Dominik Fletschinger | **Motivational** Letter

I wish to earn my Ph.D. in the field of computer vision. My interest lie in the field of computer vision and machine learning (ML), with a focus on perception and scene understanding. 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 previous research, conducted at the Research Center for Information Technology (FZI) under the guidance of Dr. Ömer Sahin Tas and in the context of my master's thesis under the supervison of Prof. Christoph Stiller, I worked on self-supervised learning using masked image modelling, 2D and 3D object detection on camera and LiDAR input, and multimodal fusion. I am keen to engage in independent research while working in collaborating with fellow researchers to exchange ideas and insights. I am also interested in mentoring students and assisting them with their research projects.

Background

During my Bachelor's degree at the Karlsruhe Institute of Technology (KIT), I focused my studies in computational mechanics. This enabled me to develop a comprehensive understanding of tensor algebra, tensor analysis and optimisation. In my Bachelor's thesis, supervised by Prof. Thomas Böhlke, I conducted research into algorithms for generating higher-order irreducible tensor representations in the context of mechanical texture development. I explored the field of computer vision and machine learning through the exceptional course on machine vision taught by Dr. Martin Lauer. It was immediately apparent that the mathematical methods employed in continuum mechanics were highly transferable, which inspired me to pursue a career in this field.

Motivated by the course on Machine Vision I applied for the internship at SICK AG, where I contributed to the embedded application layer of a smart 3D-ToF camera. The camera is used to detect obstacles on mobile agents within indoor environments using classical ML methods. In discussions with colleagues, it was determined that the lack of labelled point cloud data were 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 period 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. In the context of my research project, entitled "Energy Consumption Prediction at High Granularity", I competed at the lecture accompanying project contest, where I was placed in the top three. I gained valuable experience in the practical application of both classical regression and forecasting methods and recurrent neural network approaches. In the second Data Driven Engineering course, my team and I worked on particle velocity and uncertainty estimation using convolutional autoencoders based on a variance attenuation loss.

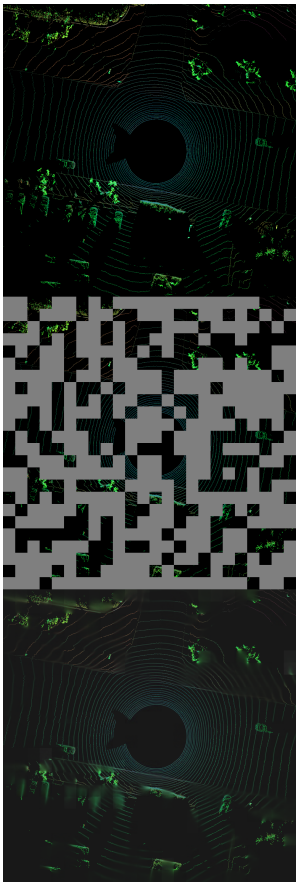

Figure 1: Masked Autoencoding (MAE) of Bird's Eye View Point Cloud Representation on Waymo Open Perception Dataset. From top to bottom: Input, Masked, Reconstruction

Past Research

I elected to delve deep into the field of perception with my master's thesis. The thesis investigated the impact of pretraining via Masked Image Modeling, with a particular focus on the Masked Autoencoding [1] technique, for improving the downstream task of object detection. In order 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 was able to gain a comprehensive overview of the field of object detection. The practical challenge of training a vision transformer on a small dataset, with large image size, and with a model with little inductive biases compared to CNN-based models, became apparent. To mitigate these challenges, I concentrated my efforts on the larger Waymo Open Perception dataset, I used bird's eye view representations of the lidar point clouds. Moreover, I used the hierarchical vision transformer Hiera, which incorporates additional inductive biases, such as local attention in early layers and

token pooling. In Figure 1 an exemplary MAE reconstruction of a bird's eye view point cloud representation on the Waymo Open Perception dataset is presented. I 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 into the hierarchical transformer. The multimodal approach to fuse bird's eye view and spherical view of the lidar point cloud is illustrated in Figure 2. A fused representation is obtained by jointly masking and reconstructing the embeddings of the two modalities. 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 train on the Jülich Supercomputing Centre, has enhanced my professional approach to machine learning at FZI.

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 demonstrates that a theoretically optimal frame-by-frame object detector still requires optimal dissection and accumulation of the point clouds and annotations in order to perform well on the KITTI-360 object detection benchmark.

Future Research

My future research will continue to focus on computer vision, with a particular emphasis on scene understanding and perception. Additionally, I intend to expand my research scope to encompass new fields beyond object detection. In addition to investigating the potential of masked modelling, I am interested in examining other self-supervised and semi-supervised learning techniques.

The development of pixelNeRF [2] opened the line of research of utilising NeRFs for conditioned scene rendering. The development of 3D Gaussian splatting, and more recently methods such as MVSplat [3] 3D Gaussian representations are obtained with single forward pass, while also rendering them in real-time. The 3D Gaussians are a unique latent space for 3D scene representations, offering many new research avenues to explore. To illustrate the authors of S4C [4] employ a pixelNeRF 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.

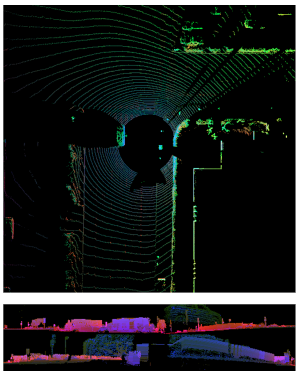

Figure 2: Multi view fusion with bird's eye view and spherical view

Methodically 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] present an impressive multimodal fusion pipeline, while also focusing on novel training strategies via additional objectives, such as denoising, in the pretraining phase.

ELLIS PhD Program

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 are closely aligned with their work on perception and scene understanding.

The opportunity to collaborate with leading laboratories across Europe, under the mentorship of renowned researchers in machine learning and its applications such as computer vision, is unique within the ELLIS Program. The 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.