# Lecture 22, Nov 25, 2025

## 3D Object Detection

- 3D object detection aims to identify the 3D position, orientation, and bounding box extent for every relevant object, using LIDAR and RGB measurements
- To quantify detection performance for all thresholds, the average precision (AP) metric can be used, which measures the area under the ROC curve
- The AVOD-FPN architecture uses proposal generation from an image feature map and a birds-eye-view LIDAR feature map, trained on the KITTI dataset
  - Proposals were anchored to the ground to significantly reduce the search area
  - Achieved 10fps at the top of the KIITI benchmark
- Training both the vision and LIDAR detectors were difficult; one method of fusing the visual and LIDAR information was to first use the 2D object detector to get masks, which are projected into the pointcloud and used to paint points with bounding box or class information
- Detection accuracy drops off sharply with distance in the LIDAR pointcloud as it becomes very sparse, so fusing 2D and 3D information can achieve better results
  - This can be done in different stages of the network, early stage (e.g. pointcloud painting), intermediate stage (e.g. fusion of feature maps) or late stage (e.g. running detections on both independently and fusing in the end)
  - Dense Voxel Fusion is an early fusion approach that uses visual detections to colour voxels in the scene to get denser data
- 3D object detection is also possible using only monocular input, where the network tries to estimate depth
  - One approach is to try to reconstruct the object using the depth prediction and comparing this with 3D models of objects constructed using LIDAR depth completion in the training data
  - Instead of generating a single depth, uncertainty can be incorporated by generating a depth distribution for each pixel (i.e. a histogram)
- Stereo matching can be enhanced with object detection, by performing instance segmentation in the two images and associating them, and performing stereo matching on the individual instances
  - Similarly this can be used to reconstruct objects and trained against completed 3D LIDAR data
  - Learned stereo depth maps tend to be smooth and have points halfway in between the background and foreground, since this results in smaller loss than having sharp separations that might be slightly wrong
- How can we get probabilistic uncertainty estimates (and not just confidences) from 3D object detection, which would allow us to fuse it properly with other sensors?
  - Instead of non-maximum suppression, we can use clustering and Bayesian fusion and treating each detection as a measurement
  - The network also regresses a variance
  - The idea is to use dropout to get different predictions, and using the difference between predictions as an estimate of uncertainty (or difficulty of estimating an object), and fusing these to get a distribution
  - Other approaches either directly generate uncertainty or output a set of samples directly
- *Domain adaptation* involves transferring knowledge from a labelled source dataset to an unlabelled target dataset, e.g. making a network trained mainly on clear weather handle adverse weather conditions
  - We can create pseudo labels generated using the teacher network
  - The student network is taught to ignore variations between the datasets
  - We can also use foundation models such as DINOv2