(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118429958 A
(43)申请公布日 2024.08.02

(21)申请号 202410534127.5

(22)申请日 2024.04.30

(71)申请人 天津理工大学
地址 300384 天津市西青区宾水西道391号

(72)发明人 周冕 田雪媛 吴少清

(74)专利代理机构 天津耀达律师事务所 12223
专利代理师 邵洪军

(51)Int.Cl.
G06V 20/64(2022.01)
G06V 10/80(2022.01)
G06V 10/82(2022.01)
G06N 3/045(2023.01)
G06N 3/0464(2023.01)

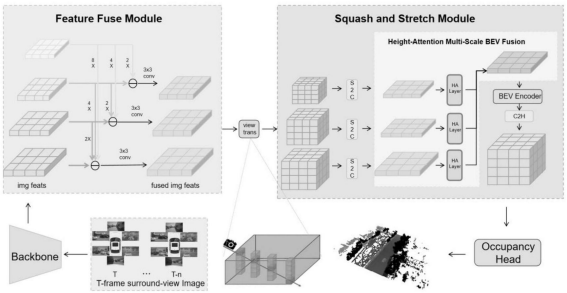
权利要求书2页 说明书4页 附图1页

(54)发明名称

一种基于三维占用预测的纯视觉三维感知方法

(57)摘要

本发明涉及一种基于三维占用预测的纯视觉三维感知方法,属于计算机视觉技术领域;现有模型很难脱离实验环境进行部署,并且忽视了数据在高度维度的分布的不均衡性是目前需要解决的问题;本方法通过图像特征融合模块融合多尺度图片特征,使用2D-3D空间转换模块把图像特征投影为体素特征,再将体素特征压缩到bev层面进行特征提取和融合,并使用高度注意力模块对BEV特征的不同高度特征通道进行权重预测和加权计算,最后定义任务头进行三维占有预测;本发明只需要输入图像,无需昂贵的雷达设备,大大减少了感知成本;使用轻量的二维卷积,无需使用注意力机制,深度估计,三维卷积等需要大量资源的方法;实验证明,本发明可以达到很好的性能。



CN 118429958 A

1.一种基于三维占用预测的纯视觉三维感知方法,其特征在于,包括:

使用环视相机采集图像;

从环视图像中获取图像特征;

将图像特征投影到预定义好的三维体素空间得到体素特征;

将体素特征压缩到bev层面进行特征提取和融合;

定义三维占有预测的任务头,构建整体的感知网络模型;

使用感知网络模型对新的输入数据进行三维占有预测,获得三维世界中每个体素单元的占用状态和类别。

2.根据权利要求1所述的一种基于三维占用预测的纯视觉三维感知方法,其特征在于,还包括:

采用图像增强策略对图像进行增强用于训练感知网络模型;

采用BEV空间增强策略对BEV特征进行增强用于训练感知网络模型;

采用时间融合策略引入多帧图像用于训练感知网络模型。

3.根据权利要求1所述的一种基于三维占用预测的纯视觉三维感知方法,其特征在于,从环视图像中获取图像特征的具体步骤为:

将六张环视图像输入到backbone模块,再经过FPN模块,再经过图像特征融合模块获得多尺度的图像特征。

4.根据权利要求1所述的一种基于三维占用预测的纯视觉三维感知方法,其特征在于,将图像特征投影到预定义好的三维体素空间得到体素特征的具体步骤为:

利用相机的内外参,提前计算好2D图像特征位置和3D体素位置的对应关系,再按照这个对应关系将多尺度的2D图像特征投影到对应的多尺度的3D体素的位置,得到预定义好的三维体素空间中的多尺度的3D体素的特征。

5.根据权利要求1所述的一种基于三维占用预测的纯视觉三维感知方法,其特征在于,将体素特征压缩到bev层面进行特征提取和融合的具体步骤为:

将多尺度体素特征输入到一个压缩-拉伸模块,把多尺度体素特征通过S2C操作压缩到BEV层面得到多尺度BEV特征,再经过一个高度注意力层对BEV特征的不同高度特征通道进行权重预测和加权计算,再使用特征上采样和二维卷积对加权后的多尺度BEV特征的提取和融合,再把融合后的bev特征经过一个bev编码器进行特征的提取,再通过一个C2H操作拉伸回体素层面的体素特征。

6.根据权利要求1所述的一种基于三维占用预测的纯视觉三维感知方法,其特征在于,定义三维占有预测的任务头,构建整体的感知网络模型的具体步骤为:

三维占有预测的任务头使用一个分类网络,输入每个体素单元的体素特征,输出每个体素单元预测的占有状态和语义分类。

7.根据权利要求2所述的一种基于三维占用预测的纯视觉三维感知方法,其特征在于,采用图像增强策略对图像进行增强用于训练感知网络模型具体包括:

对输入图像进行随机裁剪,随机翻转,随机旋转。

8.根据权利要求2所述的一种基于三维占用预测的纯视觉三维感知方法,其特征在于,采用BEV空间增强策略对BEV特征进行增强用于训练感知网络模型具体包括:

对输入BEV特征进行随机翻转。

9.根据权利要求2所述的一种基于三维占用预测的纯视觉三维感知方法,其特征在于,采用时间融合策略引入多帧图像用于训练感知网络模型的具体步骤为:

先将多帧图片对齐到当前帧,然后将多帧图片和当前帧图片同时输入网络进行特征提取,再同时投影到预定义好的三维体素空间得到体素特征。

10.根据权利要求5所述的一种基于三维占用预测的纯视觉三维感知方法,其特征在于,经过一个高度注意力层对BEV特征的不同高度特征通道进行权重预测和加权计算的具体步骤为:

先将BEV特征的长宽维度通过全局平均池化进行压缩,再用全连接层预测高度维度中各通道的重要性,再根据预测结果获得不同的通道权重对之前的BEV特征重新加权。

一种基于三维占用预测的纯视觉三维感知方法

技术领域

[0001] 本发明属于计算机视觉技术领域,涉及一种基于三维占用预测的纯视觉三维感知方法。

背景技术

[0002] 自动驾驶算法技术框架的核心分为环境感知、决策规划、控制执行三部分,其中准确而全面地感知三维环境对自动驾驶系统至关重要,是下游决策和控制任务的基础。最近,继BEV(Bird Eye View,鸟瞰图)感知之后,一种新的纯视觉的感知任务三维占有预测受到了广泛的关注,这项新任务的目标是从输入的多视角图像序列中估计场景中每个体素的占用状态和语义标签。

[0003] 但是,目前学术界提出的纯视觉三维占有预测算法存在以下弊端:

[0004] 1.现有模型大多使用需要专用芯片支持的注意力机制,或使用需要进行对加速不友好的体素池化操作的深度预测模块,或使用需要大量计算资源和存储资源的三维卷积操作,都使模型很难脱离实验环境进行实际部署。

[0005] 2.现在使用的nuscenes数据集存在严重的数据空间分布不平衡问题,尤其是在高度维度上,但是目前几乎没有模型关注这方面的问题,导致某些类别的预测效果不佳。

发明内容

[0006] 为了解决目前利用纯视觉三维占有预测进行环境感知时存在的部署难度大,计算资源和存储资源需求量大,并且忽视数据在高度维度分布不均衡,导致性能不佳的问题,本发明公开了一种基于三维占用预测的纯视觉三维感知方法,主要的技术亮点是使用轻量的二维卷积,无需特定芯片或巨大的计算和存储资源的支持,使模型易于部署在除实验环境以外的其他平台;同时使用高度注意力机制对数据空间分布不平衡问题进行处理,使检测效果更为精准。

[0007] 实现发明目的的技术方案如下:一种基于三维占用预测的纯视觉三维感知方法,包括:

[0008] 使用环视相机采集图像;

[0009] 从环视图像中获取图像特征;

[0010] 将图像特征投影到预定义好的三维体素空间得到体素特征;

[0011] 将体素特征压缩到bev层面进行特征提取和融合;

[0012] 定义三维占有预测的任务头,构建整体的感知网络模型;

[0013] 使用感知网络模型对新的输入数据进行三维占有预测,获得三维世界中每个体素单元的占用状态和类别。

[0014] 进一步的,还包括:

[0015] 采用图像增强策略对图像进行增强用于训练感知网络模型;

[0016] 采用BEV空间增强策略对BEV特征进行增强用于训练感知网络模型;

[0017] 采用时间融合策略引入多帧图像用于训练感知网络模型。

[0018] 优选地,从环视图像中获取图像特征的具体步骤为:

[0019] 将六张环视图像输入到backbone模块,再经过FPN模块,再经过图像特征编码模块获得多尺度的图像特征。

[0020] 优选地,将图像特征投影到预定义好的三维体素空间得到体素特征的具体步骤为:

[0021] 利用相机的内外参,提前计算好2D图像特征位置和3D体素位置的对应关系,再按照这个对应关系将多尺度的2D图像特征投影到对应的多尺度的3D体素的位置,得到预定义好的三维体素空间中的多尺度的3D体素的特征。

[0022] 优选地,将体素特征压缩到bev层面进行特征提取和融合的具体步骤为:

[0023] 将体素特征压缩到bev层面进行特征提取和融合的具体步骤为:

[0024] 将多尺度体素特征输入到一个压缩-拉伸模块,把多尺度体素特征通过S2C操作压缩到BEV层面得到多尺度BEV特征,再经过一个高度注意力层对BEV特征的不同高度特征通道进行权重预测和加权计算,再使用特征上采样和二维卷积对加权后的多尺度BEV特征的提取和融合,再把融合后的bev特征经过一个bev编码器进行特征的提取,再通过一个C2H操作拉伸回体素层面的体素特征。

[0025] 其中,经过一个高度注意力层对BEV特征的不同高度特征通道进行权重预测和加权计算的具体步骤为:

[0026] 先将BEV特征的长宽维度通过全局平均池化进行压缩,再用全连接层预测高度维度中各通道的重要性,再根据预测结果获得不同的通道权重对之前的BEV特征重新加权。

[0027] 优选地,定义三维占有预测的任务头,构建整体的感知网络模型的具体步骤为:

[0028] 三维占有预测的任务头使用一个分类网络,输入每个体素单元的体素特征,输出每个体素单元预测的占有状态和语义分类。

[0029] 优选地,采用图像增强策略对图像进行增强用于训练感知网络模型具体包括:

[0030] 对输入图像进行随机裁剪,随机翻转,随机旋转。

[0031] 优选地,采用BEV空间增强策略对BEV特征进行增强用于训练感知网络模型具体包括:

[0032] 对输入BEV特征进行随机翻转。

[0033] 优选地,采用时间融合策略引入多帧图像用于训练感知网络模型的具体步骤为:

[0034] 先将多帧图片对齐到当前帧,然后将多帧图片和当前帧图片同时输入网络进行特征提取,再同时投影到预定义好的三维体素空间得到体素特征。

[0035] 技术效果

[0036] 与现有技术相比,本发明至少具有如下有益效果:

[0037] 易于部署,不需要特定的芯片的支持;

[0038] 仅需要较少的计算资源和存储资源;

[0039] 对数据在高度维度上的空间分布不平衡问题进行处理,使检测效果更为精准;

[0040] 通过性能对比,本发明可以在以上优势的基础上做到有竞争力的性能。

附图说明

[0041] 为了更清楚地说明本申请实施例的技术方案,下面将对实施例中所需要使用的附图作简单地介绍:

[0042] 图1为本发明的三维感知方法的流程示意图;

[0043] 图2为本发明的三维感知方法与现有技术的性能比较;

具体实施方式

[0044] 下面通过实施例并结合说明书附图对本发明做进一步说明。需要说明是,本发明还可以采用其他不同于在此描述的其他方式来实施,因此,本发明的保护范围并不受下面公开的具体实施例的限制。

[0045] 本发明的一个具体实施例,如图1,公开了一种基于三维占用预测的纯视觉三维感知方法,具体方法如下:

[0046] 步骤1:使用环视相机采集图像。

[0047] 步骤2:从环视图像中获取图像特征。

[0048] 步骤2.1:将六张环视图像经过图像增强,这里的图像增强包括随机裁剪,随机翻转,随机旋转,再经过BEV空间增强,包括随机翻转,再输入到backbone模块(resnet50),获得原始的多尺度图像特征(4层)。

[0049] 步骤2.2:经过FPN模块,获得融合后的多尺度图像特征(4层)。

[0050] 步骤3:经过图像特征融合模块对多尺度特征进行融合,获得编码后的三层图像特征。

[0051] 步骤3.1:对最小的s1分辨率进行特征的下采样2倍到与下一层的s2分辨率层的特征相同的尺寸,再将其与s2分辨率层的特征进行拼接,再经过一个33卷积,得到融合后的最大分辨率的特征的第一层特征a1。

[0052] 步骤3.2:把s1分辨率的特征进行下采样4倍到与s3分辨率层的特征相同的尺寸,把s2分辨率的特征进行下采样2倍到与s3分辨率层的特征相同的尺寸,再将它们与s3分辨率层的特征进行拼接,再经过一个33卷积,得到融合后的最大分辨率的特征的第二层特征a2。

[0053] 步骤3.3:把s1分辨率的特征进行下采样8倍到与s4分辨率层的特征相同的尺寸,把s2分辨率的特征进行下采样4倍到与s4分辨率层的特征相同的尺寸,把s3分辨率的特征进行下采样2倍到与s4分辨率层的特征相同的尺寸,再将它们与s4分辨率层的特征进行拼接,再经过一个33卷积,得到融合后的最大分辨率的特征的第三层特征a3;

[0054] 步骤4:将多尺度图像特征投影到预定义好的三维体素空间得到多尺度体素特征。

[0055] 步骤4.1:在三维体素空间中预定义好三个尺度的体素单元的位置,三个尺度分别是100x100x8,150x150x8,200x200x8。

[0056] 步骤4.2:利用相机的内外参,提前计算好2D图像特征位置和3D体素位置的对应关系,再按照这个对应关系将多尺度的2D图像特征投影到对应的多尺度的3D体素的位置,得到预定义好的三维体素空间中的多尺度的3D体素的特征。

[0057] 步骤4.3:利用相机的内外参数,将对齐后的历史帧的2D环视图片也进行步骤4.2的操作,把包括当前帧和历史帧的所有关联到的图像特征堆叠起来,得到的多尺度的3D体

素特征。

[0058] 步骤5:将体素特征压缩到bev层面进行特征提取和融合。

[0059] 步骤5.1:把多尺度体素特征通过S2C操作压缩到BEV层面,S2C操作具体是把H×W×Z×C尺寸的4维多尺度体素转换成H×W×(ZC)的3维多尺度BEV特征表示。

[0060] 步骤5.2:将多尺度BEV特征的长宽维度通过全局平均池化进行压缩,从H×W×(ZC)压缩到1×1×(ZC),再用全连接层预测ZC中各通道的重要性,再根据预测结果获得不同的通道权重,对之前的多尺度BEV特征重新加权。

[0061] 步骤5.3:使用特征上采样和1×1卷积进行多尺度特征的提取和融合。

[0062] 步骤5.4:把融合后的bev特征经过一个bev编码器进行特征的提取。

[0063] 步骤5.5:通过一个C2H操作拉伸回体素层面的体素特征,C2H操作具体是把H×W×(ZC)的bev特征表示转换回H×W×Z×C的体素表示。

[0064] 步骤6:定义三维占有预测的任务头,构建整体的感知网络模型。

[0065] 步骤6.1:三维占有预测的任务头使用一个分类网络,输入每个体素单元的体素特征,输出每个体素单元预测的占有状态和语义分类。

[0066] 步骤6.2:对模型进行训练使用的损失函数是:Loss=λ1lovasz loss+λ2Cross Entropy Loss,其中λ1=λ2=1。

[0067] 步骤7:使用感知网络模型对新的输入数据进行三维占有预测,获得三维世界中每个体素单元的占用状态和类别。

[0068] 本发明实际效果的验证,具体内容如下。

[0069] 本发明在Occ3D-nuScenes数据集上对ConvOcc进行了评估,该数据集包含850个自动驾驶场景,每个场景以2Hz的频率捕获20秒的注释感知数据。该数据集分为700个训练场景和150个验证场景。每个样本沿X和Y轴的范围为-40米至40米,沿Z轴的范围为-1米至5.4米,体素大小为0.4米x0.4米x 0.4米。我们只使用环绕摄像机数据作为输入,其中有六个视图:前左、前、前右、后左、后、后右。输出有16个普通类别和一个附加的通用对象类别。在评估指标方面,我们报告了所有类别中的平均交并比(mIoU)。

[0070] 本发明使用ResNet-50作为骨干,输入图像大小为256x704,在3090GPU上进行了24个epoch的训练。

[0071] 本发明在Occ3D-nuScenes数据集中与各种方法对比,结果如图2,可以看出,本发明在仅使用小特征提取器ResNet-50和小图片输入尺寸256x704的情况下,性能超过了很多使用大特征提取器ResNet-101和大图片输入尺寸900x1600的模型。本发明在计算资源和存储资源的低需求下,实现了三维占有预测任务中优秀的性能表现。

[0072] 以上所述,仅为本发明较佳的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,可轻易想到的变化或替换,都应涵盖在本发明的保护范围之内。

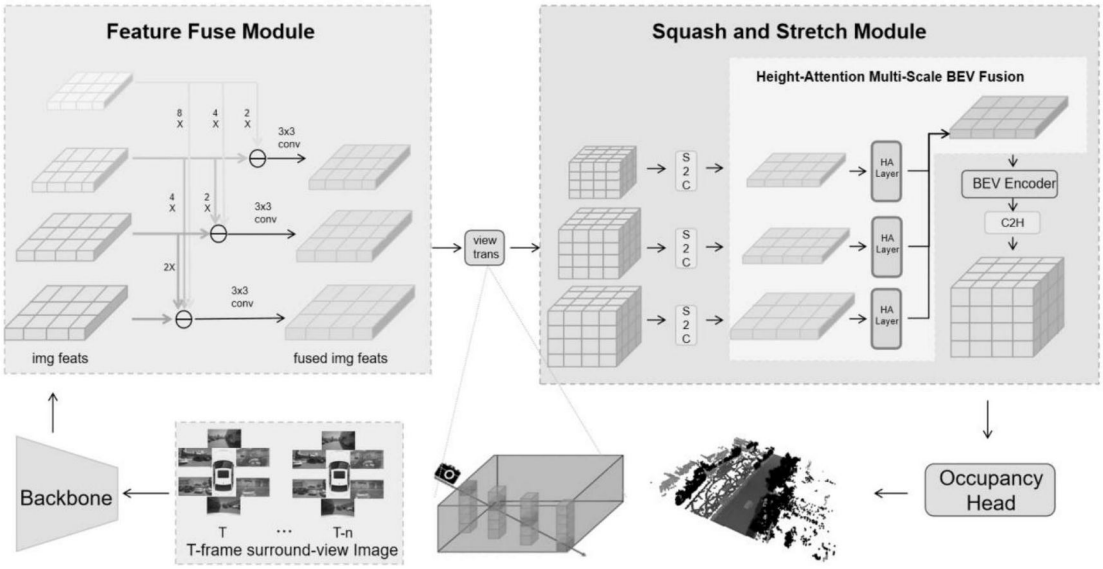


图1

Method	Input Modality	Image Backbone	Image Size	Epoch	Visible Mask	mIoU (%)
MonoScene [3]	Camera	ResNet-101	928 × 600	24	✘	6.06
TPVFormer [4]	Camera	ResNet-101	928 × 1600	24	✘	27.83
OccFormer [5]	Camera	ResNet-101	928 × 1600	24	✘	21.93
BEVFormer [6]	Camera	ResNet-101	928 × 1600	24	✘	26.88
CTF-Occ [1]	Camera	ResNet-101	928 × 1600	24	✘	28.53
BEVFormer [6]	Camera	ResNet-101	928 × 1600	24	✔	39.24
VoxFormer [7]	Camera	ResNet-101	900 × 1600	24	✔	40.7
Ours	Camera	ResNet-50	256 × 704	24	✔	36.11

图2