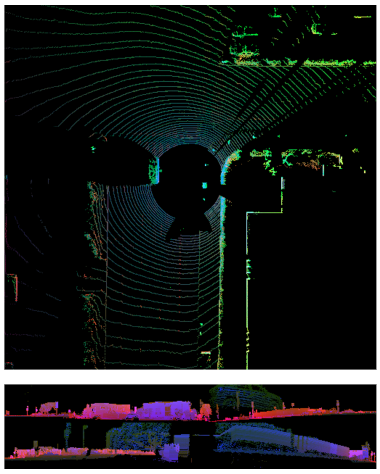
Dominik Fletschinger | Motivational Letter

Motivation



Machine Vision, ML Data driven, Internship at Sick AG - Coming from theoretical side of mechanical --> Tensor algebra, Tensor analysis and optimization, generalize well on Machine Learning and Deep Learning research. At Prof Böhlke who comes from mathematical side of mechanical engineering, I learned rigourous mathematical thinking and how to apply it to real world problems.

- challenge drives me
- ML has high relevance ELLIS, Why Ellis: - ELLIS is a great opportunity to work with the best in the field, to learn from them and to contribute to the field. - ELLIS is international, oppurtunities

Past research

As part of my Information Technology major field, I had the opportunity to participate in the Data Driven Engineering I/II course series by Dr. Cihan Ates, where I obtained a solid foundation in machine learning and optimization. With my first machine learning research project "Energy Consumption Prediction at High Granularity" I competed at the lecture accompanying project contest, where I placed in the top three . I gained valuable experience in the practical application of classical regression methods, such as Linear regression, Support Vector Regression and Random Forest models. In a second step I compared the performance of the classical methods with Recurrent Neural Networks and LSTM based neural network models. [TODO: make broader and less] In Data Driven Engineering II, my team and I worked on particle velocity and uncertainty estimation using convolutional autoencoders based on a variance attenuation loss. The excellent course in Machine Vision by Dr. Martin Lauer inspired me to focus more on perception and in particular object detection.

This led to my master's thesis titled "Masked Autoencoding as Pre-Training for Traffic Participant Detection," reviewed by Prof. Dr.-Ing. Christoph Stiller. The goal of the thesis was to investigate self-supervised learning as a pretext task for downstream 2D and 3D object detection in the context of autonomous driving. For pretraining Masked Image Modeling in particular the popular Masked Autoencoding [1] technique was used. To achieve real-time performance within a flexible detection framework we chose CenterNet. It reframes object detection as a keypoint estimation task and additionally regress bounding box parameters from the predicted keypoints. By performing 2D object detection on monocular images, 3D object detection on monocular images, and 3D object detection on bird's eye view representations of the lidar point cloud I was able to get an overview of the object detection field. First we benchmarked the detector on KITTI using a ResNet backbone. After replacing the ResNet with a vision transformer we met the practical challenge of small dataset size, large image size, and model size when training a vision transformer with little inductive biases compared to CNN-based models. While I saw decent results when using a pretrained vision transformer in the detectron-2 framework, training a vision transformer from scratch on KITTI is not practical. To mitigate these challenges we focussed on the larger Waymo Open Perception dataset using bird's eye view representations of the lidar point cloud. Further we investigated the impact of other vision transformers and found the hierarchical vision transformer Hiera [2]. It comes with inductive biases like local attention in early layers and token pooling. Through the clever application of mask units Hiera allows for the use of Masked Autoencoding (MAE) as a pretext task. In Figure 1 a exemplary MAE reconstruction of a bird's eye view point cloud representation on the Waymo Open Perception dataset is shown. We showed that the MAE pretext task is beneficial for the downstream task of 3D object detection evaluated on the Waymo Open Perception validation set. regarding augmentation our findings align with prior research namely that pointcloud rotation augmentation increases the performance and mirror augmentation decreases the performance.

I continued my research at the Research Center for Information Technology (FZI), where I designed data and training pipelines to incorporate video sequence input as well as multimodal inputs to the Hiera model. The multimodal approach to fuse bird's eye view and spherical view of the lidar point cloud as shown in Figure 2 by jointly masking and reconstructing the embeddings of the two views. For that I designed separate encoders for the modalities similar to the [3]. During MAE pretraining both modalities are reconstructed and during downstream 3D object detection only the bird's eye view was used for the detection task. Ultimately we confirmed that the MAE task is beneficial for the downstream task of 3D object detection. Furthermore MAE can be used to fuse multimodal inputs in a self supervised way. Finally by benchmarking against a ResNet backbone we showed that in the low data regime CNN-based architectures are still highly competitive in object detection. The opportunity to utilize large public like Waymo Open Perception in addition to training on the Jülich Supercomputing Centre, has enhanced my professional approach to machine learning. In a second project at FZI I processed accumulated

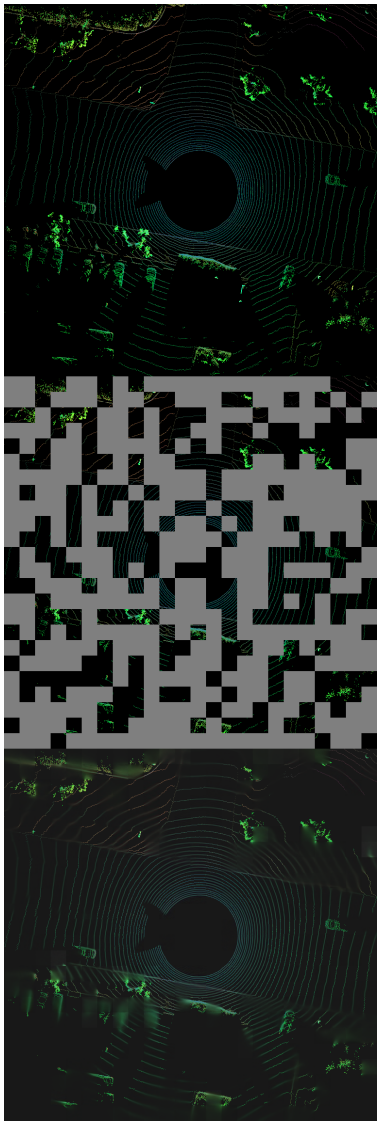


Figure 1: 3D Object Detection on Waymo Open using Hiera

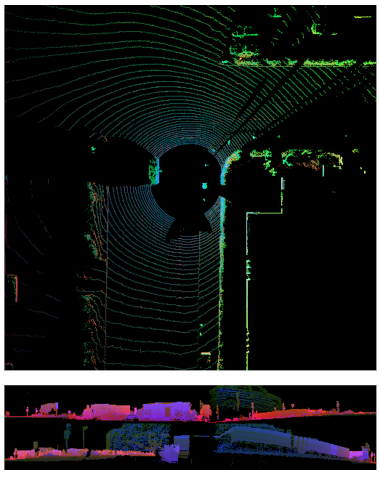


Figure 2: 3D Object Detection on Waymo Open using Hiera

Future research agenda

- Coming from perception

Line of research NeRF

- Currently following the line of research of nerf [4] pixel nerf, gaussian splatting and mvsplat (technical, perception) - Both NeRF scene representation and 3D Gaussian representation can represent scenes in a hybrid way. First arbitray 2D views of the scene can be rendered which is the main goal of novel view synthesis. Second, the 3D representation can be used for 3D reconstruction and scene understanding. NerF can directly be queried for 3D points and their color. Depending on the target representation the 3D Gaussian representation can be mapped to voxel or point cloud representations. With the emergence of methods to conditionally create NeRF and 3D Gaussian representations from images, the line of research can be extended to.

- Self supervised representation learning and pretraining (conceptual) --> Masked modeling for point clouds, voxel grids, and images (technical)

- Curriculum learning in combination with regularizing the learning process (conceptual) with multi task learning (technical)

social

- Looking forward to collaborate with other researchers but independent work is also important especially when coding

Student mentoring

- Looking forward Thesis students. Connect and network with other students.

Advisors

- Valada and Geiger

References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000--16009, 2022.

[2] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441--29454. PMLR, 2023.

[3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348--367. Springer, 2022.

[4] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99--106, 2021.