Hung-Jen (Daniel) Tu 杜弘仁

AI Software Engineer

New Taipei City, Taiwan
danghoangnhan.1@gmail.com
danghoangnhan.github.io
github.com/danghoangnhan
+886 906-276-967
ORCID: 0009-0005-8703-9956

Taiwan-born, Vietnam-raised, Taiwanese citizen. Vietnamese name Đặng Hoàng Nhân — hence `danghoangnhan`*; also Daniel Du.*

Education

Fu Jen Catholic University — New Taipei City, Taiwan
Master of Science in Computer Science — *Sep 2022 - Aug 2024*

Thesis: Enhancing Medical Diagnosis Document Analysis with Layout-Aware Multitask Models

Research areas: Optical Character Recognition, Document Understanding, Large Language Models, Federated Learning, Object Detection

Fu Jen Catholic University — New Taipei City, Taiwan
Bachelor of Engineering, Software Engineering — *Sep 2018 - Aug 2022*

Thesis: Reinforcement Learning Approach for the Restaurant Meal Delivery Problem using Value-inserting Method

Technical Skills

- **Deep Learning**: PyTorch, TensorFlow, PaddlePaddle, ONNX, OpenCV, LoRA/PEFT, DPO
- **Inference/Edge**: TensorRT, CUDA, NVIDIA Triton, NVIDIA Jetson, llama.cpp
- **MLOps**: Kubeflow Pipelines, Kubeflow Trainer, KServe, Model Registry, MLflow, Knative
- **Programming Language**: Python, C++, Java, Go, Rust
- **Backend**: FastAPI, gRPC-Gateway, Spring
- **DevOps**: Docker, Kubernetes, Helm, kubectl, Rancher, GitLab CI/Runner, nginx, Envoy, GCP
- **Security**: Trivy, OWASP ZAP, Dependency-Track, Harbor, CycloneDX
- **Database**: PostgreSQL, GDAL/PostGIS, Microsoft SQL Server, MongoDB, Redis, InfluxDB, GeoServer
- **Message Queue/Streaming**: Kafka, RabbitMQ, Pathway
- **Architecture**: SOLID, Dependency Injection, Event-Driven Architecture, Design Patterns, Multithreaded Programming

Experience

ThinkTron Ltd. — Siyuan St, Zhongzheng District, Taipei City, Taiwan
AI Software Engineer — *May 2025 - Present*

- **Project: Satellite Imagery AI Automatic Identification System**:
  - Oriented Object Detection: Applied **YOLOv11-OBB** to perform **oriented bounding box detection**, capturing the heading of aircraft, ships, and vehicles.
  - Small Object Enhancement: Integrated **Slicing Aided Hyper Inference (SAHI)** to slice remote sensing images and fan the tiles across **KServe** replicas, lifting small-target recall from [X]% to [Y]%. Added a Super-Resolution stage ahead of detection to improve subsequent detection and annotation quality.
  - GPU Performance Optimization: Compiled **TensorRT** engines with FP16 quantization and layer fusion, cutting inference to [X] ms/image ([N]x versus FP32).
  - AI Serving & Deployment (Kubeflow Ecosystem): Ran multi-GPU distributed training with **Kubeflow Trainer** (PyTorchJob, **Distributed Data Parallel**), promoted checkpoints through the **Kubeflow Model Registry**, and orchestrated train-to-deploy as **Kubeflow Pipelines**. Served models as **KServe** InferenceServices on NVIDIA Triton behind an nginx ingress with **Envoy** as the Knative data plane, autoscaled on concurrency with scale-to-zero to release idle GPUs.
  - Experiment Management: Used **MLflow** to track model versions, hyperparameters, and performance metrics.
  - Parallel Processing: Automated data workflows with **Pathway**'s incremental streaming engine, sharding large aerial images across parallel workers (Pansharpening, COG generation, PostGIS Raster writing, GeoServer publishing).
  - Resource Management: Integrated GDAL to build partition indexes at each stage and relied on Pathway's backpressure-aware scheduler to optimize CPU/IO allocation.
- **Platform: On-Premise Camera Edge Inference Appliance**:
  - Edge Runtime: Built a reusable on-premise inference appliance on **NVIDIA Jetson Orin Nano**, running [N] live camera feeds at [X] FPS through whichever vision model a deployment calls for, entirely on device.
  - Device-Side Optimization: Exported models through **ONNX** with dynamic input shapes and compiled **TensorRT** engines on the target board using FP16 precision and layer fusion, sized to the Orin's unified memory budget.
  - Monocular Depth Estimation: Benchmarked **MiDaS** against **Depth Anything V2** for depth quality and on-device latency, and shipped **Depth Anything V2** as the depth backbone behind a common runtime interface.
  - Model Delivery: Promoted checkpoints from the **Kubeflow Model Registry** into versioned per-device TensorRT engines, so one trained artifact serves both **KServe** in-cluster and the edge fleet.
- **Platform: Secure CI/CD Delivery (DevSecOps)**:
  - Delivery Pipeline: Built Docker images on self-hosted **GitLab** Runners, pushed them to **Harbor**, and released to Kubernetes as **Helm** umbrella charts versioning the full stack as one release.

- Security Gates: **Trivy** scanned on push and Harbor's prevent-vulnerable-images-from-running policy blocked flagged tags from Kubernetes; **OWASP ZAP** ran DAST across the FastAPI/gRPC gateway and **CycloneDX** SBOMs fed **Dependency-Track**.
- Team Visibility: Tracked **GitLab CI/CD** builds, scans, DAST, CVEs, and Harbor pulls in Microsoft Teams.

Fubon Life Insurance Co., Ltd. Neihu District, Taipei City, Taiwan
Data Scientist Intern *Mar 2024 - Sep 2024*

- **Project: Insurance AI**:
  - Image Preprocessing: Enhanced document clarity with **OpenCV** and applied **Non-Maximum Suppression (NMS)** to mitigate stamp occlusion, improving downstream text detection accuracy.
  - **OCR Engine Development:**
    - **Text Detection:** Built the detector on the **PaddleOCR** library over the **PaddlePaddle** framework, fine-tuning **ICDAR benchmark models (DBNet, EAST)** for text localization and evaluating on **Recall, Precision, and F1-score**.
    - **Text Recognition:** Integrated **CRNN, SVTR, and TROCR** models for character transcription, leveraging open data from **ESUN.AI**. Validated model performance using **Character Error Rate (CER)**.
    - **Data Augmentation:** Used **Text Renderer** (an OCR generator tool) to programmatically generate a large-scale synthetic training dataset, improving generalization across varied document fonts and formats.
  - Designed microservices on Kubernetes, with Grafana for model performance metrics and Kibana for log analysis.
  - **Natural Language Processing (NLP) & Document Intelligence:**
    - Enhanced the LLM's OCR capabilities by integrating **layout-aware embedding** techniques with **LayoutXLM**, built on the **PaddleNLP** library over the **PaddlePaddle** framework.
    - Applied the **TAIDE** encoder for Traditional Chinese, cutting character segmentation errors by 45%; fine-tuned **Llama 2.0** for entity extraction with **DPO** and **LoRA** (**PEFT**), reaching 3x faster inference at 92% accuracy.
    - Optimized and deployed the fine-tuned model for inference using **llama.cpp**.

Knowledge Engineering Lab, Fu Jen Catholic University New Taipei City, Taiwan
*Research Assistant, Supervisor: **Jia-Lien Hsu*** *Jul 2022 - Sep 2024*

- **Form Understanding Project Certificate of Diagnosis**: Architected end-to-end medical document analysis system on ALTOS AI workstation using Kubernetes orchestration. Integrated OCR algorithms and document understanding models in containerized microservices architecture. Earned project scholarship award for innovative system design with Fubon Life Insurance (link)

Jasslin Technology Co., Ltd. Taishan District, New Taipei City, Taiwan
Part-time Software Engineer *Jun 2019 - Sep 2022*

- **Project: Fleet Management System**:
  - Built a high-throughput event-driven platform on Apache Kafka for real-time data feeds of 10,000+ devices.
  - Used Java Socket.IO for bi-directional communication between server and devices.
  - Optimized large-scale insert paths to 10,000 inserts/sec, a 5x improvement, across SQL and NoSQL stores.
- **Project: Quality License Management**: Developed a Quality License Management (QLM) system with automated alerts for expiring licenses, integrated with the company's built-in ERP system.

Projects

- **CB-STM Test Implementation**: Android application for early Mild Cognitive Impairment (MCI) detection, with real-time behavior tracking in SQLite and on-device inference via ONNX. Processed 1000+ assessments; demonstrated at Taiwan Healthcare Expo 2022 (link)

Open Source Contributions

- **Pathway — Real-Time Data Framework (github)**: Authored two **merged** native connectors to Pathway, an open-source Python/Rust streaming-data framework.
  - **Microsoft SQL Server connector** (`pw.io.mssql`): pure-Rust TDS implementation via `tiberius` (no ODBC drivers); static-polling and Change Data Capture (CDC) streaming reads, snapshot / change-stream writes, and full SQL Server type mapping (PR #205).
  - **RabbitMQ Streams connector** (`pw.io.rabbitmq`): native read/write on `rabbitmq-stream-client` with TLS, AMQP 1.0 message keys, and parallel readers; shipped an end-to-end RabbitMQ → Pathway → PostgreSQL ETL example (PR #207).
  - Authored Python ↔ Rust bindings (**PyO3**); the MSSQL connector now underpins further community contributions.

Publications

- Enhancing Medical Diagnosis Document Analysis with Layout-Aware Multitask Models, **Hung-Jen Tu**, Jia-Lien Hsu, **Diagnostics 2025**, paper
- Adaptive Model Transfers and Aggregations for Efficient Federated Learning in IoT Edge Systems with Non-IID Data, Jenn-Wei Lin, Taewoon Kim, **Hung-Jen Tu**, Po-Hsien Kuo, **DSIT 2023**, paper

Certificates

- TensorFlow Developer Certificate, Google
- Deep Learning Specialization, DeepLearning.AI

Languages

- **Vietnamese**: Native
- **Chinese**: Proficient
- **English**: Proficient