3D Object Proposals using Stereo Imagery for Accurate Object Class Detection

使用立体图像生成 3D 目标提案用于精确目标类别检测

论文引用：

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, 2017:1-1.

ABSTRACT

The goal of this paper is to perform 3D object detection in the context of autonomous driving. Our method aims at generating a set of high-quality 3D object proposals by exploiting stereo imagery. We formulate the problem as minimizing an energy function that encodes object size priors, placement of objects on the ground plane as well as several depth informed features that reason about free space, point cloud densities and distance to the ground. We then exploit a CNN on top of these proposals to perform object detection. In particular, we employ a convolutional neural net (CNN) that exploits context and depth information to jointly regress to 3D bounding box coordinates and object pose. Our experiments show significant performance gains over existing RGB and RGB-D object proposal methods on the challenging KITTI benchmark. When combined with the CNN, our approach outperforms all existing results in object detection and orientation estimation tasks for all three KITTI object classes. Furthermore, we experiment also with the setting where LIDAR information is available, and show that using both LIDAR and stereo leads to the best result.

摘要

本文的目标是在自动驾驶的环境中执行 3D 物体检测。我们的方法旨在通过利用立体图像生成一组高质量的 3D 对象提案。我们将问题表述为最小化能量函数，该函数编码物体尺寸先验，物体在地平面上的放置以及几个已知深度特征，这些特征推理自由空间，点云密度和到地面的距离。然后，我们在这些提案之上利用 CNN 来执行对象检测。特别地，我们使用卷积神经网络（CNN），其利用上下文和深度信息来共同回归到 3D 边界框坐标和对象位姿。我们的实验显示，在具有挑战性的 KITTI 基准测试中，现有的 RGB 和 RGB-D 对象建议方法的性能显着提升。当与 CNN 结合使用时，我们的方法优于所有三个 KITTI 对象类的对象检测和方向估计任务中的所有现有结果。此外，我们还尝试了 LIDAR 信息可用的设置，并显示同时使用 LIDAR 和立体图像可获得最佳效果。

1. Introduction

AUTONOMOUS driving is receiving a lot of attention from both industry and the research community. Most self-driving cars build their perception systems on expensive sensors, such as LIDAR, radar and high-precision GPS. Cameras are an appealing alternative as they provide richer sensing at a much lower cost. This paper aims at high performance 2D and 3D object detection in the context of autonomous driving by exploiting stereo imagery.

With impressive advances in deep learning in the past few years, recent efforts in object detection exploit object proposals to facilitate classifiers with powerful, hierarchical visual representation [1], [2]. Compared with traditional sliding window based methods [3], the pipeline of generating object proposals that are combined with convolutional neural networks has lead to more than 20% absolute performance gains [4], [5] on the PASCAL VOC dataset [6].

1. 引文

自主驾驶正受到业界和研究界的广泛关注。大多数自动驾驶汽车都在昂贵的传感器上构建感知系统，例如激光雷达，雷达和高精度 GPS。相机是一种吸引人的选择，因为它们以更低的成本提供更丰富的感应。本文旨在通过利用立体图像在自动驾驶的背景下进行高性能的 2D 和 3D 物体检测。

随着过去几年在深度学习方面取得的令人印象深刻的进步，最近在对象检测方面的努力利用对象提议来促进具有强大的分层视觉表示的分类器[1]，[2]。与传统的基于滑动窗口的方法相比[3]，与卷积神经网络相结合的生成对象提议的流水线导致 PASCAL VOC 数据集[6]的绝对性能增益超过 20％[4]，[5]。

Object proposal methods aim at generating a moderate number of candidate regions that cover most of the ground truth objects in the image. One typical approach is to perform region grouping based on superpixels using a variety of similarity measures [7], [8]. Low-level cues such as color contrast, saliency [9], gradient [10] and contour information [11] have also been exploited in order to select promising object boxes from densely sampled windows. There has also been some recent work on learning to generate a diverse set of region candidates with ensembles of binary segmentation models [12], parametric energies [13] or CNN-based cascaded classifiers [14].

对象提议方法旨在生成覆盖图像中的大多数实际真值对象的中等数量的候选区域。一种典型的方法是使用各种相似性度量来执行基于超像素的区域分组[7]，[8]。还利用了诸如颜色对比度，显着性[9]，梯度[10]和轮廓信息[11]之类的低级线索，以便从密集采样的窗口中选择有希望的对象框。最近还有一些关于学习用二元分割模型[12]，参数能量[13]或基于 CNN 的级联分类器[14]的集合生成多样化区域候选者的工作。

The object proposal methods have proven effective on the PASCAL VOC benchmark. However, they have very low achievable recall on the autonomous driving benchmark KITTI [15], which presents the bottleneck for the state-of-the-art object detector R-CNN [4], [16] on this benchmark. On one hand, the PASCAL VOC dataset uses a loose overlap criteria for localization measure, i.e., a predicted box is considered to be correct if its overlap with the ground-truth box exceeds 50%. For self-driving cars, however, object detection requires a stricter overlap criteria to enable correct estimates of the distance of vehicles from the ego-car. Moreover, objects in KITTI images are typically small and many of them are heavily occluded or truncated. These challenging conditions limit the performance of most existing bottom-up proposals that rely on intensity and texture for superpixel merging and window scoring.

对象提案方法已证明对 PASCAL VOC 基准有效。然而，他们对自动驾驶基准 KITTI [15]的可实现召回率非常低，这为该基准测试提供了最先进的物体探测器 R-CNN [4]，[16]的瓶颈。一方面，PASCAL VOC 数据集使用松散重叠标准进行定位测量，即，如果预测框与真值框的重叠超过 50％，则认为该框是正确的。然而，对于自动驾驶汽车，物体检测需要更严格的重叠标准，以便能够正确估计车辆距离自驾车的距离。此外，KITTI 图像中的对象通常很小，并且其中许多被严重遮挡或截断。这些具有挑战性的条件限制了大多数现有的自下而上建议的性能，这些提案依赖于强度和纹理来进行超像素合并和窗口评分。

In this paper, we propose a novel 3D object detection approach that exploits stereo imagery and contextual information specific to the domain of autonomous driving. We propose a 3D object proposal method that goes beyond 2D bounding boxes and is capable of generating high quality 3D bounding box proposals. We make use of the 3D information estimated from a stereo camera pair by placing 3D candidate boxes on the ground plane and scoring them via 3D point cloud features. In particular, our scoring function encodes several depth informed features such as point densities inside a candidate box, free space, visibility, as well as object size priors and height above the ground plane. The inference process is very efficient as all the features can be computed in constant time via 3D integral images. Learning can be done using structured SVM [17] to obtain class-specific weights for these features. We also present a 3D object detection neural network that takes 3D object proposals as input and predict accurate 3D bounding boxes. The neural net exploits contextual information and

在本文中，我们提出了一种新颖的 3D 物体检测方法，该方法利用立体图像和上下文信息特定于自动驾驶的领域。我们提出了一种 3D 对象提议方法，该方法超越了 2D 边界框，并且能够生成高质量的 3D 边界框提议。我们利用从立体相机对估计的 3D 信息，将 3D 候选框放置在地平面上，并通过 3D 点云特征对其进行评分。特别是，我们的评分功能可以编码多个深度信息特征，例如候选框内的点密度，自由空间，可见性，以及物体尺寸先验和地平面以上的高度。推理过程非常有效，因为所有特征都可以通过 3D 积分图像在恒定时间内计算。学习可以使用结构化 SVM [17]来获得这些特征的类特定权重。我们还提出了一个 3D 对象检测神经网络，它将 3D 对象提案作为输入并预测精确的 3D 边界框。神经网络利用上下文信息并使用多任务损失来共同回归到边界框坐标和对象方向。

uses a multi-task loss to jointly regress to bounding box coordinates and object orientation.

We evaluate our approach on the challenging KITTI detection benchmark[15]. Extensive experiments show that: 1) The proposed 3D object proposals achieve significantly higher recall than the state-of-the-art across all overlap thresholds under various occlusion and truncation levels. In particular, compared with the state-of-the-art RGB-D method MCG-D [18], we obtain 25% higher recall with 2K proposals. 2) Our 3D object detection network combined with 3D object proposals outperforms all published results on object detection and orientation estimation for Car, Cyclist and Pedestrian. 3) Our approach is capable of producing accurate 3D bounding box detections, which allows us to locate objects in 3D and infer the distance and pose of objects from the ego-car. 4) We also apply our approach to LIDAR point clouds with more precise, but sparser, depth estimation. When combining stereo and LIDAR data, we obtain the highest 3D object detection accuracy.

我们评估了我们对具有挑战性的 KITTI 检测基准的方法[15]。大量实验表明：1）在各种遮挡和截断水平下，所提出的 3D 对象建议实现了比所有重叠阈值的现有技术更高的召回率。特别是，与最先进的 RGB-D 方法 MCG-D [18]相比，我们通过 2K 提案获得了高出 25％的召回率。2）我们的 3D 物体检测网络结合 3D 对象建议优于所有公布的汽车，自行车和行人的物体检测和方向估计结果。3）我们的方法能够生成精确的 3D 边界框检测，这使我们能够在 3D 中定位对象并推断出自我车的物体的距离和姿势。4）我们还将我们的方法应用于 LIDAR 点云，其具有更精确但更稀疏的深度估计。当组合立体图像和 LIDAR 数据时，我们获得最高的 3D 物体检测精度。

A preliminary version of this work was presented in[19]. In this manuscript, we make extensions in the following aspects: 1) A more detailed description of the inference process of proposal generation. 2) The 3D object proposal model is extended with a class-independent variant. 3) The detection neural network is extended to a two-stream network to leverage both appearance and depth features. 4) We further apply our model to point clouds obtained via LIDAR, and provide comparison of the stereo, LIDAR and the hybrid settings. 5) We extensively evaluate the 3D bounding box recall and 3D object detection performance. 6) Our manuscript includes ablation studies of network design, depth features, as well as ground plane estimation.

这项工作的初步版本在[19]中提出。在本手稿中，我们在以下方面进行了扩展：1）更详细地描述了提案生成的推理过程。2）3D 对象提案模型使用与类无关的变体进行扩展。3）检测神经网络扩展到双流网络，以利用外观和深度特征。4）我们进一步将我们的模型应用于通过 LIDAR 获得的点云，并提供立体图像，激光雷达和混合设置的比较。5）我们广泛评估 3D 边界框召回和 3D 物体检测性能。6）我们的手稿包括网络设计，深度特征以及地平面估计的消融研究。

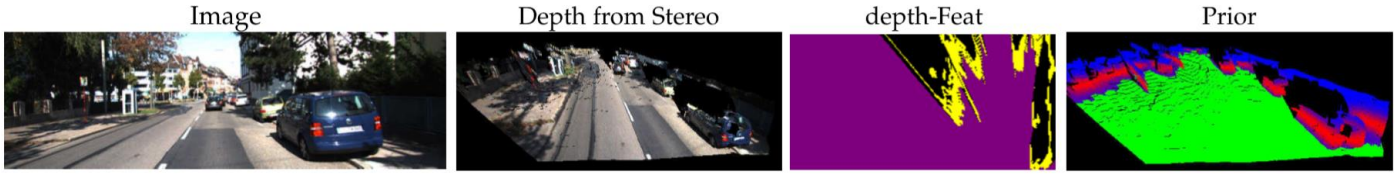

Figure 1. ***Features in our model (from left to right):*** *left camera image, stereo 3D reconstruction, depth-based features and our prior. In the third image, occupancy is marked with yellow (P in Eq. (1)) and purple denotes free space (F in Eq. (2)). In the prior, the ground plane is green and blue to red indicates increasing prior value of object height.*

图 1. 我们的模型中的功能（从左到右）：左相机图像，立体 3D 重建，基于深度的功能和我们的先验。在第三幅图像中，占用率用黄色标记(P in Eq. (1))，紫色表示自由空间(F in Eq. (2))。在先验中，地平面是绿色，蓝色到红色表示物体高度的先验值增加。