

# NEIWEN LING

Arthur K. Watson Hall, 51 Prospect Street, New Haven, CT 06511, U.S.

Email:[neiwen.ling@yale.edu](mailto:neiwen.ling@yale.edu) ◊ Homepage:<https://neawhen.github.io/neiwen.github.io/>

## RESEARCH INTERESTS

---

Edge AI, System for AI, Smart City, Mobile Computing, On-device Deep Learning

## EDUCATION

---

**The Chinese University of Hong Kong** 08/2018-07/2022  
Ph.D., Information Engineering  
Supervisor: [Prof. Guoliang Xing](#)  
**Northwestern Polytechnical University** 09/2014-07/2018  
Bachelor's Degree, Electronics and Information Engineering

## RESEARCH EXPERIENCE

---

**Yale University** 11/2023-Present  
Postdoctoral Associate, Department of Computer Science  
Supervisor: [Prof. Lin Zhong](#)  
**The Chinese University of Hong Kong** 08/2022-10/2022  
Postdoctoral Fellow, Department of Information Engineering  
**Shenzhen Institute of Artificial Intelligence and Robotics for Society** 11/2019-01/2020  
Visiting Ph.D. student

## PUBLICATIONS

---

- [1] **Neiwen Ling**, Xuan Huang, Zhihe Zhao, Nan Guan, Zhenyu Yan, and Guoliang Xing, “BlastNet: Exploiting Duo-Blocks for Cross-Processor Real-Time DNN Inference”, The 20th Conference on Embedded Networked Sensor Systems, ACM **SenSys 2022**, 15 pages double column (Acceptance ratio: 52/209=24.88%)  
**Best Paper Award Finalist**
- [2] **Neiwen Ling**, Kai Wang, Yuze He, Guoliang Xing, and Daqi Xie, “RT-mDL: Supporting real-time mixed deep learning tasks on edge platforms”, The 19th Conference on Embedded Networked Sensor Systems, ACM **SenSys 2021**, 14 pages double column (Acceptance ratio:25/139=17.98%)
- [3] Zhehao Jiang\*, **Neiwen Ling**\*, Xuan Huang, Shuyao Shi, Chenhao Wu, Xiaoguang Zhao, Zhenyu Yan, and Guoliang Xing, “CoEdge: A Cooperative Edge System for Distributed Real-Time Deep Learning Tasks”, The 22nd ACM/IEEE Conference on Information Processing in Sensor Networks, ACM/IEEE **IPSN 2023**, 14 pages double column (Acceptance ratio:22/83=26.51%)
- [4] Zhihe Zhao, **Neiwen Ling**, Nan Guan, and Guoliang Xing, “Miriam: Exploiting Elastic Kernels for Real-time Multi-DNN Inference on Edge GPU”, the 21th Conference on Embedded Networked Sensor Systems, ACM **SenSys 2023**, 14 pages double column (Acceptance ratio: 34/179=18.99%)
- [5] Bufang Yang, Lixing He, **Neiwen Ling**, Zhenyu Yan, Guoliang Xing, Xian Shuai, Xiaozhe Ren and Xin Jiang, “EdgeFM: Leveraging Foundation Model for Open-set Learning on the Edge”, the 21th Conference on Embedded Networked Sensor Systems, ACM **SenSys 2023**, 14 pages double column (Acceptance ratio: 34/179=18.99%)
- [6] Zhihe Zhao, Kai Wang, **Neiwen Ling**, and Guoliang Xing, “Edgempl: An automl framework for real-time deep learning on the edge”, The 6th ACM/IEEE Conference on Internet of Things Design and Implementation, ACM/IEEE **IoTDI 2021**, 12 pages double column (Acceptance ratio: 19/74=25.7%)
- [7] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, **Neiwen Ling**, Guoliang Xing, Jiayu Zhou, and Jianwei Huang, ”Harmony: Heterogeneous Multi-Modal Federated Learning through Disentangled Model Training”,

---

<sup>1</sup>\* Equal contribution

The 21st ACM International Conference on Mobile Systems, Applications, and Services, ACM **MobiSys 2023**, 14 pages double column (Acceptance ratio: 41/198=20.71%)

[8] Wenjing Xie, Tao Hu, **Neiwen Ling**, Guoliang Xing, Shaoshan Liu, and Nan Guan, “Timely Fusion of Surround Radar/Lidar for Object Detection in Autonomous Driving Systems”, 2023 Design, Automation & Test in Europe Conference & Exhibition, IEEE **DATE 2023**, 6 pages double column

[9] Zhihe Zhao, Xian Shuai, **Neiwen Ling**, Nan Guan, Zhenyu Yan, and Guoliang Xing, ”Moses: Exploiting Cross-device Transferable Features for On-device Tensor Program Optimization”, The 24th International Workshop on Mobile Computing Systems and Applications 2023, ACM **HotMobile 2023**, 7 pages double column

[10] **Neiwen Ling\***, Yuze He\*, Nan Guan, Heming Fu, and Guoliang Xing, “Dataset: An Indoor Smart Traffic Dataset and Data Collection System”, The 5th International SenSys/BuildSys Workshop on Data, ACM **DATA 2022**, 4 pages double column

[11] Zhihe Zhao, Zhehao Jiang, **Neiwen Ling**, Xian Shuai, and Guoliang Xing, “Demo Abstract: Ecart: an edge computing system for real-time image-based object tracking”, The 16th ACM Conference on Embedded Networked Sensor Systems, ACM **SenSys 2018**, 2 pages double column

[12] Zhihe Zhao, **Neiwen Ling**, Nan Guan, and Guoliang Xing, “Poster Abstract: Aaron: Compile-time Kernel Adaptation for Multi-DNN Inference Acceleration on Edge GPU”, The 20th Conference on Embedded Networked Sensor Systems, ACM **SenSys 2022**, 2 pages double column

### **Best Poster Award**

[13] Zhihe Zhao, **Neiwen Ling**, Kaiwei Liu, Nan Guan, and Guoliang Xing, “Poster Abstract: Unifying On-device Tensor Program Optimization through Large Foundation Model”, The 21th Conference on Embedded Networked Sensor Systems, ACM **SenSys 2023**, 2 pages double column

## **WORK IN PROGRESS**

---

[1] Shuyao Shi\*, **Neiwen Ling\***, Zhehao Jiang\*, Xuan Huang\*, Yuze He, Xiaoguang Zhao, Bufang Yang, Chen Bian, Jingfei Xia, Zhenyu Yan, and Guoliang Xing, ”Soar: Design and Deployment of A Smart Roadside Infrastructure System for Autonomous Driving”

Full-length paper under review by a conference.

## **HONORS & AWARDS**

---

- Best Paper Award Finalist, ACM SenSys (top 3.3%) 2022
- Best Poster Award, ACM SenSys (top 1.25%) 2022
- N2Women Young Researcher Fellowship, ACM SenSys 2021
- Postgraduate Scholarship, The Chinese University of Hong Kong 2018-2022
- Undergraduate Excellent Graduation Project, Northwestern Polytechnical University 2018

## **SUMMARIES OF RESEARCH**

---

### **Supporting Real-Time Mixed Deep Learning Tasks on Edge Platforms.**

Edge platforms need to execute a set of real-time mixed Deep Learning (DL) tasks concurrently to support real-time applications, e.g., autonomous driving. Such an application paradigm poses major challenges due to the huge compute workload of deep neural network models, diverse performance requirements of different tasks, and the lack of real-time support from existing DL frameworks. We design a framework named RT-mDL to optimize the mixed DL task execution by exploiting unique compute characteristics of DL tasks. Our implementation on an F1/10 autonomous driving testbed shows that, RT-mDL can enable multiple concurrent DL tasks to achieve satisfactory real-time performance. This work is published at SenSys’21.

### **Exploiting Duo-Blocks for Cross-Processor Real-Time DNN Inference.**

Deep Neural Network (DNN) has been increasingly adopted by a wide range of time-critical applications running on edge platforms with heterogeneous CPU-GPU multiprocessors. Such a cross-processor real-time DNN inference

paradigm poses major challenges due to the inherent performance imbalance among different processors and the lack of real-time support for cross-processor inference from existing deep learning frameworks. We propose a system named BlastNet that exploits *duo-block - a new model inference abstraction* to support highly efficient cross-processor real-time DNN inference. Each duo-block has a dual model structure, enabling efficient fine-grained inference alternatively across different processors. We have implemented BlastNet based on NNI and LibTorch. Extensive evaluations on an indoor autonomous driving platform and three popular edge platforms show the effectiveness of our approach. This work is published at SenSys'22 and is selected as the Best Paper Award Finalist.

### **A Cooperative Edge System for Distributed Real-Time Deep Learning Tasks.**

Recent years have witnessed the emergence of a new class of *cooperative edge systems* in which a large number of edge nodes can collaborate through local peer-to-peer connectivity. We design a novel cooperative edge system named CoEdge that can support concurrent data/compute-intensive deep learning (DL) models for distributed real-time applications such as city-scale traffic monitoring and autonomous driving. The design of CoEdge is based on a hierarchical DL task scheduling framework, which implements global task dispatching and batched DNN execution on local edge nodes. CoEdge also includes a new edge containerization approach that enables concurrent DL task execution on edge GPU nodes. We have implemented CoEdge based on KubeEdge, ROS2 and TensorRT. We also extensively evaluate CoEdge on a self-deployed smart lamppost testbed on a university campus. This work is accepted by IPSN'23.

### **Design and Deployment of a Smart Roadside Infrastructure System for Autonomous Driving.**

Smart roadside infrastructure (SRI) has demonstrated the potential of achieving fully autonomous driving systems. To explore the potential of infrastructure-assisted autonomous driving, we design and deploy the first end-to-end SRI system specifically designed for supporting autonomous vehicles. This system consists of carefully designed software and hardware components to overcome various system and physical challenges, which allows leveraging the existing operational infrastructure like street lampposts and hence lowers the barrier of adoption. This system also features a hierarchical DL task management framework that integrates cluster-level task dispatching and a local opportunistic DNN execution mechanism. This design achieves desirable load balancing among nodes, enabling them to collaborate efficiently to run multiple data-intensive autonomous driving applications. We deployed this system on a total of 18 existing outdoor lampposts on campus, which have been operational for over two years. This work is under review by a top conference.

## **PROFESSIONAL SERVICES**

---

- **Technical Program Committee**

The Eighth IEEE/ACM Conference on Connected Health: Applications, Systems, and Engineering Technologies (CHASE 2023)

Poster & Demo, The ACM/IEEE International Conference on Internet of Things Design and Implementation 2023 (IoTDI 2023 Poster & Demo)

- **Organizer**

N2Women (Networking Networking Women) Meeting in the 19th ACM Conference on Embedded Networked Sensor Systems (SenSys 2021)

- **Invited reviewer**

IEEE Transactions on Mobile Computing (TMC)

Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT/UbiComp)

The First Workshop on DL-Hardware Co-Design for AI Acceleration in the 37th AAAI Conference on Artificial Intelligence (DCAA 2023)

SCIENCE CHINA Information Sciences (SCIS)

ACM Transactions on Sensor Networks (TOSN)

IEEE International Conference on Computer Communications (INFOCOM 2024)

IEEE Network Magazine

ACM Transactions on Computing for Healthcare (HEALTH)

## **PROFESSIONAL SKILLS**

---

- **Programming Language**

Python, C/C++ Language, Verilog HDL, Assembly Language

- **Development Experience on Hardware**

NVIDIA Jetson TX2, NVIDIA AGX Xavier, NVIDIA Jetson Nano, FPGA, Arduino, 51 SCM, STM32F1, STM32F4

- **Development Experience on Software**

PyTorch, LibTorch, NNI, TensorRT, KubeEdge, ROS2

## **TEACHING ASSISTANT**

---

ENGG1100: Introduction to Engineering Design, 2018 Fall, CUHK

IERG4230: Introduction to Internet of Things, 2019 Spring, CUHK

ENGG1110: Problem Solving by Programming, 2019/2020/2021 Fall, CUHK

IERG2602: Engineering Practicum, 2020/2021 Spring, CUHK