Energy- and Latency-Efficient Architectures, Chips, and Integrated Systems towards Ubiquitous Edge Intelligence

Research Overview

Problem Statement. The recording-breaking performance of artificial intelligence (AI) algorithms, especially deep neural networks (DNNs), has motivated a growing demand for bringing powerful AI-powered intelligent functionalities onto edge devices, e.g., virtual reality/augmented reality (VR/AR) and medical devices, towards ubiquitous edge intelligence. However, the powerful performance of AI algorithms comes with much increased computational complexity and memory storage requirements, which stand at odds with the limited compute/storage resources on edge devices. Additionally, the stringent application-specific requirements, including real-time response (i.e., high throughput/low latency), high energy efficiency, and small form factor, further aggravate the aforementioned gap.

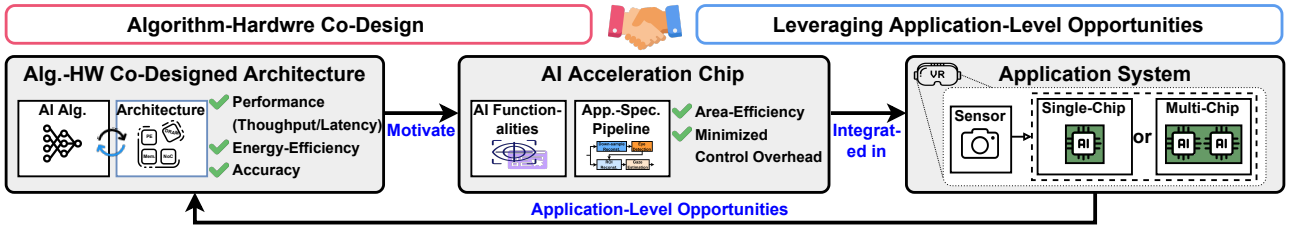


Figure 1: My holistic solutions from efficient architectures, to chips, and to integrated systems.

Research Goals and Deliveries. My research aims at closing the above-mentioned gap between the tremendous resource requirements of powerful AI algorithms and the constrained resources on edge devices to enable ubiquitous edge intelligence. To tackle this, my research delivers holistic solutions from energy- and latency-efficient **architectures** [1–3], to **chips** [4, 5], and to **integrated systems** [4, 6, 7]. Furthermore, my research works share the same underlying design insight, which is to advocate simultaneously *harmonizing dedicated algorithms and architectures/chips/integrated-systems via* ***algorithm-hardware co-design*** and ***leveraging application-level opportunities*** *to minimize redundant computations and/or data movements in the processing pipeline and thus boost the achievable efficiency*, as categorized below:

- **Algorithm-Hardware Co-designed Architecture:** One major efficiency bottleneck in AI acceleration is the massive and high-cost data movements [8]. Our algorithm-hardware co-design work, called *SmartExchange* [1, 2], trades higher-cost memory storage/accesses for lower-cost computations to boost the energy- and latency-efficiency.
- **AI Acceleration Integrated Chip:** Motivated by the promising efficiency achieved by *SmartExchange*, we further validated its co-designed architecture by designing an AI acceleration chip prototype, while taking the chip's area efficiency and control overhead into design considerations.
- **Single-Chip Integrated System:** To demonstrate the real-efficiency of the above *SmartExchange* architecture and its chip prototype, we integrated the AI acceleration chip with a lensless camera (i.e., FlatCam [9]) to develop the first-of-its-kind real-time eye-tracking system, called *i*-FlatCam [4], targeting eye tracking in next-generation VR/AR devices [10], where the application-level opportunities were leveraged to reduce both spatial and temporal redundancy in eye images.
- **Multi-Chip Integrated System:** We also built a scaled-up eye-tracking system, called EyeCoD [6], utilizing multiple AI acceleration chips for enabling more general eye-tracking solutions.

Publications and Highlights. My research has led to over 20 publications (**11 as the 1st author**) in top-tier **computer architecture and circuit design conferences/journals**, including **ISCA**, **VLSI**, **MICRO**, HPCA, ICCAD, ICASSP, FPGA, TNNL, TCAD, TVLSI, etc, and **won 1st place demonstration at the 32nd ACM SIGDA University Demonstration at DAC'22** [7]. Additionally, I have been a significant contributor to **5 research projects** (4 funded by *NSF* and 1 funded by *NIH*), and am a recipient of the **2020 Cadence Women in Technology Scholarship** (1 of 13 winners nationwide) [11]. Finally, I have served as a regular reviewer for top-tier journals, e.g., TCAS-II and TNNL, and was one of the four key contributors/speakers in the **AutoDL tutorial** at MICRO'22 [12].

Completed Research

***SmartExchange* - Algorithm-Hardware Co-Designed Architecture (ISCA'20):** The huge amount of parameters and intermediate data in DNNs need external DRAM for storage in edge devices. However, the prohibitive cost of massive DRAM accesses, whose unit energy is two orders of magnitude higher than the corresponding computation operations, limits the achievable acceleration efficiency of DNNs, calling for innovations to minimize the required data movements and boost efficiency. To tackle this, we proposed an algorithm-hardware co-design technique dubbed *SmartExchange* [1], which trades higher-cost memory storage/accesses for lower-cost computations, to boost the acceleration efficiency of both DNN inference and training. Specifically, on the algorithm level, a hardware-friendly DNN weight structure was enforced, where only a subset of parameters is stored in DRAM for each layer, and the remaining majority of weights can be recovered from lower-cost computations when needed (see Fig. 2); On the hardware level, I further designed a dedicated architecture to fully leverage the *SmartExchange*-enforced algorithm structure, including minimizing the overhead of rebuilding weights and taking advantage of the structured sparsity to improve energy- and latency-efficiency. Experiments show that *SmartExchange* can boost energy- and latency-efficiency by up to 6.7× and 19.2×, respectively, over 4 state-of-the-art (SOTA) DNN accelerators when benchmarked on 7 DNN models and 3 datasets.

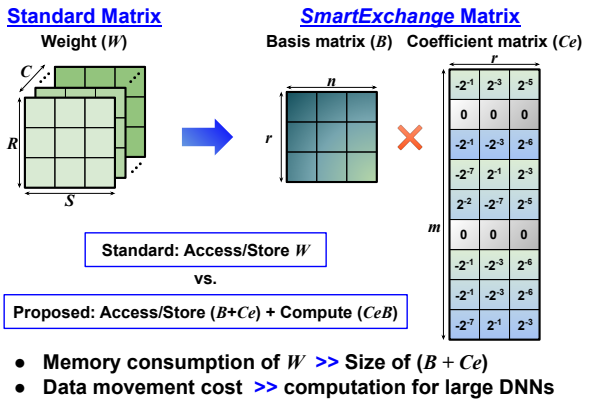


Figure 2: The proposed *SmartExchange*'s weight representation.

AI Acceleration Chip (VLSI'22): Motivated by the promising results of the above-mentioned *SmartExchange*, I implemented its architecture into an AI acceleration chip prototype. First, our acceleration chip focuses on optimizing depth-wise convolution layers (DWs), which are widely used in parameter-efficient DNNs but may not lead to real hardware efficiency due to their low hardware utilization [4]. Second, our acceleration chip takes area efficiency and control overhead into consideration to balance the area overhead with hardware efficiency while minimizing the control overhead. To achieve these goals, our AI chip features *(i)* a dedicated heterogeneous dataflow for both general convolutional layers (CONVs) and DWs to enhance hardware utilization, *(ii)* a sequential-write-parallel-read (SWPR) input activation feature map (IFM) buffer design to provide a 2× higher FM memory bandwidth for better leveraging the structure sparsity in the *SmartExchange* algorithm with a negligible area overhead, and *(iii)* a customized instruction set architecture (ISA) to support the above optimizations. Specifically, *(i)* for the hybrid dataflow, the acceleration chip leverages the intra-channel dataflow for DWs while using the inter-channel dataflow for other layers (e.g., CONVs and point-wise convolution layers), boosting the compute resource utilization by 75∼87.5% for DWs. A reconfigurable feature map (FM) global buffer (GB) storage and weight buffer designs are developed to support the dataflow. *(ii)* The structure sparsity in the *SmartExchange* algorithm allows for skipping both corresponding computations and GB weight accesses but at the cost of a higher FM GB bandwidth, which will increase the chip area and leads to bandwidth waste when processing other layers. The SWPR IFM buffer design is inserted between the FM GB and compute resource to provide a 2× higher bandwidth, incurring a negligible area overhead of 0.58%. *(iii)* The customized ISA explores the parallel and repetitive processing operations in DNN processing to support all the optimizations adopted by the chip efficiently.

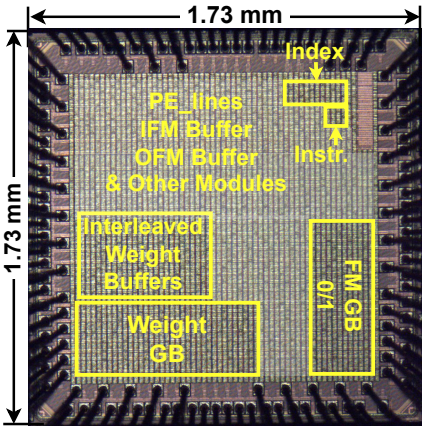
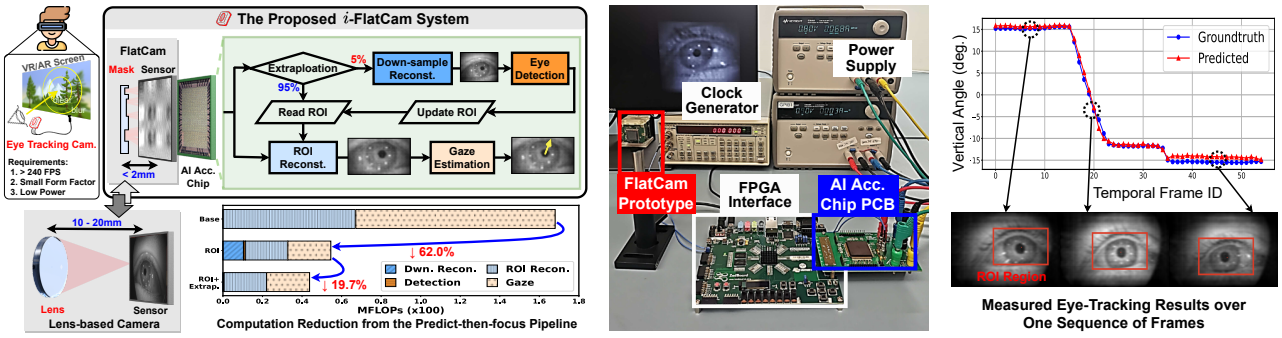


Figure 3: The die photo of our AI acceleration chip.



(a) (b) (c)

Figure 4: *i*-FlatCam's (a) predict-then-focus pipeline, (b) measured setup, and (c) measured results.

System Integration of Our AI Acceleration Chip for a Real-world Application (VLSI'22): We integrated the above AI acceleration chip with a lensless camera (i.e., FlatCam [9]) to build an eye-tracking system, called *i*-FlatCam, for enabling a highly demanded AI-powered functionality, i.e., eye tracking which estimates the gaze directions of human eyes, in VR/AR devices. Real-hardware measurements show that our *i*-FlatCam system is the first to simultaneously meet all three requirements of eye tracking required by next-generation VR/AR devices, including real-time throughput (i.e, >240FPS), milli-watt power consumption, and small form factor as recently pointed out by Meta [10]. As a highlight, our *i*-FlatCam system further *leverages the application-level opportunity* that spatial and temporal redundancy exists in input eye images to reduce redundant computations and costly data movements. Specifically, to reduce spatial redundancy in eye images, we proposed a dedicated predict-then-focus pipeline that first extracts region-of-interests (ROIs), which comprise only 24% (average) of the original eye images for gaze estimation, to reduce unnecessary spatial information, using an *eye detection* model. To reduce temporal redundancy, the temporal correlation of eyes across frames is leveraged so that only 5% of the frames require ROIs adjustment over time. To enable this predict-then-focus pipeline, our chip integrates dedicated instructions in its ISA. Notably, our *i*-FlatCam system won **1st place demonstration at the 32nd ACM SIGDA University Demonstration at Design Automation Conference (DAC), 2022 [7]**.

Scaled-Up System Integration of Multiple AI Acceleration Chips (EyeCoD, ISCA'22): In addition to the single-chip *i*-FlatCam system, we took another big leap towards accelerating *eye segmentation*-based eye tracking functionality for enabling more general eye tracking in VR/AR. The motivation is that segmentation can provide richer features to enable more follow-up tasks needed in VR/AR towards practical and general VR/AR uses [13]. However, the higher-complexity segmentation model brings about three challenges for the underlying hardware: *(i)* a much larger model complexity, i.e., 80.1× more parameters and 186.7× more operations than the previously-adopted detection model, *(ii)* a more diverse model structure, i.e., U-Net [14] type for segmentation versus a MobileNet [15] type for detection, and *(iii)* a higher input image resolution. To address these challenges, we developed a scaled-up architecture and multi-chip system focusing on orchestrating the required segmentation and gaze estimation models, without significantly increasing the required chip and memory bandwidth, and thus the power consumption. Specifically, the multi-chip architecture includes a dedicated: *(i)* workload orchestration between the eye segmentation and gaze estimation models and *(ii)* activation FM partition method and activation FM GB storage arrangement for all involved layer types (e.g., downsampling and upsampling layers) to favor the workload orchestration and reduced activation FM GB size. Our scaled-up system can achieve the required real-time throughput requirement (i.e., >240FPS), with only 2.68×/2.17× area/power of that in the single-chip system for enabling general VR/AR uses [13].

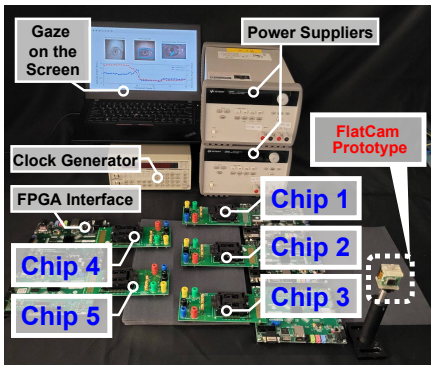


Figure 5: EyeCoD's multi-chip setup.

Future Research

Looking forward, I strongly believe that our holistic perspective from **architectures**, to **chips**, and to **integrated systems** will be one key driving force to enable ubiquitous edge intelligence, and my research principle of marrying *algorithm-hardware co-design techniques* with *application-level opportunities* can lead to much-boosted efficiency and thus enable **more emerging AI-powered intelligent applications**. Additionally, my research principle can be extended to consider more real-world application **metrics**, e.g., **both robustness and efficiency**, and provide indispensable insights for designing **automation tools** that can reduce the design complexity, time-to-market, labor cost, and risks. Looking ahead, I am very excited to expand my research principle and expertise into the following research directions towards ubiquitous AI-powered edge intelligence.

Enabling More Emerging AI-powered Applications: AI algorithms continue to revolutionize an ever wider range of emerging applications with an unprecedented amount of computations, parameters, and intermediate data [16], while their corresponding AI-powered applications demand stringent energy- and/or latency-efficiency to be deployed edge devices for unleashing their big promise. My research principle can be expanded to enable emerging AI-powered applications. For example, 3D reconstruction, i.e., the reconstruction of a 3D scene to generate images of any arbitrary views given images from a set of sparsely sampled views of a scene, is a representative application that has recently achieved breakthroughs in rendering quality thanks to the adopted AI algorithm called Neural Radiance Field (NeRF). My prior and ongoing works in accelerating NeRF are as follows:

- **Real-Time On-Device NeRF Inference (RT-NeRF; ICCAD'22)**: To enable immersive real-time (>30FPS) NeRF inference, we first profile and analyze the throughput bottlenecks of a SOTA efficient NeRF algorithm and then propose RT-NeRF [16], the first algorithm-hardware co-design technique to tackle the identified bottlenecks. Specifically, on the algorithm level, RT-NeRF integrates an efficient rendering pipeline for largely alleviating irregular memory accesses by directly computing the geometry of pre-existing points; On the hardware level, we proposed *(i)* a hybrid encoding scheme, aiming to maximize the storage savings and thus reduce the required DRAM accesses and *(ii)* a high-density sparse search unit and a dual-purpose bi-direction adder & search tree to coordinate the hybrid encoding scheme. Extensive experiments on 8 datasets consistently validate the real-time performance of RT-NeRF while maintaining the rendering quality.
- **Instant-NeRF for Instant On-Device NeRF Training via Near-Memory Processing**: To tackle the memory-bounded training time bottleneck in NeRF unveiled by our profiling, we propose Instant-NeRF [17], the first near-memory processing (NMP) architecture for NeRF training via algorithm-hardware co-design. Specifically, our Instant-NeRF's algorithm adopts a locality-sensitive hashing function to enhance the locality of hash table lookups and a ray-first point processing sequence to replace the original random sequence; Our instant-NeRF's architecture integrates a dedicated mapping scheme optimized for Instant-NeRF's algorithm and a heterogeneous inter-bank parallelism design, orchestrating the diverse computation and memory patterns in NeRF's heterogeneous training steps, to minimize the inter-bank data movement overhead. Extensive experiments on 8 datasets verify that Instant-NeRF provides a 22.0$\sim$266.1$\times$ speedup over SOTA edge GPUs.

Enabling Both Robustness and Efficiency Towards Real-World Intelligent Systems: Real-world applications require both efficiency and robustness, the latter of which is because real-world systems are vulnerable to malicious hardware modifications (i.e., hardware Trojans), erroneous inputs, and execution-time errors. Specifically, for AI-powered applications, hardware Trojans can lead a system to malfunction after being triggered; Adversarial examples can fool the models to degrade accuracy or even crash the systems. My prior work inspired and prepared me for this direction is as follows:

- **Memory Trojan Attack on DNN Accelerators (Memory Trojan; TCAD'20, DATE'19)**: The development of practical attacks is the prerequisite for defending robustness in AI-powered systems. Previous works introduce hardware Trojan attacks in the scope of DNN accelerators, which require a strict assumption that the adversary has access to the DNN models, toolchains (i.e., algorithm-to-hardware mapping tools), and hardware accelerators. In this work [18, 19], we developed a novel

hardware Trojan inserted within the memory controller, where the Trojan can only monitor the memory access patterns and modifies the data written back to external storage after being triggered. Such a hardware Trojan is much more practical and even works well with environmental noise and/or preprocessing on the original images, potentially inspiring more practical robust defense and robustness-efficiency co-designed methods.

Developing Automated Tools to Facilitate Fast Development of Efficient AI Solutions: Fast and accurate performance estimation of AI accelerators with various hardware optimization techniques is one key enabler [20] for developing automated co-design and co-search tools. My prior work that has inspired and prepared me for in this direction is as follows:

- ***DNN-Chip Predictor* (ICASSP'20)**: I developed and released the first-of-its-kind analytical simulator of AI accelerator chips called *DNN-Chip Predictor* [3] by bridging hardware design principles and mathematical models, which can efficiently and effectively predict an accelerator's energy, latency, and resource consumption prior to its actual implementation and has been deployed by following-up work [20]. *DNN-Chip Predictor* features two highlights: *(i)* its analytical performance formulation of DNN ASIC/FPGA accelerators facilitates fast design space exploration and optimization and *(ii)* it supports DNN accelerators with different dataflows and hardware architectures. Experiment results based on two DNN models and three different ASIC/FPGA implementations show that our *DNN-Chip Predictor*'s predicted performance differs from those of chip measurements of FPGA/ASIC implementation by no more than 17.66% when using different DNN models, hardware architectures, and dataflows.

References

[1] **Y. Zhao**, X. Chen, Y. Wang, C. Li, H. You, Y. Fu, Y. Xie, Z. Wang, and Y. Lin, "*SmartExchange*: Trading Higher-cost Memory Storage/Access for Lower-cost Computation," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, 2020.

[2] X. Chen, **Y. Zhao**, Y. Wang, P. Xu, H. You, C. Li, Y. Fu, Y. Lin, and Z. Wang, "*SmartDeal*: Remodeling Deep Network Weights for Efficient Inference and Training," *IEEE Transactions on Neural Networks and Learning Systems (TNNL)*, 2021.

[3] **Y. Zhao**, C. Li, Y. Wang, P. Xu, Y. Zhang, and Y. Lin, "*DNN-Chip Predictor*: An Analytical Performance Predictor for DNN Accelerators with Various Dataflows and Hardware Architectures," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[4] **Y. Zhao**, Z. Li, Y. Fu, Y. Zhang, C. Li, C. Wan, H. You, S. Wu, X. Ouyang, V. Boominathan *et al.*, "*i*-FlatCam: A 253 FPS, 91.49 μJ/Frame Ultra-Compact Intelligent Lensless Camera for Real-Time and Efficient Eye Tracking in VR/AR," in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI)*, 2022.

[5] **Y. Zhao**, Y. Zhang, Y. Fu, X. Ouyang, C. Wan, S. Wu, A. Banta, M. M. John, A. Post, M. Razavi *et al.*, "e-G2C: A 0.14-to-8.31 μJ/Inference NN-based Processor with Continuous On-chip Adaptation for Anomaly Detection and ECG Conversion from EGM," in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI)*, 2022.

[6] H. You, C. Wan, **Y. Zhao**, Z. Yu, Y. Fu, J. Yuan, S. Wu, S. Zhang, Y. Zhang, C. Li *et al.*, "EyeCoD: Eye Tracking System Acceleration via Flatcam-based Algorithm & Accelerator Co-Design," in *2022 ACM/IEEE 49th Annual International Symposium on Computer Architecture (ISCA)*, 2022.

[7] DAC2022, "32nd ACM SIGDA University Demonstration at Design Automation Conference (DAC), 2022," https://www.sigda.org/sigda-events/ubooth/.

[8] Z. Du, R. Fasthuber, T. Chen, P. Ienne, L. Li, T. Luo, X. Feng, Y. Chen, and O. Temam, "ShiDianNao: Shifting Vision Processing Closer to the Sensor," in *Proceedings of the 42nd Annual International Symposium on Computer Architecture (ISCA)*, 2015.

[9] M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk, "FlatCam: Thin, Lensless Cameras using Coded Aperture and Computation," *IEEE Transactions on Computational Imaging*, 2016.

[10] C. Liu, A. Berkovich, S. Chen, H. Reyserhove, S. S. Sarwar, and T.-H. Tsai, "Intelligent Vision Systems–Bringing Human-Machine Interface to AR/VR," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019.

[11] Cadence, "2020 Cadence Women in Technology Scholarship, 2020," https://community.cadence.com/cadence_blogs_8/b/can/posts/meet-the-2020-women-in-technology-scholarship-recipients.

[12] Y. Lin, Y. Zhang, Y. Fu, C. Li, and **Y. Zhao**, "AutoDL: Automated Tools for Fast Development of Deep Learning Networks and Accelerators at 55th IEEE/ACM International Symposium on Microarchitecture (MICRO), 2022," https://sites.google.com/rice.edu/auto-dl/home?authuser=0.

[13] Meta, "OpenEDS Challenge's Semantic Segmentation Challenge, 2020," https://research.facebook.com/openeds-challenge/.

[14] A. K. Chaudhary, R. Kothari, M. Acharya, S. Dangi, N. Nair, R. Bailey, C. Kanan, G. Diaz, and J. B. Pelz, "RITNet: Real-Time Semantic Segmentation of the Eye for Gaze Tracking," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.

[16] C. Li, S. Li, **Y. Zhao**, W. Zhu, and Y. Lin, "RT-NeRF: Real-Time On-Device Neural Radiance Fields Towards Immersive AR/VR Rendering," in *2022 International Conference on Computer-Aided Design (ICCAD)*, 2022.

[17] **Y. Zhao**, W. Shang, J. Zhang, S. Li, C. Li, and Y. Lin, "Instant-NeRF: Instant On-Device Neural Radiance Field Training via Algorithm-Accelerator Co-Designed Near-Memory Processing," 2023.

[18] **Y. Zhao**, X. Hu, S. Li, J. Ye, L. Deng, Y. Ji, J. Xu, D. Wu, and Y. Xie, "Memory Trojan Attack on Neural Network Accelerators," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2019.

[19] X. Hu, **Y. Zhao**, L. Deng, L. Liang, P. Zuo, J. Ye, Y. Lin, and Y. Xie, "Practical Attacks on Deep Neural Networks by Memory Trojaning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2020.

[20] C. Li, Z. Yu, Y. Fu, Y. Zhang, **Y, Zhao**, H. You, Q. Yu, Y. Wang, and Y. Lin, "HW-NAS-Bench: Hardware-Aware Neural Architecture Search Benchmark," *Initional Conference on Learning Representations, (ICLR)*, 2021.