

SHIVAM PANDEY

≈ 2 years of research and internship experience in AI Research, and Software Engineering

 shivampr21.github.io

 shivampr21

 +91-7974326386

 Kanpur, UP, India

 shivampr21@gmail.com



EXPERIENCE

Computer Vision Engineer

Quidich Innovation Labs

Apr'24–Present

Foundation Models Transformers ViT VLM Python CUDA ZeroMQ

IPC GDS FAISS zarr Docker ffmpeg PyTorch C++

- Innovating AI for the domain of sports broadcast & analytics.
- Development and training of transformer models for interaction modeling and state forecasting, for on-ground real-time deployment.
- Responsible for developing a Multi-Camera real-time Player Tracking and Fusion system from scratch, with model quantization and compilation.
- Developed GPU-direct (GDS) based media storage and streaming pipeline for end-to-end low latency and high throughput system for large streaming data processing.
- Built monocular planar transform tracer with GPU-accelerated optical flow and re-localization and matching through DNN, and fast embedding search, sustaining $\geq 300FPS$ on RTX 4090 systems.

AI Research Engineer

Manifest AI

Feb'24–Apr'24

LLM JAX Optax Triton XLA CUDA MPI Huggingface zarr

Nsight Compute Nsight Systems GlusterFS

- Responsible for developing highly parallel code (infrastructure, architecture, and kernel) for efficient and effective training and evaluation of Foundation Models.
- Highly parallel implementation including lower-level kernels for transformers to train for larger contexts on multiple GPUs and nodes.
- Trained LLMs - Linear Attention Transformers for context scaling laws on 4x8 H100 GPUs.
- Implemented a LoRA like mechanism for training LLMs to derive the scaling law to compute against context length.
- Implemented both Data and Model Sharding approaches to scale the model training across the GPUs and compute nodes, along with the assessment of communication overhead.
- Implemented compute-communication overlap within the model architecture for latency hiding through multiple CUDA streams.
- Implemented custom reduction and matrix operations to scale across multiple GPUs and nodes for faster training by enforcing computation and communication overlap.
- Processed Red-Pajama-v2 dataset of 30T tokens on GCP, to carve out sequences of large context lengths, and store final dataset in tokenized form efficient usage in context law experiments.

Research Engineer Intern

Five AI (Robert Bosch & BCAI)

Aug'22–Oct'22

Trajectory Prediction GNNs Deep Learning Python C++

- Motion Planning & Prediction Team
- Research work on vehicle trajectory prediction.
- Implementation of **GNN** based trajectory prediction system, with improvements in optimization towards multi-modal goal-set prediction.
- Improved SOTA under the quantitative explanation for training efficiency with end-to-end training mechanism.

EDUCATION AND ACHIEVEMENTS

 Master of Technology

 CPI 10.0/10

Geo-Informatics, IIT Kanpur

2020–Jan 2024

Research Focus: Efficient and Robust Discriminative Manifold Learning and Optimization

Thesis: 3D Multi-Modal Multi-Object Tracking

 Bachelor of Technology

 CPI 7.1/10

Civil Engineering, IIT Kanpur

2017–Jan 2024

 JEE Advanced 2017:

 Gen AIR 3315

Joint Engineering Entrance Exam

2017

 English Proficiency Test

 CEFR Level C1

International Test for English Proficiency

2022

PUBLICATIONS

- KERNELIZED (BLOG Series) (profile page)
- RMS-ICP: Robust Multi-Scale ICP (Paper Link).
- Contrastive Learning & 3D MOT (Paper Link).
- 3D Multi-Modal MOT (MS Thesis Link)

POSITION OF RESPONSIBILITY

Teaching Assistant

Inertial And Multi-Sensor Navigation: CE677B

Concept explanation & conduction of labs.

Teaching Assistant

Geoinformatics: CE331

Responsible for conducting discussion hours.

Event Coordinator

ISSTF Open House, IITK

Responsible for the overall management of the Science Fair event (ISSTF).

TECHNICAL SKILLS

C++ C Python Rust CUDA Triton XLA

MPI OpenMP Eigen LLDB Valgrind LLVM

MLIR NCCL ROS2 OpenCV Open3D Keras

Tensorflow PyTorch JAX TorchRL ONNX WandB

Hydra LLM LangChain OpenAI Gym PCL Docker

Kubernetes Spark Conan CMake Bazel Spacy

NLTK Numpy SciKit-Learn SymPy Seaborn

Pandas Matplotlib Plotly Flask Streamlit SQL

NoSQL Vector DB FAISS HF Transformers HF TRL

HF Diffusers Qdrant MongoDB Pinecone CI/CD

LaTeX AWS GCP Linux Git SLURM