

CHIRAG SHETTY

Phone: +91-9092866690 Email: chirag4798@gmail.com Website: chirag4798.github.io

PROFESSIONAL SUMMARY

Software Development Engineer with 5+ years building production AI/ML systems across generative AI, computer vision, and backend engineering. Delivered image-generation and analytics workflows for consumer platforms, and optimized inference for GPUs and edge devices with Python, C++, FastAPI, and TensorRT.

SKILLS

Generative AI	Stable Diffusion, Flux, vLLM, Super Resolution GAN, multimodal LLMs, LangChain, LangGraph, LlamaIndex, Hugging Face, RAG, LoRA/PEFT
Computer vision / edge	You Only Look Once (YOLO), PyTorch, TensorFlow, OpenCV, NVIDIA DeepStream, NVIDIA Jetson devices (TX2 NX), Raspberry Pi
Engineering	Python, C++, FastAPI, Flask, REST APIs, PostgreSQL/SQL, GCP, Docker, Kubernetes, Linux, Git, CI/CD, JWT/OAuth2
Inference / operations	TensorRT, ONNX Runtime, PTQ/QAT, INT8, sparsity, logging, monitoring, debugging, incident response

EXPERIENCE

SDE-2 (ML Engineer) @ Jio Fynd September 2023 - Present

- Architected the Impetus AI fashion-design platform for design generation, variation, and trend exploration; built agentic and retrieval-augmented workflows with LangChain and multimodal LLMs, integrating Flux, Stable Diffusion, and vLLM.
- Led generative image delivery for Ajio, JioMart, and Tira, combining automated tagging, Stable Diffusion backgrounds, U2Net foreground extraction, and Super Resolution GAN upscaling; contributed to a **5% increase in product engagement**.
- Built Python/FastAPI orchestration and applied model quantization, LoRA/PEFT adaptation, and token/pipeline efficiency work to reduce inference latency, GPU memory use, and operating cost.

Machine Learning Engineer @ Adagrad AI September 2021 - August 2023

- Built real-time retail analytics and people-analytics systems for footfall, behavior, person counting, re-identification, and semantic matching; delivered C++ YOLO/PyTorch/TensorRT inference at **225 FPS on NVIDIA T4** GPUs.
- Applied PTQ/QAT to INT8 detection models for distributed deployments and Fortune 500 clients; developed edge Automated License Plate Recognition on NVIDIA Jetson TX2 NX at **50 FPS and 90-95% accuracy**, connected to traffic-enforcement workflows.

Machine Learning Engineer @ Optisol Business Solutions November 2020 - September 2021

- Built TensorFlow/OpenCV power-pole inspection and defect-detection systems, reducing model latency by **50%** with quantization-aware training and sparsity; also built FastAPI/Flask ML workflows and a Q-learning website-testing system.

OPEN-SOURCE CONTRIBUTION

BYTETrack: Implemented and contributed a C++ port of the multi-object tracking algorithm, then integrated it with NVIDIA DeepStream for GPU-accelerated video analytics. Other projects: Shop the Look fashion recommendations using image similarity and ranking; Mercedes-Benz Kaggle bench-time prediction ranked in the top 100; locality clustering using data mining and web scraping. **Profiles**: [GitHub](#) and [LinkedIn](#).

TECHNICAL WRITING

[NVIDIA DeepStream 101: A beginner's guide to real-time computer vision](#)
[NVIDIA DeepStream 101: Creating your first DeepStream application](#)

EDUCATION

Vellore Institute of Technology, Vellore, Bachelor of Technology, Mechanical Engineering, 2016 - 2020
Cumulative Grade Point Average: 8.33/10