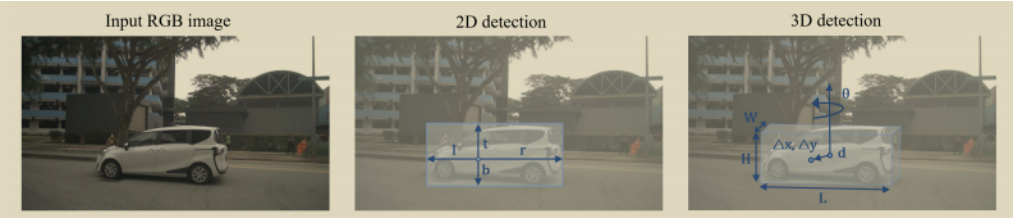
Fully Convoluional One-Stage Monocular 3D Objecion Detection

1. 2D detection & 3D detection□□□



2D检测：在2D检测中，我们使用RGB图像作为输入，目标是检测图像中的物体并生成一个边界框，用于表示物体的位置和形状。2D检测中的一种常见方法是"anchor-free"检测器，它不需要预定义的锚框（anchor boxes）。这种检测器需要预测从**前景点到边界框四条边的距离**，以确定物体在图像中的位置。 **3D检测**：在单目3D物体检测中，我们不仅要检测物体的位置和形状，还要推断物体的三维属性，如其在世界坐标系中的中心位置、大小以及朝向（XYZ轴上的位置，绕XYZ轴的旋转角度，目标尺寸大小）。这是一个更为复杂的任务，因为我们需要理解物体在三维空间中的属性。

2. 概要：3D单目相机物体检测因其价格较低的优势广泛应用于自动驾驶领域，但因其**缺乏深度信息**相较于2D物体检测难度较大。本文提出了一个单阶段出了一个用于目标检测的通用框架**FCOS3D(（Fully Convolutional One-Stage Detector)**，它能够将三维目标信息与图像领域的2D信息结合起来，通过考虑目标的2D尺度和中心位置，使目标分布到不同的特征级别，以此来训练模型。这个框架重新定义了中心性，将中心性定义为基于三维中心的2D高斯分布，以更好地适应三维目标的表示。最重要的是，它通过这些方法，实现了简单而高效的目标检测，无需预先了解2D检测或2D-3D对应的信息。

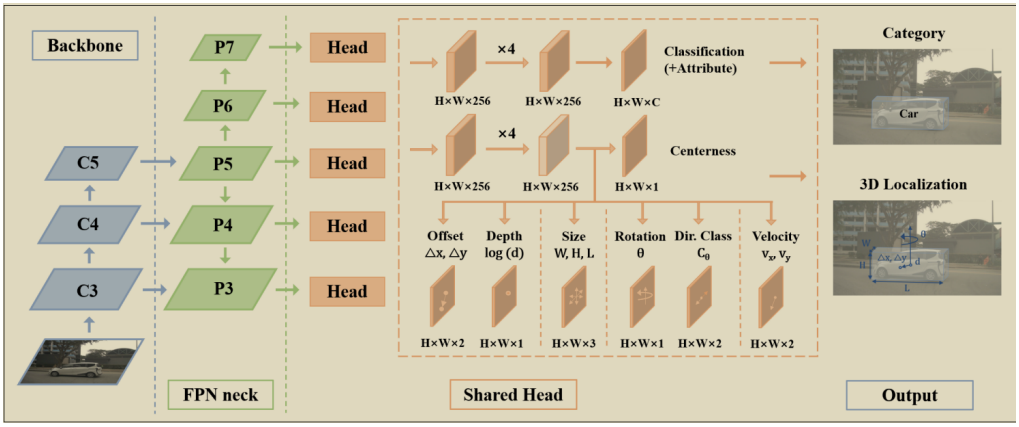
3. 2D Objection 模型：Faster R-CNN，RetinaNet，FCOS。 Anchor-based：基于锚点的方法，通常会预定义不同尺度的宽高比的锚点，更容易进行回归，更容易地确定物体的目标形状和位置，但需要调整许多超参数来获得最佳性能。 Anchor-free：无锚点的方法，不需要预先设点锚点，不需要调整大量的超参数，更加简介，通常通过其它方式来确定目标的位置和形状，直接预测目标的边界框。本文基于FCOS方法，它是一种无锚点的探测器，能够处理重叠的目标和尺度变化等问题。

4. 通过LiDARs point clouds雷达点云的方式可以获得精确的三维信息，但是其价格较为昂贵对于自动驾驶汽车来说。相反，基于单目摄像机的3D

Objecion Detection较为简单便宜，更受欢迎。

5. 过去的3D Objection Detection （1）在2D Objecion Detection的基础上，添加额外的阶段来预测物体的三维属性。首先预测2D边界框（2D boxes）的基础上，进一步回归出物体的三维属性。这些方法通过在2D物体中心和感兴趣的区域上进行3D属性的预测来完成这一任务。 （2）同时获取2D和3D的边界框，并且每个2D的锚点都对应有3D的先验信息，帮助模型了解2D位置与3D属性之间的关联，这种方法可以同时获取2D和3D的信息。 （3）使用冗余的3D信息，例如额外的keypoints，利用这些关键点更精确的表示物体的3D属性，以达到更优化的结果。基本思想都是将3D目标转化到2D范围，然后利用2D与3D之间的联系来解决问题。2D Objection Detection 和3D Objection Detection二者有着紧密的联系，相互发展促进。 （4）基于子网络的方法实现三维物体探测，这些子网络如深度估计网络、2D物体检测网络，这些子网络有时需要额外的数据和预训练，会使整个系统变得负责，也会过度依赖于子网络性能。 （5）将输入的RGB图像转化为体素或点云，这种方法依旧依赖于密集深度信息，且处理大规模点云数据复杂困难。 （6）学习2D Objection Detection采用端到端的设计，简化模型，提升性能，有基于锚点的方法，需要一致的2D和3D锚点定义并进行多个阶段训练和优化，还有不基于锚点的方法，不需要对数据进行统计分析，更加方便，本文也采用不基于锚点的方法。实际上，将2D检测器应用于单目三维物体检测依旧面临困难，本文将讨论二者之间的深入联系，为将典型的2D检测器框架应用于与之密切相关的任务时提供一个参考。

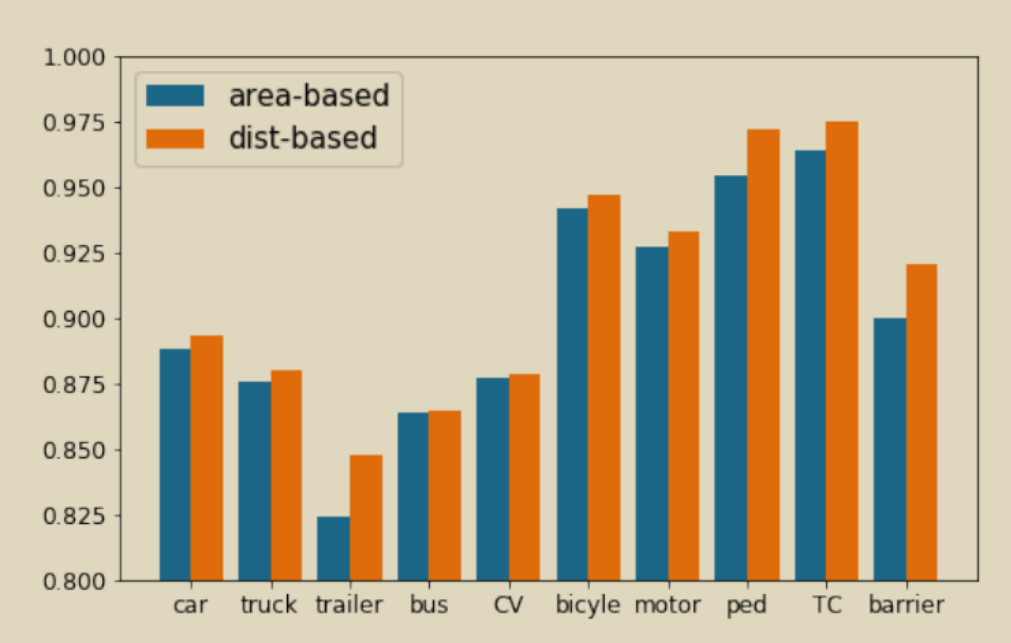
6. 方法概览□□□



借助已有的2D特征提取器，沿用了2D检测器的典型骨干和颈部设计。利用backbone提取出原始数据的不同层级特征。 然后输入到FPN neck网络中，将来自于backbone网络的不同层级特征融合在一起，以便于在多个尺度上检测物体。 检测头部分，采用了中心为基础的方式重新构建3D目标，以实现多任务学习，包括分类是什么样的物体，中心点确定，

各种坐标角度方位信息。 最终输出物体种类和三维物体位置信息。 为了处理重叠的真实目标和尺度差异问题，还相应地调整了多级别目标分配和中心采样策略。这一流程的目标是提高单目三维物体检测的性能和鲁棒性。

7. 具体网络和技术 backbone：采用了经过预训练的 ResNet101 模型，并在特征提取过程中引入了可变卷积。固定了第一个卷积块的参数，以避免额外的内存开销，减少模型的内存需求，提高效率。 neck：采用FPN特征金字塔网络，用于在不同尺度上检测对象和提取特征。文章将特征图从第3级到第7级分别表示为 P3 到 P7，采用了原始的FCOS方法来获取P3到P5，并通过两个卷积块对P5进行下采样以获取P6和P7。这五个特征图负责在接下来的任务中进行不同尺度的目标预测。 detection head：首先将目标分配到不同的特征级别和不同的点上。其次在网络架构上依旧采用每个共享头部由4个共享卷积块和不同目标的小头部组成。对于不同度量的回归目标，构建额外的分离头部更加有效，因此为每个回归目标设置了一个小头部。 基于距离的目标分配模糊准则：过去，当一个点在同一特征级别内位于多个真实目标框之内时，会根据2D边界框的面积进行选择。面积较小的框被选为该点的目标框。我们将这种方案称为基于面积的准则，这种方案有一个明显的缺点：大型目标在这种处理方式下会受到较少的关注。提出了一种基于距离的准则，即选择距离更近的中心作为回归目标框。这种方案与定义回归目标的中心为基础的机制一致。此外，这也是合理的，因为靠近目标中心的点可以获得更全面和平衡的局部区域特征，从而更容易产生更高质量的预测结果。□□□



具体实现细节

（1）网络架构：

使用ResNet101作为特征提取的骨干网络，用于从输入图像中提取特征。然后，使用特征金字塔网络（Feature Pyramid Networks，FPN）生成多层级的预测。有助于在不同尺度上检测目标。

多个级别的特征图共享一个检测头部，共同用于生成目标检测的结果。为了区分不同尺度的目标，框架使用三个尺度因子，分别应用于最终回归结果，包括偏移、深度和尺寸等。

卷积模块包括基本的局安居、批量归一化和激活层

网络的权重参数采用正态分布进行初始化

整体框架构建在MMDetection3D之上

（2）训练参数不需要提前预训练，从随机初始化的网络开始进行训练，端到端使用SGD随机梯度下降法更新模型参数

（3）数据增强使用图像翻转的方法

测试结果（采用nuScenes数据集）□□□

Table 1: Results on the nuScenes dataset.

Methods	Dataset	Modality	mAP	mATE	mASE	mAOE	mAVE	mAAE	NDS
CenterFusion [22]	test	Camera & Radar	0.326	0.631	0.261	0.516	0.614	0.115	0.449
PointPillars [14]	test	LiDAR	0.305	0.517	0.290	0.500	0.316	0.368	0.453
MEGVII [39]	test	LiDAR	**0.528**	0.300	0.247	0.379	0.245	0.140	**0.633**
LRM0	test	Camera	0.294	0.752	0.265	0.603	1.582	0.14	0.371
MonoDIS [30]	test	Camera	0.304	0.738	0.263	0.546	1.553	0.134	0.384
CenterNet [37] (HGLS)	test	Camera	0.338	0.658	0.255	0.629	1.629	0.142	0.4
Noah CV Lab	test	Camera	0.331	0.660	0.262	0.354	1.663	0.198	0.418
FCOS3D (Ours)	test	Camera	**0.358**	0.690	0.249	0.452	1.434	0.124	**0.428**
CenterNet [37] (DLA)	val	Camera	0.306	0.716	0.264	0.609	1.426	0.658	0.328
FCOS3D (Ours)	val	Camera	**0.343**	0.725	0.263	0.422	1.292	0.153	**0.415**

□□□

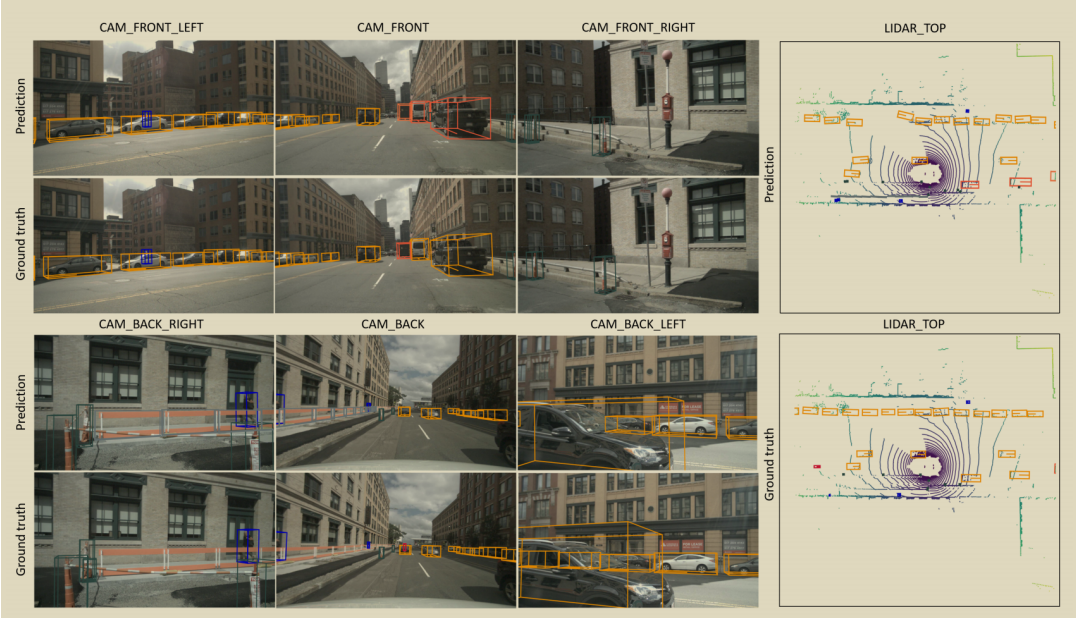
ble 2: Average precision for each class on the nuScenes test benchmark. CV and TC are abbreviation of construction hicle and traffic cone in the table.

Methods	car	truck	bus	trailer	CV	ped	motor	bicycle	TC	barrier	mAP
LRM0	0.467	0.21	0.17	0.149	0.061	0.359	0.287	0.246	0.476	0.512	0.294
MonoDIS [30]	0.478	0.22	0.188	0.176	0.074	0.37	0.29	0.245	0.487	0.511	0.304
CenterNet [37] (HGLS)	0.536	0.27	0.248	0.251	0.086	0.375	0.291	0.207	0.583	0.533	0.338
Noah CV Lab	0.515	0.278	0.249	0.213	0.066	0.404	0.338	0.237	0.522	0.49	0.331
FCOS3D (Ours)	0.524	0.27	0.277	0.255	0.117	0.397	0.345	0.298	0.557	0.538	**0.358**

Table 3: Ablation studies on the nuScenes validation 3D detection benchmark.

主要关注官方指标mAP（mean Average Precision）和NDS（Normallized Detection Score）mAP评估指标下，同为RGB图像输入的前提下，FCOS3D 0.358（测试集） 和 0.343（验证集）最佳表现，但雷达数据取得0.528更好的表现。

NDS评估指标下，同为RGB图像输入的前提下，FCOS3D 0.428（测试集） 和0.415（验证集）最佳表现，但与雷达数据相比由于引入多帧的点云数据，可以更加高效的测量速度。未来如何在单、连续帧图像中测量速度信息是一个值得探索的方向。□□□



从6个视角进行验证，预测和实际基本符合，不同的种类用不同颜色的边界框标记。

有以下几种情况：右后视图中未标记的障碍物被该模型检测到。

一些被遮挡的障碍物未被检测到。

从俯视图中可以看到对于深度和方向的预测并不那么准确。