Tae-Hoon Yong

yongtaehoon@gmail.com | LinkedIn | GitHub

FIELD OF INTERESTS

- Production-grade ML Serving Systems (Real-time & Batch)
- High-reliability AI Services under SLA Constraints
- GPU-accelerated Inference & Model Optimization
- Scalable Retrieval & Inference Pipelines for Enterprise AI
- Cloud-native ML Systems & MLOps (Azure, CI/CD)

EXPERIENCE

KT — AI Engineer

Seoul, South Korea | Apr 2025 – Present

- **Production AI Service Development (Financial Domain):** Designed and operated a production-grade AI service for financial sales assistance, supporting 100+ concurrent users in a restricted (closed-network) Azure environment.
- **Scalable Retrieval & Inference Pipeline:** Built a large-scale document ingestion and retrieval pipeline by parsing, chunking, and embedding 300K+ insurance policy documents, enabling low-latency hybrid search using Azure AI Search.
- **Asynchronous Processing & Traffic Isolation:** Designed queue-based asynchronous workflows to isolate heavy indexing workloads and prevent performance degradation of Azure Functions under peak traffic.
- **ML Service API Development:** Implemented FastAPI-based service endpoints supporting both bulk and single-file ingestion, ensuring predictable latency and stable request handling in production.
- **Secure Cloud Infrastructure:** Integrated Azure Functions, Durable Functions, Key Vault, Cosmos DB, and Private Endpoints with GitLab CI/CD to operate AI services in a security-constrained enterprise environment.

OSSTEM IMPLANT Co., Ltd. — AI Research Engineer / Team Leader

Seoul, South Korea | Aug 2021 – Mar 2025

- Led an AI development team of 8+ members, owning end-to-end delivery from research prototyping to commercial deployment
- Delivered multiple production AI systems spanning computer vision, 3D modeling, and on-device inference, tightly integrated into commercial dental software pipelines

Selected Contributions

- Led the development of on-device ML inference modules designed to run on low-spec Windows environments (Windows 7 and earlier) with strict hardware constraints
- Optimized end-to-end inference latency by reducing CUDA initialization overhead and applying model compilation using TensorRT and ONNX Runtime
- Applied pruning, quantization, and knowledge distillation to balance accuracy, throughput, and resource usage for real-time deployment
- Designed lightweight serving logic to ensure stable response times despite limited compute and memory resources
- Developed and deployed production AI modules integrated into commercial dental software products (OneGuide, One3, OneOrtho, OneClick, V-Ceph)

EDUCATION

Seoul National University — M.S. in Applied Bio-Engineering

Seoul, South Korea | 2019 – 2021

- Thesis: GAN-based Quantitative Cone-Beam CT for Bone Mineral Density (Advisor: Won-Jin Yi)

Hongik University — B.S. in Computer Engineering

Seoul, South Korea | 2014 – 2019

PUBLICATIONS (SELECTED)

Journal Papers (SCI/SCIE)

- **T.H. Yong*** et al., QCBCT-NET for Bone Mineral Density Measurement from Quantitative Cone-beam CT, Scientific Reports, 2021.
- O. Kwon*, **T.H. Yong*** et al., Automatic diagnosis for cysts and tumors of both jaws on panoramic radiographs, DMFR, 2020.

International Conferences

- D. Lee*, **T.H. Yong** et al., Revolutionizing Dental Image Segmentation (Filter Pruning & KD), IEEE ICASSP, 2024.
- **T.H. Yong*** et al., Automatic Detection of Inferior Alveolar Nerve on CBCT, IPIU, 2023. (Oral presentation & Best paper award)
- H.G. Ahn*, **T.H. Yong*** et al., Deep high-resolution landmark detection, MICCAI Workshop, 2022.

CHALLENGES & AWARDS (SELECTED)

- MICCAI Grand Challenge (2023): 8th Prize — CBCT Segmentation Challenge
- Best Paper Award, IPIU 2023 — Inferior Alveolar Nerve Detection (Oral)
- MICCAI Grand Challenge (2022): 6th Prize — 3D Teeth Scan Segmentation and Labeling

TECHNICAL STACK (SELECTED)

- **Languages/DL:** Python, C++, PyTorch, TensorFlow, ONNX, TensorRT
- **AI/LLM:** Azure OpenAI, Vector DB (Azure AI Search, Chroma)
- **Cloud/DevOps:** Azure, Docker, GitLab CI/CD