Dominik Fletschinger | **Motivational** Letter

Motivation

I am interested in computer vision and machine learning (ML) with a focus on perception and scene representation. I am particularly interested in self-supervised and semi-supervised learning techniques since they are important for real-world application with limited annotation budget. My prior research during my work at Research Center for Information Technology (FZI) under Dr. Ömer Sahin Tas and during my master's thesis under Prof. Christoph Stiller touched on self-supervised learning using masked image modeling, 2D and 3D object detection on camera and lidar input as well as multimodal fusion. I wish to earn a Ph.D. in the field of computer vision and machine learning.

I focused on computational mechanics in my bachelors where I obtained my fundamental understanding of tensor algebra, tensor analysis as well as optimization. In my bachelor's thesis supervised by Prof. Thomas Böhlke I researched algorithms to generate higher order irreducible tensor representations in the context of mechanical texture development. I delved into the field of computer vision and machine learning with the exceptional course on Machine Vision by Dr. Martin Lauer. There I immediately recognized the high transferability of the mathematical methods from continuum mechanics and was inspired to pursue a career in the field.

Motivated by the course on Machine Vision I applied for the internship at SICK AG, where I worked on the embedded application layer of a smart 3D-ToF Camera, used to detect obstacles in indoor scenes using classical ML methods. In many discussions the lack of labeled point cloud data as well as the difficulty to adapt to new domains using off-the-shelf detectors were mentioned as the reason for the slow implementation of deep learning. This emphasizes importance of self-supervised and unsupervised learning techniques. During my internship and time as a working student, I developed my skills as a software engineer and became a more effective team player through hands-on experience and collaboration.

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 built a strong foundation in machine learning and optimization. With my 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 and recurrent nueral network approaches such as RNNs and LSTMs. 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.

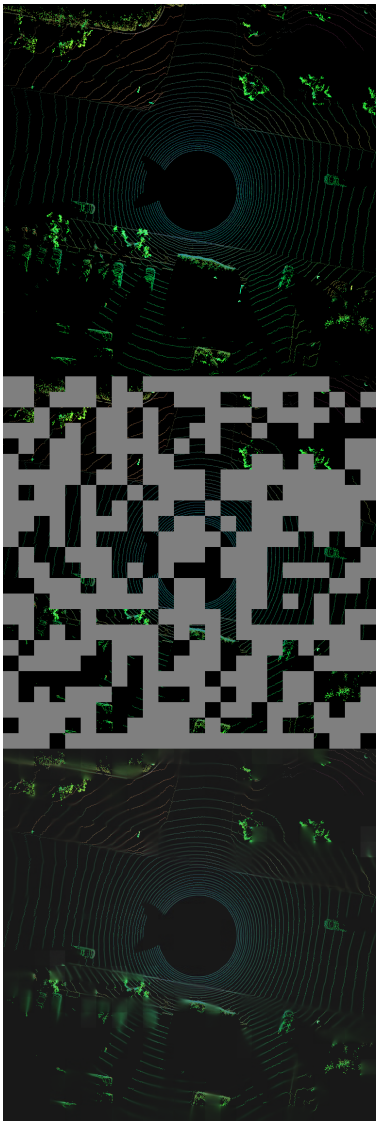

Figure 1: Masked Autoencoding (MAE) of Bird's Eye View Point Cloud Representation on Waymo Open Perception Dataset. The top shows the input, the middle the masked input and the bottom the reconstruction.

Past research

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 point clouds and annotations in KITTI-360 towards a format suitable for frame by frame object detection similar to the KITTI dataset. The project shows that a theoretically optimal frame by frame object detector still needs optimal dissection and accumulation of the point clouds and annotations to perform well on the KITTI-360 object detection benchmark.

Future research agenda

In the future I plan to continue my research in the field of machine learning and computer vision especially perception, object detection and the general machine learning methods. Today representation learning by masking is a technique applied not just in natural language or image modeling but also in domains like point cloud representation, trajectory prediction and others. In my future research I am keen to not only explore the possibilities of masked modeling, but also other self-supervised and semi supervised learning techniques. I believe that innovative training strategies such as regularizing via multi task and curriculum learning will further lead to novel research. For example the authors of DAFormer [4] show that seemingly simple techniques such as rare class sampling, warmup scheduling and additional distillation supervision improve the performance of their domain adaptation model. While the authors of ProFusion3D [5] showcase an impressive multimodal fusion pipeline, they also focus on training strategies via additional objectives in the pretraining phase. Noise prediction, cross-modal depth prediction and cross-modal intensity prediction are used on top of the MAE reconstruction objective.

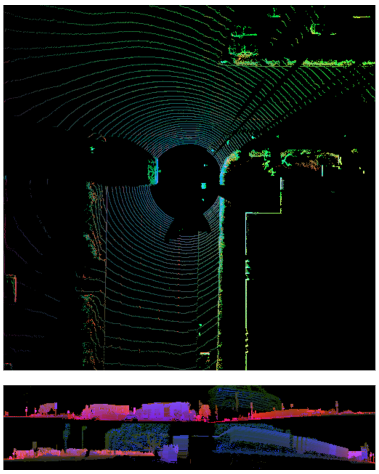

Figure 2: Multi View Fusion with Bird's Eye (top) and Spherical View (bottom)

Since the development of pixelNeRF [6], I am following the line of research of NeRF and 3D Gaussian representations. While pixelNeRF solved the problem of conditioning NeRFs on images in real-time, 3D Gaussian splatting on the other hand allows for fast rendering. With the recent development of methods like MVSplat [7] 3D Gaussian representations are obtained with only one forward pass. The 3D Gaussians are a unique latent space for 3D scene representations. The authors of S4C [8] use a NeRF to perform 3D semantic scene completion without the need for 3D annotations in a smart self-supervised manner. Similarly I would like to explore the use of 3D Gaussian representations to perform real-time 2D semantic segmentation 3D semantic scene completion within one model. While the 2D semantic segmentation can be rendered from the 3D Gaussian representation, the 3D semantic scene completion can be derived directly from the 3D Gaussian representation.

Closing

With the support of the ELLIS PhD Program, I am eager to delve into independent research while collaborating with fellow researchers to exchange ideas and insights. I look forward to mentoring students and assisting them with their research projects. The opportunity to engage in intensive research at the lab of my future ELLIS advisor excites me.

In particular I am interested in working with Prof. Dr. Abhinav Valada and Prof. Dr.-Ing. Andreas Geiger as my research interests strongly align with their work on perception and scene understanding.

The opportunity to collaborate with top-tier laboratories across Europe under the mentorship of renowned researchers in machine learning and its applications such as computer vision is unique withing the ELLIS Program. Its environment is ideal for pursuing novel fundamental and interdisciplinary research. With my solid research background, my software engineering expertise, and closely aligned interests with the program's faculty, I am confident that I am a suited fit to contribute and excel within the ELLIS PhD program.

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] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9924--9935, 2022.

[5] Rohit Mohan, Daniele Cattaneo, Abhinav Valada, and Florian Drews. Progressive multi-modal fusion for robust 3d object detection. In *8th Annual Conference on Robot Learning*, 2024.

[6] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578--4587, 2021.

[7] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370--386. Springer, 2025.

[8] Adrian Hayler, Felix Wimbauer, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. S4c: Self-supervised semantic scene completion with neural fields. In *2024 International Conference on 3D Vision (3DV)*, pages 409--420. IEEE, 2024.