西南大学

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 4 月 15 日

目　录

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： 本文首先对时下自动驾驶领域现有的感知模型做了概括性综述，分别分析了各类算法的优缺点及研究现状，随后重点介绍并剖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、优势与不足，接着在MSeg3D的基础上提出了三个可以优化与改进方向：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，经过一系列实验证明验证了改进方案的可行性（改进模型在mIoU指标上取得了最先进成绩，达到了82.5），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

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 multimodal 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 achieved state-of-the-art results among the mIoU indicator, reaching 82.5). 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（C）**

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

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]。

(2). **基于LiDAR的感知算法（L）**

LiDAR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，能够捕获周遭环境的三维点云数据，使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。

(3). **基于雷达的感知算法（R）**

雷达早在第一次世界大战时就已被人类发明，用以在反空袭战中搜寻敌机。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。

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。

(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1，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。

MSeg3D当然也存在着若干缺陷，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。

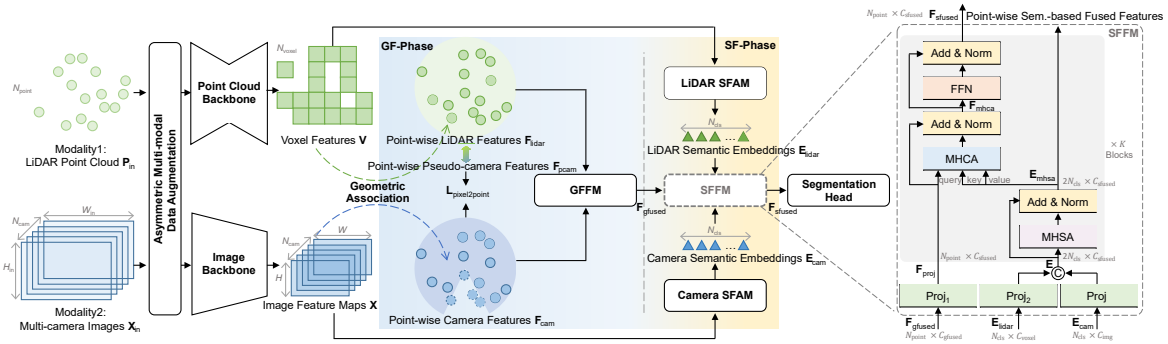

图 1: MSeg3原理

如图所示，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backbone网络分别对点云和图像进行单模态特征提取，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（如HRNet）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，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，以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 F_{lidar} 预测出一个伪相机特征 F_{pcam} ，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 F_{pcam} 与真实 F_{cam} 之间的均方差损失 $\mathcal{L}_{pixel2point}$ ，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_{sfused}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 改进思路

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。

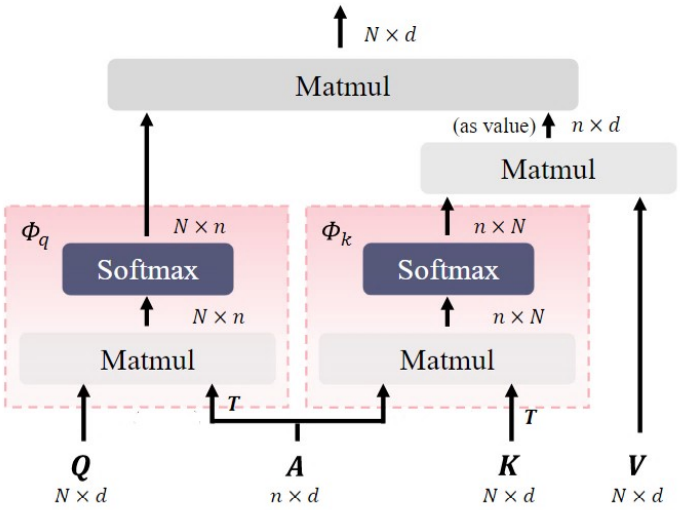

图 2: Softmax注意力机制实例，图片源自文献 [17]

(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。

上述问题导致的最主要后果是算力需求增加。当然，这对于有着充足计算资源的实验室或大企业而言或许算不上什么，但如果个人或小型企业希望将MSeg3D模型进行微调后用于其他用途，算力紧张就会是一个令人头疼的问题了。为此，本文对MSeg3D的某些部分进行了更改与调整，使之更适应低算力机器。

3.2 改进方案

针对上述问题，本文在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。

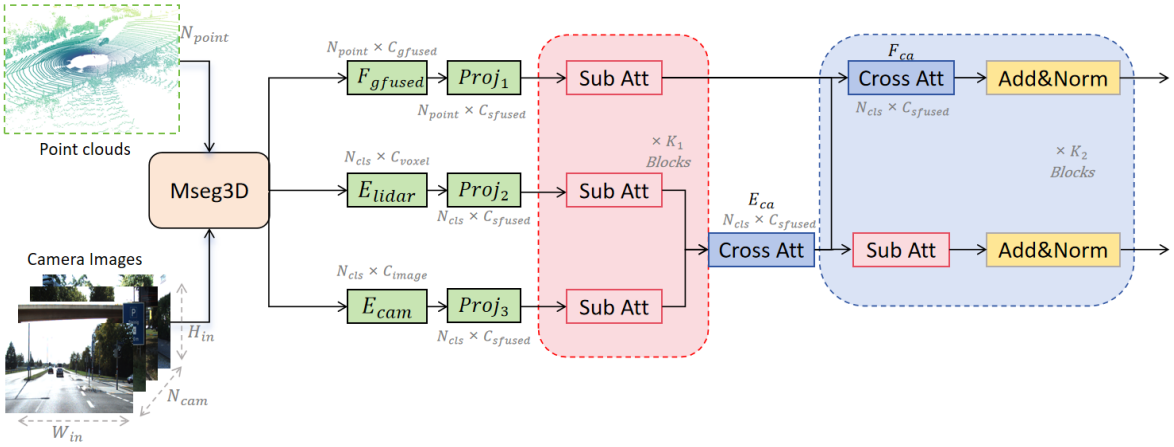

图 3: 修改后的MSeg3框架

修改后的主要运作逻辑为：

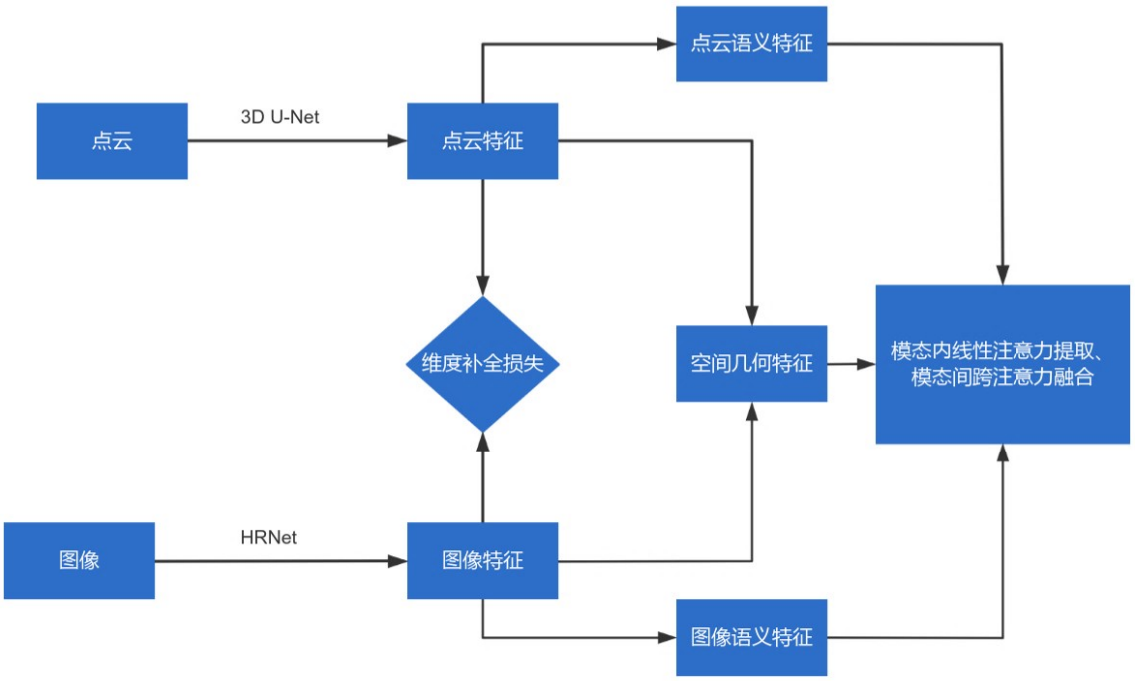

图 4: 主要运作逻辑

具体的改动点同样分为三条，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

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). **线性注意力机制**

针对第二个问题，本文建议将Softmax注意力机制更换为一种线性注意力机制，在后文中以“Sub Attention”代称。新定义的代理令牌 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。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，为此需要提出一种新的注意力机制以平衡计算量与性能。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，这让模型在保持全局上下文建模能力的同时，实现了相对于 $\mathcal{O}(Nnd)$ 输入特征的数量 N 的线性计算复杂度。

具体的token S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token S，而Embeddings的token S则通过全局平均池化来提取。

$$S_F = \text{Conv}\left(Q\right)$$

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

其中 $Q, K, V \in \mathbb{R}^{N_{cls} \times C_{sfused}}, \quad S \in \mathbb{R}^{n \times C_{sfused}}$ 。

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：

$$\text{SubAtt}\left(E_n\right) = \text{Softmax}\left(QS_E^T\right) \cdot \text{Softmax}\left(S_E K^T\right) \cdot V + DWC\left(V\right)$$

$$\text{SubAtt}\left(F\right) = \text{Softmax}\left(QS_F^T\right) \cdot \text{Softmax}\left(S_F K^T\right) \cdot V + DWC\left(V\right)$$

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。

总的来说，结合Softmax和线性注意力的优点，本文提出的这一模块具有以下优点：

(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：

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

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：

$$E_{ca} = \mathrm{MHCA}\left(E_{slidar}, E_{scam}, E_{scam}\right)$$

$$F_{ca} = \mathrm{Norm}\Big(F_{proj} + \mathrm{MHCA}\left(F_{proj}, E_{ca}, E_{ca}\right)\Big)$$

$$F_{fused} = \mathrm{Norm}\Big(F_{ca} + \mathrm{FFN}\left(F_{ca}\right)\Big)$$

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_{lidar} 和 E_{cam} 的融合，以及语义关系建模的逐点投影特征 F_{proj} 和语义嵌入 E_{ca} 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。

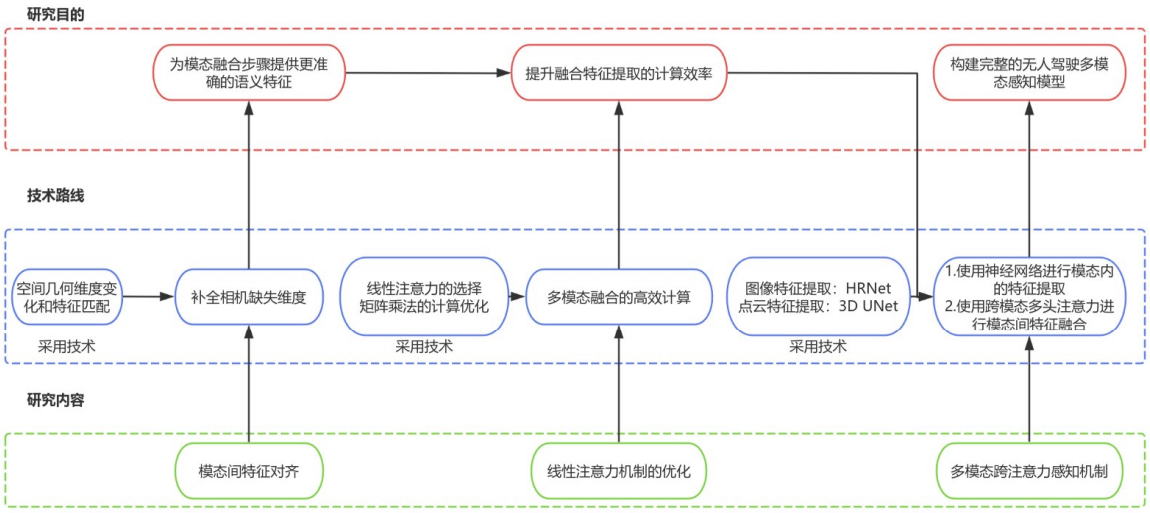

图 5: 改进工作

4 改进效果展示与对比

4.1 效果展示

这里给出四份输入输出样例以对分割效果做可视化演示。三份样例均随机抽取自公开数据集SemanticKITTI [16]，这是一个大规模的真实场景LiDAR点云序列逐点注释的数据集，标注了28类语义，共计22个sequences与43000 scans。

SemanticKITTI采用不同的颜色以对不同的语义分类进行可视化，如下图所示，本文将采取同样的可视化方案。

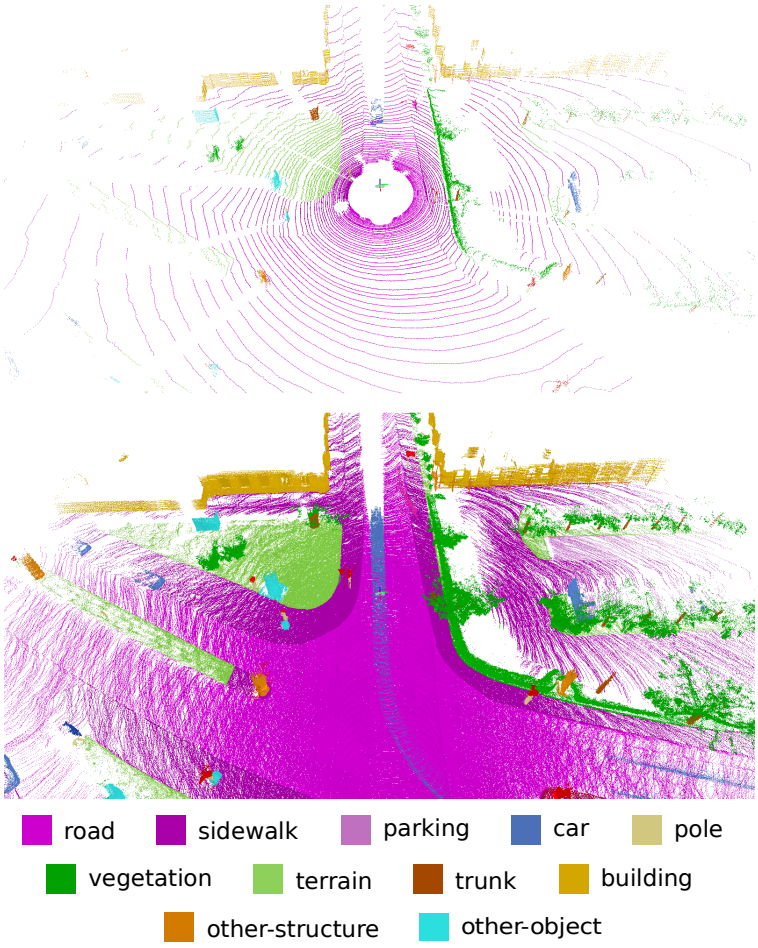

图 6: SemanticKITTI语义分割

样例一

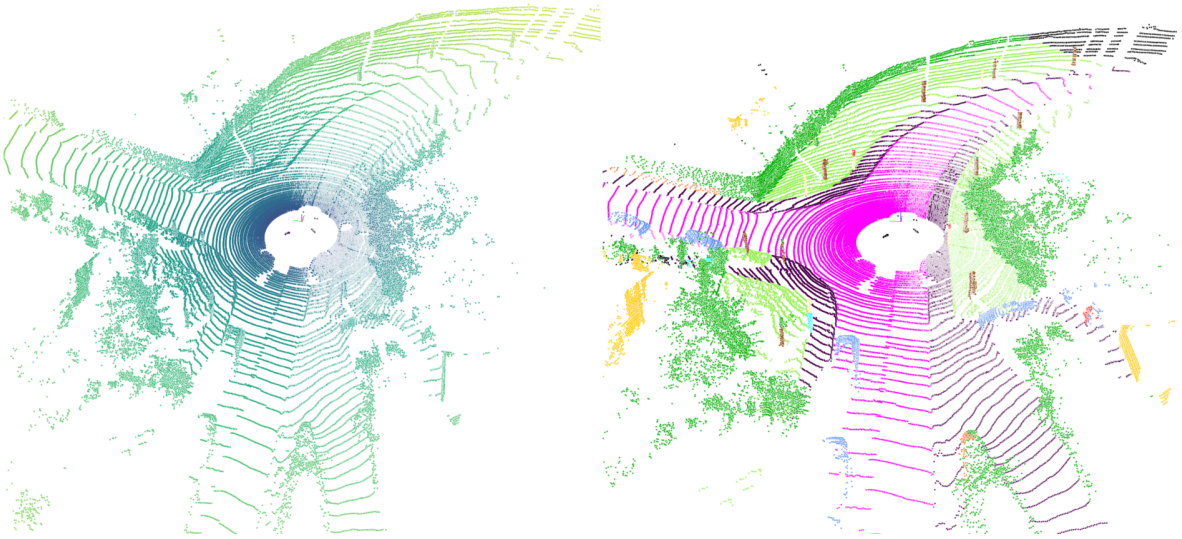

图 7: 可视化案例(i)

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

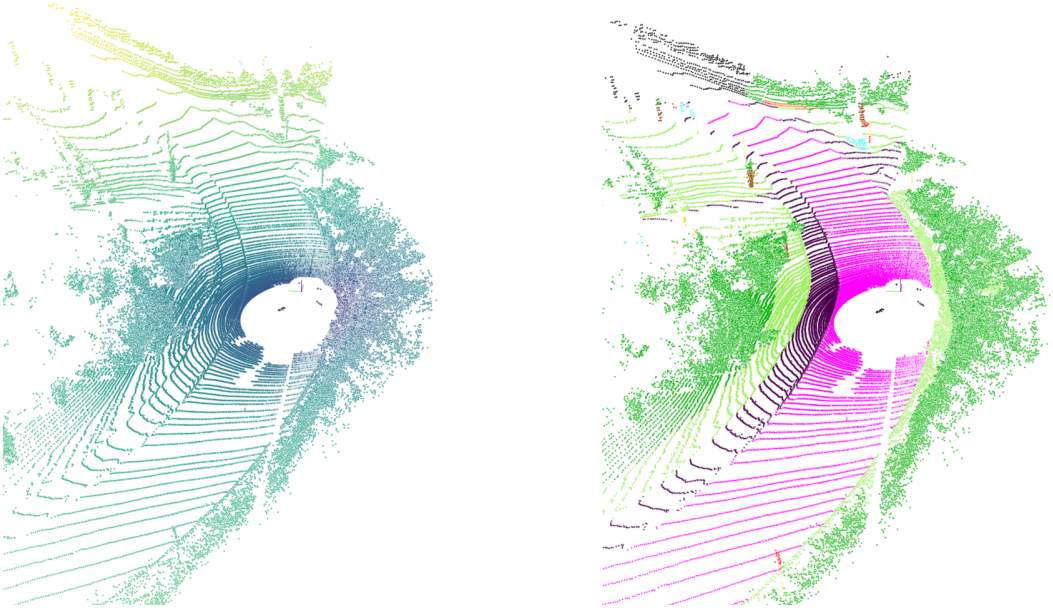

图 8: 可视化案例(ii)

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

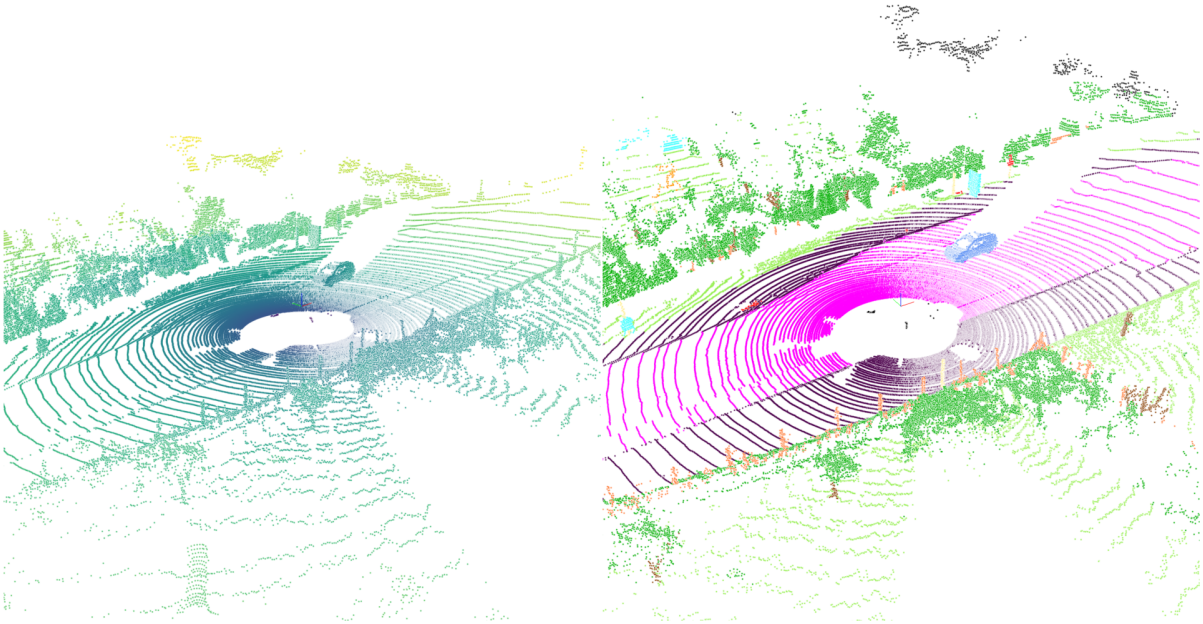

图 9: 可视化案例(iii)

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。

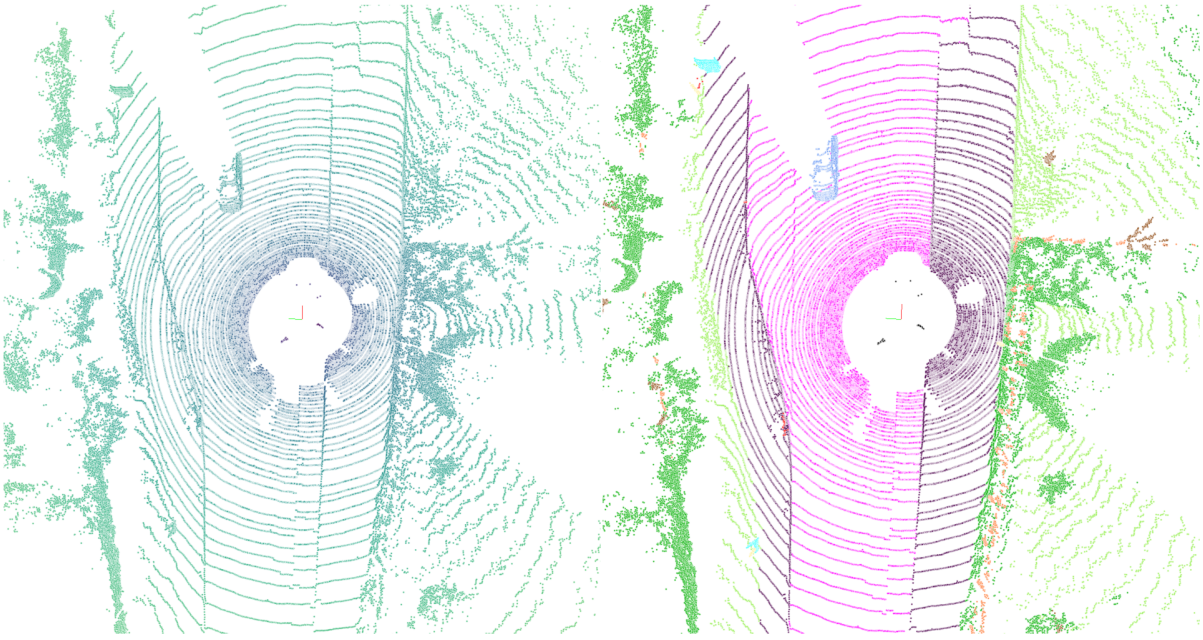

图 10: 可视化案例(iv)

这是俯视图，可以看出模型的识别是足够准确的。

4.2 横向对比

下为本文改进后的的新模型与包括MSeg3D在内的三个高性能模型间的性能评估测试中具体指标的对比表格，评估所用数据源自公开数据集Waymo与SemanticKITTI：

表 1: 性能评估对比

Methods	Modality	mIoU	Barrier	Bicycle	Bus	Car	C-Vehicle	Motorcycle	Pedestrian	Traffic Cone	Trailer	Truck	D-Surface	Other	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
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

可以看出，MSeg3D相较于先前存在的若干主流模型具有巨大优势，在全部的16项评估指标中，有9项都从13个模型中脱颖而出，达到了业内最先进水准，有4项仅以较小差距稍逊于其他模型。

经本文改进过后的模型，即表中的“Ours”，评估表现可被总结为：

- 有3项均明显优于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%）；
- 有7项几乎与MSeg3D没有差别，差距均在0.5%以内；
- 只有3项与MSeg3D存在较显著距离，但仍位于或超过业内平均水平；
- 按业内通用的语义分割综合评价指标mIoU（均交并比），记 X_i 为样本是否为第 i 类的真实标签、Y_i 为样本被判别是否为第 i 类的标签，TP_i 为正确判别第 i 类元素类别的次数（真阳性数，$|X_i \cap Y_i|$ ）、FP_i 为错将其他类别元素判别为第 i 类的次数（假阳性数，$|X_i \cap \overline{Y_i}|$ ）、TN_i 为正确判别非第 i 类元素不属于第 i 类的次数（真阴性数，$|\overline{X_i} \cap Y_i|$ ）、FN_i 为错将第 i 类元素判别为其他类别的次数（假阴性数，$|\Omega_i \cap \overline{X_i \cup Y_i}|$ ），p_{ij} 为总混淆矩阵的第 i 行、第 j 列的元素（即模型将第 i 类元素判断为第 j 类的次数），

$$
\begin{aligned}
mIoU &= \hat{\underset{0\leqslant i\leqslant k}{\mathbb{E}}} \frac{|X_i \cap Y_i|}{|X_i \cup Y_i|} \\
&= \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP_i}{FP_i + FN_i + TP_i} \\
&= \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{k=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}
\end{aligned}
$$

在mIoU评估标准下，改进后的模型具有业界最先进水平，甚至超越了原模型MSeg3D。以上均发生在大大降低了时间复杂度的前提下。

考虑到本文提出的改进模型相较于MSeg3D已经节约了不少的计算量，评估中能取得这样的成绩已属难得的成果，甚至在某些意义上超过了MSeg3D。总的来说，**改进后的模型比MSeg3D仅需要更少的算力，却达到了更佳的效果。**

5 不足与展望

目前，本文提出的改进模型已在实验中取得了优秀的效果，但受限于设备、条件与资源所困，本文没有对模型进行更广泛的测试乃至考察其在异常复杂的实际应用场景下的表现。

改进模型中可能存在的问题与后续可能的再改进方向有：

(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). 本文提出的降维操作有更多的模型可以验证使用，有助于提升多模态融合效率。

(2). 本文提出的注意力机制有更多的模型可以验证使用，有助于提升深层特征提取效率。

本文认为，完全有理由相信改进模型具有极大的潜力；即使是在一些计算能力稍差的设备上，模型也能顺利运行并得到理想效果。

参考文献

[1] Li, Jiale, Hang Dai, Hao Han and Yong Ding. 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, Yen-Chun, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng and Jingjing Liu. UNITER: Learning UNiversal Image-TExt Representations[J]. arXiv preprint arXiv:1909.11740,2019.

[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：

不知道应该讲些什么，仿佛昨天我还是半只脚刚踏入校门的小孩，只一刹那就要身着学士服、拍着毕业照试图抓住在西大的最后一刻了。

就这样毕业了，西大四年、一路走来，总的来说还是挺开心的，喜欢西大的自由散漫，更喜欢本科四年里认识的大家。

这四年，有诸多想做的事，可惜大多都停留在“想”的阶段，并没有真正去“做”。一回首，似乎除了考研，没有太多成功，或许平平淡淡就是世事常态吧。

接下来我将会在重庆大学继续学业，攻读硕士。我知道，路途漫且长，我想我们都应该好好地享受每一分每一秒。

最后，太飘渺的愿望都难以实现，我只有一个心愿：望家人们、朋友们、老师们都健健康康的吧。

余乐扬

2024 年 4 月 15 日 11:31

于西南大学 李园 一舍