Dominik Fletschinger | Motivational Letter

Motivation and Background

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, as they are crucial for real-world applications with limited data annotation. In 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 I worked on self-supervised learning using masked image modeling, 2D and 3D object detection on camera and LiDAR input as well as multimodal fusion. I would like to earn a Ph.D. in the field of computer vision.

I focused on computational mechanics in my bachelor where I built 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. The camera is used to detect obstacles on mobile agents in indoor scenes using classical ML methods. In discussions with colleagues the lack of labeled point cloud data were found to be the main reason for the slow implementation of deep learning, further emphazising the 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. With my 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 as well as recurrent neural network approaches. 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.

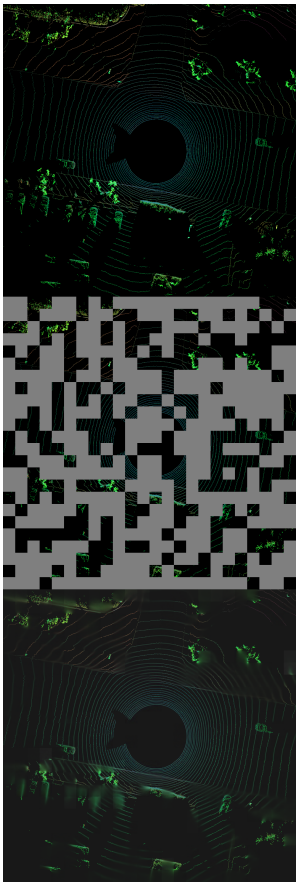

Figure 1: Masked Autoencoding (MAE) of Bird's Eye View Point Cloud Representation on Waymo Open Perception Dataset. From top to bottom: Input, Masked, Reconstruction

Past Research

I chose to delve deep into the field of perception with my master's thesis. During the thesis I investigated the impact of pretraining via Masked Image Modeling in particular the Masked Autoencoding [1] technique on improving the downstream task of object detection. To achieve real-time performance within a flexible detection framework we chose CenterNet, reframing object detection as a keypoint estimation. By working on 2D object detection on monocular images, and 3D object detection on monocular images and on bird's eye view representations of LiDAR point clouds, I gained a comprehensive overview of the field of object detection. Using vision transformer I was confronted with 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. To mitigate these challenges we focused on the larger Waymo Open Perception dataset using bird's eye view representations of the lidar point clouds. Furthermore I used the hierachical vision transformer which comes with additional inductive biases like local attention in early layers and token pooling. 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 confirmed that the MAE pretext task is beneficial for the downstream task of 3D object detection evaluated on the Waymo Open Perception validation set.

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 tot he hierarchical transformer. The multimodal approach to

fuse bird's eye view and spherical view of the lidar point cloud is shown in Figure 2. A fused representation is obtained by jointly masking and reconstructing the embeddings of the two modalities. Ultimately we confirmed that the MAE task is beneficial for the downstream task of 3D object detection. 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. At FZI the opportunity to train on the Jülich Supercomputing Centre, has enhanced my professional approach to machine learning.

In a second project at FZI, I processed global point clouds and global annotations from KITTI-360, converting them into 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

In the future, I plan to continue my research in computer vision especially scene understanding and perception while broadening my focus to explore new fields outside of object detection. I am keen to not only explore the possibilities of masked modeling, but also other self-supervised and semi supervised learning techniques.

The development of pixelNeRF [2] opened the line of research of using NeRFs for conditioned scene rendering. With the development of 3D Gaussian splatting and recently methods such as MVSplat [3] 3D Gaussian representations are obtained with only one forward pass while also rendered in real-time. The 3D Gaussians are a unique latent space for 3D scene representations with many new research avenues to explore. For example the authors of S4C [4] use a NeRF to perform 3D semantic scene completion without the need for 3D annotations in a self-supervised manner. I am interested in developing similar methods that make use of the fast rendering of 3D Gaussians for scene understanding.

Methododically I am particular interested in developing and focusing on training strategies such as regularizing via multi task and curriculum learning. For example the authors of ProFusion3D [5] showcase an impressive multimodal fusion pipeline, while they also focus on novel training strategies via additional objectives in the pretraining phase.

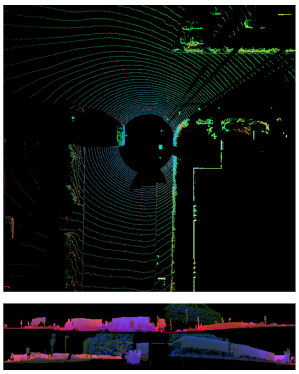

Figure 2: Multi view fusion with bird's eye view and spherical view

ELLIS PhD Program and Advisor

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 conduct intensive research at the lab of my future ELLIS advisor is exciting.

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 well 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] 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.

[3] 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.

[4] 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.

[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.