Boxuan Zhang | **Research** Statement

Research Priorities

My research focuses on crafting trustworthy AI and ML models that prioritize efficiency and safety, thereby ensuring intelligent systems are equipped with decision-making capabilities that are both efficient and reliable in real-world environments.

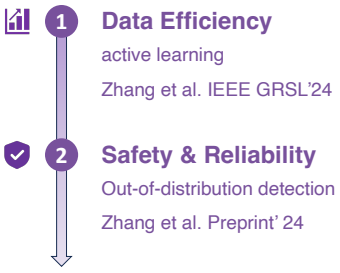
I have graduated with a Master's degree from Wuhan University, specializing in machine learning and computer vision. In the burgeoning epoch of Artificial Intelligence, propelled by swift advancements in machine Learning technologies, the emergence of Foundation Models (FMs) is progressively altering the way we navigate our daily lives, marking a significant shift in our interaction with intelligent systems. The powerful capabilities of AI models, while promising, confront us with dual issues of data utilization efficiency and reliability in real-world deployments. For instance, the quest for efficient data utilization is particularly pronounced in scenarios where labeled data resources are scarce, such as in medical diagnostics and remote sensing, where the need for accurate AI models is heightened but the availability of training data is limited. Meanwhile, the reliability of AI models is a pressing concern, particularly in high-stakes environments like autonomous driving, where models may exhibit overconfident yet incorrect predictions for out-of-distribution classes, such as misidentifying a cow as a car, leading to potentially catastrophic consequences. Furthermore, the potential for generative models, notably large language models (LLMs), to produce misleading or harmful content that deviates from human values, poses a threat to informed decision-making in dynamic, open-world scenarios. This highlights the urgent need to bolster the robustness and trustworthiness of AI systems across various domains.

To address the above challenges, my research endeavors to enhance the data utilization efficiency and ensure the safety and reliability of AI models, laying the groundwork for an efficient and trustworthy AI ecosystem. In the realm of **data utilization efficiency** of AI/ML models, I am focused on the challenge of scarce labeled resources in the field of **remote sensing object detection**, where the complexity and variability of natural landscapes coupled with the vast data scale make manual annotation both time-consuming and resource-intensive. By integrating active learning (Fig. B) with semi-supervised learning, my research aims to train a high-performance detector using only a minimal amount of annotated imagery. This innovative approach significantly reduces the model's dependence on labeled data, making it exceptionally suitable for applications in remote sensing [1]. Concentrating on the **safety and reliability** of AI/ML models, I have delved into the critical issue of **out-of-distribution (OOD) detection** within trustworthy machine learning. OOD detection (Fig. C) is crucial for ensuring that AI/ML systems can reliably identify and reject data that deviates from what they were trained on, termed as OOD data, thereby maintaining robust performance and preventing catastrophic failures in sensitive applications. My research has identified a unique challenge in this domain: the difficulty of extracting robust and discriminative features from a single input, which has been overlooked in many existing approaches. To overcome this, I have developed an innovative scoring framework with expanded input dimensionality, which improves the ability of OOD detector to accurately identify anomalies [2].

During my two-year graduate research journey, I authored two academic papers as the first author. The first paper, published at IEEE GRSL 2024 and under review at NeurIPS 2024, respectively. In the following sections, I will discuss my contributions and future research plan in building efficient and reliable AI/ML models in more detail.

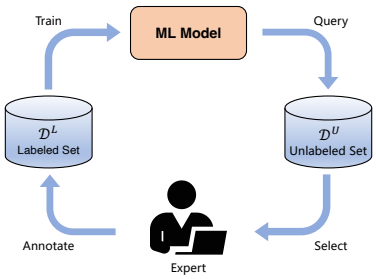
A. Research Overview

My research tackles two areas of efficiency and safety to improve intelligent systems.



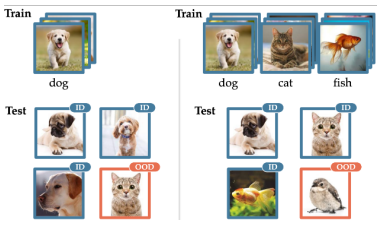
B. Definition of Active Learning

Active Learning uses a step-by-step process to identify the most valuable and informative data



C. Definition of OOD Detection

Traditional OOD detection faces two classes of data in the test time: ID classes (i.e., training data) and OOD classes (i.e., unknown test data.)



Research Accomplishments

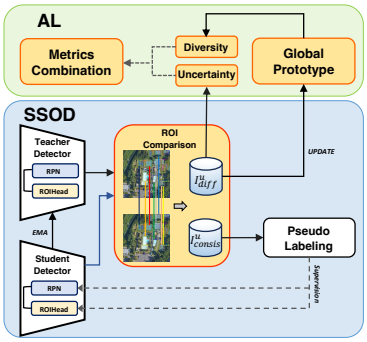
Part I: Leveraging Active Learning for Data Utilization Efficiency.

The first part of my research is situated at data utilization efficiency of AI models through active learning techniques in remote sensing scenarios. The success of deep learning approaches heavily relies on large-scale datasets with accurately labeled data, which are typically annotated by human experts. Unlike natural scenes images, remote sensing images often contain objects with various orientations and scales, which further complicates the object-level labeling process. Hence, the availability of labeled images for object detection is usually limited. To address the problem of limited labeled images, I noticed two promising techniques in machine learning, Semi-Supervised Learning (SSL) and Active Learning (AL). However, existing methods usually ignore the redundancy in the RoIs, resulting in more images are selected to improve the performance for Semi-supervised Object Detection(SSOD).

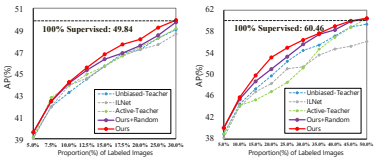
Following this, I developed SSOD-AT, a novel method to boost SSOD with AL based on the teacher-student network, which can provide both **confident pseudo-labels** and **informative images**. In SSOD-AT, A RoI comparison module(RoICM) is introduced by comparing the RoIs generated by teacher and student network to alleviate the influence of noisy RoIs for SSL and help AL to evaluate the uncertainty of images. I further incorporates the global class prototype for the diversity of selected images. The combination of the two sampling strategies maximizes the effectiveness of AL process. Compared with the best performance in the SOTA methods, SSOD-AT achieves 2% improvement at most cases in the whole active learning. This work is accepted by IEEE GRSL 2024 for its innovative perspective to mitigate the reliance on labeled resources in remote sensing scenarios.

D. Part I Overview

My main contributions are highlighted in the following image.



Detection results on the two remote-sensing datasets, DIOR (left) and DOTA (right).



Part II: Input Dimensionality Expansion for OOD Detection.

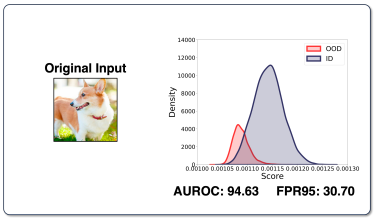
The second part of my research explores the critical and timely area of AI safety, specifically in the out-of-distribution detection for AI/ML model's reliability. Out-of-distribution (OOD) detection aims to identify OOD inputs from unknown classes, which is important for the reliable deployment of machine learning models in the open world. Various scoring functions are proposed to distinguish it from in-distribution (ID) data. However, existing methods generally focus on excavating the discriminative information from a single input, which implicitly limits its representation dimension, thus leaving some hard-to-distinguish OOD samples with features similar to ID samples fail to be identified. A series of works propose to eliminate the effects through the perspective of feature representation. However, achieving that is demonstrated to be hard. Adopting some agnostic corruptions on the single input may result in worse performance. Therefore, it naturally motivates the following critical research question: *What if we expand the dimension of representation for the original inputs to enhance OOD discriminative representations?*

Based on the above, I introduced a novel perspective to investigate that, i.e., employing the common corruptions in the input space. Through a systematical comparison, I revealed an interesting phenomenon termed *confidence mutation*, where the confidence of OOD data can decrease significantly under the corruptions, while ID data shows higher confidence expectation considering different input dimensions. To this end, I developed **Co***nfidence a***Ver***age* (CoVer), a new scoring framework that simultaneously considers the original and expanded input dimensions with a simple but effective average operation. CoVer can capture the dynamic differences by simply averaging the scores obtained from different corrupted inputs and the original ones, making the OOD and ID distributions more separable in detection tasks. I have conducted extensive experiments to

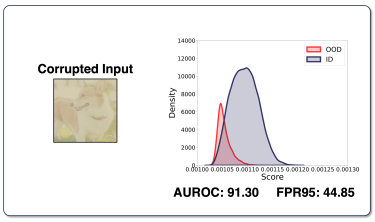
E. Part II Overview

CoVer performs a better ID-OOD separability with multiple inputs.

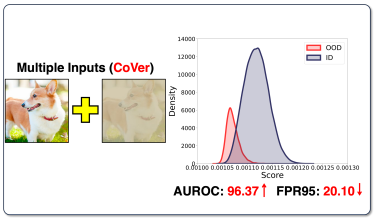
Single original input.



Single corrupted input.



Our CoVer with multiple inputs.



present its effectiveness and compatibility with different methods. I believe that this work is potentially a new generalized framework and provides new insights into OOD detection from a different perspective.

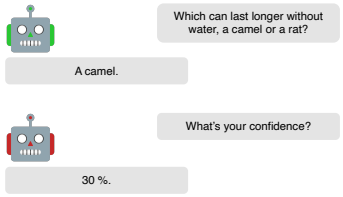
Research Plan

My future research plans remain anchored in the domain of trustworthy machine learning, with a keen focus on extending its applications to the burgeoning field of foundation models. My research agenda is divided into two distinct yet interlinked phases, reflecting both short-term and long-term aspirations.

Direction I: Uncertainty Quantification in LLMs.

In the near term, I am poised to immerse myself in the realm of uncertainty quantification within Large Language Models (LLMs). One of the biggest open questions about whether large language models (LLMs) can benefit society and reliably be used for decision making hinges on whether or not they can accurately represent uncertainty over the correctness of their output [3]. However, LLMs struggle to assign reliable confidence estimates to their generations (Fig. F). This inherent unpredictability, particularly in high-stakes scenarios, is a critical concern that demands a rigorous and meticulous approach. My research will endeavor to develop sophisticated methodologies that can quantify the uncertainties intrinsic to LLMs, thereby enhancing their reliability and utility in real-world applications.

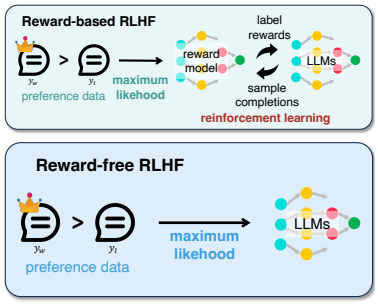
F. UQ in LLMs.
Large language models struggle to assign reliable confidence estimates to their generations.



Direction II: Align LLMs with Human Preferences.

My long-term academic trajectory is set to delve into the feild of LLMs, with the ultimate goal of aligning the generations of large language models with human preferences. LLMs, exemplified by ChatGPT and LLaMa, have showcased remarkable proficiency across a series of applications. However, their training on vast and varied datasets can sometimes result in the inadvertent propagation of misinformation and the generation of content that may be considered harmful. Aligning LLMs with human preferences is crucial for enhancing their utility in terms of helpfulness, truthfulness, safety, harmlessness, and interestingness. Various methods have been developed for fine-tuning LLMs, with the most widely used method called Reinforcement Learning from Human Feedback (RLHF, Fig. G) [4]. My research agenda is to align LLMs with human objectives through reward-free RLHF techniques, thereby fortifying their alignment with ethical standards and amplifying their practical utility in real-world scenarios.

G. RLHF in LLMs.
RLHF can be categorized into reward-based (pioneered by OpenAI) and reward-free (e.g., DPO, RRHF, and PRO) methods.



Through these endeavors, I aspire to contribute to the evolution of machine learning, steering it towards a future where AI is not just a tool but a trustworthy partner.

References

[1] Boxuan Zhang, Zengmao Wang, and Bo Du. Boosting semi-supervised object detection in remote sensing images with active teaching. *IEEE Geoscience and Remote Sensing Letters*, 2024.

[2] Boxuan Zhang, Jianing Zhu, Zengmao Wang, Tongliang Liu, Bo Du, and Bo Han. What if the input is expanded in ood detection? In *Preprint*, 2024.

[3] Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In *Arxiv*, 2024.

[4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. 2022.