TANUSHREE BANERJEE STATEMENT OF PURPOSE

The human mind infers extensive 3D information about the world from 2D photographs. Further, it can generalize this understanding to unseen situations. Can we emulate these abilities in machines? A possible approach is to use priors learned from physical world observations to complete missing information in 2D. My current work applies this principle by **formulating perception as inverse generation**, i.e. deducing 3D information in the input by learning to generate it. Leveraging rich generative priors, this approach yields generated images as interpretable explanations for inferred information, rendering it amenable to safer deployment in applications like robotics and autonomous driving. I want to advance such ideas toward safer, generalizable machines reflecting human cognition during my PhD. My research on **bias in vision models, LLM self-refinement**, and **inverse generation** at Princeton University has prepared me for this challenge.

I was introduced to research by **Prof. Olga Russakovsky**, under whom I analyzed bias in skin lesion classification [5] and hallucination in visual question answering (VQA) [4]. I designed methods identifying questions unrelated to an image, improving accuracy by 39% over the random baseline, to reduce hallucination in VQA. Later, under **Prof. Karthik Narasimhan**, I explored human-in-the-loop and LLM self-refinement methods to detect lies in Diplomacy game conversations. I designed a bootstrapping framework that generated feedback to refine initial LLM predictions, rivaling previous supervised learning methods, resulting in a **first-author arXiv preprint** [3]. These experiences in various AI research areas made me reflect upon the fundamental limitations of state-of-the-art models: often relying on feed-forward prediction, they are un-interpretable, making it hard to reason about their failure modes. Could we re-think these architectures to expose their underlying reasoning?

In my undergraduate thesis under **Prof. Felix Heide**, I explored a potential approach to this: **formulating perception as an inverse generation task**. I chose monocular 3D perception as a representative problem. Deducing 3D information from 2D images is an inherently challenging task. Moreover, canonical feed-forward prediction-based perception methods are uninterpretable and struggle to generalize to unseen data. Recasting perception as an inverse problem addresses these challenges. I have investigated multiple perspectives on this idea.

The first is a graphics perspective: assuming vision inherently arises from an expressive 3D representation provided by graphics rendering pipelines, we approached perception as an inverse rendering problem. We chose 3D monocular Multi-Object Tracking (MOT) as a representative task. Given a generative 3D object representation, we found latents best reproducing observed object images via inverse rendering to match the observation with previous ones for tracking. The image generated from these latents yields an interpretable explanation for tracked objects. We outperformed existing dataset-agnostic and dataset-specific learned approaches operating on the same detections. This resulted in a **co-first author arXiv preprint** [1] **under review at *Nature Machine Intelligence***. I received the **Outstanding Computer Science Senior Thesis Prize** awarded to 6 of 216 graduating CS students for my work. I designed our optimization mechanism, ran all experiments, validated our method on 3D MOT benchmarks, and conducted extensive ablations. I am most excited about this method's explainable nature, enabling the analysis of failure modes via the rendered images from optimized latents. Conceptually, I was

intrigued by how learning to represent the object in general enabled MOT by exploiting generative priors.

My current project [2] investigates a second perspective on inverse generation for perception, hinging on generative features as strong world representations. We solve perception by explicitly extracting structured 3D information from unstructured latents containing embedded 3D information, choosing 3D object detection as a representative task. To this end, we train a Variational Auto-Encoder to learn a prior distribution over object appearance, class, and pose. The encoder decomposes the observed image into pose and object features, while the decoder, given object features, generates images explicitly conditioned on the pose. To solve 3D detection, we recover the pose that best generates the observed image by optimizing over the decoder's latent space. We can examine detected objects from novel views via images generated from recovered features. I designed the training and inverse pose optimization method. Pose optimization proved non-trivial; using gradient-based optimization to minimize reconstruction loss between the generated and observed image pair failed due to convergence at local minima. Realizing optical flow estimates the warp between this image pair, we compute the relative shift and scale, recovering the optimal pose in a single step. Qualitative results are promising; our method detects objects despite occlusions. Currently, I am investigating different objectness prediction methods to improve precision.

I want to continue my work on **leveraging generative models for explainable perception**. I am excited by this paradigm's potential to transfer advances in generative modeling to predictive vision. While my past work explores *analysis-by-synthesis* – inferring physical properties of the space by learning to generate it – as one approach under this paradigm, I am keen to explore alternatives. This could enable a unified perception framework, simultaneously solving several tasks by mimicking human reasoning. Refining these ideas in a PhD program is the natural next step toward my goal of building machines that reflect human cognition.

Following graduate school, I aim to pursue an academic career advancing machine perception to comprehend the 3D world behind 2D photographs. A PhD is a crucial step toward these goals.

References

* denotes equal contribution.

[1] Julian Ost*, **Tanushree Banerjee***, Mario Bijelic, Felix Heide, *"Inverse Neural Rendering for Explainable Multi-Object Tracking,"* **arXiv preprint (under review at Nature Machine Intelligence)**, April 2024. Project page: https://tinyurl.com/inr-track.

[2] **Tanushree Banerjee***, Julian Ost*, Maolin Mao, Mario Bijelic, Felix Heide, *"OD-VAE: Inverting Generation for 3D Object Detection,"* **ongoing project**. Manuscript draft: https://tinyurl.com/odvae-draft

[3] **Tanushree Banerjee**, Richard Zhu, Runzhe Yang, Karthik Narasimhan, *"LLMs are Superior Feedback Providers: Bootstrapping Reasoning for Lie Detection with Self-Generated Feedback,"* **arXiv preprint**. Paper: https://arxiv.org/abs/2408.13915

[4] **Tanushree Banerjee**, Advisor: Prof. Olga Russakovksy, *"Reducing Object Hallucination in Visual Question Answering,"* **Independent Work Project**, Spring 2022. Report: https://tinyurl.com/hallucination-vqa

[5] **Tanushree Banerjee**, Advisor: Prof. Olga Russakovksy, *"Bias in Skin Lesion Classification,"* **Independent Work Project**, Spring 2021. Report: https://tinyurl.com/bias-skin-lesion