CubeDETR: End-to-End 3D Object Detection with Transformers

02501 Advanced Deep Learning in Computer Vision

Dimitrios Papadopoulos
Associate Professor, DTU Compute

12 March, 2024

Keywords/lectures: Transformers, Object Detection

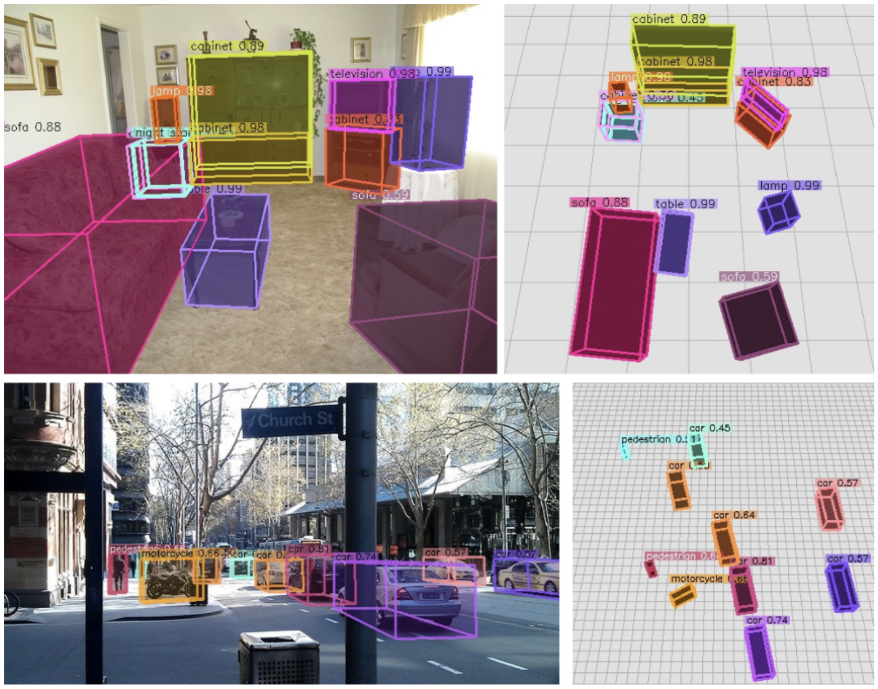

Figure 1: **3D Object Detection task.**

1 Project description

Understanding objects and their properties from single images is a longstanding problem in computer vision with applications in robotics and AR/VR. In the last decade, 2D object recognition [6, 7, 9, 4, 10] has tremendously advanced in predicting objects on the image plane with the help of large datasets [5, 8]. However, the world and its objects are three-dimensional laid out in 3D space. Perceiving objects in 3D from 2D visual inputs poses new challenges framed by the task of 3D object detection. Here, the goal is to estimate a 3D location and 3D extent of each object in an image in the form of a tight-oriented 3D bounding box.

Cube R-CNN [1] is a general and simple 3D object detector, inspired by advances in 2D and 3D recognition in recent years. Cube R-CNN extends the Faster R-CNN model and detects all objects and their 3D location, size, and rotation end-to-end from a single image of any domain and for many object categories. Cube R-CNN shows strong generalization and outperforms prior works for indoor and urban domains with one unified model.

The goal of this project is to build a Cube DETR model by combining the ideas from the Cube R-CNN model with the successful DETR model [2] used in 2D object detection.

2 Data

In this project, you can use either the SUN RGB-D [11] or the KITTI dataset [3], as standard for the 3D object detection task. More options for datasets for 3D object detection: ARKitScenes, Objectron, Hypersim, nuScenes, Omni3D.

3 Tasks

In this project, you could work on the following tasks:
Task 1: Train and evaluate the Cube R-CNN model.
Task 2: Build a CubeDETR model. Start from the implementation of DETR and modify it to work for 3D object detection (extra cube head).
Task 3: Evaluate your model and visualize the outputs.
Task 4: ...
Task 5: ...

References

[1] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164, 2023.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. 2012.

[4] R. Girshick. Fast R-CNN. 2015.

[5] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. 2019.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. pages 2980–2988, 2017.

[7] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[9] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. 2016.

[10] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. 2015.

[11] S. Song, S. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. 2015.