Multi-Task Multi-Sensor Fusion for 3D Object Detection
多任务多传感器融合的 3D 目标检测

论文引用：

Liang M, Yang B, Chen Y, et al. Multi-Task Multi-Sensor Fusion for 3D Object Detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2019: 7345-7353.

ABSTRACT

In this paper we propose to exploit multiple related tasks for accurate multi-sensor 3D object detection. Towards this goal we present an end-to-end learnable architecture that reasons about 2D and 3D object detection as well as ground estimation and depth completion. Our experiments show that all these tasks are complementary and help the net-work learn better representations by fusing information at various levels. Importantly, our approach leads the KITTI benchmark on 2D, 3D and bird's eye view object detection, while being real-time.

摘要

在本文中，我们提出利用多个相关任务进行精确的多传感器三维物体检测。为实现这一目标，我们提出了一种端到端的可学习架构，该架构可以解释 2D 和 3D 物体检测以及地面估计和深度补全。我们的实验表明，所有这些任务都是互补的，有助于网络通过融合各种级别的信息来学习更好的表示。重要的是，我们的方法引领了 KITTI 对 2D，3D 和鸟瞰视觉对象检测的基准，同时是实时的。

1. Introduction

Self-driving vehicles have the potential to improve safety, reduce pollution, and provide mobility solutions for otherwise underserved sectors of the population. Fundamental to its core is the ability to perceive the scene in real-time. Most autonomous driving systems rely on 3dimensional perception, as it enables interpretable motion planning in bird's eye view.

1. 引文

自动驾驶车辆有可能提高安全性，减少污染，并为人口中服务不足的部门提供出行解决方案。其核心的基础是能够实时感知场景。大多数自动驾驶系统依赖于三维感知，因为它可以在鸟瞰视图中实现可解释的运动规划。

Over the past few years we have seen a plethora of methods that tackle the problem of 3D object detection from monocular images [2, 31], stereo cameras [4] or LiDAR point clouds [36, 34, 16]. However, each sensor has its challenge: cameras have difficulty capturing fine-grained 3D information, while LiDAR provides very sparse observations at long range. Recently, several attempts [5, 17, 12, 13] have been developed to fuse information from multiple sensors. Methods like [17, 6] adopt a cascade approach by using cameras in the first stage and reasoning in LiDAR point clouds only at the second stage. However, such cascade approach suffers from the weakness of each single sensor. As a result, it is difficult to detect objects that are occluded or far away. Others [5, 12, 13] have proposed to fuse multi-sensor features instead. Single-stage detectors [13] fuse multi-sensor feature maps per LiDAR point, where local nearest neighbor interpolation is used to densify the correspondence. However, the fusion is still limited when LiDAR points become extremely sparse at long range. Two-stage detectors [5, 12] fuse multi-sensor features per object

在过去几年中，我们已经看到了大量的方法来解决单目图像[2,31]，立体相机[4]或 LiDAR 点云[36,34,16]的 3D 物体检测问题。然而，每个传感器都面临着挑战：摄像机难以捕获细粒度的 3D 信息，而 LiDAR 则提供远距离非常稀疏的观测。最近，已经开发了若干尝试[5,17,12,13]来融合来自多个传感器的信息。类似[17,6]的方法采用级联方法，在第一阶段使用摄像机，仅在第二阶段使用 LiDAR 点云进行推理。然而，这种级联方法受到每个传感器的弱点的影响。结果，很难检测到被遮挡或远距离的物体。其他人[5,12,13]提出将多传感器特征融合在一起。单阶段检测器[13]为每个 LiDAR 点融合多传感器特征图，其中使用局部最近邻插值来密集对应关系。然而，当 LiDAR 点在远距离变得非常稀疏时，融合仍然是有限的。两阶段探测器[5,12]为每个目标感兴趣区域（ROI）融合多传感器特征。然而，融合过程通常很慢（因为它涉及数千个 ROI）和不精确（使用固定尺寸的锚或忽略物体方向）。

region of interest (ROI). However, the fusion process is typically slow (as it involves thousands of ROIs) and imprecise (either using fix-sized anchors or ignoring object orientation).

In this paper we argue that by solving multiple perception tasks jointly, we can learn better feature representations which result in better detection performance. Towards this goal, we develop a multi-sensor detector that reasons about 2D and 3D object detection, ground estimation and depth completion. Importantly, our model can be learned end-to-end and performs all these tasks at once. We refer the reader to Figure 1 for an illustration of our approach.

在本文中，我们认为通过联合解决多个感知任务，我们可以学习更好的特征表示，从而获得更好的检测性能。为实现这一目标，我们开发了一种多传感器探测器，可以解决 2D 和 3D 物体检测，地面估计和深度补全问题。重要的是，我们的模型可以端到端学习，并立即执行所有这些任务。我们通过参考图 1 向读者说明我们的方法。

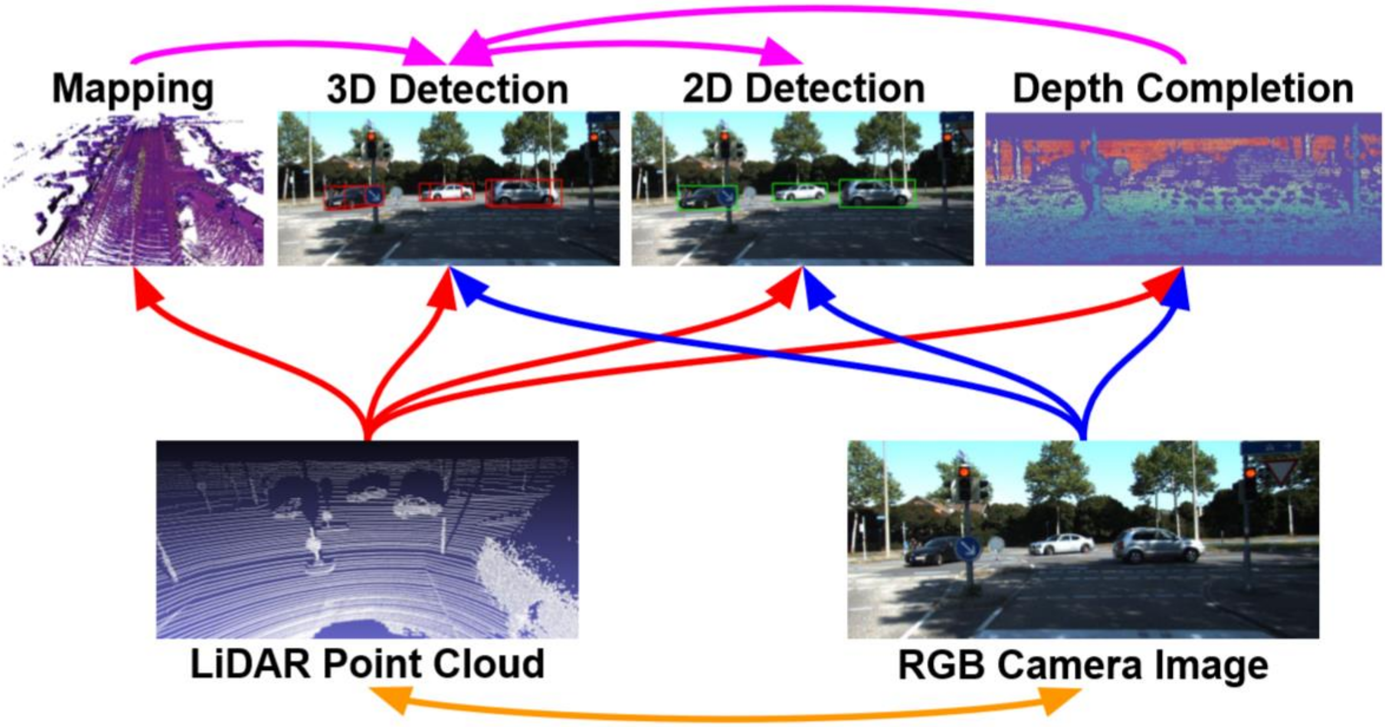


Figure 1. Different sensors (bottom) and tasks (top) are complementary to each other. We propose a joint model that reasons on two sensors and four tasks, and show that the target task - 3D object detection can benefit from multi-task learning and multi-sensor fusion.

图 1.不同的传感器（底部）和任务（顶部）互为补充。我们提出了一个联合模型，它推导出两个传感器和四个任务，并表明目标任务 - 三维物体检测可以从多任务学习和多传感器融合中获益。

We propose a new multi-sensor fusion architecture that leverages the advantages from both point-wise and ROI-wise feature fusion, resulting in fully fused feature representations. Knowledge about the location of the ground can provide useful cues for 3D object detection in the context of self-driving vehicles, as the traffic participants of interest are restrained to this plane. Our detector estimates an accurate voxel-wise ground location online as one of its auxiliary tasks. This in turn is used by the bird's eye view (BEV) backbone network to reason about relative location. We also exploit the task of depth completion to learn better cross-modality feature representation and more importantly, to achieve dense point-wise feature fusion with pseudo

我们提出了一种新的多传感器融合架构，它利用了逐点和逐 ROI 特征融合的优势，从而实现了完全融合的特征表示。关于地面位置的知识可以在自动驾驶车辆的背景下为 3D 物体检测提供有用的提示，因为感兴趣的交通参与者被限制在该平面上。我们的探测器在线估计精确的体素地面位置，作为其辅助任务之一。这反过来被鸟瞰（BEV）骨干网用于推断相对位置。我们还利用深度补全的任务来学习更好的跨模态特征表示，更重要的是，利用密集深度的伪 LiDAR 点实现密集的逐点特征融合。

LiDAR points from dense depth.

We demonstrate the effectiveness of our approach on the KITTI object detection benchmark [8] as well as the more challenging TOR4D object detection benchmark [34]. On the KITTI benchmark, we show very significant performance improvement over previous state-of-the-art approaches in 2D, 3D and BEV detection tasks. Meanwhile, the proposed detector runs over 10 frames per second, making it a practical solution for real-time application. On the TOR4D benchmark, we show detection improvement from multi-task learning over previous state-of-the-art detector.

我们证明了我们的方法对 KITTI 物体检测基准[8]的有效性以及更具挑战性的 TOR4D 物体检测基准[34]。在 KITTI 基准测试中，我们展示了与 2D，3D 和 BEV 检测任务中先前最先进方法相比非常显着的性能改进。同时，所提出的探测器每秒运行超过 10 帧，使其成为实时应用的实用解决方案。在 TOR4D 基准测试中，我们展示了与先前最先进的探测器相比，多任务学习的检测改进。

2. Related Work

2. 相关工作

We review related works that exploit multi-sensor fusion and multi-task learning to improve 3D object detection.

我们回顾了利用多传感器融合和多任务学习来改进 3D 对象检测的相关工作。

3D detection from single modality: Early approaches to 3D object detection focus on camera based solutions with monocular or stereo images [3, 2]. However, they suffer from the inherent difficulties of estimating depth from images and as a result perform poorly in 3D localization. More recent 3D object detectors rely on depth sensors such as LiDAR [34, 36]. However, although range sensors provide precise depth measurements, the observations are usually sparse (particularly at long range) and lack the information richness of images. It is thus difficult to distinguish classes such as pedestrian and cyclist with LiDAR-only detectors.

单模态的 3D 检测：早期的 3D 物体检测方法主要集中在基于摄像头的单目或立体图像解决方案[3,2]。然而，它们遭受了图像估计深度的固有困难，因此在 3D 定位中表现不佳。更新近的 3D 物体探测器依赖于深度传感器，例如 LiDAR [34,36]。然而，尽管距离传感器提供精确的深度测量，但观测通常是稀疏的（特别是在远距离）并且缺乏图像的信息丰富度。因此，仅限 LiDAR 的探测器难以将诸如行人和骑车人的类别区分开来。

Multi-sensor fusion for 3D detection: Recently, a variety of 3D detectors that exploit multiple sensors (e.g. LiDAR and camera) have been proposed. F-PointNet [17] uses a cascade approach to fuse multiple sensors. Specifically, 2D object detection is done first on images, 3D frustums are then generated by projecting 2D detections to 3D and PointNet [18, 19] is applied to regress the 3D position and shape of the bounding box. In this framework the overall performance is bounded by each stage which is still using single sensor. Furthermore, object localization from a frustum in LiDAR point cloud has difficulty dealing with occluded or far away objects as LiDAR observation can be very sparse (often with a single point on the far away object). MV3D [5] generates 3D proposals from LiDAR features, and refines the detections with ROI feature fusion from LiDAR and image feature maps. AVOD [12] further extends ROI feature fusion to the proposal generation stage to improve the object proposal quality. However, ROI feature

用于 3D 检测的多传感器融合：最近，已经提出了利用多个传感器（例如，LiDAR 和相机）的各种 3D 检测器。F-PointNet [17]使用级联方法融合多个传感器。具体而言，2D 物体检测首先在图像上完成，然后通过将 2D 检测投影到 3D 来生成 3D 平截锥体，并且应用 PointNet [18,19]来回归边界框的 3D 位置和形状。在这个框架中，整体性能受到仍在使用单个传感器的每个阶段的限制。此外，LiDAR 点云中的平截锥体的物体定位难以处理被遮挡或远处的物体，因为 LiDAR 观察可能非常稀疏（通常在远处的物体上有一个点）。MV3D [5]从 LiDAR 特征生成 3D 提议，并通过 LiDAR 和图像特征映射的 ROI 特征融合来重新确定检测结果。AVOD [12]进一步将 ROI 特征融合扩展到提议生成阶段，以提高对象提议质量。但是，ROI 特征融合仅在高级特征映射中发生。此外，它仅融合选定对象区域的特征而不是特征图上的密集位置。为了克服这个缺点，ContFuse [13]使用连续卷积[30]来融合来自每个传感器的多尺度卷积特征图，其中图像和 BEV 空间之间的对应关系是通过投影 LiDAR fusion happens only at high-level feature maps. Furthermore, it only fuses features at selected object regions instead of dense locations on the feature map. To overcome this drawback, ContFuse [13] uses continuous convolution [30] to fuse multi-scale convolutional feature maps from each sensor, where the correspondence between image and BEV spaces is achieved through projection of the LiDAR points. However, such fusion is limited when LiDAR points are very sparse. To address this issue, we propose to predict dense depth from multi-sensor data, and use the predicted depth as pseudo LiDAR points to find dense correspondences between multi-sensor feature maps.

3D detection from multi-task learning: Various auxiliary tasks have been exploited to help improve 3D object detection. HDNET [33] exploits geometric ground shape and semantic road mask for BEV vehicle detection. SBNet [21] utilizes the sparsity in road mask to speed up 3D detection by > 2 times. Our model also reasons about a geometric map. The difference is that this module is part of our detector and thus end-to-end trainable, so that these two tasks can be optimized jointly. Wang et al. [29] exploit depth reconstruction and semantic segmentation to help 3D object detection. However, they rely on 3D rendering, which is computationally expensive. Other contextual cues such as the room layout [23, 26], and support surface [24] have also been exploited to help 3D object reasoning in the context of indoor scenes. 3DOP [3] exploits monocular depth estimation to refine the 3D shape and position based on 2D proposals. Mono3D [2] uses instance segmentation and semantic segmentation as evidence, along with other geometric priors to reason about 3D object detection from monocular images. Apart from geometric map estimation, we also exploit depth completion which brings two benefits: it guides the network to learn better cross-modality feature representations, and its prediction serves as pseudo LiDAR points for dense fusion between image and BEV feature maps.

3. Multi-Task Multi-Sensor Detector

One of the fundamental tasks in autonomous driving is to perceive the scene in real-time. In this paper we propose a multi-task multi-sensor fusion model for the task of 3D object detection. We refer the reader to Figure 2 for an illustration of the model architecture. Our approach has the following highlights. First, we design a multi-sensor architecture that combines point-wise and ROI-wise feature fusion. Second, our integrated ground estimation module reasons about the geometry of the road. Third, we exploit

点来实现的。然而，当 LiDAR 点非常稀疏时，这种融合是有限的。为了解决这个问题，我们建议从多传感器数据预测密集深度，并使用预测深度作为伪 LiDAR 点，以找到多传感器特征图之间的密集对应关系。

来自多任务学习的 3D 检测：已经利用各种辅助任务来帮助改进 3D 对象检测。HDNET [33]利用几何地形和语义道路掩模进行 BEV 车辆检测。SBNet [21]利用道路掩模中的稀疏性将 3D 检测速度提高了 2 倍以上。我们的模型也推理几何图。不同之处在于，该模块是我们探测器的一部分，因此端到端可训练，因此可以共同优化这两个任务。Wang 等人[29]利用深度重建和语义分割来帮助 3D 对象检测。然而，它们依赖于 3D 渲染，这在计算上是昂贵的。其他上下文线索，例如房间布局[23,26]和支撑平面[24]也被用于在室内场景的背景下帮助 3D 对象推理。3DOP [3]利用单目深度估计来基于 2D 提议重新确定 3D 形状和位置。Mono3D [2]使用实例分割和语义分割作为证据，以及其他几何先验来推断单目图像的三维物体检测。除了几何图估计之外，我们还利用深度补全带来两个好处：它引导网络学习更好的跨模态特征表示，并且其预测用作图像和 BEV 特征图之间的密集融合的伪 LiDAR 点。

3. 多任务多传感器探测器

自动驾驶的基本任务之一是实时感知场景。在本文中，我们提出了一个多任务多传感器融合模型，用于三维物体检测任务。我们将读者引用到图 2 中以获得模型体系结构的图示。我们的方法有以下亮点。首先，我们设计了一种多传感器架构，它结合了逐点和逐 ROI 特征融合。其次，我们的综合地面估算模块有关道路几何形状的推理。第三，我们利用深度补全的任务来学习更好的多传感器特征并实现密集的逐点特征融合。因此，可以通过利用多任务损失来端到端地学习整个模型。

the task of depth completion to learn better multi-sensor features and achieve dense point-wise feature fusion. As a result, the whole model can be learned end-to-end by exploiting a multi-task loss.

In the following, we first introduce the architecture of the multi-sensor 2D and 3D detector with point-wise and ROI-wise feature fusion. We then show how we exploit the other two auxiliary tasks to further improve 3D detection. Finally we provide details of how to train our model end-to-end.

在下文中，我们首先介绍了具有逐点和逐 ROI 特征融合的多传感器 2D 和 3D 探测器的架构。然后，我们将展示如何利用其他两个辅助任务来进一步改进 3D 检测。最后，我们提供了如何对端到端训练模型的详细信息。

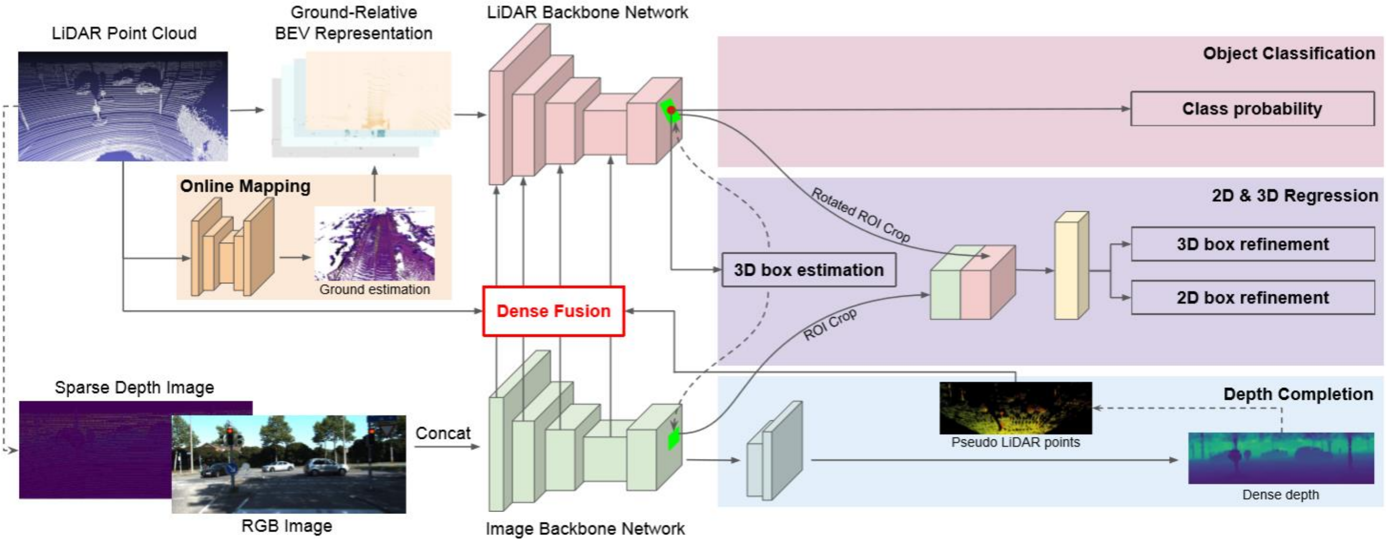


Figure 2. The architecture of the proposed multi-task multi-sensor fusion model for 2D and 3D object detection. Dashed arrows denote projection, while solid arrows denote data flow. Our model is a simplified two-stage detector with densely fused two-stream multi-sensor backbone networks. The first stage is a single-shot detector that outputs a small number of high-quality 3D detections. The second stage applies ROI feature fusion for more precise 2D and 3D box regression. Ground estimation is explored to incorporate geometric ground prior to the LiDAR point cloud. Depth completion is exploited to learn better cross-modality feature representation and achieve dense feature map fusion by transforming predicted dense depth image into dense pseudo LiDAR points. The whole model can be learned end-to-end.

图2.用于 2D 和3D 物体检测的所提出的多任务多传感器融合模型的架构。虚线箭头表示投影，而实线箭头表示数据流。我们的模型是一个简化的两级探测器，具有密集融合的双流多传感器骨干网络。第一阶段是单发探测器，可输出少量高质量3D 探测器。第二阶段应用ROI 特征融合，以实现更精确的2D 和3D 框回归。探索地面估计以在LiDAR 点云之前合并几何地面。利用深度补全来学习更好的跨模态特征表示，并通过将预测的密集深度图像变换为密集的伪LiDAR 点来实现密集的特征图融合。整个模型可以端到端地学习。

3.1 Fully Fused Multi-Sensor Detector

Our multi-sensor detector takes a LiDAR point cloud and an RGB image as input. The backbone network adopts the two-stream structure, where one stream extracts image feature maps, and the other extracts LiDAR BEV feature maps. Point-wise feature fusion is applied to fuse multiscale image features to BEV stream. The final BEV feature map predicts dense 3D detections per BEV voxel via 2D convolution. After Non-Maximum Suppression (NMS) and score thresholding, we get a small number of high-quality 3D detections and their projected 2D detections (typically fewer

3.1 多传感器完全融合探测器

我们的多传感器探测器采用 LiDAR 点云和 RGB 图像作为输入。骨干网采用双流结构，其中一个流提取图像特征映射，另一个流提取 LiDAR BEV 特征映射。逐点特征融合用于将多尺度图像特征融合到 BEV 流。最终的 BEV 特征图通过 2D 卷积预测每个 BEV 体素的密集 3D 检测。在非最大抑制（NMS）和得分阈值之后，我们获得少量高质量 3D 检测及其投影的 2D 检测（在 KITTI 数据集上测试时通常少于 20）。然后，我们通过逐 ROI 特征融合应用 2D 和 3D 盒子改进，其中我们结合了来自图像 ROI 和 BEV 导向 ROI 的特征。完成后，探测器

than 20 when tested on KITTI dataset). We then apply a 2D and 3D box refinement by ROI-wise feature fusion, where we combine features from both image ROIs and BEV oriented ROIs. After the refinement, the detector outputs accurate 2D and 3D detections.

输出准确的 2D 和 3D 检测。

Input representation: We use the voxel based LiDAR representation [13] due to its efficiency. In particular, we voxelize the point cloud into a 3D occupancy grid, where the voxel feature is computed via 8-point linear interpolation on each LiDAR point. This LiDAR representation is able to capture fine-grained point density clues efficiently. We consider the resulting 3D volume as BEV representation by treating the height slices as feature channels. This allows us to reason in 2D BEV space, which brings significant efficiency gain with no performance drop. We simply use the RGB image as input for the camera stream. When we exploit the auxiliary task of depth completion, we additionally add a sparse depth image generated by projecting the LiDAR points to the image.

输入表示：由于其效率，我们使用基于体素的 LiDAR 表示[13]。特别地，我们将点云体素化为 3D 占用网格，其中通过每个 LiDAR 点上的 8 点线性插值来计算体素特征。这种 LiDAR 表示能够有效地捕获细粒度的点密度线索。我们将高度切片视为特征通道，将得到的 3D 体积视为 BEV 表示。这使我们能够在 2D BEV 空间中进行推理，这样可以在不降低性能的情况下获得显着的效率增益。我们只是使用 RGB 图像作为相机流的输入。当我们利用深度补全的辅助任务时，我们另外添加通过将 LiDAR 点投影到图像而生成的稀疏深度图像。

Network architecture: The backbone network follows a two-stream architecture [13] to process multi-sensor data. Specifically, for the image stream we use the pre-trained ResNet-18 [10] until the fourth convolutional block. Each block contains 2 residual layers with number of feature maps increasing from 64 to 512 linearly. For the LiDAR stream, we use a customized residual network which is deeper and thinner than ResNet-18 for a better trade-off between speed and accuracy. In particular, we have four residual blocks with 2, 4, 6, 6 residual layers in each, and the numbers of feature maps are 64, 128, 192 and 256. We also remove the max pooling layer before the first residual block to maintain more details in the point cloud feature. In both streams we apply a feature pyramid network (FPN) [14] with 1 × 1 convolution and bilinear up-sampling to combine multi-scale features. As a result, the final feature maps of the two streams have a down-sampling factor of 4 compared with the input.

网络架构：骨干网遵循双流架构[13]来处理多传感器数据。具体而言，对于图像流，我们使用预先训练的 ResNet-18 [10]直到第四个卷积块。每个块包含 2 个残留层，其中多个特征映射线性地从 64 增加到 512。对于 LiDAR 流，我们使用比 ResNet-18 更深更薄的定制残差网络，以便在速度和准确度之间进行更好的权衡。特别地，我们有四个残差块，每个块有 2,4,6,6 个残差层，特征映射的数量是 64,128,192 和 256.我们还在第一个残差块之前移除最大池化层以维持点云功能的更多细节。在两个流中，我们应用特征金字塔网络（FPN）[14]和 1×1 卷积和双线性上采样来组合多尺度特征。因此，与输入相比，两个流的最终特征图具有 4 的下采样因子。

On top of the last BEV feature map, we simply add a 1 × 1 convolution to perform dense 3D object detection. After score thresholding and oriented NMS, a small number of high-quality 3D detections are projected to both BEV space and 2D image space, and their ROI features are cropped from each stream's last feature map via precise ROI feature extraction. The multi-sensor ROI features are fused together and fed into a refinement module with two 256-dimension fully connected layers to predict the 2D and 3D

在最后一个 BEV 特征图之上，我们只需添加 1×1 卷积即可执行密集的 3D 对象检测。在评分阈值和定向 NMS 之后，将少量高质量 3D 检测投射到 BEV 空间和 2D 图像空间，并且通过精确的 ROI 特征提取从每个流的最后特征图中裁剪它们的 ROI 特征。多传感器 ROI 功能融合在一起，并送入具有两个 256 维全连接层的改善模块，分别预测每个 3D 检测的 2D 和 3D 盒子重建。

box refinements for each 3D detection respectively.

Point-wise feature fusion: We apply point-wise feature fusion between the convolutional feature maps of LiDAR and image streams (as shown in Figure reffig: point). The fusion is directed from image steam to LiDAR steam to augment BEV features with information richness of image features. We gather multi-scale features from all four blocks in the image backbone network with a feature pyramid network. The resulting multi-scale image feature map is then fused to each block of the LiDAR BEV backbone network.

逐点特征融合：我们在 LiDAR 的卷积特征图和图像流之间应用逐点特征融合（如图所示：点）。融合从图像流引导到 LiDAR 流，以增强具有图像特征信息丰富度的 BEV 特征。我们通过功能金字塔网络从图像骨干网络中的所有四个块收集多尺度特征。然后将得到的多尺度图像特征图融合到 LiDAR BEV 骨干网的每个块。

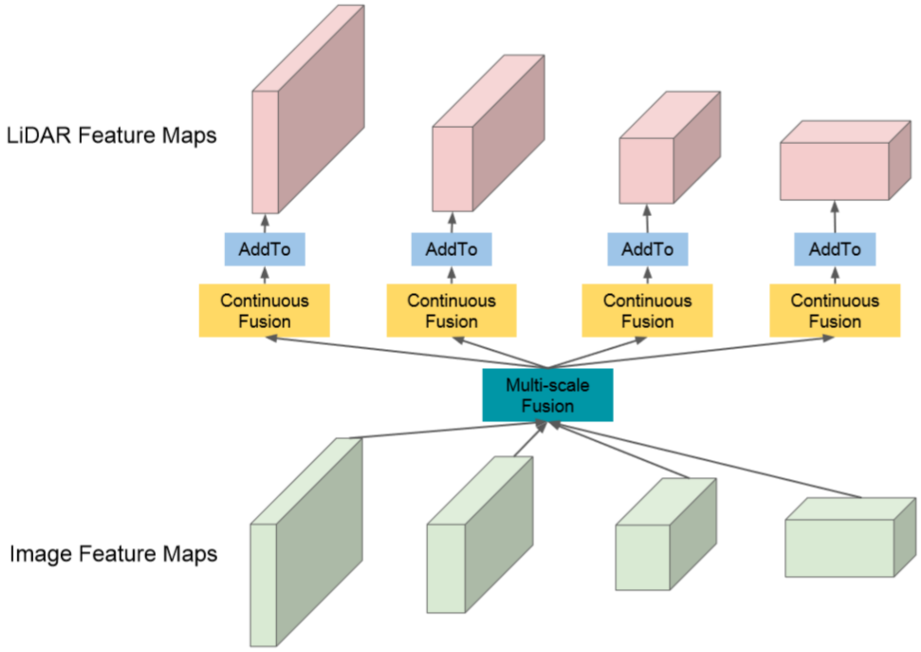


Figure 3. Point-wise feature fusion between LiDAR and image backbone networks. A feature pyramid network is used to combine multi-scale image feature maps, followed with a continuous fusion layer to project image feature map into BEV space. Feature fusion is implemented by element-wise summation.

图3. LiDAR 与图像骨干网络之间的逐点特征融合。特征金字塔网络用于组合多尺度图像特征图，随后是连续融合图层以将图像特征图投影到BEV 空间中。特征融合通过逐元素求和来实现。

To fuse image feature map with BEV feature map, we need to find the pixel-wise correspondence between the two sensors. Inspired by [13], we use LiDAR points to establish dense and accurate correspondence between the image and BEV feature maps. For each pixel in the BEV feature map, we find its nearest LiDAR point and project the point onto the image feature map to retrieve the corresponding image feature. We compute the distance between the BEV pixel and LiDAR point as the geometric feature. Both retrieved image feature and the BEV geometric feature are passed into a Multi-Layer Perceptron (MLP) and the output is fused to BEV feature map by element-wise addition. Note that such point-wise feature fusion is sparse by nature of LiDAR observation. Later we explain how we exploit dense depth as pseudo LiDAR points to provide dense correspondence for dense point-wise fusion.

为了将图像特征图与 BEV 特征图融合，我们需要找到两个传感器之间的像素对应关系。受[13]的启发，我们使用 LiDAR 点在图像和 BEV 特征图之间建立密集和精确的对应关系。对于 BEV 特征图中的每个像素，我们找到其最近的 LiDAR 点并将该点投影到图像特征图上以检索相应的图像特征。我们计算 BEV 像素和 LiDAR 点之间的距离作为几何特征。检索到的图像特征和 BEV 几何特征都被传递到多层感知器（MLP）中，并且通过逐元素添加将输出融合到 BEV 特征图。注意，这种逐点特征融合本质上是稀疏的 LiDAR 观察。稍后我们将解释如何利用密集深度作为伪 LiDAR 点来为密集点式融合提供密集对应。

ROI-wise feature fusion: The motivation of the ROI-wise feature fusion is to further refine the localization precision of the high-quality detections in 2D and 3D spaces respectively. Towards this goal, the ROI feature extraction itself needs to be precise so as to properly predict the relative box refinement. By projecting a 3D detection onto the image and BEV feature maps, we get an axis-aligned image ROI and an oriented BEV ROI. We adopt ROIAlign [9] to extract features from an axis-aligned image ROI.

逐 ROI 特征融合：逐 ROI 特征融合的动机是进一步分别在 2D 和 3D 空间中重新定义高质量检测的定位精度。为实现这一目标，ROI 特征提取本身需要精确，以便正确预测相关的盒子改进。通过将 3D 检测投影到图像和 BEV 特征图上，我们得到轴对齐图像 ROI 和定向 BEV ROI。我们采用 ROIAlign [9]从轴对齐图像 ROI 中提取特征。

For oriented BEV ROI feature extraction, however, we observe two new issues (as shown in Figure 4). First, the periodicity of the ROI orientation causes the reverse of feature order around the cycle boundary. To solve this issue, we propose an oriented ROI feature extraction module with anchors. Given an oriented ROI, we first assign it to one of the two orientation anchors, 0 or 90 degrees. Each anchor has a consistent feature extraction order. The two anchors share the refinement net except for the output layer. Second, when the ROI is rotated, its location offset has to be parametrized in the rotated coordinates as well. In practice, we rotate the axis-aligned location offset to be aligned with the ROI orientation axes. Similar to ROIAlign[9], we extract bilinear interpolated feature into a n × n regular grid for the BEV ROI (in practice we use n = 5).

然而，对于定向 BEV ROI 特征提取，我们观察到两个新问题（如图 4 所示）。首先，ROI 方向的周期性导致周期边界周围的特征顺序的反转。为了解决这个问题，我们提出了一个带锚的定向 ROI 特征提取模块。给定定向 ROI，我们首先将其分配给两个定向锚之一，0 或 90 度。每个锚具有一致的特征提取顺序。除输出层外，两个锚点共享改进网络。其次，当旋转 ROI 时，其位置偏移也必须在旋转的坐标中进行参数化。在实践中，我们旋转轴对齐的位置偏移以与 ROI 方向轴对齐。与 ROIAlign [9]类似，我们将双线性插值特征提取到 BEV ROI 的 n×n 规则网格中（实际上我们使用 n = 5）。

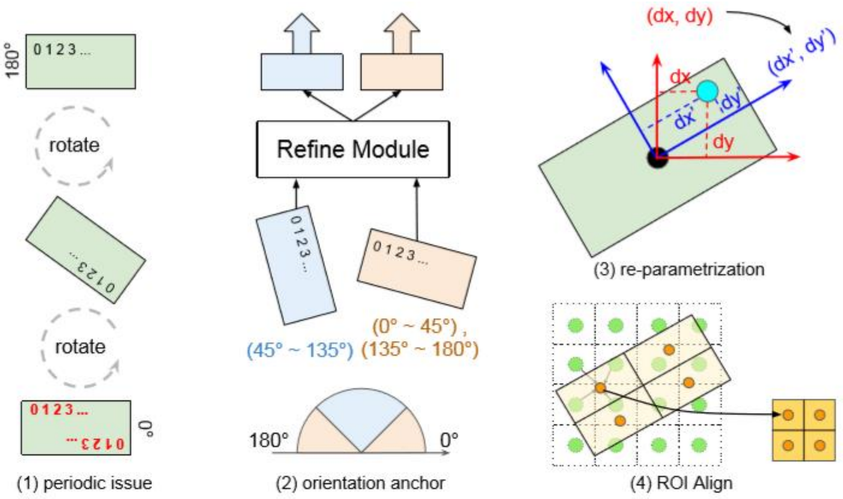


Figure 4. Precise rotated ROI feature extraction that takes orientation cycle into account. (1) The rotational periodicity causes reverse of order in feature extraction. (2) ROI refine module with two orientation anchors. An ROI is assigned to 0◦ or 90◦. They share most refining layers except for the output. (3) The regression target of relative offsets are re-parametrized with respect to the object orientation axes. (4) A n × n sized feature is extracted using bilinear interpolation (we show an example with n = 2).

图 4.精确旋转的 ROI 特征提取，将定向周期考虑在内。（1）旋转周期性导致特征提取中的顺序颠倒。（2）具有两个定向锚的 ROI 调整模块。ROI 分配为 0°或 90°。除输出外，它们共享大多数重新定义的层。（3）相对偏移的回归目标相对于物体定向轴重新参数化。（4）使用双线性插值提取 n×n 大小的特征（我们示出 n = 2 的示例）。

3.2 Multi-Task Learning for 3D Detection

In this paper we exploit two auxiliary tasks to improve 3D object detection, namely ground estimation and depth completion. They help in different ways: ground estimation provides geometric prior to canonicalize the LiDAR point clouds. Depth completion guides the image network to learn better cross-modality feature representations, and further facilitates dense point-wise feature fusion.

3.2.1 Ground estimation

Mapping is an important task for autonomous driving, and in most cases the map building process is done offline. However, online mapping is appealing for that it decreases the system's dependency on offline built maps and increases the system's robustness. Here we focus on one specific subtask in mapping, which is to estimate the road geometry on-the-fly from a single LiDAR sweep. We formulate the task as a regression problem, where we estimate the ground height value for each voxel in the BEV space. This formulation is more accurate than plane based parametrization [3, 1], as in practice the road is often curved especially when we look far ahead.

Network architecture: We apply a small fully convolutional U-Net [25] on top of the LiDAR BEV representation to estimate the normalized voxel-wise ground heights at an inference time of 8 millisecond. We choose the U-Net architecture because it outputs prediction at the same resolution as the input, and is good at capturing global context while maintaining low-level details.

Map fusion: Given a voxel-wise ground estimation, we first extract point-wise ground height by looking for the point index during voxelization. We then subtract the ground height from each LiDAR point's Z axis value and generate a new LiDAR BEV representation (relative to ground), which is fed to the LiDAR backbone network. In the regression part of the 3D detection, we add the ground height back to the predicted Z term. Online ground estimation eases 3D object localization as traffic participants of interest all lay on the ground.

3.2.2 Depth completion

LiDAR provides long range 3D information for accurate 3D object detection. However, the observation is sparse especially at long range. Here, we propose to densify LiDAR points by depth completion from both sparse LiDAR observation and RGB image. Specifically, given the projected (into the image plane) sparse depth from the

3.2 多任务学习的 3D 检测

在本文中，我们利用两个辅助任务来改进三维物体检测，即地面估计和深度补全。它们以不同的方式提供帮助：地面估计在规范化 LiDAR 点云提供几何先验。深度补全指导图像网络学习更好的跨模态特征表示，并进一步促进密集的逐点特征融合。

3.2.1 地面估计

建图是自动驾驶的一项重要任务，在大多数情况下，地图构建过程是离线完成的。但是，在线建图很有吸引力，因为它降低了系统对离线建图的依赖性，并提高了系统的稳健性。在这里，我们关注建图中的一个特定子任务，即从单个 LiDAR 扫描估计道路几何形状。我们将任务表示为回归问题，我们在其中估计 BEV 空间中每个体素的地面高度值。这种公式比基于平面的参数化更准确[3,1]，因为在实践中，道路通常是弯曲的，特别是当我们向前看时。

网络架构：我们在 LiDAR BEV 表示之上应用一个小的完全卷积 U-Net [25]，以在 8 毫秒的推断时间内估计标准化的体素地面高度。我们选择 U-Net 架构是因为它以与输入相同的分辨率输出预测，并且擅长捕获全局上下文，同时保持低级细节。

地图融合：给定体素地面估计，我们首先通过在体素化期间寻找点指数来提取逐点地面高度。然后，我们从每个 LiDAR 点的 Z 轴值中减去地面高度，并生成新的 LiDAR BEV 表示（相对于地面），将其馈送到 LiDAR 骨干网络。在 3D 检测的回归部分中，我们将地面高度添加回预测的 Z 项。在线地面估算可以简化三维物体的定位，因为所有有兴趣的交通参与者都在地面上。

3.2.2 深度补全

LiDAR 提供长距离 3D 信息，以实现精确的 3D 物体检测。然而，观察是稀疏的，特别是在远距离。在这里，我们建议通过稀疏 LiDAR 观察和 RGB 图像的深度完成来增加 LiDAR 点的密度。具体而言，假设从 LiDAR 点云和相机图像投影（进入图像平面）稀疏深度，模型以与输入图像相同的分辨率输出密集深度。

LiDAR point cloud and a camera image, the model outputs dense depth at the same resolution as the input image.

Sparse depth image from LiDAR projection: We first generate a three-channel sparse depth image from the LiDAR point cloud, representing the sub-pixel offsets and the depth value. Specifically, we project each LiDAR point (x,y,z) to the camera space, denoted as $(x_{cam},y_{cam},z_{cam})$ (the Z axis points to the front of the camera), where z_{cam} is the depth of the LiDAR point in camera space. We then project the point from camera space to image space, denoted as (x_{im},y_{im}). We find the pixel (u,v) closest to (x_{im},y_{im}), and compute $(x_{im} - u, y_{im} - v, z_{cam}/10)$ as the value of pixel (u,v) on the sparse depth image. For pixel locations with no LiDAR point, we set the pixel value to zero. The resulting sparse depth image is then concatenated with the RGB image and fed to the image backbone network.

来自 LiDAR 投影的稀疏深度图像：我们首先从 LiDAR 点云生成三通道稀疏深度图像，表示子像素偏移和深度值。具体来说，我们将每个 LiDAR 点(x,y,z)投影到相机空间，表示为$(x_{cam},y_{cam},z_{cam})$（Z 轴指向相机的前部），其中 z_{cam} 是 LiDAR 点的深度在相机空间。然后我们将点从相机空间投影到图像空间，表示为(x_{im},y_{im})。我们找到最接近(x_{im},y_{im})的像素(u,v)，并计算$(x_{im} - u, y_{im} - v, z_{cam}/10)$作为稀疏深度图像上的像素(u,v)的值。对于没有 LiDAR 点的像素位置，我们将像素值设置为零。然后将得到的稀疏深度图像与 RGB 图像连接并馈送到图像骨干网络。

Network architecture: The depth completion network shares the feature representation with the image backbone network, and applies four convolutional layers accompanied with two bilinear up-sampling layers afterwards to predict the dense pixel-wise depth at the same resolution as the input image.

网络架构：深度补全网络与图像骨干网络共享特征表示，然后应用四个卷积层，随后伴随两个双线性上采样层，以与输入图像相同的分辨率预测密集像素方式的深度。

Dense depth for dense point-wise feature fusion: As mentioned above, the point-wise feature fusion relies on LiDAR points to find the correspondence between multisensor feature maps. However, since LiDAR observation is sparse by nature, the point-wise feature fusion is sparse. In contrast, the depth completion task provides dense depth estimation per image pixel, and therefore can be used as “pseudo” LiDAR points to find dense pixel correspondence between multi-sensor feature maps. In practice, we use true and pseudo LiDAR points together in fusion and resort to pseudo points only when true points are not available.

密集深度的逐点特征融合：如上所述，逐点特征融合依赖于 LiDAR 点来找到多传感器特征图之间的对应关系。然而，由于 LiDAR 观察本质上是稀疏的，因此逐点特征融合是稀疏的。相反，深度完成任务提供每个图像像素的密集深度估计，因此可以用作“伪”LiDAR 点以找到多传感器特征图之间的密集像素对应。在实践中，我们在融合中使用真实和伪 LiDAR 点，并且仅当真实点不可用时才使用伪点。

3.3 Joint Training

We employ multi-task loss to train our multi-sensor detector end-to-end. The full model outputs object classification, 3D box estimation, 2D and 3D box refinement, ground estimation and dense depth. During training, we have detection labels and dense depth labels, while ground estimation is optimized implicitly by the 3D localization loss. There are two paths of gradient transmission for ground estimation. One is from the 3D box output where ground height is added back to predicted Z term. The other goes through the LiDAR backbone network to the LiDAR

3.3 联合训练

我们采用多任务损失来训练我们的多传感器端到端检测器。完整模型输出对象分类，3D 盒估计，2D 和 3D 盒子重建，地面估计和密集深度。在训练期间，我们有检测标签和密集深度标签，而地面估计则由 3D 定位损失隐含地优化。用于地面估计的梯度传输有两条路径。一个来自 3D 盒子输出，其中地面高度被添加回预测的 Z 项。另一个通过 LiDAR 骨干网络到 LiDAR 体素化层，其中从每个 LiDAR 点的 Z 坐标减去地面高度。

voxelization layer where ground height is subtracted from the Z coordinate of each LiDAR point.

For object classification loss L_{cls}, we use binary cross entropy. For 3D box estimation loss L_{box} and 3D box refinement loss L_{r3d}, we compute smooth ℓ_1 loss on each dimension of the 3D object $(x, y, z, \log(w), \log(l), \log(h), \theta)$, and sum over positive samples. For 2D box refinement loss L_{r2d}, we similarly compute smooth ℓ_1 loss on each dimension of the 2D object $(x, y, \log(w), \log(h))$, and sum over positive samples. For dense depth prediction loss L_{depth}, we sum ℓ_1 loss over all pixels. The total loss is defined as follows:

对于对象分类损失 L_{cls}，我们使用二进制交叉熵。对于 3D 盒估计损失 L_{box} 和 3D 盒重建损失 L_{r3d}，我们计算 3D 对象的每个维度上的平滑 ℓ_1 损失 $(x, y, z, \log(w), \log(l), \log(h), \theta)$，并且对正样本求和。对于 2D 盒修复损失 L_{r2d}，我们类似地计算 2D 对象的每个维度上的平滑 ℓ_1 损失 $(x, y, \log(w), \log(h))$，并且对正样本求和。对于密集深度预测损失 L_{depth}，我们在所有像素上总和 ℓ_1 损失。总损失定义如下：

$$Loss = L_{cls} + \lambda(L_{box} + L_{r2d} + L_{r3d}) + \gamma L_{depth}$$

where λ,γ are the weights to balance different tasks.

其中 λ,γ 为平衡不同任务的权重。

A good initialization is important for faster convergence. We use the pre-trained ResNet-18 network to initialize the image backbone network. For the additional channels of the sparse depth image at the input, we set the corresponding weights to zero. We also pre-train the ground estimation network on TOR4D dataset [34] with offline maps as labels and ℓ2 loss as objective function [33]. Other networks in the model are initialized randomly. We train the model with stochastic gradient descent using Adam optimizer [11].

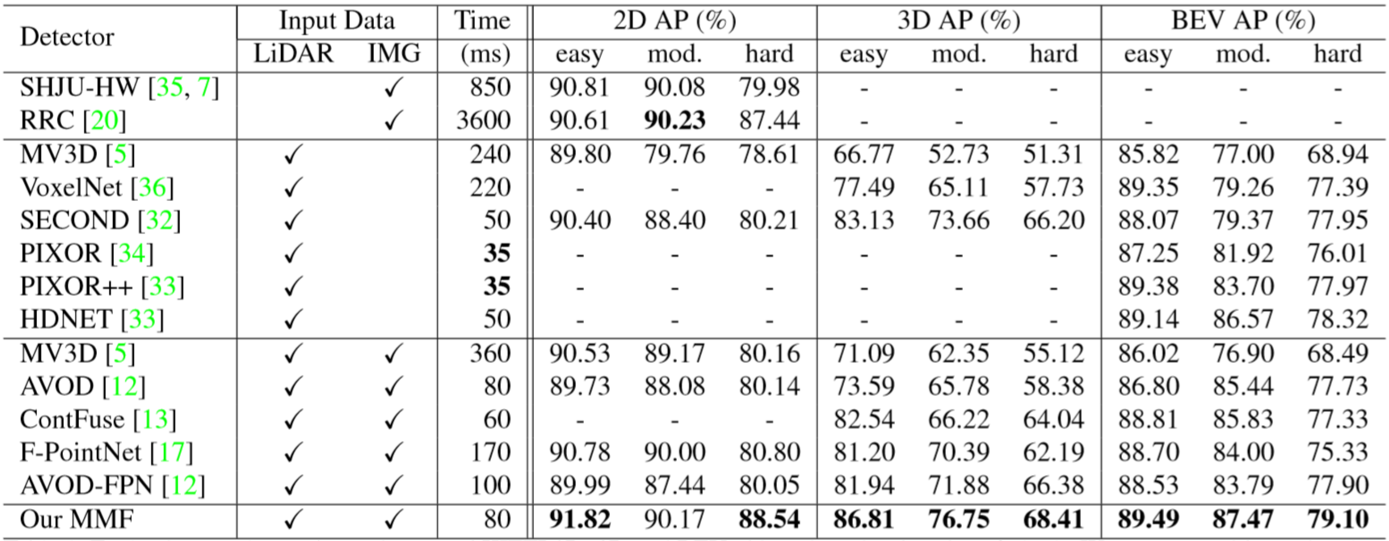
良好的初始化对于更快的收敛非常重要。我们使用预先训练的 ResNet-18 网络来初始化图像骨干网络。对于输入处的稀疏深度图像的附加通道，我们将相应的权重设置为零。我们还在 TOR4D 数据集[34]上预先训练地面估算网络，具有离线地图作为标记，ℓ2 损失作为目标函数[33]。模型中的其他网络随机初始化。我们使用 Adam 优化器[11]训练具有随机梯度下降的模型。

4. Experiments

4. 实验

One of the fundamental tasks in autonomous driving is to perceive the scene in real-time. In this paper we propose a multi-task multi-sensor fusion model for the task of 3D object detection. We refer the reader to Figure 2 for an illustration of the model architecture. Our approach has the following highlights. First, we design a multi-sensor architecture that combines point-wise and ROI-wise feature fusion. Second, our integrated ground estimation module reasons about the geometry of the road. Third, we exploit the task of depth completion to learn better multi-sensor features and achieve dense point-wise feature fusion. As a result, the whole model can be learned end-to-end by exploiting a multi-task loss.

自动驾驶的基本任务之一是实时感知场景。在本文中，我们提出了一个多任务多传感器融合模型，用于三维物体检测任务。我们将读者引用到图 2 中以获得模型体系结构的图示。我们的方法有以下亮点。首先，我们设计了一种多传感器架构，它结合了逐点和逐 ROI 特征融合。其次，我们的综合地面估算模块有关道路几何形状的推理。第三，我们利用深度补全的任务来学习更好的多传感器特征并实现密集的逐点特征融合。因此，可以通过利用多任务损失来端到端地学习整个模型。



Detector	Input Data		Time	2D AP (%)			3D AP (%)			BEV AP (%)		
	LiDAR	IMG	(ms)	easy	mod.	hard	easy	mod.	hard	easy	mod.	hard
SHJU-HW [35, 7]		✓	850	90.81	90.08	79.98	-	-	-	-	-	-
RRC [20]		✓	3600	90.61	**90.23**	87.44	-	-	-	-	-	-
MV3D [5]	✓		240	89.80	79.76	78.61	66.77	52.73	51.31	85.82	77.00	68.94
VoxelNet [36]	✓		220	-	-	-	77.49	65.11	57.73	89.35	79.26	77.39
SECOND [32]	✓		50	90.40	88.40	80.21	83.13	73.66	66.20	88.07	79.37	77.95
PIXOR [34]	✓		**35**	-	-	-	-	-	-	87.25	81.92	76.01
PIXOR++ [33]	✓		**35**	-	-	-	-	-	-	89.38	83.70	77.97
HDNET [33]	✓		50	-	-	-	-	-	-	89.14	86.57	78.32
MV3D [5]	✓	✓	360	90.53	89.17	80.16	71.09	62.35	55.12	86.02	76.90	68.49
AVOD [12]	✓	✓	80	89.73	88.08	80.14	73.59	65.78	58.38	86.80	85.44	77.73
ContFuse [13]	✓	✓	60	-	-	-	82.54	66.22	64.04	88.81	85.83	77.33
F-PointNet [17]	✓	✓	170	90.78	90.00	80.80	81.20	70.39	62.19	88.70	84.00	75.33
AVOD-FPN [12]	✓	✓	100	89.99	87.44	80.05	81.94	71.88	66.38	88.53	83.79	77.90
Our MMF	✓	✓	80	**91.82**	90.17	**88.54**	**86.81**	**76.75**	**68.41**	**89.49**	**87.47**	**79.10**

Table 1. Evaluation results on the testing set of KITTI 2D, 3D and BEV object detection benchmark (car). We compare with previously published detectors on the leaderboard ranked by Average Precision (AP) in the moderate setting.

4.1 2D/3D/BEV Object Detection on KITTI

Dataset and metric: KITTI's object detection dataset has 7,481 frames for training and 7,518 frames for testing. We evaluate our approach on "Car" class. We apply the same data augmentation as [13] during training, which applies random translation, orientation and scaling on LiDAR point clouds and camera images. For multi-task training, we also leverage the dense depth labels from the KITTI's depth dataset [28]. KITTI's detection metric is defined as Average Precision (AP) averaged over 11 points on the Precision-Recall (PR) curve. The evaluation criterion for cars is 0.7 Intersection-Over-Union (IoU) in 2D, 3D or BEV. KITTI also divides labels into three subsets (easy, moderate and hard) according to the object size, occlusion and truncation levels, and ranks methods by AP in the moderate setting.

Implementation details: We detect objects within 70 meters forward and 40 meters to the left and right of the ego-car, as most of the labeled objects are within this region. We voxelize the cropped point cloud into a volume of size 512 × 448 × 32 as the BEV representation. We also center-crop the images of different sizes into a uniform size of 370 × 1224. We train the model on a 4 GPU machine with a total batch size of 16 frames. Online hard negative mining [27] is used during training. We set the initial learning rate to 0.001 and decay it by 0.1 after 30 and 45 epochs respectively. The training ends after 50 epochs. We train two models on KITTI: one without depth completion auxiliary task, and one full model. We submit the former one to the test server for fair comparison with other methods that are trained on KITTI object detection dataset only. We evaluate the full model in ablation study to showcase the performance gain brought by depth completion and dense fusion.

4.1 KITTI 上的 2D/3D/BEV 目标检测

数据集和度量：KITTI 的对象检测数据集有 7,481 个用于训练的帧和 7,518 个用于测试的帧。我们评估我们对“汽车”类别的表现。我们在训练期间应用与[13]相同的数据增强，其在 LiDAR 点云和相机图像上应用随机平移，定向和缩放。对于多任务训练，我们还利用 KITTI 深度数据集中的密集深度标签[28]。KITTI 的检测指标定义为准确率-召回率（PR）曲线上 11 个点的平均精度（AP）平均值。汽车的评估标准是在 2D，3D 或 BEV 中的 0.7 交叉联合（IoU）。KITTI 还根据对象大小，遮挡和截断级别将标签划分为三个子集（简单，中等和硬），并在适度设置中按 AP 对方法进行排名。

实施细节：我们检测向前 70 米和自我车左右 40 米的物体，因为大多数标记的物体都在这个区域内。我们将裁剪的点云体素化为体积为 512×448×32 的体积作为 BEV 表示。我们还将不同尺寸的图像中心裁剪成 370×1224 的均匀尺寸。我们在 4 GPU 机器上训练模型，总批量为 16 帧。在训练期间使用在线硬负挖掘[27]。我们将初始学习率设置为 0.001，并在 30 和 45 个时期之后将其衰减 0.1。培训在 50 个时期后结束。我们在 KITTI 上训练两个模型：一个没有深度补全辅助任务，一个完整模型。我们将前一个提交给测试服务器，以便与仅在 KITTI 对象检测数据集上训练的其他方法进行公平比较。我们评估消融研究中的完整模型，以展示深度补全和密集融合带来的性能增益。

Evaluation results: We compare our approach with previously published state-of-the-art detectors in Table 1, and show that our approach outperforms competitors by a large margin in all 2D, 3D and BEV detection tasks. In 2D detection, we surpass the best image detector RRC [20] by 1.1% AP in the hard setting, while being 45× faster. Note that we only use a small ResNet-18 network as the image stream backbone network, which shows that 2D detection benefits a lot from exploiting the LiDAR sensor and reasoning in 3D. In BEV detection, we outperform the best detector HDNET [33], which also exploits ground estimation, by 0.9% AP in moderate setting. The improvement mainly comes from multi-sensor fusion. In the most challenging 3D detection task (as it requires 0.7 3D IoU), we show an even larger gain over competitors. We surpass the best detector SECOND [32] by 3.09% AP in moderate setting, and outperform the previously best multi-sensor detector AVOD-FPN [12] by 4.87% AP in moderate setting. We believe the large gain mainly comes from the fully fused feature representation and the proposed ROI feature extraction for precise object localization.

评估结果：我们将我们的方法与表 1 中先前发布的最先进的探测器进行了比较，并表明我们的方法在所有 2D，3D 和 BEV 检测任务中都大大优于竞争对手。在 2D 检测中，我们在硬设置中超过 1.1％AP 的最佳图像检测器 RRC [20]，同时快 45 倍。请注意，我们仅使用小型 ResNet-18 网络作为图像流骨干网络，这表明 2D 检测在利用 LiDAR 传感器和 3D 推理方面受益匪浅。在 BEV 检测中，我们优于最佳探测器 HDNET [33]，它也利用地面估计，在中等设置下为 0.9％AP。改进主要来自多传感器融合。在最具挑战性的 3D 检测任务中（因为它需要 0.7 个 3D IoU），我们显示出比竞争对手更大的收益。我们在中等设置下超过最佳检测器 SECOND [32] 3.09％AP，并且在中等设置下优于先前最佳的多传感器检测器 AVOD-FPN [12] 4.87％AP。我们认为，大增益主要来自完全融合的特征表示和提出的 ROI 特征提取，用于精确的对象定位。

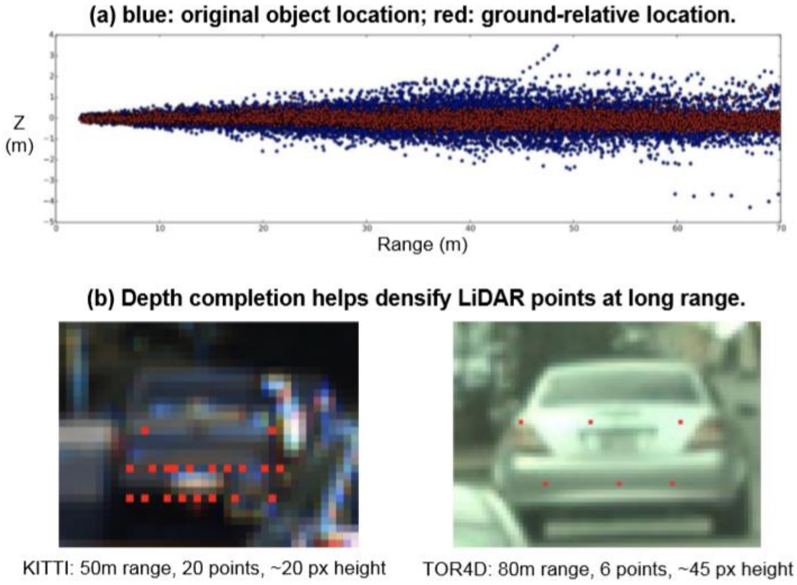
Ablation study: To analyze the effects of multi-sensor fusion and multi-task learning, we conduct an ablation study on KITTI training set. We use four-fold cross validation and accumulate the evaluation results over the whole training set. This produces stable evaluation results for apple-to-apple comparison. We show the ablation study results in Table 2. Our baseline model is a single-shot LiDAR only detector. Adding image stream with point-wise feature fusion brings over 5% AP gain in 3D detection, possibly because image features provide complementary information on the Z axis in addition to the BEV representation of LiDAR point cloud. Ground estimation improves 3D and BEV detection by 1.9% and 1.4% AP respectively in moderate setting. This gain suggests that the geometric ground prior provided by online mapping is very helpful for detection at long range (as shown in Figure 5 a), where we have very sparse 3D LiDAR observation. Adding the refinement module with ROI-wise feature fusion brings consistent improvements on all three tasks, which purely come from more precise localization. This proves the effectiveness of the proposed orientation aware ROI feature extraction. Lastly, the model further benefits in BEV detection from the depth completion task with better feature representations and dense fusion, which suggests that depth completion provides complementary information in BEV space. On KITTI we do not see much gain from dense point-wise fusion using estimated depth. We hypothesize that this

模型简化研究：为了分析多传感器融合和多任务学习的效果，我们对 KITTI 训练集进行了模型简化研究。我们使用四重交叉验证并在整个训练集上累积评估结果。这为逐个特征的比较产生了稳定的评估结果。我们在表 2 中显示了模型简化研究结果。我们的基线模型是单发 LiDAR 检测器。添加具有逐点特征融合的图像流在 3D 检测中带来超过 5％的 AP 增益，这可能是因为除了 LiDAR 点云的 BEV 表示之外，图像特征在 Z 轴上提供补充信息。在中等设置下，地面估计可分别提高 1.9％和 1.4％AP 的 3D 和 BEV 检测。这一增益表明，在线建图提供的几何地面先验对于远距离探测非常有用（如图 5a 所示），其中我们有非常稀疏的 3D LiDAR 观测。使用 ROI 智能特征融合添加改进模块可以对所有三个任务进行持续改进，这纯粹来自更精确的本地化。这证明了所提出的方向感知 ROI 特征提取的有效性。最后，该模型在深度完成任务的 BEV 检测中进一步受益，具有更好的特征表示和密集融合，这表明深度完成提供了 BEV 空间中的补充信息。在 KITTI，我们没有看到使用估计深度的密集点融合获得多少收益。我们假设这是因为在 KITTI 中捕获的图像在远距离处具有相当的 LiDAR 分辨率（如图 5b 所示）。因此，从图像特征中挤出的汁液并不多。然而，在我们拥有更高分辨率相机图像的 TOR4D 基准测试中，我们在下一节中展示深度完成不仅有助于多任务学习，还有助于密集的特征融合。

is because in KITTI the captured image is at equivalent resolution of LiDAR at long range (as shown in Figure 5 b). Therefore, there isn't much juice to squeeze from image feature. However, on TOR4D benchmark where we have higher resolution camera images, we show in next section that depth completion helps not only by multi-task learning, but also dense feature fusion.

Model	Multi-Sensor		Multi-Task			2D AP (%)			3D AP (%)			BEV AP (%)		
	pt	roi	map	dep	depf	easy	mod.	hard	easy	mod.	hard	easy	mod.	hard
LiDAR only						93.44	87.55	84.32	81.50	69.25	63.55	88.83	82.98	77.26
+image	✓					+2.95	+1.97	+2.76	+4.62	+5.21	+3.35	+0.70	+2.39	+1.25
+map	✓		✓			+3.06	+2.20	+3.33	+5.24	+7.14	+4.56	+0.36	+3.77	+1.59
+refine	✓	✓	✓			+3.94	**+2.71**	+4.66	**+6.43**	+8.62	+12.03	+7.00	+4.81	+2.12
+depth	✓	✓	✓	✓		**+4.69**	+2.65	+4.64	+6.34	**+8.64**	**+12.06**	+7.74	+5.16	+2.26
full model	✓	✓	✓	✓	✓	+4.61	+2.67	**+4.68**	+6.40	+8.61	+12.02	**+7.83**	**+5.27**	**+2.34**

Table 2. Ablation study on KITTI object detection benchmark (car) training set with four-fold cross validation. *pt*: point-wise feature fusion. *roi*: ROI-wise feature fusion. *map*: online mapping. *dep*: depth completion. *depf*: dense fusion with dense depth.


(a) blue: original object location; red: ground-relative location.

(b) Depth completion helps densify LiDAR points at long range.

KITTI: 50m range, 20 points, ~20 px height

TOR4D: 80m range, 6 points, ~45 px height

Figure 5. Object detection benefits from ground estimation and depth completion.

图5.来自地面估计和深度补全的物体检测效益。

4.2 BEV Object Detection on TOR4D

Dataset and metric: The TOR4D BEV object detection benchmark [34] contains over 5,000 video snippets with a duration of around 25 seconds each. To generate the training and testing dataset, we sample video snippets from different scenes at 1 Hz and 0.5 Hz respectively, leading to around 100,000 training frames and around 6,000 testing frames. To validate the effectiveness of depth completion in improving object detection, we use images captured by camera with long-focus lens which provide richer information at long range (as shown in Figure 5 b). We evaluate on multi-class BEV object detection (i.e. vehicle, pedestrian and bicyclist) with a distance range of 100 meters from the ego-car. We use AP at different IoU thresholds as the metric for multi-class object detection. Specifically, we look at 0.5 and 0.7 IoUs for vehicles, 0.3 and 0.5 IoUs for the pedestrians and

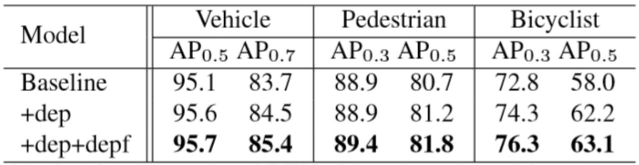
4.2 TOR4D 鸟瞰物体检测

数据集和度量：TOR4D BEV 物体检测基准[34]包含5,000 多个视频片段，每个片段的持续时间约为 25 秒。为了生成训练和测试数据集，我们分别以 1 Hz 和 0.5 Hz 的频率对来自不同场景的视频片段进行采样，从而产生大约 100,000 个训练帧和大约 6,000 个测试帧。为了验证深度完成在改善物体检测方面的有效性，我们使用具有长焦距镜头的相机拍摄的图像，其在远距离提供更丰富的信息（如图 5b 所示）。我们评估距离自我车 100 米的距离范围的多级 BEV 物体检测（即车辆，行人和骑车人）。我们使用不同 IoU 阈值的 AP 作为多类对象检测的度量。具体来说，我们看车辆为 0.5 和 0.7 IoU，行人和骑车者为 0.3 和 0.5 IoU。

cyclists.

Evaluation results: We re-produce the previously state-of-the-art detector ContFuse [13] on our TOR4D dataset split. Two modifications are made to further improve the detection performance. First, we follow FAF [16] to fuse multi-sweep LiDAR point clouds together at the BEV representation. Second, following HDNET [33] we incorporate semantic and geometric HD map priors to the detector. We use the improved ContFuse detector as the baseline, and apply the proposed depth completion with dense fusion on top of it. As shown in Table 3, the depth completion task helps in two ways: multi-task learning and dense feature fusion. The former increases the bicyclist AP by an absolute gain of 4.2%. Since bicyclists have the fewest number of labels in the dataset, having additional multi-task supervision is particularly helpful. In terms of dense fusion with estimated depth, the performance on vehicles is improved by over 5% in terms of relative error reduction (i.e. 1-AP) at 0.7 IoU. The reason for this gain may be that vehicles receive more additional feature fusion compared to the other two classes.

评估结果：我们在 TOR4D 数据集拆分中重新生成了先前最先进的检测器 ContFuse [13]。进行了两种修改以进一步提高检测性能。首先，我们遵循 FAF [16]在 BEV 表示中将多扫描 LiDAR 点云融合在一起。其次，在 HDNET 之后[33]，我们将语义和几何高清地图先验结合到探测器中。我们使用改进的 ContFuse 探测器作为基线，并在其顶部应用建议的深度完井和密集融合。如表 3 所示，深度完成任务有两种方式：多任务学习和密集特征融合。前者使自行车 AP 增加了 4.2%的绝对增益。由于自行车骑手在数据集中具有最少数量的标签，因此具有额外的多任务监督特别有用。就具有估计深度的密集融合而言，在 0.7IoU 的相对误差减小（即 1-AP）方面，车辆的性能提高了超过 5%。这种增益的原因可能是与其他两个类相比，车辆接收到更多额外的特征融合。



Model	Vehicle		Pedestrian		Bicyclist	
	$AP_{0.5}$	$AP_{0.7}$	$AP_{0.3}$	$AP_{0.5}$	$AP_{0.3}$	$AP_{0.5}$
Baseline	95.1	83.7	88.9	80.7	72.8	58.0
+dep	95.6	84.5	88.9	81.2	74.3	62.2
+dep+depf	**95.7**	**85.4**	**89.4**	**81.8**	**76.3**	**63.1**

Table 3. Ablation study of BEV object detection with multi-task learning on TOR4D benchmark. The baseline detector is based on [13], with multi-sweep LiDAR and HD maps added to the input for better performance. *dep*: depth completion. *depf*: dense fusion using estimated dense depth.

4.3 Qualitative Results and Discussion

We show qualitative 3D object detection results of the proposed detector on KITTI benchmark in Figure 6. The proposed detector is able to produce high-quality 3D detections of vehicles that are highly occluded or far away from the ego-car. Some of our detections are un-labeled cars in KITTI dataset. Previous works [5, 12] often follow state-of-the-art 2D detection framework (like two-stage Faster RCNN [22]) to solve 3D detection. However, we argue that it may not be the optimal solution. With thousands of pre-defined anchors, the feature extraction is both slow and inaccurate. Instead we show that by detecting 3D objects in BEV space, we can produce high-quality 3D detections via a single pass of the network (as shown in Table 2 by model variants without refinement), given that we fully fuse the multi-sensor feature maps via dense fusion.

4.3 定性结果和讨论

我们在图 6 中的 KITTI 基准测试中显示了所提出的探测器的定性 3D 物体探测结果。所提出的探测器能够产生高度遮挡或远离自主车的高质量车辆 3D 检测。我们的一些检测是 KITTI 数据集中未标记的汽车。以前的作品[5,12]经常遵循最先进的 2D 检测框架（如两阶段 Faster RCNN [22]）来解决 3D 检测问题。但是，我们认为它可能不是最佳解决方案。有数千个预先定义的锚点，特征提取既缓慢又不准确。相反，我们通过检测 BEV 空间中的 3D 对象，我们可以通过网络的单次传递产生高质量的 3D 检测（如表 2 所示，模型变体无需改进），因为我们完全融合了多传感器功能通过密集融合的地图。

Cascade approaches [17, 6] assume that 2D detection is solved better than 3D detection, and therefore use 2D

级联方法[17,6]假设 2D 检测比 3D 检测更好地解决，因此使用 2D 检测器来生成 3D 提议。但是，我们认为 3D detector to generate 3D proposals. However, we argue that 3D detection is actually easier than 2D. As we detect objects in 3D metric space, we do not have to handle the problems of scale variance and occlusion reasoning that would otherwise arise in 2D. Our model, which uses a pre-trained ResNet-18 as image backbone network and is trained from thousands of frames, surpasses F-PointNet [17], which exploits two orders of magnitude more training data (i.e. COCO dataset [15]), by over 7% AP in hard setting of KITTI 2D detection. Multi-sensor fusion and multi-task learning are highly interleaved. In this paper we provide a way to combine them together under the same hood. In the proposed framework, multi-sensor fusion helps learn better feature representations to solve multiple tasks, while different tasks in turn provide different types of clues to make feature fusion deeper and richer.

检测实际上比 2D 更容易。当我们在 3D 度量空间中检测对象时，我们不必处理在 2D 中出现的比例差异和遮挡推理的问题。我们的模型使用经过预先训练的 ResNet-18 作为图像骨干网络并经过数千帧训练，超越了 F-PointNet [17]，它利用了两个数量级的训练数据（即 COCO 数据集[15]），在 KITTI 2D 检测的硬设置中，超过 7%的 AP。多传感器融合和多任务学习是高度交错的。在本文中，我们提供了一种在同一个引擎盖下将它们组合在一起的方法。在所提出的框架中，多传感器融合有助于学习更好的特征表示来解决多个任务，而不同的任务又提供不同类型的线索，使特征融合更深入，更丰富。

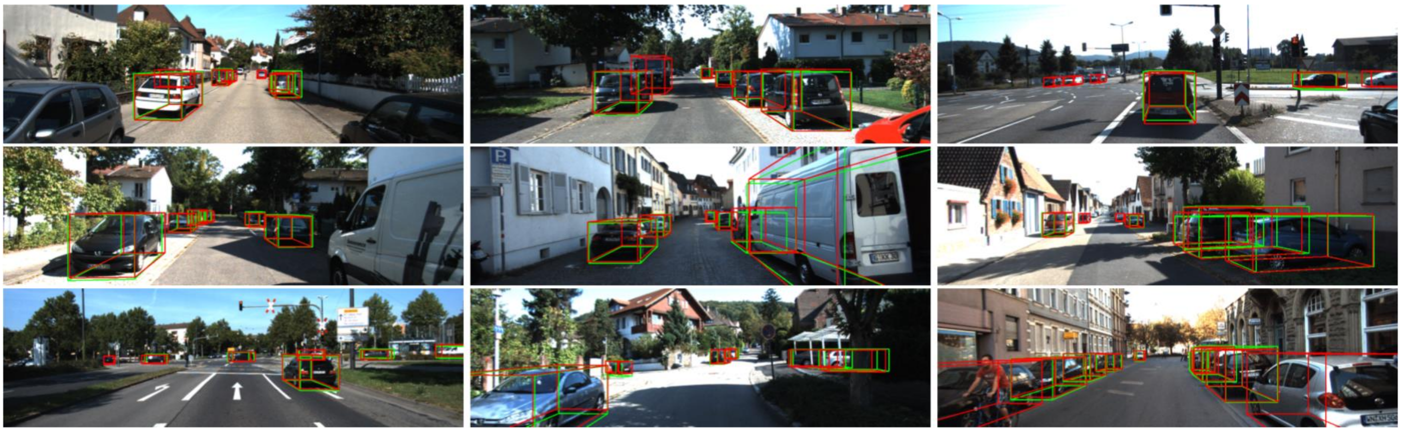


Figure 6. Qualitative results of 3D object detection (car) on KITTI benchmark. We draw object labels in green and our detections in red.

图 6. KITTI 基准测试中 3D 物体检测（汽车）的定性结果。我们用绿色绘制对象标签，用红色绘制检测。

5. Conclusion

We have proposed a multi-task multi-sensor detection model that jointly reasons about 2D and 3D object detection, ground estimation and depth completion. Point-wise and ROI-wise feature fusion are applied to achieve full multi-sensor fusion, while multi-task learning provides additional map prior and geometric clues enabling better representation learning and denser feature fusion. We validate the proposed method on KITTI and TOR4D benchmarks, and surpass the state-of-the-art methods in all detection tasks by a large margin. In the future, we plan to expand our multi-sensor fusion approach to exploit other sensors such as radar as well as temporal information.

5. 结论

我们提出了一种多任务多传感器检测模型，该模型共同推导了 2D 和 3D 物体检测，地面估计和深度补全。逐点和逐 ROI 特征融合应用于实现多传感器完全融合，而多任务学习提供额外的地图先验和几何线索，从而实现更好的表示学习和更密集的特征融合。我们在 KITTI 和 TOR4D 基准测试中验证了所提出的方法，并在所有检测任务中大大超越了最先进的方法。在未来，我们计划扩展我们的多传感器融合方法，以利用其他传感器，如雷达和时间信息。

6. References

[1] Beltrán J, Guindel C, Moreno F M, et al. Birdnet: a 3D object detection framework from LiDAR information[C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 3517-3523.

[2] Chen X, Kundu K, Zhang Z, et al. Monocular 3d object detection for autonomous driving[C]//Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2147-2156.

[3] Chen X, Kundu K, Zhu Y, et al. 3d object proposals for accurate object class detection[C]//Advances in Neural Information Processing Systems (NIPS). 2015: 424-432.

[4] Chen X, Kundu K, Zhu Y, et al. 3d object proposals using stereo imagery for accurate object class detection[J]. IEEE transactions on pattern analysis and machine intelligence (PAMI), 2017, 40(5): 1259-1272.

[5] Chen X, Ma H, Wan J, et al. Multi-view 3d object detection network for autonomous driving[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 1907-1915.

[6] Du X, Ang M H, Karaman S, et al. A general pipeline for 3d detection of vehicles[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 3194-3200.

[7] Fang L, Zhao X, Zhang S. Small-objectness sensitive detection based on shifted single shot detector[J]. Multimedia Tools and Applications, 2018: 1-19.

[8] 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 (CVPR). IEEE, 2012: 3354-3361.

[9] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision (ICCV). 2017: 2961-2969.

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2016: 770-778.

[11] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[12] 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). IEEE, 2018: 1-8.

[13] Liang M, Yang B, Wang S, et al. Deep continuous fusion for multi-sensor 3d object detection[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 641-656.

[14] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2117-2125.

[15] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision (ECCV). Springer, Cham, 2014: 740-755.

[16] Luo W, Yang B, Urtasun R. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 3569-3577.

[17] Qi C R, Liu W, Wu C, et al. Frustum pointnets for 3d object detection from rgb-d data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 918-927.

[18] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 652-660.

[19] Qi C R, Yi L, Su H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[C]//Advances in Neural Information Processing Systems (NIPS). 2017: 5099-5108.

[20] Ren J, Chen X, Liu J, et al. Accurate single stage detector using recurrent rolling convolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 5420-5428.

[21] Ren M, Pokrovsky A, Yang B, et al. Sbnet: Sparse blocks network for fast inference[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 8711-8720.

[22] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems (NIPS). 2015: 91-99.

[23] Ren Z, Sudderth E B. Three-dimensional object detection and layout prediction using clouds of oriented gradients[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 1525-1533.

[24] Ren Z, Sudderth E B. 3d object detection with latent support surfaces[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 937-946.

[25] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, Cham, 2015: 234-241.

[26] Schwing A G, Fidler S, Pollefeys M, et al. Box in the box: Joint 3d layout and object reasoning from single

images[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2013: 353-360.

[27] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 761-769.

[28] Uhrig J, Schneider N, Schneider L, et al. Sparsity invariant cnns[C]//2017 International Conference on 3D Vision (3DV). IEEE, 2017: 11-20.

[29] Wang S, Fidler S, Urtasun R. Holistic 3d scene understanding from a single geo-tagged image[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 3964-3972.

[30] Wang S, Suo S, Ma W C, et al. Deep parametric continuous convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 2589-2597.

[31] Xu B, Chen Z. Multi-level fusion based 3d object detection from monocular images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 2345-2353.

[32] Yan Y, Mao Y, Li B. Second: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[33] Yang B, Liang M, Urtasun R. Hdnet: Exploiting hd maps for 3d object detection[C]//Conference on Robot Learning(CoRL). 2018: 146-155.

[34] Yang B, Luo W, Urtasun R. Pixor: Real-time 3d object detection from point clouds[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 7652-7660.

[35] Zhang S, Zhao X, Fang L, et al. Led: Localization-Quality Estimation Embedded Detector[C]//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018: 584-588.

[36] Zhou Y, Tuzel O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 4490-4499.