KAZUKI MINEMURA

Email: kazuki.minemura@gmail.com, Mobile: +81-8023704256

PROFESSIONAL SUMMARY

Bilingual AI engineer and software performance consultant with over 6 years of hands-on experience in AI inference, compiler tooling, and GPU optimization. Specialized in Python-based ML workloads, inference engine migration (e.g., vLLM to SYCL for Intel XPU), and performance tuning with tools like Intel VTune and oneDNN. Proven ability to deliver technical enablement for customers, design scalable validation systems, and support the deployment of AI pipelines across enterprise environments. Fluent in English and Japanese with experience supporting both R&D and production teams.

TECHNICAL SKILLS

- Languages: Python, C/C++, SYCL, CUDA, FORTRAN
- AI/ML: PyTorch, TensorFlow, OpenVINO, vLLM, oneDNN, LLM inference
- ML Ops & Cloud: Docker, Jenkins, Ansible, VTune, Vertex AI (introductory), GCP (ongoing)
- Data/Infra: BigQuery, JSON, XML, MongoDB
- OS: Linux (Ubuntu), Kernel-level debugging, Yocto BSP
- Tools: Intel VTune, GDB, Git, Jira, SYCL, MLFlow (introductory)

PROFESSIONAL EXPERIENCE

Software Technical Consultant Engineer

Intel Kabushiki Kaisha, Tokyo, Japan — July 2022 – Present

Field-facing architect & optimization specialist across AI inference, compilers, and performance engineering.

- Acted as technical liaison between Intel and enterprise clients, providing architecture-level consultation, model optimization, and toolchain debugging across HPC and AI/ML stacks.
- Spearheaded the SYCL-based port of vLLM, enabling generative AI workloads on Intel XPU and expanding OSS compatibility.
- Designed and delivered over a dozen enterprise trainings on compiler internals, oneDNN tuning, VTune profiling, and AI performance best practices.
- Conducted performance tuning of transformer-based inference workloads (vLLM) on Intel GPU (XPU) using VTune and SYCL. Optimized multithreading and memory access patterns to improve runtime efficiency.
- Investigated and resolved OS-level regressions (kernel panic, CPU-specific bottlenecks) for next-gen mobility use cases.

Key Achievements:

- Built reusable SYCL migration templates and GPU optimization guides.
- Enabled model benchmarking and inference on Intel accelerators (GPU/XPU).
- Actively contributed to Intel’s developer ecosystem and OSS community.

Computer Vision Engineer

Intel Microelectronics (M), Penang, Malaysia — Jan 2019 – Jun 2022

Lead AI PoC implementation and system validation for large-scale CV/AI projects.

- Developed a multi-view object detection system for China’s Ministry of Education, utilizing OpenVINO for CPU inference at scale.
- Designed an automated CI/CD validation framework with over 6,000 test cases for multi-platform inference benchmarking.
- Led a cross-functional team of engineers in validating models on edge devices, optimizing for inference speed and reliability.

Key Achievements:

- Delivered infrastructure to support scalable AI deployment across GPU and CPU backends.
- Mentored junior engineers in QA automation and system architecture.
- Supported client-side engineers with hands-on PoC delivery and benchmarking.

Software Validation Engineer / Graduate Trainee
Intel Microelectronics (M), Penang, Malaysia **Jan 2016 – Dec 2018**

- Built LiDAR-based perception models and detection pipelines for autonomous systems.
- Presented AI-related research at international venues and represented Intel at public technical events.
- Automated regression and validation infrastructure to reduce manual testing effort by 80%.

EDUCATION

Ph.D. in Computer Science Feb 2013 – Jan 2017
University of Malaya, Kuala Lumpur, Malaysia

- Conducted research on "Sketch - An Investigation into Feature Extraction in Compressed Domain"
- Published 2 ISI-indexed journal articles, 1 book chapter, and 8 peer-reviewed conference papers

M.S. in Electrical and Electronics Engineering Apr 2010 - Mar 2012
Shinshu University, Nagano, Japan

B.S. in Electrical and Electronics Engineering Apr 2006 - Mar 2010
Shinshu University, Nagano, Japan

PUBLICATIONS (Last 6 years)

ISI Indexed Journal

J1. Raphaël C.-W. Phan, Yin-Yin Low, KokSheik Wong, **Kazuki Minemura**, “Strengthening speech content authentication against tampering”. Speech Communication. Vol 6. 2021, (IF 2.017)

J2. **Kazuki Minemura**, KokSheik Wong, C.-W Phan, Kiyoshi Tanaka, “A novel sketch attack for H.264/AVC format-compliant encrypted video”. IEEE Transactions on Circuits and Systems for Video Technology. Jul. 2016, (IF 9.9)

J3. **Kazuki Minemura**, KokSheik Wong, Xiaojun Qi and Kiyoshi Tanaka, “A Scrambling Framework for Block Transform Compressed Image,” Multimedia Tools and Application, Feb. 2016, (IF 2.313)

Peer Reviewed Conference Paper

C1. **Kazuki Minemura**, Hengfui Liau, Abraham Monrroy and Shinpei Kato, “LMNet: Real-time Multiclass Object Detection on CPU using 3D LiDAR”, IEEE Conference on Intelligent Robot Systems (ACIRS), pp. 28-34, 2018.

C1. Yiqi Tew, **Kazuki Minemura** and KokSheik Wong, “HEVC selective encryption using transform skip signal and sign bin”, Asia-Pacific Signal and Information Processing Association (APSIPA), pp. 963-970, 2015.

C2. Masaya Moriyama, **Kazuki Minemura** and KokSheik Wong, “Moving Object Detection in HEVC Video by Frame Sub-sampling,” IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pp. 48-52, 2015.