SHIVAM PANDEY

≈ 1.5 years of experience in AI Research, and Software Engineering

shivampr21.github.io | shivampr21 | shivampr21 | shivampr21 | shivampr21
pandeyshivam2023robotics@gmail.com | +91-7974326386 | Kanpur, UP, India

EXPERIENCE

Sr. Computer Vision Engineer

Quidich Innovation Labs — Apr'25–Present

CUDA | IPC | RDMA | RoCEv2 | NVMe-oF | GDS | NCCL | MPI | Docker | PyTorch | Python | C++

- Responsible for designing and developing large-scale model training and inference compute infrastructure for both in-house model training and on-site deployment.
- Implemented **RoCEv2(RDMA)** with **Nvidia Spectrum** Switche & **Connect-X NIC** for online video stream decoding, real-time NVMe-oF storage operations, and network separation to serve AI inference output external to cluster.

Computer Vision Engineer

Quidich Innovation Labs — Apr'24–Apr'25

Foundation Models | Transformers | ViT | VLM | Python | CUDA | ZeroMQ | IPC | RDMA | RoCEv2 | NVMe-oF | GDS | NCCL | FAISS | QDrant | MPI | 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, 3D Pose Estimation, and Fusion system from scratch, including model quantization and compilation implementations.
- Developed **GPU-direct Storage (GDS)**, **NVMe-oF**, and **Remote Direct Memory Access (RDMA): RoCEv2** based multi-node media storage and streaming pipeline for end-to-end low-latency and high-throughput AI training, inference, and intelligence broadcast engines.
- 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.
- Solved the model training problems with **Detection Transformers (DeTr)** with compute scaling making them effective in sports domain and surpassing CNN models like YOLO in FP reduction leveraging larger context window of attention mechanism, while maintaining real-time inference.
- Designed unified online video-saliency detection architecture based on **infini-attention** mechanism solving the constant temporal context constraints, and the output frame delay problem by increasing the context window to near-infinite, while reducing compute requirements.
- Developed person Re-id for global context understanding in cricket for distributed camera system with GNNs and Transformers.
- Developed an LLM agentic system that uses vision-extracted data to generate coherent streaming commentary.

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.

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

12th Board Exam: — **82.2%**
UP Board — Jun'16

10th Board Exam: — **90.0%**
UP Board — Jun'14

PUBLICATIONS

1. RMS-ICP: Robust Multi-Scale ICP (Paper Link).
2. Contrastive Learning & 3D MOT (Paper Link).
3. 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).

INTERESTS

Compute Scaling | Infinite Context | Self-Driving Cars | Reinforcement Learning | AI Interpretability | Embodied AI | AI for Science

- 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.
- Profiling and Debugging of the highly optimized CUDA kernel calls for latency hiding opportunities.
- Developed multiprocessing visualization platform using Streamlit for seamless comparison of results.

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.

Visiting Artificial Intelligence Researcher

DeepKapha AI Labs Feb'22–Jun'22

Deep Learning | Signal Processing | Template Matching | PyTorch

- Worked on deep learning based pattern matching, and segmentation.
- Designed **GAN** with two generators to cope with both FP & FN errors, and a unified discriminator for small-object detection.
- Innovated **Template Matching** in gamma ray signal logs, to find similarity b/w spatially correlated yet different locations, with use of **Generative model** to learn robust embeddings.

SR. Student Research Associate

IIT Kanpur & Science and Engineering Research Board Sep'21–July'22

Computer Vision | Deep Learning | Pytorch | LiDAR | Python

- Designed an expandable 3D multi-modal multi-object tracking system.
- Achieved 50% decrease in ID switching, and MOTA increase by 5% w/ image and point-cloud fusion & **self-supervised representation learning**.
- Improved Contrastive Learning (InfoNCE) loss function for faster convergence with the definition of ideal contrastive loss.
- Defined formal implementation structure of a tracking system for heterogeneous (sensor & track dimensionality) and expandable setup.

Software Engineering Intern

Bosch Global Software Technologies Jun'21–Jan'22

Simulation | CARLA | ROS2 | C++ | Python

- Developed **ROS2** based integration of **Carla** with navigation stack.
- Developed Carla integration for unified **SIL** simulation framework **Cloe** for **simulation** based testing of Autonomous Driving stack.

[1]Online

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

SELECTED COURSES & CERTIFICATES

- Sequence Models[1]
- Reinforcement Learning Specialization[1]
- Machine Learning for Signal Processing
- Machine Learning
- Algorithmic Toolbox and Data Structures[1]
- Machine Processing Of Remotely Sensed Data
- Self-Driving Cars Specialization[1]
- Automatic Control Of Aircraft and Rockets
- Controls for Mobile Robotics[1]
- Reference Frames, Coordinate Systems
- Physical Geodesy
- Environmental Geodesy
- Laser Scanning And Photogrammetry

BLOGS AND OPEN SOURCE

Kernelized: Blog series on Computation for AI

shivampr21.github.io Jan'25-Present

CUDA | C++ | Python | Triton | PTX | NCCL | MPI | Assembly | SIMD | SIMT | Nanobind | uv | Docker | Git

- Blog1: Max Reduction Kernel: Forward and Backward Pass Derivation
- Blog2: SoftMax Kernel: Forward and Backward Pass Derivation
- Blog3: Flash Attention Kernel: The preliminary exploration of Compute.
  1. Forward and Backward Pass Derivation through Tensor Differentiation with Einstein Index Notations.
  2. Analysis of parallelism with the data dependency graph (DDG).
  3. Identification of parallelization constraints through strongly connected components (SCC) identification in DGG.
  4. Loop transformation analysis and tiling opportunities identification in a generalizable manner.