ANSH MITTAL

Mountain View, CA | (213) 573-9188 | manshm1811@gmail.com | linkedin.com/in/mittalansh/ | github.com/AnshMittal1811 | medium.com/@anshm18111996

SUMMARY

5+ yrs of ML experience bridging 3D Computer Vision and LLM optimization, spanning foundational research to enterprise production. Proven track record of shipping optimized spatial pipelines and architecting large-scale agentic AI systems. Skilled in translating cutting-edge prototypes (NeRFs, Gaussian Splats) and fine-tuned Language modelsz into scalable, production-ready workflows for complex computing challenges.

EXPERIENCE

Guidewire Software Inc | *San Mateo, CA* **April 2024–Present**

- **Machine Learning Engineer II - GenAI, Research**
  - Led design and deployment of multi-stage OCR–RAG–LLM pipelines across 5 business lines in the AMER region, coordinating research, experimentation, and productionization for enterprise-scale risk modeling workflows
  - Finetuned and Deployed Qwen-3-4B claim summarization QLoRA adapter with 0.76 score (6% ↑)) (ROUGE-L, BERTScore, Domain-based) based on comparative analysis with GEMMA 3-4B QLoRA adapter and base models
  - Innovated auto-annotation pipeline for community outline segmentation with hazard-based SatelliteMAE Optical/SAR modelling (Avg Dice Coef.: 0.92 (10% ↑)) associated with late fusion for public safety buildings (from news text/images)
  - Developed Agentic AI workflow with tool calling, Chain-of-Thoughts, Tree-of-Thoughts, Graph-of-Thoughts (branched SFT DocLLM, OCR-RAG-LLM, BDA) to reduce False Positive Rates in Year Loss Tables for Cyber Risk Modeling
  - Architected a Digital Twin simulation infrastructure to model complex cyber risks on network infrastructure, analyzing phase shift signals and packet captures to simulate and predict multi-stage attack vectors

Remix Inc | *Palo Alto, CA* **September 2023–December 2023**

- **Senior Computer Vision Engineer (Founding)**
  - Integrated 3D dynamic view reconstruction with LeRF-joint training and human Gaussian-splatting for semantically 3D-aware panoramic systems, achieving 38ms latency in real-time novel scenes
  - Compiled custom Gstreamer-CUDA-OpenCV pipelines across ARM64/AMD64/X86 architectures and optimized NeRF architectures for parallel multi-GPU processing and camera-Jetson-Unity synchronization

KNOW Corp | *Los Angeles, CA* **January 2023–May 2023**

- **AI/ML Research Engineer (Co-op)**
  - Personalized GenerativeQA (GPT-NeoX) using CoT, PEFT LoRA, and Rotary Position Embeddings; compared metric performance (BLEU, ROUGE, STS) and implemented GitHub CI/CD monitoring workflows
  - Prepared PoC for 3D Neural Human Avatars with synchronized SpeechT5 Text-to-Speech and deployed via FastAPI for real-time interactive agents

USC Research Labs, USC ISI & IMSC labs | *Marina Del Rey, CA* **January 2022–July 2024**

- **AI/ML CV Research Engineer & Graduate Assistant**
  - Published a survey paper NeRFs (Neural Radiance Fields): Past, Present, and Future and applied 3D reconstruction for Astronomy, Geospatial, and Medical domains
  - Classified street objects using fine-tuned MobileNet/ViT (83% accuracy (23 ms)), and engineered 3DMM feature regression baselines (62-D) without explicit face landmark detection and I-SPLIT algorithm for edge-devices via modified GradCAM

Sociometrik | *Delhi, India* **October 2020–August 2021**

- **Data Scientist (GIS, Computer Vision)**
  - Led Metal Roof-Detection and Super-Resolution pipelines on AWS (EC2/Lambda) using PyTorch Horovod, achieving 0.698 IoU and 0.87 Precision for multi-label terrain classification
  - Orchestrated car segmentation and feature extraction workflows using Jenkins CI/CD and RESTful APIs, facilitating automated GIS data analysis from S3 buckets

Indian Institutes of Technology Delhi | *Delhi, India* **January 2020–October 2020**

- **Research Scientist (Game, Reinforcement Learning)**
  - Hosted Unity apps for Cybersecurity educational games using seq2seq chatbot & NAF-based actor-critic RL for adaptive NPC
  - Implemented user feedback clustering (76% accuracy) and behavioral-tree systems, increasing average user interactivity by 27% based on action-feedback analytics with Firebase storage

TECHNICAL SKILLS

- **Languages & frameworks**: Python, C++, Kotlin DSL (CI/CD)
- **Libraries**: Pandas, Matplotlib, PySpark, TensorFlow/Keras, PyTorch, ONNX, OpenCV, LangChain, HuggingFace, DBT
- **Development expertise**: AWS (EC2, ECS, S3, EMR, RDS, Redshift, Lambda, Glue, Athena, Sagemaker), Teamcity, Docker, AirFlow, Kubernetes, Terraform/Terragrunt, PostgreSQL
- **ML & GPU Systems Expertise**: NLP (BERT, CLIP, RAG, RLHF), Computer Vision (Diffusion, NeRF, GAN), Computer Graphics, DeepSpeed, TensorRT-LLM, vLLM, Megatron-LM, cuSPARSE, cuBLAS, cuRAND, PTX, OpenGL-CUDA interop

PUBLICATIONS [Scholar]

Detecting pneumonia using convolutions & dynamic capsule routing... | *Sensors, MDPI 2020* (cited: 163)(paper)

- Detected: Pneumonia in Chest X-rays; Used: Capsule Networks in conjunction with CNNs; Accuracy Obtained: **96.36% (2.8%↑)**

Cybersecurity Enhancement through Blockchain Training (CEBT)... | *IJIM Data Insights, Elsevier 2021* (cited: 82(paper)

- Developed NAF-based (Actor-Critic Variant) Cont. NPC Adaptiveness Algorithm to augment Game Design & Feedback Mechanics

Data Augmentation Based Morphological Classification of Galaxies... | *ESIN, Springer 2019* (cited: 42)(paper)

- Classified: Galaxy Morphologies; Used: Machine Learning, CNNs with L1 and L2 Regularization; Accuracy: 97.92% (0.4%↑)

NeRFs (Neural Radiance Fields): Past, Present, and Future | *Arxiv Preprint 2023* (cited: 30)(paper)

- Classified various 3D Computer Vision & Generative AI with History of model-based and image-based Novel View Synthesis
- Surveying various MPI & NeRF architectures, models, and extensions in terms of objective functions, datasets & evaluation metrics

AiCNNs (Artificially-integrated Convolutional Neural Networks) for Brain Tumor Prediction | *EAI 2019* (cited: 13)(paper)

- Classified: 3 types of brain tumors; Used: ML Models, Ensemble CNNs (Regularization); Accuracy: 99.49% (3.72%↑).

Guess who?-A serious game for cybersecurity professionals | *9th International Conference, GALA 2020* (cited: 10)(paper)

- Developed a cyber security threat game with clustering based on user feedback (Firebase DB)

SAVCHOI: Detecting Suspicious Activities using Dense Video Caption... | *Arxiv Preprint 2022* (cited: 4)(paper)

- Detected: Suspicious activities (Videos); Used: BMT and DETR (Human-Object Interactions) (ResNet50 bb) + Genetic Algo & **BERT**; BLEU@1: 14.78, BLEU@2: 12.73, BLEU@3: 10.91, BLEU@4: 7.11, METEOR: 16.27, Detection Accuracy: 96.84%

FuNet-40: fundus disease/abnormality classification using ensemble of fine-tuned pretrained convns | *T&F'24* (cited: 4)(paper)

- Trained 56 models for 15-ensemble models for 40 different fundus diseases/abnormalities; Classification Accuracy: 99.6 %;
- Visualizations created in 3-D and 2-D for each image using T-SNE and PCA to depict distinct decision boundaries between diseases

On Multi-Agent Deep Deterministic Policy Gradients and their Explainability for SMARTS Environment | (cited: 1)(paper)

- Compared MAPPO and MADDPG-based RL (with Priority-based Replay Buffers) (on-policy and off-policy RL)

PROJECTS [GitHub]

CUDAx365: 365 Days of GPU, HPC, and AI Engineering | *CUDA (PTX/SASS), Triton, LLMs* (CUDAx365) **Jan 2025–Present**

- Mastered GPU internals–developed custom CUDA kernels for memory patterns and warp-level optimizations for rasterization
- Engineered 2D MHD solvers using MUSCL (achieving near-peak throughput via shared memory padding & graph replays)

Real-Time CUDA-Q Quantum Error Correction Decoder | *CUDA-Q, cuBLAS/cuRAND, PyTorch, C++17* **May 2025–Dec 2025**

- Building GPU-accelerated surface-code decoder with CUDA-Q & transformer (2.3× latency reduction vs. CPU baseline)

Research Paper Implementations (NLP) | *PyTorch, TensorFlow, DeepSpeed, HuggingFace* (Gist) **May 2023–March 2024**

- Implemented LoRA (QLoRA, LongQLoRA), Attention, CoT (& ToT), RAG, HNSW for various Transformer architectures
- Leveraged HF-transformers, bitsandbytes, and Peft for LongNet, LLAMA, RetNet, RWKY

Multi-Agent Deep Deterministic Policy Gradients and Explainability for SMARTS (paper) **December 2023–February 2024**

- Proposed a Multi-Agent DDPG with priority-based buffers against top-15 submissions for safer autonomous driving

Real-Time Data Streaming with Object Detection | *PySpark, Kafka, StreamLit* (project) **May 2022–June 2022**

- Deployed real-time data pipeline using Kafka to capture and stream data, enabling immediate analysis of incoming information

Few Shot Learning Approach to Dynamic Intent Satisfaction | *GPT3, GPT-Neo, NLP* (video) **January 2022–May 2022**

- Proposed LLM-based problem–Intent Satisfaction (fulfill complex intents in slot filling systems) (BLEU (GPT_{dv}): 0.73)
- Curated function definition and docstrings benchmark dataset of 500,000 tokens for Intent Satisfaction

EDUCATION

University of Southern California | *Los Angeles, CA* **August 2021-May 2023**

- ***MS in Computer Science**; Courses: Artificial Intelligence, Algorithms, Deep Learning, Applied NLP, Advanced Computer Vision, Autonomous Cyber Physical System, Information Retrieval & Web Search Engines*

Guru Gobind Singh Indraprastha University | *Delhi, India* **August 2015-May 2019**

- ***Bachelors of Technology in Computer Science and Engineering**; Rank: 6; Courses: Operating Systems, Compiler Design, Computer Networks, Artificial Intelligence, Data Mining, Machine Learning (Python), Discrete Mathematics, C++*

Summer School

- Qiskit Global Summer School (QGSS'21) (certificate) **May 2021–June 2021**
- Introduction to Quantum Computing (QxQ By Coding School) **October 2020–May 2021**

HONORS & AWARDS

- Viterbi Grad Hack Winners (DeepFake Detection (w/wo LipSync) Chrome Extension) (project) (video) **April 2023**
- Smart India Hackathon'19 3rd Runner-Up (INSUrance Recommendations with Integrated Mixed-reality) **April 2019**
- Smart India Hackathon'18 2nd Runner-Up (LOgical Graph Augmented Virtual Indian Map) (project) **April 2018**