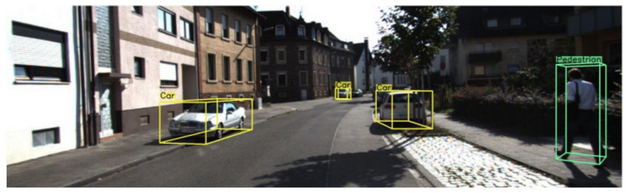
MONOCULAR 3D OBJECT DETECTION

Esteban Zamora (s222383), Jonathan Mikler (s222962), Lucas M. Sandby (s223093), Rolando Esquivel-Sancho (s230025)
DTU Autonomous Systems, Technical University of Denmark

BACKGROUND

3D object detection is crucial for having situational awareness in AVs. Traditionally, stereo vision and LiDAR have been used, but monocular cameras offer a simpler, more cost-effective and easier to integrate alternative. Hence, work in 3D object recognition from this source is relevant.



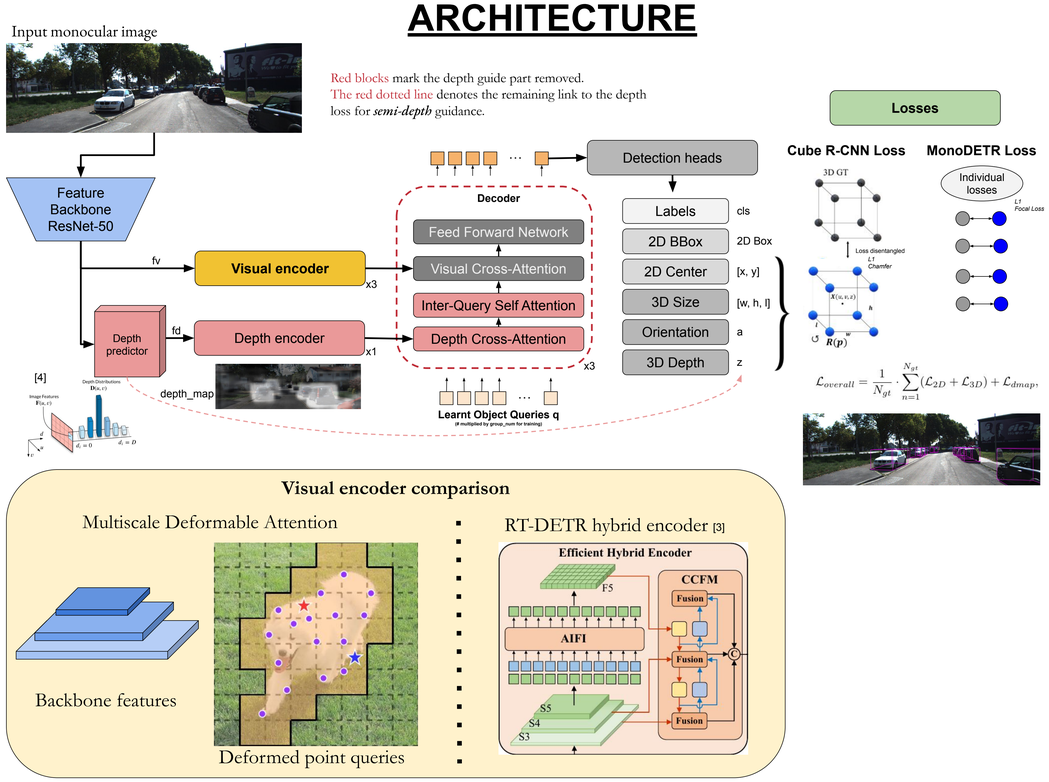
This works investigates modification to the MonoDETR architecture, a model aimed to make monocular camera 3D object detections.
The approaches taken:

- Modifying the information channels in the original model
- Replacing certain parts of the model to assess its impact on inference time, model size and performance (depth guidance, RT-DETR)
- Implementing a different set of losses leveraging on the idea of *disentangled loss.* (Cube R-CNN)

Main contributions:

- Investigation of backbone and depth guidance impact.
- Integration of RT-DETR hybrid visual encoder for improved speed.
- Excluding depth guidance (semi and fully disconnected)
- Training optimization with Cube R-CNN 3D loss
- Inference evaluation with integrated DeepSpeed Profiler.

ARCHITECTURE



TRAINING

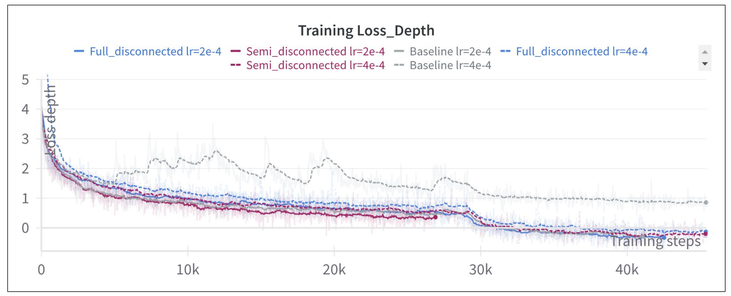
Dataset: ~7500 KITTI rgb images (375x1242)

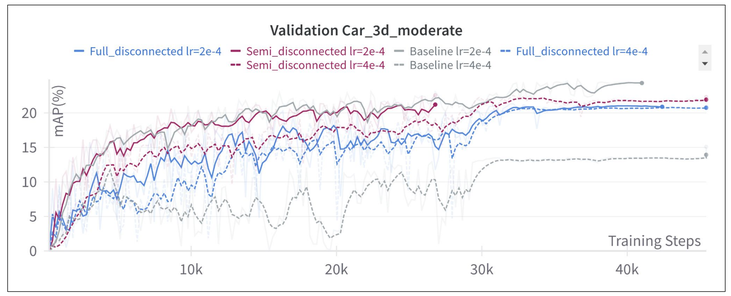
- Training: 49.6%, Validation: 50.4%

Infrastructure:

- DTU's HPC + Weights & Biases + Hydra

Batch Size	16	Epochs	195
Learning Rates	2e-4 4e-4	Weight Decay	1e-4



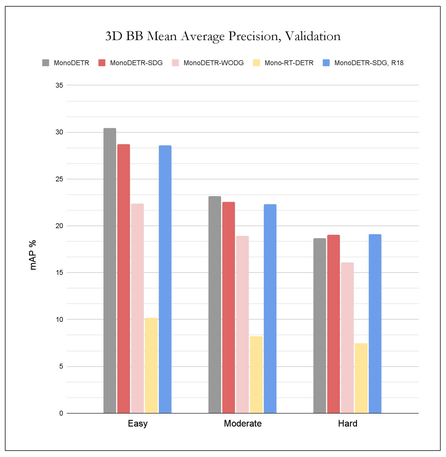


RESULTS - Quantitative

Model validation performance 3D Bounding Box

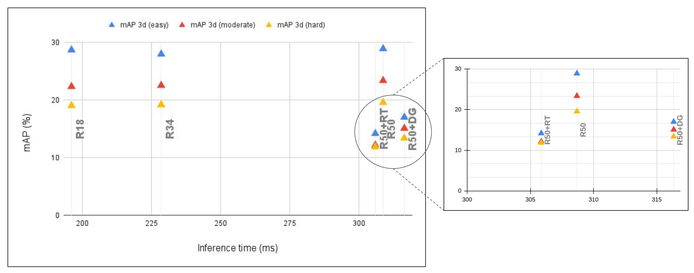
Method (from papers)	Val, AP 3D BB Easy - Mod - Hard
Cube R-CNN [2]	23.59 - 15.01 - 12.56
MonoDETR [1]	25.00 - 16.47 - 13.58

Method (ours)	Val, AP 3D BB Easy - Mod - Hard
MonoDETR *(lr=2e-4)*	30.46, 23.18, 18.65
MonoDETR *(lr=4e-4)*	18.75 - 14.96 - 13.70
MonoDETR-SDG *(lr=2e-4)* (w. semi depth guidance)	28.73 - 22.55 - 19.06
MonoDETR-SDG *(lr=4e-4)*	27.22 - 22.30 - 18.90
MonoDETR-WODG (lr=2e-4) (w/o depth guidance)	22.35 - 18.94 - 16.11
MonoDETR-WODG (lr=4e-4)	25.46 - 20.89 - 18.19
Mono-RT-DETR (lr=4e-4)	10.19 - 08.21 - 07.47
MonoDETR-SDG Resnet18 (lr=4e-4)	28.61 - 22.32 - 19.09
MonoDETR-SDG Resnet34 (lr=4e-4)	25.72 - 22.60 - 19.19

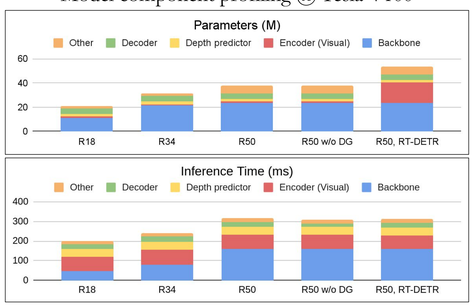


RESULTS - Inference

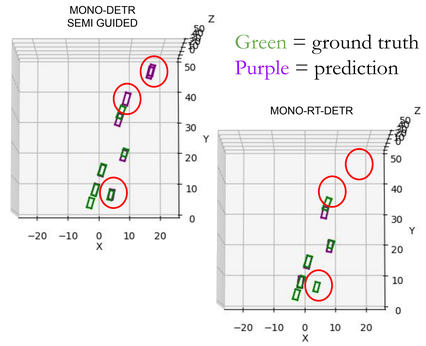
Precision versus inference performance



Model component profiling @ Tesla V100



RESULTS - Qualitative



CONCLUSIONS

- Successfully trained and evaluated monocular 3D object detection models based on MonoDETR.
- The semi-depth guided model is more stable in training across learning rates with similar best precision (mAP). This is related to the depth predictor's training loss gradient paths.
- The ResNet 50 backbone represented ~50% of the parameters and inference time in the baseline.
- ResNet 34 and 18 greatly reduces inference time without significantly affecting precision (mAP).
- The bipartite matching involved in the loss makes DETR models training unstable.
- The RT-DETR hybrid visual encoder results in worse visual queries, reducing performance.

REFERENCES

[1] Zhang, R., Qiu, H., Wang, T., Xu, X., Guo, Z., Qiao, Y., Gao, P., & Li, H. (2022). *MonoDETR: Depth-guided Transformer for Monocular 3D Object Detection*. ICCV 2023.
[2] Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., & Gkioxari, G. (2023). *Omni3D: A Large Benchmark and Model for 3D Object Detection in the Wild*. arXiv preprint arXiv:2207.10660
[3] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2023). *DETRs Beat YOLOs on Real-time Object Detection*. arXiv preprint arXiv:2304.08069.
[4] Reading, C., Harakeh, A., Chae, J., & Waslander, S. L. (2021). *Categorical Depth Distribution Network for Monocular 3D Object Detection*. arXiv preprint arXiv:2103.01100.