

PRANAV SHINDE

Computer Vision Engineer | ML Systems, Annotation Infrastructure, MLOps

shinde.a.pranav@gmail.com | +91 9623623596 | Pune, India

linkedin.com/in/pranaavshinde | shindepranav.site

PROFESSIONAL SUMMARY

Computer Vision engineer building large-scale detection and segmentation systems in production. Built a 500,000+ image annotation pipeline with a custom SAM 2.1 boundary refiner (mAP50 ~0.74, 13-class instance segmentation) and INT8/TensorRT inference pipelines cutting latency by over 5x. Currently at a private R&D firm (defence and government clients), working on object detection and tracking. AWS Certified Developer Associate.

TECHNICAL SKILLS

Computer Vision & ML: PyTorch, YOLO (v8/v11), SAM 2.1, ONNX, TensorRT, OpenCV, custom loss functions, P2 multi-scale detection heads, multi-GPU training, Kalman Filters, ByteTrack, BoT-SORT, ReID

Data & MLOps: CVAT (bare-metal deployment & REST API automation), 500k+ image dataset curation, Docker, Kubernetes, CI/CD, bare-metal Linux

Model Optimization: ONNX/TensorRT export, INT8 & dynamic quantization (TorchAO), Nvidia Nsight profiling, batch inference optimisation

Backend & Cloud: Python, TypeScript, Node.js, REST APIs, PostgreSQL, Redis, AWS (EC2, Lambda, RDS, S3), Nginx

LLM & NLP: OpenAI/Gemini APIs, LangChain, RAG, prompt engineering, VLM-based classification (Qwen-VL)

EXPERIENCE

Badawe Engineers Pvt. Ltd (Junior Engineer) | Private R&D Firm (Defence & Govt. Clients) 01/2026 – Present

- Built and maintained a 500,000+ image annotation pipeline with a custom SAM 2.1 boundary refiner to automate mask correction at scale.
- Trained a 13-class YOLO instance segmentation model (mAP50 ~0.74) using a custom class-weighted loss to handle severe dataset imbalance.
- Engineered a dual-GPU training/validation pipeline resolving CUDA context conflicts, cutting validation turnaround by 18%.
- Migrated inference from PyTorch to ONNX/TensorRT with INT8 & dynamic quantization, cutting model memory footprint by 78% (9.2GB → 2GB) and inference latency from ~23ms to ~3.7ms; profiled cold-start vs. warm inference with Nvidia Nsight Systems/Compute.
- Implemented P2 (high-resolution feature pyramid) detection heads to improve separability between visually similar object classes in cluttered backgrounds.
- Built a 4-stage automated dataset survey pipeline for Hugging Face Hub (metadata harvesting, deduplication, VLM-based classification with majority voting, annotation schema detection), screening 500k+ items in under 5 hours.
- Automated the CVAT annotation workflow end-to-end via REST API, replacing several hours of daily manual effort.

Full Stack Intern (ML Backend) | Vital Vistara Pvt. Ltd.

01/2025 – 04/2025

- Designed AWS infrastructure for ML workloads (VPC, RDS, Lambda, S3, CloudWatch) and implemented CI/CD with Docker and ECS Fargate.

Frontend Intern | Oneqid Technologies, IIT Delhi

06/2024 – 08/2024

- Modernized frontend using Next.js, improving load time by 35% and Lighthouse score from 65 to 92.

KEY PROJECTS

Quantized KV Cache Preallocation for Long-Context LLM Serving (mlx-lm)

- Diagnosed macOS kernel panics and Metal OOM crashes serving a 27B model at 64k context on 24GB unified memory, tracing the cause to dynamic KV cache reallocation churn and a double-buffering bug.
- Patched mlx-lm's cache and server layers to preallocate quantized (4-bit) KV cache buffers and enforce single-instance cache ownership, achieving stable 65,536-token serving with zero crashes across a 10-turn stress test.
- Benchmarked 2.4–3x faster prefill and 30–60% faster decode versus a comparable GGUF/llama-server setup.

Athlete Connect — AI Sports Talent Assessment Platform

08/2024 – 10/2024

- Top 10 National Finalist, Smart India Hackathon 2024** (50,000+ competing teams). On-device pose estimation (Google ML Kit) on Android with a Firebase-backed pipeline for athlete profiling and leaderboards.

Skill Learn — Distributed Compute Pooling Platform

09/2024 – 12/2024

- Built a web platform to pool compute/RAM across multiple low-end machines into one usable resource for 3D rendering, game dev, and AI inferencing; prototyped with aggregated AWS EC2 instances and sandboxed Docker execution, with PgBouncer connection pooling cutting overhead by 70%.

EDUCATION & CERTIFICATIONS

B.Tech in AI and Data Science | MMCOE, Pune University, India

2023 - 2026

Diploma in Computer Technology | BVJNIOT, Pune, India

2020 – 2023

Certifications: AWS Certified Developer Associate (2024), Microsoft Azure Fundamentals (AZ-900)