西南大學

# 本科毕业论文(设计)

## 题 目 一种多模态自动驾驶感知模型的改进

| 学 |   |   | 院 | 数学与统计学院         |
|---|---|---|---|-----------------|
| 专 |   |   | 业 | 统计学             |
| 年 |   |   | 级 | 2020级           |
| 学 |   |   | 号 | 222020314011081 |
| 姓 |   |   | 名 | 余乐扬             |
| 指 | 导 | 教 | 师 | 王建军             |
| 成 |   |   | 绩 |                 |

2024年5月19日

### 独创声明

本人郑重声明:所呈交的毕业论文(设计),是本人在指导老师的指导下,独立进行研究工作所取得的成果,成果不存在知识产权争议。尽我所知,除文中已经注明引用的内容外,本论文(设计)不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体均已在文中以明确方式标明。

声明的法律后果由本人承担。

## 作者签名: 点\*物

2014年05月19日

## 毕业论文(设计)使用授权声明

本人完全了解西南大学关于收集、保存、使用毕业论文(设计)的相关规定。

本人愿意按照学校要求提交论文(设计)的印刷本和电子版,同意学校 保存论文(设计)的印刷本和电子版,或采用影印、数字化或其它复制手段 保存论文(设计);同意学校在不以营利为目的的前提下,建立目录检索与 阅览服务系统,公布论文(设计)的部分或全部内容,允许他人依法合理 使用。

(保密论文在解密后遵守此规定)

## 作者签名: 点\*物

2014年05月19日

#### 目 录

i

## 摘要

| 摘  |                   | 1         |
|----|-------------------|-----------|
| Ał | bstract           | 1         |
| 1  | <b>绪论</b>         | <b>2</b>  |
|    | 1.1 研究背景          | 2         |
|    | 1.2 感知算法综述        | 2         |
| 2  | 多模态3D语义分割模型MSeg3D | <b>5</b>  |
|    | 2.1 模型简介及总结       | 5         |
|    | 2.2 工作原理          | 6         |
| 3  | 改进措施              | 8         |
|    | 3.1 问题提出与改进动机     | 8         |
|    | 3.2 改进措施          | 8         |
| 4  | 改进效果展示及对比         | <b>13</b> |
|    | 4.1 分类效果可视化       | 13        |
|    | 4.2 改进前后性能对比      | 16        |
|    | 4.3 改进前后效率对比      | 17        |
| 5  | 不足与展望             | 18        |
| 参  | 考文献               | 18        |
| 索  | 引                 | 19        |
| 致  | 谢                 | 20        |

## 一种多模态自动驾驶感知模型的改进

余乐扬

西南大学数学与统计学院,重庆400715

**摘要**: 本文首先对当下自动驾驶领域现有的感知模型分类做了概括性综述,逐 个列举了各类算法的优缺点及研究现状,随后重点介绍并分析了一种新的多模 态语义分割感知模型MSeg3D,客观评价了模型的创新点、优势及不足,随后 以MSeg3D为基础提出了三个可以优化改进方向:①跨注意力取代拼接、②线性 注意力机制、③模态内特征的深层提取和中期融合策略,经过了一系列实验证 明验证了改进方案的可行性(mIoU达到82.5,略高于MSeg3D;修改注意力机制 后,响应延迟缩短至原先的26.3%),最后对改进模型做了整体总结并对其未来做 出展望。

关键词: 自动驾驶; MSeg3D; 多模态; 语义分割; 注意力

### An Improvement with a Multi-modal Perception for Autonomous Driving

#### Yu Leyang

School of Mathematics & Statistics, Southwest University, Chongqing 400715

Abstract: This paper first gives a general review of the existing perception models in the field of autonomous driving, and analyzes both the advantages and disadvantages of various algorithms and the current research status. Then, the paper focuses on introducing and analyzing a new multi-modal semantic segmentation perception model MSeg3D with objective evaluations of the model's innovations, merits and drawbacks. And next, this paper proposes three optimization and improvement directions based on MSeg3D: ①cross-attention instead of splicing, ②linear attention mechanism, ③deep extraction of intra-modal features and mid-term fusion strategy. The feasibility of the improvement plan is verified through a series of experiments (The improved model's mIoU reaches 82.5, which is slightly higher than that of MSeg3D; after modifying the attention mechanism, the improved model's response latency is reduced to 26.3% of the original one). Finally, an overall summary of the improved model is made and its future prospects are given.

Key words: autonomous driving; MSeg3D; multi-modal; semantic segmentation; attention

#### 1 **绪论**

#### 1.1 研究背景

自动驾驶技术的发展日新月异,其快速发展为我们的日常生活出行带来了巨大便利。目前,不少电动汽车品牌均已配置部署了一定程度的智能驾驶模块,这其中较为 闻名的有特斯拉、比亚迪与华为(国内厂商一般将之称为辅助驾驶)。

感知系统,作为自动驾驶最为核心的成分之一,主要负责收集与处理周围环境的 信息,蕴含了道路边界监测、车辆检测、行人检测等关键技术。当前量产车型中的感 知算法实现多依赖于集成了视觉处理功能的自动驾驶芯片,同时也有广泛运用基于英 伟达等公司提供的高性能芯片开发自有算法的趋势。感知技术的发展经历了从传统视 觉算法到深度学习的转变,这一与时俱进的跃进在客观上极大地推动了自动驾驶技术 的进步。一般来说,感知系统与决策系统是自动驾驶汽车体系结构中的两个主要组成 成分,其中感知系统又可以被细分为定位、静态障碍物测绘、移动障碍物检测与跟踪 等多个子系统。感知是决策的前提,因此在某种意义上,感知系统的发展阶段决定了 整个自动驾驶技术的发展前景。

然而,如此重要的感知系统却至今都没有一个足够堪用的解决方案,现有的算法 都有着各自十分明显的局限性,因此自动驾驶感知模型的研究在学术界与工业界中均 为炙手可热的热门课题。

#### 1.2 感知算法综述

过去的自动驾驶感知技术往往不能高效的融合多种模态信息,在信息利用率与识别精度等等方面都迟迟无法突破瓶颈,所以现在最新的感知算法研究主要聚焦于多模 套感知算法,但这也面临着诸多未知的挑战。

当下,感知算法主要可被划分为以下四个种类:

#### (1). 基于相机的感知算法

利用二维图像信息进行分类任务是最早被人们考虑的办法,通过充当"人眼"的 相机捕获图像,再利用充当"人脑"的深度学习模型对图像实现识别、分割与分 类。

- 优点:由于相机拍摄图像分辨率较高且颜色信息丰富,图像数据蕴含着大量信息,因而在理想条件下可以检测各种物体,包括车道标记和交通标志。
- 缺点:考虑到相机的特性,这种算法对光照条件、遮挡物与气象条件(例如雨、雾、雪天气)极度敏感。

由于相机的表现受条件限制太大,现在少见基于相机的单模态感知算法,而是 常常与LiDAR配合作为多模态感知算法的输入(这是自动驾驶任务的需求与性质

#### 第2页共20页

所决定的)。基于相机感知算法的传统代表模型主要为用于对象检测和语义分割的CNN (卷积神经网络),如著名的AlexNet、VGGNet和ResNet;最新研究进展包括:用于对象检测和实例分割的Vision Transformers (ViTs) [6] [5]、鱼眼图像自适应卷积技术 [3] 与域泛化技术3DLabelProp [4]。

在后文的表格中简记为C。

#### (2). 基于LiDAR的感知算法

LiDAR,即激光雷达,作为一种集激光、GPS(全球定位系统)与INS(惯性导航 系统)于一体的先进技术,能够准确地捕获周遭环境的三维点云数据,使得模型 在分割时能够充分利用空间立体信息。

- 优点: 精确的3D点云数据、优良的光照条件鲁棒性和准确的深度测量。
- •缺点:数据稀疏、范围有限且难以检测小型或低反射率物体。

在当下的单模态感知算法中,基于LiDAR的感知算法是被研究最多、应用最广的。基于这类算法的传统代表模型主要为用于物体检测和分割的欧几里得聚类、RANSAC与基于体素的方法;最新研究进展包括:用于3D物体检测和语义分割的SalsaNext [7]、CENet [2]和其他流行的点云深度学习模型。

在后文的表格中简记为L。

(3). 基于雷达的感知算法

雷达早在第一次世界大战时就已被人类发明,是一项被广泛应用的成熟技术。雷达的工作依靠声波的反射,其原理与蝙蝠"听声辩位"相同。

- 优点:对天气条件具有鲁棒性,能够检测移动的物体并能提供相应的速度信息。
- 缺点:信号分辨率低、范围有限且难以区分物体类型,难以成为感知系统的 主要探测手段,故通常仅在多模态系统中起辅助作用。

目前的自动驾驶感知系统中的绝大多数雷达都被用于多模态感知算法,为相机 与LiDAR提供补充信息。用于目标检测和跟踪的恒定误报率(CFAR)方法是基于 雷达的感知算法的经典应用。

在后文的表格中简记为R。

(4). 多模态感知算法

多模态算法能充分结合、利用多种感受器捕获的信息,其出现标志着感知算法不 再单一或大部分地依靠某一类传感器。

- 优点:利用不同传感器模态的互补优势充分提高感知准确性和鲁棒性,例如LiDAR能够精确提供距离与形状信息,而相机则能捕捉颜色和纹理信息,通过对各种不同类型数据的统筹兼顾、充分利用各个传感器的优势,使车辆能够更加全面地理解附近复杂多变的道路环境,即使是在极端恶劣的天气条件下也能进行部分工作。
- 缺点:算法与模型的设计较为复杂,对算力的需求尤为强烈,目前传感器校 准问题和数据融合的复杂性都为多模态感知算法带来巨大挑战。

多模态学习的概念早在上世纪80年代就被提出,但直到2010年后随着深度学习的 崛起,尤其是近几年大模型研究突飞猛进之时才展现出其巨大威力。多模态感知 算法是所有感知算法中发展最为迅速的,而在多模态学习的实现方法上又可以被 进一步划分为依靠深度卷积神经网络进行特征提取和依靠跨注意力机制进行模态 融合(单流跨模态注意力机制 [9] [10],多流跨模态注意力机制 [11] [12]与多流单 流相结合的跨模态注意力机制 [13] [14]),前者的代表作有FusionNet,后者的代表 作有TransFuser,二者均为融合了多模态数据的端到端深度学习模型。

在各类多模态感知算法中,又数Transformer及其各种衍生模型风头最盛,是 当下最具潜力的先进技术(这里又以融合方式不再局限于简单的Cross-modal Attention和多模态Self-Attention、而是也可以借助Token本身来进行交互的新兴融 合机制为最前沿技术[15])。本文将要介绍并改进的模型MSeg3D(多模态3D语义 分割模型,Muti-modal 3D Semantic Segmentation)[1]则提出了一种新的跨模态 方法,可见多模态感知算法生命力之旺盛。

#### 2 多模态3D语义分割模型MSeg3D

由于本文的工作是基于多模态3D语义分割模型MSeg3D展开的,这里先对MSeg3D做 一个整体性介绍。这部分的内容主要参考了首次提出MSeg3D的文献 [1]。

#### 2.1 模型简介及总结

MSeg3D是一种新的基于相机传感器与LiDAR传感器的多模态3D语义分割模型, 其特色在于通过联合内部模态特征提取和跨模态特征融合来解决模态异质性问题, 从而为自动驾驶系统提供更全面的环境感知能力。MSeg3D包括基于几何的特征融合 (GF-Phase)、跨模态特征补全和基于语义的特征融合(SF-Phase)三个模块,这使得 模型能够在所有可见点上进行特征融合。除此以外,MSeg3D还通过对激光雷达点云和 多摄像头图像分别应用不对称变换来重新激活多模态数据增强,进一步提高了模型训 练的多样性。

MSeg3D在多个数据集上均取得了最先进的结果,实验表明,即便是在多摄像 头输入故障和多帧点云输入情景下仍然显示出了鲁棒性。目前,MSeg3D的代码已 在GitHub上公开,参见:https://github.com/jialeli1/lidarseg3d,可免费下载。

**现有多模态感知模型存在的问题**:回顾文献 [1],文献首先指出现有的多模态分割 模型,主要导致了以下三个关键性问题,这也是MSeg3D研究团队所着眼解决的问题。

- (1). 模态异质性问题:由于激光雷达点稀疏和像素密集,点云特征提取器和图像特征 提取器被分别开发,导致模态之间存在差异;
- (2).有限的传感器视场交集:只有落在传感器视场交集中的点才与多模态数据几何相关联,而仅考虑交集中的多模态数据对于实际应用来说是不够的;
- (3). 多模态数据增强:现有的多模态分割工作放弃了许多有用的点云增强变换,牺牲 了感知性能。

MSeg3D的特色与优势:为解决上述问题,研究团队进行了一系列创新性的工作, 其中最重要的贡献可以被概括为以下三点,这也是MSeg3D最大的亮点与特色。

- (1). 模态异质性: MSeg3D通过联合优化内部模态特征提取和跨模态特征融合, 最大限 度地提高了异质模态之间的相关性和互补性;
- (2). 特征融合: MSeg3D的多模态融合包括基于几何的特征融合、跨模态特征补全和基于语义的特征融合,这些特征融合在所有可见点上进行;
- (3). 数据增强: MSeg3D通过对激光雷达点云和多摄像头图像分别应用不对称变换来重 新激活多模态数据增强,增加了训练样本的多样性。

研究团队在MSeg3D在nuScenes、Waymo和SemanticKITTI三个公开数据集上对MSeg3D模型均进行了性能评估,同时横向对比了其他流行模型,结果表明MSeg3D均实现了最优异的综合表现,尤其是在挑战性小物体(如行人和交通锥)的性能表现显著优于仅使用激光雷达的方法,并且即便在故障的多摄像机输入和多帧点云输入下仍表现出鲁棒性。具体的性能评估对比见后文的表格2,在以总计16种一般路况下常见实体为指标的感知任务中,对比另外12种主流模型,MSeg3D一共在9个指标上取得了最佳成绩,并且达到了mIoU指标下的第一甲,这些都印证了MSeg3D的确是一个强大的自动驾驶感知模型。

总的来说,MSeg3D的研究为自动驾驶领域的3D语义分割提供了新的视角和技术 路线,展示了多模态数据融合在提高分割准确性和鲁棒性方面的潜力。随着自动驾驶 技术的不断进步,完全有理由相信MSeg3D等先进的感知模型将为我们实现更安全、更 可靠的自动驾驶系统做出重要贡献。

当然,根据实际需求,MSeg3D也可以"因地制宜"地做部分改进,这些内容将 在下一章中引出,本文的研究目标便是在MSeg3D原本的基础上尝试克服其部分缺点。

2.2 工作原理

下图(即图1)参考了文献 [1],较为形象地表明了MSeg3D的工作流程,稍后将会对图示做简明的文字诠释。



图 1: MSeg3原理,参考自文献 [1]

如图1所示, MSeg3D模型的输入包括两个模态:

- (1). LiDAR点云  $P_{in} \in \mathbb{R}^{N_{point} \times C_{in}}$ ,其中 $N_{point}$ 表示点的数量、 $C_{in}$ 表示每个点的输入 特征维数(如3D坐标和反射率);
- (2). 多摄像头RGB图像  $X_{in} \in \mathbb{R}^{N_{cam} \times 3 \times H_{in} \times W_{in}}$ ,其中  $N_{cam}$  是摄像头数量, $H_{in}$  和  $W_{in}$ 则 分别为图像高度和宽度。

首先,MSeg3D通过两个并行的预训练backbone网络分别对点云和图像进行单模态

特征粗提取,其目的是解决两种模态的异质性问题。对于LiDAR点云,MSeg3D采用基于体素的3D U-Net对点云进行特征提取。具体而言:

- 3D U-Net会先对输入点云进行体素化,将非空体素 $V_{in} \in \mathbb{R}^{N_{voxel} \times C_{in}}$ 作为稀疏张量输入到点云backbone网络中,从而得到表达性的体素特征 $V \in \mathbb{R}^{N_{voxel} \times C_{voxel}}$ 。
- 对于多摄像头图像,则采用可训练的图像backbone(这里考虑HRNet-w48)对 $X_{in}$ 进行非线性投影,得到下采样的图像特征图 $X \in \mathbb{R}^{N_{cam} \times C_{img} \times H \times W}$ 。

紧接着,是多模态特征融合的三个主要阶段:

- (1). 基于几何的特征融合阶段(GF-Phase):将体素特征V和图像特征图X通过几何 关联分别投影到每个点上,插值得到逐点的LiDAR特征 $F_{lidar} \in \mathbb{R}^{N_{point} \times C_{voxel}}$ 和逐 点的相机特征 $F_{cam} \in \mathbb{R}^{N_{point} \times C_{img}}$ 。随后,通过一个几何特征融合模块(GFFM) 将 $F_{lidar}$ 和 $F_{cam}$ 进行融合:使用一个全连接层将 $F_{lidar}$ 与 $F_{cam}$ 投影到同一维度 下并对两个特征矩阵进行拼接,得到基于几何信息融合的逐点特征 $F_{gfused} \in \mathbb{R}^{N_{point} \times C_{gfused}}$ 。
- (2). 跨模态特征完成:为补偿GF-Phase中未能完整融合点云和图像特征的缺陷,提出 了一个跨模态特征完成模块。该模块从 Flidar 预测出一个伪相机特征 Fpcam,用于 替换位于摄像头视野外的点的缺失相机特征。通过最小化 Fpcam 与真实 Fcam 之间 的均方差损失 Cpixel2point,实现了从LiDAR到相机特征的迁移。
- (3). 基于语义的特征融合阶段(SF-Phase): 首先通过LiDAR语义特征聚合模块 (LiDAR SFAM)和相机语义特征聚合模块(Camera SFAM)分别将体素特 征V和图像特征图X聚合为语义嵌入 $E_{lidar} \in \mathbb{R}^{N_{cls} \times C_{voxel}} 和 E_{cam} \in \mathbb{R}^{N_{cls} \times C_{img}}$ , 其中 $N_{cls}$ 是语义类别数。然后,通过语义特征融合模块对这些语义嵌入进行显式 的关系建模、利用多头注意力机制学习点与类别之间的语义关系,实现基于语义 信息的多模态特征融合,得到最终的逐点融合特征 $F_{sfused} \in \mathbb{R}^{N_{point} \times C_{sfused}}$ 。

最后,将F<sub>sfused</sub>输入到分类头中进行3D语义分割。以上即为MSeg3D的运行逻辑。

#### 3 改进措施

接下来是本文的工作,对MSeg3D进行了部分改良以适应某些情况下的实际需要。

#### 3.1 问题提出与改进动机

尽管MSeg3D性能表现优秀,但经实际测试后发现仍存在有一些可以根据实际应用 需求进行优化、改良的地方,主要表现为:

#### (1). 模态间特征拼接导致的计算损耗

MSeg3D在处理不同模态的数据时,采用了拼接的方式将不同模态间的数据进行整合。这种方式虽然能够保持每个模态内部数据的完整性,确保了信息的不丢失, 但同时整个数据集的维度会迅速增加,这会导致后续的操作中面临矩阵维度过高 的问题。高维度的矩阵不仅会增加计算的复杂性,还会对存储和处理能力提出更 高的要求,从而增加了计算资源的损耗。高维度的矩阵也会带来"维度灾难",即 随着维度的增加,所需样本的数量呈指数级增长,以保持模型的性能。

#### (2). Softmax注意力机制导致的计算损耗

Softmax注意力机制在多模态特征融合中也起着重要的作用。它通过为不同的特征 分配不同的注意力权重,从而强调重要特征,抑制不重要的特征。然而,这种机 制的计算复杂度通常与输入特征的数量成正比,当处理大规模点云数据时,计算 量会显著增加。当MSeg3D尝试将点云数据与图像的多模态数据进行特征提取和融 合时,由于点云的庞大数据量和Softmax注意力机制,导致整个过程的计算时间冗 长。这不仅影响模型的训练效率,还导致该模型的响应延迟非常高(达到445ms), 限制了其在现实中的使用。

#### (3). 过早地融合多模态数据导致融合效率降低

有效的模态融合应该是基于对每个模态深入理解和分析的基础上进行的,这样可以更好地捕捉和利用模态间的关联信息。MSeg3D过早地融合,可能导致某些重要信息的丢失,无法充分利用每个模态的独特信息,从而影响融合效果。

如分析所述,上面列举问题所导致的最显著后果是融合效率不佳与延迟延迟较高 (计算量较大),本文将在下一章提出具体的改进举措以一一解决上述问题。

#### 3.2 改进措施

针对上述问题,本文主要在MSeg3D模型的基础上修改了其中的SFFM模块,修改 后的新模型框架如图2所示。



图 2: 修改后的MSeg3框架

具体的改动点同样分为三项,主要聚焦于MSeg3D的特征融合阶段,分别对应上文中提出的改动前MSeg3D模型的三个问题:

#### (1). 跨注意力取代拼接

针对第一个问题,使用跨注意力机制(Cross-Attention)取代拼接操作,降低了特征矩阵的维度,减小了整整一倍的计算量 ( $2N_{cls} \times C \rightarrow N_{cls} \times C$ )。其主要实现逻辑为:

- (a) 把查询向量与键向量映射到不同的空间
- (b) 计算出查询向量与键之间的相似度,从而得到原始的关联度分布
- (c) 把各关联度与值向量分别相乘后再相加,得到融合了两个输入序列信息的跨 注意力表示
- (d)标准化,得到权重的最终分布

原本的MSeg3D将语义embeddings投影后拼接,得到包含 $E_{cam}$ 和 $E_{lidar}$ 信息的,之后通过多头自注意力机制得到提取后的语义embeddings,过早的拼接 $E_{cam}$ 和 $E_{lidar}$ 导致了在注意力模型中矩阵乘法的计算复杂度为( $2N_{cls}C^2$ )。使用跨注意力机制后,计算复杂度降为( $N_{cls}C^2$ )。

值得一提的是,虽然跨注意力机制能够在减小计算量的前提下融合多个来源的信息并处理跨模态数据,但同时也可能带来注意力偏置等一系列问题。

(2). Sub Attention机制

针对第二个问题,受清华大学黄高团队所提出的Agent Attention机制的启发 [17],本文在Agent Attention的基础上做了部分修改,将其特化为一种较之MSeg3D中的Softmax注意力更轻量化、更适合用于自动驾驶的注意力模型,并将其命名为 "Sub Attention"。



图 3: 三种注意力机制的构造,引用自文献 [17]

文献 [17]指出,虽然Softmax注意力在性能表现上有着很好的效果,但鉴于其计算成本高昂,需要提出一种新的注意力机制Agent Attention以平衡计算量与性能。新定义代理令牌*S*(token S),*S*本质上用作*Q*的代理,聚合来自*K*和*V*的全局信息,然后将其提取的特征通过矩阵乘法回传给*Q*;按照文献的思路,本文将代理令牌的数量*S*设置为一个较小的超参数。

而具体的代理令牌S需要通过不同的下采样方式得到。针对 $F_{gfused}$ 特征中包含的 点云空间信息,本文采用3D-UNet提取 $F_{gfused}$ 的代理令牌S; embeddings的代理 令牌S则通过全局平均池化来提取。

$$S_F = \operatorname{Conv}(Q)$$

 $S_E = \text{Pooling}(Q)$ 

其中 $Q, K, V \in \mathbb{R}^{N_{cls} \times C_{sfused}}$ ,  $S \in \mathbb{R}^{n \times C_{sfused}}$ 。 为了保证提取特征的多样性,本文专门设计了额外的空洞卷积操作:

SubAtt  $(E_n)$  = Softmax  $(QS_E^T) \cdot$  Softmax  $(S_EK^T) \cdot V + DWC(V)$ 

SubAtt (F) = Softmax  $(QS_F^T) \cdot$  Softmax  $(S_FK^T) \cdot V + DWC(V)$ 

本文将代理令牌S的矩阵大小设为n(n远小于C),这使得模型在保持全局上下文建模能力的同时,实现了相对于输入特征的数量N的线性计算复杂度(N<sub>cls</sub>nC),而原本的Mseg3D模型的计算复杂度为(2N<sub>cls</sub>C<sup>2</sup>),本文在保证全局信息提取能力的同时降低了计算的复杂度。

本文将上述机制称为Sub Attention。就理论上来说,Sub Attention结合了Softmax和 线性注意力的长处,是Agent Attention针对自动驾驶任务的特化与升级版本。Sub Attention具有以下优点:

- (a) 高效计算和高表达能力。以前的工作通常将Softmax注意力和线性注意力视为两种不同的注意力范式,旨在解决它们各自的局限性。作为这两种注意力形式的无缝集成,本文的代理注意力自然继承了两者的优点,同时具有较低的计算复杂度和高模型表达能力。
- (b) 大感受野。该模块可以在保持相同数量的计算量的同时采用较大的感受野。 得益于线性复杂度,改进后的模型可以在保持相同计算量的同时享受大甚至 全局感受野的优势。

#### (3). 模态内特征的深层提取和中期融合策略

针对第三个问题,为了深化模态内的特征提取、充分利用模态内信息,本文采用 了中期融合策略。本文对project后的每个矩阵( $F_{gfused}$ 、 $E_{cam}$ 、 $E_{lidar}$ )先进行 多头Sub Attention深层提取特征:

$$E_{s(modal)} = \left[ \text{SubAtt}(E_n) \right]_{n=1}^{n=N_h}$$

为使模态信息更充分,根据文献 [14],考虑在k1层Sub Attention深层提取模态内特征后再进行embeddings的融合:

$$E_{ca} = \text{MHCA}(E_{slidar}, E_{scam}, E_{scam})$$

$$F_{ca} = \operatorname{Norm}(F_{proj} + \operatorname{MHCA}(F_{proj}, E_{ca}, E_{ca}))$$

$$F_{fused} = \operatorname{Norm}(F_{ca} + \operatorname{FFN}(F_{ca}))$$

本文使用多头跨注意力(MCHA)在公式中用于不同模态语义嵌入 *E*<sub>lidar</sub> 和 *E*<sub>cam</sub> 的 融合,以及语义关系建模的逐点投影特征 *F*<sub>proj</sub> 和语义嵌入 *E*<sub>ca</sub> 之间进行进一步融合,前馈网络(FNN)用于等式中特征的前向传播。



图 4: 修改后模型的主要运作逻辑

#### 4 改进效果展示及对比

#### 4.1 分类效果可视化

这里给出三份实验输入输出样本以对分类效果做可视化演示,均为本文使用 改进后的模型运行得出。三份样本随机抽取自公开数据集SemanticKITTI [16]的验 证集部分,模型则在训练集部分上进行训练。SemanticKITTI是一个大规模的真实 场景LiDAR点云序列逐点注释的数据集,标注28类语义,共计22个sequences与43000 scans,是MSeg3D研究团队在实验中选择的数据集之一,也是本文在稍后的实机实验 中会用到的数据集之一。

SemanticKITTI采用不同的颜色以对不同的语义分类进行可视化,颜色对应类别的 映射如图5的图例所示,本文将采取同样的可视化方案。



图 5: SemanticKITTI可视化示例,引用自文献 [16]

下面的三份实验输入输出样本中, 左半部分图片均为输入样例可视化, 右半部分 均为实机输出分类可视化。



图 6: 可视化案例(i)

可以看出,改进模型精确地区分开了路面与边界,对边缘的识别十分准确,特别 是正确分类了属于建筑物部分的少量点云。

### 样例二



图 7: 可视化案例(ii)

第14页共20页

这里同样准确区分开了各语义的边界,重点是正确识别到了混杂在路边植被附近 的行人。

样例三



图 8: 可视化案例(iii)

这份样例中有较多的行人,路况要素较丰富,模型仍正确分开了道路、车辆、行 人与植被。



图 9: 可视化案例(iv)

第15页共20页

这是俯视图,可以看出模型的识别是足够准确的。

#### 4.2 改进前后性能对比

下表为不同模型间的性能评估对比表,以"Ours"表示经本文改进后的MSeg3D模型。采用与文献 [1]相同的评估标准,其中Methods为模型名称,Modality为模型所利用的信息,mIoU代表均交并比,其他标签均为实验中模型对该类别分类的准确率 (accuracy)。所有的数值取值均在0至100之间,越接近100代表模型性能愈强劲。

表中除了"Ours"外的全部数据均获得自文献 [1],是浙江大学丁勇教授研究团队在nuScenes、Waymo和SemanticKITTI三个数据集对包括MSeg3D在内的若干模型进行实验得到的数据;受条件限制,本文对且仅对"Ours"在Waymo与SemanticKITTI上进行了实验,并将结果与文献 [1]中的数据做横向对比。

|                             |          |       |         |         |       | 11    | 1. 13     | モビロ        | 日日日        | 10 \1        |         |       |           |       |                     |         |         |            |
|-----------------------------|----------|-------|---------|---------|-------|-------|-----------|------------|------------|--------------|---------|-------|-----------|-------|---------------------|---------|---------|------------|
| Methods                     | Modality | mloU  | Barrier | Bicycle | Bus   | Car   | C-Vehicle | Motorcycle | Pedestrian | Traffic Cone | Trailer | Truck | D-Surface | Other | $\mathbf{Sidewalk}$ | Terrain | Manmade | Vegetation |
| PolarNet                    | L        | 69.42 | 72.16   | 16.81   | 77.01 | 86.53 | 51.14     | 69.65      | 64.80      | 54.11        | 69.70   | 63.53 | 96.64     | 67.14 | 77.70               | 72.13   | 87.13   | 84.47      |
| JS3C-Net                    | L        | 73.60 | 80.14   | 26.15   | 87.79 | 84.54 | 55.17     | 72.56      | 71.28      | 66.26        | 76.79   | 71.11 | 96.80     | 64.47 | 76.86               | 74.09   | 87.48   | 86.10      |
| Cylinder3D                  | L        | 77.16 | 82.76   | 29.75   | 84.34 | 89.41 | 63.03     | 79.29      | 77.21      | 73.40        | 84.55   | 69.17 | 97.66     | 70.24 | 80.29               | 75.51   | 90.41   | 87.55      |
| AMVNet                      | L        | 77.27 | 80.64   | 31.96   | 81.73 | 88.93 | 67.07     | 84.33      | 76.11      | 73.48        | 84.87   | 67.30 | 97.37     | 67.37 | 79.41               | 75.45   | 91.45   | 88.69      |
| SPVNAS                      | L        | 77.35 | 80.00   | 29.98   | 91.92 | 90.81 | 64.68     | 78.99      | 75.62      | 70.94        | 81.01   | 74.64 | 97.44     | 69.23 | 79.95               | 76.10   | 89.28   | 87.06      |
| Cylinder3D++                | L        | 77.86 | 82.76   | 33.89   | 84.34 | 89.41 | 69.63     | 79.42      | 77.26      | 73.40        | 84.55   | 69.41 | 97.66     | 70.24 | 80.29               | 75.51   | 90.42   | 87.55      |
| AF2S3Net                    | L        | 78.34 | 78.87   | 52.21   | 89.93 | 84.17 | 77.42     | 74.30      | 77.32      | 71.95        | 83.88   | 73.78 | 97.13     | 66.47 | 77.51               | 74.01   | 87.69   | 86.80      |
| SPVCNN++                    | L        | 81.12 | 86.35   | 43.13   | 91.90 | 92.18 | 75.90     | 75.72      | 83.44      | 77.31        | 86.82   | 77.36 | 97.69     | 71.22 | 81.08               | 77.19   | 91.67   | 88.98      |
| LIFusion*                   | LC       | 75.74 | 58.13   | 36.30   | 86.67 | 84.28 | 59.96     | 79.69      | 80.30      | 77.77        | 83.23   | 68.74 | 97.18     | 68.19 | 77.04               | 74.45   | 91.03   | 88.95      |
| PMF                         | LC       | 77.03 | 82.11   | 40.33   | 80.94 | 86.42 | 63.72     | 79.22      | 79.75      | 75.86        | 81.17   | 67.05 | 97.28     | 67.69 | 78.05               | 74.48   | 89.94   | 88.46      |
| $\operatorname{CPFusion}^*$ | LCR      | 77.72 | 83.67   | 37.03   | 89.02 | 86.24 | 70.08     | 77.47      | 78.07      | 74.53        | 82.78   | 67.94 | 96.64     | 68.24 | 79.53               | 74.91   | 90.47   | 86.95      |
| 2D3DNet                     | LC       | 79.96 | 83.01   | 59.35   | 87.99 | 85.09 | 63.70     | 84.39      | 81.95      | 75.96        | 84.79   | 71.93 | 96.88     | 67.35 | 79.81               | 75.96   | 92.05   | 89.18      |
| MSeg3D                      | LC       | 81.14 | 83.11   | 42.46   | 94.92 | 92.01 | 67.10     | 78.58      | 85.66      | 80.47        | 87.53   | 77.32 | 97.74     | 69.82 | 81.22               | 77.83   | 92.35   | 90.07      |
| Ours                        | LC       | 82.5  | 85.1    | 53.4    | 93.2  | 91    | 78.5      | 78.9       | 84.3       | 77.5         | 86.8    | 76.2  | 97.8      | 66.6  | 81.4                | 77.3    | 91.2    | 87.7       |

表 1: 模型性能评估对比

可以看出, MSeg3D在实验的条件下相较于若干主流模型具有一定的优势。在全部的16项评估指标中, 有9项表现突出, 有3项则表现出明显的不足。

经本文改进过后的模型,即表中的"Ours",与MSeg3D的性能表现对比可以总结为:

- 有3项表现明显优于MSeg3D,特别注意的是有2项的提升均超过15%(分别达到了17.98%与17.00%),而这部分恰是MSeg3D极不擅长的;
- 有7项可以认为在表现上与MSeg3D没有差别,得分差距均在0.5%以内;
- 有3项与MSeg3D相比存在较显著劣势,分类能力有所下降;
- 总得来看,改进后模型的均交并比mIoU略高于MSeg3D,这说明改进后的模型至 少有着不差于MSeg3D的综合表现。经过对表格的初步分析,改进后模型mIoU的 提升可能源自于MSeg3D有着部分明显不擅长的应用场景,恰好改进后的模型相对 地补足了这些短板。

结论1:改进后模型的综合性能表现稍强于MSeg3D,特别是针对一些MSeg3D并 不擅长的场景。

#### 4.3 改进前后效率对比

下表为针对本文改进后的模型,控制其他条件不变、仅分别应用不同的注意 力机制,实验并记录三种情况下模型的综合性能指标(以mIoU为标准)、参数数量 (Params)以及响应延迟(Latency)的对比表,其中"Sub Attention"代表本文的改 进方案,用参数数量与响应延迟对实验中的"计算量"进行量化。

| 表 2: 三种注意力机制应用效果对比 |      |                            |               |  |  |  |  |  |  |
|--------------------|------|----------------------------|---------------|--|--|--|--|--|--|
| Transformer        | mIoU | $\operatorname{Params}(M)$ | M) Latency(s) |  |  |  |  |  |  |
| Softmax Attention  | 81.4 | 87.34                      | 0.445         |  |  |  |  |  |  |
| Linear Attention   | 75.3 | 48.84                      | 0.108         |  |  |  |  |  |  |
| Sub Attention      | 82.5 | 51.85                      | 0.117         |  |  |  |  |  |  |

三种注意力机制在模型上的应用表现可以总结为:

- Softmax Attention具有不错的性能表现,但同时也有着最大的参数数目与最高的 响应延迟(最大的计算量),其平均延迟分别约为Linear Attention的4.1倍与Sub Attention与3.8倍;
- Linear Attention具有最少的参数数目与最低的响应延迟(最小的计算量),但其性能相对逊色,mIoU均明显低于Softmax Attention与Sub Attention;
- Sub Attention具有最佳的整体表现,在三种注意力机制中取得最高mIoU成绩的同时,相较于最简单的Linear Attention并未增加太多参数与响应延迟,融合了Softmax Attention与Linear Attention的长处,在性能与效率间取得了令人满意的平衡。

结论 2: 改进后的模型相较于MSeg3D提升了效率,在综合性能没有变差的前提 下减小了一定的计算量。

#### 5 不足与展望

本文的完成(包括实验在内)历时约五个月半,恰逢毕业季需要处理的事务较多,时间上相对紧张,加之作者能力有限,改进难免存在着一些疏漏。

改进模型中可能存在的问题与后续潜在的再改进方向有:

- (1). 本文只对MSeg3D模型的特征融合部分进行了改进,而其特征提取阶段可能还有更大的潜在优化空间。
- (2). 在跨注意力中没有考虑注意力偏置的问题。
- (3). 受资源限制,本文没有在nuScenes数据集上对改进后的模型进行实验。

但本文所提出的改进思想也可以为MSeg3D的后续改进或其他模型的性能优化提供新的思路:

- (1).本文为降低计算量而提出的一系列降维操作可能有更多的多模态模型可以使用, 有助于提升多模态融合效率。
- (2). 本文提出的基于Agent Attention特化、修改而得的注意力机制Sub Attention可能 有更多的语义分割模型可以使用,有助于提升深层特征提取效率。

综上所述,虽然改进还存在着若干缺陷和瑕疵,但本文认为改进后的模型具有一定的应用潜力,并且改进的思路也可以被应用于其他符合条件的模型上。

#### 参考文献

- Li, Jiale, Hang Dai, Hao Han, et al. MSeg3D: Multi-Modal 3D Semantic Segmentation for Autonomous Driving[C]. Proceedings of the 2023 IEEE Conference on the Computer Vision and Pattern Recognition, 2023;21694-21704
- [2] Cheng H X, Han X F, Xiao G Q. Cenet: Toward Concise and Efficient Lidar Semantic Segmentation for Autonomous Driving[C].Proceedings of the 2022 IEEE International Conference on Multimedia and Expo.2022: 01-06.
- [3] Playout C, Ahmad O, Lecue F, et al. Adaptable Deformable Convolutions for Semantic Segmentation of Fisheye Images in Autonomous Driving Systems[J]. arXiv preprint arXiv:2102.10191, 2021.
- [4] Sanchez J, Deschaud J E, Goulette F. Domain Generalization of 3D Semantic Segmentation in Autonomous Driving[C].Proceedings of the 2023 IEEE/CVF Conference on Computer Vision. 2023: 18077-18087.
- [5] Ando A, Gidaris S, Bursuc A, et al. Rangevit: Towards Vision Transformers for 3D Semantic Segmentation in Autonomous Driving[C].Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 5240-5250.
- [6] Jiang F, Tu C, Zhang G, et al. Revisiting Multi-modal 3D Semantic Segmentation in Real-world Autonomous Driving[J]. arXiv preprint arXiv:2310.08826, 2023.
- [7] Cortinhal T, Tzelepis G, Aksoy E E. Salsanext: Fast Semantic Segmentation of Lidar Point Clouds for Autonomous Driving[J]. arXiv preprint arXiv:2003.03653, 2020, 3(7): 2.
- [8] Chen Y C, Li L, Yu L, et al. Uniter: Universal image-text representation learning[C]Proceedings of the European Conference on Computer Vision.2020: 104-120.
- [9] Kim W, Son B, Kim I. Vilt: Vision-and-language Transformer Without Convolution or Region Supervision[C].Proceedings of the 2021 IMLS Conference on Machine Learning.2021: 5583-5594.
- [10] Tsai Y H H, Bai S, Liang P P, et al. Multimodal Transformer for Unaligned Multimodal Language Sequences[C].Proceedings of the 2019 ACL Conference on Natural Language Processing. 2019: 6558-6569.
- [11] Lu J, Batra D, Parikh D, et al. Vilbert: Pretraining Task-agnostic Visiolinguistic Representations for Vision-and-language Tasks[J]. Advances in neural information processing systems, 2019, 32.
- [12] Li R, Yang S, Ross D A, et al. Ai Choreographer: Music Conditioned 3D Dance Generation with Aist++[C].Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. 2021: 13401-13412.
- [13] Recasens A, Lin J, Carreira J, et al. Zorro: the Masked Multimodal Transformer[J]. arXiv preprint arXiv:2301.09595, 2023.
- [14] Nagrani A, Yang S, Arnab A, et al. Attention Bottlenecks for Multimodal Fusion[J]. Advances in Neural Information Processing Systems, 2021, 34: 14200-14213.
- [15] Wang Y, Chen X, Cao L, et al. Multimodal Token Fusion for Vision Transformers[C].Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12186-12195.
- [16] Behley J, Garbade M, Milioto A, et al. A Dataset for Semantic Segmentation of Point Cloud Sequences[J]. arXiv preprint arXiv:1904.01416, 2019, 2(3): 12.
- [17] Han D, Ye T, Han Y, et al. Agent Attention: On the Integration of Softmax and Linear Attention[J]. arXiv preprint arXiv:2312.08874, 2023.

致谢:

当重大疾病悄然降临,这一小概率事件恰好发生在21岁的我身上,才意识到人的脆弱与不自 由。西大四年,谢谢有大家的陪伴,特别是本科毕业前夕、被迫住院的这段时光里。

特别感谢指导老师王建军教授、系主任李婷婷老师和荣腾源同学,你们在我住院期间给予我鼓励与支持,协助我顺利通过了论文的答辩。没有你们的鼎力相助,能否顺利毕业或许都是未知数。

感谢高中好友曾光强、卜英博,感谢杨虎城学弟,感谢来自重庆邮电大学的同学刘俊杰、冯宇 祥,在我对病症焦虑时给了我及时的宽慰。

感谢西南大学吧的朋友们,包括小朱、小叶、小张\*2、小毛、小汪、小白、小崔等等。

感谢亲人,除父母外,特别是表哥表姐(小白小黑小花)、姑妈姑父等等。治疗的过程,我无 法离开你们的关心与帮助。

接下来,如果一切顺利,我将在虎溪、在重庆大学继续学业。嘉陵泱泱、缙云巍然,谢谢一路 上的大家。由衷祝愿所有的好朋友与亲人们身体健康、不必受病痛之苦,也祝自己早日康复、生活 尽快回到正轨。

> 余乐扬 2024年5月19日09:44 于绵阳市第一人民医院