

Bilel Khlaifia

AI / Machine Learning Engineer • Computer Vision & Distributed Training

Tunis, Tunisia | khlaifiabilel@icloud.com | +216 48 077 300 | linkedin.com/in/khlaifiabilel | github.com/khlaifiabilel

SUMMARY

- AI / Machine Learning Engineer with 5+ years building, optimizing, and deploying production-scale deep learning systems (computer vision, geospatial, edge) on NVIDIA DGX, A100, and H100 GPUs.
- Expert in PyTorch, CUDA C++, TensorRT, distributed training (DDP, NCCL, mixed precision), model compression (pruning, quantization, ONNX), and MLOps (Docker, Kubernetes, CI/CD, MLflow, Weights & Biases).
- Scaled and led a 25+ member international engineering organization; delivered \$860K+ in enterprise contracts and cloud credits (AWS, Azure) while cutting training cycles up to 75% and inference latency by 66%.
- PyTorch contributor and former Assistant Professor; translate research (super-resolution, transformers, few-shot learning) into reliable, low-latency production deployments for energy, defense, and agriculture.

TECHNICAL SKILLS

Languages: Python, C/C++, CUDA C++, Bash, SQL

ML / DL Frameworks: PyTorch, Hugging Face Transformers, TensorRT, JAX, ONNX Runtime, scikit-learn

Models & Methods: Vision Transformers (ViT), CNNs, GANs, Diffusion (familiar), LLM fine-tuning (LoRA, PEFT), Few-shot learning, Self-supervised learning, RAG (familiar)

Distributed Training: PyTorch DistributedDataParallel (DDP), NCCL, Automatic Mixed Precision (AMP), Gradient & activation checkpointing, Sharded data loaders, Slurm

GPU & Acceleration: CUDA kernels, Streams, cuDNN, cuBLAS, Nsight Systems / Compute, TensorRT engine tuning, Triton Inference Server

Model Optimization: Structured pruning, Quantization (INT8 / FP16), ONNX export, Knowledge distillation, Latency & memory profiling

MLOps & Deployment: Docker, Kubernetes, FastAPI, gRPC, MLflow, Weights & Biases, Git, GitHub Actions / CI-CD, DVC, Model registry, Experiment tracking, A/B testing

Cloud & Infra: AWS (SageMaker, S3, EC2), Azure ML, Linux, Slurm clusters, NVIDIA DGX, H100 / A100

Data Engineering: Reproducible data pipelines, Apache Airflow, Spark (familiar), GeoPandas, Rasterio

Edge & Embedded: NVIDIA Jetson (AGX Xavier), Low-power CV (<15 W), On-device perception, ARKit / LiDAR

Geospatial / Remote Sensing: Sentinel-1 SAR, Sentinel-2 multispectral, Tiling, Cloud masking, Temporal modeling, Super-resolution

EXPERIENCE

Machine Learning Engineer Lead | NextAV (Geospatial AI Startup)

Sep 2021 – Present

- Designed and productionized a methane / oil leak detection system (multispectral satellite ingestion; 50 TB+ processed), reducing p95 inference latency from 950 ms to 320 ms via structured pruning and TensorRT engine optimization.
- Delivered multi-generation Sentinel-2 super-resolution (10 m to 1.5 m; >95% SSIM, +0.12 PSNR) sustaining 1000+ 512×512 tiles/hour through mixed precision, fused augmentation, and asynchronous I/O; improved downstream land-cover segmentation IoU by +5.1 percentage points.
- Scaled distributed PyTorch training (NVIDIA DGX A100 + Azure H100 clusters) using DistributedDataParallel (DDP), NCCL, and Automatic Mixed Precision (AMP), achieving 3.9× throughput and 75% shorter training cycles via sharded data loaders, gradient checkpointing, and cached tensor pipelines.
- Reduced per-epoch GPU cost by 48% through Nsight Systems profiling, memory reuse, minimized host-to-device transfer, activation checkpointing, and selective precision tuning.
- Closed \$360K+ in enterprise contracts and secured \$500K+ in cloud credits (AWS, Azure) via benchmark-driven presales, architecture validation, and performance evidence for energy-sector stakeholders.
- Built and mentored a 25+ member hybrid engineering organization (6 core engineers + 20+ interns; 95% conversion), instituting profiling standards, reproducible evaluation harnesses, CI/CD pipelines, and code-review quality gates.

Fractional Staff Research Engineer | AtriaGem (Luxury Tech)

Sep 2025 – May 2026

- Lead R&D from concept to prototype, translating cutting-edge research into scalable, production-ready solutions.
- Developed on-device AI / 3D vision models for iOS ring sizing using the LiDAR depth sensor (ARKit), including depth point-cloud processing, geometric calibration, and uncertainty estimation, achieving ~1 mm end-to-end measurement accuracy.

AI Research Consultant | Startupbootcamp (Startup Accelerator)

Jun 2024 – Present

- Standardized MLOps and GPU-acceleration templates (Docker, CI/CD, experiment tracking), accelerating ML model deployment timelines across 5+ portfolio companies.
- Supported cohort growth through expert feedback in workshops and mentor sessions, connecting founders with relevant investors and partners and helping refine go-to-market and fundraising narratives.

Autonomous Systems Research Lead | AVIONAV (Ultralight Aircraft Manufacturer)

Jan 2019 – Feb 2020

- Prototyped an R&D autonomous quadcopter platform from scratch, integrating Pixhawk (PX4) flight control with an NVIDIA Jetson AGX Xavier companion computer to enable GPS / vision-assisted navigation, real-time perception, and closed-loop autonomy (mission planning, tracking, and landing).

- Developed an embedded C++ / CUDA perception stack achieving 60 fps object detection under <15 W on NVIDIA Jetson-class hardware, enabling reliable autonomous landing.

Mar 2020 – Aug 2021: Independent research and academic transition (see Teaching Experience) – applied AI / geospatial curriculum design and remote-sensing R&D collaborations.

EDUCATION

Private Engineering School of Gafsa (ESIP Gafsa)

Engineering Degree, Computer Science (Software Engineering)

Sep 2018 – Aug 2021

Higher Institute of Technological Studies of Gafsa (ISET Gafsa)

Bachelor's Degree, Computer Networks

Sep 2015 – Aug 2018

SELECTED PROJECTS

- **PyTorch Core Reliability Fix:** Resolved a CUDA autograd memory leak impacting 10K+ users by isolating a minimal reproducible C++ / CUDA scenario. [Issue #152063](#)
- **Satellite Super-Resolution GAN (Sentinel-1 / Sentinel-2):** Built an enhancement pipeline (Sentinel-1 SAR 4×; Sentinel-2 10 m to 1.5 m) sustaining 1000+ tiles/hour with mixed precision and optimized tiling. [\[SAR repo\]](#) [\[Multispectral repo\]](#)
- **Autonomous UAV Visual Navigation Stack:** Implemented a CUDA-accelerated GPS-denied navigation stack (60 fps; 89% autonomous landing success) on NVIDIA Jetson AGX Xavier. [\[GitHub\]](#)

PUBLICATIONS

- **Few-Shot Crop Mapping with Vision Transformers** (ISPRS 2022, Oral). Few-shot Vision Transformer (ViT) with temporal token reduction and spectral embeddings; 94.2% macro accuracy with <15% labels (+7.8 pp vs. CNN+LSTM); 30% fewer convergence epochs.
- **Multi-Scale Satellite Super-Resolution** (Under Review, 2026). Cross-scale attention with spectral-consistency loss; 10 m to 1.5 m (6.7×), >95% SSIM, +0.12 PSNR, 42 ms/tile (TensorRT), +5.1 pp land-cover IoU uplift.

TEACHING EXPERIENCE

Assistant Professor, AI & Data Science | Military Aviation School of Borj El Amri (EABA)

Dec 2020 – Nov 2022

- Delivered applied AI / geomatics curriculum (remote sensing, spatial analytics) for defense mission planning.
- Led GIS and ML research improving operational situational-awareness workflows.

Assistant Professor, Data Science | National Agronomic Institute of Tunisia (INAT)

Nov 2021 – Nov 2022

- Taught precision-agriculture AI (spectral analytics, time-series modeling, yield prediction) and supervised graduate research.

Assistant Professor, AI & Data Science | Virtual University of Tunis (UVT)

Oct 2021 – Nov 2022

- Delivered remote AI instruction (200+ students/year) focused on geospatial modeling and agritech applications.

AWARDS & TALKS

- **AfricArena Summit 2024 & 2025** (2026 upcoming) – Selected as exhibitor and speaker at Africa's leading tech-ecosystem accelerator, showcasing AI-powered geospatial solutions to investors and tech leaders. *Cape Town, South Africa.*
- **World LPG Challenge 2025 (Winner)** – Won for an AI-powered methane leak detection solution using multispectral satellite imagery, competing against 50+ international teams with a production-ready detection system. *Global Competition.*
- **GITEX Morocco 2023** – Presented geospatial AI innovations for energy-sector applications at North Africa's premier technology exhibition, engaging 100+ enterprise clients and government stakeholders. *Marrakech, Morocco.*
- **NVIDIA GTC 2021** – Participated in the Jetson Educators Roundtable at the GPU Technology Conference, sharing best practices for teaching edge AI and autonomous systems on NVIDIA Jetson platforms. *Virtual.*
- **IADE Tunisia 2020** – Exhibited autonomous UAV navigation research at the International Aerospace & Defence Exhibition, demonstrating GPS-denied flight systems to defense and aviation industry leaders. *Tunis, Tunisia.*

CERTIFICATIONS

- **NVIDIA Deep Learning Institute (DLI) Certified Instructor & Jetson AI Specialist** (2021 – Present) – Delivered applied CUDA and deployment workshops (profiling, TensorRT optimization, memory efficiency).
- **DeepLearning.AI Ambassador** (2021 – Present) – Led sessions on transformer adaptation, few-shot generalization, and production readiness.

COMMUNITY LEADERSHIP

- **GitHub Campus Advisor** (2021 – Present) – Established branching, PR review, and CI standards, plus repository governance for academic engineering cohorts.
- **GitKraken Ambassador** (2021 – Present) – Improved repository hygiene through semantic commit norms, reduced integration conflicts, and a structured release cadence.