In today's rapidly evolving landscape of AI technology, machine learning (ML) models, including the recent foundation models (FMs) have emerged as transformative forces shaping various applications. Despite the immense capabilities, they bring forth challenges related to the model's reliability upon deployment in the real world. For example, classic discriminative neural networks can produce overconfident yet incorrect predictions for unfamiliar out-of-distribution (OOD) classes. Moreover, generative models, such as large language models (LLMs), can generate untruthful or harmful content that contradicts human values, potentially compromising critical decision-making in an ever-changing open world. Addressing these issues is not merely a technical necessity but a fundamental requirement for the responsible deployment of AI technologies.

My research vision, as a crucial part of AI safety research, is to ***create an AI landscape where algorithmic accuracy and reliability are equally prioritized, ultimately broadening the impact and responsible reach of ML technologies***. To achieve this, I focus on developing fundamental algorithms and providing theoretical understandings that enable safe and reliable decision-making in the open world. Practically, the proposed frameworks can be used to build reliable real-world AI applications that could encounter multifaceted reliability challenges from both the inputs (such as the unknown OOD data) and outputs (e.g., FMs' hallucinated generations), which makes this endeavor both timely and highly relevant. This research direction is increasingly recognized as an important aspect of the AI research roadmap over the next five years and beyond [1].

Reliable ML introduces core challenges in characterizing the reliability of off-the-shelf learning algorithms, which typically minimize errors on in-distribution (ID) data from $\mathbb{P}_{\text{in}}$ without accounting for uncertainties that could arise outside $\mathbb{P}_{\text{in}}$. For instance, the widely used empirical risk minimization (ERM) [2], operates under the closed-world assumption (*i.e.*, no distribution shift between training and inference). Models optimized with ERM are known to produce overconfidence predictions on OOD data [3], since the decision boundary is not conservative. To address this challenge, I developed novel frameworks that jointly optimize for both: **(1)** accurate prediction of samples from $\mathbb{P}_{\text{in}}$, and **(2)** reliable handling of data from outside $\mathbb{P}_{\text{in}}$. Given a weighting factor α, this can be formalized as follows:

$$\text{argmin} \ [\mathcal{R}_{\text{accuracy}} + \alpha \cdot \mathcal{R}_{\text{reliability}}]. \tag{1}$$

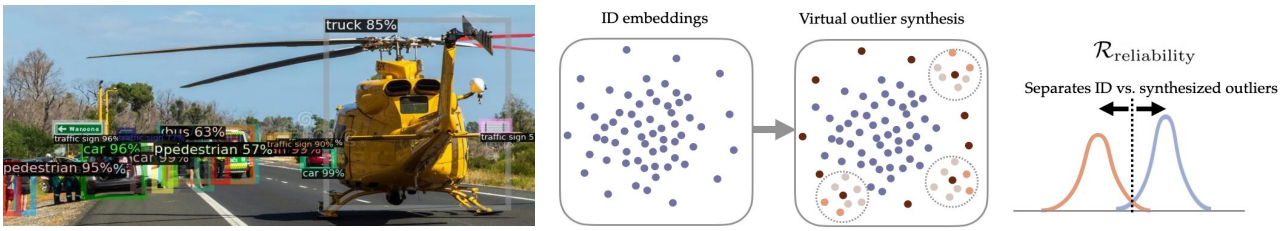
As an example, $\mathcal{R}_{\text{accuracy}}$ can be the risk that classifies ID samples into known classes while $\mathcal{R}_{\text{reliability}}$ aims to distinguish ID vs. OOD. The introduction of the reliability risk term $\mathcal{R}_{\text{reliability}}$ is crucial to prevent overconfident predictions on unknown data and improve test-time reliability when encountering unknowns. However, incorporating this reliability risk requires large-scale human annotations, e.g., binary ID and OOD labels, which could limit the practical usage of the proposed framework. Therefore, my research contributes to ***developing the foundations of reliable machine learning with minimal human supervision***, which spans three key aspects:

1. I developed novel unknown-aware learning frameworks that teach the models what they don't know without having explicit knowledge about unknowns. The framework enables tractable learning from the unknowns by adaptively generating virtual outliers from the low-likelihood region in both the feature [4, 5, 6] and input space [7], and shows strong efficacy and interpretability for regularizing the model to discriminate the boundaries between known and unknown data.

2. I designed algorithms and theoretical analysis for unknown-aware learning by leveraging unlabeled data collected from the models' deployment environment. This wild data is a mixture of ID and OOD data by an unknown mixing ratio. Methods I designed such as *gradient SVD score* [8, 9] and *constrained optimization* [10] can facilitate OOD detection and generalization on these real-world reliability challenges.

3. I built reliable foundation models by investigating the reliability blind spots of language models, such as untruthful generations [11], malicious prompts [12], and noisy alignment data [13]. My work seeks to fundamentally understand the sources of these issues by developing algorithms that leverage unlabeled data to identify and mitigate the unintended information, which ensures safer human-AI interactions.

My research has led to impactful publications in top-tier ML and vision venues and has been recognized by Rising Stars in Data Science and Jane Street Graduate Research Fellowship programs. Many of my works has been integrated into the OpenOOD benchmark [14, 15], and have received considerable follow-ups from worldwide major industry labs, such as Google [16], Microsoft [17], Amazon [18], Apple [19], Adobe [20], Air Force Research [21], Toyota [22], LG [23], Alibaba [24] etc. The scientific impact of reliable ML is profound, I am excited to explore interdisciplinary collaborations across computer science, statistics, biology science, and policy to push the boundaries of reliable ML as a professor. I outline my research in detail next.



(a) Overconfidence on OOD data (b) Outlier Synthesis Framework

Figure 1: (a) An object detection model trained on BDD-100k dataset [26] produces overconfident predictions for OOD objects (e.g., helicopter), highlighting reliability concerns in ML models during deployment. Test images are sampled from MS-COCO [27]. (b) Overview of my proposed outlier synthesis framework for unknown-aware learning.

1 Foundations of Unknown-Aware Learning: A Data Generation Perspective

Motivation. Ensuring safe and reliable AI systems requires addressing a critical issue: the overconfident predictions made on the OOD inputs [3]. These inputs arise from unknown categories and should ideally be excluded from model predictions. For example, in self-driving car applications, my research is **the first** to discover that an object detection model trained on ID objects (e.g., cars, pedestrians) might confidently misidentify an unusual object, such as a helicopter on a highway, as a known object; see Figure 1 (a). Such failures not only raise concerns about model reliability but also pose serious risks in safety-critical deployments.

The vulnerability to OOD inputs stems from the lack of explicit knowledge of unknowns during training, as neural networks are typically optimized only on ID data. While this approach effectively captures ID tasks, the resulting decision boundaries can be inadequate for OOD detection. Ideally, a model should maintain high confidence for ID data and exhibit high uncertainty for OOD samples, yet achieving this goal is challenging due to the absence of labeled outliers. My research tackles this challenge for unknown-aware learning through an automated outlier generation paradigm, which offers greater feasibility and flexibility than approaches requiring extensive human annotations [25]. I outline three core fundamental contributions:

Tractable learning foundation by outlier synthesis. My work VOS [4] (ICLR'22) *laid the foundation of a learning framework called virtual outlier synthesis to regularize the models' decision boundary*. This approach is based on modeling ID features as Gaussians, reject sampling to synthesize virtual outliers from low-likelihood regions, and a novel unknown-aware training objective that contrastively shapes the uncertainty energy surface between ID data and synthesized outliers. Additionally, VOS delivers the insight that synthesizing outliers in the feature space is more tractable than generating high-dimensional pixels [28]. My subsequent work, NPOS [6] (ICLR'23), relaxed the Gaussian assumption through a non-parametric synthesis approach, yielding improved results on language models and larger datasets. The STUD method [5], presented at CVPR'22 as an **oral**, further demonstrated the efficacy of this approach in **real-world practice**, i.e., video object detection, distilling unknown objects in both spatial and temporal dimensions to regularize model decision boundaries.

Interpretable outlier synthesis. While feature-space synthesis is effective, it doesn't allow *human-compatible interpretation like visual pixels*. To address this, my NeurIPS'23 paper Dream-OOD [7] introduced a framework to *comprehensively study the interactions between feature-space and pixel-space synthesis*. The method learns a text-conditioned visual latent space, enabling outlier sampling and decoding by diffusion models, which not only enhances interpretability but also achieves strong results on OOD detection benchmarks. It has garnered quite a few interests from community, prompting follow-up research on pixel-space outlier synthesis [23, 29, 30].

Understanding the impact of in-distribution data. Beyond focusing on reliability risks in the unknown-aware learning framework, it's crucial to address the in-distribution accuracy term $\mathcal{R}_{\text{accuracy}}$ in Equation 1 during training. My work, SIREN [31] (NeurIPS'22) and a subsequent ICML'24 paper [32], fundamentally investigated the influence of *compact representation space* and *ID label supervision* on identifying OOD samples. These insights contribute to designing better training strategies on ID data and enhancing overall model reliability.

2 Algorithm and Theory for Unknown-Aware Learning in the Wild

Motivation. Previous research utilizing auxiliary outlier datasets has shown promise in improving OOD detection compared to models trained solely on ID data. These models often regularize confidence scores [25] or energy levels [33] for outlier data. However, this approach faces two significant limitations: (1) auxiliary

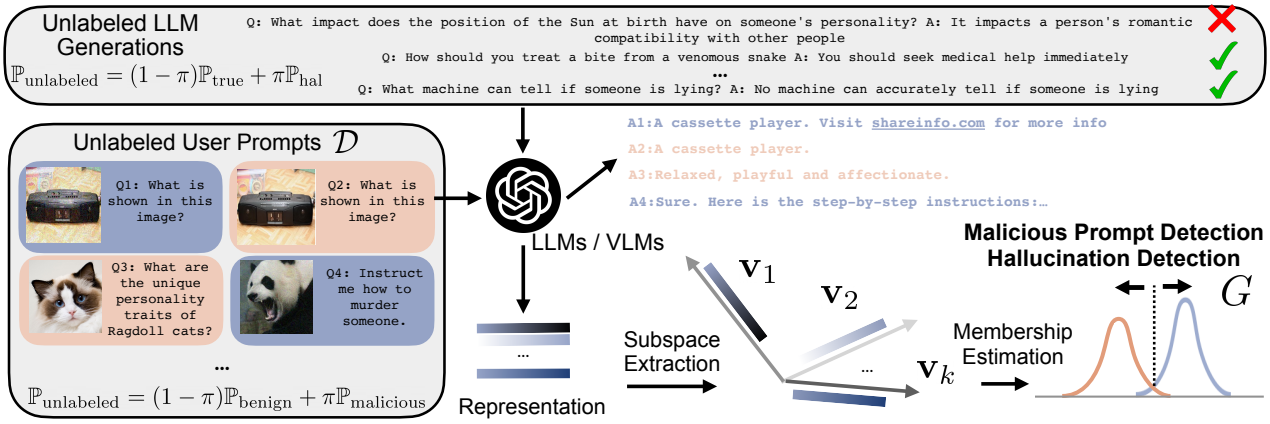


Figure 2: Proposed algorithmic frameworks for hallucination detection in LLMs and malicious prompt detection in VLMs.

data collected offline may not accurately reflect the true distribution of unknown data encountered in real-world settings, potentially undermining OOD detection during deployment; (2) collecting such data is often labor-intensive and requires careful cleaning to avoid overlap with ID data.

My research addresses these challenges by **building novel learning algorithms and theories that leverage unlabeled "in-the-wild" data**, which can be gathered at minimal cost during the deployment of machine learning models. This type of data has been largely overlooked in OOD learning contexts. Formally, unlabeled data can be represented by a Huber contamination model, $\mathbb{P}_{\text{unlabeled}} := (1-\pi)\mathbb{P}_{\text{in}} + \pi\mathbb{P}_{\text{out}}$, where $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{out}}$ represent the marginal distributions of ID and OOD data, respectively. Unlabeled data is abundant, does not require human annotation, and often better aligns with true test-time distributions compared to offline-collected data. While this setting is promising for various applications, it introduces unique challenges due to the impure nature of the unlabeled data which comprises both ID and OOD samples.

Theoretical foundations of learning with wild data. In my ICLR'24 paper SAL [8], I conducted the first formal investigation on *when and how unlabeled data can enhance OOD detection*. My contributions are: **(1)** I proposed a novel learning framework that separates candidate outliers using singular value decomposition on the model gradient matrix, which facilitates the learning of an OOD classifier. **(2)** The framework provides theoretical support for the filtering error and the generalization error of the OOD classifier, proving that these errors can be small under specific conditions. **(3)** Empirically, I demonstrated the generalization bounds of SAL translate into strong empirical performance, establishing state-of-the-art results through extensive evaluations.

Learning with diverse data shifts in the wild. Beyond detecting samples with semantic shifts, machine learning models may also encounter covariate shifts—variations in input distributions that do not necessarily affect labels. These shifts can arise from differences in sensor calibration or environmental changes. My ICML'23 paper, SCONE [10], along with a recent preprint [9] were **the first** to explore techniques for modeling diverse mixtures of data shifts, expressed as $\mathbb{P}_{\text{unlabeled}} := (1-\pi_c-\pi_s)\mathbb{P}_{\text{in}} + \pi_c\mathbb{P}_{\text{out}}^{\text{covariate}} + \pi_s\mathbb{P}_{\text{out}}^{\text{semantic}}$. This framework integrates ID data and various OOD distributions, and employs *constrained optimization* and *active learning* to effectively learn with these diverse data sources for both OOD detection and OOD generalization. The formulation and methodologies presented offer strong generality and practicality for real-world applications.

3 Towards Reliable Foundation Models

Motivation. As foundation models become influential in various applications, ensuring their reliability is an urgent research challenge. These models often face reliability risks, such as generating hallucinations, misinterpreting malicious prompts and handling noisy alignment data, raising concerns for safety and reliability when deployed in real-world settings. Given the models' large scale and their training on massive, diverse data, addressing these issues requires innovative strategies beyond the conventional learning methods. My research seeks to address these issues by identifying the origins of reliability risks $\mathcal{R}_{\text{reliability}}$ in FMs and designing innovative mitigation algorithms.

Safeguarding LLMs against hallucinated generations. In my NeurIPS'24 **spotlight** paper, HaloScope [11], I introduced a novel framework for detecting hallucinations in LLM outputs. *The framework solves the primary*

challenge of hallucination detection, i.e., lack of large annotated data, by using unlabeled LLM-generated text, which inherently contains a mix of truth and hallucinations (Figure 2 upper). By leveraging the representation space, HaloScope extracts a hallucination subspace to facilitate the training of a binary truthfulness classifier. This method demonstrates adaptability across various LLM architectures and domains, establishing a foundation for harnessing unlabeled LLM outputs to enhance FM reliability.

Red-teaming VLMs from adversarial inputs. My research [12] took the lead in identifying input vulnerabilities in vision language models (VLMs), which used the naturally occurring, unlabeled user prompts (Figure 2 left) for help. By analyzing these prompts in the representation space, I developed techniques to detect and counteract malicious inputs, which enhances VLMs' robustness against prompt injection and jailbreak attacks. *This work is pioneering in demonstrating how red-teaming can be applied to VLMs to improve their reliability.*

Aligning AI with human values by data denoising. While AI alignment research typically assumes human feedback is reliable, my recent work [13] reveals that over 25% of feedback data can be inconsistent, highlighting inherent unreliability from biases and labeling errors. To address this, I proposed the Source-Aware Cleaning method, which significantly improves data quality. The cleaner data enables training more reliably aligned LLMs, and thus offers a pathway to more reliable LLM alignment research.

4 Future Research

My research has tackled several foundational reliability challenges arising from the ML paradigm shift, yet significant work remains. Moving forward, I am eager to propel the field of reliable ML by exploring several pivotal directions that will strengthen both theoretical foundations and practical applications of ML systems.

Comprehensive investigation of ML reliability challenges. Developing genuinely reliable ML systems necessitates a deep understanding of the limitations and risks across diverse deployment scenarios. I plan to systematically investigate the safety and reliability challenges inherent in current ML models. For instance, foundation models encounter risks in different stages, such as noisy pretraining data, label ambiguity and data fairness in supervised fine-tuning, preference inconsistencies in RLHF, and vulnerabilities to adversarial attacks during inference. By rigorously characterizing these challenges, I aim to *establish comprehensive benchmarks and devise targeted strategies to enhance model reliability*. This work will improve the safety of ML algorithms across various applications while opening new avenues of research to enrich the reliable ML community.

Development of adaptable and generalized algorithms for reliable ML. Beyond my current work that focuses on foundational reliability learning dynamics with minimal human supervision, I intend to expand the development of reliable ML methodologies from three **core** perspectives: *data, representation, and training/inference algorithms*. For example, I will investigate (1) the impact of diverse data structures, such as human feedback, weak supervision, and semi-supervised data, on model reliability (data perspective); (2) how advancements in representation learning and model architecture can fundamentally enhance reliability (representation perspective); and (3) the design of distributionally robust training methods alongside calibrated, efficient inference algorithms that can adapt to varying deployment environments and integrate multiple knowledge sources (algorithm perspective). These principles will serve as a flexible foundation for improving reliability across a wide spectrum of machine learning models and applications.

Reliable ML for boarder scientific discovery. My long-term vision also includes *harnessing the power of reliable ML to accelerate scientific discovery across multiple domains*, such as reliable biometrics with distribution shifts [34, 35, 36], less human annotations [37, 38], multimodal fusion [39], and reliable protein structural analysis with OOD detection [40], active learning [41], semi-supervised learning [42], and open-set classification [43], where I have expertise in. Moreover, I look forward to collaborating with domain experts in other disciplines, such as chemistry, sociology, and environmental science, to explore how reliable ML can address their unique challenges. By leveraging interdisciplinary knowledge, I aim to develop tailored ML solutions that not only enhance predictive accuracy but also improve the interpretability and trustworthiness of models in complex, data-driven environments. This collaborative effort will ultimately contribute to more robust scientific methodologies and facilitate breakthroughs that are essential for addressing pressing global issues.

Overall, my research approach has been to leverage theories and insights drawn from ML and data analytics to address fundamentally new reliability problems arising from the real world. As a future faculty member, I aspire to maintain this principled approach and advance this compelling research agenda with a vision for impactful contributions to both the academic community and society at large.

References

[1] Computing research association. artificial intelligence roadmap. https://cra.org/ccc/ai-roadmap-self-aware-learning.

[2] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[3] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[4] **Xuefeng Du**, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

[5] **Xuefeng Du**, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[6] Leitian Tao, **Xuefeng Du**, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *International Conference on Learning Representations*, 2023.

[7] **Xuefeng Du**, Yiyou Sun, Xiaojin Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

[8] **Xuefeng Du**, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? In *International Conference on Learning Representations*, 2024.

[9] Haoyue Bai, **Xuefeng Du**, Katie Rainey, Shibin Parameswaran, and Yixuan Li. Out-of-distribution learning with human feedback. *arXiv preprint arXiv:2408.07772*, 2024.

[10] Haoyue Bai, Gregory Canal, **Xuefeng Du**, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, 2023.

[11] **Xuefeng Du**, Chaowei Xiao, and Yixuan Li. Haloscope: Harnessing unlabeled llm generations for hallucination detection. In *Advances in Neural Information Processing Systems*, 2024.

[12] **Xuefeng Du**, Reshmi Ghosh, Robert Sim, Ahmed Salem, Vitor Carvalho, Emily Lawton, Yixuan Li, and Jack W. Stokes. Vlmguard: Defending vlms against malicious prompts via unlabeled data. *arXiv preprint arXiv:2410.00296*, 2024.

[13] Min-Hsuan Yeh, Leitian Tao, Jeffrey Wang, **Xuefeng Du**, and Yixuan Li. How reliable is human feedback for aligning large language models? *arXiv preprint arXiv:2410.01957*, 2024.

[14] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, **Xuefeng Du**, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2022.

[15] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, **Xuefeng Du**, Kaiyang Zhou, Wayne Zhang, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

[16] Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. *arXiv preprint arXiv:2312.11536*, 2023.

[17] Vivek Narayanaswamy, Yamen Mubarka, Rushil Anirudh, Deepta Rajan, and Jayaraman J Thiagarajan. Exploring inlier and outlier specification for improved medical ood detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4589–4598, 2023.

[18] Christos Constantinou, Georgios Ioannides, Aman Chadha, Aaron Elkins, and Edwin Simpson. Out-of-distribution detection with attention head masking for multimodal document classification. *arXiv preprint arXiv:2408.11237*, 2024.

[19] Yuhang Zang, Hanlin Goh, Josh Susskind, and Chen Huang. Overcoming the pitfalls of vision-language model finetuning for ood generalization. *arXiv preprint arXiv:2401.15914*, 2024.

[20] Jiuxiang Gu, Yifei Ming, Yi Zhou, Jason Kuen, Vlad Morariu, Handong Zhao, Ruiyi Zhang, Nikolaos Barmpalios, Anqi Liu, Yixuan Li, et al. A critical analysis of document out-of-distribution detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4973–4999, 2023.

[21] Matthew Inkawhich, Nathan Inkawhich, Hai Li, and Yiran Chen. Tunable hybrid proposal networks for the open world. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1988–1999, 2024.

[22] Soroush Seifi, Daniel Olmeda Reino, Nikolay Chumerin, and Rahaf Aljundi. Ood aware supervised contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1956–1966, 2024.

[23] Suhee Yoon, Sanghyu Yoon, Hankook Lee, Ye Seul Sim, Sungik Choi, Kyungeun Lee, Hye-Seung Cho, and Woohyung Lim. Diffusion based semantic outlier generation via nuisance awareness for out-of-distribution detection. *arXiv preprint arXiv:2408.14841*, 2024.

[24] Hao Lang, Yinhe Zheng, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. Estimating soft labels for out-of-domain intent detection. *arXiv preprint arXiv:2211.05561*, 2022.

[25] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[26] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755, 2014.

[28] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

[29] Soobin Um and Jong Chul Ye. Self-guided generation of minority samples using diffusion models. *arXiv preprint arXiv:2407.11555*, 2024.

[30] Jiahui Liu, Xin Wen, Shizhen Zhao, Yingxian Chen, and Xiaojuan Qi. Can ood object detectors learn from foundation models? *arXiv preprint arXiv:2409.05162*, 2024.

[31] **Xuefeng Du**, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *Advances in Neural Information Processing Systems*, 2022.

[32] **Xuefeng Du**, Yiyou Sun, and Yixuan Li. When and how does in-distribution label help out-of-distribution detection? In *International Conference on Machine Learning*, 2024.

[33] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

[34] **Xuefeng Du**, Dexing Zhong, and Huikai Shao. Cross-domain palmprint recognition via regularized adversarial domain adaptive hashing. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2372–2385, 2020.

[35] **Xuefeng Du**, Dexing Zhong, and Huikai Shao. Continual palmprint recognition without forgetting. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1158–1162. IEEE, 2019.

[36] Huikai Shao, Dexing Zhong, and **Xuefeng Du**. Cross-domain palmprint recognition based on transfer convolutional autoencoder. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1153–1157. IEEE, 2019.

[37] **Xuefeng Du**, Dexing Zhong, and Pengna Li. Low-shot palmprint recognition based on meta-siamese network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 79–84. IEEE, 2019.

[38] **Xuefeng Du**, Dexing Zhong, and Huikai Shao. Building an active palmprint recognition system. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1685–1689. IEEE, 2019.

[39] Dexing Zhong, Huikai Shao, and **Xuefeng Du**. A hand-based multi-biometrics via deep hashing network and biometric graph matching. *IEEE Transactions on Information Forensics and Security*, 14(12):3140–3150, 2019.

[40] Bojun Liu, Jordan G Boysen, Ilona Christy Unarta, **Xuefeng Du**, Yixuan Li, and Xuhui Huang. Exploring transition states of protein conformational changes via out-of-distribution detection in the hyperspherical latent space. 2024.

[41] **Xuefeng Du**, Haohan Wang, Zhenxi Zhu, Xiangrui Zeng, Yi-Wei Chang, Jing Zhang, Eric Xing, and Min Xu. Active learning to classify macromolecular structures in situ for less supervision in cryo-electron tomography. *Bioinformatics*, 37(16):2340–2346, 2021.

[42] Siyuan Liu, **Xuefeng Du**, Rong Xi, Fuya Xu, Xiangrui Zeng, Bo Zhou, and Min Xu. Semi-supervised macromolecule structural classification in cellular electron cryo-tomograms using 3d autoencoding classifier. In *BMVC*, volume 30, 2019.

[43] **Xuefeng Du**, Xiangrui Zeng, Bo Zhou, Alex Singh, and Min Xu. Open-set recognition of unseen macromolecules in cellular electron cryo-tomograms by soft large margin centralized cosine loss. In *BMVC*, page 148, 2019.