Prasoon Kumar

San Jose, CA • prasoonkumar.23702@gmail.com • 925-997-8644 • LinkedIn

SUMMARY

Applied AI Engineer with 4+ years of experience in developing and deploying computer vision and AI systems, with a focus on real-time performance and high-impact product features. Passionate about blending photography, computational imaging, and AI to deliver magical user experiences. Skilled in converting research-grade models into production-ready solutions using PyTorch, ONNX, and CoreML. Strong foundation in C++, Python, and GPU programming. Adept at cross-platform deployment for desktop, cloud, and mobile ecosystems.

EXPERIENCE

MLOps Intern — *Vimaan Robotics Inc*

Feb 2025 – Present

- Designed containerized vision inference pipelines with PyTorch + FastAPI, deployed via AWS Lambda with latency <100ms.
- Built CI/CD workflows for cloud-deployed AI models, enabling automatic retraining and version tracking.
- Worked with vision sensor data to detect anomalies and image quality issues under environmental drift.
- Wrote automation tools in Python and Bash to monitor and benchmark model performance in production.

ML Engineer — *Sprouts AI*

Oct 2023 – Mar 2024

- Developed full-stack LLM-powered tools for visual-text content analysis and image-to-text summarization.
- Created custom image annotation and evaluation pipelines for fine-tuning vision models using low-shot data.
- Optimized model inference and memory efficiency using quantization and ONNX export for web deployment.
- Collaborated with design teams to improve visual rendering and alignment for generated image content.

ML Engineer — *Automation Anywhere*

Mar 2024 – Sep 2024

- Integrated real-time summarization modules into front-end interfaces, adapting outputs for low-latency constraints.
- Worked on PySpark + Kafka pipelines for image/text-based bot logs and modeled patterns in automation behavior.
- Deployed models via AWS SageMaker and used Terraform + Docker for robust infra-as-code deployment.
- Explored GPU-based acceleration to reduce render time of visual model outputs.

Graduate Research Assistant — *University of Texas at San Antonio*

Aug 2021 – Oct 2023

- Researched 3D scene understanding and semantic segmentation in robotics using U-Net, YOLOv5, and ORB-SLAM.
- Built robust, real-time image processing and SLAM pipelines for indoor/outdoor navigation.
- Evaluated vision model generalization under varying lighting and occlusion conditions in physical environments.
- Converted models into deployable C++/Open3D modules for embedded SLAM testing.

EDUCATION

MS in Computer Science

University of Texas at San Antonio — *May 2023* | GPA: 3.8/4.0

B.Tech in Computer Science

Institute of Engineering and Management, Kolkata — *May 2021* | GPA: 8.6/10

TECHNICAL SKILLS

- **Languages:** Python, C++, Shell, JavaScript
- **Vision & Imaging:** Computer Vision, Image Segmentation, Object Detection, 3D SLAM, HDR Tuning
- **Frameworks**: PyTorch, OpenCV, Open3D, TensorFlow, Hugging Face, YOLO, U-Net
- **Model Conversion:** ONNX, CoreML, TorchScript
- **Deployment:** FastAPI, Docker, AWS Lambda, Terraform, Cloud + Edge Deployment
- **Mobile/Platform:** iOS CoreML (basic), Android (familiar), Mac/Windows image stack integration
- **Optimization**: Model Quantization, GPU Acceleration, Low-latency Inference
- **Collaboration:** Cross-functional teamwork, technical documentation, iterative feature tuning