RUI FANG

rfang@arbor.ee.ntu.edu.tw (+86)13400725807

EDUCATION

National Taiwan University, Graduate Institute of Communication Engineering — 2021 - Now
Ph.D. at Network Database Laboratory. — Supervisor: Ming-Syan Chen

National Taipei University of Technology, Graduate Institute of Manufacturing Technology — 2018 - 2021
M.S. at Advanced Microsystems and Device Laboratory. — Supervisor: Chih-Cheng Lu

National Taipei University of Technology, Collage of Mechanical & Electrical Engineering — 2014 - 2018
B.S. Mechanical Engineering.

PUBLICATIONS & PREPRINTS

A visually interpretable detection method combines 3-D ECG with a multi-VGG neural network for myocardial infarction identification [link]
R. Fang, C. C. Lu, C. T. Chuang, W. H. Chang.
Computer Methods and Programs in Biomedicine (SCI Q1, JCR Q2+Top), 2022.

Dual-Triangular QR Decomposition with Global Acceleration and Partially Q-Rotation Skipping [link]
R. Fang, S. Jiang, H. W. Chen, W. Ding, M. S. Chen.
International Conference on Field-Programmable Technology (ICFPT, CCF-C), 2022.

BiLEE: Bi-Level Early Exiting for Generative Document Retrieval [link]
R. Fang, C. Y. Yeh, H. W. Chen, M. S. Chen.
European Conference on Artificial Intelligence (ECAI, CCF-B), 2024.

Dual Alignment Framework for Few-shot Learning with Inter-Set and Intra-Set Shifts [link]
S. Jiang, **R. Fang**, H. W. Chen, W. Ding, M. S. Chen.
Annual Conference on Neural Information Processing Systems (NeurIPS, CCF-A), 2025.

LoGIC: Multi-LoRA Guided Importance Consensus for Multi-Task Pruning in Vision Transformers [link]
Y.-H. Chou*, **R. Fang***, H.-W. Chen, M.-S. Chen.
Proceedings of the AAAI Conference on Artificial Intelligence (AAAI, CCF-A), 2026.

Learning What to Write: Write-Gated KV for Efficient Long-Context Inference [link]
Y.-C. Huang, **R. Fang**, M.-S. Chen, P.-C. Hsiu.
arXiv preprint arXiv:2512.17452, 2025.

PROFESSIONAL ACTIVITY

Conference Reviewer for CVPR, ICML, IJCAI, ECAI, ECCV, etc.

RESEARCH INTERESTS

Efficient Deep Learning for Transformers (2024–Present)
Efficient computation for Transformers to enable scalable, low-latency inference without sacrificing performance. Including dynamic LLM inference *(BiLEE, ECAI 2024)*, pruning ViT with LoRA-guided importance consensus *(LoGIC, AAAI 2026)*, and KV-cache compression for long-context inference *(WGKV, arXiv 2025)*.

Few-Shot Learning under Distribution Shifts (2023–2024)
Dual alignment and OT-based feature calibration for robust few-shot adaptation under inter-/intra-set shifts *(DuAL, NeurIPS 2025)*.

Interpretable ECG Diagnosis (2019–2021)
Interpretable 3-D ECG deep models for myocardial infarction detection *(CMPB 2022)*.

*Equal contribution.