Devendra Vyas

Germany vyasdevms@gmail.com LinkedIn Github Google Scholar

Summary

Solution architecture for AI-enabled workflows, automation systems, and production ML platforms. Experience in building production AI systems, including edge and on-premises inference infrastructure using Docker, GPU-accelerated deployment, and low-latency model serving. Designed and deployed deep learning infrastructure from edge perception pipelines to low-latency inference. Comfortable deploying LLM/VLM and embedding-based systems in restricted or private environments without reliance on public cloud services. Expertise spans MLOps, Computer Vision, and communicating complex ML concepts to diverse stakeholders.

Professional Experience

Sr. ML Research Engineer **Dec 2023 – Dec 2025**

Verses AI *Remote(Contractor)*

- Designed and deployed an edge-ready 3D perception stack for reliable warehouse object detection and depth-aware navigation on Jetson-class hardware.
- Developed an YOLOv10n perception pipeline for warehouse-specific objects using a BlenderProc2-curated synthetic dataset, improving mAP from 0.15 to 0.62 at 640×640 and reducing the sim-to-real gap by retraining the synthetic-initialized model on cleaned, real warehouse data to reach 0.58 mAP
- Optimized the end-to-end detection stack (pre/post-processing + YOLOv10n inference) with TensorRT on Jetson Orin NX and desktop GPUs, achieving 26 ms latency at 640×640 for real-time edge deployment
- Integrated a DepthAnythingv2-based depth estimation module into the TensorRT perception pipeline, cutting depth latency from 1 s to 50 ms and delivering a combined depth + detection stack at 10 FPS (640×640) and 13 FPS (378×378) on Jetson Orin NX

ML Tech Lead **Dec 2023 – Mar 2026**

Verses Global BV (Project: dAIEdge) *Remote(Contractor)*

- Led development of an Active Inference–based routing agent leveraging the perception pipeline to perform obstacle avoidance, optimizing the planning stack(Jax based) and reducing planning time from 7 minutes to 21 seconds
- Developed a PyMDP-based saccading agent using a Tapo security camera for active visual exploration, enabling person detection and tracking and later forming the basis of a peer-reviewed conference publication
- Designed and maintained a 3D simulation pipeline in NVIDIA Isaac Omniverse to develop and evaluate routing agents in realistic warehouse environments before real-world deployment
- **Publication**: Towards smart and adaptive agents for active sensing on edge devices, **D. Vyas**, M. De Prado and T. Verbelen, HiPeac 2025 (Link)

AI Strategy Consultant **Jan 2024 – Present**

Dev AI Ventures (Self-Employed) *Remote*

- Advised Docai.tools(prev. TenderFix) on migration from OCR-heavy document processing to RAG-based knowledge systems, including retrieval architecture, chunking/indexing strategy, embedding model selection, and private deployment considerations.
- Conducted SOTA Vision-Language Model research for Drizz.dev, delivering preliminary findings and technical recommendations
- Consulted Eventgraphia on deep learning approaches for image culling, developing high-level strategy and architecture planning

Machine Learning Engineer **May 2020 – Jan 2022**

Edge Case Research GmbH (Working Student) *Munich, Germany*

- Deployed an end-to-end ADAS perception pipeline with Multiple Object Tracking (MOT), extending 2D object detection to 3D bounding boxes on AWS
- Integrated monitoring and alerting for production ML models using Weights & Biases, with automated performance tracking across model versions to support data-driven iteration

Machine Learning Engineer **Oct 2019 – Apr 2020**

TerraLoupe GmbH (Working Student) *Munich, Germany*

- Designed and automated an end-to-end MLOps pipeline for multi-terabyte aerial imagery segmentation using Docker and CI/CD, integrating Sacred/Omniboard for experiment tracking and reproducibility across 20+ model iterations
- Achieved 40% reduction in experimentation time through toolchain optimization
- Benchmarked GPU vs TPU setups on throughput, memory usage, and cost per inference, and used these results to define the production infrastructure strategy for large-scale segmentation

ADAS Engineer **Jan 2019 – Aug 2019**

KPIT Technologies GmbH *Munich, Germany*

- Architected Vehicle State Monitor for high-frequency multi-sensor data processing pipeline, handling 100+ Hz sampling rates
- Developed a GUI dashboard to monitor and visualise real-time vehicle health from streaming telemetry
- Developed and Automated comprehensive testing framework for vehicle telemetry system

Software Engineer **Jan 2017 – Aug 2018**

CNRS (XLIM Lab) *Poitiers, France*

- Optimized radio propagation simulation engine through algorithm profiling and memory-efficient implementations
- Achieved 30% runtime reduction and improved numerical stability, enabling larger-scale IoT network simulations
- **Publication:** CupCarbon: A new platform for the design, simulation and 2D/3D visualization of radio propagation and interferences in IoT networks (Link)

Research Experience — Google Scholar Profile

Fraunhofer AISEC Research (Robustness/Privacy) **Feb 2022 – Sep 2023**

Fraunhofer AISEC *Munich, Germany*

- Led robustness/privacy research on CNNs, Lipschitz constants, and differential privacy threats; co-authored pre-print with automated containerized CI/CD pipelines for reproducible evaluation.
- **Pre-Print:** Gradient Masking and the Underestimated Robustness Threats of Differential Privacy in Deep Learning, F Boenisch, P Sperl, **D Vyas**, K Böttinger (Link)

Synthetic Adversarial Generation [Master's Thesis] **Feb 2023 – Aug 2023**

Fraunhofer AISEC — Grade: 1.0/5.0 *Munich, Germany*

- Generated DDPM-based synthetic adversarial datasets for CIFAR-10 robustness re-training, improving ResNet18 accuracy by 8-10%

Pedestrian Segmentation & LiDAR PoC **Nov 2020 – Mar 2021**

Warp Automotive, UnternehmerTUM *Munich, Germany*

- Developed semantic segmentation for urban pedestrians and LiDAR point clouds; refactored pipeline with MLOps (multi-stage Docker, experiment tracking, Go-Task automation)

3D Medical Image Reconstruction **Apr 2020 – Jul 2020**

CAMP Chair, TUM *Munich, Germany*

- Built U-Net for multi-coil MRI 3D reconstruction (3rd place globally); researched NeRFs for CT-from-XRay reconstruction
- **Publication:** Multi-Coil MRI Reconstruction Challenge, *Frontiers in Neuroscience* (Link)

Education

Technical University of Munich **Oct 2019 – Aug 2023**

M.Sc. Informatics | Majors: ML and CV *Munich, Germany*

The LNM Institute of Information Technology **Jul 2012 – Jun 2016**

B.Tech Computer Science *Jaipur, India*

Technical Skills

Gen AI:	vLLM, Ollama, n8n, LoRA, RAG pipelines, agent workflows
Enterprise Integration:	REST APIs, knowledge systems, documentation workflows, Jira/Confluence integrations
MLOps and Infra:	FastAPI, WandB/MLFlow, Docker, Git, CI/CD(GitHub Actions, Jenkins), Kubernetes
Cloud:	Google Cloud Engine, AWS
Databases:	PostGres, MongoDB
Libraries:	PyTorch, Jax, Tensorflow, Pandas, OpenCV ROS/ROS2
Languages:	Python, C++, MATLAB, Java, bash

Misc.

- **Languages:** English(Bilingual/Native), Hindi(Native), German(Intermediate), French(Basics)
- Stand-up comedian, (ex)Podcaster, Mentor TUM.ai