一种多层多模态融合3D目标检测方法

周治国,马文浩
(北京理工大学集成电路与电子学院,北京 100081)

摘 要: 在自动驾驶感知系统中视觉传感器与激光雷达是关键的信息来源,但在目前的3D目标检测任务中大部分纯点云的网络检测能力都优于图像和激光点云融合的网络,现有的研究将其原因总结为图像与雷达信息的视角错位以及异构特征难以匹配,单阶段融合算法难以充分融合二者的特征. 为此,本文提出一种新的多层多模态融合的3D目标检测方法:首先,前融合阶段通过在2D检测框形成的锥视区内对点云进行局部顺序的色彩信息(Red Green Blue, RGB)涂抹编码;然后将编码后点云输入融合了自注意力机制上下文感知的通道扩充PointPillars检测网络;后融合阶段将2D候选框与3D候选框在非极大抑制之前编码为两组稀疏张量,利用相机激光雷达对象候选融合网络得出最终的3D目标检测结果. 在KITTI数据集上进行的实验表明,本融合检测方法相较于纯点云网络的基线上有了显著的性能提升,平均mAP提高了6.24%.

关键词: 自动驾驶;多传感器融合;3D目标检测;点云编码;自注意力机制
基金项目: 装备预研领域基金(No.61403120109)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2024)03-0696-13
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220593

3D Object Detection Based on Multilayer Multimodal Fusion

ZHOU Zhi-guo, MA Wen-hao
(*School of Integrated Circuits and Electronics*, *Beijing Institute of Technology*, *Beijing* 100081, *China*)

Abstract: Camera and lidar are the key sources of information in autonomous vehicles (AVs) . However, in the current 3D object detection tasks, most of the pure point cloud network detection capabilities are better than those of image and laser point cloud fusion networks. Existing studies summarize the reasons for this as the misalignment of view between image and radar information and the difficulty of matching heterogeneous features. Single-stage fusion algorithm is difficult to fully fuse the features of both. For this reason, a nova 3D object detection based on multilayer multimodal fusion (3DMMF) is presented. First, in the early-fusion phase, point clouds are encoded locally by Frustum-RGB-PointPainting (FRP) formed by the 2D detection frame. Then, the encoded point cloud input is combined with the self-attention mechanism context-aware channel to expand the PointPillars detection network. In the later-fusion phase, 2D and 3D candidate boxes are coded as two sets of sparse tensors before they are not greatly suppressed, and the final 3D target detection result is obtained by using the camera lidar object candidates fusion (CLOCs) network. Experiments on KITTI datasets show that this fusion detection method has a significant performance improvement over the baseline of pure point cloud networks, with an average mAP improvement of 6.24%.

Key words: auto-driving; multi-sensor fusion; 3D target detection; point cloud coding; self-attention mechanism
Foundation Item(s): Equipment Pre-Research Field Foundation (No.61403120109)

1 引言

随着相关技术的发展,汽车产业正在朝着智能化、无人化的方向快速发展. 而环境感知技术需要为无人驾驶汽车提供准确、可靠的目标类别和位置信息,这是实现自主导航、避障的基础,也是进一步完成路径规划及车辆控制决策的先决条件[1]. 目前路面目标检测手段主要包括2D图像目标检测、3D点云目标检测、图像与雷达点云融合的目标检测等[2]. 从传感器的特性来说,视觉传感器所采集的图像信息提供了细颗粒度的上下文信息配合感知算法可以提供环境目标的分类结果,但其固有的深度模糊性导致不能获取到目标的距离和运动状态,且成像效果容易受到光照变化的影响;激光雷达可

收稿日期:2022-05-23;修回日期:2022-11-09;责任编辑:梅志强

以获取目标的位置及速度,几乎不受光照的影响,但不能提供颜色、纹理等视觉信息,且随着目标距离的增加,点云的稀疏性也就越明显. 激光雷达与视觉传感器信息所具有的较强互补性,利用多传感器融合的手段实现传感器的优势互补以达到更高精度的目标检测应该是比较理想的技术路线. 但研究现状显示:在目前KITTI数据集[3]的检测能力排行中,PointPillars[4],PointRCNN[5]和VoxelNet[6]这类纯激光点云的处理算法,大都优于现有的Frustum-PointNets[7]、MV3D(Multi-View 3D Object Detection Network)[8]和AVOD(Aggregate View Object Detection)[9]等多模态融合算法的表现.

根据融合多模态传感器数据的方式,将现有的图像与雷达点云融合的网络大致分为三类:(1)前融合网络;(2)后融合网络;(3)深度融合网络. 具体而言,前融合的方法通常利用单独的感知算法来处理多模态原始传感器数据,可以看作是一种"数据级别的融合". 这类算法需要多模数据的精确对齐,如果原始传感器数据在早期没有很好地配准或者某一类传感器出现了故障,其造成的特征错位将导致检测性能严重下降. 目前比较典型的前融合网络有PointPainting[10]和PI-RCNN(Point cloud Image RCNN)[11]. 尽管这类前融合方法实现了图像语义向点云空间的传递,但是也同时将一个模态中的噪声传递到另一个模态,这种噪声与点云中的物体形状特征对齐和组合,明显破坏某些物体的点云空间特征的突出性. 基于后融合的方法仅在决策层融合已处理的特征,因为点云和图像之间的空间和模式差异在这一阶段被大大减小,因此有时也被归类为"决策级别的融合",但是这类方法对原始数据信息融合的影响不大,如何合理的利用两种模态生成方案的置信度也是一类课题性研究. 基于深度融合的方法一般比较灵活且复杂,网络构造各不相同:有类似Frustum-PointNets这类两阶段级联式融合,也有类似MV3D、AVOD这类并联方式的特征融合,这些方法都利用原始和高级语义信息,但是目前来看大都运算量大,检测指标也不是很理想.

总结出目前多模态融合的3D目标检测算法表现不佳的主要原因与挑战:图像视角与激光点云视角的错位、异构数据格式差异等带来的特征不匹配. 目前大多数的3D目标检测网络,会将点云数据压缩或投影到鸟瞰图后进行后续处理,这样会避免在主视角投影产生的深度模糊的现象,以达到更好的检测效果. 显然,激光雷达可以很容易的完成到鸟瞰图的投影,但是在主视角的图像信息却因为深度模糊特性很难做到这一点;另外,目前的融合算法中为了融合来自异构模态的特征向量而使用的裁剪和调整大小操作可能破坏来自每个传感器的特征结构. 图像是高分辨率的密集数据,雷达点云是低分辨率的稀疏数据,连接、聚合二者向量的同时完成合理的特征匹配是有挑战的. 要解决以上问题,一次性将两种模态的特征进行充分的对齐融合是困难的,并且在不同融合的阶段,色彩信息(Red Green Blue,RGB)与激光雷达(Light Detection and Ranging,LIDAR)特征的置信度也在动态变化:在浅层的融合阶段,图像的RGB所蕴含的语义和上下文信息置信度更高;在深度编码后,点云的空间上下文信息变得更有价值. 由此,本文提出了一种多层多模态融合3D目标检测方法(3D Object Detection based on Multilayer Multimodal Fusion,3DMMF),通过图像特征与点云特征在多个阶段充分融合,且在各融合阶段对不同模态的特征置信度有所侧重,最终提高融合后网络的检测精度.

在KITTI数据集上对本方法进行实验验证,结果显示本多模态多层融合检测方法效果明显,相较于纯点云网络的基线指标上有了显著的性能提升,且各级部件对指标都有所增益.

2　多层多模态融合3D目标检测框架

如图1所示,3DMMF方法为顺序串行结构,主要包含三个层级:前融合采用改进自PointPainting方法的局部顺序融合编码方法(Frustum-RGB-PointPainting,FRP),将点云信息扩充为8个维度,将图像RGB信息中的浅层语义与初始点云对齐;3D目标检测主干网络采用了集成全局自注意机制(Full Self-Attention Module,FSA)上下文感知模块[12]的PointPillars,以解决点云扩充编码后参数量倍增,全局点云特征难以提取带来的高虚警率等问题;后融合采用了相机激光雷达候选目标融合网络(Camera-LiDAR Object Candidates fusion,CLOCs)[13],利用2D与3D目标检测结果的几何空间和语义的一致性,提高总体的3D目标检测的准确率. 我们提出的算法主要流程如下:

步骤1　由2D图像目标检测算法提出2D候选框.

步骤2　在非极大抑制(Non-Maximum Suppression,NMS)后的候选框形成的锥视区内进行局部顺序的彩色点云编码,将空间点云向图像投影,在锥视区对点云进行彩色涂抹,将点云在局部区域内进行顺序编码融合,在点云"反射率"通道后附加一个"推荐通道"和三个"颜色通道".

步骤3　将编码后点云输入通道扩充的PointPillars网络,利用卷积神经网络模型提取空间特征,同时引入全局自注意机制提取上下文特征,将两种特征串联后送入SSD检测头,提出3D目标检测候选框.

步骤4　最后再次利用步骤1产生2D候选框与步骤3产生的3D候选框在非最大抑制之前编码为两组稀疏张量,输入相机激光雷达对象候选融合网络得出最

终的3D目标检测结果.

这种方法在多个阶段实现了两种模态浅层特征和高级特征的融合,并且只使用了一次2D目标检测算法,使其同时直接作用于前融合和后融合两个阶段的跨模态信息融合,在提升检测精度的同时保证了算法效率.

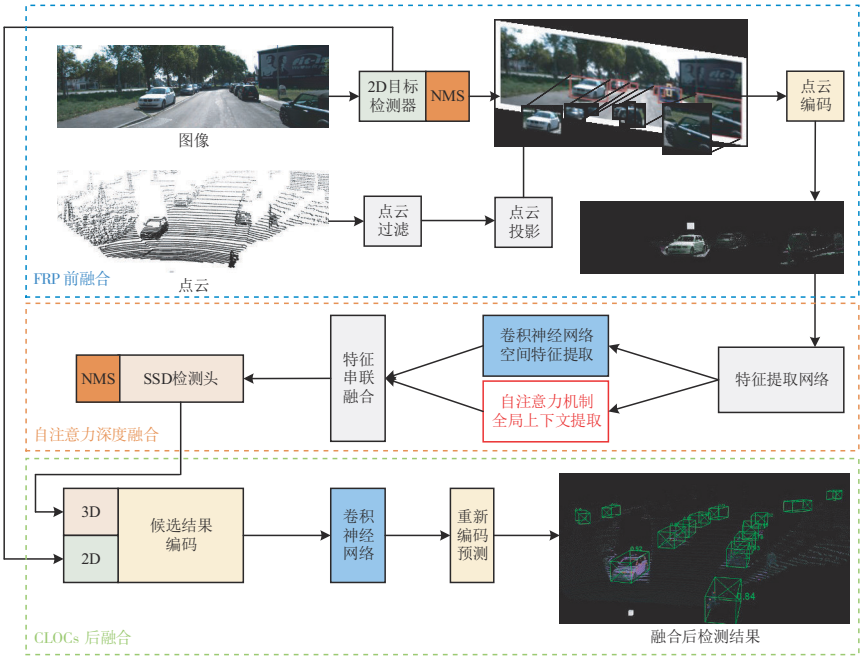


图1 多层多模态融合3D目标检测

2.1 锥视区点云彩色涂抹(Frustum-RGB-Point-Painting)前融合方法

为了改善前融合阶段图像视角与激光雷达视角错位的问题,PointPainting的工作创新性的提出了一种简单而有效的激光雷达点云与图像信息的顺序融合方法:将每个激光雷达点投影到图像语义分割网络的输出中,并将信道附加到每个激光雷达点的反射强度通道之后. 编码后的激光雷达点可以用于任何激光雷达检测方法,无论是鸟瞰图处理或前视图处理,且大多数的纯点云网络在经过涂抹后检测性能都有所提升. PointPainting解决了以往融合概念的缺点:它没有对三维检测体系结构添加任何限制;它不会出现特征或深度模糊;它不需要计算伪点云;也不限制最大召回率. 虽然在PointPainting方法实际运行的过程中,也暴露出一些问题:(1)通常涉及2D分割算法的网络结构都比较复杂,导致前传等待时间较长. 为了改善此问题,作者提出了一种异步匹配的方法:采用临近时间帧的图像分割结果来为本帧的雷达点云进行涂抹,但对于高速移动的路面目标检测,鲁棒性有所下降;(2)3D检测网络输入增加的通道数量和执行语义分割的类别数量一致,在KITTI数据集中需要扩充4个通道,在类别细分较多的nuScenes数据集[14]中需要扩充11个通道,明显增加了计算成本;(3)每次的点云"涂抹",都要进行全局的点云编码,在占最大面积的无用背景类别中消耗计算量较大;(4)图像的语义掩膜特征非常突出,分割后的噪声点会传递到点云模态中,很容易在目标的后方位置出现虚警. 但是这种创新性的将点云与图像语义信息进行数据级初级融合,通过对点云进行附加通道编码来提高每个点云信息量,最终提高3D检测精度的方法,是有效果的. 在PointPainting的思想基础上,针对其存在的问题,本文提出了锥视区点云彩色涂抹(Frustum-RGB-PointPainting,FRP)——一种图像与点云的局部融合方法,如图2所示.

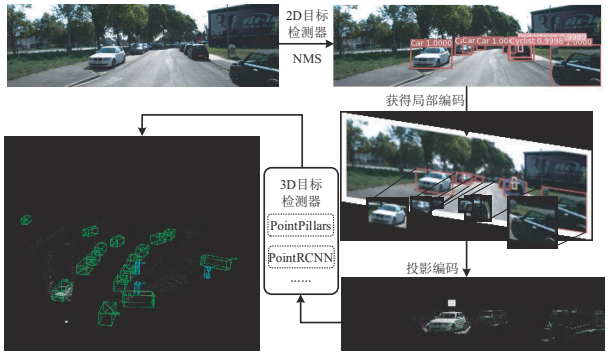


图2 Frustum-RGB-PointPainting方法处理流程

这种方法将在2D目标检测框架下进行,利用图像检测结果在点云中形成投影后的目标所在锥视区,在此局部区域内对为每个点云附加四个通道的编码,将原有的

KITTI点云信息由(x,y,z,r)扩充编码为(x,y,z,r,S,R,G,B),其中x,y,z为点云的空间位置信息,r为激光雷达反射强度,S为推荐通道,R,G,B为点云对应的颜色通道. 激光雷达点云经过均匀变换后,向2D图像上投影,利用KITTI数据集中自带的内、外参数据可以完成这一工作:

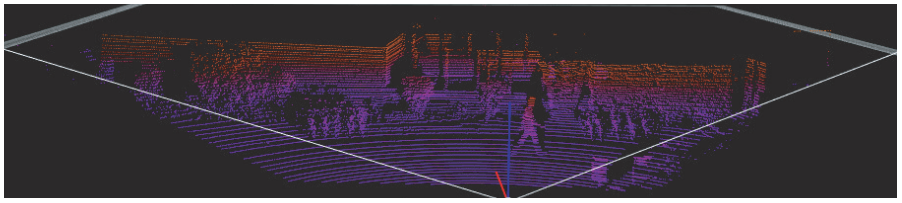
$$\boldsymbol{s}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{v}\\ 1\end{bmatrix}=\begin{bmatrix}\alpha_x & 0 & u_0 & 0\\ 0 & \alpha_y & v_0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\Phi} & \boldsymbol{\Delta}\\ \boldsymbol{0}^{\mathrm{T}} & 1\end{bmatrix}\begin{bmatrix}x_{\mathrm{L}}\\ y_{\mathrm{L}}\\ z_{\mathrm{L}}\\ 1\end{bmatrix}=\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{X}_{\mathrm{L}} \quad (1)$$

式中,$\alpha_x=f/w,\alpha_y=f/h$,矩阵$\boldsymbol{M}_1$为相机内参矩阵,与每个相机的传感器物理指标而不同,$\boldsymbol{M}_2$为外参齐次变换矩阵,是描述相机到激光雷达物理间的旋转平移关系,由多传感器联合标定取得. 最后可以得到激光雷达坐标系上的点云$\boldsymbol{P}(x_{\mathrm{L}},y_{\mathrm{L}},z_{\mathrm{L}})$到其图像坐标系的位置$\boldsymbol{p}(u,v)$的映射关系.

推荐通道S采用了如图3中Frustum-PointPillars的"概率掩膜". 主要依据目标分布概率信息:2D检测框的中心附近更有可能被对象占据,中心区域附近的投影三维点也更有可能属于对象,而不是背景杂波,因此将属于对象的点的概率定义为高斯函数:

$$L(\bar{x},\bar{y})=\exp\left(-\frac{(\bar{x}-\bar{x}_0)^2}{2w^2}-\frac{(\bar{y}-\bar{y}_0)^2}{2h^2}\right) \quad (2)$$

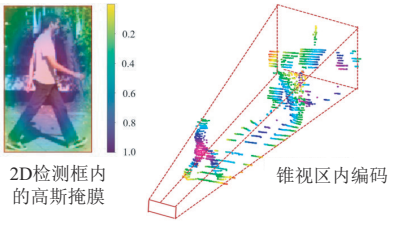
其中,$\bar{x}$、$\bar{y}$是图像平面上的点云投影,$\bar{x}_0$、$\bar{y}_0$是中心坐标,w、h是2D检测框的宽度和高度. $\alpha=w^2$、$\beta=h^2$定义似然函数的曲率. 将推荐似然值作为附加特征向量添加到原始点云的反射强度r后. 如果一个点由多个二维边界框共享,则选择最大似然值. Frustum-RGB-PointPainting方法实现见算法1.



(*a*) 原始点云

(*b*) PointPainting

(*c*) Frustum-PointNets



(*d*) Frustum-PointPillars

(*e*) Frustum-RGB-PonitPainting
(背景点云并没有被删除)

图3　各融合方法对点云处理的直观区别

将原有PointPainting方法的语义分割网络框架替换成检测网络框架的原因之一是分割任务的总体难度要高于检测任务. 对于比较简单的分割任务只需要一个布尔值来指示对象是否存在,但是如果需要分割多个类别,它就会变得更加复杂. 例如,在KITTI数据集中可以将路面场景分为4类:背景、车辆、行人和自行车骑手,需要将图像中的每个像素点进行四选一的分类;而在nuScenes数据集中,类别信息上升到了11种,分割任务的复杂度则大大增加. 因为图像分割的处理时间的问题,PointPainting折中提出了如图4这种"流水线"的处理方案,减少因为3D检测网络对分割结果等待而开销太多的前传时间,虽然这对nuScenes数据集这样带时间标签,且图像前后比较连贯的数据集有所改善,但是在KITTI数据这样没有时间标签的数据集上就无法减少前传时间;更重要的是,根据图5引用的PointPainting文中的消融实验结果显示,如果前后帧图像分割结果的mIOU下降到50%以下,对3D目标的检测能力将会急剧下降. 在车辆行驶过程中,如果自身运动速度较快,或者视野内有移动的小目标,导致物体在前后两帧图像中的位置差别较大,而3D检测网络接收到的还是利用前几帧的图像分割结果来涂抹编码的点云数据,就会造成误检或者漏检,导致算法的鲁棒性降低.

将前向串行网络替换为2D检测网络的另一个原

算法 1 Frustum-RGB-PointPainting 点云编码

输入:激光雷达点云 $L \in \mathbb{R}^{N,D}$(N为雷达反射点数目,D为点云维度,KITTI中$D=4$).

推荐通道 $S \in \mathbb{R}^{W,H}$

颜色通道 $C \in \mathbb{R}^{W,H,3}$ 分别对应 R,G,B 三个颜色通道.

外参齐次变换矩阵 $\boldsymbol{M}_2 \in \mathbb{R}^{4,4}$.

相机内参矩阵 $\boldsymbol{M}_1 \in \mathbb{R}^{3,4}$.

输出:局部彩色涂抹编码后的雷达点云 $\boldsymbol{P} \in \mathbb{R}^{N,D+1+3}$

方法:

FOR $\boldsymbol{l} \in L$:

1. 点云投影:

$\boldsymbol{l}_{\text{image}} = \text{投影}\left(\boldsymbol{M}, \boldsymbol{T}, \boldsymbol{l}_{\text{point}(x,y,z)}\right)$ ▷ $\boldsymbol{l}_{\text{image}} \in \mathbb{R}^2$

2. 推荐通道获取:

IF 雷达点云在2D检测框:

$S = L(\bar{x}, \bar{y})$

ELSE:

$S = 0$

$\boldsymbol{s} = S(\boldsymbol{l}_{\text{image}}[0], \boldsymbol{l}_{\text{image}}[0])$ ▷ $\boldsymbol{s} \in \mathbb{R}$

3. 颜色通道获取:

$\boldsymbol{c} = C(\boldsymbol{l}_{\text{image}}[0], \boldsymbol{l}_{\text{image}}[0], :)$ ▷ $\boldsymbol{c} \in \mathbb{R}^3$

4. 点云编码:

$\boldsymbol{p} = \text{连接编码}(\boldsymbol{l}, \boldsymbol{s}, \boldsymbol{c})$ ▷ $\boldsymbol{p} \in \mathbb{R}^{D+1+3}$

END FOR

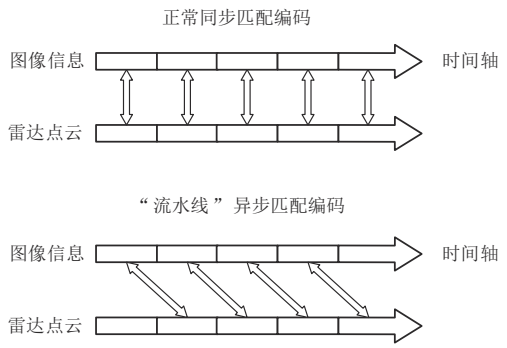


图4 “流水线”方法解决等待图像分割时间过长的问题

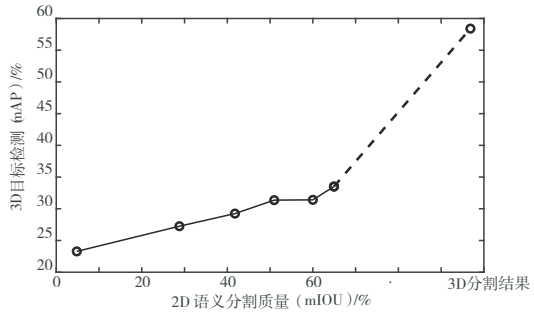


图5 2D分割质量对3D目标检测的影响

因,则是图像上的分割结果投影到3D点云空间上,相较于纯点云的基线指标提升不够明显.在经典的融合网络Frustum-PointNets试验利用2D图像分割掩膜对点云目标进行预测,发现效果并不如3D分割后的预测结果;甚至依靠2D和3D的共同分割结果,2D分割结果造成的累积误差还会拉低3D分割后预测的准确率.因为在典型2D掩膜预测的任务中,虽然估计的2D掩膜在RGB图像上显示的质量较好,但2D掩膜中仍然存在大量杂波和前景点,这在由2D空间投影向3D空间后会更加明显.如图6所示,像素级的分割误差,可能会造成3D点云空间中几十米的距离误差.而传统的PointPainting方法对多通道进行布尔编码,将雷达点云向图像进行投影后,分辨其对应的像素位置的掩膜类别,将对应类别通道编码为“1”,同像素的其他类别通道的编码为“0”.PointPillars将点云分割为柱体网格后,利用简化PointNet进行特征提取和点云编码,每一个柱体为一个编码单位.由于点云的无序性,超出同一柱体范围以外的点云,将很难被再次关联起来.而2D语义掩膜的特征非常明显,如果某一柱体内的点云有被某类别的2D掩膜捕捉到,那么这个柱体大概率会被归类为这个类别目标.这就导致了两个现象:(1)对行人和远处目标检测准确度提升非常明显,因为这些点云大都在一个柱体内被编码,2D掩膜特征将会非常有效.(2)实验结果显示,对于车辆等目标,由于2D掩膜的“毛刺噪声”被投影到点云空间,会在目标的身后出现大量虚警,反而降低检测的“精确率”.而在FRP的方法中,由2D检测框形成的建议区域,及在锥视区内的局部点云编码,为目标区域点云增加了额外的信息量,对3D检测网络学习有所指引;同时保留了建议区域内的RGB颜色信息,虽然在信息增维的过程中加入了小部分不属于目标的背景点云,但是目标的边界信息可以通过颜色的落差更加准确的体现出来;此外,虽然雷达点云在图像上的投影相当于对二维空间的图像进行稀疏采样,但是目标的颜色、材质和纹理都在一定程度上得到保留,实验表明,即便散落在各个柱体网格中,目标特定的RGB属性对3D检测还是有所增益.

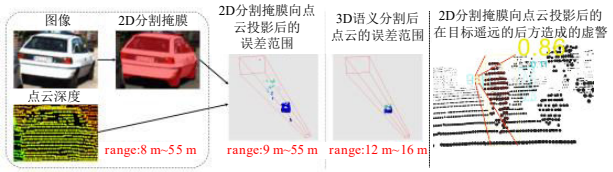


图6 2D掩膜噪声在3D空间遥远的后方位置造成的误检

本文的2D检测网络采用YOLOv3网络[15],其检测精度,稳定性以及速度都表现的比较出色,是工程界首选的检测算法之一.YOLOv3是在YOLOv1[16],YOLOv2[17]的基础上改进而来,重点优化了小目标的检测精度,相较于主流的SSD和Faster R-CNN网络,在保持了相同精度的同时,将检测速度提高了两倍以上.超快

的处理速度,更加符合FRP这种前融合的串行数据处理结构,避免其在2D图像处理阶段消耗过多的时间,导致出现和PointPainting一样前传时间过久的问题.

FRP需要2D图像目标检测器在融合阶段提供较为精准的目标位置语义,候选框要位置精准且尽可能少.因此在完成前融合编码前,需要将YOLOv3的检测结果引入非极大抑制,在候选框的领域空间进行局部最优解搜索,在2D目标有可能出现的最高概率区域形成锥视区.而将未引入NMS前的2D候选框结果,将直接传递给相机激光雷达对象候选网络中.

2.2　3D目标检测主干网络

如图7所示,3DMMF中采用的3D目标检测的主干网络,是在PointPillars的基础上修改而来的,PointPillars首先将点云以竖状"柱体"的形式转换为稀疏伪图像,再利用2D卷积网络对伪图像进行检测并预测3D检测框.而在前融合FRP编码阶段,扩充了原始点云的通道维度,需要对应的修改3D目标检测器的输入尺寸.此外,编码后的点云通道数扩充了一倍,使得点云在鸟瞰方向形成的"柱体"张量提取特征形成稀疏伪图像的过程中,伪图像的单个位置所代表的特征信息量也增加了一倍,这会一定程度的模糊物体点云的空间形状特征,造成一些高置信度的虚警.为了适应这种点云的额外属性,必须增加特征提取卷积层的深度,以使原本的卷积核的感受野足够大以捕获编码后点云信息,但是输入的点云特征相对重要性随着网络加深而下降、且跨多个层协调参数优化以捕获数据中的模式是有挑战性的.为此,我们为PointPillars引入了上下文感知自注意力机制模块:自注意的核心思想是获取全局点云信息,将所有点云位置的特征加权求和到目标位置,通过嵌入空间中各位置特征之间的相似度函数动态计算相应的权重.这种特征提取器可以学习全局点云之间的相关性,从而产生更强大、更独特和更健壮的特征.例如,在同一车道上行驶的汽车、自行车的方向特征之间,行人的高度和身着颜色特征之间就有明显的相关性,这种上下文感知自注意力机制就可以提取路面目标的这类特征的相关性,更加充分利用彩色点云信息,以产生更准确的检测,特别是对于距离较远的汽车,较小的行人目标等.

如图8所示,PointPillars网络会对输入点云进行划分表示,将尺寸为0.16 m×0.16 m的$H\times W$个网格均匀且连续的分布在Oxy平面上的网格中,形成一组"柱体"集合.同时在这个阶段中,前融合中被FRP编码后的8维点云(x,y,z,r,S,R,G,B)将被增强表示为13维$(x,y,z,r,x_c,y_c,z_c,x_p,y_p,S,R,G,B)$,其中$x,y,z$为原始点云位置坐标,$r$为反射率,$x_c,y_c,z_c$为"柱体"中所有点的算数平均中心坐标,$x_p,y_p$表示算数中心与"柱体"中

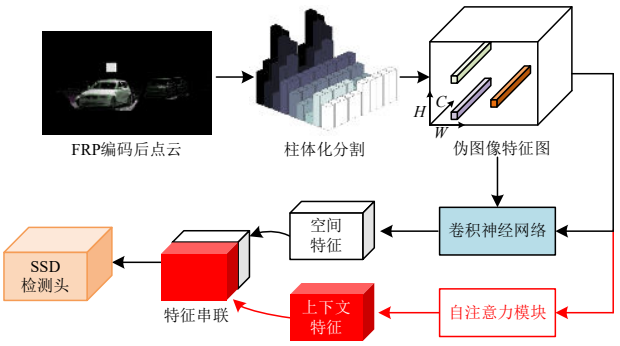


图7　PointPillars主干网络引入自注意力机制

心的偏移,S为推荐通道,R,G,B为点云对应的颜色通道.通过限制每个样本的非空"柱体"数P和每个"柱体"中的点个数N,使得点柱数据保持0~97%的稀疏度,如"柱体"中的数据太多,则进行随机采样,如数据太少,则用"0"进行填充,以此建立了一个大小为(D,P,N)的张量,其中D代表了每个点云的特征维度,本文FRP编码后为13(原始PointPillars为9),P代表了所有非空的柱体,N代表了每个柱体中最多点云数目.接着采用一个简化的端对端网络PointNet对张量化的点云数据进行进一步处理和特征提取,即对每一个点都运用一个线性层+ BatchNorm层+ReLU层以生成一个大小为(C,P,N)的张量,再在通道上进行最大池化,得到一个大小为(C,P)的张量.最后,将编码后特征按照原始柱的位置组合堆积起来,散布回原始的"点柱"位置,创建大小为(C,H,W)的伪图像特征图X,其中C为通道数,H和W表示伪图像的高度和宽度.

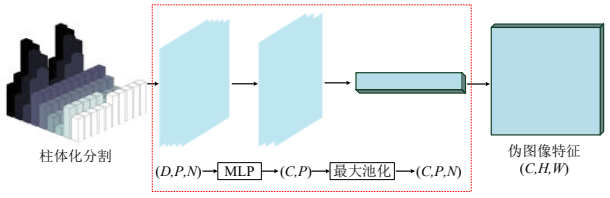


图8　伪图像生成过程

其中线性层也称为全连接层,常用于特征映射,输入张量在经过网络后会输出特征新的特征张量,本文中二维卷积算子表示为Conv2D(c_{in},c_{out},k,s),其中c_{in}和c_{out}是输入和输出通道的数量,k和s分别是卷积核大小和步长.本文选用了1×1的卷积层,主要是为了提升特征维度的同时保持计算效率.线性层卷积为Conv2D(13, 64, (1, 1), 1)以适应原始点云数据的通道数扩充,将$(13,P,N)$特征扩充为$(64,P,N)$.

形成伪图像后,可以采用比较先进的2D特征处理办法来提取点云特征,其主要分为两个部分:基于CNN骨干网络的空间特征提取器与基于自注意力机制模块的上下文特征提取器,提取更加全面有效的全局点云特征.

如图9所示，CNN骨干网络采用类似FPN结构分为下采样连接网络和上采样连接网络，共进行了3层的卷积操作. 下采样网络通过卷积，提取更抽象高层的伪图像特征，同时减小空间分辨率，从而捕获不同尺度下的特征信息；上采样网络通过反卷积，保持特征图像尺寸与原始伪图像对齐，从而将不同尺度特征信息融合.

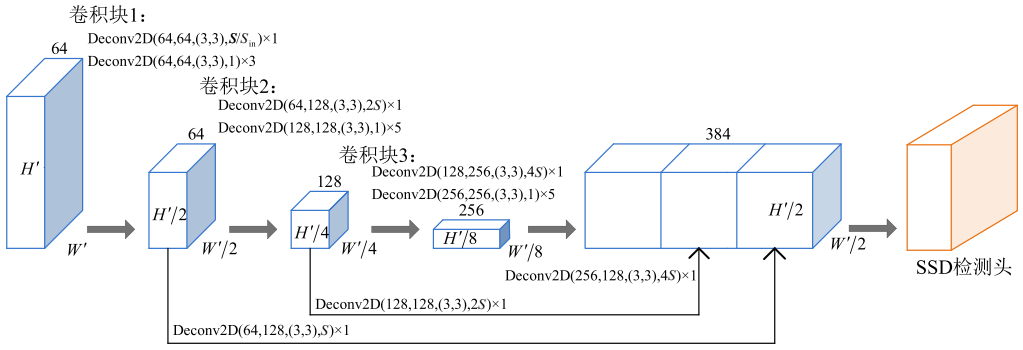


图9 基于CNN骨干网络的空间特征提取器

下采样网络由(S,H,O)的全卷积块来表征，其中S为步长，H为3×3的二维卷积的层数，O为输出通道数. 同样的，每一层卷积后与BatchNorm(BN)层和Rectified Linear Unit(ReLU)层相连，卷积层内的第一个步幅需要设为S/S_{in}，才能确保网络块操作在接收到步长S_{in}输入后，仍保持为S运行，且层中后续卷积的步长均为1. 伪图像输入的特征通道$C=64$，因此三层下采样块为：Block1$(S,4,64)$，Block2$(2S,6,128)$，Block3$(4S,6,256)$. 根据目标大小特征控制输入块步幅：汽车类目标$S=2$，行人和骑手这类小目标$S=1$.

上采样可以表示为$(S_{\text{in}},S_{\text{out}},f)$，$S_{\text{in}}$是初始步长，$S_{\text{out}}$为最终输出步长，得到$f$个输出特征. 使用2D反卷积得到$f$，添加BN层和ReLU层提取特征，三层上采样块为：Up1$(S,S,128)$，Up2$(2S,S,128)$，Up3$(4S,S,128)$. 最终的输出特性是来自各个步长的特性相组合连接，输入后续SSD检测模块.

如图10所示，在全局自注意力机制模块提取点云上下文特征前，需要将由“柱体”提取后的伪图像特征$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$在数学上进一步表示形成一个连接图$\boldsymbol{G}=(v,\varepsilon)$，图中的结点集合$v=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\in\mathbb{R}^C)$由伪图像特征构成，将结点两两相连构成边集合$\varepsilon=(\boldsymbol{r}_{ij}\in\mathbb{R}^{N_h},i=1,\cdots,n;j=1,\cdots,n)$，边$\boldsymbol{r}_{ij}$表示第$i$个节点和第$j$个节点之间的关系，$N_h$空间为多头(Multi-Head)头数，即为注意力图的总数，自注意力模块获取特征结点集v后会计算边集ε. 这种将点云特征表示为连接图中的结点的方式，会把对全局上下文信息聚合的过程转化为类似于追溯连接图上的结点之间的信息传递的过程，这将更有利于提取全局点云之间的高阶交互.

如图11所示，基于全局自注意力机制模块通过线性层将处理后的特征图向查询向量$\boldsymbol{Q}$、键向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$映射，其中$\boldsymbol{V}$表示输入特征，$\boldsymbol{K}$和$\boldsymbol{Q}$用于计算自注意图中特征的重要性. 查询元素$\boldsymbol{q}_i$与每一个键$\boldsymbol{k}_{j=1:n}$进行相似度点积计算得到权重：

$$\begin{cases} f(\boldsymbol{Q},\boldsymbol{K}_i),i=1,2,\cdots,n \\ f(\boldsymbol{Q},\boldsymbol{K}_i)=\dfrac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}_i}{\sqrt{D}} \end{cases} \tag{3}$$

使用Softmax函数对这些权重进行归一化，转换为注意力权重$\boldsymbol{w}_i$：

$$\boldsymbol{w}_i=\frac{\mathrm{e}^{f(\boldsymbol{Q},\boldsymbol{K}_i)}}{\sum_{j=1}^{m}f(\boldsymbol{Q},\boldsymbol{K}_j)},i=1,2,\cdots,m \tag{4}$$

使用注意力权重计算连接图中的成对交互项$\boldsymbol{r}_{ij}=\boldsymbol{w}_{ij}\boldsymbol{v}_j$. 每个结点向量$\boldsymbol{a}_i$的累积全局上下文信息就是这些成对交互项的总和$\boldsymbol{a}_i=\sum_{i=1:n}\boldsymbol{r}_{ij}$，因此多个并行的注意力头可以独立地提取通道依赖关系. 将这种累积全局上下文向量跨注意力头的连接起来，生成节点i的最终输出$\boldsymbol{a}_i^{h=1:N_h}$，将其通过一个线性层，用层归一化(LayerNormalization)进行归一化处理，然后用原输入特征图中的$\boldsymbol{x}_i$进行残差计算连接$F(\boldsymbol{x}_i)+\boldsymbol{x}_i$得出全局自注意力机制模块的输出.

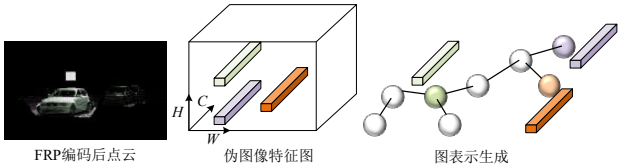


图10 图卷积生成

空间特征提取器的输出特征将于上下文提取器的输出特征连接，合并输入检测头SSD网络中. SSD是一种单步检测算法，具有检测速度快、精度高、多尺度适应性好等优点，保证了PointPillars网络的高效、高速的

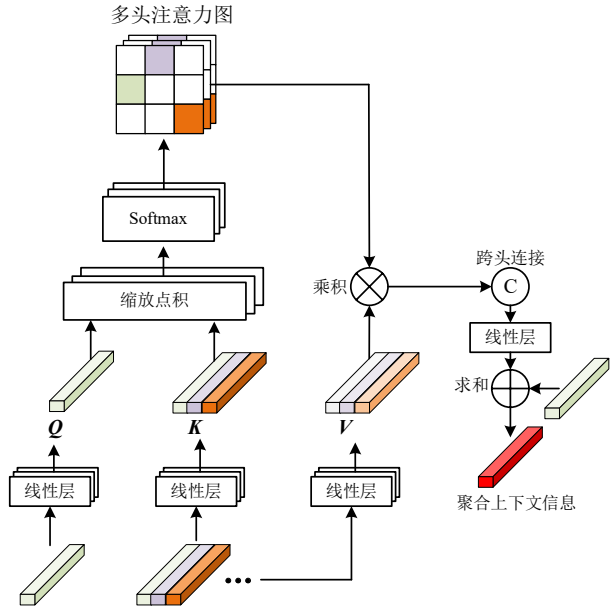


图11　上下文感知自注意力机制模块

检测能力. SSD算法会预测物体的种类、空间位置、三维大小与物体朝的锚框与真实框gt在BEV俯瞰视角进行匹配,而忽略考虑高度信息匹配,主要有两个原因:(1)KITTI数据集中所有的物体都是在同一个Oxy平面上的,没有某车辆位置升高的情况;(2)路面上的主要目标的高度差别不大,使用SmoothL1回归就得到了比较好的结果. 而SSD的预测结果将在非极大抑制(NMS)之前送入后融合网络CLOCs中.

本方法中的损失函数有回归、分类两部分组成,回归损失用于给出目标的位置、大小和方向,分类损失用于给出目标的类别. 其中3D边界框由参数(x,y,z,w,l,h,θ)定义,其中x、y、z为边界框中心坐标,w、l、h为边界框的分别为边界框的宽度、长度和高度,θ为物体朝向.

目标和锚框之间的回归损失为:

$$\begin{cases}\Delta x=\dfrac{x_{\mathrm{gt}}-x_a}{d_a},\Delta y=\dfrac{y_{\mathrm{gt}}-y_a}{d_a},\Delta z=\dfrac{z_{\mathrm{gt}}-z_a}{h_a}\\ \Delta w=\log\dfrac{w_{\mathrm{gt}}}{w_a},\Delta l=\log\dfrac{l_{\mathrm{gt}}}{l_a},\Delta h=\log\dfrac{h_{\mathrm{gt}}}{h_a}\\ \Delta\theta=\sin(\theta_{\mathrm{gt}},\theta_a)\end{cases}\tag{5}$$

其中,下标gt为真实框中的值,下标a为预测框中的值,d_a的计算方法为:

$$d_a=\sqrt{(w_a)^2+(l_a)^2}\tag{6}$$

真实值和预测值之间的定位误差L_{loc}为:

$$L_{\mathrm{loc}}=\sum_{b\in(x,y,z,w,l,h,\theta)}\mathrm{SmoothL1}(\Delta b)\tag{7}$$

其中,L1平滑函数为:

$$\mathrm{smoothL1}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & \text{其他}\end{cases}\tag{8}$$

总的分类损失为:

$$L_{\mathrm{cls}}=-\alpha_a(1-P_a)^{\gamma}\log P_a\tag{9}$$

式中P_a为锚框的分类概率;超参数设定为$\alpha=0.25$,$\gamma=2$.

由于角度定位损失无法区分翻转的3D候选框,在离散化方向上使用Softmax分类损失L_{dir},使网络能够学习正确的行驶方向,最后通过加权和的方式得到总的损失函数:

$$L=\frac{1}{N_{\mathrm{pos}}}(\beta_{\mathrm{loc}}L_{\mathrm{loc}}+\beta_{\mathrm{cls}}L_{\mathrm{cls}}+\beta_{\mathrm{dir}}L_{\mathrm{dir}})\tag{10}$$

其中,N_{pos}为正确框数目;超参数设定为$\beta_{\mathrm{loc}}=2$,$\beta_{\mathrm{cls}}=1$,$\beta_{\mathrm{dir}}=0.2$.

2.3　基于相机激光雷达对象候选网络(CLOCs)的后融合方法

3DMMF中采用的后融合方法为相机激光雷达对象候选网络CLOCs,CLOCs使用FRP阶段的2D图像检测器与3D目标检测主干网络输出的候选框在非极大抑制之前对组合输出候选对象进行操作训练,通过利用跨模态信息,它可以保留可能被某一单模态方法错误地抑制的候选检测结果,根据2D候选框与3D候选框的几何和语义的一致性来产生更加精确的3D目标检测结果,如图12所示.

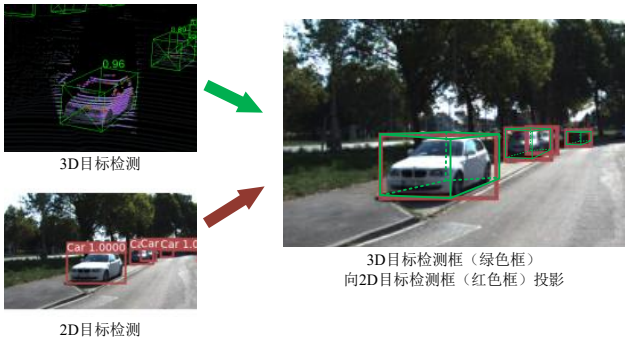


图12　3D候选框与2D候选框的几何和语义一致性

几何一致性认为:如果在2D检测和3D检测上都同时检测到目标,那么在图像平面上将至少有一个高度重叠的边界框. 而单一模态的虚警目标很难同时在两个模态下有相同的边界框. 这种几何一致性可以利用2D候选框与3D候选框投影到2D平面后的交并比(IOU)进行量化. 语义一致性则可以认为:每一个模态的检测器都存在多种类别的输出,CLOCs只对同类别的目标做关联融合.

因为后融合阶段决策器不会再创建新的检测框,而是对已有两种模态的检测结果的选择,由此在CLOCs输入阶段需要对两种单模态检测器进行调节以

其召回率最大化，而不是精确度最大化：高召回率意味着可能伴随更多的误检，但是尽可能检测到每一个应该被发现的目标．具体来说就是需要舍弃非极大抑制（NMS）算法或者把IOU阈值尽可能降低，因为这可能会在某一个模态中错误地抑制真实检测框．

CLOCs的主要网络结构如图13所示，主要分为四个阶段：（1）2D和3D目标检测器分别提出非NMS的候选框；（2）将这两种模态的候选结果编码成稀疏张量（3）对非空元素采用二维卷积完成对应的特征融合；（4）通过最大池化（max-pooling）将处理后的张量映射到期望的学习目标，即概率得分映射．

将两种模态的候选结果编码为稀疏张量的过程，需要将2D与3D检测候选对象转化为同一目标的联合检测候选对象，共同送入融合网络中进行处理．2D目标检测器的输出是图像平面上的一组2D检测框和相应的置信度．对于一幅图像中的候选结果，定义如下：

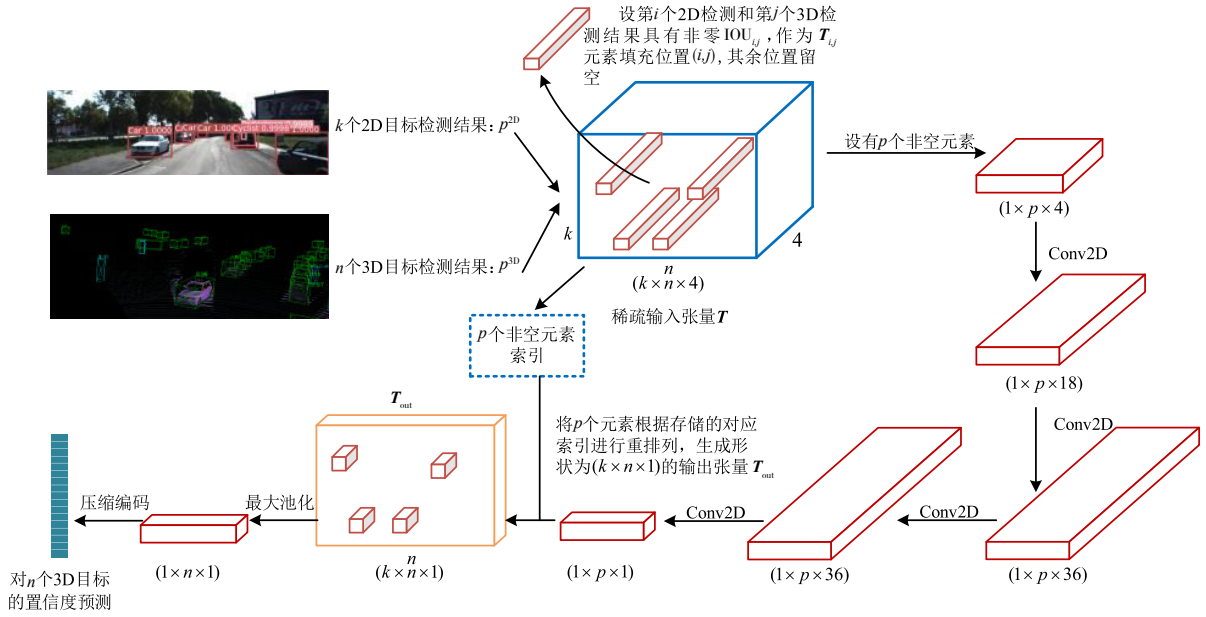


图13 CLOCs后融合网络结构

$$P^{2D}=\left\{p_1^{2D},p_2^{2D},\cdots,p_k^{2D}\right\} \tag{11}$$

$$p_i^{2D}=\left\{[x_{i1},y_{i1},x_{i2},y_{i2}],s_i^{2D}\right\} \tag{12}$$

其中，P^{2D}是一幅图像中所有k个检测候选结果的集合，对于第i个检测结果p_i^{2D}，$x_{i1},y_{i1},x_{i2},y_{i2}$为2D包围框左上角和右下角点的像素坐标，$s_i^{2D}$为对应框的置信度．

同样的，3D目标检测器输出为点云空间坐标中的3D检测边界框和置信度，由于相对坐标系与世界坐标系之间的转化方式不同，3D检测边界框的方法在不同的数据集中有不同的编码方式，在KITTI数据集中使用了7维向量，定义如下：

$$P^{3D}=\left\{p_1^{3D},p_2^{3D},\cdots,p_n^{3D}\right\} \tag{13}$$

$$p_i^{3D}=\left\{[h_i,w_i,l_i,x_i,y_i,z_i,\theta_i],s_i^{3D}\right\} \tag{14}$$

其中，P^{3D}是一幅图像中所有n个检测候选结果的集合，对于第i个检测结果p_i^{3D}，$[h_i,w_i,l_i,x_i,y_i,z_i,\theta_i]$为三维检测框的高度、宽度、长度、中心点坐标、中心旋转角度，s_i^{3D}为对应3D框的置信度．

对于k个2D候选结果与n个3D候选结果，会重新编码为一个$k\times n\times 4$的张量，每个元素$\boldsymbol{T}_{ij}$对应四个通道：

$$\boldsymbol{T}_{ij}=(\mathrm{IOU}_{ij},s_i^{2D},s_i^{3D},d_j) \tag{15}$$

其中，IOU_{ij}表示为第i个2D候选结果与投影到2D平面后的第j个3D候选结果之间的交并比，是几何一致性的集中表现．s_i^{2D}为2D检测置信度，s_i^{3D}为3D检测置信度，d_j表示第j个3D候选边界框到地面Oxy的归一化距离．这种表示方法会很容易的剔除IOU很低的张量元素，因为不满足几何一致性的融合条件．剔除空元素后的张量具有稀疏属性，因为对于每个投影的3D检测框，只有很少的2D检测框与它相交．如果有某些目标只被3D检测器提出检测结果，为了不浪费这些候选结果，会将$\boldsymbol{T}_{kj}$中的IOU_{kj}项和s_k^{2D}项设为“−1”，不设为“0”是为了和其他几何不一致的检测结果区分开来．

CLOCs融合网络依次使用了四个卷积层，且在前三个卷积层之后，都使用ReLU函数激活．由于稀疏张量$\boldsymbol{T}_{ij}$在特征融合网络中是为了概率驱动的信息融合而不是图形感知，所以卷积采用1×1的核大小．二维卷积分别为第1层：Conv2D(4, 18, (1, 1), 1)，第2层：Conv2D(18, 36, (1, 1), 1)，第3层：Conv2D(36, 36, (1, 1), 1)，第

4层:Conv2D(36, 1, (1, 1), 1),这产生了一个大小为$1\times p\times 1$的张量,其中p是输入张量$\boldsymbol{T}$中非空元素的数量. 参照输入张量的位置索引(i,j),来用p个输出元素进行填充,构建一个$k\times n\times 1$形状的张量$\boldsymbol{T}_{\text{out}}$. 最后通过对张量$\boldsymbol{T}_{\text{out}}$第个一维度进行最大池化(max-pooling)将其与$1\times n$的概率分数的期望目标相映射,选择最终的融合结果.

后融合CLOCs采用焦点损失(Focal Loss)[18]函数进行目标分类,这个损失函数在标准的交叉熵标准上添加了一个因子$(1-p_{\text{t}})^{\gamma}$,使模型更加集中于小数量的目标类别且增加分类错误的样本权重,以解决目标和背景的类别不平衡问题. 标准的二元分类的交叉熵(CE, Cross Entropy)损失可以表示为:

$$\text{CE}(p,y)\begin{cases}-\log(p), & y=1\\ -\log(1-p), & \text{其他}\end{cases} \tag{16}$$

其中,y指定真实类,正类为"1",负类为"0". $p\in[0,1]$是模型对标签$y=1$类的预测概率. 为了便于标记,本文定义p_{t}为:

$$p_{\text{t}}\begin{cases}p, & y=1\\ 1-p, & \text{其他}\end{cases} \tag{17}$$

则CE可以表示为:

$$\text{CE}(p_{\text{t}})=-\alpha_{\text{t}}\log(p_{\text{t}}) \tag{18}$$

引入因子$(1-p_{\text{t}})^{\gamma}$和平衡变量的焦点损失为:

$$\text{FL}(p_{\text{t}})=-\alpha_{\text{t}}(1-p_{\text{t}})^{\gamma}\log(p_{\text{t}}) \tag{19}$$

在实验中设置超参数$\alpha_{\text{t}}=0.25$,$\gamma=2$,进行训练以提高后融合方法的泛化能力.

3　实验校验

3.1　实验环境

实验的硬件平台采用了GTX1080Ti的11 GB显存GPU,16 GB内存. 所有的检测实验结果由OpenPCDet中的KITTI结果评估插件生成,数据预处理及训练参数中严格控制变量. KITTI数据集作为目前世界最大的面向自动驾驶的计算机视觉公开评测数据集之一,包含市区、乡村和高速公路等场景的真实图像数据,整个数据集由389对立体图像和光流图,39.2 km视觉测距序列以及超过20万的3D标注物体的图像组成,以10 Hz的频率采样及同步. 这目前是验证立体图像,光流,视觉测距,3D目标检测和3D跟踪等计算机视觉技术在车载环境下性能的最权威平台之一. 在3D目标检测类别中提供了两个140万像素的FL2-14S3C-C彩色相机,两个140万像素的FL2-14S3M-C灰度相机,一个Velodyne HDL-64线激光雷达以及配准的雷达到不同相机的转换矩阵,其中包括激光雷达到0号灰度相机的外参矩阵、0号相机的矫正旋转矩阵、其他相机的内参矩阵等.

在KITTI数据集对原始点云进行数据预处理和数据增强后采用标准训练测试流程,训练集3 712张,验证集3 769张. 本文的主干方法是PointPillars,设置柱体的尺寸为0.16 m×0.16 m×4 m,柱内最大点云数目为32个,最大柱体数目在训练集为16 000,验证集中为40 000. 采用初始学习率为3×10^{-3}的Adam优化器根据历史梯度的震荡情况和过滤震荡后的真实历史梯度对变量进行更新,避免后期学习率过大而导致结果不收敛. 由于全局自注意力机制会大量消耗GPU显存资源,因此批数(Batch Size)设为1,最大训练80轮.

3.2　检测指标结果及分析

本文的3DMMF与新近主流的3D目标检测算法在汽车类相比较结果如表1所示.

表1　各算法的3D检测能力指标横向对比　　单位:%

模型	全类别平均正确率(mAP)	车辆检测平均正确率		
		简单	普通	困难
MV3D	62.85	71.09	62.35	55.12
MV3D(LIDAR)	56.94	66.77	52.73	51.31
F-PointNet	71.26	81.20	70.39	62.19
AVOD	65.92	73.59	65.78	58.38
PI-RCNN	76.41	84.37	74.82	70.03
SECOND	74.33	83.13	73.66	66.20
VoxelNet	66.77	77.47	65.11	57.73
PointRCNN	76.25	85.29	75.08	68.38
PointPillars	73.59	80.36	73.64	66.79
本文3DMMF	**79.83**	**87.42**	**77.36**	**74.72**

由结果表明,本文提出的方法相较于单阶段融合的检测网络MV3D、F-PointNet、AVOD、PI-RCNN在全部三种难度指标中都有着绝对优势,其中在简单指标与困难指标中都有着最好的成绩. 相较于纯点云网络的性能指标也有一定优势,尤其是较于基线方法PointPillars本文的方法则在全部指标都有大幅提升,平均mAP提高了6.24%.

3.3　实时性对比

而在处理时间方面,本文将3D目标检测的处理时间分为两个部分:一个为前传等待时间,即3D检测前的点云数据处理时间,在PointPainting中是在非"流水线"的条件下DeeplabV3+算法完成2D图像语义分割与全局编码的时长,在FRP中为YOLOv3算法完成2D图像目标检测与局部编码的时长;另一个为3D检测时间,即串行的3D检测网络从接收编码后的点云到完成3D目标检测任务的时长,具体结果如表2所示.

本文提出的FRP是双阶段融合网络中检测效率最高的,而3DMMF则在检测的准确率指标与检测效率之间有一个良好的平衡性. FRP相较于PointPainting的改善是非常显著的,平均前传等待时间减少了95.35%,这是由图像分割任务转为图像检测,前传任务减负带来的

表2 各算法处理时间横向对比 单位:s

模型	前传等待时间	3D检测时间	总处理时间
PointPillars	N/A	0.020 2	0.020 2
SECOND	N/A	0.062 6	0.062 6
F-PointNets	0.020 4	0.130 4	0.150 8
MV3D	N/A	0.447 3	0.447 3
AVOD	N/A	0.134 7	0.134 7
PointPainting	0.438 5	0.056 3	0.494 8
本文FRP	0.020 4	0.030 8	0.051 2
本文3DMMF	0.020 4	0.057 7	0.078 1

改善. 而3DMMF也能摆脱“流水线”的方法,这是由于只使用了一次2D目标检测算法,使其同时直接作用于前融合和后融合两个阶段的跨模态信息融合,在提升检测精度的同时,没有延长太多的前传时间和后处理时间.

如图14所示,是本文3DMMF方法逐步优化的过程,在PointPainting方法中,右侧小车的投影正后方很明显出现了小车的目标虚警. 这是由于图像的语义掩膜特征非常突出,2D分割后的噪声会传递到点云模态中,很容易在目标的后方位置出现虚警. 采用FRP方法后,小车后方的虚警消失,这是由于保留了锥视区内的RGB颜色信息,目标的边界信息可以通过颜色的落差更加准确的表现. 但是小车的前方出现一个虚警,这是由于通道扩充PointPillars伪图像特征图单个位置的代表的信息量倍增造成的虚警,这种虚警在引入全局自注意力机制的3DMMF中得到改善.

而如图15和图16所示,分别对比PointPillars网络在原始PointPainting方法编码后与3DMMF方法的生成的检测框在3D点云场景和对应图像下的检测结果. 可以直观的发现在PointPainting方法中较高的虚警率和置信度在3DMMF中都有所改善.

3.4 消融实验

首先如图17的局部消融实验所示,在FRP前融合阶段,随着2D图像检测精度的提升,3D目标检测的指标也随之上升,并在采用“真实框”形成锥视区的前融

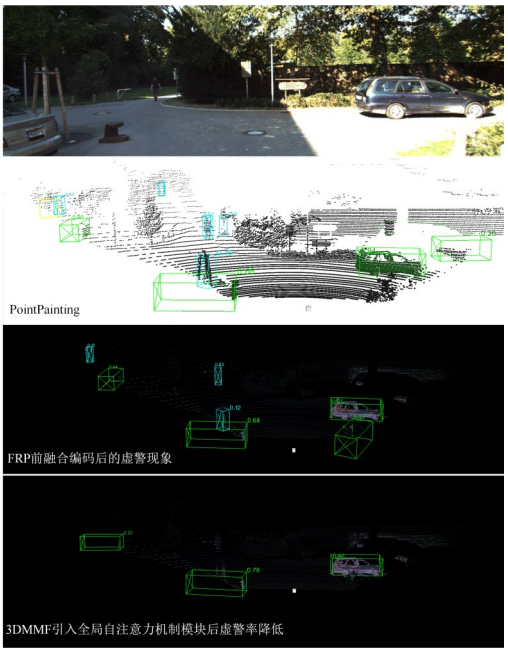


图14 引入全局自注意力机制虚警率降低

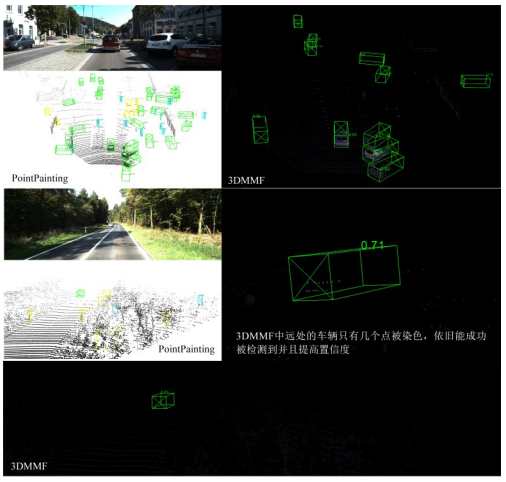


图15 3DMMF虚警率降低同时对真阳值的预测置信度提高

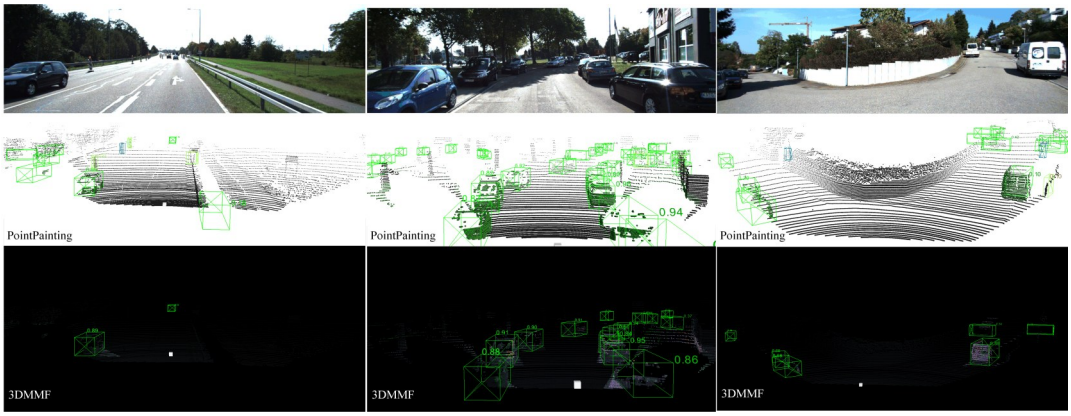


图16 3DMMF检测结果可视化

合编码时实现最优指标. 而使用PointPainting后的对车辆的检测指标反而降低了(PointPainting在对行人等远小目标的检测更有优势),这也符合第2.1节中的分析与预期结果.

同时我们注意到,随着2D检测精度的持续下降,对FRP的3D目标检测指标影响是有限的. 在相同的图像检测条件下,传统的F-PointNet、F-PointPillars等算法的3D检测指标一定会低于图像中的检出率:当2D目标检测失效时,3D目标检测算法也无法进行. 而FRP没有对三维检测体系结构添加任何限制,只是为点云附加了额外的通道信息,因此即便2D目标检出失败,3D目标检测网络还是有一定概率成功检测到目标. 虽然这造成少许在2D图像中没有出现,仅在3D点云空间中的虚警,但是这同时保证了在恶劣光照条件下融合3D目标检测算法的相对鲁棒性.

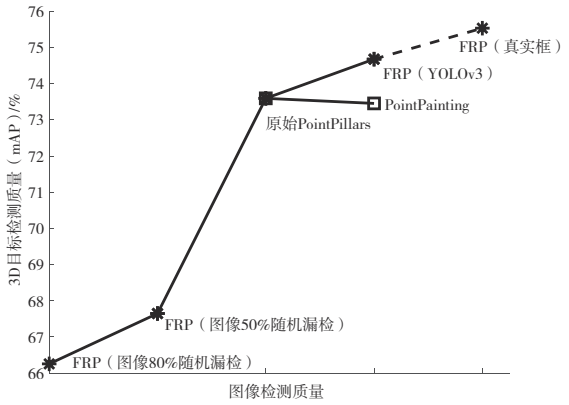


图17　2D图像检测结果对3D检测能力的影响

而本文提出的多层多模态融合3D目标检测方法,旨在多个阶段融合2D与3D信息,在实验中我们发现除了前融合阶段的FRP局部编码以外,3D目标检测中的自注意力机制提取上下文信息,后融合阶段的CLOCs的概率驱动的决策融合对于3D目标指标改善都是有意义的. 与纯点云的基线方法PointPillars或者是单阶段融合方法的PointPainting相比,多阶段融合跨模态信息能够有效改善视角错位与特征不匹配的问题. 在各模块消融实验结果中予以证明,如表3所示.

表3　消融实验中3D检测能力指标横向对比　　单位:%

模型	FRP	FSA	CLOCs	全类别平均正确率(mAP)	车辆检测平均正确率		
					简单	普通	困难
PointPillars				73.59	80.36	73.64	66.79
PointPainting				73.46	79.42	73.67	67.28
本文3DMMF	√			74.68	80.33	74.12	69.59
		√		77.49	85.09	75.18	72.21
			√	75.62	82.26	76.12	68.49
	√	√		79.49	87.14	77.06	74.28
	√	√	√	79.83	87.42	77.36	74.72

4　结论

本文面向自动驾驶场景,为了改善单阶段图像与激光雷达信息融合难以解决的视角错位及异构特征不匹配的问题. 提出了一种多层多模态融合的3D目标检测算法,通过在多个阶段融合浅层、深层多模特征信息,提升3D目标检测能力. 首先利用2D目标检测算法进行图像检测;再将向图像投影的雷达点云进行锥视区彩色涂抹编码,再将编码后的点云输入增加输入通道的PointPillars网络,为了加深对编码后彩色点云的理解,加入自注意力机制提取上下文信息. 最后通过概率驱动匹配2D、3D的检测结果的几何方位特征和类别语义特征,利用CLOCs网络提升3D目标检测精度. 在KITTI公开路面数据集中进行实验,结果证明了此方法的有效性:相较于PointPillars基线方法,3D目标检测精度明显提升,平均mAP提高6.24%. 而检测效率相较于PointPainting方法提高了89.65%,摆脱了"流水线"这种不稳定的方法,使检测精度与效率达到比较合适的平衡. 与其他单阶段融合算法或纯点云算法相比,本文提出的融合方法具有明显优势,这种多阶段多模态融合方法为传感器融合算法的开发提供了一种新思路.

参考文献

[1] 沈峘, 李舜酩, 柏方超, 等. 路面车辆实时检测与跟踪的视觉方法[J]. 光学学报, 2010, 30(4): 1076-1083.
SHEN H, LI S M, BAI F C, et al. Visual method for real-time detection and tracking of road vehicles[J]. Acta Optica Sinica, 2010, 30(4): 1076-1083. (in Chinese)

[2] 于洁潇, 张美琪, 苏育挺. 基于双目视觉的三维车辆检测算法[J]. 激光与光电子学进展, 2021, 58(2): 0215004.
YU J X, ZHANG M Q, SU Y T. 3D vehicle detection algorithm based on binocular vision[J]. Laser & Optoelectronics Progress, 2021, 58(2): 0215004. (in Chinese)

[3] 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: IEEE, 2012: 3354-3361.

[4] LANG A H, VORA S, CAESAR H, et al. PointPillars: Fast encoders for object detection from point clouds[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 12689-12697.

[5] SHI S S, WANG X G, LI H S. PointRCNN: 3D object proposal generation and detection from point cloud[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 770-779.

[6] ZHOU Y, TUZEL O. VoxelNet: End-to-end learning for point cloud based 3D object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4490-4499.

[7] QI C R, LIU W, WU C X, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 918-927.

[8] CHEN X Z, MA H M, WAN J, et al. Multi-view 3D object detection network for autonomous driving[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 6526-6534.

[9] KU J, MOZIFIAN M, LEE J, et al. Joint 3D proposal generation and object detection from view aggregation[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2018: 1-8.

[10] VORA S, LANG A H, HELOU B, et al. PointPainting: Sequential fusion for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 4603-4611.

[11] XIE L, XIANG C, YU Z X, et al. PI-RCNN: An efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12460-12467.

[12] BHATTACHARYYA P, HUANG C J, CZARNECKI K. SA-Det3D: Self-attention based context-aware 3D object detection[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE, 2021: 3022-3031.

[13] PANG S, MORRIS D, RADHA H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection[C]// 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2020: 10386-10393.

[14] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: A multimodal dataset for autonomous driving[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 11618-11628

[15] REDMON J, FARHADI A. YOLOv3: An incremental improvement[EB/OL]. (2018-04-08)[2022-04-20]. https://arxiv.org/abs/1804.02767.pdf.

[16] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 779-788.

[17] REDMON J, FARHADI A. YOLO9000: Better, faster, stronge[EB/OL]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 6517-6525.

[18] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

作者简介

周治国 男, 1977 年 9 月出生于湖北省武汉市. 现为北京理工大学集成电路与电子学院医学图像与信号处理研究所副教授、硕士生导师. 主要研究方向包括智能无人系统、感知与导航和机器学习. 在国内外发表学术论文20余篇. 中国电子学会会员编号:E190015683M.
E-mail: zhiguozhou@bit.edu.cn

马文浩 男, 1996 年 6 月出生于新疆维吾尔自治区伊宁市. 现为北京理工大学集成电路与电子学院医学图像与信号处理研究所硕士, 从事无人器融合感知方面的研究工作.
E-mail: 3220190552@bit.edu.cn