

Trung Dao

trung.dt880@gmail.com | linkedin.com/in/trung-dt880 | github.com/bomcon123456

PUBLICATIONS

(*) denotes equal contribution.

- [P1] **Trung Dao***, Duc Hong Vu*, Cuong Pham and Anh Tran. "EFHQ: Multi-purpose ExtremePose-Face-HQ dataset." CVPR, 2024.
- [P2] **Trung Dao**, Thuan Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, Anh Tran. "SwiftBrushV2: Make Your One-step Diffusion Model Better Than Its Teacher." ECCV, 2025.
- [P3] Hao Phung*, Quan Dao*, **Trung Dao**, Hoang Phan, Dimitris N. Metaxas, Anh Tran. "DiMSUM: Diffusion Mamba - A Scalable and Unified Spatial-Frequency Method For Image Generation." NeurIPS, 2024.
- [P4] Quan Dao*, Hao Phung*, **Trung Dao**, Dimitris N. Metaxas, Anh Tran. "Self-Corrected Flow Distillation for Consistent One-Step and Few-Step Image Generation." AAAI, 2025.
- [P5] Viet Nguyen*, Viet Nguyen*, **Trung Dao**, Toan Tran, Anh Tran. "SNOOPI: Supercharged One-step Diffusion Distillation with Proper Guidance." ICCV, 2025.
- [P6] Anh Nguyen*, Viet Nguyen*, Duc Vu, **Trung Dao**, Chi Tran, Toan Tran, Anh Tran. "Improved Training Technique for Shortcut Models." NeurIPS, 2025.

EXPERIENCE

- **Qualcomm AI Research** Vietnam
Staff Machine Learning Engineer Nov 2025 - Current
 - Develop and optimize cutting-edge multimodal models on edge devices.
 - Architect scalable systems to transition prototypes into production, ensuring code quality and maintainability.
- **Qualcomm AI Research** Vietnam
Senior Machine Learning Engineer Nov 2024 - Nov 2025
 - Promoted to **Staff Engineer** following a performance rating **5/5** ("Far Exceeds Expectations") in the 2025 annual review.
 - **Efficient diffusion models:**
 - Pioneered a block-based distillation technique for the flux.1 model, reducing its size by 40% while maintaining over 95% performance integrity (only a 5% drop) on HPSv2 for internal benchmarks.
 - Engineered an attention-based distillation method to significantly compress text encoder requirements for DiT-based models, enabling the swap from the T5-XXL to the T5-base encoder in PixArt-alpha with less than 2% performance drop on HPSv2.
 - **Cutting-edge models on edge devices quantization and deployment:** Spearheaded the W4A16 quantization and deployment of LFM-2, the first hybrid transformer model successfully running on the Snapdragon Gen 5 chip, achieving high-performance throughputs of 9000 tok/s (prefilling) and 90 tok/s (decoding).
 - Role commenced with MovianAI, transitioning to a full Qualcomm position after the acquisition and integration in April.
- **VinAI Research** Vietnam
Research Resident March 2023 - Oct 2024
 - **Advisor:** Dr. **Anh Tran**, Dr. **Cuong Pham**.
 - **Research Focus:** Generative vision models, emphasizing GANs and diffusion models.
 - **Past works:**
 - Improved quality of one-step and few-step text-to-image diffusion models [P2, P4, P5].
 - Introduced a novel diffusion model architecture integrating Mamba for enhanced efficiency and scalability [P3].
 - Developed a large-scale extreme-view face dataset to enhance synthesis quality and benchmark face recognition [P1].
 - **Managing HPC cluster:** Managed and optimized a high-performance computing (HPC) cluster with 48 A100 GPUs, increasing real-time GPU utilization by **30x** through a novel queuing strategy.
- **VinAI Research** Vietnam
AI Engineer December 2020 - March 2023
 - **Advisor:** Dr. **Dzung Nguyen**, Dr. **Anh Tran**, Prof. **Minh Hoai Nguyen**.
 - **Face Recognition Module**
Role: Module Owner.
 - Multi-node model training on large-scale datasets (up to 60M images).
 - Created a framework for profiling, parameter tuning, and optimizing the training process on SLURM.
 - Developed Face Recognition Models for diverse applications, including masked face access control and surveillance CCTV, deployed at scale with **50K daily active identities**.

- Achieved **8th** place overall (**2nd** on Masked Dataset) in ICCV21-MFR Competition, July 2022.
- Built multiple supporting apps for Face Recognition: Model Visualization, Video Inference, Data Labeling Tool (support semi-automated interclass/ intraclass cleaning).
- Quantized and deployed a module of 3 models on Qualcomm's AIC100 (up to 30 concurrent streams), also deployed to NVIDIA's device using TensorRT and to Android using multiple inference engines (ONNX, MNN, and NCNN).
- **Face Detection Module**
Role: Module Co-Owner.
 - Trained multi-task masked-face detector for surveillance cameras, handling tiny faces, blocking artifacts and occlusions.
 - Participated in building the AI SDK. Optimized and deployed various models to run on Xilinx devices. Involved in building an asynchronous inference flow for multi-stream (using DeepStream), the final SDK can run up to **60 streams** simultaneously on Xilinx ZCU104.
 - Built an object detection visualization tool based on an open-source project to analyze data and model output.
 - Developed a framework to generate pseudo-masks for training datasets using 2D and 3D methods.
- **Traffic Sign/Light Recognition Module for Autonomous Driving**
Role: Module Co-owner.
 - Designed a novel data pipeline based on CVAT to accelerate video dataset labeling, achieving a dataset with *six superclasses and 317 child classes*.
 - Co-managed labeling team to guarantee the data's quality.
 - Developed a hierarchical multi-task model, achieving an F1-score of **98.3** on a private long-tailed dataset of **171 classes**.
 - Addressed varying lighting conditions and implemented a ReID model to enhance traffic sign tracking accuracy.
 - Quantized and deployed models using TensorRT for NVIDIA's device.
- **Other projects**
 - **Noise Cancelling on Smartphone** Responsible for converting models across various frameworks (PyTorch, TensorFlow, ONNX) into TFLite, followed by quantization and smartphone deployment. Optimized existing algorithm with FFT, achieving a **40%** runtime reduction.
 - **SmartData** Redesigned the data labeling pipeline of the backend system built with Flask. Introduced a new end-to-end multi-step labeling feature, improving labeling efficiency by **30%**.

PROFESSIONAL SERVICES

Reviewer: ICCV(2025), ICLR(2025), WACV(2025), NeurIPS(2024), CVPR(2023, 2024, 2025), ECCV(2024), ACCV(2022, 2024).

EDUCATION

Thang Long University

Bachelor of Computer Science; GPA: 9.0/10.0 (Valedictorian)

Vietnam

Aug 2016 - April 2021

CERTIFICATES, HONORS AND AWARDS

Academic Excellence Scholarship

Thang Long University

2016-2021

First Runner-up

VietAI Machine Learning Foundation Hanoi

2020

First Runner-up

Fintech Track, Junction X Hanoi

2018

Rank 76th

ICPC Asia Hanoi Regional Contest

2018

SKILLS SUMMARY

Languages: C++, Python, Unix scripting, SQL

Tools: PyTorch, TensorFlow, TensorRT, ONNX, NCNN, MNN, OpenCV, Docker, Git, Jira

REFERENCES

Dr. Dzung Nguyen: Head of Generative AI & AI Optimization Department, VinAI Research, Vietnam: v.dungnt244@vinai.io

Dr. Anh Tran: Head of Computer Vision Group of Research Department, Vietnam: v.anhtt152@vinai.io

Prof. Minh Hoai Nguyen: Deputy Director of Research Department, VinAI Research, Vietnam: v.hoainm@vinai.io