Shilong Lei

+1 765 775 7952 ◇ lei105@purdue.edu ◇ West Lafayette, INDIANA
LinkedIn ◇ Homepage

EDUCATION

Purdue University, Department of Computer Science — West Lafayette, United States
Ph.D student in Computer Science — Jul 2022 - Present

Tsinghua University, School of Information Science and Technology — Beijing, China
Bachelor of Engineering in Automation — Aug 2018 – Jul 2022

Awards:

* Technological Innovation Scholarship & Hage Foundation Scholarship of Tsinghua University — 2019 - 2021
* MobiCom'24 Best Poster Award — Sep 2024

PUBLICATIONS

(MobiCom'24) *"D-AirPatrol: A Dual-Layer Architecture for Traffic Patrol From the Sky."* Jiaxin Du, Shilong Lei, Chunyi Peng. **Best Poster Award**. Sep 2024

(CONF-CDS 2021) *"A Spatial-Temporal Adaptive Video Denoising Algorithm."* Shilong Lei. *Computing and Data Science: Third International Conference, Virtual Event, Aug 12–17, 2021, Proceedings 3.* Springer Singapore. May 2021

INTERNSHIP

Meta Platforms – Ads Training Infra — Sunnyvale, CA
Research Scientist Intern in Distributed ML Systems — *Mentor: Shuai Yang, May 2025 – Aug 2025*

· Collaborated with the FAIR PyTorch/FSDP team to co-design and implement an **Automatic Selective Unsharding Framework** atop **SimpleFSDP** (a PyTorch 2.0 compile-friendly FSDP), improving communication–computation overlap and training efficiency.
· Built a **hierarchical memory profiler** and adaptive unsharding algorithm leveraging **GPU memory budget** for selective parameter retention, achieving up to **5.5% QPS improvement** on 8×GPU ads-model benchmarks.
· Enhanced **SimpleFSDP infrastructure** with configurable communication scheduling and integrated memory–computation optimization, enabling future **adaptive distributed-training pipelines** within Meta Ads Infra.

ByteDance - Seed — Bellevue, WA
Research Scientist Intern in LLM infrastructure — *Mentor: Yanghua Peng, May 2024 - Sep 2024*

· Supported a **LLM training simulator** to profile and simulate the training latency of LLM (especially **diffusion transformer models**) on **1k - 10k** GPUs for text to video generation and customized computing kernels to achieve **95%+** accuracy.
· Supported memory profiling and distributed training simulation in frameworks like **pytorch.distributed** and **deepspeed-megatron** into the LLM training simulator and achieve **95%** accuracy.
· Supported PyTorch.FSDP & various **parallelism** techniques in Bytedance opensource distributed machine learning framework **veScale** to enhance LLM training efficiency to support Doubao large language model.

RESEARCH EXPERIENCE & PROJECTS

OOGC: Out-of-GPU-Core LLM Training System — Purdue University
Researcher & Full-stack engineer — *Advisor: Dr.Xuehai Qian, May 2023 - Dec 2023*

· Led the project with two interns to develop an Efficient **Out-of-GPU-Core** LLM training system based on **DeepSpeed ZeRO** and **Megatron-LM** for distributed LLM training.
· Fully awared of the LLMs' execution pattern and hardware resources, the system runs over-sized models by **CPU offloading** and speeds up training by better utilizing **GPU memory** and reducing **CPU-to-GPU communication**.
· With novel multi-layer **prefetching**, memory management system and pipelined optimizer stepping and optimized activation checkpointing, the system yields a **20%** training speed improvement and **15%** in MoE model training compared with baselines.

CPU Model Inference System — Purdue University
Researcher & Full-stack engineer — *Advisor: Dr.Xuehai Qian, Oct 2022 - Apr 2023*

· Designed and implemented a **sparse-aware CPU inference system** for deep neural networks, inspired by **DeepSparse**-style architectures, optimizing execution for large-core CPU clusters.
· Proposed a novel **depth-direction forwarding** strategy that maximizes on-chip cache reuse and reduces memory bandwidth pressure, achieving substantial speedups over layer-by-layer scheduling.
· Developed kernel-level optimizations in **C++** and **PyTorch backend**, enabling efficient sparse tensor operations and low-latency inference for transformer-style workloads.
· Strengthened expertise in **LLM inference infrastructure**, covering operator fusion, memory locality optimization, and CPU parallelism for model serving.

Drone-ROCKET: a Visual-Prompted Multi-Step Framework for Agentic Drones — Purdue University
Researcher — *Advisor: Dr.Chunyi Peng, Sep 2025 - Dec 2025*

· Designed and implemented a **novel embodied agent system** for drones, extending visual-temporal prompting from simulated environments to **real-world aerial platforms** for multi-step drone interactions.

- Built an end-to-end **agentic drone pipeline** integrating real-time video streaming, visual prompt construction, VLM-based step-wise planning, and closed-loop action execution.
- Designed and prototyped a **temporal context and visual prompt injection mechanism** with an **interaction-conditioned policy**, to support **fine-grained, multi-step drone interactions** with interpretable and compositional behaviors.
- The system improves interaction task success rate by **20%** over single-step OpenVLA and non-prompted baselines.

D-AirPatrol: A Dual-Layer System and Data Platform for Aerial Traffic Monitoring Purdue University
Researcher & Full-stack engineer *Advisor: Dr.Chunyi Peng, Jan 2024 - Nov 2024*

- Proposed a novel dual-layer architecture to decouple the foreground and background of aerial views to make drones air patrol for traffic monitoring more accurate and reliable. Developed an **Android App** to track cars and estimate car speed in real-time from DJI drone view. Built a public **Django** website with user & tasks management for data collection.
- Built a full-stack **WebRTC** streaming system to stream video from drones to edge server, and stream processed frames to client web browser. Presented a **demo** to view real-time speed monitoring video and results from a website.
- Gained proficiency in Python, Pytorch, node.js, Java and LaTeX, and developed expertise in optimization and testing methods.
- D-AirPatrol substantially outperforms the baseline, improving MOTA from 41% to **96.5%**. Poster accepted by **MobiCom 2024** (**Best Poster Award**).

Video Classification & Video Denoising Algorithms Los Angeles, CA & Beijing, China
Student Researcher *Advisors: Dr. Ram Nevatia, Dr. Guoqing Xiang, Jan 2021 – Sep 2021*

- Independently designed and implemented a **teacher–student video classification framework** leveraging **temporally consistent spatial augmentation** with **PyTorch** and **OpenMMLab's MMAction2**, improving cross-frame feature stability in long-horizon videos.
- Proposed a novel **spatio-temporal adaptive video denoising algorithm** by integrating space-time adaptive processing into scene-change–aware filtering, effectively overcoming temporal failure modes of conventional time-domain methods.
- Led the entire research lifecycle from conception, implementation, and experimental validation to **first-author publication** of "*A Spatio-Temporal Adaptive Video Denoising Algorithm*".

Multi-Camera 3D Pedestrian Detection & 3D Human Pose Reconstruction Tsinghua University
Student Researcher *Advisor: Dr.Jianjiang Feng, Oct 2021 - Jun 2022*

- Independently proposed and implemented a **multi-camera 3D pedestrian localization framework** for large-scale public spaces, integrating **probabilistic multi-view fusion**, **clustering**, and **CNN-based feature aggregation** to precisely estimate each individual's global position across overlapping camera views.
- Designed a **Bayesian consistency optimization module** to reconcile inter-camera depth ambiguity and improve cross-view association robustness under occlusion and perspective distortion.
- Achieved superior localization accuracy compared to CVPR multi-camera baselines on in-the-wild datasets, demonstrating substantial gains in both precision and recall.
- Further extended the pipeline to **3D human pose reconstruction** from multi-person videos, enabling high-fidelity body pose recovery under dynamic social interactions.

SKILLS

Languages: Python, C/C++, CUDA, MATLAB, Java, JavaScript, HTML, MIPS, LaTeX
Technologies: PyTorch, LLM, DeepSpeed, Megatron, Distributed Machine Learning, Ubuntu (Linux), Docker, Django, Node.js, MySQL, Qt, Wireshark, Verilog