1105.

[36] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 580−587.

[37] GIRSHICK R. Fast R-CNN[C]//IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1440−1448

[38] GIRSHICK R, DONAHUE J, DARRELL T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(1): 142–158.

[39] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2117−2125.

[40] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 779−788

[41] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 21−37.

[42] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 7263−7271.

[43] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE international conference on computer vision, 2017: 2980–2988.

[44] MEYER G P, LADDHA A, KEE E, et al. LaserNet: an efficient probabilistic 3D object detector for autonomous driving[C]//IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 12677−12686.

[45] QI C R, SU H, MO K, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 652−660.

[46] QI C R, YI L, SU H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[C]//Advances in Neural Information Processing Systems. Long

Beach, USA, 2017: 5099−5108.

[47] YANG B, LUO W, URTASUN R. PIXOR: real-time 3D object detection from point clouds[C]//IEEE Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7652−7660.

[48] ASVADI A, GARROTE L, PREMEBIDA C, et al. Multimodal vehicle detection: fusing 3D LIDAR and color camera data[J]. Pattern recognition letters, 2018, 115: 20–29.

[49] SCHLOSSER J, CHOW C K, KIRA Z. Fusing LIDAR and images for pedestrian detection using convolutional neural networks[C]//IEEE International Conference on Robotics and Automation. Stockholm, Sweden, 2016: 2198−2205.

[50] QI C R, GUIBAS L J. Frustum PointNets for 3D object detection from RGB-D data[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 918−927.

[51] GUAN D, CAO Y, YANG J, et al. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection[J]. Information fusion, 2018, 50: 148–157.

[52] YANG B, LIANG M, URTASUN R, at al. HDNET: exploiting HD maps for 3D object detection[J]. Proceedings of machine learning research, 2018, 87: 146–155.

[53] ZHOU T, JIANG K, XIAO Z, et al. Object detection using multi-sensor fusion based on deep learning[C]//COTA International Conference of Transportation. Nanjing, China, 2019: 5770−5782.

[54] CHO H, SEO Y W, KUMAR B. A multi-sensor fusion system for moving object detection and tracking in urban driving environments[C]//IEEE International Conference on Robotics and Automation. Hong Kong, China, 2014: 1836−1843.

[55] DOU J, XUE J, FANG J. SEG-VoxelNet for 3D vehicle detection from RGB and lidar data[C]//International Conference on Robotics and Automation. Montreal, Canada, 2019: 4362−4368.

[56] LIANG M, YANG B, CHEN Y, et al. Multi-task multi-sensor fusion for 3D object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7345−7353.

[57] SINDAGI V A, ZHOU Y, TUZEL O. MVX-Net: multimodal VoxelNet for 3D object detection[C]//2019 International Conference on Robotics and Automation. Montreal, Canada, 2019: 7276−7282.

[58] LIANG M, YANG B, WANG S, et al. Deep continuous fusion for multi-sensor 3D object detection[C]//The European Conference on Computer Vision. Munich, Germany, 2018: 641−656.

[59] WANG Z, ZHAN W, TOMIZUKA M. Fusing bird's eye view LIDAR point cloud and front view camera image for deep object detection[C]//IEEE Intelligent Vehicles Symposium. Changshu, China, 2018: 1−6.

[60] KIM J, KOH J, KIM Y, et al. Robust deep multi-modal learning based on gated information fusion network[C]//Asian Conference on Computer Vision. Perth, Australia, 2018: 90−106.

[61] CASAS S, LUO W, URTASUN R. IntentNet: learning to predict intention from raw sensor data[J]. Proceedings of machine learning research, 2018, 87: 947–956.

[62] KU J, MOZIFIAN M, LEE J, et al. Joint 3D proposal generation and object detection from view aggregation[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain, 2018: 5750−5757.

[63] PFEUFFER A, DIETMAYER K. Optimal sensor data fusion architecture for object detection in adverse weather conditions[C]//International Conference on Information Fusion. Cambridge, UK, 2018: 1-8.

[64] XU D, ANGUELOV D, JAIN A. PointFusion: deep sensor fusion for 3D bounding box estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 2018: 244−253.

[65] DU X, ANG M H, RUS D. Car detection for autonomous vehicle: lidar and vision fusion approach through deep learning framework[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, BC, Canada, 2017: 749−754.

[66] MATTI D, EKENEL H K, THIRAN J. Combining LiDAR space clustering and convolutional neural networks for pedestrian detection[C]//IEEE International Conference on Advanced Video and Signal Based Surveillance. Lecce, Italy, 2017: 1−6.

[67] SCHNEIDER L, JASCH M. Multimodal neural networks: RGB-D for semantic segmentation and object detection[C]//Scandinavian Conference on Image Analysis. Norrköping, Sweden, 2017: 98−109.

[68] OH S, KANG H. Object detection and classification by decision-level fusion for intelligent vehicle systems[J]. Sensors (Basel), 2017, 17(1): 207–214.

[69] KIM T, GHOSH J. Robust detection of nonmotorized road users using deep learning on optical and lidar data[C]//IEEE Iinternational Conference on Intelligent Transportation Systems. Rio de Janeiro, Brazil, 2016: 271−276.

[70] BAI M, MATTYUS G, HOMAYOUNFAR N, et al. Deep

multi-sensor lane detection[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain, 2018: 3102−3109.

[71] CALTAGIRONE L, BELLONE M, SVENSSON L, et al. LIDAR–camera fusion for road detection using fully convolutional neural networks[J]. Robotics and autonomous systems, 2019, 111: 125–131.

[72] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[73] ZHANG X, ZHOU M, LIU H, et al. A cognitively-inspired system architecture for the Mengshi cognitive vehicle[J]. Cognitive computation, 2020, 12(1): 140–149.

[74] ZOU Q, JIANG H, DAI Q, et al. Robust lane detection from continuous driving scenes using deep neural networks[J]. IEEE transactions on vehicular technology, 2020, 69(1): 41–54.

[75] ZHANG X, GAO H, GUO M, et al. A study on key technologies of unmanned driving[J]. CAAI transactions on intelligence technology, 2016, 1(1): 4–13.

作者简介：

张新钰，研究员，清华猛狮智能车团队负责人，剑桥大学访问学者，主要研究方向为智能驾驶和多模态信息融合。担任国家重点研发计划项目负责人。多次在国内无人驾驶顶级赛事获得冠亚军，获 2019 年吴文俊人工智能科技进步二等奖，发表智能驾驶领域的 SCI/EI 检索 30 篇，入选 ESI 高被引论文 1 篇。

刘华平，副教授，博士生导师，中国指挥与控制学会青年工作委员会副主任，中国人工智能学会理事，中国人工智能学会认知系统与信息处理专业委员会秘书长，IEEE 高级会员，主要研究方向为智能机器人的多模态感知、学习与控制技术。国家杰出青年基金获得者，获中国指挥与控制学会曙光创新奖和国家高技术研究发展计划 (863 计划)“十二五”科技攻关“青年创新之星”，以及 IEEE 仪器与测量协会 (IMS) 颁发的 Andy Chi Best Paper Award 。