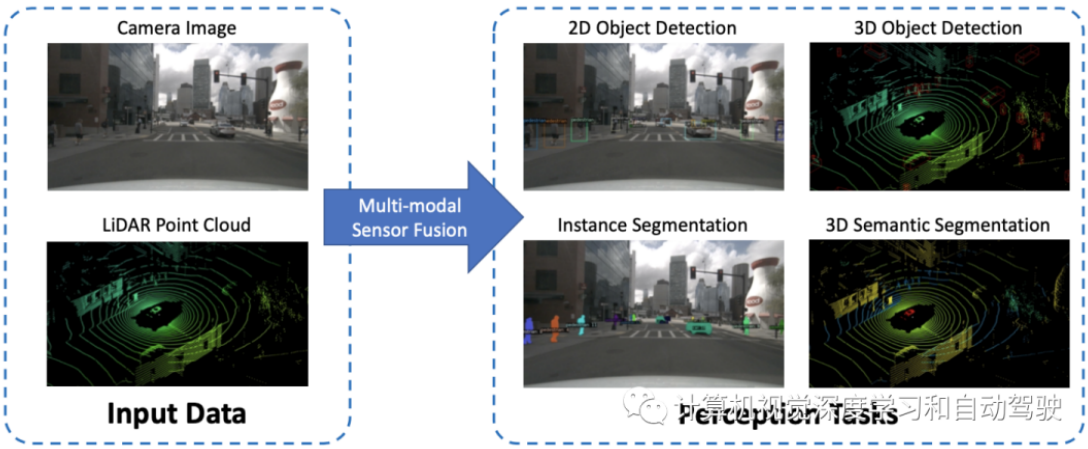
一文读懂自动驾驶多模态传感器融合

焉知智能汽车 2022-04-24 20:14

多模态融合是感知自动驾驶系统的一项基本任务，最近引起了许多研究人员的兴趣。然而，由于原始数据噪声大、信息利用率低以及多模态传感器的无对准，达到相当好的性能并非易事。本文对现有的基于多模态自动驾驶感知任务方法进行了文献综述。分析超过50篇论文，包括摄像头和激光雷达，试图解决目标检测和语义分割任务。与传统的融合模型分类方法不同，作者从融合阶段的角度，通过更合理的分类法将融合模型分为两大类，四小类。此外，研究了当前的融合方法，就潜在的研究机会展开讨论。

最近，用于自动驾驶感知任务的多模态融合方法发展迅速，其从跨模态特征表示和更可靠的模态传感器，到更复杂、更稳健的多模态融合深度学习模型和技术。然而，只有少数文献综述集中在多模态融合方法本身的方法论上，大多数文献都遵循传统规则，将其分为前融合、深度（特征）融合和后融合三大类，重点关注深度学习模型中融合特征的阶段，无论是数据级、特征级还是提议级。首先，这种分类法没有明确定义每个级别的特征表示。其次，它表明，激光雷达和摄像头这两个分支在处理过程中始终是对称的，模糊了激光雷达分支中融合提议级特征和摄像头分支中融合数据级特征的情况。综上所述，传统的分类法可能是直观的，但对于总结最近出现的越来越多的多模态融合方法来说却很落后，这使得研究人员无法从系统的角度对其进行研究和分析。

如图是自动驾驶感知任务的示意图：

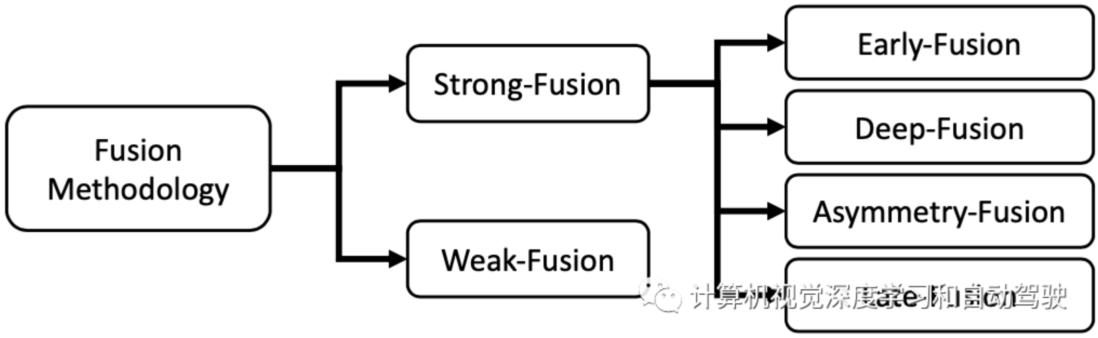


深度学习模型仅限于输入的表示。为了实现该模型，需要在数据输入模型之前，通过一个复杂的特征提取器对原始数据进行预处理。

至于图像分支，大多数现有方法保持与下游模块输入的原始数据相同的格式。然而，激光雷达分支高度依赖于数据格式，这种格式强调不同的特性，并对下游模型设计产生巨大影响。因此，这里将其总结为基于点、基于体素和基于二维映射的点云数据格式，以适应异构深度学习模型。

数据级融合或前融合方法，通过空间对齐直接融合不同模式的原始传感器数据。特征级融合或深度融合方法通过级联或元素相乘在特征空间中混合跨模态数据。目标级融合方法将各模态模型的预测结果结合起来，做出最终决策。

一种新的分类法，将所有融合方法分为强融合和弱融合，如图展示了二者之间的关系：



为性能比较，KITTI benchmark的3D检测和鸟瞰目标检测。如下两个表分别给出BEV和3D的KITTI测试数据集上多模态融合方法的实验结果。

Table 2. Survey of BEV Task Results in KITTI [26] for Test Dataset

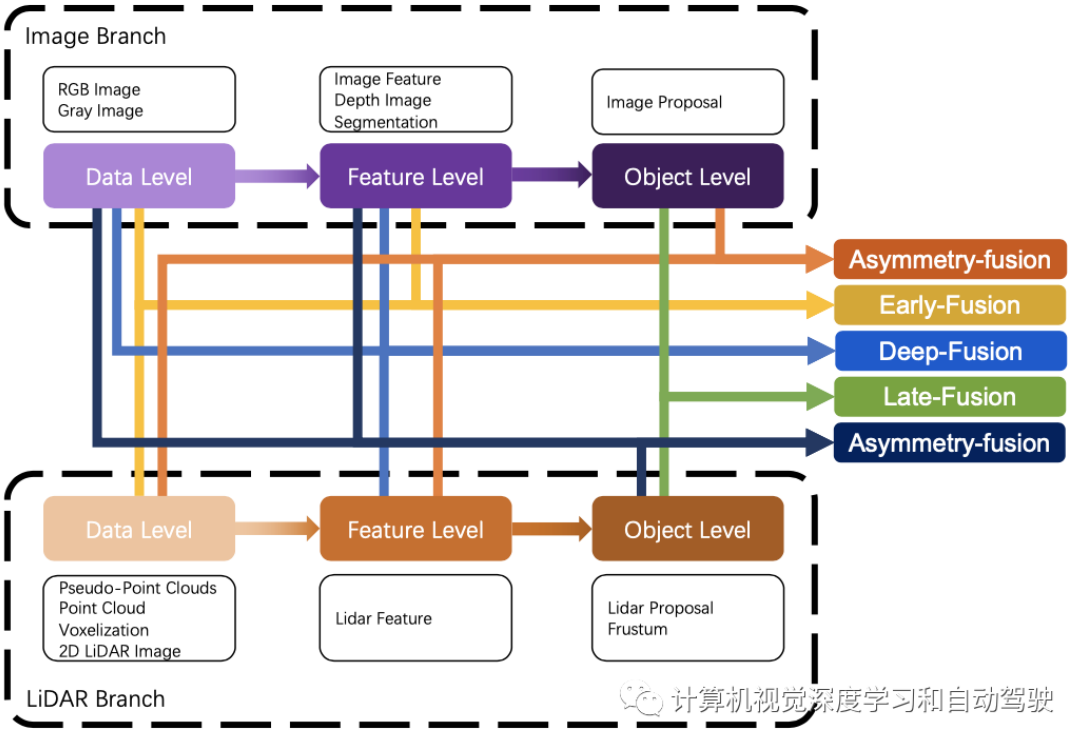
Method	Year	Car			Pedestrian			Cyclist		
		Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
Early-Fusion										
PFF3D [87]	2021	89.61	85.08	80.42	48.74	40.94	38.54	72.67	55.71	49.58
Painted PointRCNN [76]	2020	92.45	88.11	83.36	58.70	49.93	46.29	83.91	71.54	62.97
PI-RCNN [90]	2020	91.44	85.81	81.00	-	-	-	-	-	-
Complexer-YOLO [68]	2019	77.24	68.96	64.95	21.42	18.26	17.06	32.00	25.43	22.88
MVX-Net(PF) [69]	2019	89.20	85.90	78.10	-	-	-	-	-	-
Deep-Fusion										
RoIFusion [9]	2021	92.88	89.03	83.94	46.21	38.08	35.97	83.13	67.71	61.70
EPNet [32]	2020	94.22	88.47	83.69	-	-	-	-	-	-
MAFF-Net [105]	2020	90.79	87.34	77.66	-	-	-	-	-	-
SemanticVoxels [22]	2020	-	-	-	58.91	49.93	47.31	-	-	-
MVAF-Net [78]	2020	91.95	87.73	85.00	-	-	-	-	-	-
3D-CVF [102]	2020	93.52	89.56	82.45	-	-	-	-	-	-
MMF [45]	2019	93.67	88.21	81.99	-	-	-	-	-	-
ContFuse [46]	2018	94.07	85.35	75.88	-	-	-	-	-	-
SparsePool [85]	2017	-	-	-	43.33	34.15	31.78	43.55	35.24	30.15
Late-Fusion										
CLOCs [55]	2020	92.91	89.48	86.42	-	-	-	-	-	-
Asymmetry-Fusion										
VMVS [40]	2019	-	-	-	60.34	50.34	46.45	-	-	-
MLOD [17]	2019	90.25	82.68	77.97	55.09	45.40	41.42	73.03	55.06	48.21
MV3D [12]	2017	86.62	78.93	69.80	-	-	-	-	-	-
Weak-Fusion										
Faraway-Frustum [104]	2020	91.90	88.08	85.35	52.15	43.85	41.68	79.65	64.54	57.84
F-ConvNet [83]	2019	91.51	85.84	76.11	57.04	48.96	44.33	84.16	68.88	60.05
IPOD [99]	2018	86.93	83.98	77.85	60.83	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
F-PointNet [60]	2017	91.17	84.67	74.77	57.13	49.57	45.48	77.26	61.37	53.78

Table 3. Survey of 3D Task Results in KITTI [26] for Test Dataset

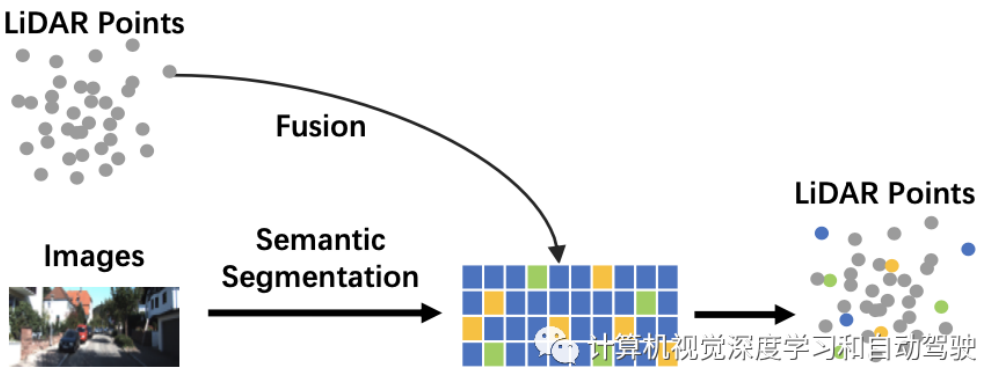
Method	Year	Car			Pedestrian			Cyclist		
		Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
Early-Fusion										
PFF3D [87]	2021	81.11	72.93	67.24	43.93	36.07	32.86	63.27	46.78	41.37
Painted PointRCNN [76]	2020	82.11	71.70	67.08	50.32	40.97	37.87	77.63	63.78	55.89
PI-RCNN [90]	2020	84.37	74.82	70.03	-	-	-	-	-	-
Complexer-YOLO [68]	2019	55.93	47.34	42.60	17.60	13.96	12.70	24.27	18.53	17.31
MVX-Net(PF) [69]	2019	83.20	72.70	65.20	-	-	-	-	-	-
Deep-Fusion										
RoIFusion [9]	2021	88.09	79.36	72.51	42.22	35.14	32.92	80.84	64.05	58.37
EPNet [32]	2020	89.81	79.28	74.59	-	-	-	-	-	-
MAFF-Net [105]	2020	85.52	75.04	67.61	-	-	-	-	-	-
SemanticVoxels [22]	2020	-	-	-	50.90	42.19	39.52	-	-	-
MVAF-Net [78]	2020	87.87	78.71	75.48	-	-	-	-	-	-
3D-CVF [102]	2020	89.20	80.05	73.11	-	-	-	-	-	-
MMF [45]	2019	88.40	77.43	70.22	-	-	-	-	-	-
SCANet [49]	2019	76.09	66.30	58.68	-	-	-	-	-	-
ContFuse [46]	2018	83.68	68.78	61.67	-	-	-	-	-	-
PointFusion [92]	2018	77.92	63.00	53.27	-	-	-	-	-	-
SparsePool [85]	2018	-	-	-	37.84	30.38	26.94	40.87	32.61	29.05
Late-Fusion										
CLOCs [55]	2020	89.16	82.28	77.23	-	-	-	-	-	-
Asymmetry-Fusion										
VMVS [40]	2019	-	-	-	53.44	43.27	39.51	-	-	-
MLOD [17]	2019	77.24	67.76	62.05	47.58	37.47	35.07	68.81	49.43	42.84
MV3D [12]	2017	74.97	63.63	54.00	-	-	-	-	-	-
Weak-Fusion										
Faraway-Frustum [104]	2020	87.45	79.05	76.14	46.33	38.58	35.71	77.36	62.00	55.40
F-ConvNet [83]	2019	87.36	76.39	66.69	52.16	43.38	38.80	81.98	65.07	56.54
IPOD [99]	2018	79.75	72.57	66.33	56.92	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
F-PointNet [60]	2018	82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01

根据激光雷达和摄像头数据表示的不同组合阶段，将强融合再分为前融合、深度融合、后融合和非对称融合四类。作为研究最多的融合方法，强融合近年来取得了许多杰出的成就。

如图所示：强融合的每个小类都高度依赖于激光雷达点云，而不是摄像头数据。



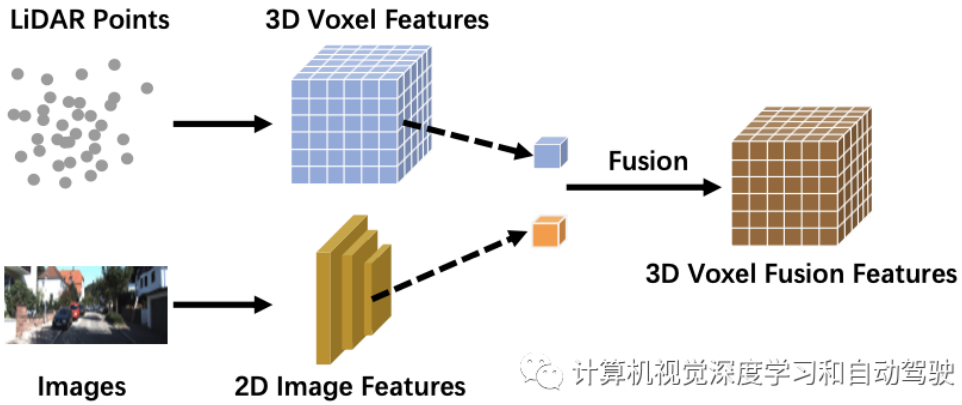
前融合。数据级融合是一种通过原始数据级的空间对齐和投影直接融合每个模态数据的方法，与之不同的是，前融合在数据级是融合激光雷达数据，在数据级或特征级则融合摄像头数据。一个例子如图所示：



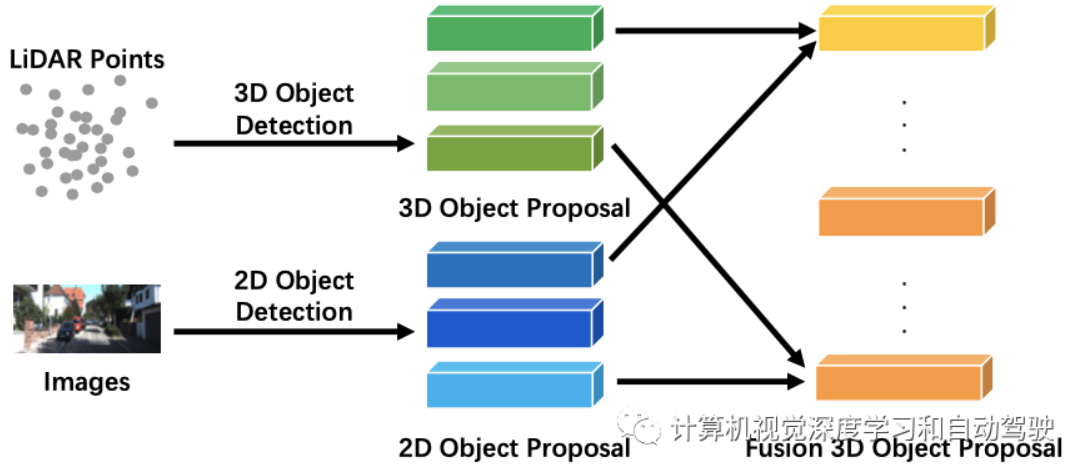
在激光雷达分支，点云可以是有反射图、体素化张量、前视图/距离视图/鸟瞰视图以及伪点云等形式。尽管所有这些数据都具有不同的内在特征，与激光雷达主干网高相关，但除了伪点云之外，大多数据通过基于规则的处理生成。此外，与特征空间嵌入相比，该阶段的数据仍然具有可解释性，因此所有这些激光雷达数据表示都直观可视。

对于图像分支，严格的数据级定义应该只包含RGB或灰度等数据，缺乏通用性和合理性。与前融合的传统定义相比，摄像头数据放松为数据级和特征级数据。特别是，这里将有利于三维目标检测的图像语义分割任务结果作为特征级表示，因为这些“目标级”特征与整个任务的最终目标级提议不同。

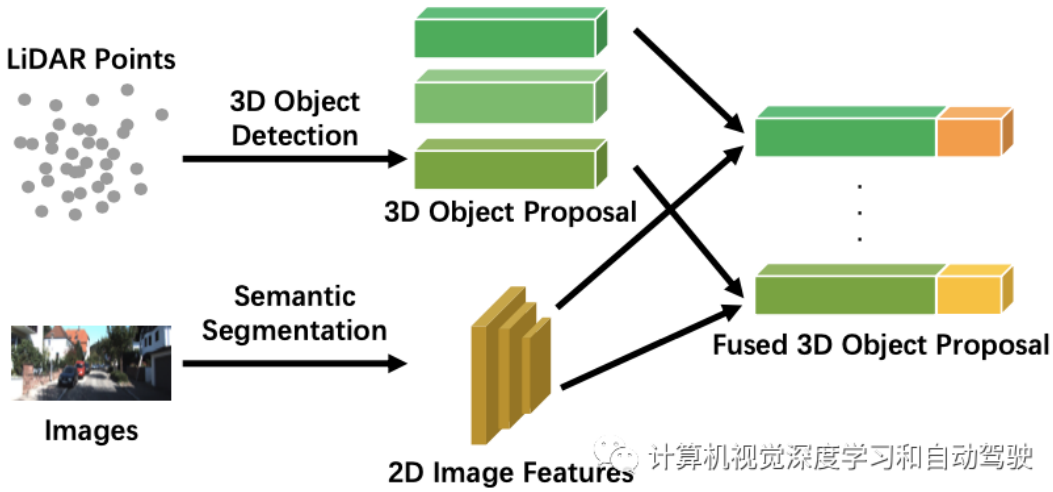
深度融合。深度融合方法在激光雷达分支的特征级对跨模态数据融合，但在图像分支的数据级和特征级做融合。例如，一些方法使用特征提取器分别获取激光雷达点云和摄像头图像的嵌入表示，并通过一系列下游模块将特征融合到两种模式中。然而，与其他强融合方法不同，深度融合有时以级联方式融合特征，这两种方法都利用原始和高级语义信息。深度融合的一个例子如图所示：



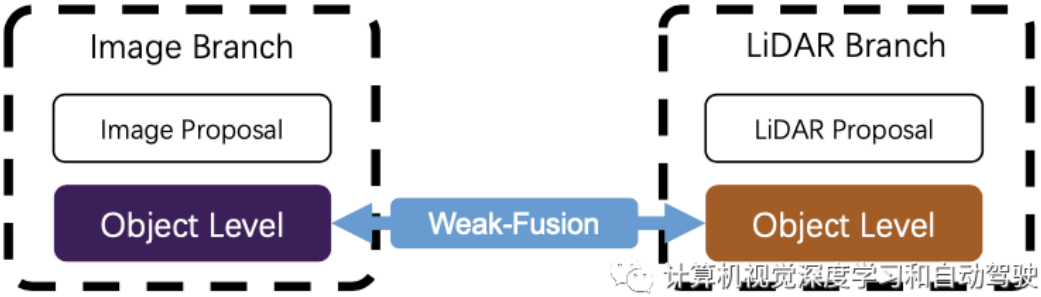
后融合。后融合，也称为目标级融合，指的是融合每个模态中流水线结果的方法。例如，一些后融合方法利激光雷达点云分支和摄像头图像分支的输出，并基于两种模式的结果进行最终预测。请注意，两个分支提议数据格式应与最终结果相同，但在质量、数量和精度上有所不同。后融合是一种多模态信息优化最终提议的成方法（ensemble method）。如图是后融合的一个例子：



非对称融合。除了早融合、深度融合和后融合外，一些方法以不同的权限处理跨模态分支，因此融合一个分支的目标级信息和其他分支的数据级或特征级信息，定义为非对称融合。强融合的其他方法将两个分支视为似乎相等的状态，非对称融合至少有一个分支占主导地位，而其他分支提供辅助信息来执行最终任务。如图是非对称融合的一个例子：可能具有提议的相同提取特征，但非对称融合只有来自一个分支的一个提议，而后融合有来自所有分支的提议。



与强融合不同，弱融合方法不会以多种方式直接从分支融合数据/特征/目标，而是以其他方式操作数据。基于弱融合的方法通常使用基于规则的方法来利用一种模态数据作为监督信号，以指导另一模态的交互。如图展示了弱融合模式的基本框架：



有可能图像分支中CNN的2D提议导致原始激光雷达点云出现截锥体（frustum）。然而，与图像特征组合非对称融合不同，弱融合直接将选择的原始激光雷达点云输入到激光雷达主干网，以输出最终提议。

有些工作不能简单地定义为上述任何类型的融合，在整个模型框架中采用多种融合方法，例如深度融合和后合的结合，也有将前融合和深度融合结合在一起。这些方法从模型设计看存在冗余，这不是融合模块的主流。

待解决的问题有一些分析。

当前的融合模型面临着错对齐和信息丢失的问题。此外，平融合（flat fusion）操作也阻止了感知任务性能的进一步提高。总结一下：

- **错对齐和信息丢失：**传统的前融合和深度融合方法利用外部标定矩阵将所有激光雷达点直接投影到相应的像素，反之亦然。然而，由于传感器噪声，这种逐像素对齐不够精确。因此，可以采取周围的信息作为补充，会产生更好的性能。此外，在输入和特征空间的转换过程中，还存在其他一些信息损失。通常，降维操作的投影不可避免地会导致大量信息丢失，例如，将3-D激光雷达点云映射到2-DBEV图像。将两个模态数据映射到另一个专门为融合设计的高维表示，可以有效地利用原始数据，减少信息损失。

- **更合理的融合操作：**级联和元素相乘这些简单的操作可能无法融合分布差异较大的数据，难以弥合两个模态之间的语义鸿沟。一些工作试图用更复杂的级联结构来融合数据并提高性能。

前视图单帧图像是自动驾驶感知任务的典型场景。然而，大多数框架利用有限的信息，没有详细设计辅助任务来进一步理解驾驶场景。总结一下：

- **采用更多的潜在信息：**现有方法缺乏对多维度和来源信息的有效利用。其中大多数都集中在前视图的单帧多模态数据上。其他有意义的信息还有语义、空间和场景上下文信息。一些模型试图用图像语义分割任务结果作为附加特征，而其他模型可能利用神经网络主干中间层的特征。在自动驾驶场景中，许多明确语义信息的下游任务可能会极大地提高目标检测任务的性能。例如车道检测、语义分割。因此，未来的研究可以通过各种下游任务（如检测车道、交通灯和标志）共同构建一个完整的城市场景的认知框架，帮助感知任务的表现。此外，当前的感知任务主要依赖于忽略时间信息的单一框架。最近基于激光雷达的方法结合了一个帧序列来提高性能。时间序列信息包含序列化的监控信号，与单帧方法相比，它可以提供更稳健的结果。

- **表征学习的自监督：**相互监督的信号自然地存在于从同一个真实世界场景但不同角度采样的跨模态数据中。然而，由于缺乏对数据的深入理解，目前无法挖掘出各模态之间的协同关系。未来的研究可以集中在如何利用多模态数据进行自监督学习，包括预训练、微调或对比学习。通过实施这些最先进的机制，融合模型将加深对数据的理解并取得更好的结果。

域偏差和数据分辨率与真实场景和传感器高相关。这些缺陷阻碍了自动驾驶深度学习模型的大规模训练和实施

- **域偏差：**在自主驾驶感知场景中，由不同传感器提取的原始数据伴随着域相关特征。不同的摄像头系统有其光学特性，而激光雷达可能因机械激光雷达和固态激光雷达而不同。更重要的是，数据本身可能是有域偏差的，例如天气、季节或地理位置。因此，检测模型无法顺利适应新的场景。由于泛化失败，这些缺陷妨碍大规模数据集的收集和原始训练数据可重用性。

- **分辨率冲突：**来自不同模式的传感器通常具有不同的分辨率。例如，激光雷达的空域密度明显低于图像的空域密度。无论采用何种投影方法，由于无法找到对应关系，一些信息被消除。这可能导致模型被一个特定模态的数据所主导，无论是特征向量的分辨率不同还是原始信息的不平衡。