

Complexer-YOLO: Real-Time 3D Object Detection and Tracking on Semantic Point Clouds

Martin Simon, Karl Amende, Andrea Kraus, Jens Honer,
Timo Sämann, Hauke Kaulbersch and Stefan Milz
Valeo Schalter und Sensoren GmbH
Ffi rstname. l astnameG@val eo. com

Horst Michael Gross
Ilmenau University of Technology
horst-mi chael . gross@tu-i lmenau. de

Abstract

Accurate detection of 3D objects is a fundamental problem in computer vision and has an enormous impact on autonomous cars, augmented/virtual reality and many applications in robotics. In this work we present a novel fusion of neural network based state-of-the-art 3D detector and visual semantic segmentation in the context of autonomous driving. Additionally, we introduce Scale-Rotation-Translation score (SRTs), a fast and highly parameterizable evaluation metric for comparison of object detections, which speeds up our inference time up to 20% and halves training time. On top, we apply state-of-the-art online multi target feature tracking on the object measurements to further increase accuracy and robustness utilizing temporal information. Our experiments on KITTI show that we achieve same results as state-of-the-art in all related categories, while maintaining the performance and accuracy trade-off and still run in real-time. Furthermore, our model is the first one that fuses visual semantic with 3D object detection.

1. Introduction

Over the last few years self-driving cars got more and more into the focus of the automotive industry as well as new mobility players. Today, commercial vehicles already offer manifold automation like assisted or automated parking, adaptive cruise control and even highway pilots. To reach the full level of automation, they require a very precise system of environmental perception, working for every conceivable scenario. Additionally, real world scenarios strictly require real-time performance.

Recent vehicles are equipped with multiple different kind of sensors like ultrasonics, radar, cameras and Lidar (light detection and ranging) as well. With the help of redundancy and sensor fusion, relevant reliability and safety can be achieved. These circumstances significantly boosted

the rapid development of sensor technology and the growth of artificial intelligence algorithms for fundamental tasks like object detection and semantic segmentation.

Many modern approaches for these tasks use camera, Lidar or combine both. Compared to camera images, there are some difficulties dealing with Lidar point cloud data. Such point clouds are unordered, sparse and have a highly varying density due to the non-uniform sampling of the 3D space, occlusion and reflection. On the other hand, they offer way higher spatial accuracy and reliable depth information. Therefore, Lidar is more common in the context of autonomous driving. In this paper, we propose Complexer-YOLO, a real-time 3D object detection and tracking on semantic point clouds (see Fig. 1, 2). The main contributions are:

- **Visual Class Features:** Incorporation of visual point-wise Class-Features generated by fast camera-based Semantic Segmentation [39].
- **Voxelized Input:** Extension of Complex-YOLO [42] processing voxelized input features with a variable depth of dimension instead of fixed RGB-maps.
- **Real 3D prediction:** Extension of the regression network to predict 3D box heights and z-offsets to treat targets in three dimensions.
- **Scale-Rotation-Translation score (SRTs):** We introduce SRTs, a new validation metric for 3D boxes, notably faster than intersection over union (IoU), considering the 3DoF pose of the detected object including the yaw angle such as width, height and length.
- **Multitarget-Tracking:** Application of an Online feature tracker decoupled from the detection network, enabling time depending tracking and target instantiation based on realistic, physical assumptions.
- **Realtime capability:** We present a complete novel tracking pipeline with an outstanding overall real-time

