Rutansh Suthar

Riverside, CA | +1-951-823-3018 | rsuth004@ucr.edu | linkedin.com/in/rutansh-suthar

EDUCATION

University of California — Riverside, CA
Master of Science in Computer Science — Dec 2025
Relevant Coursework: Advanced Operating System, Cloud Computing and Cloud Networking, GPU Architecture & Programming

Gujarat Technological University — Vadodara, India
Bachelor of Engineering in Computer Science & Engineering — Jun 2023

SKILLS

Languages: Python, C, C++, Go, Bash
Cloud & DevOps: AWS (S3, IVS, CDK, Batch, SageMaker, Step Functions, Lambda, ECS, DynamoDB, EC2), Microsoft Azure, Linux, Docker, Kubernetes, OpenFaas, CI/CD, Git, Hugging Face Spaces
AI/ML Frameworks: PyTorch, OpenCV, ONNX, CUDA, ROCm
Backend Frameworks: Django, FastAPI, REST API, Django Rest Framework
Databases: PostgreSQL, MySQL

PROJECTS

Scalable Video Analysis Pipeline (StreamInsight) | GitHub — Riverside, CA
MLOps Project — Jan 2026 - Present

- Architected an event-driven ML inference pipeline on AWS using CDK, where S3 triggers Lambda to initiate Step Functions, orchestrating Batch (Fargate) processing, and SageMaker Serverless inference with results stored across two DynamoDB tables.
- Built a containerized video processing job using FFmpeg and PySceneDetect to extract keyframes at scene boundaries, dynamically scaling extraction volume with video duration to reduce downstream inference load by ~95%.
- Deployed OpenAI CLIP (ViT-B/32) on SageMaker Serverless endpoints for zero-shot frame classification, eliminating idle compute costs with scale-to-zero infrastructure.

Aesthetic Image Curation (Theia Sense) | Project Link — Riverside, CA
Full-Stack Cloud Project — Jun 2025 - Sep 2025

- Architected a microservice-based application deployed as Azure Container Apps and Hugging Face Spaces, decoupling the FastAPI backend and the ML inference service to enable independent auto scaling.
- Developed an AI-powered curation model that analyzes and scores uploaded images for aesthetic quality and used ONNX Runtime to reduce the final container image size by 75% to improve the inference speed.
- Implemented a dynamic ranking algorithm in Python that adapts the results to the collection's overall quality, providing a personalized user experience.

PROFESSIONAL EXPERIENCE

Electrum IT Solutions Pvt. Ltd. — Vadodara, India
Software Engineer - L1 — May 2023 - Oct 2023

- Spearheaded the integration of multiple LLMs leading to a new report generation feature and engineered a containerized Docker solution to emulate and bypass OpenAI API rate limits for robust testing.
- Architected a custom AWS cloud solution using Interactive Video Service, S3 and EC2 to consolidate Video-on-Demand (VoD) streams, simplifying video asset management for clients.

Electrum IT Solutions Pvt. Ltd. — Vadodara, India
Software Engineer Intern — Feb 2023 - May 2023

- Delivered an end-to-end backend solution for an Invoice Management Application, creating a full REST API with over 20 endpoints and complex permission logic using Django Rest Framework.