摘 要

注意力机制在点云和图像融合目标检测中的研究

目前，单一传感器在感知任务中存在局限性，为了提高感知任务的准确性，许多工作采用了多传感器融合方法。在三维场景感知技术迅速发展的背景下，多模态融合在三维目标检测中已广泛应用。然而，当前的多传感器融合方法存在以下问题：对于多传感器信息的利用效率较低，难以有效解决复杂场景下成像分辨率低以及部分物体被遮挡的检测问题。此外，多模态融合需要考虑多个传感器的数据，算法鲁棒性差，容易受到传感器故障、数据缺失等因素影响。因此，当前的多传感器融合方法仍需要进一步改进，以提高感知任务的准确性和鲁棒性。

在深度学习技术中，注意力机制通过实现对不同特征的自适应选择和加权提升深度网络表征、分析和理解数据的能力。针对多模态融合中存在的问题，本文主要研究在点云和图像信息融合的目标检测任务中，注意力机制对检测结果的影响，并验证了本文算法的有效性。具体的工作内容如下：

1）基于注意力机制的单模态检测算法研究。该部分，研究单模态数据实现目标检测任务。实验采用编码-解码结构，在编码器和解码器之间插入局部-全局注意力机制模块，以获取更丰富的全局上下文信息。局部-全局注意力机制模块由局部模块、全局注意力机制模块及跳跃连接结构组成。实验结果表明，本文提出的检测算法在图像和点云单模态下都能有效地提升检测效果。

2）基于注意力机制的多模态检测算法研究。该部分，研究采用包含多种传感器数据的 NuScense 数据集。在基于第一部分的研究基础上，实验首先采用初始化目标查询机制，将提取的图像特征作为引导信息得到 Query；接着在编码-解码结构中添加交叉注意力机制融合图像特征和点云特征；最后通过两个解码器层分别预测候选框和输出目标检测结果。其中，交叉注意力机制模块分别将点云特征和图像特征的 Query 交叉计算得到融合特征。检测头中的每个解码器层之后添加前馈神经网络和监督，利用预测候选框限制交叉注意力。交叉注意力机制可以对不同传感器特征图之间进行建模，能够充分

利用特征图中的语义信息。实验结果表明，相同场景下，多模态比单模态的目标检测指标 mAP 提升了 4.1%；不同场景下，本文的多模态融合方法的网络检测性能更加稳定。

关键词：

点云，注意力机制，多模态融合，目标检测

Abstract

Attention Mechanism in the Study of Fusion Target Detection of Point Cloud and Image

Currently, single-sensor perception tasks have limitations. To improve the accuracy of perception tasks, many works have adopted multi-sensor fusion methods. With the rapid development of 3D scene perception technology, multimodal fusion has been widely applied in 3D object detection. However, the current multi-sensor fusion methods have the following problems: the utilization efficiency of multi-sensor information is low, making it difficult to effectively address the detection problems of low imaging resolution and partially occluded objects in complex scenes. In addition, multimodal fusion needs to consider data from multiple sensors, making the algorithms less robust and easily affected by factors such as sensor failure and data loss. Therefore, current multi-sensor fusion methods still need further improvement to enhance the accuracy and robustness of perception tasks.

In deep learning technologies, attention mechanisms improve the ability of deep networks to represent, analyze, and understand data by adaptively selecting and weighting different features. This thesis mainly investigates the impact of attention mechanisms on the detection results in the target detection task of point cloud and image information fusion, and verifies the effectiveness of the proposed algorithm. The specific work content is as follows:

1) Research on single-modal detection algorithms based on attention mechanisms. In this part, single-modal data is used to perform object detection tasks. The experiment adopts an encoder-decoder structure and inserts a local-global attention mechanism module between the encoder and decoder to obtain richer global context information. The local-global attention mechanism module consists of a local module, a global attention mechanism module, and a skip connection structure. The experimental results show that the proposed detection algorithm can effectively improve the detection effect under both image and point cloud single-modal

scenarios.

2) Research on multimodal detection algorithms based on attention mechanisms. In this part, the NuScense dataset containing data from multiple sensors is used. Based on the research foundation of the first part, the experiment first adopts an initialization target query mechanism, using the extracted image features as guiding information to obtain the Query. Then, a cross-attention mechanism is added to the encoder-decoder structure to fuse image features and point cloud features. Finally, candidate boxes are predicted and target detection results are output through two decoder layers. The cross-attention mechanism module separately computes the fusion features of the point cloud features and image feature queries. A feedforward neural network and supervision are added after each decoder layer in the detection head, using predicted candidate boxes to constrain cross-attention. The cross-attention mechanism can model the relationships between feature maps from different sensors, making full use of semantic information in the feature maps. Experimental results show that in the same scene, the multimodal object detection metric mAP increased by 4.1% compared with single-modal; in different scenarios, the multimodal fusion method proposed in this thesis exhibits more stable network detection performance.

Keywords:

Point Cloud, Attention Mechanism, Multi-modal Fusion, Object Detection

目　　录

第 1 章 绪论

1.1 研究背景及意义

计算机视觉是人工智能技术中一个备受关注的研究领域。在面对当今海量数据的问题时，计算机视觉技术需要发展出更快速、更有效的方法。数据驱动的深度学习技术为这个问题提供了一个有效的解决方案。深度神经网络进行数据特征的映射和逐层处理，为不同特征层之间建立联系，从而更加准确地进行决策和分析。目标检测在工业领域得到了广泛应用，是计算机视觉领域研究的重要任务之一，如图 1.1 所示，该任务是自动驾驶感知系统中重要的研究内容。

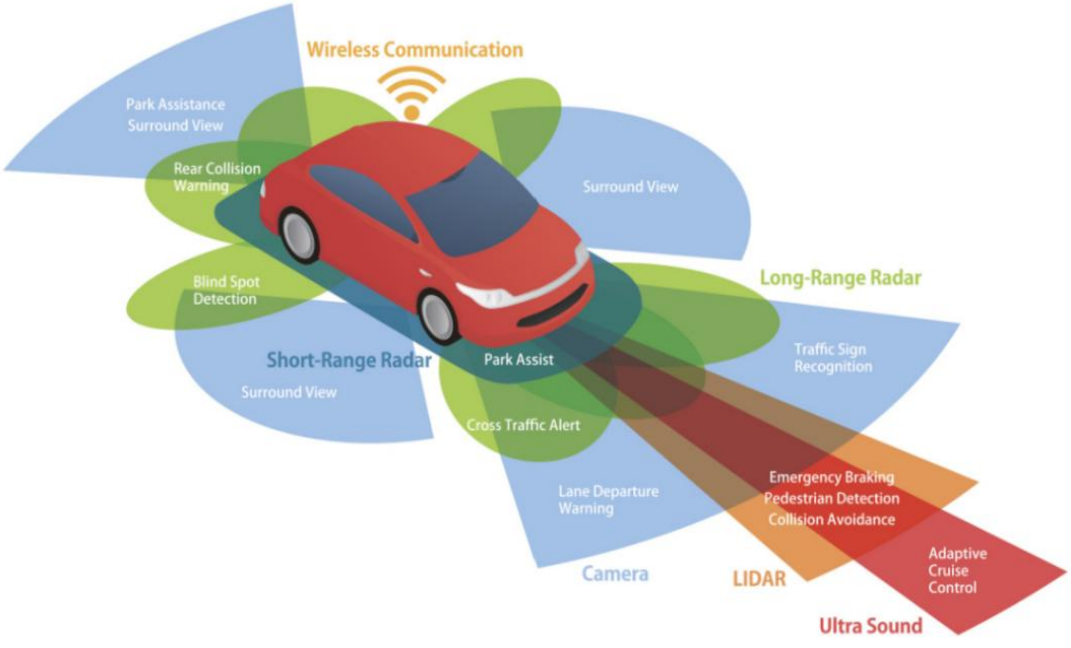

图 1.1 自动驾驶感知系统图

2015 年特斯拉公司发布了历史上第一个商用自动辅助驾驶系统 Autopilot；同年，百度公司在长沙开设了无人驾驶出租车 RoboTaxi 试点；2018 年，百度推出全球首款 Apollo 系统的阿波龙巴士，在 2022 年正式开放运营。自动驾驶感知系统需要对车辆周围环境中的静态和动态目标进行检测和追踪，以做出规划和决策[1]。深度学习在图像识别、图像分割以及目标检测等任务中取得了巨大进步。图像具有丰富的颜色和纹理信息，但易受环境光线和障碍物的影响，并且缺乏对深度信息的感知。在三维场景中，二维图

像检测技术的应用具有很大的局限性，因为三维空间中的目标检测需要准确定位检测物体的长、宽、高、物理位置和方向等信息，以提高机器与真实世界的交互能力。

基于激光雷达的三维目标检测任务存在一些难以克服的技术问题，模型的鲁棒性较差，主要由于激光传感器发射并收集返回激光束的工作原理以及点云数据标定技术不成熟等原因[2]。一方面，激光雷达在雨雪、雾以及灰尘等天气下的扫描精度会受到影响，导致检测算法极易丢失被检测目标；另一方面，受点云数据的稀疏、范围有限和激光雷达价格昂贵等因素的影响，使得自动驾驶技术的普及成为一项极具挑战的工作。通过融合多模态数据的方式可以充分利用不同传感器获取数据的优点，以此达到更准确感知周围复杂场景的目的，这也是三维空间目标检测领域未来的发展趋势[3]。目前，被研究最多和应用最广泛的多模态融合方案是点云与图像的融合。

点云和图像是场景感知中两种重要的数据源。图像可以提供对象的纹理和颜色等信息，点云数据可以提供精准的深度和几何信息[4]。随着检测目标和 LiDAR 传感器之间距离的增加，点云的密度通常会急剧下降，而图像在这种情况下作为辅助信息，充分利用其自身数据的特性为点云检测困难目标提供必要的信息。早期研究者们提出了诸多点云与图像数据的融合方案，利用数据之间互补特性以得到更加准确的检测结果。由于点云的无序、稀疏等特点，因此很难直接使用成熟卷积网络进行特征学习。目前，已有的一些融合检测方法，主要采用了两种解决方案：第一种是对图像进行特征提取，然后将目标在图像上对应位置的特征映射到点云空间中，最后推测目标的三维信息；第二种是先将点云数据转化为图像，然后通过卷积网络完成特征提取，最后将其特征与图像特征融合。

虽然以上检测方法在各大公开数据集上分别取得了不错的成绩，但仍然存在不足。例如，第一种方式对不同传感器之间的标定矩阵格外敏感，如果标定矩阵质量很差会导致前景点映射到背景点，严重影响到检测结果[5]。第二种方式将三维数据转化为二维数据，会导致特征信息损失，同时也会导致无法充分融合点云和图像特征。因此，我们需要进一步研究在多模态融合中如何保留更多的点云和图像的丰富语义信息，更高效地完成融合，进而提高算法检测的准确性以及增强网络的鲁棒性。

综上所述，基于点云和图像融合的目标检测在自动驾驶技术中是一项极具挑战性的任务。通过从点云和图像中提取有用的特征信息，将这些信息进行融合，更准确地检测出目标的位置和形状等信息。从而优化感知系统后续任务。

1.2 国内外研究现状

三维目标检测在很多领域都有广泛的应用，比如自动驾驶、机器人技术、虚拟现实等。通过对三维数据的处理和分析，我们可以更加准确地获取物体的信息，从而实现更智能化、更自动化的应用场景。国内外已经涌现出大量优秀的检测算法，根据数据源类型可分为三大类：基于图像、点云和多源传感器进行融合的三维目标检测[6]。

1.2.1 基于图像的目标检测算法

基于图像的三维目标检测算法的模型输入可以是单目图像或者双目立体图像。该算法与二维目标检测算法本质相同，但不同之处在于三维目标检测需要依赖于一些先验知识或者深度估计的方法。Chen[46]等人提出了一种仅使用单目相机实现同等三维目标检测性能的算法，该算法首先使用物体所在平面与图像平面正交作为约束条件，然后借助能量最小化方法来产生候选框，接下来，使用一些手工标注的特征先对产生的候选框打分，最后根据得分对候选框进行修正，从而提高检测结果的准确性。Chabot[8]等人提出了一种以单张图片为输入的多任务网络，该网络使用特定的二维目标检测器来预测图像中目标的二维检测框、二维坐标、部分可见性以及与 CAD 模型的相似性，能够对被检测目标的位置以及方向进行估计。Mousavian[9]等人提出了一种基于先验几何属性的三维目标检测和姿态估计网络，该网络利用物体三维坐标角点的透视投影与二维物体检测框紧密无限靠近的原理，结合一些先验的知识可以估计出被检测目标的三维框方向以及位置。Tong He[10]等人提出了更健壮的评分机制和三维框生成器 Mono3D 网络，其网络结构如图 1.2 所示，进一步提高了基于单目方法的精度。由于基于单目图像的目标检测更依赖先验的信息，可能会限制模型在复杂场景中的推广。

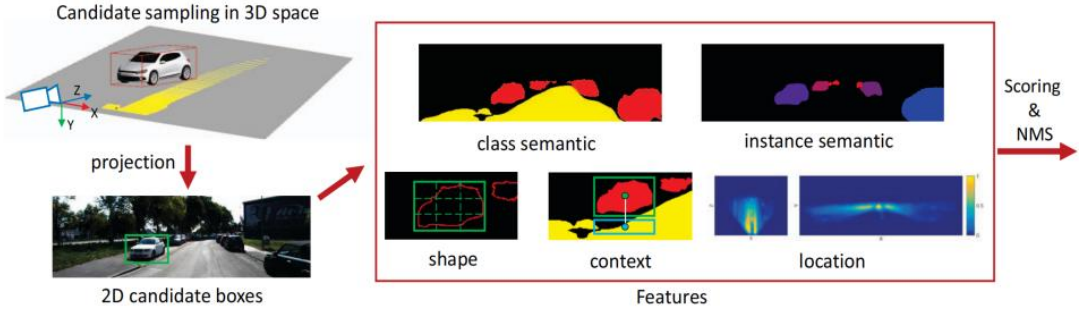

图 1.2 Mono3D 网络框架[10]

Li P[11]等人提出通过将双目图像上的候选框生成问题表述为马尔可夫随机场(MRF)的能量最小化问题，并利用上下文信息提高三维物体检测框的准确率。提出一种可以同时利用二维和三维之间的投影关系的算法。该算法利用权重共享网络分别提取左右视图的特征并生成二维检测框。接着，左边视图的特征主要用作被计算目标的一些关键点，再将关键点通过和三维检测框、二维检测框之间的映射关系计算三维估计结果。虽然该方法是当时检测准确度最高的图像检测算法，但由于图像的局限性，该方法与基于原始点云的检测算法在精度上仍有相当差距。

由于图像缺乏深度信息，想要从图像中获取深度信息，通常会采用手工的特征进行匹配计算。随着神经网络的出现，深度信息的计算逐渐被网络替代。Chang[47]等人提出，该网络主要借助卷积网络来进行立体图像的深度估计；Godard[12]等人提出一种基于自监督的深度估计方法，对采样方法和相关的损失函数进行了改进，得到更精确的深度估计结果；此外，Wang[13]等人提出将深度估计网络应用到图像中，先是对图像中的每个像素点的一些伪点云坐标进行计算，再借助点云的一些检测方法，最后对三维空间中的目标完成检测，该方法较好的提升了三维空间中目标的检测精度。

1.2.2 基于点云的目标检测算法

不同于图像数据的稠密性和规则性，点云数据具有稀疏性和不规则性，点云数据无法直接作为卷积神经网络的输入。为了解决这个问题，早期研究者们提出了许多点云数据的规则表示方法和特征提取方法，以此实现点云数据的三维目标检测任务。根据点云表示方法的不同，现有的检测方法可分为多视图法、体素法和原始点云法。

基于多视图的检测方法，需要将稀疏的点云数据转换为前视图或者鸟瞰图等密集表示的二维视图，以便于应用成熟二维目标检测框架。Li B[14]等人提出一种将点云转换为前视二维特征图的方法，转换之后的特征图可以作为扩展后的三维 FCN 网络的输入，以实现三维空间中目标的检测。由于三维空间中的物体可能处于不同的深度位置，前视图的映射重叠会导致一些点云数据的遮挡问题[15]。自上而下的鸟瞰图压缩的是物体高度，又因为所有物体几乎都处于同一水平面，所以压缩高度可以保留被检测物体的深度信息以及形状信息，且不会有物体遮挡问题。紧接着，Yang[16]等人则提出 PIXOR，该方法的策略主要是将点云变换成一种更为紧凑的二维鸟瞰图表示，有效的避免了前视图的遮挡问题，并且兼顾了计算速度和检测精度。在一定程度上这种做法也存在缺陷，比

如在点云的量化过程中可能会损失某些维度的信息，从而降低了网络特征的提取效果。

体素检测方法常将不规则的点云转化为体素表示，再通过三维卷积神经网络实现特征提取。体素法会导致信息丢失，从而降低定位精度。为了解决这个问题，Li[17]等人提出了基于二进制体素编码的 3D-FCN 网络，而 Engelcke[18]等人则提出了 Vote3deep。该算法使用 6 个统计量编码非空体素，并结合改进的卷积层结构，实现更高效的目标检测。针对不断变化的应用场景，传统的特征提取方法无法提取三维空间信息，所以如何让体素进行自动编码成了一个亟待解决的问题。Zhou[19]等人提出了一个端到端可训练的网络，该网络将点云划分为等间距的三维体素，并将体素内的点集转换为矢量，利用卷积层获取更多上下文信息，经检测头输出三维目标预测框的结果，该方法的计算复杂度随体素数量的立方速度增长，从而需要付出较高的计算成本。Yan[20]等人提出了一种基于 VEF 体素的改进网络，该网络通过稀疏卷积运算提高计算速度并减少内存消耗，并在速度和精度上取得了很大的进步。SECOND[48]网络采用稀疏三维卷积，相对于 VoxelNet[49]中的零填充体素而言计算效率有了很大的提高。受卷积操作感受野大小的影响，在多尺度特征学习方面，三维卷积具有天然的缺点，会导致模型的鲁棒性较差。

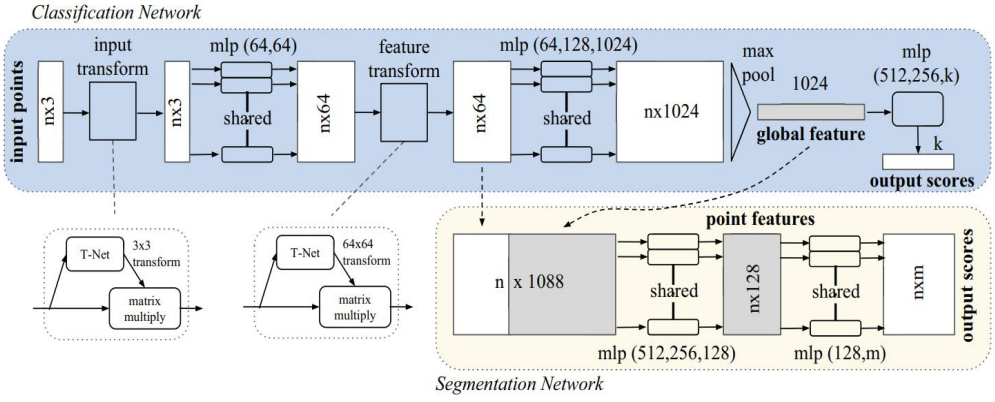

图 1.3 PointNet 网络架构图[21]

基于点云数据的目标检测方法与多视图法和体素法表示不同，它直接将无序、不规则的原始点云数据作为输入，最大限度地保留了点云数据信息。Qi[21]等人率先提出了 PointNet，这是第一个直接以原始点云数据作为输入进行学习的网络，它能够直接提取原始点云数据中的特征，网络架构见图 1.3。PointNet 及其改进版本能够自适应地捕捉点云数据的局部信息和细粒度信息，在点云稠密任务上表现出令人惊叹的性能。Shi[22]等人提出了两阶段检测模型 PointR-CNN，在第一阶段使用 PointNet 从点云中提取信息，使用集合抽取层进行下采样和特征传播层进行特征传递和上采样，使用区域建议网络生

成三维建议框。在第二阶段结合池化层和局部特征完成边界框修正和置信度评估。针对 PointNet 系列网络中 FP 层（Feature Propagation Layer）和细化模块导致的推理耗时问题，Yang[23]等人提出的 3D SSD 的改进。3D SSD 通过移除 FP 层、融合欧氏度量和特征度量等方式，在下采样期间弥补不同前景实例内部点的损失，并提出无锚点回归头来降低内存消耗，进一步提高预测准确率。3D SSD 网络的三维目标检测推理速度大于 25 FPS（Frames Per Second），是 PointRCNN 网络的两倍。

1.2.3 基于多源融合的目标检测算法

激光雷达是自动驾驶感知系统中重要的传感器之一。激光雷达易受恶劣天气条件和盲点等问题的影响，特别是远距离小物体有效点云特别稀少，现有的纯点云算法无法做到准确有效检测。立体相机或者单目相机可以提供细粒度纹理信息和 RGB 属性，但缺少深度信息，但其相比于激光雷达的优势在于价格便宜数倍。每种传感器都有其优缺点，相机采集的前视图和激光点云获取到的鸟瞰图来自不同视角，导致实现多传感器协同难度大。有研究者提出多传感器有效融合可以避免单传感器提供信息不充分的问题，基于点云和图像融合的方式通常分为 Early-Fusion，Deep-Fusion，Asymmetry-Fusion 以及 Late-Fusion，见图 1.4。

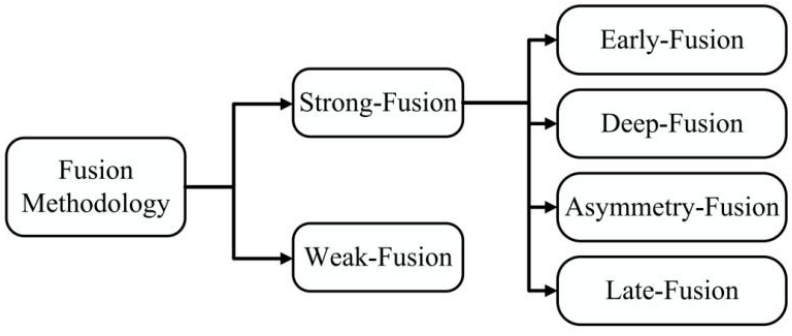

图 1.4 融合方法的分类

Late-Fusion 是利用成熟的二维检测网络生成感兴趣区域，再映射到对应的三维空间，结合点云数据的深度信息推测出三维物体的坐标和方向；Deep-Fusion 先将点云投影为鸟瞰图生成三维候选框，所有模态共享候选框完成候选框内所对应区域不同模态的融合，最后将各模态感兴趣区域融合完成三维目标检测。在点云和图像融合上，Late-Fusion 融合首先将点云和图像分别独立完成检测或分类，然后为每个传感器分配不同的权重，将两个网络的输出结果进行融合得出全局最优检测结果，这种方式又称为决策级