https://doi.org/10.1051/jnwpu/20234161190

基于 Transformer 的 3D 点云目标检测算法

刘明阳[1], 杨啟明[2], 胡冠华[2,3], 郭岩[4], 张建东[2]

(1.沈阳飞机设计研究所, 辽宁 沈阳 110035;2.西北工业大学 电子信息学院, 陕西 西安 7101291;
3.中国船舶集团有限公司系统工程研究院, 北京 100094;4.空装驻沈阳地区第一军事代表室, 辽宁 沈阳 110850)

摘 要:针对在三维目标检测中由于空间维度的增加基于锚框的方法难以部署的问题,研究了基于集合预测的点云目标检测算法。提出一种基于 Transformer 的 3D 点云目标检测算法,并结合自动驾驶场景下的点云特点,提出了改进空间调制注意力和热图初始化策略进行训练加速和查询初始化,在浅层网络下取得了良好的检测性能。在 KITTI 数据集上与其他算法进行比较,结果表明所提算法在性能上已经达到先进水平,进一步对算法中的主要组成部分进行了消融实验,验证了各个模块对检测效果的贡献。

关 键 词:Transformer;空间调制注意力机制;热图初始化;目标检测;深度学习
中图分类号:TP18 **文献标志码**:A **文章编号**:1000-2758(2023)06-1190-08

传统的基于锚框的图像目标检测算法其本质是对预定义的密集锚框进行内部物体类别的分类和边框位置的微调[1]。而将其应用于点云目标检测时由于空间维度的增加,大部分基于锚框的检测算法在部署时面临新的困难,一方面在锚框铺设时需要考虑高度因素,涉及到锚框数量、大小、角度和密度等超参数,这些参数的手动试错过程需要耗费极大的精力和时间成本;另一方面锚框角度回归的增加使得这一类算法在检测框微调时难度大大提升[2]。

作为将 Transformer[3] 应用于机器视觉领域的先驱,DETR 算法[4] 将目标检测视为一个集合预测问题。Transformer 本质上起到的是一个序列转换作用,因而可以将 DETR 视为一个将图像特征序列向目标集合序列的转换模块。具体来说,DETR 中设计了一系列的目标查询向量负责检测图像中不同位置的物体[5],每个目标查询向量都与来自卷积神经网络的空间视觉特征进行交互,并利用交叉注意力机制[6] 自适应地收集目标相关信息,用于估计检测框位置和目标类别。

在训练时,DETR 预测固定数量的目标,并对真实值和预测目标之间进行二分匹配以保证每个真实目标只有一个相匹配的对应。假设 $\boldsymbol{y}$ 表示真实目标集合,$\hat{\boldsymbol{y}} = \{\hat{y}_i\}_{i=1}^{N}$ 表示 N 个预测目标集合,一般来说 N 的取值要远远大于图像中真实目标数量。将 $\boldsymbol{y}$ 采用空元素 $\varnothing$ 填充至 N,并计算 N 个元素之间的代价最低的二分匹配结果

$$\hat{\sigma} = \arg\min \sum_{i}^{N} L_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) \tag{1}$$

式中,L_{match} 表示真实值 y_i 在组合 σ 下与预测值 $\hat{y}_{\sigma(i)}$ 的匹配代价,在 DETR 中采用匈牙利算法进行计算。匹配代价由损失函数所决定,同时包含了类别预测和与真实框的相似性。对于每个真实目标 $y_i = (T_i, b_i)$ 和预测目标 $\hat{y}_{\sigma(i)} = (T_{\sigma(i)}, b_{\sigma(i)})$,将其类别相同的概率定义为 $\hat{P}_{\sigma(i)}(T_i)$,其中类别标签可能为 $\varnothing$。则匹配代价可以表示为

$$L_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) = -1_{\{c \neq \varnothing\}} \hat{P}_{\sigma(i)}(T_i) + 1_{\{c \neq \varnothing\}} L_{\text{box}}(b_i, b_{\sigma(i)}) \tag{2}$$

本文从基于体素化表达的检测算法出发,采用 DETR 的结构构造一种新型的针对户外场景的 3D 点云目标检测算法,并且结合点云数据和 3D 目标

收稿日期:2023-01-09 **基金项目**:陕西省自然科学基础研究计划(2022JQ-593)与陕西省重点研发计划(2022GY-089)资助
作者简介:刘明阳(1986—),沈阳飞机设计研究所高级工程师,主要从事航空电子系统设计研究。
通信作者:杨啟明(1988—),西北工业大学助理研究员,主要从事人工智能与自主决策控制研究。
e-mail:yangqm@ nwpu.edu.cn

检测场景下的特点对原有的 DETR 进行了改进。其包含 3 个主要模块:①体素化及三维体素编码网络;②二维特征提取网络以及特征金字塔网络;③结合空间调制交叉注意力(spatially modulated cross-attention,SMCA)的解码器。

1　问题描述

假定原始点云输入记为 $\{\boldsymbol{p}_n = (x_n, y_n, z_n, r_n), n = 1, 2, \cdots, N\}$,其中$(x_n, y_n, z_n)$ 是点云的三维空间坐标位置,r_n 是反射强度,N 是集合中包含的点云总数。3D 目标检测任务中要求对于场景内的所有真实目标$\{y_i = (T_i, b_i), i = 1, 2, \cdots, M\}$ 进行定位和分类,T_i 和 b_i 分别表示目标的类别和检测框。检测框如图 1 所示。

三维检测框通常采用底部中心点坐标加三维尺寸和角度的表示方法,即 $b = (x_c, y_c, z_c, l, w, h, \alpha)$。其中$(x_c, y_c, z_c)$ 维底部中心点坐标,l, w, h 为目标的三维尺寸长、宽、高,α 是目标行进方向和 X 轴正方向的夹角。

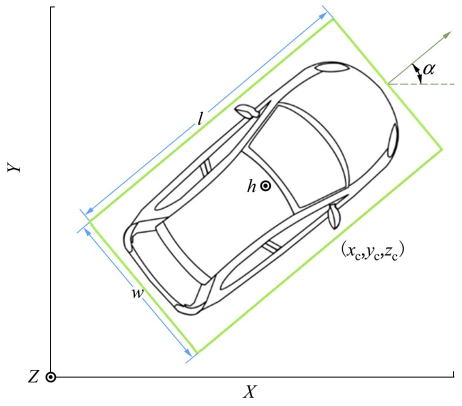


图 1　三维检测框鸟瞰示意图

2　算法设计

2.1　原理概述

本文所提出的基于 Transformer 的 3D 点云目标检测器结构如图 2 所示。与大部分三维目标检测器类似,包含 3D 体素编码网络、2D 骨干网络以及头部检测网络。本算法在 2D / 3D 特征提取过程中所

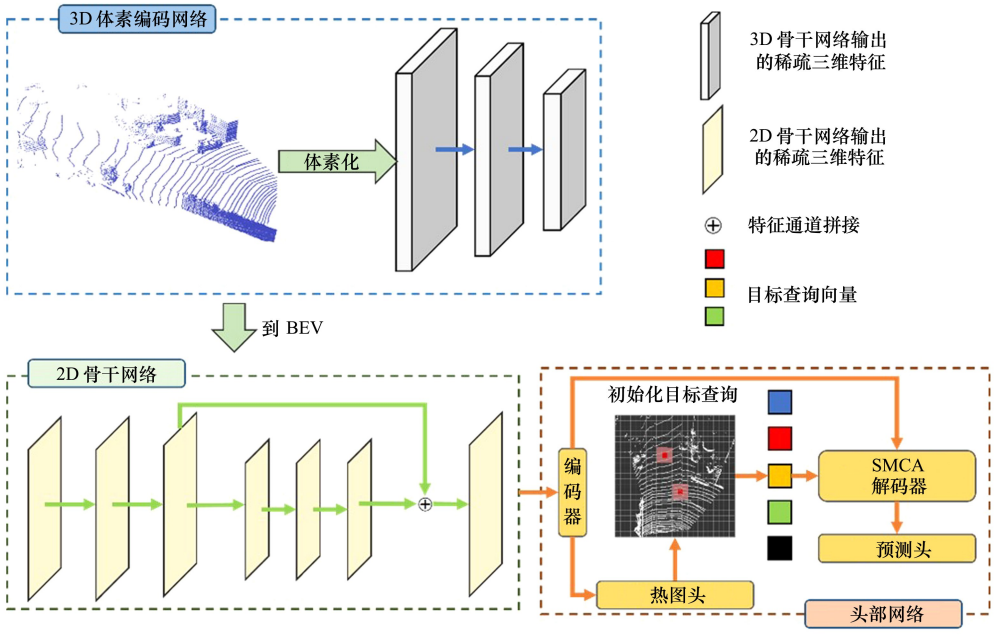


图 2　基于 Transformaer 的 3D 点云目标检测器主体结构

采用的三维体素特征编码网络和二维特征提取骨干网络与 SECOND 相同。本文的研究内容主要集中在头部检测网络,一方面采用热度图初始化目标查询向量的方法,使得编码器无需从头开始学习目标特征;另一方面改进原有的空间调制注意力机制使其与热图初始化策略相适应,极大地加快了网络的

训练速度。

2.2 热图初始化的目标查询

DETR 采用编码器-解码器的 Transformer 架构，编码器和解码器都各自级联 6 层。编码器由多头自注意力和前馈网络组成，起到的作用与卷积层类似，都是从输入中关联并提取上下文特征，只是自注意力机制更多关注全局特征，其感受野要远大于卷积网络。解码器则具有额外的多头交叉注意力，将固定数量的目标查询与编码器的输出特征进行交互，并利用交叉注意力机制自适应地聚合相关信息。

在 DETR 及其诸多改进版本中，目标查询通常采用 0 初始化或随机初始化，这使得解码器需要更多层的级联才能使目标查询具备物体地空间特征。Efficient DETR 算法[7]中指出，通过更好地初始化目标查询向量或者对解码器输出进行辅助损失计算，能够有效减少解码器的级联层数并且增强其目标感知能力。受到这一观点的启发，本算法采用了热图初始化的目标查询策略，使得初始的目标查询位于或靠近真实的目标中心，从而无需多层解码器来进行位置的细化。在本算法中，对于来自骨干网络的深度点云特征首先送入解码器进行全局特征感知，热图头接收解码器输出的全局特征并进行热图预测。

热图头预测位于潜在目标位置的关键点热图 $\hat{\boldsymbol{M}} \in \boldsymbol{R}^{H\times W\times C}$，其中 H 和 W 分别是输入鸟瞰特征图的高和宽，$\boldsymbol{C}$ 是待检测目标类别（不包含背景类）。在训练时，对于每一个被真实 3D 检测框覆盖的像素 $(x,\boldsymbol{y})$，以类高斯分布形式对其进行赋值，如(3) 式所示

$$M_{x,\boldsymbol{y},c} = \exp\left(-\frac{(x-p_x)^2+(\boldsymbol{y}-p_y)^2}{2\sigma_p^2}\right) \tag{3}$$

式中：p_x，p_y 是真实目标检测框的中心点；σ_p 是与检测框大小相关的分布半径。热图头的损失函数可以描述为

$$L_{\text{heat}} = -\frac{1}{N}\sum_{x,\boldsymbol{y},c}\cdot \begin{cases}(1-\hat{M}_{x,\boldsymbol{y},c})^{\alpha}\lg(\hat{M}_{x,\boldsymbol{y},c}), & M_{x,\boldsymbol{y},c}=1\\ (1-\hat{M}_{x,\boldsymbol{y},c})^{\beta}(\hat{M}_{x,\boldsymbol{y},c})^{\alpha}\lg(1-\hat{M}_{x,\boldsymbol{y},c}), & \text{其他}\end{cases} \tag{4}$$

式中：α 和 β 是用于训练的超参数；N 是当前样本中所有待检测目标数量。

将热图头所输出的关键点热图 $\hat{\boldsymbol{M}} \in \boldsymbol{R}^{H\times W\times C}$ 用于初始化目标查询流程如图 3 所示。

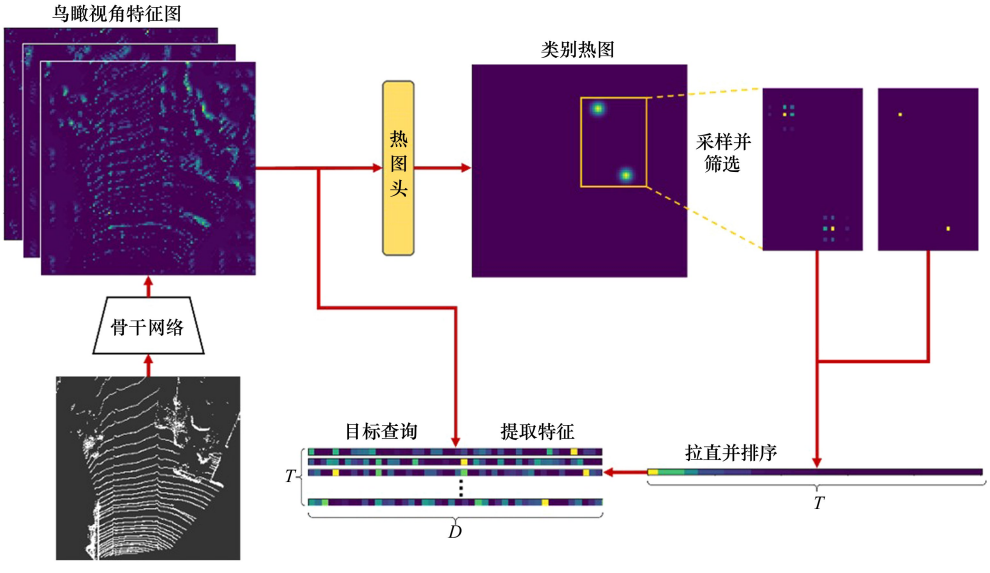


图 3 热图初始化查询向量计算流程

对于给定的关键点热图 $\hat{M} \in R^{H\times W\times C}$ 首先进行上采样和下采样

$$\hat{M}^{s} = S_{\text{up}}(S_{\text{down}}(\hat{M})) \tag{5}$$

式中，S_{up} 和 S_{down} 分别表示上采样和下采样。此处的上采样方法选用最近邻采样，$\hat{M}^{s}$ 相比于 $\hat{M}$ 保留了局部最大特征的同时，产生了更明显的梯度变化。将采样后的热图与原始热图进行比较，并逐像素筛选出梯度变化处数值相同的位置作为关键点如(6) 式所示。从而生成分布相对稀疏的热图，以避免产

生的查询在空间上过于封闭,稀疏程度可以通过调节下采样倍率来改变。

$$\hat{M}^{\text{raw}}_{x,y,c} = \hat{M}_{x,y,c} \times 1_{\{\hat{M}_{x,y,c}=\hat{M}^{\text{s}}_{x,y,c}\}} \tag{6}$$

在 CenterNet[8] 中直接采用关键点热图的局部热峰作为预测框的中心点。但是在 DETR 的解码器中,设定的解码器中目标查询数量要远大于真实目标数量,若只采用目标中心点进行特征进行查询初始化,在解码器的自注意力计算中,会产生大量待检测目标与远处背景的无意义交互,造成极大的计算开销。而在目标周边生成诸多稀疏的查询点,中心点查询也会与边界点查询进行特征交互,从而起到辅助类别判断和类似于偏移头的位置细化作用,但同时位于目标周围数量众多的查询会导致大量假阳性检测结果。对此,本节结合了 2 种查询点生成方式,用孤立的中心查询点代替一部分在上一步中获得的稀疏查询点。

在得到稀疏化的热图后,需要从各个通道的热图中选取热度值最大的 T 个位置作为初始查询的空间坐标。计算时首先将 $\hat{M}^{\text{raw}} \in \boldsymbol{R}^{H\times W\times C}$ 拉直为一维向量 $\hat{M}^{\text{raw}} \in \boldsymbol{R}^{L}$,并进行逆序排序,进而可以得到最大的 T 个热度值在一维向量内的坐标 p。其所属类别和在热图上的位置坐标可以描述为(7) 式。

$$\begin{aligned} T_t &= \lfloor p_t/(H\times W)\rfloor \\ i_t &= p_t - \lfloor p_t/(H\times W)\rfloor \times (H\times W) \end{aligned} \tag{7}$$

需要注意的是此处的空间位置坐标 i 是每个单一类别热图拉直后的位置坐标。此外对每个查询的类别属性进行独热向量编码,再通过多层感知机升维后产生类别编码。

最后将鸟瞰图下的特征图沿每个通道展开,并将所有 i_t 位置下的特征进行拼接,并叠加类别编码从而生成目标查询向量 $q_t \in \boldsymbol{R}^D$。

2.3　改进的空间调制交叉注意力机制

DETR 另一个重要的问题在于其需要比现有的目标检测器训练更长的时间才能收敛[9]。在 COCO 2014[10] 数据集上 DETR 需要进行 500 周期的训练才能完全收敛,相较于 Faster R-CNN[11] 周期长了大约 10~20 倍。这主要是在初始状态下,交叉注意模块在整个特征图上平均分配注意力。而在收敛状态下,其注意力只关注特定的空间位置,变得极为稀疏。这种显著的注意力变化需要很长的周期进行学习。另外在解码器中采用交叉注意力进行特征聚合时,巨大的键和值的数量稀释了注意力权重,最终导致输入特征的梯度模糊。

注意力机制的本质就是定位到感兴趣的信息[12],抑制无用信息。而上述的 2 个问题可以总结归纳为检测器无法快速将注意力集中在关键点信息上。而对于该问题,一种直接快速的解决方法是采用动态调制策略,受到 SMCA-DETR 的启发,本文设计了一种与热图初始化目标查询相适应的空间调制交叉注意力。在 DETR 所采用的 Transformer 中,来自骨干网络的特征图与位置嵌入相叠加,生成查询、键以及值,并输入编码器与所有空间位置的特征进行信息交互。同时为增加特征的多样性,采用多头自注意力机制

$$\begin{aligned} E_i &= \text{softmax}(K_i^{\text{T}}\, Q_i/\sqrt{d})\, V_i \\ E &= \text{Concat}(E_1, E_2, \cdots, E_{n_{\text{head}}}) \end{aligned} \tag{8}$$

获得来自编码器视觉特征 E 后,DETR 在对象查询 $O^q \in \boldsymbol{R}^{T\times D}$ 和视觉特征 $E \in \boldsymbol{R}^{L\times D}$ 之间执行交叉注意力计算

$$\begin{aligned} Q &= FC(O^q), \quad K, V = FC(E) \\ \hat{O}_t^q &= \text{softmax}(K_i^{\text{T}}\, Q_i/\sqrt{d})\, V_i \\ \hat{O}^q &= \text{Concat}(\hat{O}_1^q, \hat{O}_2^q, \cdots, \hat{O}^q_{n_{\text{head}}}) \end{aligned} \tag{9}$$

DETR 中原始的交叉注意力无法感知目标的空间位置,因此需要多次迭代才能为每个对象查询生成适当的注意图。针对这一问题,SMCA-DETR[13] 将可学习的交叉注意力特征图与手工创建的查询先验空间相结合。通过对仅嵌入位置编码的空初始化查询向量进行学习,生成先验的感知位置,并通过堆叠的编码器进行位置细化。在本文算法中,由于热图初始化目标查询机制的存在,初始的目标查询已经位于或靠近真实的目标中心,因而无需从查询中学习先验的感知位置。其计算过程描述如下。

每个查询生成其负责对象的中心和尺寸,用于生成二维高斯空间注意力权重图[14]。目标查询的中心 c^h, c^w 由热图初始化得到的位置坐标 i 计算得到,高斯分布尺寸 $\boldsymbol{\Sigma} \in \boldsymbol{R}^{2\times 2}$ 则通过学习目标查询得到。目标查询的中心点仅需在第一个解码器中进行计算,并与之后的解码器共享。

$$\begin{aligned} c_t^h &= \lfloor i_t/H\rfloor \\ c_t^w &= i_t - \lfloor i_t/H\rfloor \times H \\ \boldsymbol{\Sigma}_t &= \text{MLP}(O_t^q) \end{aligned} \tag{10}$$

三维空间场景下,不同类别的物体尺寸差别极大,并且在具备角度信息的情况下更为复杂。SMCA 模块动态生成不同的权重半径,使得较大的目标能够聚合足够信息或抑制小目标的背景杂波。

同时该模块生成反对角线非零的权重尺寸,使得权重图并非单一地沿横纵方向延展,更利于细化目标朝向。在获取到目标中心以及分布尺寸后,SMCA 生成的类高斯权重图如(11)式所示。其中 $\boldsymbol{v} \in \boldsymbol{R}^2$ 是特征图上任意一点,$\boldsymbol{c} = [c^h, c^w]$ 是中心点坐标,β 为手动调整的超参数,以确保权重图在训练开始时覆盖较大的空间范围,使网络能接收到更多的梯度信息。

$$G = \exp\left(-\frac{1}{\beta}(\boldsymbol{v}-\boldsymbol{c})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{v}-\boldsymbol{c})\right) \tag{11}$$

在得到动态生成的空间权重图 G 的情况下,利用其调制目标查询 O^q 和编码器输出特征 E 之间的交叉注意权重。

$$O = \mathrm{softmax}(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}/\sqrt{d} + \lg G)\boldsymbol{V} \tag{12}$$

空间权重图的对数与点积获得的注意力得分之间逐元素相加,然后对所有空间位置进行 softmax 归一化。解码器中的交叉注意模块会在潜在的目标中心附近增加权重,限制了注意力搜索空间,从而提高了收敛速度。

在多头注意力下,SMCA 模块会为每个注意力头预测一个共享中心的偏移量 $[\Delta c_i^h, \Delta c_i^w]$ 和高斯分布尺寸 $\boldsymbol{\Sigma}_i$。每个注意力头的类高斯权重图将由特定的中心 $[c^h + \Delta c_i^h, c^w + \Delta c_i^w]$ 和尺寸进行生成。其注意力调制过程如(13)式所示。

$$O_i = \mathrm{softmax}(\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{Q}_i/\sqrt{d} + \lg G_i)\boldsymbol{V}_i,\quad \text{for } i = 1,2,\cdots,n_{\mathrm{head}} \tag{13}$$

区别于由所有注意力头共享的空间权重,多头调制的权重能有效突出不同的特征信息。

2.4 预测头及优化目标

对于解码器输出的 T 个查询向量,分别输入到多个并列的预测头中,生成检测结果。包括底部中心点坐标 $o \in \boldsymbol{R}^{2\times T}$,底部中心点高度 $h \in \boldsymbol{R}^{1\times T}$,检测框三维尺寸 $d \in \boldsymbol{R}^{3\times T}$,检测框角度 $r \in \boldsymbol{R}^{2\times T}$ 以及分类结果 $T \in \boldsymbol{R}^{C\times T}$,其中检测框角度由该角的正弦和余弦值编码而成。

遵循 DETR 中集合预测的检测算法,首先由各个查询生成的底部中心点、三维尺寸和角度生成预测框,并赋予对应的类别。通过匈牙利算法计算真实框与预测的二分匹配,其中匹配成本由分类损失、中心点回归损失和 IoU 损失的加权和构成

$$L_{\mathrm{match}} = \lambda_1 L_{\mathrm{cls}}(p,\hat{p}) + \lambda_2 L_{\mathrm{reg}}(b,\hat{b}) + \lambda_3 L_{\mathrm{IoU}}(b,\hat{b}) \tag{14}$$

式中:L_{cls} 是二分类交叉熵损失;L_{reg} 是预测框的鸟瞰投影中心点与真实框中心之间归一化后的的 L1 损失;L_{IoU} 是预测框和真实框之间的 IoU_{3D} 损失,λ_1,λ_2,λ_3 是各损失项的加权系数。

对于所有的预测结果,分类损失采用的 Focal Loss[15],如(15)式所示,其中 $\hat{p}_c$ 是经过 softmax 归一化后的分类预测结果,α_c,γ 是超参数。

$$L_{\mathrm{cls}} = -\alpha_c(1-\hat{p}_c)^{\gamma}\lg(\hat{p}_c) \tag{15}$$

对于匹配成功的正样本对,通过 L1 损失来监督检测框中心点的回归,并计算 IoU_{3D} 损失[16]。2 个 3D 检测框的 IoU 可以表示为(16)式

$$\mathrm{IoU}_{3D} = \frac{a_{\mathrm{overlap}} \times h_{\mathrm{overlap}}}{a_{\mathrm{gt}} \times h_{\mathrm{gt}} + a_p \times h_p - a_{\mathrm{overlap}} \times h_{\mathrm{overlap}}} \tag{16}$$

式中:a 表示检测框在鸟瞰图上的投影;h 是检测框高度;overlap 是交集区域。相比于图像检测中轴对齐的预测框回归任务,带有旋转角度的 3D 检测框最主要的区别是在计算交集部分时较为复杂。其损失计算与 IoU_{2D} 类似,可以被定义为

$$L_{\mathrm{IoU}} = 1 - L_{\mathrm{IoU}_{3D}} \tag{17}$$

3 实验评估

本文主要采用 KITTI 3D Object 数据集进行有效性验证。将官方给出的训练样本集划分为 3 712 个训练样本和 3 769 个验证样本。实验中所有模型分别在训练集和验证集上进行训练和分析比较。

3.1 与相关工作的对比

在 KITTI[17] 验证集中,将本文算法与其他算法进行对比,并计算了在汽车类别中 3 种难度下的平均检测精度,平均精度的计算方式采用 Recall 11 Position。比较结果如表 1 所示。

表 1 本算法与其他算法在 KITTI 验证集上汽车类别的平均精度对比

算法		平均精度/%		
		简单	中等	困难
体素化	SECOND	89.69	79.06	77.75
	VoxelNet	81.63	66.74	62.48
	MVFNet	**90.45**	79.60	75.02
	PointPillar	89.82	77.62	75.93
	SA-SSD	89.26	79.95	**78.47**
原始点云	PointRCNN	88.77	78.63	77.38
	PV-RCNN	89.15	81.05	78.30
	3D SSD	88.81	78.58	77.47
本文		89.44	**81.12**	77.30

由表中结果可见本文所提出的检测算法在性能上已经达到先进水平,特别是在中等难度下取得了81.12%的平均精度。上述结果可以证明,将 Transformer 用于体素化的点云检测是完全可行的。此外,本算法在推理过程中速度为 16.9 frame/s,达到了较快的处理速度。而与本算法性能相近的PV-RCNN推理速度仅为 8.9 frame/s,这主要归功于本算法中完全基于体素的特征提取,避免了体素和原始点云进行交互时产生的巨大计算开销。

3.2　消融实验

本节对基于 Transformer 的 3D 目标检测器头部网络中的主要组成部分进行了消融实验,以验证每个部分对检测效果的贡献。消融实验的结果如表 2 所示。

表 2　基于 Transformer 的 3D 目标检测器头部网络模块消融实验

主要模块	方法 A	方法 B	方法 C	方法 D
热图初始化目标查询	✓		✓	✓
空间调制交叉注意力		✓	✓	✓
解码器堆叠层数	3	3	3	6
中等难度目标检测精度/%	80.94	73.14	81.12	81.30

1) 方法 C 是浅层解码器堆叠下的完整的基于 Transformer 的 3D 目标检测算法,通过空间调制交叉注意力加速训练过程,并通过关键点热图对目标查询进行初始化,以达到降低网络深度、提高推理速度的目的。

2) 方法 A 相比于方法 C 删除了用于加速训练的空间调制交叉注意力模块,2 种方法的收敛曲线对比如图 4 所示。在训练次数同为 80 轮时,方法 C 的检测精度已经趋于平缓,基本收敛完毕。相比之下,未加入空间调制交叉注意力模块的方法 A,此时的检测精度为 75.9%,且波动幅度相对较大。训练过程截止到 150 轮时,方法 A 基本完成收敛,此时 2 个模型的精度差别可以认为是随机性误差。

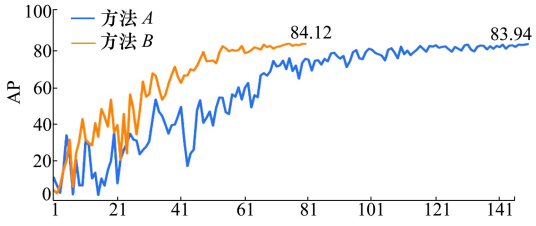


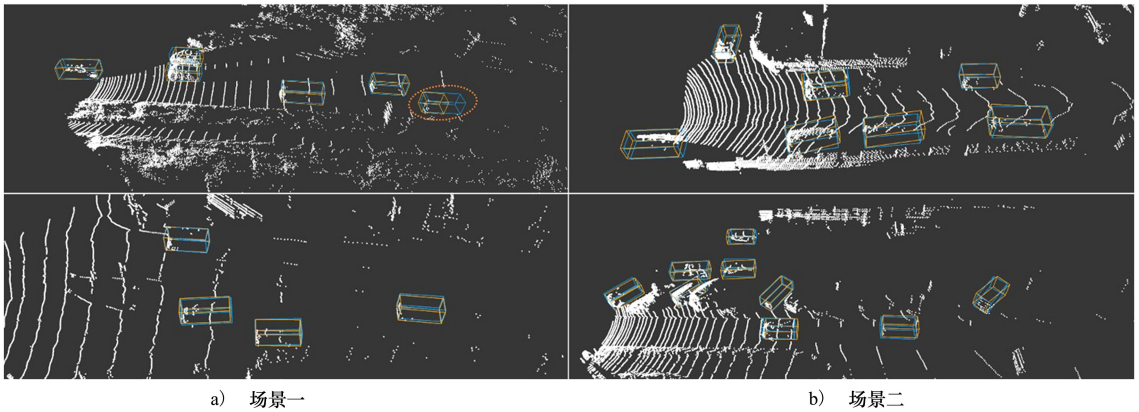
图 4　方法 A 和方法 C 的收敛曲线对比图

3) 方法 B 相比于方法 C 删除了热图初始化目标查询模块,此时解码器的目标查询采用零初始化方法。其检测精度相比于方法 C 下降了 7.98%,说明仅依靠 3 层解码器不足以将初始查询拟合至目标真实位置。采用热图初始化目标查询策略使得初始的查询点靠近潜在的目标位置,在浅层网络下取得了远超无此策略模型的性能。

4) 方法 D 相比于方法 C 将解码器层数提升至 6 层,其余模块未做变动,检测精度达到了 81.50%,推理速度为 12.7 frame/s。相比之下,计算成本的增加远远超过检测精度的提升,这也进一步证明了热图初始化查询策略的有效性。

3.3　可视化结果

图 5 是本文算法在验证集上的检测结果可视化演示。其中蓝色为目标真实框,黄色为本算法产生的预测框。图中的 4 个场景均包含有较多的车辆,第一个场景中由于最远处车辆距离较远且仅在车体尾部生成激光雷达点云,致使检测器对该目标尺寸



a)　场景一　　b)　场景二

图 5　基于 Transformer 的点云目标检测器可视化检测结果

产生误判。其余 3 个场景中车辆数量众多、停放角度复杂,而基于 Transformer 的 3D 目标检测算法依然能精确地定位并识别场景中的车辆。

4 结 论

本文提出了一种基于 Transformer 的 3D 目标检测算法。在推理过程中首先使用三维体素编码网络进行初步特征提取,而后压缩为鸟瞰图并采用二位骨干网络输出多尺度特征图,这些特征图在颈部网络中进行融合。在头部网络中,本算法采用了 DETR 中集合预测的检测思想,并采用热图初始化目标策略使得模型在浅层网络下依然能从特征图中有效汇集数据用于拟合回归。同时设计了一种改进的空间调制交叉注意模块用于训练过程中的加速。本算法在 KITTI 验证集中进行了对比和消融实验,结果均表明基于 Transformer 的 3D 目标检测算法作为一种简单而有效的算法,能够胜任三维目标检测及其他下游任务的应用。

参考文献:

[1] 李柯泉, 陈燕, 刘佳晨, 等. 基于深度学习的目标检测算法综述[J]. 计算机工程, 2022, 48(7): 1-12
LI Kequan, CHEN Yan, LIU Jiachen, et al. Survey of deep learning-based object detection algorithms[J]. Computer Engineering, 2022, 48(7): 1-12 (in Chinese)

[2] 董文轩, 梁宏涛, 刘国柱, 等. 深度卷积应用于目标检测算法综述[J]. 计算机科学与探索, 2022, 16(5): 1025-1042
DONG Wenxuan, LIANG Hongtao, LIU Guozhu, et al. Review of deep convolution applied to target detection algorithms[J]. Journal of Frontiers of Computer Science and Technology, 2022,16(5): 1025-1042 (in Chinese)

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//31st International Conference on Neural Information Processing Systems, New York, 2017: 6000-6010

[4] KIRILLOV A, USUNIER N, CARION N, et al. End-to-end object detection with transformers[C]//2020 European Conference on Computer Vision, Cham, 2020: 213-229

[5] 周全, 倪英豪, 莫玉玮, 等. FMA-DETR:一种无编码器的 Transformer 目标检测方法[J/OL]. (2023-10-16)[2023-11-30]. https://link.cnki.net/urlid/11.2406.TN.20231012.1541.014
ZHOU Quan, NI Yinghao, MO Yuwei, et al. FMA-DETR: a Transformer object detection method without encoder[J/OL]. (2023-10-16)[2023-11-30]. https://link.cnki.net/urlid/11.2406.TN.20231012.1541.014 (in Chinese)

[6] 廖峻霜, 谭钦红. 多粒度空间注意力与空间先验监督的 DETR[J/OL]. (2023-09-26)[2023-11-30]. https://link.cnki.net/urlid/50.1075.tp.20230925.0916.008
LIAO Junshuang, TAN Qinghong. DETR with multi-granularity spatial attention and spatial prior supervision[J/OL].(2023-09-26)[2023-11-30]. https://link.cnki.net/urlid/50.1075.tp.20230925.0916.008 (in Chinese)

[7] YAO Z, AI J, LI B, et al. Efficient DETR: improving end-to-end object detector with dense prior[J].(2021-08-03)[2023-01-09]. https://doi.org/10.48550/arXiv.2104.01318

[8] DUAN K, BAI S, XIE L, et al. CenterNet: keypoint triplets for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision, Piscataway, 2019: 6568-6577

[9] ZHU X, SU W, LU L, et al. Deformable DETR: deformable transformers for end-to-end object detection[C]//International Conference on Learning Representations, Montreal, 2020

[10] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//13th European Conference on Computer Vision, Piscataway, 2014: 740-755

[11] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

[12] 朱张莉, 饶元, 吴渊, 等. 注意力机制在深度学习中的研究进展[J]. 中文信息学报, 2019, 33(6): 1-11
ZHU Zhangli, RAO Yuan, WU Yuan, et al. Research progress of attention mechanism in deep learning[J]. Journal of Chinese Information Processing, 2019, 33(6): 1-11 (in Chinese)

[13] GAO P, ZHENG M, WANG X, et al. Fast convergence of DETR with spatially modulated co-attention[C]//2021 International

Conference on Computer Vision, Piscataway, 2021: 3601-3610
[14] 刘庆雯. 基于 Transformer 矢量化高清地图的构建[D]. 沈阳:辽宁大学, 2023
LIU Qingwen. Construction of vectorized HD map based on transformer[D]. Shenyang: Liaoning University, 2023 (in Chinese)
[15] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327
[16] ZHOU D, FANG J, SONG X, et al. IoU loss for 2D/3D object detection[C]//2019 International Conference on 3D Vision, Piscataway, 2019: 85-94
[17] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, 2012: 3354-3361

3D point cloud object detection algorithm based on Transformer

LIU Mingyang[1], YANG Qiming[2], HU Guanhua[2,3], GUO Yan[4], ZHANG Jiandong[2]

(1.Shenyang Aircraft Design Research Institute, Shenyang 110035, China;
2.School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China;
3.CSSC Systems Engineering Research Institute, Beijing 100094, China;
4.No.1 Military Representative Office of Equipment Department of PLA Airforce in Shenyang, Shenyang 110850, China)

Abstract: In response to the difficulty in deploying anchor box based methods in 3D object detection due to the increase in spatial dimensions, this paper studies a point cloud object detection algorithm based on set prediction. This article proposes a Transformer based 3D point cloud object detection algorithm, and combines the characteristics of point clouds in autonomous driving scenarios to propose an improved spatial modulation attention and heat map initialization strategy for training acceleration and query initialization, achieving good detection performance in shallow networks. This article compares it with other algorithms on the KITTI dataset, and the results show that our algorithm has reached an advanced level in performance. We also conducted ablation experiments on the main components of the algorithm to verify the contribution of each module to the detection effect.

Keywords: Transformer; spatial modulation attention mechanism; heat map initialization; target detection; deep learning

引用格式: *刘明阳, 杨启明, 胡冠华, 等. 基于 Transformer 的 3D 点云目标检测算法[J]. 西北工业大学学报, 2023, 41(6): 1190-1197*
LIU Mingyang, YANG Qiming, HU Guanhua, et al. 3D point cloud object detection algorithm based on Transformer[J]. Journal of Northwestern Polytechnical University, 2023, 41(6): 1190-1197 (in Chinese)