

# Joonbeom Park

---

**Seoul, Republic of Korea**

[joonb14@gmail.com](mailto:joonb14@gmail.com)

GitHub: [github.com/joonb14](https://github.com/joonb14)

---

## Summary

---

MLOps/Backend Engineer with 5+ years of experience designing and implementing scalable AI infrastructure and serverless architectures. Proven expertise in deploying LLM/VLM serving systems, building MLOps pipelines, and developing production-ready applications using AWS cloud services. Published researcher in mobile AI systems with strong foundation in computer vision and real-time inference optimization.

---

## Technical Skills

---

### Cloud & Infrastructure

AWS (Lambda, SageMaker, ECS, Fargate, Step Functions, API Gateway, CloudFront, S3, DynamoDB, OpenSearch, Bedrock, Cognito, Media Converter), IaC(CDK)

### AI/ML

PyTorch, TensorFlow, vLLM, Ray Serve, Triton Server, Computer Vision, LLM/VLM Serving, MLOps

### Backend

Python, FastAPI, TypeScript, Node.js, C++, WebSocket

### DevOps

Docker, Airflow

### Mobile

Android (Java), Unity

---

## Experience

---

**SK Telecom | Security AI Development Team**

### Synapsego: AI Video Synopsis & Goal-step Analysis

- Architected and deployed multimodal AI pipeline for automated video analysis and report generation from bodycam footage
- Implemented LLM/SLM/VLM serving infrastructure using vLLM, Ray Serve, and Triton Server with BLS
- Built end-to-end data processing pipelines with Airflow DAGs and FastAPI
- Optimized VLM inference performance using Python concurrent.futures for parallel processing

### CVOps: MLOps for AI Surveillance Cameras

- Designed and implemented automated model retraining pipeline to reduce false positives in surveillance systems
- Built scalable MLOps infrastructure using AWS serverless services (SageMaker, Step Functions, Lambda, SQS)
- Developed image streaming middleware with CloudFront, ALB, Fargate, and AWS Fathom SDK
- Created semantic image search system using OpenSearch and OpenCLIP for feature extraction
- Deployed 90% of infrastructure as code using AWS CDK

### Radio Tower Anomaly Detection (June 2023 - August 2023)

- Built batch inference service for drone-based tower inspection using SageMaker ProcessingJob
- Implemented cost-effective on-demand compute model reducing operational expenses
- Designed fault-tolerant pipeline with Step Functions for error handling and automated reporting

### Aug Team

#### Backend Engineer | September 2023 ~

#### Shot Up

- Designed and deployed serverless infrastructure for AI-powered screenshot analysis application
- Implemented semantic search using OpenSearch Serverless and AWS Bedrock with RAG architecture
- Built authentication system with Lambda@Edge and CloudFront for origin protection
- Deployed infrastructure using AWS CDK with Lambda Function URLs for streaming responses
- Product launched on [AppStore](#)

#### aug: spatial social | Aug Team

- Architected event-driven serverless infrastructure for location-based AR social network

- Built real-time direct messaging system using WebSocket API, EventBridge, Lambda, and DynamoDB
- Implemented video streaming with automatic HLS conversion using MediaConvert and CloudFront CDN
- Developed geolocation-based content search API using Mapbox quadkey indexing
- Deployed authentication with Firebase (Google, Apple login) and mobile push notifications via SNS/FCM
- Product launched on [AppStore](#)

## Tmax Inc | Multiple Teams

**Software Engineer** | Bundang, Korea | February 2021 - January 2023

### Online Meeting Team

- Ported PyTorch audio denoiser to C++ libtorch for integration with OWT media server
- Optimized real-time audio processing pipeline for online meeting solution

### Computer Vision Team

- Achieved KISA BIO Authentication certification (ISO/IEC JTC1 SC37) for HyperFace algorithm
- Deployed deep learning models on NVIDIA Triton Server with Ensemble pipeline
- Developed real-time pose estimation on smartphones using BlazePose and Unity (16 FPS on Galaxy S21)

## Yonsei University | Mobile Embedded System Lab

**Graduate Researcher** | Seoul, Korea | March 2019 - February 2021 | Advisor: Prof. Hojung Cha

- Published paper on runtime gaze tracking for smartphones at IEEE PerCom 2021
- Developed lightweight CNN for real-time gaze estimation using TensorFlow Lite
- Created data collection platform using Android ML Kit and RenderScript optimization

---

## Education

---

**M.S. in Computer Science** | Yonsei University | February 2021 **B.S. in Computer Science** | Yonsei University | February 2019

---

## Publications

---

**GAZEL: Runtime Gaze Tracking for Smartphones** J. Park, S. Park, H. Cha | IEEE International Conference on Pervasive Computing and Communications (PerCom 2021) DOI: [10.1109/PERCOM50583.2021.9439113](https://doi.org/10.1109/PERCOM50583.2021.9439113)