SHIVAM PANDEY

pprox 2 years of research and internship experience in Al Research, and Software Engineering

shivampr21.github.io in shivampr21 G pandeyshivam2023robotics@gmail.com

shivampr21 **)** +91-7974326386 ★ shivampr21 Kanpur, UP, India



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 > 300FPS on RTX 4090 systems.
- Solved the model training problems with Detection Transformers 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 constraint and output frame delay problem 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.

Al Research Engineer

Manifest Al

Feb'24-Apr'24

LLM JAX Optax Triton XLA CUDA MPI Huggingface 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.
- Profiling and Debugging of the highly optimized CUDA kernel calls for latency hiding opportunities.

EDUCATION AND ACHIEVEMENTS

■ Master of Technology

k shivampr21

P 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 Test CEFR Level C1

International Test for English Proficiency

12th Board Exam:

Y 82.2%

UP Board

Jun'16

2022

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

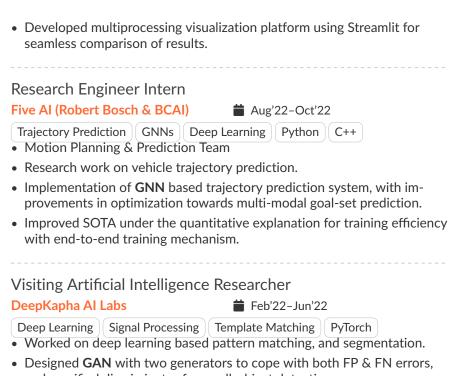
Geoinformatics: CE331

Responsible for conducting discussion hours.

Event Coordinator

ISSTF Open House, IITK

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



- 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.

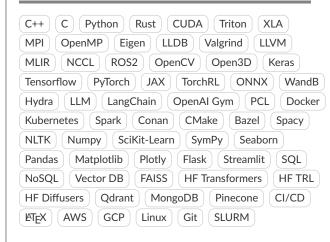
Software Engineering Intern

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.

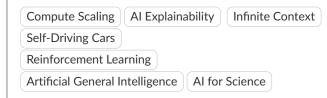
TECHNICAL SKILLS



SELECTED COURSES & CERTIFICATES

- Sequence Models¹
- Reinforcement Learning Specialization¹
- Machine Learning for Signal Processing
- Machine Learning
- Algorithmic Toolbox and Data Structures¹
- Machine Processing Of Remotely Sensed Data
- Self-Driving Cars Specialization¹
- Automatic Control Of Aircraft and Rockets
- Controls for Mobile Robotics¹
- Reference Frames, Coordinate Systems
- Physical Geodesy
- **Environmental Geodesv**
- Laser Scanning And Photogrammetry

INTERESTS



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 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.

Rank1: L1-BLAS implementation in Rust

ShivamPR21 Rust | SIMD

Machine Learning

Nov'23-Jan'24

 Implementation of Basic Linear Algebra Subprogramm 	os (DLAS) lovel 1 in DLIST
MZT: Module Zoo Torch	is (BLA3) level-1 III RO31.
ShivamPR21	■ Sep'21-Present
Deep Learning Pytorch Python Git	Supplies the suppl
A deep learning library for the research community ba	sed on DeepMind's explanation of NNs.
RESEARCH AND DEVELOPMENT PROJECT	'S
Multi-View Action Recognition	
Prof. Pranamesh Chakrabarty	昔 Feb'22-Nov'22
Computer Vision PyTorch Python	
-	n robust classification removal of segmentation requirements.
 Achieved more than 98% training, and 93% test/valida 	ation accuracy, starting from scratch training.
Globally Optimized Point Cloud Registration	
Geoinformatics Lab Prof. Salil Goel	■ May'19-May'21
	ython
 The novelty of the work reduced convergence iteratio art. 	ons by 80% , and execution time by 83.33% as compared to state-of-the-
 Reduced search volume by 98.44%, for faster sub-opt 	imal initialization.
	obust kernel, and persisting gradients for faster convergence rates and
higher reliability.	
Reinforcement Learning based Adaptive-PID	
Automatic Controls Prof. D. K. Giri	★ Aug'20-Feb'21
Reinforcement Learning TensorFlow OpenAI-gym Num	
• Deployed RL for real-world highly non-linear systems	
and racing.	dealing with highly non-linear and noisy systems like drone acrobatics,
 Developed RL-based, gain tuning policy for adaptive P 	PID controls, and trained with more than 20M training steps.
Spatio-Temporal Variations of Greenland Ice Ma	ass
Geoinformatics Lab Prof. D. Balaji	■ Aug'20-Dec'20
Deep Learning CNN-LSTM Fourier Transform TensorF	
	in Greenland, with Sinusoidal Regression and CNN-LSTM models.
	and used de-striping filter for data errors due to sensor failure. Isier consumption and adaptation to other datasets and locations.
Low-light Image Enhancement Using MSR-DCN Summer Project: EE	NN m May'19-July'19
Computer Vision Deep Learning Deep CNN TensorFlo	
	ight with multi-scale retinex method, and learnable kernels , enabled by
Deep-CNN.	· · · · · · · · · · · · · · · · · · ·
 Implemented MSR-DCNN from scratch in tensorflow. learning rate. 	. Experimented with ensemble techniques for faster training with a higher
3D Reconstruction	
Geoinformatics Lab Prof. Salil Goel	■ Jan'19-May'19
Computer Vision OpenCV RANSAC Python	
	g, feature extraction, and used RANSAC for optimal data association.
 Implemented 3D reconstruction for monocular camer scratch with essential and fundamental matrix, includ 	a setup in python with bare minimum support of OpenCV , starting from ing the constraints, decompositions, and optimization.

¹Online