

UniLION：基于线性群RNN的统一自动驾驶模型

刘哲^{1,3}、侯静华¹、叶晓青⁴、王景东⁴、IEEE会士、赵恒爽^{3,†}、IEEE会员、白翔^{2,†}、IEEE高级会员

摘要——尽管Transformer在多个领域展现出卓越能力，但其二次注意力机制在处理长序列数据时会引入显著的计算开销。本文提出了一种统一的自动驾驶模型UniLION，该模型基于线性群RNN算子（即对分组特征进行线性RNN），能高效处理大规模激光雷达点云、高分辨率多视角图像甚至时序序列。值得注意的是，UniLION作为单一通用架构，无需显式时间或多模态融合模块即可无缝支持多种专用变体（即纯激光雷达、时序激光雷达、多模态及多模态时序融合配置）。此外，UniLION在包括3D感知（例如3D目标检测、3D目标跟踪、3D空间预测、BEV地图分割）、预测（例如运动预测）和规划（例如端到端规划等）在内的广泛核心任务中，始终保持具有竞争力甚至处于前沿水平的性能表现。这种统一范式在保持卓越性能的同时，自然简化了多模态多任务自动驾驶系统的设计。我们期待UniLION能为自动驾驶领域3D基础模型的开发提供全新视角。相关代码可通过<https://github.com/happinesslz/UniLION>获取。

索引词——统一模型、线性群循环神经网络、自动驾驶、三维感知、运动预测、规划

arXiv:2511.01768v1 cs.CV 2025年11月3日 []

1 介绍

IN自动驾驶技术，能高效处理来自多视角摄像头的海量异构传感器数据，在空间和时间维度上整合激光雷达（LiDAR）传感器数据，对于实现复杂驾驶场景中的稳健感知、预测乃至规划至关重要。

如图1(a)所示，对于空间多模态融合，经典方法[1][2], [3]通过逐点或逐体素对齐，建立多视角图像与激光雷达点云之间的明确几何对应关系，以实现有效的跨模态信息交互。最近，基于BEV的融合方法[4][5]通过将异构传感器特征转换至鸟瞰视图（BEV）空间，并采用统一的空间表征策略，随后通过拼接或注意力机制融合多模态BEV表征。在时间建模方面，近期研究[6][7], [8], [9], [10], [11], [12]通过几何对齐BEV特征跨帧或采用基于注意力的查询融合来整合时间信息。因此，现有方法通常需要专门的多模态和时间融合模块，导致系统架构更为复杂。

此外，如图1(b)所示，处理诸如三维物体检测、运动预测和规划等多样化任务时，通常需要复杂的模块间依赖关系。

¹华中科技大学电子与信息学院，中国武汉。

²华中科技大学软件学院，中国武汉。

³香港大学计算机科学系，中国香港特别行政区（HKU）。

⁴中国北京百度公司。

[†]通讯作者：赵恒爽（电子邮箱：hszhao@cs.hku.hk）与白翔（电子邮箱：xbai@hust.edu.cn）。

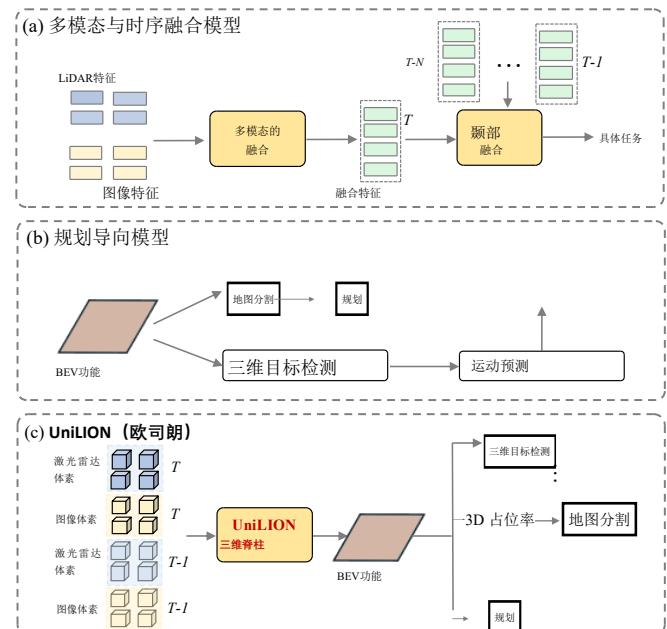


图1. (a)展示了实现多模态融合或时间序列融合的主流方法。(b)呈现了构建端到端自动驾驶系统的经典流程。(c)展示我们的UniLION方法，该方法巧妙地将多种输入模态和时间序列整合为单一通用架构。UniLION无需显式的时间或多模态融合模块，即可无缝支持多种专业配置（例如纯激光雷达、时间序列激光雷达、多模态及多模态时序融合方案）。此外，UniLION通过共享BEV特征表示，借助其3D主干网络强大的特征提取能力，实现了多个下游任务的解耦并行执行。

更好的性能。例如，UniAD[13]和VAD[14]采用具有复杂任务相互依赖性的顺序模块连接，其中下游任务依赖于上游模块的高维潜在特征，这可能导致误差传播和优化挑战。基于UniAD，FusionAD[15]通过额外的跨模态注意力机制和特征对齐模块进一步整合了相机-LiDAR融合，但该框架仍依赖于增加系统复杂性的专用融合架构。为避免顺序依赖，PARA-Drive[16]采用基于共享BEV特征的多任务并行架构，但主要侧重于规划导向的优化，导致在基础3D感知和预测任务上的表现较差。总之，这些方法要么因专用融合模块和顺序依赖性而面临架构复杂性问题，要么在不同任务间牺牲了性能平衡。

虽然Transformer凭借其灵活的注意力机制展现出构建统一框架的潜力，但在处理自动驾驶场景中常见的长序列数据时——例如包含数十万个点的密集点云、延展的时间序列甚至多模态数据——其二次复杂度会变得难以承受。这引发了一个关键问题：我们能否设计出一种统一的3D骨干网络，无需任何显式融合模块，就能在跨任务场景中以可接受的计算成本，无缝处理不同模态和时间信息？

线性循环神经网络（RNN）提供了一个极具吸引力的解决方案。其核心优势在于序列长度的线性复杂度，这与基于Transformer的注意力机制的二次复杂度形成鲜明对比。更重要的是，这种计算优势开辟了全新可能：线性复杂度使得不同模态和时间帧的标记可以直接拼接成单一扩展序列，既能实现跨模态与时间维度的深度交互，又无需人工设计融合方案，就能自动学习互补关系。

为实现这一目标，我们提出UniLION框架——一个通过基于线性RNN的三维主干网络处理多模态与时序信息的统一架构（图1(c)）。该方法本质上将范式从显式多模态与时序融合转向类似大语言模型（LLMs）的隐式统一表征学习。具体而言，异构传感器流（即多视角图像、激光雷达点云及时序序列）可通过直接的标记级（即将每个体素视为标记）拼接实现有效统一，并由UniLION的统一三维主干网络进行协同处理，从而无需依赖专用融合架构。

依托统一3D骨架的强大长距离建模能力，UniLION生成的BEV特征集虽精简却全面，通过并行多任务学习可同时支持多项下游任务。该方法消除了顺序依赖关系，同时在感知、预测和规划方面保持了具有竞争力的性能。

最后但同样重要的是，UniLION可作为单一通用架构，能够无缝适配多种

我们开发的UniLION系统无需依赖显式的时间/多模态融合模块，即可处理多种专业场景（例如纯激光雷达、时序激光雷达、多模态及多模态时序场景）。这意味着在完成多模态时序数据训练后，该系统在推理阶段即可直接应用于纯激光雷达、时序激光雷达或多模态场景，不仅能在不同传感器环境下稳定运行，还能在安全关键型应用中实现容错部署。

总之，基于我们先前的会议成果LION[17]（用于基于线性RNN的3D目标检测），该研究引入了3D空间特征描述符以增强局部空间信息捕捉，并采用体素生成策略来增强前景特征密度。本次扩展工作主要贡献如下：

- **统一异构输入：**UniLION基于线性群RNN的长距离建模优势和线性计算特性，通过直接拼接标记的方式，将多视角图像、激光雷达点云和时间信息整合到统一的3D主干网络中，省去了人工设计的融合模块，提供了一种更优雅且可扩展的解决方案。
- **统一模型：**UniLION支持跨不同输入格式的参数共享。具体而言，该模型在完成多模态时间序列数据训练后，无需重新训练即可直接部署于各类传感器配置和时间场景（例如单独激光雷达、时间序列激光雷达或多模态融合），展现出卓越的适应性，能灵活应对多样化运行环境。
- **统一输出表示：**UniLION能高效压缩异构多模态与时序信息，生成紧凑的BEV特征表示。该特征作为通用模块，通过并行多任务学习同时支持多种自动驾驶任务，既消除任务间的顺序依赖，又在感知、预测和规划任务中保持优异性能。
- **卓越性能：**UniLION在全面的自动驾驶任务中实现了具有竞争力的先进性能，包括3D感知（例如3D目标检测、跟踪、占用预测、BEV地图分割）、运动预测和端到端规划，证明了我们统一方法的泛化能力和有效性。

2 相关工作

线性RNN。循环神经网络（RNNs）最初是为解决自然语言处理（NLP）中的序列建模问题而开发的，例如时间序列预测和语音识别，通过有效捕捉序列数据中的时序依赖关系。RNNs的一个关键优势在于其处理序列特征时的线性计算复杂度，这与基于注意力机制相比，能显著降低处理长序列时的计算成本。近年来，

研究人员开发了先进的可并行化时间数据依赖的循环神经网络（本文中称为线性循环神经网络），以克服Transformer架构中固有的二次计算复杂度[18]。[19], [20], [21], [22], [23], [24], [25], [26], [27]这些现代线性RNN变体在保持理想线性复杂度的同时，实现了高效的并行训练，使其在各类任务中的表现可与Transformer相媲美甚至超越。基于这些进展，大量研究[17][28], [29], [30], [31]已探索线性RNN算子在多种二维和三维计算机视觉应用中的适应性。特别是在大规模户外三维场景中，线性循环神经网络（RNN）相较于基于Transformer的方法，在实现长距离建模时展现出更优的性能，且计算开销更低，从而显著提升了自动驾驶感知任务的性能。

多模态时间融合。多模态融合[2], [3], [4], [32], [33], [34], [35]以及时间融合[6][8], [9], [11], [12], [36], [37]是提升自动驾驶性能与鲁棒性的关键技术。对于多模态融合，现有方法可分为两大范式。第一种范式[4], [5]将点云和图像特征统一转化为BEV表示，通过BEV空间中的特征级整合实现多模态融合。第二种范式[1], [2], [33], [34], [38]该系统采用基于投影的交互机制，将激光雷达的点特征或体素特征投影至多视角图像，以促进跨模态特征交互。此外，时间融合技术通过跨时间步的丰富上下文信息，有效提升特征表征质量。早期方法[8][36]将历史与当前输入点云进行拼接以实现时间整合。较新方法[6][9]在特征层面进行时间融合（例如，BEV特征与查询特征）。为解决反复提取历史特征的计算开销问题，流式时间融合方法[11][12], [37]已被提出用于以流式方式实现对扩展序列的高效时序整合。相比之下，我们的UniLION优雅地将多种输入模态和时序序列统一到一个单一的多功能架构中。UniLION无需显式时序或多模态融合模块，即可无缝适配多种专用配置（即纯激光雷达、时序激光雷达、多模态及多模态时序融合变体）。**三维感知。**三维感知是后续预测与规划任务的基础。从目标视角来看，包括三维目标检测、三维多目标跟踪、BEV地图分割及三维空间占用预测任务。对于三维检测，基于点的方法采用原始点[39] [40], [41], [42], [43], [44], [45]作为输入并实现点网络[46][47]以获取精细的几何信息。基于体素的方法[36], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62]将输入的不规则点云量化为规则的3D体素以进行特征提取，随后将3D特征转换为BEV（鸟瞰视图）特征用于3D检测。对于3D多目标跟踪，部分方法[36][63], [64], [65]采用基于检测结果的追踪-检测范式来跟踪目标，以及其他方法[66][67]采用末端-

采用端到端范式联合优化检测与跟踪。对于BEV地图分割，现有方法[4], [9]对BEV特征进行二维卷积以预测分割掩膜。对于三维占用预测，部分方法[68][69]将多视角图像特征提升为三维体积特征，随后对三维体积特征进行三维卷积运算以预测每个体素的结果。

运动预测与规划。运动预测与规划任务分别涉及对周围物体和自车未来轨迹的预测。当前自动驾驶系统的研究可大致分为两种架构范式。第一种范式采用模块化架构[13], [14], [67], [70], [71]该方法将自动驾驶流程分解为多个组件，其中规划模块直接依赖运动预测输出。第二种范式采用并行架构[16]，通过基于鸟瞰图（BEV）表征的并行处理架构，有效缓解了累积误差问题。

多任务学习。许多研究致力于通过多任务学习实现单个模型处理更多任务（例如三维目标检测、跟踪、BEV地图分割和三维占用预测）。为融合三维检测与BEV地图分割任务，BEVFusion[4]采用独立的BEV主干网络以实现两者的平衡。当将三维检测与三维占用预测结合时，PanoOcc[69]通过共享占用表示统一特征学习与场景表征。对于涉及三维检测、BEV地图分割和三维占用预测的更复杂多任务场景，M3Net[72]引入任务导向的通道缩放机制以缓解联合优化过程中的梯度冲突。PARA-Drive[16]采用基于共享BEV特征的并行多任务架构来处理各类自动驾驶任务。然而其主要聚焦于规划导向的优化，导致在基础感知与预测任务上的表现欠佳。与之形成鲜明对比的是，本文提出的UniLION是一个综合性统一框架，能够同时处理自动驾驶任务的完整谱系（即三维感知、预测与规划）。值得注意的是，我们的方法无需额外的任务特定设计即可实现这种统一，仅通过动态多任务损失机制（详见第3.6节）达成。尽管结构简单，UniLION在所有评估任务中始终展现出与专业单任务模型相当甚至更优的性能表现。

3 方法

3.1 概述

本文提出一种基于线性群RNN的简单有效窗口统一框架（即对分组特征进行线性RNN），命名为UniLION，该框架可对数千个体素进行分组（比先前方法的数量多数十倍[60]）。[61], [73])用于特征交互。此外，UniLION可直接处理时序多模态体素，无需任何附加融合模块即可实现时序融合与多模态融合。本文展示了UniLION的流程。

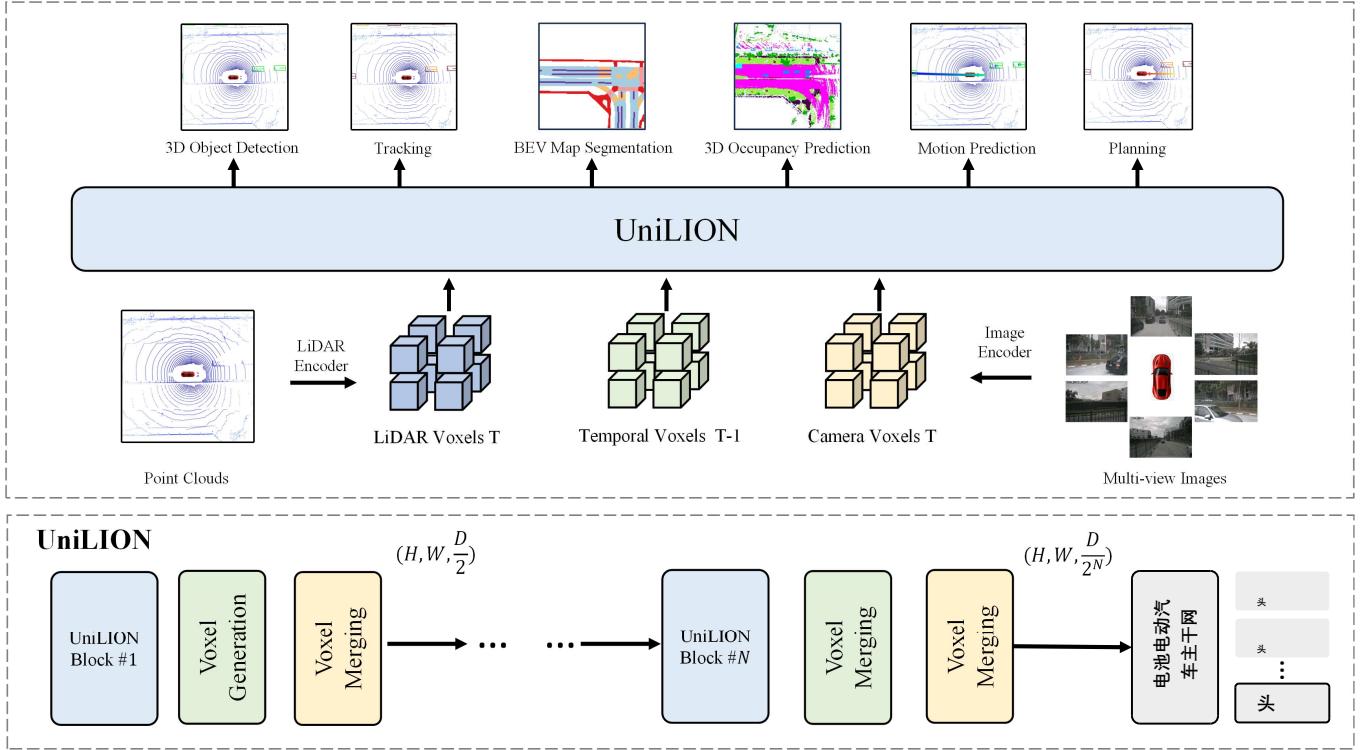


图2. 我们提出UniLION，这是一种通过线性群 RNN 在UniLION主干中实现潜在时间融合与多模态融合的统一模型，生成服务于所有自动驾驶任务（包括感知、预测和规划）的统一BEV特征。UniLION主要由 N 个UniLION模块组成，每个模块配有一个用于特征增强的体素生成器和一个用于沿高度维度下采样特征的体素合并器。 (H, W, D) 表示3D特征图的形状，其中 H 、 W 和 D 分别表示3D特征图沿X轴、Y轴和Z轴的长度、宽度和高度。 N 是UniLION模块的数量。在UniLION中，我们首先将输入的多模态体素分割为一系列等大小的组。然后，我们将这些分组特征输入UniLION 3D主干以增强其特征表示。最后，这些增强后的特征被输入BEV主干，为所有任务生成统一的BEV特征。

如图2所示。UniLION由激光雷达编码器、图像编码器、统一3D主干网络、BEV主干网络以及针对不同任务的各类任务头组成。本文的核心贡献在于基于线性群 RNN 的统一3D主干网络，该网络支持多模态与时间融合。下文将详细阐述UniLION的整体架构。

3.2 激光雷达与图像编码器

UniLION融合了激光雷达编码器和图像编码器，分别从点云和多视角图像中提取激光雷达体素和相机体素。对于激光雷达编码器，我们通过动态体素化将点云转换为体素，随后经过两个线性层生成激光雷达体素特征。对于图像编码器，我们利用成熟的视觉图像骨干网络（例如ResNet-50、SwinTiny）提取多视角图像特征。为将这些二维图像特征投影到三维空间，我们采用由三个二维卷积层组成的轻量级深度估计分支来预测像素级深度值。具体而言，我们根据深度估计置信度选择前 K 个深度候选（默认设置 $K=4$ ），然后将这些深度候选与相机矩阵结合，在统一的三维坐标系中生成相机体素。为解决多个相机体素占据相同三维位置的空间冲突，我们通过在每个空间位置对特征进行元素级求和来合并重复的相机体素。最终，提取的激光雷达体素和

摄像机体素沿体素维度直接拼接后输入UniLION三维主干网络。

3.3 三维稀疏窗口划分

UniLION采用三维稀疏窗口划分来对输入体素进行分组以实现特征交互。具体而言，我们首先将输入体素划分为形状为 (S_x, S_y, S_z) 的非重叠三维窗口，其中 S_x 、 S_y 和 S_z 分别表示窗口沿X轴、Y轴和Z轴的长度、宽度和高度。接着，我们分别沿X轴对体素进行X轴窗口划分，沿Y轴对体素进行Y轴窗口划分。最后，为降低计算成本，我们将已排序的体素按等大小分组 G ，而非按等形状窗口进行特征交互。由于Transformer的二次计算复杂度，先前基于Transformer的方法[60] [61]，[73] 仅能通过小规模组实现特征交互。相比之下，我们采用更大规模的组 G ，借助线性群 RNN 算子的线性计算复杂度，从而实现长程特征交互。

3.4 UniLION块

UniLION模块是我们方法的核心组件，该模块包含用于长程特征交互的UniLION层，以及用于捕捉局部特征的三维空间特征描述符。

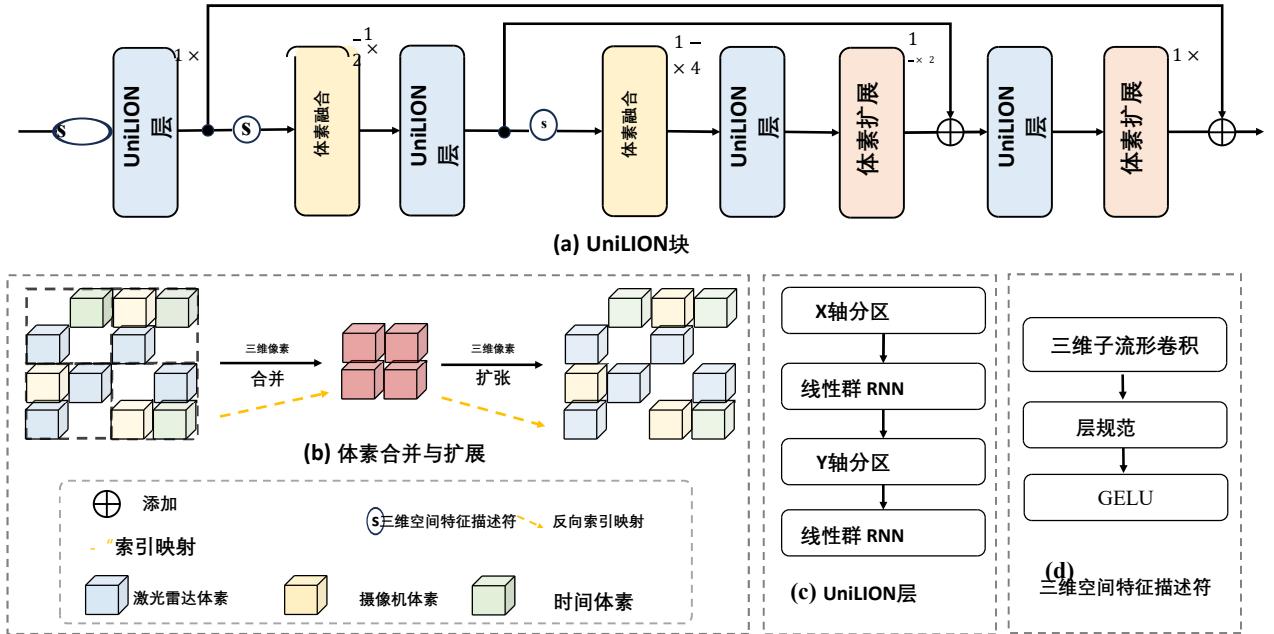


图3.(a)展示了UniLION模块的结构，该模块包含四个UniLION层、两次体素合并操作、两次体素扩展操作以及三个三维空间特征描述符。此处 $1 \times$ ， $\frac{1}{2} \times$ ，以及 $\frac{1}{4} \times$ 分别表示三维特征图的分辨率 (H, W, D) 、 $(H/2, W/2, D/2)$ 和 $(H/4, W/4, D/4)$ 。(b)展示了体素下采样时的体素合并与上采样时的体素扩展过程。我们通过体素合并将输入的LiDAR体素、相机体素和时间体素融合，实现多模态融合与时间融合。(c)展示了UniLION层的结构。(d)展示了三维空间特征描述符的细节。

如图3(a)所示，3D空间信息、用于特征降采样的体素合并以及用于特征升采样的体素扩展。此外，UniLION模块采用分层结构以更好地提取多尺度特征。接下来，我们将介绍UniLION模块的各个组成部分。

体素合并与体素扩展。为使网络能够获取多尺度特征，我们的UniLION采用分层特征提取结构。为此，需要在高度稀疏的点云中执行特征下采样和上采样操作。但需注意的是，由于三维点云具有不规则数据格式，我们无法像处理二维图像那样简单地应用最大池化、平均池化或上采样操作。因此，如图3(b)所示，我们在高度稀疏点云中采用体素合并进行特征下采样，体素扩展进行特征上采样。具体而言，体素合并通过计算下采样索引映射来合并体素；体素扩展则通过对应的逆索引映射对下采样体素进行上采样。

UniLION层。在UniLION模块中，我们通过线性群RNN算子对分组特征进行建模，以建立特征间的长距离关联。具体而言，如图3(c)所示，UniLION层由两个线性群RNN算子构成：第一个算子基于X轴窗口划分实现长距离特征交互，第二个算子则通过Y轴窗口划分提取长距离特征信息。通过这种双窗口划分机制，UniLION层能获取更充分的特征交互，从而生成更具区分度的特征表示。

三维空间特征描述符。尽管线性循环神经网络具有

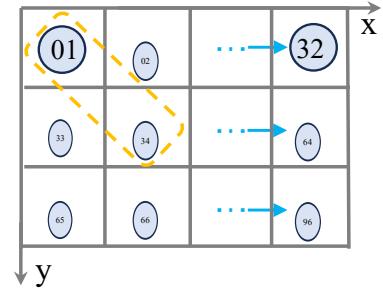


图4.一维序列格化时空间信息丢失的示意图。例如，空间位置上相邻的两个体素（索引为01和34）在一维序列中相距较远。

长距离建模具有计算成本低的优势，但不可忽视的是，当输入体素特征被展平为一维序列特征时，空间信息可能会丢失。例如，如图4所示，三维空间中有两个相邻特征（即索引为01和34）。然而，当它们被展平为一维序列特征后，它们在一维空间中的距离变得非常遥远。我们将这种现象视为三维空间信息的丢失。为解决此问题，一种可行的方法是增加体素特征的扫描顺序数量，例如VMamba[74]和Vim[75]。然而，扫描顺序过于手工设计。此外，随着扫描顺序的增加，相应的计算成本也会显著上升。因此，在大规模稀疏三维点云中采用这种方法并不合适。如图3(d)所示，我们引入了一种三维空间特征描述符，它由三维子流形卷积-

该模型包含一个归一化层 (LayerNorm) 和一个 GELU 激活函数。我们利用三维空间特征描述符为UniLION层提供丰富的三维局部位置感知信息。此外，将三维空间特征描述符置于体素合并之前，以减少体素合并过程中的空间信息丢失。

自回归体素生成。尽管我们在UniLION主干网络中使用相机体素和时间体素来补偿LiDAR体素，但在图3中实现体素融合仍存在潜在信息丢失的挑战。因此，我们提出一种利用线性群 RNN 自回归能力的体素生成策略来解决这些问题。具体而言，为方便起见，我们将选定前景体素特征 F_m 的对应坐标定义为 P_m 。如图5所示，我们首先通过沿X轴、Y轴和Z轴分别扩散 P_m 四个不同偏移量（即[-1, -1, 0]、[1, 1, 0]、[1, -1, 0]和[-1, 1, 0]）来获取扩散体素。随后，我们将扩散体素的对应特征初始化为全零。接着，将第 i 个UniLION模块的输出特征 F_i 与初始化后的体素特征拼接，再输入后续的第 $(i+1)$ 个UniLION模块。最后，得益于UniLION模块的自回归特性，基于大组内其他体素特征可有效生成扩散体素特征。该过程可表述为：

$$F_p = F_i \oplus F_{[-1, -1, 0]} \oplus F_{[1, 1, 0]} \oplus F_{[1, -1, 0]} \oplus F_{[-1, 1, 0]}, \quad (1)$$

$$F_p = \text{块}(F_p), \quad (2)$$

其中 $F_{[x, y, z]}$ 表示沿X轴、Y轴和Z轴具有 x 、 y 和 z 扩散偏移的初始化体素特征。 \oplus 和Block分别表示拼接和UniLION块。

3.5 统一特征表示

先前的方法[4], [6], [7], [10] 通常需要额外设计模块来实现多模态或时间融合。而我们的目标是将所有激光雷达体素、相机体素甚至时间体素直接输入UniLION统一的3D主干网络，无需额外的多模态或时间融合模块。得益于UniLION 3D 主干网络在长距离建模中的强大表征能力，我们能够自适应地建模激光雷达体素、相机体素和时间体素之间的关系。

多模态特征学习。在自动驾驶场景中，激光雷达点云与相机图像具有高度互补性——激光雷达提供精确的几何结构，而相机图像则贡献丰富的语义外观信息。因此，UniLION致力于通过统一的3D主干网络有效整合这些异构模态，借助线性循环神经网络强大的长距离建模能力实现相互增强。具体来说，给定输入点云和多视角图像，我们首先将点云量化为体素，并采用体素特征编码器 (VFE) 提取这些体素，获得激光雷达体素 $V_l \in RL_l \times C$ 。对于多视角图像，我们首先采用图像骨干网络提取多视角图像特征。随后，我们采用深度网络预测多视角深度图和

将图像特征转换为基于预测深度和数据集提供的校准矩阵的相机体素 $V_c \in RL_c \times C$ 。其中， C 、 L_l 和 L_c 分别表示特征通道、激光雷达体素数量和相机体素数量。随后我们将激光雷达体素 V_l 与相机体素 V_c 连接，得到多模态体素 $V_m \in R^{(L_l + L_c) \times C}$ 。值得注意的是，由于三维空间中的空间位置可能同时被激光雷达和相机体素占据，我们采用提出的体素合并策略来合并重叠的多模态体素。最后，我们将合并后的多模态体素 $V'm \in RLm \times C$ 直接输入UniLION 3D主干网络，以进一步在三维空间中提取多模态特征，其中 L_m 表示合并后的多模态体素数量。**时间特征学习。**时间信息提取对于自动驾驶系统中的精确运动预测和轨迹规划至关重要。因此，UniLION进一步旨在借助线性RNN强大的长程建模能力，将时间信息整合到我们的统一3D骨架中。具体而言，给定当前帧的多模态体素 $V_T \in RLT \times C$ ，当可用时，我们从时间记忆库中获取历史多模态体素 $V_{T-1} \in RLT_{-1} \times C$ 。为确保时间帧间的空间一致性，我们通过数据集提供的变换矩阵进行空间对齐，将历史体素转换为当前帧的坐标系。随后，我们将 V_{T-1} 与当前体素 V_T 拼接，构建时间体素 $V_p \in R^{(LT-1+LT) \times C}$ 。此处 L_{T-1} 和 L_T 分别表示 $T-1$ 帧与 T 帧中的体素数量。同样地，我们采用体素合并策略来合并时间序列体素，因为多个体素可能在时间帧中占据相同的三维位置。最后，我们将合并后的体素直接输入UniLION 3D主干网络，以自适应地学习时间信息。

3.6 动态多任务损失

作为统一模型，UniLION通过整合点云、多视角图像和历史信息，生成用于自动驾驶感知、预测与规划的统一BEV表征。基于紧凑的BEV特征，我们部署任务专用头以同时输出各任务结果。得益于模块化并行架构，UniLION可选择性执行不同任务以降低推理时的计算开销。然而在多任务训练中，需考虑多任务平衡问题。为尽可能保持各任务性能，我们采用动态损失平衡策略。具体而言，给定检测损失 \mathcal{L}_{det} 、占用损失 \mathcal{L}_{occ} 、BEV地图分割损失 \mathcal{L}_{map} 、运动预测损失 \mathcal{L}_{mot} 和规划损失 \mathcal{L}_{plan} ，我们计算动态损失权重以使各任务损失 \mathcal{L}_{task} 与 \mathcal{L}_{det} 对齐：

$$w_{task} = \frac{\mathcal{L}_{det}}{\mathcal{L}_{task} + 1e^{-5}} \quad (3)$$

最终损失可表述为：

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{det} + \lambda_2 \cdot w_{map} \cdot \mathcal{L}_{map} + \lambda_3 \cdot w_{occ} \cdot \mathcal{L}_{occ} + \lambda_4 \cdot L_{mot} + \lambda_5 \cdot L_{plan}, \text{ 其中 } \lambda_1, \lambda_2, \lambda_3, \lambda_4 \text{ 以及 } \lambda_5 \text{ 为损失权重。} \quad (4)$$

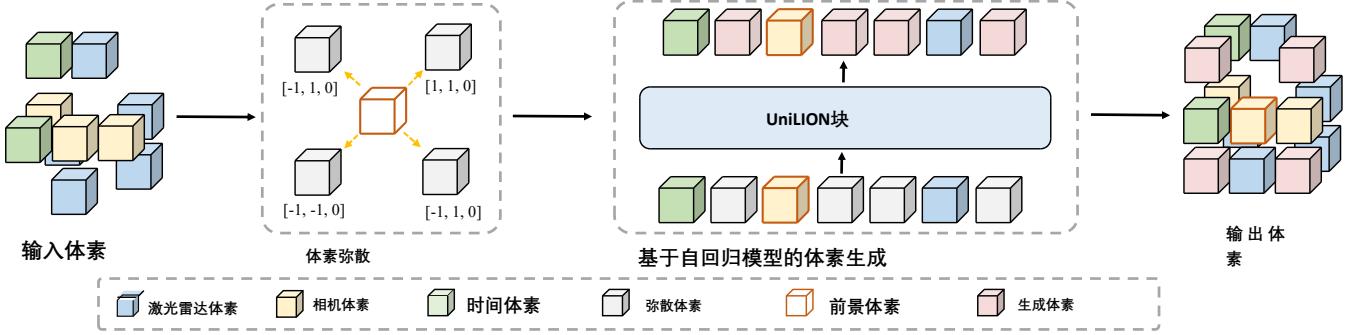


图5. 体素生成示意图。我们首先从激光雷达体素、相机体素和时间体素中筛选出前景体素，并沿不同方向进行扩散。随后，将扩散体素的对应特征初始化为零值，利用后续UniLION模块的自回归特性生成扩散特征。

4 实验

4.1 数据集与评估指标

数据集。我们在nuScenes[76]上进行实验，该数据集是公认的自动驾驶基准测试集，具有50米的感知范围，场景标注频率为2赫兹。数据集包含1000个场景，分为750个训练场景、150个验证场景和150个测试场景。nuScenes提供了全面的多模态数据，包括激光雷达生成的点云和来自6个环绕摄像头的多视角图像。该数据集支持涵盖感知、预测和规划领域的多样化自动驾驶任务，包括3D目标检测、多目标跟踪、BEV地图分割、3D空间占用预测、运动预测和轨迹规划。

评估指标。我们采用特定任务的指标，遵循既定协议：3D目标检测采用平均精度（mAP）和NuScenes检测分数（NDS）；跟踪性能采用AMOTA；BEV地图分割采用平均交并比（mIoU）[4]；3D占用预测采用RayIoU[68]；运动预测采用最小平均位移误差（minADE）[66]；规划评估采用L2距离与碰撞率[13]。

4.2 实施细节

网络细节。我们评估了UniLION的四种配置以展示其多功能性：仅LiDAR(L)、LiDAR-相机(LC)、仅LiDAR带时间融合(LT)以及LiDAR-相机带时间融合(LCT)。所有变体均采用单一统一的3D骨干架构，能够同时处理多项任务。我们将体素网格分辨率设置为(0.3m, 0.3m, 0.25m)，并使用N=4个UniLION模块，其窗口尺寸逐步细化： $(S_x, S_y, S_z) = (13, 13, 32), (13, 13, 16), (13, 13, 8), (13, 13, 4)$ 及对应的组大小G=4096, 2048, 1024, 512。在体素生成过程中，我们将采样率r=0.2设定为在精度与计算效率之间实现最佳平衡。对于相机处理，我们采用两种骨干网络配置：以输入分辨率384×1056的Swin-Tiny[77]作为基础骨干网络，以及ResNet-50[78]

作为分辨率 256×704 的轻量级骨干网络。两种图像骨干网络均在nuImages[76]上进行预训练。在时间建模方面，我们采用流式处理方式，依次处理4个连续帧，并将提取的特征依次输入我们的UniLION骨干网络。

任务特定实现细节。我们采用既定架构实现各任务以确保公平评估。对于3D目标检测，我们采用DSVT[61]和TransFusion[38]的检测头。对于多目标跟踪，我们采用CenterPoint[36]的关联策略。对于BEV地图分割，我们使用BEVFusion地图头[4]。对于3D占用预测，我们整合FlashOcc头[79]。对于运动预测，我们利用通过单个Transformer解码器层处理的检测查询，并采用SparseDrive[67]的六轨迹锚点。类似地，对于轨迹规划，我们采用单个Transformer解码器层来预测未来来自车轨迹。为确保所有任务间公平比较并防止信息泄露，我们在训练和评估过程中默认严格避免使用任何自车状态信息[80]。

训练过程细节。在训练阶段，我们采用标准数据增强策略，包括水平翻转、旋转、平移、缩放和真实样本采样[36], [88]。为增强模型鲁棒性并缓解过拟合问题，我们在训练UniLION时采用了多阶段训练策略，并精心设计了数据增强方案。在单帧训练阶段，我们以单帧激光雷达或激光雷达相机数据作为输入。首先，我们采用类别平衡分组采样(CBGS)[89]结合所有数据增强技术，对检测和地图分割任务进行12轮联合训练，确保早期训练阶段的特征学习具有鲁棒性。随后，我们通过整合检测、地图分割和占用预测任务，仅使用旋转和翻转增强进行24轮感知模型训练，因为真实占用标注对几何变换施加了约束。经过这些步骤后，我们即可获得UniLION的单帧感知模型（即激光雷达单独和激光雷达相机组合两种变体）。在时序训练阶段，我们加载相应单帧预训练感知模型的权重，并将多帧输入以流式方式输入模型。

表1 nuScenes验证集上感知、预测与规划的性能比较。 \dagger 表示使用自我状态信息。 $'L'$ 、 $'C'$ 和 $'T'$ 分别表示激光雷达输入、相机输入和时间信息输入。

方法	出席	模态	Dete $\uparrow\downarrow$	NDS $\uparrow\downarrow$	节mAP	跟踪 AMOTA \uparrow	地图 mIoU \uparrow	占用率 RayIoU \uparrow	运动minADE (汽车/行人) \downarrow	规划 L2 Col. \downarrow
BEVDet [81]	arXiv 22	C	37.9	29.8	-	-	-	-	-	-
BEVFusion-C [4]	ICRA 23	C	41.2	35.6	-	57.1	-	-	-	-
VIP3D [66]	CVPR 23	CT	-	-	21.7	-	-	-	2.05 / -	-
StreamPETR [11]	ICCV 23	CT	59.2	50.4	-	-	-	-	-	-
打开[37]	ECCV 24	CT	60.6	51.9	-	-	-	-	-	-
SparseOcc [68]	ECCV 24	CT	-	-	-	-	-	35.1	-	-
OPUS [82]	神经信息处理系统会议24	CT	-	-	-	-	-	41.2	-	-
UniAD [13]	CVPR 23	CT	49.8	38.0	35.9	-	-	-	0.71 / -	0.73 0.61
VAD [14]	ICCV 23	CT	-	-	-	-	-	-	0.72 / -	0.21
BEVFormer-C [70]	TPAMI 24	CT	51.7	41.6	40.5	-	-	-	-	0.99 0.70
SparseDrive [67]	ICRA 25	CT	58.8	49.6	50.1	-	-	-	0.60 / 0.72	0.61 0.10
CenterPoint [36]	CVPR 21	L	66.5	59.2	-	-	-	-	-	-
TransFusion-L [38]	CVPR 22	L	70.1	65.5	-	-	-	-	-	-
BEVFusion-L [4]	ICRA 23	L	69.3	64.7	-	48.6	-	-	-	-
SEED [83]	ECCV 24	L	71.2	66.6	-	-	-	-	-	-
HEDNet [84]	神经信息处理系统会议第23届	L	71.4	66.7	-	-	-	-	-	-
LION-Mamba [17]	神经信息处理系统会议24	L	72.1	68.0	-	-	-	-	-	-
3DMOTFormer [64]	ICCV 23	L	-	-	71.2	-	-	-	-	-
UniLION (欧司朗)	-	L	72.3	67.5	72.6	71.7	46.8	-	-	-
PnPNet [85]	CVPR 20	LT	-	-	-	-	-	-	1.15 / -	-
MGTANet [7]	AAAI 23	LT	68.7	62.9	-	-	-	-	-	-
QTNet [6]	神经信息处理系统会议第23届	LT	70.9	66.5	-	-	-	-	-	-
UniLION (欧司朗)	-	LT	73.0	68.9	73.3	72.4	49.6	-	0.58 / 0.39	0.60 0.27
TransFusion [38]	CVPR 22	LC	71.3	67.5	-	-	-	-	-	-
BEVFusion [5]	神经信息处理系统会议第22届	LC	71.0	67.9	-	-	-	-	-	-
DeepInteraction [3]	神经信息处理系统会议第22届	LC	72.6	69.9	-	-	-	-	-	-
UniTR-Det [86]	ICCV 23	LC	73.3	70.5	-	-	-	-	-	-
UniTR-Map [86]	ICCV 23	LC	-	-	-	74.7	-	-	-	-
BEVFusion [4]	ICRA 23	LC	71.4	68.5	-	63.0	-	-	-	-
CMT [87]	ICCV 23	LC	72.9	70.3	-	-	-	-	-	-
DAL-大 [88]	ECCV 24	LC	74.0	71.5	-	-	-	-	-	-
EagerMOT [65]	ICRA 21	LC	-	-	71.2	-	-	-	-	-
AlphaTrack [63]	IROS 21	LC	-	-	73.3	-	-	-	-	-
M3Net [72]	AAAI 25	LC	72.4	69.0	-	70.4	-	-	-	-
BEVFormer-M [70]	TPAMI 24	LC	73.2	71.2	-	-	-	-	-	-
DeepInteraction++ [71]	TPAMI 25	LC	73.3	70.6	-	-	-	-	-	-
UniLION (欧司朗)	-	LC	74.9	72.2	76.2	72.3	50.8	-	-	-
QTNet [6]	神经信息处理系统会议第23届	LCT	73.1	70.3	-	-	-	-	-	-
BEVFusion4D [10]	arXiv 23	LCT	73.5	72.0	-	-	-	-	-	-
FusionAD [15]	arXiv 23	LCT	64.6	57.4	50.1	-	-	-	0.39 / -	1.03 0.21
DeepInteraction++† [71]	TPAMI 25	LCT	66.0	55.7	-	-	-	-	0.34 / -	0.71 0.19
UniLION (欧司朗)	-	LCT	75.4	73.2	76.5	73.3	51.3	-	0.57 / 0.37	0.65 0.18

用于训练24个周期以生成时间感知模型的方式。随后，我们加载并冻结预训练的时间感知模型权重，无需任何数据增强即可训练36个周期的运动预测与规划任务，从而得到时间变体（即时间仅LiDAR模型和时间LiDAR-Camera模型）。 λ_1 、 λ_2 、 λ_3 、 λ_4 和 λ_5 分别设置为1、0.5、1、1和1。

推理过程细节。在推理阶段，与传统方法[4]不同，[10] 针对需要为不同输入模态或时间融合设计独立模型的场景，我们采用的UniLION单一模型通过多帧多模态输入，能够无缝支持多种专业变体（例如纯激光雷达、时间序列纯激光雷达、激光雷达-相机组合及时间序列激光雷达-相机组合模型），无需单独构建模型架构或设置显式的时间/多模态融合模块。这种统一范式在保持模型简洁性的同时，仍能在不同配置下实现高性能表现。

4.3 与最先进的方法的比较

总体结果。我们在nuScenes数据集上全面评估了UniLION的四个变体在六项自动驾驶任务中的表现：3D目标检测、跟踪、BEV地图分割、3D空间占用预测、运动预测和规划。需要注意的是，我们仅在整合时间输入时评估运动预测和规划任务。如表1所示，UniLION在单一统一模型中实现了所有任务的领先性能。对于纯激光雷达配置，UniLION在nuScenes验证集上取得72.3%的NDS和67.5%的mAP用于3D目标检测，72.6%的AMOTA用于跟踪，71.7%的mIoU用于地图分割，46.8%的RayIoU用于空间占用预测。当整合时间融合时，所有任务的性能进一步提升：检测任务达到73.0%的NDS和68.9%的mAP，跟踪任务为73.3%的AMOTA，地图分割任务为72.4%的mIoU，空间占用预测任务为49.6%的RayIoU，车辆运动预测任务为0.58的minADE，行人运动预测任务为0.39的minADE，

表2

在nuScenes验证集上的检测性能。‘T.L.’、‘C.V.’、‘Ped.’、‘M.T.’、‘T.C.’和‘B.R.’分别代表拖车、施工车辆、行人、机动车、交通锥和障碍物的缩写。

方法	出席	模态	NDS↑	mAP↑	汽车	卡车	总线	T.L.	C.V.	土壤自然结构体	M.T.	自行车	T.C.	B.R.
CenterPoint [36]	CVPR 21	L	66.5	59.2	84.9	57.4	70.7	38.1	16.9	85.1	59.0	42.0	69.8	68.3
VoxelNeXt [90]	CVPR 23	L	66.7	60.5	83.9	55.5	70.5	38.1	21.1	84.6	62.8	50.0	69.4	69.4
Uni3DETR [91]	神经信息处理系统会议第23届	L	68.5	61.7	—	—	—	—	—	—	—	—	—	—
TransFusion-L [38]		L	70.1	65.5	86.9	60.8	73.1	43.4	25.2	87.5	72.9	57.3	77.2	70.3
DSVT [61]	CVPR 23	L	71.1	66.4	87.4	62.6	75.9	42.1	25.3	88.2	74.8	58.7	77.9	71.0
HEDNet [84]	神经信息处理系统会议第23届	L	71.4	66.7	87.7	60.6	77.8	50.7	28.9	87.1	74.3	56.8	76.3	66.9
LION-Mamba [17]		L	72.1	68.0	87.9	64.9	77.6	44.4	28.5	89.6	75.6	59.4	80.8	71.6
UniLION (欧司朗)	—	L	72.3	67.5	88.2	63.8	78.0	45.7	28.4	89.1	75.7	56.0	79.3	70.3
MGTANet [7]	AAAI 23	LT	68.7	62.9	87.0	59.6	72.3	40.1	21.5	86.3	69.3	51.4	73.4	67.8
QTNet [38]	神经信息处理系统会议第23届	LT	70.9	66.5	87.2	61.5	75.8	43.0	25.7	87.8	75.5	61.5	75.4	71.4
UniLION (欧司朗)		LT	73.0	68.9	89.0	65.2	79.7	46.7	30.8	89.6	77.9	59.5	79.2	71.1
MVP [35]	神经信息处理系统会议21	LC	70.8	67.1	—	—	—	—	—	—	—	—	—	—
TransFusion [38]		LC	71.3	67.5	87.7	32.2	75.4	43.7	27.3	87.7	75.5	63.5	77.9	74.2
AutoAlignV2 [33]	ECCV 22	LC	71.2	67.1	—	—	—	—	—	—	—	—	—	—
BEVFusion [5]	神经信息处理系统会议第22届	LC	72.1	69.6	89.1	66.7	77.7	42.6	30.9	89.4	79.0	67.5	79.3	73.5
DeepInteraction [3]		LC	72.6	69.9	88.5	64.4	79.2	44.5	30.1	88.9	79.0	67.8	80.0	76.4
BEVFusion [4]	ICRA 23	LC	71.4	68.5	—	—	—	—	—	—	—	—	—	—
CMT [87]	ICCV 23	LC	72.9	70.3	—	—	—	—	—	—	—	—	—	—
SparseFusion [92]	ICCV 23	LC	72.8	70.4	—	—	—	—	—	—	—	—	—	—
UniTR [86]	ICCV 23	LC	73.3	70.5	—	—	—	—	—	—	—	—	—	—
DAL-大 [88]	ECCV 24	LC	74.0	71.5	—	—	—	—	—	—	—	—	—	—
M3Net [72]	AAAI 25	LC	72.4	69.0	—	—	—	—	—	—	—	—	—	—
DeepInteraction++ [71]	TPAMI 25	LC	73.3	70.6	80.0	65.2	80.0	44.7	30.4	89.3	80.3	69.4	80.6	77.2
UniLION (欧司朗)	—	LC	74.9	72.3	89.5	67.9	80.9	49.3	33.3	91.5	81.8	70.0	84.1	74.2
QTNet [38]	神经信息处理系统会议第23届 arXiv 23	LCT	73.1	70.3	88.4	64.7	79.0	44.8	29.4	89.4	80.5	70.6	79.7	76.1
BEVFusion4D [7]		LCT	73.5	72.0	90.6	70.3	81.5	47.1	32.9	90.2	81.5	73.0	80.9	71.6
UniLION (欧司朗)	—	LCT	75.4	73.2	90.3	69.6	81.7	49.4	35.6	91.8	83.5	71.6	84.8	73.8

表3
在nuScenes验证集上追踪性能表现。

方法	出席	模态	AMOTA↑	AMOTP↓	召回↑	IDS↓
VIP3D [66]	CVPR 23	CT	21.7	1.625	0.363	—
UniAD [13]	CVPR 23	CT	35.9	1.320	0.467	906
SparseDrive [67]	ICRA 25	CT	50.1	1.085	0.601	632
CenterPoint [36]	CVPR 21	L	66.5	0.567	0.562	562
SimpleTrack [93]	arXiv 21	L	69.6	0.547	0.602	—
3DMOTFormer [64]	ICCV 23	L	71.2	0.515	—	341
UniLION (欧司朗)	—	L	72.6	0.542	0.764	510
UniLION (欧司朗)	—	LT	73.3	0.515	0.765	537
EagerMOT [65]	ICRA 21	LC	71.2	0.569	0.752	899
AlphaTrack [63]	IROS 21	LC	73.3	—	—	—
UniLION (欧司朗)	—	LC	76.2	0.499	0.783	711
UniLION (欧司朗)	—	LCT	76.5	0.477	0.796	613

在规划任务中，碰撞率为0.27%。对于同时利用激光雷达和摄像头输入的多模态配置，UniLION在四项核心感知任务中表现卓越：检测任务的NDS达到74.9%，mAP为72.2%；跟踪任务的AMOTA为76.2%；地图分割任务的mIoU为72.3%；占用预测任务的RayIoU为50.8%。时间多模态版本作为我们最具竞争力的配置，在所有评估任务中均达到顶尖或高度竞争性的表现：检测任务的NDS为75.4%，mAP为73.2%；跟踪任务的AMOTA为76.5%；地图分割任务的mIoU为73.3%；占用预测任务的RayIoU为51.3%；车辆运动预测的minADE为0.57分钟，行人运动预测的minADE为0.37分钟，规划任务的碰撞率低至0.18%。需要说明的是，我们在规划任务中未使用自我状态信息。

三维目标检测结果。表2展示了UniLION在nuScenes验证集上的三维目标检测详细结果。我们的UniLION在四种配置下均达到SOTA性能：纯激光雷达、激光雷达-时间序列、多模态及时间序列多模态模式。这些结果验证了UniLION作为三维目标检测统一框架的优越性，表明我们的方法不仅能超越单一任务的专用模型，还能同时处理多项自动驾驶任务。

三维多目标跟踪结果。表3展示了UniLION在nuScenes验证集上的详细跟踪结果。我们的方法在所有配置下均保持竞争力：仅使用激光雷达时AMOTA为72.6%（较先前SOTA方法3DMOTFormer[64]提升1.4%），

表4
在nuScenes验证集上的地图分割性能。

方法	出席	模态	mIoU↑	可驱动的	穿越	步道	停止线	卡帕克	分隔器
OFT [94]	BMVC 19	C	42.1	74.0	35.3	45.9	27.5	35.9	33.9
LSS [95]	ECCV 20	C	44.4	75.4	38.8	46.3	30.3	39.1	36.5
CVT [96]	CVPR 22	C	40.2	74.3	36.8	39.9	25.8	35.0	29.4
M2BEV [97]	arXiv 22	C	—	77.2	—	—	—	—	40.5
BEVFusion [4]	ICRA 23	C	56.6	81.7	54.8	58.4	47.4	50.7	46.4
CenterPoint [36]	ECCV 21	L	48.6	75.6	48.4	57.5	36.5	31.7	41.9
DSVT [61]	CVPR 23	L	68.0	87.6	67.2	72.7	59.7	62.7	58.2
UniLION (欧司朗)	—	L	71.7	90.0	73.2	77.0	64.2	61.1	64.5
UniLION (欧司朗)	—	LT	72.4	90.5	73.7	78.1	64.9	62.4	65.0
MVP [35]	神经信息处理系统会议21	LC	49.0	76.1	48.7	57.0	36.9	33.0	42.2
BEVFusion [4]	ICRA 23	LC	62.7	85.5	60.5	67.6	52.0	57.0	53.7
M3Net [72]	AAAI 25	LC	70.4	90.3	69.6	75.8	63.4	62.3	61.1
UniLION (欧司朗)	—	LC	72.3	90.2	73.1	76.8	64.6	64.4	64.8
UniLION (欧司朗)	—	LCT	73.3	91.2	74.5	78.4	65.4	64.9	65.5

表5
在Occ3D-nuScenes验证集上的占用率预测性能。

方法	出席	模态	RayIoU↑	RayIoU _{1m}	RayIoU _{2m}	RayIoU _{4m}
RenderOcc [98]	ICRA 24	C	19.5	13.4	19.6	25.5
SimpleOcc [93]	TIV 24	C	22.5	17.0	22.7	27.9
BEVFormer [9]	ECCV 22	CT	32.4	26.1	32.9	38.0
BEVDet-Occ [81]	arXiv 22	CT	32.6	26.6	33.1	38.2
FB-Occ [99]	arXiv 23	CT	33.5	26.7	34.1	39.7
FlashOcc [79]	arXiv 23	CT	—	—	—	—
SparseOcc [68]	ECCV 24	CT	36.1	30.2	36.8	41.2
OPUS [82]	神经信息处理系统会议24	CT	41.2	34.7	42.1	46.7
UniLION (欧司朗)	—	L	46.7	43.1	47.3	49.9
UniLION (欧司朗)	—	LT	49.6	46.0	50.2	52.7
UniLION (欧司朗)	—	LC	50.8	47.2	51.3	53.9
UniLION (欧司朗)	—	LCT	51.3	47.7	51.9	54.4

表6
nuScenes验证集上的运动预测性能。

方法	出席	模态	minADE _{Car} (m) ↓	minADE _{Ped} (m) ↓	minFDE _{Car} (m) ↓	MR _{Car} ↓	EPA _{Car} ↑
VIP3D [66]	CVPR 23	CT	2.05	—	2.84	0.246	0.226
UniAD [13]	CVPR 23	CT	0.71	—	1.02	0.151	0.456
SparseDrive [67]	ICRA 25	CT	0.60	0.72	0.96	0.132	0.555
PnPNet [85]	CVPR 20	LT	1.15	—	1.95	0.226	0.222
UniLION (欧司朗)	—	LT	0.58	0.39	1.02	0.166	0.647
FusionAD [15]	arXiv 23	LCT	0.39	—	0.62	0.086	0.626
DeepInteraction++ [71]	TPAMI 25	LCT	0.34	—	0.54	0.047	—
UniLION (欧司朗)	—	LCT	0.57	0.37	0.97	0.163	0.678

激光雷达时间模式 AMOTA 率为 76.2%，多模态模式 AMOTA 率为 76.2%（较先前 SOTA 方法 AlphaTrack[63]提升 2.9%），时间多模态模式 AMOTA 率为 76.5%，创下新的 SOTA 纪录。这些结果充分证明了 UniLION 在三维多目标跟踪中的卓越性能。

BEV地图分割结果。表4展示了UniLION在nuScenes验证集上的详细地图分割结果。我们的方法在所有配置下均保持竞争力：仅使用激光雷达时 mIoU 为 71.7%（较 DSVT [61] 提升 3.7%），激光雷达-时间序列结合时 mIoU 为 72.4%，多模态方法时 mIoU 为 72.3%（较先前 SOTA 多任务方法 M3Net[72] 提升 1.9%），时间序列多模态方法时 mIoU 为 73.3%。这些结果证明了

UniLION 作为地图分割的统一框架。**占用率预测结果。**表5展示了 UniLION 在 Occ3D-nuScenes [100] 验证集上的占用率预测性能。我们的方法在不同场景下均取得优异结果：仅使用 LiDAR 时 RayIoU 为 46.7%，LiDAR- 时间序列结合时为 49.6%，多模态输入时为 50.8%，时间多模态输入时为 51.3%，创下新的前沿基准。值得注意的是，UniLION 显著优于现有 SOTA 方法，分别以 15.2% 和 10.1% 的 RayIoU 超越 SparseOcc [68] 和 OPUS [82]。这些发现验证了 UniLION 在占用率预测中的有效性。**运动预测结果。**表6评估了 UniLION 在 nuScenes 验证集上的运动预测能力。与许多仅关注

在车辆运动评估中，我们将分析范围扩展至行人轨迹。在仅使用激光雷达（LiDAR）的配置下，UniLION模型对车辆的平均绝对误差（minADE）为0.58分钟，对行人的minADE为0.39分钟。在多模态场景中，该模型对车辆的minADE达到0.57分钟，对行人的minADE为0.37分钟。这些结果凸显了UniLION作为跨物体类别运动预测综合方法的泛化能力。

规划任务评估结果。针对规划任务，我们在nuScenes验证集上对UniLION进行了测试，具体数据如表7所示。需要指出的是，整合自我状态信息会导致信息泄露^[80]。为此，UniLION特意排除了自我状态信息以缓解该问题。在激光雷达配置下，UniLION实现了0.70米的L2误差和0.27%的碰撞率。当采用多模态输入时，UniLION性能进一步提升至0.65米L2误差和0.18%的碰撞率。得益于UniLION增强的表征能力，我们的方法显著优于BEVFormer-M，将碰撞率从0.68%大幅降低至0.18%。这些结果充分证明了UniLION在规划任务中的有效性与优越性。

通用模型。UniLION是一种统一架构，通过单一模型框架支持多种模态和任务。该设计本质上实现了跨异构输入格式的参数共享。为验证此能力，我们在推理过程中通过系统性研究，选择性禁用基于多模态时间数据训练的UniLION模型（行g）中的特定模态或时间输入（如表8所示）。当仅禁用时间信息时（行f），我们的模型在保持训练-测试输入一致性的标准配置下（行e）仍保持可比性能。值得注意的是，即使同时禁用时间与相机输入（行b），我们的模型仍以70.6 NDS 的性能表现，超越了TransFusion-L（70.1 NDS）和BEVFusion-L（69.3 NDS）。这些结果表明，单个UniLION模型在基于多模态时间数据训练后，可无缝部署于不同传感器配置和时间场景（例如纯激光雷达、时间激光雷达或多模态融合）而无需重新训练。训练。该能力展现出对不同操作条件的卓越适应性，同时显著增强了我们统一框架的稳健性和通用性。

不同的图像骨干网络。在表9中，我们提供了轻量级版本UniLION，该版本采用ResNet-50^[78]作为图像骨干网络，图像分辨率较小，为 256×704 。与基础模型（以Swin-tiny^[77]作为图像骨干网络，图像分辨率 384×1056 ）相比，轻量级版本仍取得了令人鼓舞的性能：3D检测的NDS为73.6%，mAP为70.8%；多目标跟踪的AMOTA为75.0%；地图分割的mIoU为71.8%；3D占用率的RayIoU为50.2%。

不同的线性 RNN 算子。为验证框架的灵活性，我们评估了另一种代表性线性 RNN 算子 RWKV^[20]，如表9所示。虽然UniLION-RWKV 的性能略逊于UniLION-Mamba，但在多项自动驾驶任务中仍取得优异表现，充分证明了本框架的灵活性。

4.4 消融研究

为验证UniLION的有效性，我们在nuScenes验证集上对其开展消融实验。为快速验证，我们以采用ResNet50图像骨干网络和 256×704 分辨率的单帧多模态UniLION作为默认模型，并评估其在三维感知任务（例如三维目标检测、跟踪、地图分割及空间占用）中的性能表现。

UniLION的每个组件。我们验证了UniLION的各个组件，包括空间特征描述符和体素生成模块，如表10所示。当将3D空间特征描述符整合到基线模型（行a）中时，该设置（行b）带来了0.7%的NDS、0.8%的mAP、1.9%的AMOTA、0.5%的mIoU和1.1%的RayIoU的性能提升。这证明了我们提出的3D空间特征描述符在弥补线性RNN有限空间建模能力方面的有效性。此外，体素生成模块（行c）通过增强前景体素特征表示，使性能较基线模型提升了0.6%的NDS、1.1%的mAP、2.7%的AMOTA、0.1%的mIoU和0.3%的RayIoU。最后，当所有组件结合使用时（行d），UniLION实现了73.6%的NDS、70.8%的mAP、75.0%的AMOTA、71.8%的mIoU和50.2%的RayIoU，较基线模型分别提升了0.7%的NDS、1.8%的mAP、3.1%的AMOTA、1.6%的mIoU和1.8%的RayIoU。

动态损失机制的有效性。表11展示了验证动态损失机制有效性的实验结果。采用动态损失机制后，多数任务均取得稳定提升：检测任务NDS提升0.3%，跟踪任务AMOTA提升0.9%，地图分割任务mIoU提升0.6%。但三维空间占用率性能出现小幅下降。我们认为这是由于动态损失机制促使UniLION优先考虑整体任务平衡，这可能以牺牲单个任务优化为代价，尤其对占用率预测任务而言。

多任务学习。表12展示了联合训练对不同任务性能的影响。当同时训练3D检测和地图分割时，地图分割任务的性能显著提升（71.7%平均交并比 vs. 68.3%平均交并比）。若进一步整合占用预测任务，虽然检测性能略有下降，但占用预测任务的性能提升达2.7%的雷伊交并比，这得益于检测任务对3D占用估计的全面增强。总体而言，我们的联合训练方法性能与单任务模型相当甚至更优，充分证明了UniLION 3D主干网络提取的紧凑BEV特征表示的有效性。

窗口大小与组大小的鲁棒性。UniLION的一个根本优势在于通过整合线性RNN实现长程依赖建模的能力。为评估我们方法的泛化能力和参数敏感性，我们在推理过程中对不同窗口大小和组大小进行了全面的鲁棒性分析，如表13所示。具体而言，我们评估了UniLION（窗口大小为{[13, 13, 32], [13, 13, 16], [13, 13, 8], 以及[13, 13, 4]}和组大小为{4096, 2048, 1024, 512}）在不同窗口大小下的表现

表7 nuScenes验证集上的规划性能。 t 表示使用自我状态信息。

方法	出席	模态	$L2(m) \downarrow$				$Col.(\%) \downarrow$			
			1s	2s	3s	平均数	1s	2s	3s	平均数
UniAD [13] VAD [14] PARA-Drive [16] BEVFormer-C [70]	CVPR 23	CT	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61
	ICCV 23	CT	0.41	0.70	1.05	0.72	0.03	0.19	0.43	0.21
	CVPR 24	CT	0.26	0.59	1.12	0.66	0.00	0.12	0.65	0.26
	TPAMI 24	CT	-	-	-	0.99	-	-	-	0.70
FF [101] EO [102] UniLION (欧司朗)	CVPR 21	LT	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
	ECCV 22	LT	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
	-	LT	0.35	0.67	1.09	0.70	0.01	0.20	0.60	0.27
FusionAD [15] BEVFormer-M [70] DeepInteraction++† [71] UniLION (欧司朗) UniLIONt (欧司朗)	arXiv 23	LCT	-	-	-	1.03	0.25	0.13	0.25	0.21
	TPAMI 24	LCT	-	-	-	0.95	-	-	-	0.68
	TPAMI 25	LCT	0.36	0.67	1.06	0.70	0.05	0.15	0.38	0.19
	-	LCT	0.33	0.62	0.99	0.65	0.01	0.12	0.42	0.18
UniLIONt (欧司朗)	-	LCT	0.18	0.37	0.65	0.55	0.01	0.02	0.14	0.06

表8
在推理过程中禁用多模态融合与时间融合的实验。我们未采用任何掩码模态训练策略。

#	训练方式	测试方式	NDS↑	检测 mAP↑	跟踪 AMOTA↑	地图 mIoU↑	占用率 RayIoU↑
<i>a</i>	L	L	72.3	67.5	72.6	71.7	46.8
<i>b</i>	LCT	L	70.6	64.6	70.2	68.6	43.4
<i>c</i>	LT	LT	73.0	68.9	73.3	72.4	49.6
<i>d</i>	LCT	LT	70.3	64.0	70.6	68.7	42.0
<i>e</i>	LC	LC	74.9	72.2	76.2	72.3	50.8
<i>f</i>	LCT	LC	74.9	72.3	76.2	72.2	50.7
<i>g</i>	LCT	LCT	75.4	73.2	76.5	73.3	51.3

表9
不同图像骨架与线性 RNN 算子的比较

操作员	图像后端	分辨率	NDS↑	检测 mAP↑	跟踪 AMOTA↑	地图 mIoU↑	占用率 RayIoU↑
UniLION-RWKV [20]	ResNet50 [78]	256×704	73.7	69.9	74.0	71.9	49.7
UniLION-Mamba [25]	ResNet50 [78]	256×704	73.6	70.8	75.0	71.8	50.2
UniLION-RWKV [20]	Swin-Tiny [77]	384×1056	74.3	71.4	75.1	72.0	49.9
UniLION-Mamba [25]	Swin-Tiny [77]	384×1056	74.9	72.2	76.2	72.3	50.8

表10 UniLION各组件的消融研究。我们在nuScenes数据集上使用R50图像骨架和 256×704 图像尺寸训练UniLION。

#	三维空间特征描述符	体素生成	检测 NDS ↑ mAP ↑	跟踪 AMOTA↑	地图 mIoU↑	占用率 RayIoU↑	
<i>a b</i>	-	-	72.9	69.0	71.9	70.2	48.4
<i>b</i>	✓	-	73.6	69.8	73.8	70.7	49.5
<i>c d</i>	-	✓	73.5	70.1	74.6	70.3	48.7
<i>d</i>	✓	✓	73.6	70.8	75.0	71.8	50.2

表 11
动态损失消融研究

动态损耗	NDS↑	检测 mAP↑	跟踪 AMOTA↑	地图 mIoU↑	占用率 RayIoU↑
	73.3	70.7	74.1	71.2	50.4
✓	73.6	70.8	75.0	71.8	50.2

以及群体规模。我们的实证研究结果表明，尿素代谢对多种下游任务具有调节作用。这表明UniLION展现出显著的稳定性和一致性。UniLION具有良好的外推能力，无需依赖手工构建的先验信息，即可在不同窗口和组大小配置下保持稳定表现。

表 12
不同任务影响的消融研究

#	检测	地图分割	占用率	NDS↑	检测 mAP↑	地图 mIoU↑	占用率 RayIoU↑
a	✓	—	—	73.9	71.4	—	—
b	—	✓	—	—	—	68.3	—
c	—	—	✓	—	—	—	47.5
d	✓	✓	—	74.0	71.4	71.7	—
e	✓	✓	✓	73.6	70.8	71.8	50.2

表 13
不同窗口大小与组别规模稳健性的消融研究

#	窗口大小	群组大小	NDS↑	检测 mAP↑	跟踪 AMOTA↑	地图 mIoU↑	占用率 RayIoU↑
a b	[7, 7, 32], [7, 7, 16], [7, 7, 8], [7, 7, 4] [13, 13, 32], [13, 13, 16], [13, 13, 8], [13, 13, 4]	[4096, 2048, 1024, 512] [4096, 2048, 1024, 512]	73.7 73.7	70.8 70.8	74.9 74.9	71.7 71.7	50.3 50.0
c d	[25, 25, 32], [25, 25, 16], [25, 25, 8], [25, 25, 4] [13, 13, 32], [13, 13, 16], [13, 13, 8], [13, 13, 4]	[4096, 2048, 1024, 512] [2048, 1024, 512, 256]	73.6 73.7	70.8 70.8	75.0 75.3	71.8 71.2	50.2 49.7
e	[13, 13, 32], [13, 13, 16], [13, 13, 8], [13, 13, 4]	[4096, 2048, 1024, 512]	73.6	70.8	75.0	71.8	50.2
f	[13, 13, 32], [13, 13, 16], [13, 13, 8], [13, 13, 4]	[8192, 4096, 2048, 1024]	73.6	70.6	75.1	71.7	50.0

表 14

UniLION鲁棒性消融研究。第一行为仅含LiDAR数据且无错位的模型，其余实验为采用多模态输入的UniLION模型。‘无’表示对齐模型。

错位程度	NDS↑	检测 mAP↑	跟踪 AMOTA↑	地图 mIoU↑	占用率 RayIoU↑
否（仅激光雷达）	72.3	67.5	72.6	71.7	46.8
高（多模态）	72.8	69.5	74.0	71.5	48.8
中间（多模态）	73.3	70.2	74.7	71.6	49.1
低（多模态）	73.5	70.6	74.9	71.7	50.0
否（多模态）	73.6	70.8	75.0	71.8	50.2

传感器错位的鲁棒性。传感器错位问题可能出现在大多数自动驾驶系统中。因此，探索传感器错位的鲁棒性对于确保自动驾驶系统的安全性至关重要。为验证我们UniLION的鲁棒性，我们遵循FBMNet[103]来模拟激光雷达与摄像头模态之间的传感器错位。具体而言，‘低’、‘中’、‘高’三个错位级别分别表示摄像头外在矩阵沿垂直方向旋转1.5°、3.0°、5.0°，并分别平移0.15米、0.30米、0.50米。在低错位级别下，UniLION在不同任务中的表现与对齐模型相当。此外，我们的UniLION在适度退化（NDS 0.8%、mAP 1.3%、AMOTA 1.0%、mIoU 0.3%、RayIoU 1.4%）下仍表现出良好性能，并在高错位水平下展现出强鲁棒性。值得注意的是，尽管存在摄像头-激光雷达错位，多模态UniLION始终优于仅含激光雷达的版本。这些实验最终证明了UniLION对传感器错位挑战的鲁棒性。

Top-K in UniLION Backbone.在表15中，我们研究了UniLION 3D骨干网络中top-K深度候选点对生成相机体素的影响。总体而言，实验结果表明随着K的增加，3D目标检测的性能持续提升，这表明

引入更多深度候选参数可增强模型的表征能力。然而，为在计算效率与模型性能间保持良好平衡，我们通过实证确定 $K=4$ 作为本框架的默认设置。

4.5 可视化分析

在本节中，我们对UniLION在多个自动驾驶任务上的全面定性分析进行了展示，包括nuScenes验证集上的3D目标检测、跟踪、地图分割、占用预测、运动预测和规划。**1)3D目标检测：**如图6(a)和(b)所示，我们分别展示了多视角图像和BEV上的预测结果。得益于UniLION的全面特征表示，UniLION能够成功检测到难以检测的目标（例如，CAM前视图图像中的远处目标）。**2)3D目标跟踪：**如图6所示，我们使用唯一ID区分不同目标来可视化跟踪结果。UniLION能准确关联图6(b)和图6(c)中的跨帧目标。**3)占用预测：**如图6(d)所示，我们可视化了预测的占用情况。基于统一的BEV特征，UniLION能获得准确的3D占用结果。**4)地图分割：**如图6(e)所示，UniLION能准确分割地图

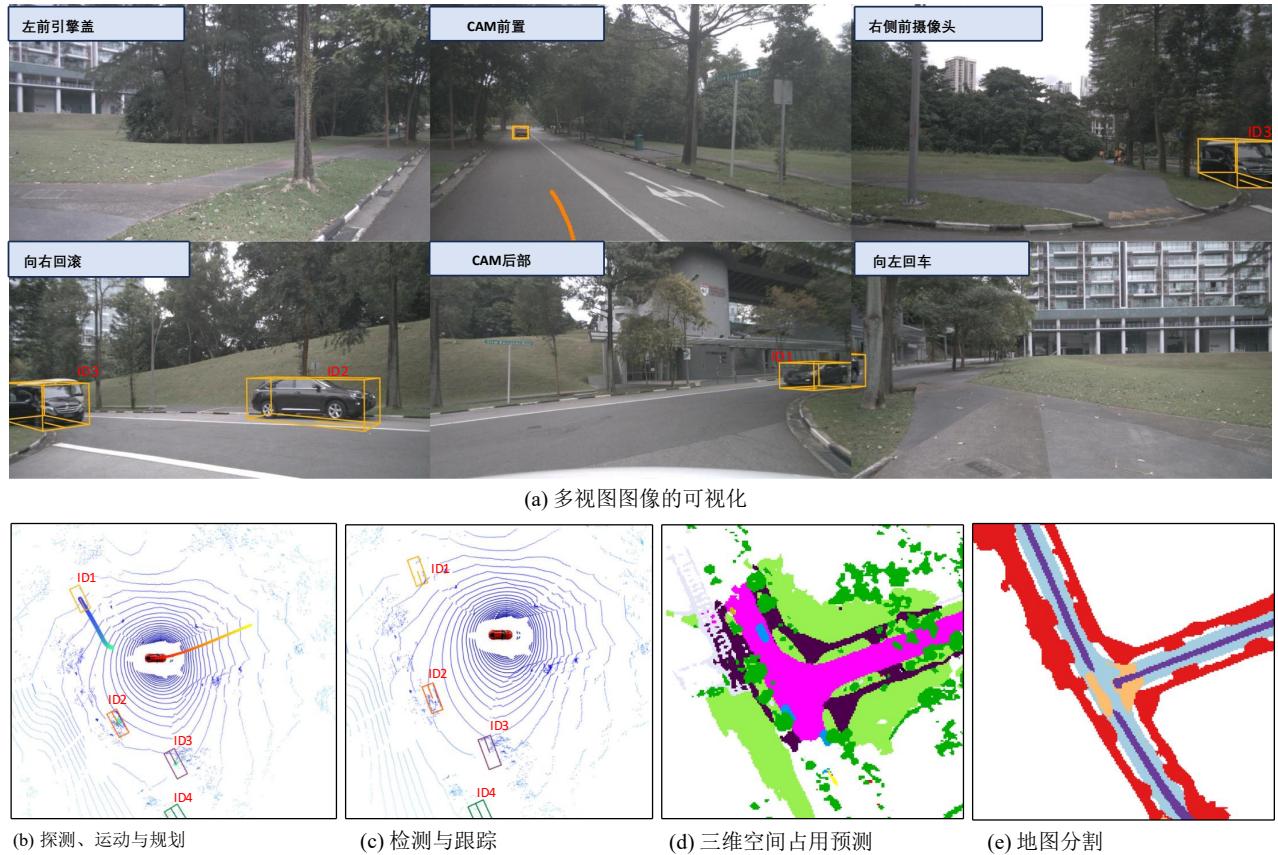


图6展示了UniLION在nuScenes验证集上进行三维物体检测、跟踪、地图分割、占用预测、运动预测及规划的定性结果。第一行和第二行分别对应多视角图像与BEV的预测结果。

表 15
UniLION主干 Top-K 消融研究

K	检测		AMOTA↑	地图 mIoU↑	占用率 RayIoU↑
	NDS↑	mAP↑			
1	73.2	70.3	74.7	71.4	49.5
2	73.5	70.7	74.8	71.5	50.2
4	73.6	70.8	75.0	71.8	50.2
8	73.8	71.1	75.3	71.7	50.0

元素（例如车道线、可行驶区域）并提供丰富的地图信息。

5) **运动与规划：**如图6(b)所示，我们可视化了运动预测与规划的预测结果。在运动预测方面，UniLION能准确区分移动与静止物体；在规划方面，UniLION可生成合理轨迹以避免碰撞。

5 结论

本文提出了一种基于线性 RNN 的框架UniLION，该框架作为统一的三维主干网络，无需任何显式融合模块即可无缝处理不同模态和时间信息。得益于这种简洁优雅的架构，该统一三维主干网络能够将多种信息压缩为紧凑统一的BEV表征，作为通用特征，实现对多样化自动驾驶场景的无缝适配。

通过并行多任务学习实现任务处理。大量实验证明了我们的UniLION在特征表示方面的优越性。最终，UniLION在包括3D感知（3D目标检测、跟踪、占用预测、BEV地图分割）、运动预测和规划在内的综合自动驾驶任务中，取得了具有竞争力甚至达到最先进水平的性能，充分证明了我们统一方法UniLION的泛化能力和有效性。

参考文献

- [1] 黄涛、刘泽、陈旭和白旭，《Epnnet：基于图像语义增强点特征用于三维物体检测》，ECCV 2020。
- [2] Z. Liu、T. Huang、B. Li、X. Chen、X. Wang 和 X. Bai，《Epnnet++：用于多模态三维目标检测的级联双向融合》，2022 年。
- [3] Z. Yang、J. Chen、Z. Miao、W. Li、X. Zhu 和 L. Zhang，“深度交互：通过模态交互实现三维物体检测”，收录于NeurIPS 2022。