成功检测远距离目标，将点云与RGB图像结合，谷歌&Waymo提出新算法：4D-Net

新机器视觉 2022-02-26 16:30

本文选自Google Blog，作者：AJ Piergiovanni 等
转自机器之心
编辑：陈萍、杜伟

> 来自谷歌的研究者提出了一种利用 3D 点云和 RGB 感知信息的 3D 物体检测方法：4D-Net。4D-Net 能够更好地使用运动线索和密集图像信息，成功地检测遥远的目标。

如今自动驾驶汽车和机器人能够通过激光雷达、摄像头等各种传感捕获信息。作为一种传感器，LiDAR 使用光脉冲测量场景中目标的 3D 坐标，但是其存在稀疏、范围有限等缺点——离传感器越远，返回的点就越少。

这意味着远处的目标可能只得到少数几个点，或者根本没有，而且可能无法单独被 LiDAR 采集到。同时，来自车载摄像头的图像输入非常密集，这有利于检测、目标分割等语义理解任务。凭借高分辨率，摄像头可以非常有效地检测远处目标，但在测量距离方面不太准确。

自动驾驶汽车从 LiDAR 和车载摄像头传感器收集数据。每个传感器测量值都会被定期记录，提供 4D 世界的准确表示。然而，很少有研究算法将这两者结合使用。当同时使用两种传感模式时会面临两个挑战

1) 难以保持计算效率

2) 将一个传感器的信息与另一个传感器配对会进一步增加系统复杂性，因为 LiDAR 点和车载摄像头 RGB 图像输入之间并不总是直接对应。

在发表于 ICCV 2021 的论文《 4D-Net for Learned Multi-Modal Alignment 》中，来自谷歌、Waymo 的研究者提出了一个可以处理 4D 数据（3D 点云和车载摄像头图像数据）的神经网络：4D-Net。

这是首次将 3D LiDAR 点云和车载摄像头 RGB 图像进行结合的研究。此外，谷歌还介绍了一种动态连接学习方法。最后，谷歌证明 4D-Net 可以更好地使用运动线索（motion cues）和密集图像信息来检测远处目标，同时保持计算效率。

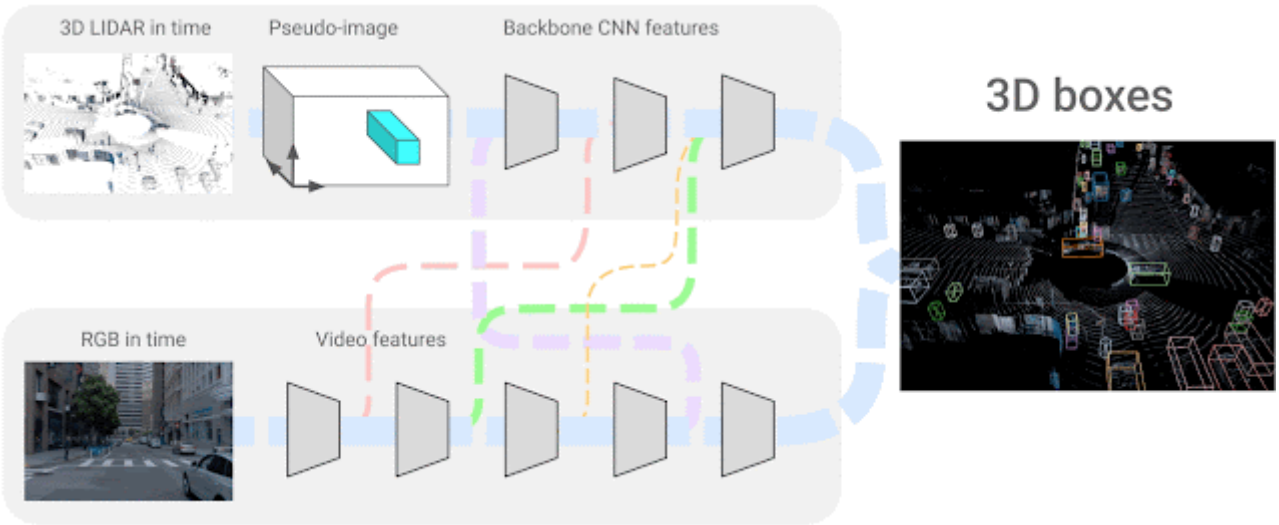
4D-Net for Learned Multi-Modal Alignment

AJ Piergiovanni, Google Research
Vincent Casser, Waymo LLC
Michael S. Ryoo, Robotics at Google
Anelia Angelova, Google Research

论文地址：https://openaccess.thecvf.com/content/ICCV2021/papers/Piergiovanni_4D-Net_for_Learned_Multi-Modal_Alignment_ICCV_2021_paper.pdf

4D-Net

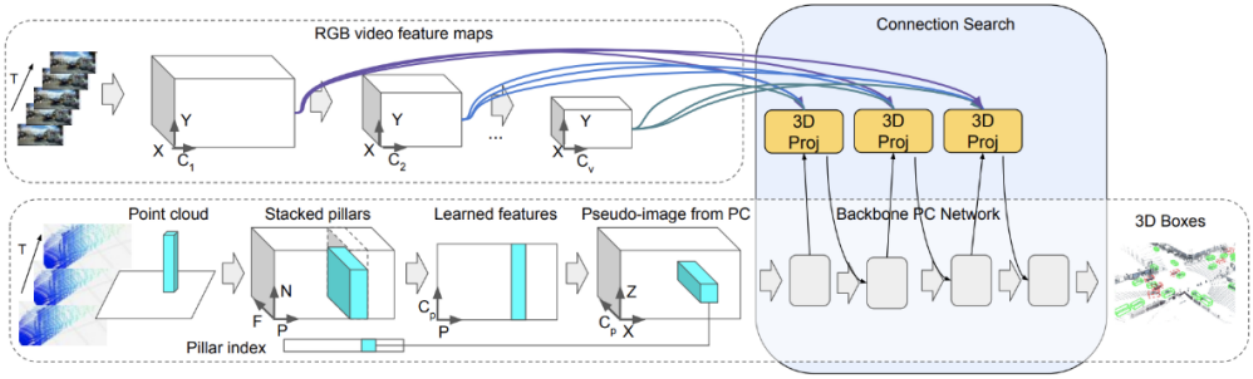
谷歌使用 4D 输入进行目标 3D 边框检测。4D-Net 有效地将 3D LiDAR 点云与 RGB 图像及时结合，学习不同传感器之间的连接及其特征表示。



谷歌使用轻量级神经架构搜索来学习两种类型的传感器输入及其特征表示之间的联系，以获得最准确的 3D 框检测。在自动驾驶领域，可靠地检测高度可变距离的目标尤为重要。

现代 LiDAR 传感器的检测范围可达数百米，这意味着更远的目标在图像中会显得更小，并且它们最有价值的特征将在网络的早期层中，与后面的层表示的近距离目标相比，它们可以更好地捕捉精细尺度的特征。

基于这一观察，谷歌将连接修改为动态的，并使用自注意力机制在所有层的特征中进行选择。谷歌应用了一个可学习的线性层，它能够将注意力加权应用于所有其他层的权重，并学习当前任务的最佳组合。



连接学习方法示意图。

结果

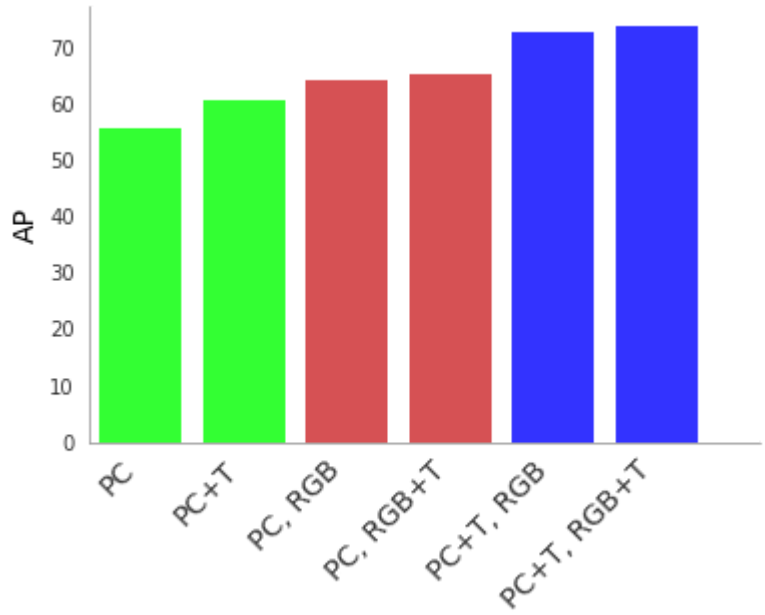
谷歌在 Waymo Open Dataset 基准中进行了测试，之前的模型只使用了 3D 点云，或单个点云和相机图像数据的组合。4D-Net 有效地使用了两种传感器输入，在 164 毫秒内处理 32 个点云和 16 个 RGB 帧，与其他方法相比性能良好。相比之下，性能次优的方法效率和准确性较低，因为它的神经网络计算需要 300 毫秒，而且比 4D-Net 使用更少的传感器输入。

3D 场景的结果。上图：与检测到的车辆相对应的 3D 框以不同颜色显示；虚线框代表丢失的目标。底部：出于可视化目的，这些框显示在相应的摄像机图像中。

检测远处的目标

4D-Net 的另一个优点是，它既利用了 RGB 提供的高分辨率，可以准确地检测到图像上的目标，又利用了点云数据提供的精确深度。因此，点云方法无法探测到的远距离目标可以被 4D-Net 探测到。这是由于相机数据的融合，能够探测到遥远的目标，并有效地将这一信息传播到网络的 3D 部分，以产生准确的探测。

为了了解 4D-Net 带来的优势，谷歌进行了一系列消融研究。实验发现，如果至少有一个传感器输入是及时流的，则可以显着提高检测准确率。及时考虑两个传感器输入可以最大程度地提高性能。

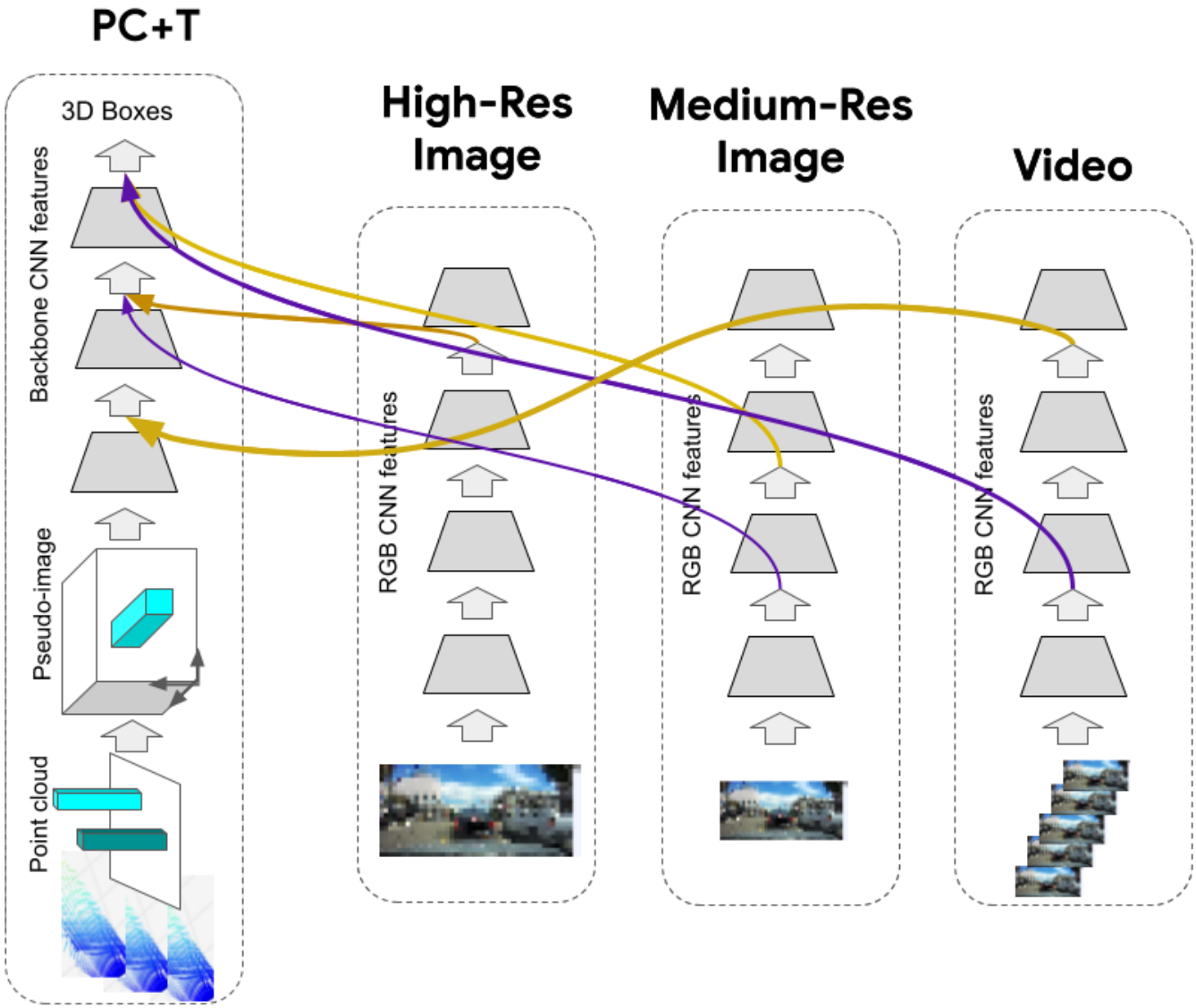


使用点云 (PC)、时间点云 (PC + T)、RGB 图像输入 (RGB) 和时间 RGB 图像 (RGB + T) 时，以平均精度 (AP) 测量 3D 目标检测的 4D-Net 性能。

多流 4D-Net

由于 4D-Net 动态连接学习机制是通用的，因此谷歌并没有局限于仅将点云流与 RGB 视频流结合起来。实际上，谷歌发现提供一个高分辨率单图像流以及一个与 3D 点云流输入结合的低分辨率视频流非常划算。因此，谷歌在下图中展示了四流（four-stream）架构的示例，该架构比具有实时点云和图像的两流架构性能更好。

动态连接学习选择特定的特征输入以连接在一起。依托多个输入流，4D-Net 必须学习多个目标特征表示之间的连接，这一点很好理解，因为算法没有改变并且只需要从输入中选择特定的特征。这是一个使用可微架构搜索的轻量级过程，可以发现模型架构内部新的连接，并进而高效地找到新的 4D-Net 模型。



多流 4D-Net 架构包含一个实时 3D 点云流以及多个图像流（高分辨率单图像流、中分辨率单图像流和更低分辨率视频流图像）。

谷歌展示了 4D-Net 是一种高效的目标检测方法，尤其适合检测远距离目标。研究者希望这项工作为未来的 4D 数据研究提供珍贵的资源。

原文链接：

https://ai.googleblog.com/

—THE END—

People who liked this content also liked

基于激光雷达点云的3D检测方法汇总(LiDAR only)

新机器视觉