Manli Shu

College Park, MD 20740 • (240)714-2447 • manlis@umd.edu • azshue.github.io

Education

University of Maryland, College Park — *Expected: 12/2023*
Ph.D. candidate in Computer Science, Department of Computer Science — *GPA: 4.0*

University of Science and Technology of China — *09/2015 – 07/2019*
B.Eng. in Information Security, School of Information Science & Technology — *GPA: 3.8*

Technical Skills

- **Coding/Programming**: Python (PyTorch, TensorFlow, JAX), Go (gRPC), SQL, C/C++.
- **Software and Tools**: Git, Docker, GCP, OpenCV, Open3D, LaTeX, MySQL.
- **Machine Learning (AI/ML)**: Trustworthy AI/ML, Vision-language models, Large language models, Generative diffusion models, Out-of-distribution generalization, Self-supervised learning.

Work Experience

Google, Research Intern — *06/2023 - Present*
Mountain View, CA

○ **Text-to-Image diffusion models**: for face anti-spoofing.

• Conceptualized a solution for face anti-spoofing using generative diffusion models.
• Designed proof-of-concept experiments on the application of face anti-spoofing.
• Validated the proposal with its zero-shot performance comparable to supervised baselines, with the additional benefit of providing explainable visualizations from the diffusion model.

Salesforce, Research Intern — *06/2022 - 11/2022*
Palo Alto, CA

○ **3D transformer detectors**: balancing the performance across object scales.

• Identified that existing 3D transformer detectors are biased against small objects.
• Proposed novel attention mechanisms to improve the precision of 3D object detectors.
• Improved the state-of-the-art transformer-based 3D detector on indoor benchmarks with an overall increase of 2.0% in mAP and 3.5% relative improvements on small objects.

Nvidia, Research Intern — *01/2022 - 05/2022*
Remote, U.S.

○ **Zero-shot generalization in vision-language models**: via prompt tuning.

• Established a new way of prompt tuning without downstream data or annotations.
• Developed a self-supervised optimization objective for prompt tuning on a single test sample.
• Refined the test-time optimization algorithm with a confidence calibration step.
• Increased the out-of-distribution accuracy of a pre-trained vision-language model by 5.6%.

Research Experience

UMD Center for Machine Learning, Graduate Research Assistant — *08/2019 - Present*
College Park, MD

○ **Instruction tuning for large language models**: from a safety perspective.

• Investigated how training data affects the LLMs' behavior from an adversarial perspective.
• Prototyped an automated and versatile data poisoning pipeline that allows adversaries to use LLMs to generate low-perplexity poison training examples with customized contexts.
• Demonstrated that the proposed attack is 8 times more effective than template-based baselines. Revealed a new type of safety concern for the responsible deployment of LLMs.

○ **Explainability for vision and language models**: a saliency tool on model parameters.

• Examined the correlation between parameter gradients and their effects on model outputs.
• Developed a generic tool for profiling parameter saliency on data samples using their gradients.
• Applied the saliency tool to image classifiers and a language model (BERT). The saliency profile can effectively retrieve samples on which the model made semantically similar mistakes.

○ **Robustness to distribution shifts**: an adversarial approach for domain generalization.

• Proposed adversarial batch normalization for simulation of novel feature distributions.
• Visualized the novel feature distribution in image space, validating the effect of the method.
• Evaluated the method on image classification and semantic segmentation and achieved consistent improvement on over ten domains with a maximum of 9.0% performance boost.

Selected Publications and Preprints

[1] **M. Shu**, J. Wang, C. Zhu, J. Geiping, C. Xiao, T. Goldstein. On the Exploitability of Instruction Tuning. In *Conference on Neural Information Processing Systems (NeurIPS), 2023.*

[2] N. Jain, K. Saifullah, Y. Wen, J. Kirchenbauer, **M. Shu**, A. Saha, M. Goldblum, J. Geiping, T. Goldstein. Bring Your Own Data! Self-Supervised Evaluation of Large Language Models. (*Preprint. Under review.*)

[3] J. Kirchenbauer, J. Geiping, Y. Wen, **M. Shu**, K. Saifullah, K. Kong, K. Fernando, A. Saha, M. Goldblum, T. Goldstein. On the Reliability of Watermarks for Large Language Models. (*Preprint. Under review.*)

[4] **M. Shu**, L. Xue, N. Yu, R. Martín-Martín, C. Xiong, T. Goldstein, J. C. Niebles, R. Xu. Model-Agnostic Hierarchical Attention for 3D Object Detection. (*Preprint. Under review.*)

[5] **M. Shu**, W. Nie, D. Huang, Z. Yu, T. Goldstein, A. Anandkumar, C. Xiao. Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models. In *Conference on Neural Information Processing Systems(NeurIPS), 2022.*

[6] A. Ghiasi, H. Kazemi, E. Borgnia, S. Reich, **M. Shu**, M. Goldblum, A. G. Wilson, T. Goldstein. What Do Vision Transformers Learn? A Visual Exploration. (*Preprint. Under review.*)

[7] R. Levin*, **M. Shu***, E. Borgnia*, F. Huang, M. Goldblum, T. Goldstein. Where do models go wrong? Parameter-space saliency maps for explainability. In *Conference on Neural Information Processing Systems (NeurIPS), 2022.*

[8] R. Ni*, **M. Shu***, H. Souri, M. Goldblum, T. Goldstein. The Close Relationship between Contrastive Learning and Meta Learning. In *International Conferences on Learning Representations (ICLR), 2022.*

[9] **M. Shu**, Z. Wu, M. Goldblum, T. Goldstein. Encoding Robustness to Image Style via Adversarial Feature Perturbations. In *Conference on Neural Information Processing Systems (NeurIPS), 2021.*

[10] Y. Shen, L. Zheng, **M. Shu**, W. Li, T. Goldstein, M. Lin. Gradient-Free Adversarial Training against Image Corruption for Learning-based Steering. In *Conference on Neural Information Processing Systems (NeurIPS), 2021.*

[11] **M. Shu**, Y. Shen, M. Lin, T. Goldstein. Adversarial Differentiable Data Augmentation for Autonomous Systems In *International Conferences on Robotics and Automation (ICRA), 2021.*

[12] C. Zhu, Z. Xu, A. Shafahi, **M. Shu**, A. Ghiasi, and T. Goldstein. Towards Accurate Quantization and Pruning via Data-free Knowledge Transfer. *Sparsity in Neural Networks (SNN) Workshop, 2021.*

[13] A. Abdelkader, M. Curry, L. Fowl, T. Goldstein, A. Schwarzschild, **M. Shu**, C. Studer, C. Zhu. Headless Horseman: Adversarial Attacks on Transfer Learning Models. In *ICASSP, 2020.*