Elior Benarous web: *https://ebenarous.github.io* ebenarous@ethz.ch

Statement of Purpose

Neural scaling laws demonstrate that models perform best by leveraging greater quantities of data. In line with these findings, self-supervised learning has proven remarkably successful to alleviate the need for labels, enabling AI systems to learn from vast amounts of data and capture more subtle and generalizable patterns. As an MSc student at ETH Zürich, and graduate researcher at both Harvard and Caltech-NASA Jet Propulsion Laboratory, I have focused my previous research on this paradigm. During my PhD, I am eager to further explore self-supervised learning through approaches like synthetic data, self-play mechanisms, and the development of more generalizable representations.

Inductive Biases and Synthetic Data. As a first-year **ETH** MSc student under Prof. Thomas Hofmann, I investigated self-supervised pre-training strategies to improve the representations of convolutional encoders. Deep neural networks learn the simplest solution to a problem, often exploiting spurious shortcuts that generalize poorly out-of-distribution. An example is focusing on local textures rather than global shapes. To enhance the versatility of representations across different modalities and tasks, it is essential to design strategies that encourage feature extractors to capture a wide range of properties. Inspired by learnable positional embeddings in vision transformers (ViTs), I developed an objective to predict the relative location of two randomly sampled and visually distorted patches in an image. Our method, described in a **co-first authored paper** [1], transformed the features from predominantly style-oriented to shape-focused, significantly enhancing generalization in few-shot classification.

Intrigued by label-free pre-training methods, I extended my research to self-supervised learning with simple-to-generate, yet unrealistic synthetic images. With increasing demand for larger datasets, synthetic data offers a scalable alternative to real data, making it crucial to understand its effectiveness-driving properties. Procedurally generated images, like fractals, hold promise in this area. However, their lack of realism poses difficulties to visually assess their potential. Accordingly, I proposed evaluating the quality of such data by holistically analysing the evolution of ViTs' and CNNs' inductive biases over training. I collaborated with a team of 3 researchers to establish a novel interpretation of a model's shape bias – defined as its reliance on shapes over textures for recognition – as a tool for estimating the diversity of its training dataset. Furthermore, we uncovered the intricate relationship between shape bias, dataset diversity and sample naturalism in the context of generalization. Our findings, presented in my **first-author paper at NeurIPS** [2], pave the way for robust evaluation metrics and effective curation strategies for synthetic datasets. Through these projects on self-supervised learning, I gained insights into overcoming common pitfalls of vision models like preferring textures over shapes, with the aim of enhancing their generalizability.

Self-Play in Image Generation. Building on my previous work, I set out to enhance self-play methods for synthetic data generation, under the advisory of Prof. Yilun Du and Prof. Heng Yang at **Harvard**. The limited control over the output of text-to-image diffusion models significantly constrains their applicability in use-cases demanding precision and consistency. Specifically, they struggle to 1) perform accurate edits of a subject in complex scenes while leaving unrelated areas of the image intact, and 2) leverage a visual prompt to capture subtle visual nuances. With closer instruction alignment, their high-quality samples could enable substantial dataset expansion in data-scarce domains. Given that the standard diffusion reconstruction loss is not well-suited for training editing models, I proposed a reinforcement learning approach with AI feedback. It features an objective designed for structural alignment to preserve content, and a semantic-focused component to ensure the generated image accurately reflects the visual prompt. In my **first-authored paper** [3], we demonstrated substantial precision improvements in generated samples. Some applications include weather editing on crowded road images, and extend to a robotics setting where the diffusion model edits simulation data to closely resemble real-world data, significantly enhancing few-shot performance. Overall, tackling accuracy challenges in generative models taught me the value of examining problems from unconventional perspectives, like self-play. This approach informed me how to design objectives that align model performance with intended outcomes.

Multi-Modal Generalizable Representations. Interested in the applications of self-supervised learning to embodied intelligent agents, I joined the **Caltech-NASA Jet Propulsion Laboratory** as a visiting researcher under the guidance of Roboticist Deegan Atha and Prof. Marco Hutter. My research focused on improving wheeled autonomy in extreme off-road terrains, where embedded vision models struggle with semantic and geometric understanding. Enhancing their adaptability to outliers is crucial for enabling robots to generalize effectively in diverse and dynamic environments. In our research, I developed an on-the-fly clustering method to enhance zero-shot visual understanding. This approach utilized multiple modalities, including 2D and 3D visual data, and vehicle suspension time-series information. Based on a self-supervised model, our method adapted its representations to recognize unknown classes and sub-classes from encountered broader categories. We also improved the understanding of variations in terrain elevation over long distances and in high-grass areas, where many sensors offer sparse and unreliable ground-truth data. This enabled the vehicle to navigate robustly and safely around these new instances and terrains. Through this research, I became acquainted with multimodal approaches for creating better representations, deepening my understanding of building robust systems suited to embodied intelligent agents capable of effectively interacting in dynamic environments.

As a PhD student, I am excited to explore areas shaped by the breadth of my past research experiences. Notably, I am eager to align self-supervised representations more closely with human-like semantic conceptualization, focusing on purpose over shape. Purpose-oriented models could greatly improve the performance of embodied agents. I'm also excited to advance self-play strategies for generative models. With sufficient instruction alignment, they could serve as valuable training data generators, much like how simulation environments are leveraged in robot learning.

References

[1] Elior Benarous*, Dustin. Brunner*, Jonathan Manz*, Felix Yang*, and Thomas Hofmann. Enforcing style invariance in patch localization, 2022.

[2] Elior Benarous, Sotiris Anagnostidis, Luca Biggio, and Thomas Hofmann. Harnessing synthetic datasets: The role of shape bias in deep neural network generalization. In *Advances in Neural Information Processing Systems (NeurIPS), Workshop on Synthetic Data Generation with Generative AI*, 2023.

[3] Elior Benarous, Yilun Du, and Heng Yang. Image-editing specialists: A multi-reward approach for diffusion models, 2024.

Links to all papers are included in my CV and on my personal website.