

KARTHIK PALYAKERE SURESH

Mannheim, Germany | +49 1746301323 | Karthikps174@gmail.com | karthikps.online

PROFESSIONAL SUMMARY

AI/ML Engineer and Data Scientist with 2+ years of experience developing applied machine learning, computer vision, generative AI, and retrieval-based systems across research and enterprise environments. Strong expertise in Python, SQL, PyTorch, deep learning, multimodal AI, Retrieval-Augmented Generation (RAG), embeddings, vector databases, and agentic workflow orchestration. Experienced in building scalable AI pipelines for structured and unstructured data, from research prototyping and model evaluation to containerized deployment and cloud-based production workflows. Published researcher at ICCV 2025 with a strong interest in semantic retrieval, knowledge grounding, context engineering, and AI agents operating on complex enterprise data.

CORE SKILLS

Programming & Data: Python, SQL, C++, C, Git, Linux, Data Processing, ROS2

Generative AI & Retrieval: Retrieval-Augmented Generation (RAG), Agentic RAG, Multimodal RAG, LLMs, Embeddings, Vector Databases, Semantic Retrieval, Hybrid Search, Vector Similarity Search, Context Engineering, Prompt Engineering, Knowledge Grounding

Agentic AI: AI Agents, Multi-Agent Systems, LangChain, LangGraph, Agentic Workflow Orchestration, Agentic Reasoning

Machine Learning & Deep Learning: PyTorch, TensorFlow, scikit-learn, CNNs, Vision Transformers, Transfer Learning, Model Fine-Tuning, Model Evaluation, Statistical Analysis, Deep Learning

Knowledge & Semantic Systems: Knowledge Representation, Semantic Modeling, Structured and Unstructured Data, Metadata, Knowledge Graph Concepts, Graph Databases, RDF Concepts, Ontology Fundamentals

Computer Vision & Multimodal AI: Object Detection, Instance Segmentation, Semantic Segmentation, Multimodal Learning, CLIP, DINO, SAM2, SAM3, Mask R-CNN, YOLO, Image Classification, 2D/3D Vision

MLOps & Cloud: Docker, FastAPI, AWS (EC2, S3, Lambda), CI/CD, Model Monitoring, ONNX Runtime, TensorRT, OpenVINO, CUDA, Scalable ML Pipelines

Languages: English (C2, Fluent) | German (B1, Actively Learning)

PROFESSIONAL EXPERIENCE

Data Scientist Intern / Werkstudent — Merck Group KGaA, Darmstadt March 2025 – Sep 2026

- Developed and evaluated machine learning and deep learning models for high-resolution SEM and TEM microscopy data, working with structured and unstructured scientific datasets.
- Fine-tuned and evaluated foundation and computer vision models including Mask R-CNN, YOLOv11, SAM2, and SAM3 for instance segmentation and image understanding.
- Architected end-to-end, scalable ML pipelines on AWS with containerized model serving using Docker, improving inference latency by 40% for high-resolution imagery.
- Engineered production-ready AI workflows with CI/CD automation, reducing deployment cycles by 50% and enabling continuous model evaluation and improvement through feedback loops.
- Prototyped LLM-based information extraction and knowledge retrieval workflows using RAG, embeddings, vector similarity search, and semantic retrieval.
- Developed synthetic data generation pipelines that improved model generalization and reduced overfitting by 35%, addressing limited labelled data challenges.
- Evaluated model performance across latency, throughput, memory consumption, and prediction quality, including mixed-precision inference benchmarking against FP32 baselines.
- Collaborated with interdisciplinary research, engineering, and domain teams in agile environments, translating research problems into scalable AI solutions with clear documentation and test coverage.

Research Assistant — DFKI – German Research Center for AI, Kaiserslautern July 2023 – Feb 2025

- Designed and deployed a real-time multi-sensor perception system for robotic platforms, achieving 92% detection accuracy in autonomous scenarios
- Developed ROS2-based perception nodes for synchronized camera streams and object detection on Jetson embedded hardware, achieving sub-200ms inference latency
- Implemented low-light and degraded-vision enhancement pipelines, improving object detection mAP50 by 15% in challenging environments
- Optimized PyTorch models for production deployment, focusing on inference performance, memory efficiency, and real-time constraints in resource-limited environments

- Built modular ML components and scalable data processing pipelines for structured and unstructured sensor data, ensuring robustness and reproducibility across distributed systems
- Developed monitoring and logging infrastructure for deployed AI services, enabling continuous performance tracking and automated alerting
- Published peer-reviewed research at ICCV 2025 demonstrating scalable, production-ready perception systems

PROJECTS

Real-Time Multi-Camera Perception System

- Built a real-time perception pipeline processing 4 camera streams at 30 FPS, integrating object detection, tracking (DeepSORT), and ROS2 communication for autonomous navigation
- Deployed on Jetson embedded hardware with GPU-optimized model serving and sub-200ms end-to-end latency

Enterprise Multimodal Agentic RAG Platform

- Designed and developed a multimodal, multi-agent RAG platform using LangGraph and LangChain to intelligently process and retrieve information from structured and unstructured data, including documents, images, and tables.
- Implemented agentic orchestration and semantic retrieval using LLMs, embeddings, vector databases, hybrid search, and reranking to deliver context-aware, grounded, and reliable responses.

CLIP-DINO Hybrid Vision Transformer

- Developed a hybrid vision foundation model combining self-supervised (DINO) and language-aligned (CLIP) representations for zero-shot object recognition
- Deployed as a containerized inference service using FastAPI and Docker with automated CI/CD pipeline achieving sub-200ms response time

Adaptive Low-Light Detection and Segmentation

- Engineered a U-Net-based image enhancement pipeline integrated with object detection models (RTMDet-tiny, YOLOv3, DeeplabV3+), improving detection mAP50 by 15%
- Evaluated and benchmarked models on dark-image datasets, achieving 92% accuracy through systematic optimization

CNN Quantization and Optimization Pipeline

- Applied Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT) techniques with mixed-precision strategies (INT8/FP16)
- Achieved 2-3x inference speedup with controlled accuracy degradation, benchmarked against FP32 baselines

PUBLICATIONS

- ICCV 2025 | TorchAdapt: Towards Light-Agnostic Real-Time Visual Perception
- IJSRD, Vol.7, Issue 6, 2022 | Intellectual Interactive System
- Active open-source contributor in multimodal and computer vision research

EDUCATION

M.Sc. Computer Science (AI & Computer Vision Specialization) — Technical University of Kaiserslautern, Germany

Apr 2023 – Sep 2026

Relevant Coursework: Machine Learning Algorithms, Deep Learning, Image Processing, NLP, Web Mining

B.E. Computer Science — Nitte Meenakshi Institute of Technology, Bangalore, India

Jul 2018 – Aug 2022

Relevant Coursework: Machine Learning, Artificial Intelligence

CERTIFICATIONS

- Oracle Cloud Infrastructure – Generative AI Professional (2025)
- Oracle Cloud Infrastructure – Data Science Professional (2025)