Shuhang Tan

+1 (518) 203-8249 | shuhangtan962@gmail.com | Homepage | Google Scholar

Education

Rensselaer Polytechnic Institute (RPI) — Troy, NY, USA

Ph.D. in Computer Science (Expected April 2028) — *Aug. 2023 – Present*

- Advisor: Radoslav Ivanov
- **Research Interests:** Efficient LLMs, model compression and distillation, explainable AI, neuro-symbolic AI.

University of Science and Technology of China (USTC) — Hefei, Anhui, China

Master of Computer Technology — *Sept. 2020 – June 2023*

- Research Areas: Autonomous vehicle perception and driving risk analysis.

China University of Mining and Technology — Xuzhou, Jiangsu, China

Bachelor of Electrical Engineering and Automation — *Sept. 2016 – June 2020*

University of Wisconsin–Madison — Madison, WI, USA

Exchange Program in Electrical Engineering — *Aug. 2017 – May 2018*

Technical Skills

Programming: Python, C/C++, MATLAB, Java, SQL, Bash, TypeScript

Machine Learning / LLMs: PyTorch, Hugging Face Transformers, lm-evaluation-harness, verl, scikit-learn

LLM Research: Model pruning, knowledge distillation, SFT, on-policy distillation, LLM evaluation

Robotics / Perception: ROS, OpenCV, OpenPCDet, LiDAR

Systems & Tools: Linux, Git, Docker, CUDA, CMake

Research Experience

Research Assistant, Rensselaer Polytechnic Institute — May 2024 – Present

- **Distillation-Aware Pruning for LLMs**
  - Formulated LLM pruning as **post-distillation student design**, showing that minimizing immediate pruning damage does not necessarily produce the strongest student after distillation.
  - Developed **Distillation-Aware Pruning (DAP)**, a second-order pruning framework that estimates irrecoverable damage by jointly modeling immediate pruning damage and recoverability under the intended teacher and distillation procedure.
  - Evaluated DAP on **Qwen3-8B** and **DeepSeek-R1-Distill-Llama-8B** under supervised fine-tuning (SFT) and on-policy distillation (OPD), improving average post-distillation performance by up to **3.9 percentage points** over the strongest pruning baseline under matched sparsity and training budgets.
- **Post-Training Neural Network Pruning Using Graph Curvature** (GitHub)
  - Developed an interpretable post-training pruning method using graph Ricci curvature to identify critical neural-network information flows and structurally redundant connections, and evaluated the method on CNN and VGG-style architectures across MNIST, CIFAR-10, and CIFAR-100.
  - Accelerated curvature computation by implementing the shortest-path stage on GPU using dynamic programming and matrix-based operations, improving the scalability of graph-based neural-network analysis.
  - Extended the graph-curvature framework beyond the published vision-model experiments to **Llama 3** and **IBM Granite** LLMs by constructing graph representations of transformer information flow for large-scale model analysis and pruning.

- **Analyzing Neural Network Robustness Using Graph Curvature**
  - Introduced neural data graphs and neural Ricci curvature to identify structural bottlenecks associated with neural-network robustness, showing that less robust predictions exhibit stronger bottleneck structures.

Research Assistant, University of Science and Technology of China September 2020 – June 2023

- **Autonomous Driving Risk Assessment** (GitHub)
  - Developed a driving safety field model that jointly represents road geometry, lane boundaries, vehicles, and pedestrians for autonomous-driving risk estimation.
  - Proposed a cosine-similarity-based method for estimating relative motion between the ego vehicle and surrounding obstacles.
- **LiDAR-Based 3D Object Detection for Autonomous Vehicles** (GitHub)
  - Designed a LiDAR-only 3D object detection architecture that combines voxel- and range-view representations to improve small-object detection.
  - Developed a hybrid architecture inspired by Voxel-RCNN, CenterPoint, and RangeRCNN to fuse complementary geometric representations from LiDAR point clouds.

Selected Publications

Prune What the Teacher Can Repair: Distillation-Aware Pruning for LLMs *Under review at ICLR 2027*

Post-Training Neural Network Pruning Using Graph Curvature *TMLR, 2026*

Analyzing Neural Network Robustness Using Graph Curvature *ICAA, 2024*

RCP-RF: A Comprehensive Road-Car-Pedestrian Risk Management Framework Based on Driving Risk Potential Field *IET Intelligent Transport Systems, 2024*

Industry Experience

Perception Research Engineer, Shineauto (Startup–Lab Collaboration) Feb. 2022 – Jan. 2023

- Developed and deployed C++/Python LiDAR perception pipelines for autonomous driving, including point-cloud clustering and 3D bounding-box estimation for real-time inference.
- Integrated 3D non-maximum suppression (NMS) into a Scanline clustering pipeline and deployed Voxel-RCNN on a real autonomous vehicle using ROS for on-road data collection and system testing.

Technology Intern, Morgan Stanley, Shanghai, China July 2022 – Sept. 2022

- Developed an internal trade-data management platform using Angular, TypeScript, Java, and SQL to support data querying, validation, revision, and record management.
- Built frontend tools for data filtering, inspection, and ID-mapping visualization, along with backend modules for database access and validation.

Service

Academic Service

- Reviewer: *IET Intelligent Transport Systems*, *IEEE Transactions on Intelligent Vehicles*, *Artificial Intelligence and Applications (AIA)*, *AAAI (2026, 2027)*, *ICCPS 2026*, *Pattern Recognition*, *ICLR 2027*

Teaching Assistant

- *Data Structures* (Fall 2023), *Foundations of Computer Science* (Spring 2024, Spring 2026), *Advanced Operating Systems* (Fall 2026)