# Citation Evidence Report

EB-1A Petition — Original Contributions of Major Significance

8 CFR § 204.5(h)(3)(v) · Criterion 5

## Tien-Ju Yang

Massachusetts Institute of Technology

Google Scholar profile

**Generated 2026-06-08 by CiteMap.** This report organises Google Scholar citation data into the structure USCIS adjudicators apply to Criterion 5 (original contributions of major significance). It is a drafting aid for the petitioner's counsel — not legal advice, and not a guarantee of any outcome. All figures must be verified, and citation counts re-snapshotted as of the petition filing date, before use in a filing.

## A. Overview & Filtering Statement

| 4,097 | 4,274 | 30 | 19 |
|---|---|---|---|
| Citing papers mapped | Citation edges | Home papers mapped | h-index (GS) |

**Filtering statement — methodology & limits**

Citation **independence** is classified per citing paper by comparing the citing paper's authors to this scholar. *Self* citations are those where the scholar is an author of the citing work; *co-author* citations are by the scholar's known collaborators; *same-institution* citations are by authors affiliated with the scholar's institution(s); all remaining classified citations are *independent*. Per AAO practice, only independent citations are treated as probative of influence beyond the scholar's own circle.

**Known limitations — counsel must verify.** (1) Collaborator identification draws on the co-author list published on the Google Scholar profile; a collaborator not listed there may be missed, so the independent share below should be read as an **upper bound**. (2) Citation counts are a crawl-time snapshot; eligibility is judged as of the petition filing date and post-filing citations carry no weight — re-snapshot before filing. (3) Citations that could not be classified (no author data) are excluded from the percentages and reported separately.

## B. Citation Independence

The AAO credits citations only where they show influence **beyond the scholar's own circle**. Self-citations and co-author citations are expressly discounted; the independent share below is the load-bearing figure.

**97.6% independent** of 3,742 classified citing papers

| Citation type | Count |
|---|---|
| Independent | 3,653 |
| Self-citation | 25 |
| Co-author | 64 |
| Same-institution | 0 |

355 citing papers could not be classified (no author data) and are excluded from the percentages above.

## C. Significant Contributions & Their Citation Evidence

Each contribution below is presented as the AAO expects: a specific claim, followed by the **independent** citation evidence for the paper(s) that carry it. Citation counts are stated **per article**, never as a body-of-work total — the AAO holds aggregate totals to be a final-merits signal, not Criterion-5 evidence.

Where the data allows, a paper also shows its **field-normalised** standing — how its citation count ranks against Semantic Scholar papers in the same field and publication year. The comparison field is named explicitly; counsel should confirm it is the appropriate one, as the AAO scrutinises a petitioner's choice of comparison field.

## Contribution 1

### Claim — Contribution 1

*The researcher established a foundational framework for efficient deep neural network processing, subsequently advancing mobile-specific hardware accelerators and platform-aware adaptation techniques.*

CLAIM: The researcher's contribution centers on the efficient processing of deep neural networks, anchored by the seminal 2017 survey and tutorial. This core work is extended by subsequent publications addressing flexible hardware acceleration and platform-aware adaptation for mobile applications.

ORIGINALITY: The titles suggest a progression from general efficiency principles to specialized solutions for emerging constraints. The 2017 core paper appears to address the broad challenge of processing efficiency, while the 2018 and 2019 follow-ups indicate a targeted innovation in adapting these networks for mobile devices through flexible accelerators and platform-aware methods.

SIGNIFICANCE: The core paper has accumulated over 6,000 citations, indicating substantial uptake. With 97.6% of classified citations originating from independent researchers, the work demonstrates broad, field-wide impact beyond the researcher's immediate circle. The follow-up papers, with over 1,400 and 800 citations respectively, further confirm the sustained relevance of this research line.

**INDEPENDENT CITATIONS FOR THIS CONTRIBUTION: 2,177** · **163 flagged influential by Semantic Scholar**

**CORE PAPER**

### Efficient processing of deep neural networks: A tutorial and survey

2017 · 6,021 citations (GS)

Field-normalised: 3,531 Semantic Scholar citations place it in the top 1% of Computer Science papers from 2017 indexed by Semantic Scholar, by citation count.

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 1 | A survey of the recent architectures of deep convolutional neural networks | Pakistan Institute of Engineering and Applied Sciences, Pakistan Institute of Engineering & Applied Sciences, PIEAS | Pakistan | — |
| 2 | CNN variants for computer vision: History, architecture, application, challenges and future scope | Linnaeus University, Parul University, Sankalchand Patel University | India, Sweden | — |
| 3 | Deep reinforcement learning: An overview | — | — | — |
| 4 | Vision-language models for edge networks: A comprehensive survey | Mohamed bin Zayed University of Artificial Intelligence | United Arab Emirates | — |
| 5 | Deep learning methods for autonomous driving scene understanding tasks: A review | Gachon University, National University of Sciences and Technology, Sungkyunkwan University | Pakistan, South Korea | — |
| 6 | Photonics for artificial intelligence and neuromorphic computing | Princeton University, Queen's University, University of Exeter | Canada, Germany, United Kingdom | — |
| 7 | Memory devices and applications for in-memory computing | IBM Research - Zurich | Switzerland | — |
| 8 | Federated learning in mobile edge networks: A comprehensive survey | Hong Kong University of Science and Technology, | Australia, China, Hong Kong | — |

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| | | Nanyang Technological University, Phenikaa (Vietnam) | | |
| 9 | Objects as points | Google DeepMind, The University of Texas at Austin | United Kingdom, United States | — |
| 10 | Edge intelligence: Paving the last mile of artificial intelligence with edge computing | Arizona State University, Sun Yat-sen University | China, United States | — |
| 11 | Pruning and quantization for deep neural network acceleration: A survey | University of Science and Technology Beijing | China | Influential |
| 12 | A state-of-the-art survey on deep learning theory and architectures | Comcast, Lawrence Livermore National Laboratory, Saint Louis University | United States | — |
| 13 | Application and theory gaps during the rise of artificial intelligence in education | Education University of Hong Kong, Hong Kong Polytechnic University, Lingnan University | China, Hong Kong, Taiwan | — |
| 14 | Convolutional neural network: a review of models, methodologies and applications to object detection | National Institute of Technology Kurukshetra | India | Influential |
| 15 | Deep learning in mobile and wireless networking: A survey | Imperial College London, Microsoft, University of Edinburgh | United Kingdom | — |
| 16 | Model compression and hardware acceleration for neural networks: A comprehensive survey | Massachusetts Institute of Technology, Tsinghua University, University of California, Irvine Medical Center | China, United States | — |
| 17 | Human Activity Recognition (HAR) Using Deep Learning: Review, Methodologies, Progress and Future Research Directions: P. Kumar et al. | National Institute of Technology Hamirpur | India | — |
| 18 | An optical neural chip for implementing complex-valued neural network | Nanyang Technological University | Singapore | — |
| 19 | A survey on deep learning for multimodal data fusion | Dalian University of Technology | China | — |
| 20 | A survey on deep learning for big data | Dalian University of Technology, Hainan University, St. Francis Xavier University | Canada, China | — |
| 21 | Bridging biological and artificial neural networks with emerging neuromorphic devices: fundamentals, progress, and challenges | Tsinghua University, University of California, Irvine Medical Center, University of Massachusetts | China, United States | — |
| 22 | Deep learning with spiking neurons: Opportunities and challenges | Robert Bosch GmbH | Germany | — |
| 23 | A survey on kolmogorov-arnold network | Texas State University | United States | — |
| 24 | A comprehensive review of model compression techniques in machine learning: PV Dantas et al. | The University of Manchester, Universidade Federal do Amazonas | Brazil, United Kingdom | — |

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 25 | Efficient acceleration of deep learning inference on resource-constrained edge devices: A review | Analog Devices (United States), Texas Tech University, University of Missouri | United States | **Influential** |
| 26 | Deep learning with edge computing: A review | University of California, Irvine Medical Center | United States | — |
| 27 | Review of Lightweight Deep Convolutional Neural Networks: F. Chen et al. | Lanzhou University | China | — |
| 28 | Efficient deep learning: A survey on making deep learning models smaller, faster, and better | Google Research | United States | — |
| 29 | The history began from alexnet: A comprehensive survey on deep learning approaches | Lawrence Livermore National Laboratory, Saint Louis University, St. Jude Children's Research Hospital | United States | — |
| 30 | Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey | Institute of Informatics of the Slovak Academy of Sciences, Instituto de Física de Cantabria | Slovakia, Spain | — |

**Showing the 30 most-cited of 1,045 independent citing papers.**

Independent citing papers only; self- and co-author citations excluded. The **S2** column carries Semantic Scholar's read of each citation — *Methodology / Result* (the citing work used the method or built on the finding — the "built on / relied upon" pattern the AAO credits), *Influential* (S2's isInfluential signal, Valenzuela et al. 2015), or *Background* (a passing mention).

**FOLLOW-UP WORK**

### Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices

2019 · 1,472 citations (GS)

Field-normalised: 986 Semantic Scholar citations place it in the top 1% of Computer Science papers from 2019 indexed by Semantic Scholar, by citation count.

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 1 | Memory devices and applications for in-memory computing | IBM Research - Zurich | Switzerland | — |
| 2 | Model compression and hardware acceleration for neural networks: A comprehensive survey | Massachusetts Institute of Technology, Tsinghua University, University of California, Irvine Medical Center | China, United States | — |
| 3 | Binary neural networks: A survey | Beihang University, ETH Zurich, University of Electronic Science and Technology of China | China, Italy, Switzerland | — |
| 4 | Efficient acceleration of deep learning inference on resource-constrained edge devices: A review | Analog Devices (United States), Texas Tech University, University of Missouri | United States | **Influential** |
| 5 | Review of Lightweight Deep Convolutional Neural Networks: F. Chen et al. | Lanzhou University | China | — |

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 6 | A comprehensive survey on model quantization for deep neural networks in image classification | Universitätsklinikum Freiburg, University of Zanjan | Germany, Iran | — |
| 7 | Brain-inspired computing: A systematic survey and future trends | Graz University of Technology, Peking University, Purdue University | Austria, China, United States | — |
| 8 | Hardware approximate techniques for deep neural network accelerators: A survey | Karlsruhe Institute of Technology, National Technical University of Athens | Germany, Greece | — |
| 9 | Neuromorphic computing chip with spatiotemporal elasticity for multi-intelligent-tasking robots | Southeast University, Tsinghua University | China | — |
| 10 | Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks | ETH Zurich, ETH Zürich, Institute of Science and Technology Austria | Austria, Switzerland | — |
| 11 | Rigging the lottery: Making all tickets winners | DeepMind, Google Brain | United Kingdom, United States | — |
| 12 | A Review on the emerging technology of TinyML | University of Thessaly | Greece | — |
| 13 | A survey on efficient convolutional neural networks and hardware acceleration | Soongsil University | South Korea | — |
| 14 | Efficient deep learning infrastructures for embedded computing systems: a comprehensive survey and future envision | Nanyang Technological University | Singapore | — |
| 15 | Loss-aware automatic selection of structured pruning criteria for deep neural network acceleration | Soongsil University | South Korea | — |
| 16 | Pre-defined sparsity for low-complexity convolutional neural networks | Intel Labs, University of Southern California | China, United States | — |
| 17 | Demystifying map space exploration for npus | Georgia Institute of Technology, NVIDIA | United States | — |
| 18 | Subgraph stationary hardware-software inference co-design | Georgia Institute of Technology | United States | — |
| 19 | Eyelet: A Cross-Mesh NoC-Based Fine-Grained Sparse CNN Accelerator for Spatio-Temporal Parallel Computing Optimization | China Aerospace Science and Technology Corporation, Harbin Institute of Technology | China | — |
| 20 | A charge domain SRAM computing-in-memory macro with quantized interval-optimized ADC and input bit-level sparsity-optimized P2O-DAC for 8-b MAC operation | Chinese Academy of Sciences | China | — |
| 21 | Flexible parallel sliding window and data flow methods for deep neural networks on processing-in-memory architectures | Islamic Azad University, Shahid Bahonar University of Kerman | Iran | — |
| 22 | A survey on deep learning hardware accelerators for heterogeneous hpc platforms | Alma Mater Studiorum Università di Bologna, INFN, Politecnico di Milano | Italy | — |

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 23 | Counting carbon: A survey of factors influencing the emissions of machine learning | Hugging Face, University of Osnabrück | Canada, Germany | — |
| 24 | Full stack optimization of transformer inference: a survey | University of California, Irvine Medical Center | United States | — |
| 25 | Edge intelligence: Empowering intelligence to the edge of network | Hong Kong University of Science and Technology, Huazhong University of Science and Technology, Hunan University | China, Finland, United Kingdom | — |
| 26 | Enable deep learning on mobile devices: Methods, systems, and applications | Massachusetts Institute of Technology, MIT, Moscow Institute of Thermal Technology | Russia, United States | — |
| 27 | Edge intelligence: A review of deep neural network inference in resource-limited environments | Dong-A University, Korea National University of Transportation | South Korea | — |
| 28 | Fpga-based deep learning inference accelerators: Where are we standing? | RWTH Aachen University, University of Lübeck | Germany | — |
| 29 | Edge intelligence: Architectures, challenges, and applications | Huazhong University of Science and Technology, Hunan University, Tsinghua University | China, Finland | **Influential** |
| 30 | Hardware-assisted machine learning in resource-constrained IoT environments for security: review and future prospective | Mediterranean University | Montenegro | — |

**Showing the 30 most-cited of 1,029 independent citing papers.**

Independent citing papers only; self- and co-author citations excluded. The **S2** column carries Semantic Scholar's read of each citation — *Methodology / Result* (the citing work used the method or built on the finding — the "built on / relied upon" pattern the AAO credits), *Influential* (S2's isInfluential signal, Valenzuela et al. 2015), or *Background* (a passing mention).

**FOLLOW-UP WORK**

### Netadapt: Platform-aware neural network adaptation for mobile applications

2018 · 876 citations (GS)

Field-normalised: 557 Semantic Scholar citations place it in the top 1% of Computer Science papers from 2018 indexed by Semantic Scholar, by citation count.

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 1 | MobileNetV4: Universal models for the mobile ecosystem | Google, National University of Singapore | Singapore, United States | — |
| 2 | A survey of convolutional neural networks: analysis, applications, and prospects | Hohai University | China | — |
| 3 | Efficientnet: Rethinking model scaling for convolutional neural networks | Google Research | United States | — |
| 4 | Depgraph: Towards any structural pruning | Huawei Technologies Ltd., National University of Singapore, Zhejiang University | China, Singapore | — |

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 5 | Lightweight deep learning for resource-constrained environments: A survey | Foxconn (Cayman Islands), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Multimedia University | Cayman Islands, China, Malaysia | Influential |
| 6 | A survey of quantization methods for efficient neural network inference | UC Berkeley, University of California, Irvine Medical Center | United States | — |
| 7 | Facial Age Estimation Using Multi-Stage Deep Neural Networks | IKERBASQUE, Université Polytechnique Hauts-de-France, University of Biskra | Algeria, Finland, France | — |
| 8 | Resource-Aware Neural Network Pruning Using Graph-based Reinforcement Learning | — | — | — |
| 9 | Resource-Efficient Convolutional Networks: A Survey on Model-, Arithmetic-, and Implementation-Level Techniques | Chung-Ang University, Queen's University Belfast, University of Cambridge | South Korea, United Kingdom | — |
| 10 | Towards Hardware-Specific Automatic Compression of Neural Networks | Heidelberg University | Germany | Influential |
| 11 | Automated Runtime-Aware Scheduling for Multi-Tenant DNN Inference on GPU | George Mason University, Microsoft, University of Maryland, Baltimore County | United States | — |
| 12 | Bringing AI To Edge: From Deep Learning's Perspective | HP Inc., Nanyang Technological University | Singapore, United States | — |
| 13 | HAQ: Hardware-Aware Automated Quantization | Massachusetts Institute of Technology, Moscow Institute of Thermal Technology | Russia, United States | — |
| 14 | Topology-Aware Network Pruning using Multi-stage Graph Embedding and Reinforcement Learning | Iowa State University, Technische Universität Darmstadt | Germany, United States | — |
| 15 | Generative Model for Models: Rapid DNN Customization for Diverse Tasks and Resource Constraints | Beijing Institute of Technology, Beijing University of Posts and Telecommunications, Tsinghua University | China | Influential |
| 16 | 2D and 3D Palmprint and Palm Vein Recognition Based on Neural Architecture Search | Hefei University of Technology | China | Influential |
| 17 | Exploring the Effectiveness of Lightweight Architectures for Face Anti-Spoofing | Advanced Technologies Application Center, Tecnológico de Monterrey | Cuba, Mexico | — |
| 18 | AN ENHANCEMENT FOR THE CONSISTENT DEPTH ESTIMATION OF MONOCULAR VIDEOS USING LIGHTWEIGHT NETWORK | Novosibirsk State University, Suez University | Egypt, Russia | — |
| 19 | GAT TransPruning: progressive channel pruning strategy combining graph attention network and transformer | Feng Chia University | Taiwan | — |

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 20 | Pruning neural networks: is it time to nip it in the bud? | University of Edinburgh | United Kingdom | — |
| 21 | FP-NAS: Fast Probabilistic Neural Architecture Search | Facebook AI | United States | — |
| 22 | ANNETTE: Accurate Neural Network Execution Time Estimation With Stacked Models | TU Wien | Austria | — |
| 23 | Fire detection in video surveillance using superpixel-based region proposal and ESE-ShuffleNet | East China University of Science and Technology | China | — |
| 24 | ePerceptive—Energy Reactive Embedded Intelligence for Baeryless Sensors | — | — | — |
| 25 | Atrous Space Bender U-Net (ASBU-Net/LogiNet) | Logitech (Switzerland) | Switzerland | — |
| 26 | NetCut: Real-Time DNN Inference Using Layer Removal | Northeastern University | United States | — |
| 27 | A Gradient-based Architecture HyperParameter Optimization Approach | — | — | **Influential** |
| 28 | Dynamic DNNs and Runtime Management for Efficient Inference on Mobile/Embedded Devices | University of Southampton | United Kingdom | — |
| 29 | Leaf Disease Segmentation and Detection in Apple Orchards for Precise Smart Spraying in Sustainable Agriculture | — | — | — |
| 30 | Efficient Deep Learning Approach for the Classification of Pneumonia in Infants from Chest X-Ray Images | — | — | — |

**Showing the 30 most-cited of 103 independent citing papers.**

Independent citing papers only; self- and co-author citations excluded. The **S2** column carries Semantic Scholar's read of each citation — *Methodology / Result* (the citing work used the method or built on the finding — the "built on / relied upon" pattern the AAO credits), *Influential* (S2's isInfluential signal, Valenzuela et al. 2015), or *Background* (a passing mention).

## Contribution 2

### Claim — Contribution 2

*The researcher established critical design and evaluation frameworks for efficient deep neural networks on processing-in-memory accelerators, challenging standard metrics and optimizing hardware efficiency.*

The researcher's core contribution centers on the 2019 paper 'Design considerations for efficient deep neural networks on processing-in-memory accelerators,' which lays the groundwork for optimizing neural network execution on specialized hardware. This work is supported by subsequent publications that refine both the architectural design and the methodological standards for assessing such systems.

This line of work appears to address a critical gap in how deep neural network processors are designed and measured. The 2020 follow-up, 'How to evaluate deep neural network processors: Tops/w (alone) considered harmful,' suggests a challenge to conventional efficiency metrics, while 'Efficient processing of deep neural networks' indicates a broader effort to improve computational efficiency. Together, these titles imply a shift from simplistic power-performance ratios to more nuanced evaluation criteria and architectural optimizations.

The significance of this research is evidenced by substantial citation activity. The core paper has received 73 citations, while the follow-up works have garnered 118 and 409 citations respectively. Notably, 97.6% of the 3,742 classified citations for this scholar originate from independent researchers, indicating that this framework has been widely adopted and validated by the broader scientific community beyond the researcher's immediate circle.

**INDEPENDENT CITATIONS FOR THIS CONTRIBUTION: 185** · **11 flagged influential by Semantic Scholar**

**CORE PAPER**

## Design considerations for efficient deep neural networks on processing-in-memory accelerators

2019 · 73 citations (GS)

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 1 | Analysis and mitigation of parasitic resistance effects for analog in-memory neural network acceleration | Sandia National Laboratories, The University of Texas at Austin | United States | — |
| 2 | Designing a Deep Neural Network engine for LLC block reuse prediction to mitigate Soft Error in Multicore | Aliah University, Indian Institute of Engineering Science and Technology, Shibpur, Indian Institute of Technology Delhi | India | — |
| 3 | 4K-memristor analog-grade passive crossbar circuit | University of California, Santa Barbara | United States | — |
| 4 | Cluster Workload Allocation: A Predictive Approach Leveraging Machine Learning Efficiency | University of Westminster | United Kingdom | — |
| 5 | On the Accuracy of Analog Neural Network Inference Accelerators [Feature] | Infineon Technologies, Sandia National Laboratories | Germany, United States | — |
| 6 | Improving the Robustness of Neural Networks to Noisy Multi-Level Non-Volatile Memory-based Synapses | CEA Grenoble | France | **Influential** |
| 7 | Evaluating complexity and resilience trade-offs in emerging memory inference machines | Sandia National Laboratories | United States | — |
| 8 | An Energy-Efficient Ring-Based CIM Accelerator using High-Linearity eNVM for Deep Neural Networks | National United University, National Yang Ming Chiao Tung University | Taiwan | — |
| 9 | Fast-OverlaPIM: A Fast Overlap-Driven Mapping Framework for Processing In-Memory Neural Network Acceleration | IEEE, University of California San Diego | United States | — |
| 10 | In-Memory Computing with Resistive Memory Circuits: Status and Outlook | Politecnico di Milano | Italy | — |
| 11 | SPCIM: Sparsity-Balanced Practical CIM Accelerator With Optimized Spatial-Temporal Multi-Macro Utilization | Tsinghua University, University of California, Irvine Medical Center | China, United States | — |
| 12 | Conductance variations and their impact on the precision of in-memory computing with resistive switching memory (RRAM) | Politecnico di Milano | Italy | — |
| 13 | The Impact of Analog-to-Digital Converter Architecture and Variability on Analog Neural Network Accuracy | Arizona State University, Sandia National Laboratories | United States | — |

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 14 | Energy-efficient Mott activation neuron for full-hardware implementation of neural networks | University of California San Diego | United States | — |
| 15 | SDP: Co-Designing Algorithm, Dataflow, and Architecture for In-SRAM Sparse NN Acceleration | Tsinghua University, University of California, Irvine Medical Center | China, United States | — |
| 16 | AR-PIM: An Adaptive-Range Processing-in-Memory Architecture | American Rock Mechanics Association, University of Michigan | United States | — |
| 17 | Are SNNs Really More Energy-Efficient Than ANNs? an In-Depth Hardware-Aware Study | — | — | — |
| 18 | A Construction Kit for Efficient Low Power Neural Network Accelerator Designs | CSEM SA, ETH Zurich | Switzerland | — |
| 19 | Analog architectures for neural network acceleration based on non-volatile memory | Sandia National Laboratories | United States | — |
| 20 | Forward Target Propagation: A Forward-Only Approach to Global Error Credit Assignment via Local Losses | Sandia National Laboratories, Texas A&M University | United States | — |
| 21 | Real-Time Compressed Sensing for Joint Hyperspectral Image Transmission and Restoration for CubeSat | National Cheng Kung University | Taiwan | — |
| 22 | Device Quantization Policy and Power-Performance- Area Co-Optimization Strategy in Variation-Aware In-memory Computing Design | National Yang Ming Chiao Tung University | Taiwan | — |
| 23 | Device-aware inference operations in SONOS nonvolatile memory arrays | Cypress Semiconductor Corporation (Belgium), Sandia National Laboratories | Belgium, United States | — |
| 24 | Dependability of Alternative Computing Paradigms for Machine Learning: hype or hope? | — | — | — |
| 25 | Modeling-Based Design of Memristive Devices for Brain-Inspired Computing | Peking University | China | — |
| 26 | Advanced Integration-Inspired Process-in-Memory: A Comprehensive Review of Design, Challenges, and Future Prospects | — | — | — |
| 27 | Device quantization policy in variation-aware in-memory computing design | — | — | — |
| 28 | Ferroelectric FET-Based Time-Mode Multiply-Accumulate Accelerator: Design and Analysis | — | — | — |
| 29 | A compute-in-memory chip based on resistive random-access memory | — | — | — |
| 30 | Oxide-based filamentary RRAM for deep learning | — | — | — |

**Showing the 30 most-cited of 44 independent citing papers.**

Independent citing papers only; self- and co-author citations excluded. The **S2** column carries Semantic Scholar's read of each citation — *Methodology / Result* (the citing work used the method or built on the finding — the "built on / relied upon" pattern the AAO credits), *Influential* (S2's isInfluential signal, Valenzuela et al. 2015), or *Background* (a passing mention).

**FOLLOW-UP WORK**

### How to evaluate deep neural network processors: Tops/w (alone) considered harmful

2020 · 118 citations (GS)

Field-normalised: 67 Semantic Scholar citations place it in the top 10% of Computer Science papers from 2020 indexed by Semantic Scholar, by citation count.

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 1 | Designing Object Detection Models for TinyML: Foundations, Comparative Analysis, Challenges, and Emerging Solutions | INSA Rennes, Khalifa University of Science and Technology | France, United Arab Emirates | — |
| 2 | Towards Hardware-Specific Automatic Compression of Neural Networks | Heidelberg University | Germany | **Influential** |
| 3 | BOLD: Boolean Logic Deep Learning | Huawei Technologies (France) | France | — |
| 4 | A framework for measuring the training efficiency of a neural architecture | Trinity College Dublin | Ireland | — |
| 5 | Union: A Unified HW-SW Co-Design Ecosystem in MLIR for Evaluating Tensor Operations on Spatial Accelerators | Georgia Institute of Technology, IBM Research, NVIDIA | Japan, United States | — |
| 6 | A Survey of Design and Optimization for Systolic Array-based DNN Accelerators | National University of Defense Technology | China | — |
| 7 | Boolean Variation and Boolean Logic Back-Propagation | Huawei Technologies | United Kingdom | — |
| 8 | Lightweight Deep Learning for Resource-Constrained Environments: A Survey | Foxconn (Cayman Islands), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Multimedia University | Cayman Islands, China, Malaysia | — |
| 9 | Neural Architecture Search and Hardware Accelerator Co-Search: A Survey | Brno University of Technology | Czech Republic | — |
| 10 | Episodic memories make goal directed action selection context-aware and explainable | — | — | — |
| 11 | Benchmarking Ultra-Low-Power µNPUs | Imperial College London, University of Cambridge | United Kingdom | — |
| 12 | Multi-Bit Compute-In Memory Architecture Using a C-2C Ladder Network | — | — | — |
| 13 | Single-chip photonic deep neural network with forward-only training | Luminous Computing Inc., MIT, Nokia Corporation | United States | — |
| 14 | Invited: The Magnificent Seven Challenges and Opportunities in Domain-Specific Accelerator Design for Autonomous Systems | Boston University, Dartmouth College, Harvard University | United States | — |
| 15 | Algorithm-hardware Co-optimization for Energy-efficient Drone Detection on Resource-constrained FPGA | GRASP Lab, University of Pennsylvania | — | — |

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 16 | Task parallelism-based architectures on FPGA to optimize the energy efficiency of AI at the edge | — | — | — |
| 17 | On the Basis of Brain: Neural-Network-Inspired Change in General Purpose Chips | — | — | — |
| 18 | Low Rank Optimization for Efficient Deep Learning: Making A Balance between Compact Architecture and Fast Training | University of Electronic Science and Technology of China | China | — |
| 19 | HCM: Hardware-Aware Complexity Metric for Neural Network Architectures | Ben-Gurion University of the Negev, Ruppin Academic Center, Technion – Israel Institute of Technology | Israel | — |
| 20 | Parameter Estimation using Random 1 bit Streams (PEERS) | — | — | — |
| 21 | Automated HW/SW Co-design for Edge AI: State, Challenges and Steps Ahead: Special Session Paper | — | — | — |
| 22 | Compute-Efficient Modelling of Multi-NPU Inference on Edge MPSoCs for Energy-Aware Online Workload Allocation | — | — | — |
| 23 | An 8-bit Single Perceptron Processing Unit for Tiny Machine Learning Applications | — | — | — |
| 24 | Distilling knowledge for low-energy AIoT | — | — | — |
| 25 | Power and Area Efficient Processing Element (PE) of CNN Accelerator for Object Detection | Berhampur University, Maharaja Engineering College | India | — |
| 26 | Reconfigurable Network-on-Chip based Convolutional Neural Network Accelerator | Institute for Research in Fundamental Sciences, Islamic Azad University, Science and Research Branch, Trinity College Dublin | Iran, Ireland | — |
| 27 | Iterative neural networks for adaptive inference on resource-constrained devices | Ghent University | Belgium | — |
| 28 | Chiplet-Gym: Optimizing Chiplet-Based AI Accelerator Design With Reinforcement Learning | Auburn University | United States | — |
| 29 | From Quantum Materials to Microsystems | AREA Science Park, Istituto di Fotonica e Nanotecnologie | Italy | — |
| 30 | Accurate and energy efficient ad-hoc neural network for wafer map classification | Institut polytechnique de Grenoble, STMicroelectronics (France) | France | — |

**Showing the 30 most-cited of 60 independent citing papers.**

Independent citing papers only; self- and co-author citations excluded. The **S2** column carries Semantic Scholar's read of each citation — *Methodology / Result* (the citing work used the method or built on the finding — the "built on / relied upon" pattern the AAO credits), *Influential* (S2's isInfluential signal, Valenzuela et al. 2015), or *Background* (a passing mention).

**FOLLOW-UP WORK**

## Efficient processing of deep neural networks

2020 · 409 citations (GS)

Field-normalised: 263 Semantic Scholar citations place it in the top 5% of Computer Science papers from 2020 indexed by Semantic Scholar, by citation count.

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 1 | Signed Binarization: Unlocking Efficiency Through Repetition-Sparsity Trade-Off | Microsoft; Georgia Institute of Technology | — | Influential |
| 2 | HYTE: Flexible Tiling for Sparse Accelerators via Hybrid Static-Dynamic Approaches | — | — | — |
| 3 | CODEBench: A Neural Architecture and Hardware Accelerator Co-Design Framework | Princeton University, Stanford University | United States | Influential |
| 4 | A2Q: Aggregation-Aware Quantization for Graph Neural Networks | — | — | — |
| 5 | A Bit-Serial, Compute-in-SRAM Design Featuring Hybrid-Integrating ADCs and Input Dependent Binary Scaled Precharge Eliminating DACs for Energy-Efficient DNN Inference | Apple (United States), The University of Texas at Austin | United States | — |
| 6 | MTIA: First Generation Silicon Targeting Meta's Recommendation Systems | Meta (United States), Vanderbilt University | United States | Influential |
| 7 | A Survey of Design and Optimization for Systolic Array-based DNN Accelerators | National University of Defense Technology | China | — |
| 8 | Improving the Robustness of Neural Networks to Noisy Multi-Level Non-Volatile Memory-based Synapses | CEA Grenoble | France | — |
| 9 | A Construction Kit for Efficient Low Power Neural Network Accelerator Designs | CSEM SA, ETH Zurich | Switzerland | — |
| 10 | Fast and Scalable Multicore YOLOv3-Tiny Accelerator Using Input Stationary Systolic Architecture | Bandung Institute of Technology, National Taiwan University of Science and Technology | Indonesia, Taiwan | — |
| 11 | Kraken: An Efficient Engine with a Uniform Dataflow for Deep Neural Networks | University of Moratuwa, University of Moratuwa, Florida International University | Sri Lanka, Sri Lanka, USA | — |
| 12 | FLAT: An Optimized Dataflow for Mitigating Attention Bottlenecks | Georgia Institute of Technology, Google, Google Research | United States | — |
| 13 | Binarized Neural-Network Parallel-Processing Accelerator Macro Designed for an Energy Efficiency Higher Than 100 TOPS/W | Tokyo Institute of Technology | Japan | — |
| 14 | Energy-efficient user allocation and cache updating in mobile edge computing networks based on user geographical aggregation | Chongqing University, Shenzhen Metro (China), Tianjin University | China | — |
| 15 | PowerGS: Display-Rendering Power Co-Optimization for Foveated Radiance-Field Rendering in Power-Constrained XR Systems | — | — | — |
| 16 | Bit-serial systolic accelerator design for convolution operations in convolutional neural networks | Qinghai Normal University, University of Electronic Science and Technology of China | China | — |

| No. | Citing paper | Citing institution(s) | Country | S2 |
|---|---|---|---|---|
| 17 | Resource-efficient VLSI Architecture of Softmax Activation Function for Real-time Inference in Deep Learning Applications | Birla Institute of Technology and Science - Hyderabad Campus, IIEST Shibpur, Indian Institute of Engineering Science and Technology, Shibpur | India | — |
| 18 | Research on Wastewater Treatment Monitoring Algorithms Based on Deep Convolutional Neural Networks | Guangdong Province Environmental Monitoring Center | China | — |
| 19 | AI Accelerator Survey and Trends | MIT Lincoln Laboratory | United States | — |
| 20 | AccelTran: A Sparsity-Aware Accelerator for Dynamic Inference With Transformers | Princeton University | United States | — |
| 21 | Energy Complexity of Fully-Connected Layers | Czech Academy of Sciences, Institute of Computer Science, Université de Versailles Saint-Quentin-en-Yvelines | Czech Republic, France | — |
| 22 | Deep Learning-Based Screening Test for Cognitive Impairment Using Basic Blood Test Data for Health Examination | Nihon University, Robotics Research (United States), The University of Tokyo | Japan, United States | — |
| 23 | VISTA: Optimizing GPU Scheduling through Versatile Locality-Aware Data Sharing | Universitat Politècnica de Catalunya | Spain | — |
| 24 | A Survey and Comparative Analysis of Number Systems for Deep Neural Networks | Khalifa University of Science and Technology | United Arab Emirates | — |
| 25 | Exploring the Potential of Wireless-enabled Multi-Chip AI Accelerators | Universitat Politècnica de Catalunya | Spain | — |
| 26 | Benchmarking of hardware-efficient real-time neural decoding in brain–computer interfaces | — | — | — |
| 27 | A Low-Cost Accelerator for License Plate Character Recognition Using Convolutional Neural Networks | Universidade do Vale do Itajaí | Brazil | — |
| 28 | Commercial Evaluation of Zero-Skipping MAC Design for Bit Sparsity Exploitation in DL Inference | Carnegie Mellon University, MediaTek USA Inc. | United States | — |
| 29 | ISimDL: Importance Sampling-Driven Acceleration of Fault Injection Simulations for Evaluating the Robustness of Deep Learning | New York University Abu Dhabi, TU Wien | Austria, United Arab Emirates | — |
| 30 | Learning Silhouettes with Group Sparse Autoencoders | Harvard University | United States | — |

**Showing the 30 most-cited of 81 independent citing papers.**

Independent citing papers only; self- and co-author citations excluded. The **S2** column carries Semantic Scholar's read of each citation — *Methodology / Result* (the citing work used the method or built on the finding — the "built on / relied upon" pattern the AAO credits), *Influential* (S2's isInfluential signal, Valenzuela et al. 2015), or *Background* (a passing mention).

## D. Citing-Institution Prestige & Geography

### Top citing institutions

| Institution | Country | World ranking | Citing papers |
|---|---|---|---|
| Tsinghua University | China | SCImago #8 · THE 12 · QS =17 | 109 |
| University of California, Irvine Medical Center | United States | — | 86 |
| Massachusetts Institute of Technology | United States | SCImago #41 · THE 2 · QS 1 | 84 |
| Chinese Academy of Sciences | China | SCImago #2 | 74 |
| Georgia Institute of Technology | United States | SCImago #270 · THE =41 · QS =123 | 69 |
| Shanghai Jiao Tong University | China | SCImago #10 · THE 40 · QS =47 | 62 |
| Peking University | China | SCImago #11 · THE 13 · QS 14 | 50 |
| NVIDIA | United States | — | 47 |
| Northeastern University | United States | QS 384 | 46 |
| ETH Zurich | Switzerland | THE 11 · QS 7 | 44 |
| Hong Kong University of Science and Technology | Hong Kong | SCImago #483 · THE =58 · QS 44 | 43 |
| Nanyang Technological University | Singapore | SCImago #137 | 42 |
| Arizona State University | United States | SCImago #357 · THE 201–250 · QS =173 | 42 |
| Stanford University | United States | SCImago #18 · THE =5 · QS 3 | 41 |
| Sun Yat-sen University | China | SCImago #40 · THE 201–250 · QS =276 | 38 |

### Geographic distribution of citing authors

| Country | Citing papers |
|---|---|
| United States | 1,147 |
| China | 1,001 |
| United Kingdom | 254 |
| India | 200 |
| South Korea | 197 |
| Canada | 139 |
| Germany | 130 |
| Italy | 124 |
| France | 107 |
| Switzerland | 99 |
| Taiwan | 98 |
| Singapore | 97 |

Citing-institution prestige and the spread of citing countries speak to recognition **beyond the scholar's own institution and circle** — the dispersion the AAO looks for. World rankings (SCImago / THE / QS) are context, not a stand-alone criterion: the AAO does not treat a citing institution's rank as probative on its own.

## F. AAO Precedent Considerations

### Pre-filing self-check (AAO denial patterns)

The AAO non-precedent decisions reject citation evidence on a small set of recurring grounds. Confirm the petition addresses each before filing:

- ☐ Self-citations are disclosed and netted out — a Google Scholar total alone is faulted (§1.1).
- ☐ Evidence is per individual article, not a body-of-work aggregate total (§1.2).
- ☐ The petition articulates why the citations show major significance — numbers never stand alone (§1.5).
- ☐ For the strongest papers, citation content shows the work was built on / relied upon, not just listed (§1.6, §2.2).
- ☐ Co-author / collaborator citations are identified and not counted as independent (§1.7).
- ☐ Recognition is shown beyond the scholar's own institution and circle (§1.8).
- ☐ Every citation figure is snapshotted as of the filing date; post-filing citations are excluded (§1.9).
- ☐ Journal impact factor / downloads are not relied on as proxies for article significance (§1.10, §1.12).
- ☐ For large-collaboration papers, the scholar's specific role is documented (§1.13).
- ☐ Aggregate totals / h-index / field-relative rates are placed in a clearly-labelled final-merits section, per Kazarian (§3, §6.1.7).

**Disclaimer**

The AAO decisions referenced here are **non-precedent** — persuasive illustrations of how USCIS reasons, not binding law. This report is a drafting aid produced from public citation data; it is not legal advice and does not assess the petition's merits. All analysis must be reviewed by qualified immigration counsel.

## G. Citation Evidence Index

Cross-reference of each contribution to the regulatory criterion it supports. Counsel should map these to the petition's exhibit numbers.

| Contribution | Core paper | Indep. cites | Supports |
|---|---|---|---|
| Contribution 1 | Efficient processing of deep neural networks: A tutorial and survey | 2,177 | 8 CFR 204.5(h)(3)(v) — Criterion 5 |
| Contribution 2 | Design considerations for efficient deep neural networks on processing-in-memory accelerators | 185 | 8 CFR 204.5(h)(3)(v) — Criterion 5 |