

YITING ZHANG

EDGE AI & COMPUTER VISION ENGINEER

Dublin, Ireland · +353 89 255 4343 · job@yiting.ie · linkedin.com/in/yitingzhang-cv · yiting.ie

PROFILE

Computer Vision engineer with **8 years of industry experience** owning the full path from model to production: training pipelines, quantisation and compression, and deployment onto edge hardware under hard latency and cost constraints — shipped in airport security screening and industrial site safety. Equally comfortable below the application layer, with deep Linux administration and a **self-hosted production estate I run end to end** — virtualisation, storage, networking and a mail server that stays deliverable. Since 2023 the same optimisation discipline applied to LLM systems: quantisation, LoRA fine-tuning and multi-agent orchestration. Currently completing an MSc in Computer Science at Griffith College Dublin.

TECHNICAL SKILLS

Computer Vision & Deep Learning — Object detection (YOLO family, SSD, Faster R-CNN), human pose and keypoint estimation, OCR, image denoising, GANs; PyTorch, TensorFlow, Keras, OpenCV

Edge & Embedded AI — Nvidia Jetson (Nano, NX), DeepStream, TensorRT, ONNX, INT8 quantisation, MobileNet v1–v3, ARM NEON SIMD optimisation, C++ deployment modules

LLM Engineering — RAG pipeline design (LlamaIndex, Elasticsearch as vector store), multi-agent architectures (LangChain, LangGraph, function calling), PEFT fine-tuning (LoRA, QLoRA) on ChatGLM / Llama 2 / Whisper, quantisation (AWQ, GPTQ), Hugging Face

Infrastructure & Backend — Linux administration (Debian, Ubuntu, CentOS), Proxmox VE, Docker / LXC, ZFS, NFS, Elasticsearch / ELK, PostgreSQL, MongoDB, Kafka, Redis, Node.js APIs, Vue.js

Languages — Python, JavaScript / Node.js, Bash, MATLAB, C++

PROFESSIONAL EXPERIENCE

Founder & Technical Lead — Beijing Houzhen Visual Design Co., Ltd.

Nov 2023 – Jan 2026

Bootstrapped product venture · Beijing, China

Founded and single-handedly built "**File IO**", a WeChat mini-program for natural-language file conversion and processing, backed by an LLM agent system. Owned product, architecture, implementation and infrastructure end to end.

- Rebuilt the assistant from a single-agent to a **multi-agent architecture** (LangChain, LangGraph, function calling), resolving tool-selection degradation and hallucination at scale; **cut token consumption per task by roughly 80%** and raised task completion rate.
- Engineered a **RAG pipeline** over a 100,000+ word bilingual knowledge base using LlamaIndex with **Elasticsearch as the vector store**, exposed as a first-class tool to the agent layer.
- Introduced **Kafka and Redis** for event-stream tracing and response caching; built ELK-based observability over agent execution traces.
- Delivered the full stack alone: Vue.js / uni-app frontend, Node.js API backend, deployed on self-managed Proxmox VE bare-metal infrastructure.
- Benchmarked **LoRA vs. P-Tuning v2** on ChatGLM2 and Llama 2 for a domain multi-turn QA dataset; fine-tuned **Whisper-large-v2** with LoRA for domain speech recognition.
- Architected a **hybrid ingress deployment** for all company services — website, file server, authoritative DNS and a **self-hosted mail server**. Compute and storage stayed on owned hardware; a minimal cloud VPS fronted it as public ingress, tunnelled back to an OpenWrt gateway with dynamic DNS, working around residential blocking of ports 80/443. **Cut hosting cost to a fraction of an equivalent fully-cloud deployment.**

Algorithm Engineer — Beijing Institute of Radio Metrology and Measurement

May 2022 – Nov 2023

Smart Municipal Infrastructure business line · Beijing, China — civil construction-site safety

- Built an end-to-end **real-time worker safety monitoring system** on Nvidia Jetson + DeepStream: fall detection, PPE compliance (hard hat, hi-vis workwear), smoking detection and virtual geofencing of restricted zones.
- Developed **perimeter-intrusion alerting** using human keypoint estimation, with automated localisation, frame capture and alarm dispatch — replacing continuous manual CCTV monitoring.
- Owned the full pipeline: data collection, annotation strategy, model training, on-device optimisation, field deployment and **false-positive tuning under real site conditions** (lighting, occlusion, camera angle, weather).
- Authored technical proposals and solution documentation for public-sector smart-city tenders, translating CV system capability into client-facing specifications.

Algorithm Engineer — Lalamove Mar 2020 – Nov 2021

Shenzhen Easy Van Technology (Lalamove) · Platform risk control · Shenzhen, China

- Built a **multimodal fraud-detection system** combining computer vision and speech recognition to assess driver and cargo compliance across the platform, reducing manual audit load and financial loss.
- Engineered an **ALPR (licence plate recognition)** system from detection + OCR that **outperformed leading commercial APIs** (Tencent, Alibaba, Megvii) on the company's internal benchmark set.
- Trained speech models to detect account sharing and fraudulent order patterns from in-app call audio.
- Built a **face-embedding search index** over a large identity database for target retrieval and verification.
- Developed edge-side detection models for in-vehicle IoT devices (forward collision warning, plate reading) running directly on embedded hardware.

Computer Vision Engineer — NUCTECH (Tsinghua Tongfang) May 2018 – Feb 2020

Aviation security screening · Beijing, China

- Designed and deployed **one-stage detection models** (YOLO family, SSD) for prohibited-item detection on millimetre-wave and terahertz imaging equipment at civil airports — firearms, blades, lighters.
- Compressed a human keypoint model from over 70 MB to 2.6 MB using MobileNet backbones, SE modules and quantisation, while improving accuracy and holding millisecond-level inference.
- Reformulated keypoint localisation as a regression task: substantially reduced localisation error and removed the need for separate gender-specific models.

Computer Vision Engineer — Beijing Institute of New Technology Application

Jul 2016 – May 2018

Embedded biometrics · Beijing, China

- Developed a **palmpoint recognition and authentication system** end to end — hand-designed Gabor filter banks for feature extraction, bitwise matching against an enrolled database — deployed on low-power embedded devices in C++.
- **Decomposed and restructured the Gaussian filter pipeline and parallelised it with ARM NEON SIMD**, materially reducing inference latency on target hardware.
- Applied early deep-learning methods to biometric liveness classification, outperforming the incumbent statistical approach.

SELECTED PERSONAL ENGINEERING PROJECTS

Infrastructure and services I run at home — owned hardware, self-managed from the network layer up.

- **Self-hosted platform and public ingress**
Proxmox VE on bare-metal hardware running services in LXC containers, behind an OpenWrt gateway with dynamic DNS. Because the residential ISP blocks ports 80/443, public traffic enters through a minimal cloud VPS tunnelled back to the gateway. Nginx Proxy Manager centralises reverse proxying and TLS across every service, and a WireGuard overlay carries routed access between Ireland and China.
- **Elasticsearch as event store and security telemetry**
- **Document and knowledge tooling**
paperless-ngx with Chinese and English OCR, turning everyday paperwork and receipts into a searchable, auto-classified archive. Alongside it: Trilium for structured notes, replacing a legacy LDAP-era setup; AFFine for infinite-canvas work; Memos for quick capture; Kavita for the ebook library and Jellyfin for media.

PUBLICATION

Zhang Y., Wang C., Fan M., Cai W., Shao X. "Quantitative Analysis of Heavy Metal Ions Based on Portable NIR Spectrometer." Spectroscopy and Spectral Analysis, 36(12), 4100–4104, 2016. <https://www.gpxygpx.com/CN/Y2016/V36/I12/4100>

First author. Built multivariate calibration models on near-infrared diffuse-reflectance spectra to quantify Ni^{2+} and Cu^{2+} simultaneously in dilute solution. Benchmarked spectral pre-processing pipelines (continuous wavelet transform, MSC, SNV, Savitzky–Golay derivatives), selected latent-factor count by leave-one-out cross-validation, and confirmed generalisation on a held-out validation set — $r = 0.98$ and 0.97 respectively, at concentrations down to $3.0 \text{ mg} \cdot \text{L}^{-1}$.

EDUCATION

MSc, Computer Science — Interactive Digital Media · Griffith College Dublin, Ireland · Jan 2026 – 2027 (expected)

MEng, Materials Engineering · Nankai University, China · 2013 – 2016 *Thesis: "Software Design for Portable NIR Spectrometer and Its Application in Quantitative and Discriminant Analysis"* — Research Center for Analytical Sciences, College of Chemistry

BSc, Safety Science and Engineering · China University of Labour Relations, China · 2009 – 2013

ADDITIONAL

Work authorisation — Currently Stamp 2 (student): eligible to work 20 hours per week during term and 40 hours per week during holiday periods. Eligible for Stamp 1G graduate permission on completion of the MSc, giving full labour-market access with no employment permit required.

Languages — Mandarin Chinese (native), English (professional working proficiency), Japanese (N5)