A Survey of Chip Design and AI Integration: CNNs, Transformers, LLMs, and Semiconductor Innovations

www.surveyx.cn

Abstract

The integration of artificial intelligence (AI) with chip design represents a pivotal advancement in semiconductor technology, enhancing computational efficiency, processing power, and intelligent functionalities. This survey explores the interdisciplinary convergence of AI technologies, such as Convolutional Neural Networks (CNNs), Transformers, and Large Language Models (LLMs), with hardware engineering, underscoring its transformative impact on chip design. Key innovations include 3D integration in neuromorphic computing, CNN applications in neuroanatomical mapping, and AI-driven security enhancements in quantum circuits. AI technologies have revolutionized electronic design automation (EDA), optimizing power, performance, and area (PPA) metrics, and addressing the computational demands of deep learning models. The survey highlights AI's role in semiconductor manufacturing, particularly through AI-driven optimization techniques that enhance energy efficiency and scalability. Challenges such as data scarcity, architectural constraints, and security vulnerabilities are discussed, alongside future research directions that focus on optimizing AI integration in chip design. The survey concludes by emphasizing the potential of AI to drive significant advancements in design automation, knowledge management, and process optimization, paving the way for the next wave of technological innovation in semiconductor technology.

1 Introduction

1.1 Interdisciplinary Nature of Chip Design and AI

The integration of chip design and artificial intelligence (AI) showcases innovative approaches that enhance hardware engineering through advanced AI technologies. For instance, 3D integration in neuromorphic computing exemplifies the convergence of AI and chip design, leading to the creation of more efficient processing units [1]. Additionally, Convolutional Neural Networks (CNNs) automate histological analysis, highlighting AI's role in biological sciences and its potential to enhance chip functionalities [2].

In automotive systems, the application of camera-based deep learning algorithms for perception tasks in Automated Driving systems illustrates the interdisciplinary challenges and opportunities presented by AI in chip design, particularly regarding the computational constraints of the automotive sector [3]. Furthermore, advanced AI techniques for parsing technical drawings demonstrate the fusion of engineering and AI, essential for precise communication in design specifications [4].

These interdisciplinary efforts are reshaping the chip design landscape by incorporating machine learning techniques, such as deep reinforcement learning, and innovative knowledge management approaches that enhance information retrieval during the design process. This collaborative approach addresses the stagnation in traditional chip design methods and fosters the development of new hardware systems that leverage heterogeneous technologies, ultimately driving advancements in integrated circuits and educational opportunities in the field [5, 6, 7, 8]. The convergence of AI

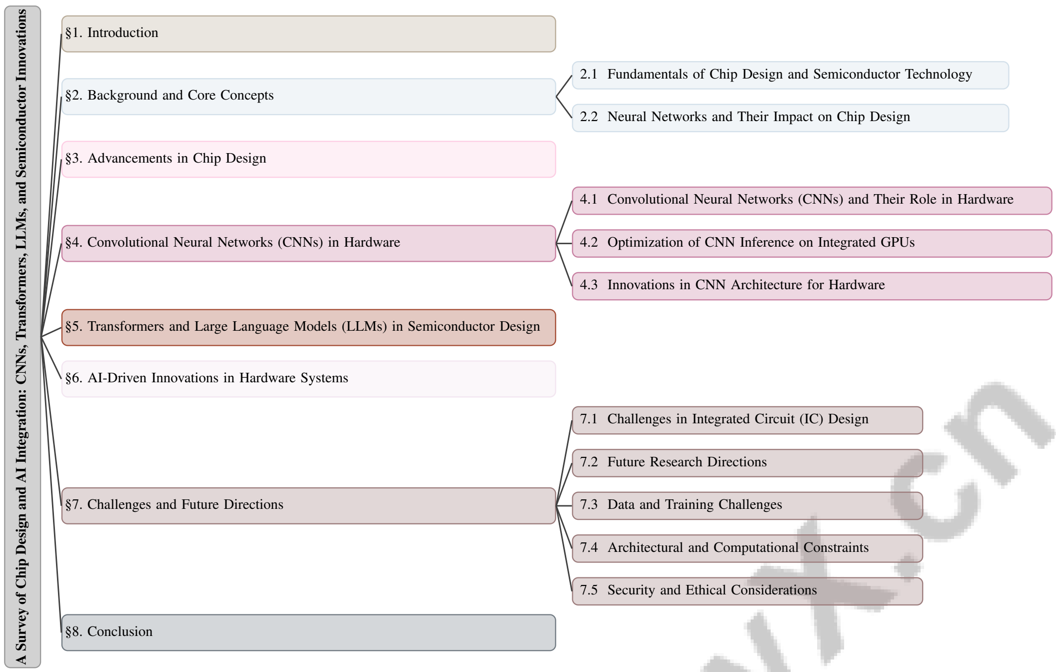

Figure 1: chapter structure

technologies with hardware engineering heralds a new era of innovation characterized by improved computational efficiency and intelligent functionalities crucial for next-generation semiconductor technologies.

1.2 Significance of AI in Hardware Systems

The integration of AI into hardware systems is pivotal for enhancing computational capabilities and efficiency across various domains. AI technologies, especially Large Language Models (LLMs), have revolutionized electronic design automation (EDA) by streamlining circuit generation and significantly improving power, performance, and area (PPA) metrics. LLMs facilitate architecture specification and hardware description language (HDL) development, making traditionally time-consuming tasks more efficient and accurate, thereby addressing the complexities of modern integrated circuits and paving the way for fully automated design processes [9, 10]. This transformation is essential for maintaining technological progress in light of diminishing returns from Moore's Law.

AI is crucial in meeting the substantial computational demands of deep learning models, particularly convolutional neural networks (CNNs), which often surpass the capabilities of conventional computing architectures. As deep learning gains traction in fields like computer vision and online education, the need for efficient processing power becomes increasingly apparent. This necessity has spurred the development of specialized hardware and AI-augmented research and development approaches that enhance productivity and facilitate the deployment of complex models, driving technological innovation and economic growth [11, 5, 12, 13]. The high energy consumption associated with deep neural network (DNN) computations, particularly within Von Neumann architectures, underscores the demand for AI-driven architectural innovations that improve analysis time and enhance power-critical test identification.

Moreover, AI's integration into hardware systems extends beyond computational efficiency to scientific breakthroughs in areas like protein folding and drug discovery, illustrating AI's potential to advance hardware capabilities [13]. The rise of heterogeneous chiplet architectures, driven by the increasing computational demands of evolving AI algorithms, exemplifies AI's role in enhancing cost efficiency and reducing design complexity in data centers [14].

AI's importance is further highlighted by its ability to achieve energy efficiency improvements exceeding 500× compared to conventional CPU-based implementations, emphasizing its critical role in advancing hardware capabilities [15]. Additionally, AI optimizes logic synthesis recipes to

AI-generated, for reference only.

enhance the area-delay product of synthesized circuits, which is essential for efficient chip design [16].

The need for new methodologies to address computational complexity in Mixed-Integer Programming (MIP) models further illustrates AI's role in solving NP-hard problems efficiently [17]. In edge-AI computing, AI addresses challenges necessitating high energy efficiency, low power consumption, and flexibility in chip design [18].

Finally, the rapid advancement of digital technology has escalated the demand for growth in the integrated circuit (IC) industry, requiring AI-driven innovations to meet these challenges [19]. AI's integration into hardware systems is crucial for overcoming the limitations of current data-starved design cycles that obscure insights due to reliance on limited data from benchmarks and performance counters [20].

1.3 Structure of the Survey

This survey is structured to comprehensively explore the integration of AI technologies with chip design, focusing on the transformative impacts of Convolutional Neural Networks (CNNs), Transformers, Large Language Models (LLMs), and semiconductor innovations. It begins with an introduction to the interdisciplinary nature of chip design and AI, highlighting their convergence and the significance of AI in enhancing hardware systems. The survey examines foundational principles in chip design, semiconductor technology, and neural networks, emphasizing recent advancements in machine learning, particularly deep learning, and their implications for computational devices in the post-Moore's Law era. Ethical considerations in AI integration within these domains are also addressed, alongside innovative applications of AI in chip design processes and the role of knowledge management in enhancing design efficiency and sustainability. This overview establishes a critical context for understanding AI's transformative role in modern hardware development [7, 21, 22, 5, 23].

Subsequent sections analyze recent advancements in chip design, emphasizing energy-efficient and high-performance computing, as well as innovations in microchip creation and optimization. The application of CNNs in hardware is explored in depth, focusing on their contributions to computational tasks and architectural innovations tailored for hardware implementation. Furthermore, the survey examines the use of Transformers and LLMs in semiconductor design, discussing their contributions to intelligent functionalities and design automation.

AI-driven innovations in hardware systems are analyzed, detailing the impact of AI on semiconductor manufacturing and optimization techniques that leverage AI to enhance performance and efficiency. The survey concludes by addressing challenges and future directions in integrating AI with chip design, identifying potential research opportunities, and discussing data, training, architectural, computational, security, and ethical considerations. This structured approach facilitates a comprehensive understanding of the interdisciplinary nature of semiconductor technology, highlighting its transformative potential through enhanced data management, knowledge extraction, and innovative design methodologies that can significantly improve chip manufacturing and application processes [21, 22, 5, 24, 8].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Fundamentals of Chip Design and Semiconductor Technology

The integration of artificial intelligence (AI) into hardware is increasingly vital as traditional CMOS scaling approaches its limits, necessitating solutions that transcend Moore's Law and Dennard Scaling to enhance chip performance and efficiency [20]. Contemporary design complexities demand significant expertise, often hindering full automation [10]. Placement, a critical process where circuit modules are arranged on the chip, significantly impacts performance and manufacturability [25]. Current methods struggle to meet evolving design requirements, particularly in macro cell coverage and optimization efficiency [26], and joint learning of placement and routing highlights limitations in existing Electronic Design Automation (EDA) approaches [27].

In semiconductor manufacturing, lithography modeling ensures manufacturable chip design masks, directly affecting pattern fidelity on silicon wafers [28]. Optimizing network-on-chip designs is crucial for low latency and power efficiency in multi-core processors, emphasizing effective data movement [29]. Predicting routing congestion and design rule checking hotspots remains vital for VLSI circuit design, requiring advanced methodologies for enhanced accuracy and efficiency [30].

Traditional processors, such as CPUs and GPUs, often lack efficiency for AI computations, prompting exploration of alternative architectures like heterogeneous neuromorphic systems-on-chip (SoC) that integrate RISC-V CPUs with neuromorphic processors for improved computational efficiency and flexibility [18]. Logic synthesis, optimizing hardware description languages into efficient implementations using Boolean logic gates, is integral to high-performance chip design. AI integration addresses technical challenges and fosters innovation, with ethical considerations in semiconductor fabrication, IoT design, and AI integration necessitating a comprehensive framework for modern chip design complexities [21].

Innovative approaches, such as the In-Memory Classifier using a standard 6T SRAM array [31], highlight memory technologies' role in AI integration within chip design. Addressing data scarcity, particularly in Verilog data for fine-tuning large language models (LLMs) for hardware description language (HDL) generation, underscores the need for comprehensive datasets to advance AI-driven chip design [32]. The introduction of a ferroelectric field effect transistor (FeFET)-based compute-in-memory (CiM) annealer for solving computationally hard combinatorial optimization problems (COPs) further exemplifies innovative strategies to enhance chip design efficiency [33].

2.2 Neural Networks and Their Impact on Chip Design

Neural networks, notably Convolutional Neural Networks (CNNs), have significantly influenced chip design, imposing substantial computational requirements and integration challenges. Their complex architectures necessitate optimizing computational resources for efficient power consumption while maintaining high accuracy, crucial for hardware-aware neural architecture search (NAS) [34]. Traditional chip architectures struggle with the intensive computational loads from operations like multiplications and inference latency, driving the need for energy-efficient solutions.

The deployment of CNNs in hardware systems is complicated by the limited computational resources of embedded systems, which face challenges accommodating high-performing CNNs due to substantial memory and power demands [35]. This is particularly evident in distributed training scenarios on mobile and edge devices, where resource and memory constraints pose significant hurdles [36]. Existing pooling methods often fail to capture second-order statistics from feature maps, limiting their effectiveness in image classification tasks [37].

Neural networks automate complex tasks, such as segmenting cytoarchitectonic areas, enhancing biological data analysis efficiency [2]. However, integrating CNNs into chip design is challenged by the inefficiencies of current architectures in achieving high accuracy while minimizing computational costs and memory usage [38]. Neural networks are also applied in predicting properties of synthesized netlists during the place-and-route stage, a process that is both time-consuming and complex [39].

The influence of neural networks on chip design is evident in enhancing CNN-based inference at edge servers, ensuring high accuracy under severe channel conditions through frameworks like the Robust Edge Intelligence Framework (REIF) [40]. The inefficiency of existing CNN accelerators in managing both convolutional (CONV) and fully-connected (FC) layers underscores the need for improved performance and energy efficiency solutions [41].

Neural networks are crucial for addressing security vulnerabilities in manycore systems, enhancing threat detection like Trojan-inserted Quantum Approximate Optimization Algorithm (QAOA) circuits. The challenge of determining the most efficient deep learning architecture and structure for various data types and applications remains central to optimizing neural networks for chip design [42].

As neural networks evolve, their impact on chip design is expected to expand, driving advancements in hardware efficiency and computational capabilities. The integration of AI with hardware systems faces challenges in managing scaled combinatorial optimization problems due to memory access and system scalability limitations, critical for advancing AI-driven chip design [33].

Recent advancements in chip design have significantly transformed the landscape of technology, particularly in the realms of energy-efficient computing, artificial intelligence (AI), and enhanced security measures. To elucidate these developments, Figure 2 provides a comprehensive illustration of the hierarchical categorization of these innovations. This figure highlights key methodologies, including neuromorphic computing, FeFET-CiM, and Chain-NN, while emphasizing the pivotal role of AI in optimizing microchip creation and security frameworks. By integrating these advancements, the figure serves as a visual representation of the intricate relationships and ongoing trends in the field, thereby enhancing our understanding of the current state and future directions of chip design.

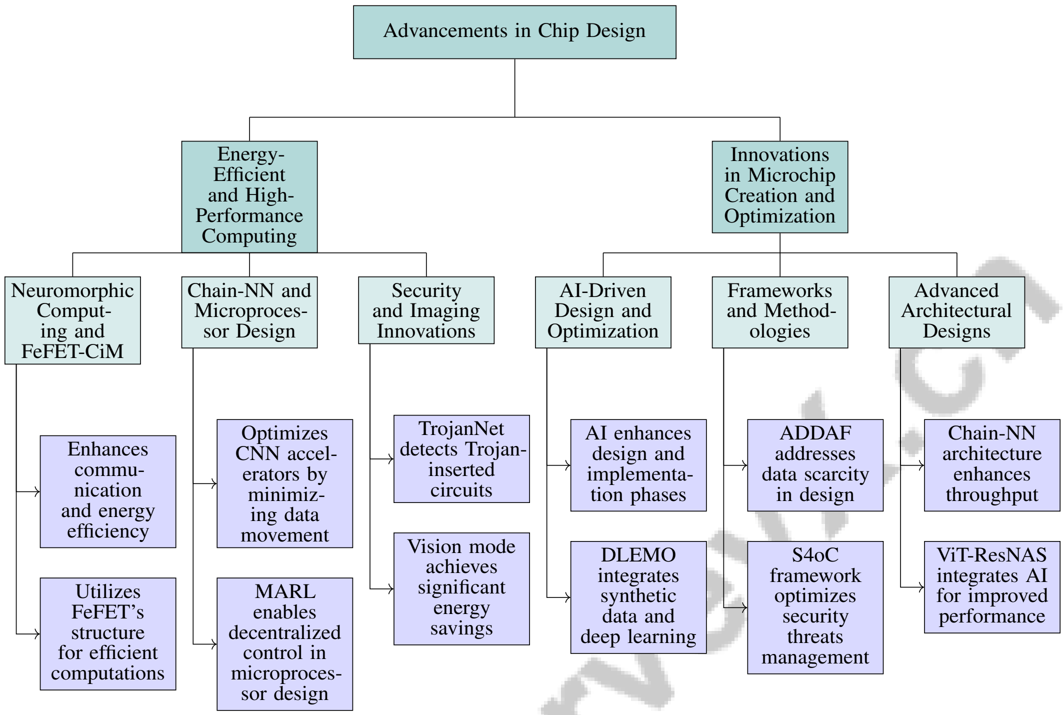

Figure 2: This figure illustrates the hierarchical categorization of recent advancements in chip design, focusing on energy-efficient computing, AI-driven innovations, and security measures. It highlights key methodologies such as neuromorphic computing, FeFET-CiM, and Chain-NN, and emphasizes the role of AI in optimizing microchip creation and security frameworks.

3 Advancements in Chip Design

3.1 Energy-Efficient and High-Performance Computing

Contemporary chip design advancements focus on balancing energy efficiency with computational performance to meet the demands of modern applications. Neuromorphic computing, through advanced interconnection topologies, enhances both communication and energy efficiency in chip architectures [1]. The FeFET-CiM methodology exemplifies this by utilizing the FeFET's three-terminal structure for in-situ vector-matrix-vector multiplication, significantly reducing energy consumption and latency [33]. Chain-NN architecture further optimizes CNN accelerators by minimizing data movement, thus enhancing resource optimization [43]. In microprocessor design, multi-agent reinforcement learning (MARL) surpasses traditional methods by enabling decentralized control over complex design spaces [44].

The Sum-EH scheme boosts energy efficiency in IoT nodes by harnessing energy from power beacons and primary transmitters, enhancing packet transmission, particularly in distributed CNN training on resource-constrained devices [45, 36]. Evaluating pruning and quantization methods, such as structured pruning and SIMD instructions, optimizes both computational efficiency and model performance [46]. The HENet architecture reduces computational costs while maintaining accuracy by optimizing feature map usage and minimizing redundant operations [38].

Security innovations like TrojanNet improve quantum computing performance by detecting Trojan-inserted circuits [47]. The S4oC framework enhances manycore systems' performance via adaptive security threat management using a multi-layer graph model and real-time optimization techniques [48]. The RMDL method combines multiple randomly generated models of DNN, CNN, and

RNN to boost classification accuracy and robustness, enhancing computational efficiency [42]. The HDLdebugger framework automates HDL debugging, achieving a superior pass rate compared to existing methods [49].

These advancements collectively highlight the ongoing pursuit of energy-efficient and high-performance computing in chip design. Figure 3 illustrates the hierarchical categorization of key advancements in this field, focusing on neuromorphic computing, microprocessor design, and security innovations. Each category highlights specific methodologies and frameworks that contribute to the enhancement of computational performance and energy efficiency, further emphasizing the importance of these developments in contemporary technology.

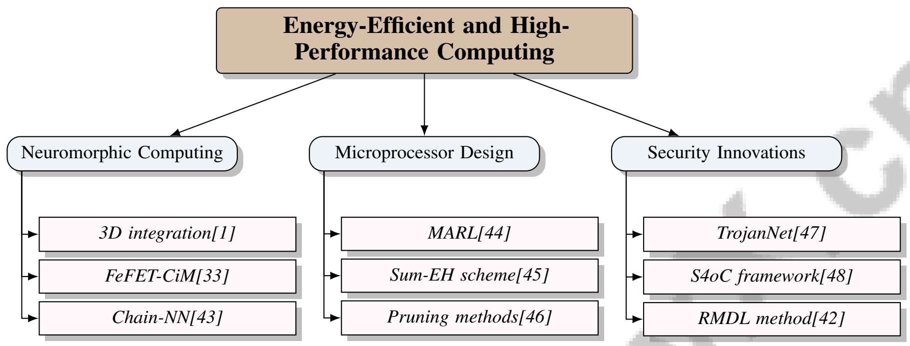

Figure 3: This figure illustrates the hierarchical categorization of key advancements in energy-efficient and high-performance computing, focusing on neuromorphic computing, microprocessor design, and security innovations. Each category highlights specific methodologies and frameworks that contribute to the enhancement of computational performance and energy efficiency.

3.2 Innovations in Microchip Creation and Optimization

AI significantly influences recent innovations in microchip creation and optimization, enhancing design and implementation phases. A novel image sensor design operates without a traditional ISP, utilizing adjustable resolution and tunable ADCs to optimize resource allocation and performance [50]. The DLEMO method demonstrates AI's transformative potential by integrating synthetic data generation, deep learning classifiers, and Bayesian optimization, improving the efficiency of traditional MIP solvers and streamlining microchip design workflows [8, 51, 17, 52].

The ADDAF framework automatically generates large-scale datasets for chip design, creating high-quality aligned natural language and Verilog/EDA script data without human intervention. This innovation addresses data scarcity in microchip design, enhancing the fine-tuning of LLMs for HDL generation and improving the efficiency and adaptability of EDA tools [9, 32, 52, 10]. A decentralized MARL approach in architectural design space exploration enhances efficiency by allowing multiple agents to independently optimize various subsystems, addressing the combinatorial explosion of design parameters and improving performance metrics like power efficiency and latency [16, 53, 54, 55, 44].

The Chain-NN architecture, utilizing a 1D systolic array of processing elements, minimizes energy consumption and enhances throughput through dual-channel processing engines and an optimized column-wise scan input pattern. Fabricated using TSMC's 28nm process, it achieves a peak throughput of 806.4 GOPS at 700 MHz with a power efficiency of 1421.0 GOPS/W, outperforming existing solutions and accelerating multiple convolutional layers in CNNs [41, 43]. The RMDL ensemble method allows simultaneous training of multiple deep learning architectures with randomly generated hyperparameters, improving robustness and accuracy in microchip design applications [5, 56, 8].

ViT-ResNAS exemplifies the integration of AI with advanced architectural designs to enhance computational efficiency and performance in microchip design, as seen in tools like AlphaChip, which generates superior chip layouts widely adopted in state-of-the-art chip production [7, 8]. Introducing biologically plausible non-linear convolution schemes within CNN architectures highlights the potential of bio-inspired methodologies to advance AI-driven microchip designs [21, 7, 8].

The S4oC framework employs a sophisticated four-layer graph model for microchip design and security, adaptively optimizing itself in real-time through reinforcement learning to counteract security threats such as hardware Trojans and side-channel attacks. By integrating heterogeneous reconfigurable processing elements and memory components, this framework enhances performance and resilience in system-on-chip designs [48, 8].

4 Convolutional Neural Networks (CNNs) in Hardware

The integration of Convolutional Neural Networks (CNNs) into hardware systems has gained prominence due to their transformative impact on applications like image processing and machine learning. Understanding CNNs' implementation and optimization within hardware contexts is crucial, particularly regarding their operational demands and innovations that enhance their deployment. The following subsections explore the intricacies of CNNs in hardware, emphasizing advancements that improve functionality and efficiency.

4.1 Convolutional Neural Networks (CNNs) and Their Role in Hardware

CNNs are pivotal in advancing hardware systems, efficiently processing complex data like images and videos through hierarchical feature extraction, essential for tasks such as image recognition and natural language processing [57]. However, integrating CNNs into hardware poses challenges due to computational and memory demands from extensive parameters and high power consumption.

As illustrated in Figure 4, the role of CNNs in hardware systems encompasses key challenges and solutions, security and efficiency methods, and energy and biological model integration. This figure categorizes advancements such as HENet for computational efficiency, TrojanNet for security, and low-power vision modes, reflecting the diverse applications and innovations in CNN hardware integration.

Innovative architectures like HENet address these challenges by optimizing accuracy, speed, and storage using group convolutions and element-wise operations, enhancing CNN efficiency in hardware [38]. The RMDL ensemble method, training multiple DNN, CNN, and RNN models in parallel, achieves superior performance, showcasing ensemble learning's potential in CNN-based hardware optimization [42].

In security, CNNs detect Trojans in Quantum Approximate Optimization Algorithm (QAOA) circuits, bolstering quantum hardware security [47]. The Volterra-based convolution method, enhancing CNN expressiveness by modeling complex visual stimuli, is crucial for advanced image processing [58].

Optimizing DSP block utilization for CNNs on FPGAs highlights CNNs' role in hardware, focusing on computational efficiency and resource use [35]. Configurable imaging pipelines operating in low-power vision modes reduce energy consumption, illustrating potential for energy-efficient CNN hardware implementations [50].

Despite advancements, CNNs often lack lateral connections typical of biological visual systems, limiting contextual information integration essential for robust object recognition. Developments like KerCNNs, incorporating biologically inspired lateral connections, enhance CNN stability against image corruptions, suggesting improved performance in global shape analysis and pattern completion tasks [59, 2, 60, 61, 62]. Continued research in CNN architecture and hardware integration promises to address performance and security challenges while optimizing resources.

4.2 Optimization of CNN Inference on Integrated GPUs

Optimizing CNN inference on integrated GPUs is vital for enhanced computational performance and energy efficiency, crucial for real-time applications like robotics and environmental monitoring. CascadeCNN exemplifies a strategic approach, employing a two-stage architecture combining low- and high-precision processing units to maximize CNN inference performance [63].

KerCNN's biologically inspired lateral connections, defined by structured kernels, enhance CNN adaptability to varying input conditions during image classification [61]. Thermal throttling significantly impacts CNN performance, necessitating thermal management strategies for optimized CNN performance on integrated GPUs [64].

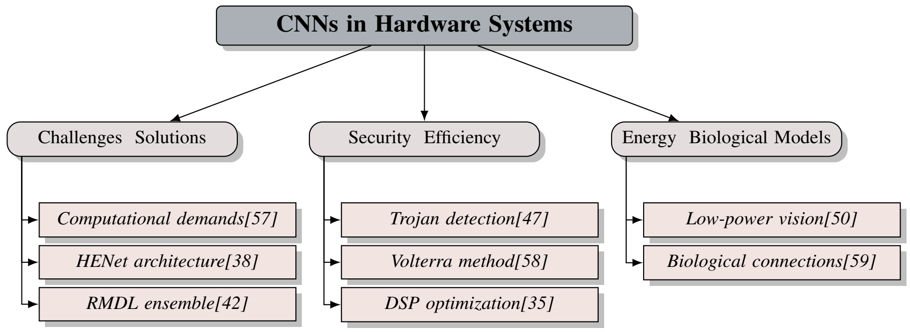

Figure 4: This figure illustrates the role of Convolutional Neural Networks (CNNs) in hardware systems, highlighting key challenges and solutions, security and efficiency methods, and energy and biological model integration. It categorizes advancements such as HENet for computational efficiency, TrojanNet for security, and low-power vision modes, reflecting the diverse applications and innovations in CNN hardware integration.

Architectures like Chain-NN achieve remarkable energy efficiency, significantly outperforming existing solutions by minimizing data movement to optimize resource utilization [43]. The Deep Convolutional Decision Jungle (CDJ) method enhances interpretability and accuracy by utilizing soft routing, crucial for optimizing CNN inference on integrated GPUs [65].

Distributed CNN training partitions forward and backward passes into independently processed tiles, minimizing communication and optimizing integrated GPU performance [36]. The Volterra-based convolution method, evaluated on CIFAR datasets, shows improved image classification performance, suggesting potential CNN inference efficiency enhancements [58]. HENet's integration of element-wise operations enhances network efficiency, benefiting integrated GPU implementations [38].

Holistic Design Space Exploration (HDSE) optimizes DSP utilization in CNN implementations on FPGAs, offering insights relevant to optimizing CNN inference on integrated GPUs [35].

These advancements focus on maximizing computational efficiency and GPU utilization, crucial for enhancing application-level throughput and ensuring a favorable return on investment. Efforts emphasize energy savings and reliable performance through innovative methodologies integrating hardware and software optimizations. Machine learning-based scheduling and unified intermediate representations significantly improve inference performance across integrated GPU architectures, maintaining flexibility for new model adoption. NeoCPU illustrates joint optimization at operation and graph levels, reducing CNN inference latency on CPUs, broadening efficient deployment in edge devices and cloud environments [66, 67, 68, 69].

4.3 Innovations in CNN Architecture for Hardware

Innovations in CNN architectures for hardware focus on enhancing computational efficiency and adaptability while maintaining high performance in feature extraction and classification tasks. Ker-CNN exemplifies such innovations by integrating structured lateral connections, improving object recognition under challenging conditions [61].

Meta-SpikeFormer introduces a meta architecture integrating spike-driven self-attention and novel designs, enhancing performance compared to CNN-based Spiking Neural Networks (SNNs) [70]. This highlights the potential of combining CNN architectures with advanced attention mechanisms for efficient neural processing on hardware platforms.

The Deep Convolutional Decision Jungle (CDJ) method applies class purity measures from decision forests to all CNN layers, allowing dynamic activation and improved interpretability without increasing model complexity [65]. This enhances interpretability, crucial in hardware implementations where transparency and reliability are paramount.

ViT-ResNAS integrates residual spatial reduction and neural architecture search (NAS) to enhance Vision Transformers for computer vision tasks [71]. This trend leverages NAS to optimize CNN architectures for specific tasks, ensuring efficient resource utilization in hardware systems.

HENet varies group convolution groups based on input channels, improving network performance compared to existing methods [38]. This innovation enhances CNN architecture flexibility and efficiency, suitable for hardware implementations with limited computational resources.

The holistic approach to optimizing DSP block utilization for CNNs on FPGAs introduces the TPR/DSP metric, measuring CNN implementation efficiency by relating classification performance to DSP resource usage [35]. This metric provides valuable insights into performance and resource utilization trade-offs, guiding efficient CNN architecture design for hardware.

Recent advancements in CNN architectures highlight efforts to enhance design and optimization for contemporary hardware environments. Techniques like automated architecture search using genetic algorithms and hill climbing methods, and frameworks like NeoCPU for optimizing inference on CPUs, pave the way for efficient and robust processing capabilities. Tools like CNN EXPLAINER facilitate deeper CNN understanding among non-experts, democratizing access to deep learning technologies. These developments improve CNN performance across benchmarks and reduce computational resources for training and inference, ensuring modern hardware's full potential is leveraged [72, 69, 38, 68, 12]. These advancements address computational challenges and enhance performance in various applications, underscoring architectural innovations' critical role in advancing CNN implementation in hardware systems.

As shown in Figure 5, recent CNN integration into hardware has seen significant advancements, particularly in designing architectures optimized for hardware. This figure illustrates key innovations in CNN architecture for hardware, categorized into efficiency improvements, advanced techniques, and hardware optimization strategies. The efficiency improvements focus on enhancing computational speed and reducing power consumption, while advanced techniques incorporate new architectural designs for improved performance. Hardware optimization strategies emphasize resource utilization and architecture search for better integration into hardware systems. Notable examples include reusing weight matrices in neural networks and converting Caffe models into hardware description formats for FPGA implementation. The first explores reusing weight matrices across multiple stages, optimizing computational efficiency through popcount operations and similarity checks. The second provides a flowchart for transforming a Caffe model into a hardware description format (Caph) for FPGA deployment using the Hadoc tool, underscoring seamless CNN model integration into hardware environments. These innovations demonstrate potential performance and energy efficiency improvements in CNN hardware implementations, paving the way for broader CNN adoption in real-world applications with critical hardware constraints [73, 35].

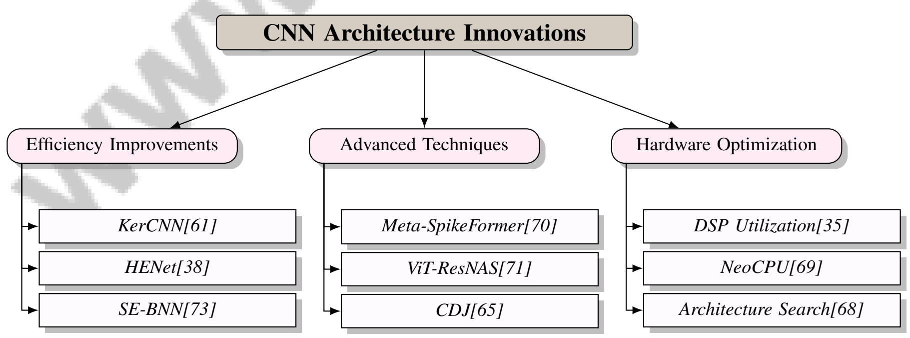

Figure 5: This figure illustrates key innovations in CNN architecture for hardware, categorized into efficiency improvements, advanced techniques, and hardware optimization strategies. The efficiency improvements focus on enhancing computational speed and reducing power consumption, while advanced techniques incorporate new architectural designs for improved performance. Hardware optimization strategies emphasize resource utilization and architecture search for better integration into hardware systems.

5 Transformers and Large Language Models (LLMs) in Semiconductor Design

5.1 Transformers, Large Language Models (LLMs), and Their Application in Semiconductor Design

Transformers and Large Language Models (LLMs) are transforming semiconductor design, enhancing methodologies and processes through applications like engineering assistant chatbots, EDA script generation, and bug summarization, as demonstrated by the ChipNeMo benchmark [52]. These models streamline design workflows by automating complex tasks and offering intelligent support throughout the design cycle. The increasing demand for computation throughput, memory, and communication bandwidth driven by LLMs has outpaced advancements in chip designs, prompting the need for innovative solutions [74]. LLMs' performance in coding assistance for chip design is evaluated through benchmarks assessing efficacy and total cost of ownership compared to state-of-the-art models [75].

In electronic design automation (EDA), LLMs automate testbench generation and enhance bug detection in RTL designs, utilizing feedback from EDA tools to improve verification processes [76]. This capability is crucial for ensuring the reliability of semiconductor designs, streamlining verification, and reducing testing time and resources. Transformer-based architectures, such as the Spike-Driven Transformer, hold promise for advancing neuromorphic chip designs, showcasing potential enhancements in intelligent functionalities [70]. Vision Transformers, particularly those enhanced by architectures like ViT-ResNAS, significantly improve performance and efficiency in tasks requiring high-level reasoning and decision-making [71].

The HDLdebugger framework exemplifies the integration of data generation, a search engine, and LLM fine-tuning to streamline HDL code debugging [49]. This illustrates the transformative impact of LLMs on semiconductor design, facilitating more efficient debugging and verification processes. Furthermore, LLMs in IC design automate HDL code generation, simplifying the design process and mitigating the complexity and time associated with traditional methodologies [77]. These advancements underscore the transformative potential of Transformers and LLMs in semiconductor design, paving the way for innovation and efficiency in design processes. By leveraging AI technologies, the semiconductor industry can achieve significant advancements in design automation, knowledge management, and process optimization.

5.2 Knowledge Graphs and Intelligent Functionalities

Transformers and LLMs significantly enhance the integration of knowledge graphs and intelligent functionalities in semiconductor design, improving efficiency and accuracy. The Oracle-Checker scheme exemplifies this by using background facts to improve knowledge graph accuracy generated by LLMs, enhancing intelligent functionalities within design processes [78]. This approach supports robust, context-aware knowledge generation, essential for the complex demands of semiconductor design.

As illustrated in Figure 6, the integration of knowledge graphs and AI technologies in semiconductor design highlights key enhancements in knowledge graph accuracy, the role of domain-specific large language models (LLMs), and efforts to reduce barriers in training foundation models. Domain-specific LLMs, as evidenced by the ChipNeMo benchmark, outperform general-purpose models on specialized tasks, underscoring the importance of domain-specific tailoring for optimal performance [52]. This specialization enables more effective LLM utilization in generating and managing knowledge graphs, enhancing decision-making and problem-solving capabilities. Technological advancements that lower barriers to foundation model training democratize access to these powerful tools, enabling broader applications and innovations in semiconductor design [79]. By reducing training costs, a wider array of stakeholders can leverage Transformers and LLMs, fostering innovation and collaboration.

These developments highlight the transformative potential of integrating knowledge graphs and intelligent functionalities with advanced AI technologies in semiconductor design. By harnessing the capabilities of Transformers and LLMs, the semiconductor industry can enhance design automation, streamline knowledge management, and improve intelligent functionalities. This integration facilitates the automation of critical processes such as architecture specification development and HDL generation, leading to more efficient and innovative design workflows. Implementing smart knowledge graphs can enhance information retrieval and visibility of past design failures, optimizing the design process and reducing time and resource consumption in the complex landscape of electronic design automation (EDA) [8, 9, 10].

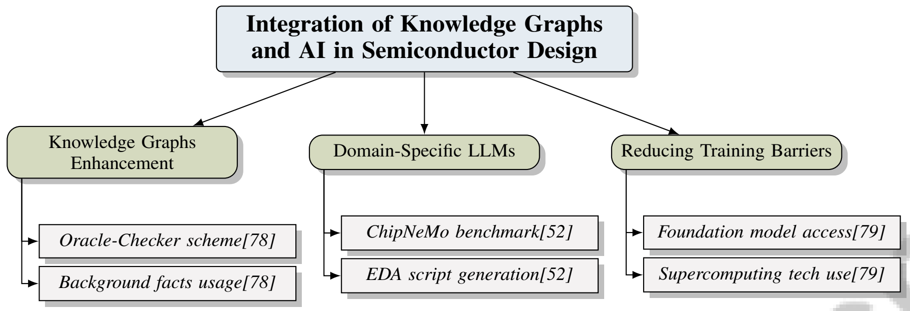

Figure 6: This figure illustrates the integration of knowledge graphs and AI technologies in semiconductor design, highlighting key enhancements in knowledge graph accuracy, the role of domain-specific large language models (LLMs), and efforts to reduce barriers in training foundation models.

6 AI-Driven Innovations in Hardware Systems

Category	Feature	Method
AI-Driven Innovations in Semiconductor Manufacturing	Feedback and Scheduling	N/A[80], LLM-HDL[77], ADDAF[32]
	Distributed and Real-Time Optimization	S4oC[48], DCNN-T[36]
	Memory and Computation Efficiency	IMC[31], FeFET-CiM[33], MTNPEO[3]
AI-Driven Optimization Techniques	Parallel and Collaborative Processing	MPNA[41], FLNet[54]
	Learning and Feedback Systems	DRL-CP[55], LLM-TG-BD[76]
	Statistical and Experimental Methods	TDCNN[81], MSN[82]
	Model Enhancement Techniques	CNN[45], RMDL[42]

Table 1: This table summarizes the various AI-driven innovations and optimization techniques applied in semiconductor manufacturing and hardware systems. It categorizes the advancements into AI-driven innovations in semiconductor manufacturing and AI-driven optimization techniques, detailing specific features and methods employed to enhance efficiency, performance, and scalability. The table serves as a comprehensive overview of the state-of-the-art methodologies and their respective contributions to the field.

The infusion of artificial intelligence (AI) into hardware systems has propelled significant advancements in efficiency, performance, and scalability, particularly within semiconductor manufacturing. Table 2 provides a comprehensive overview of AI-driven innovations and optimization techniques in semiconductor manufacturing and hardware systems, highlighting the key features and methods that contribute to these advancements. This section delves into specific AI applications that illustrate these transformative effects, highlighting how AI technologies are redefining design and production methodologies.

6.1 AI-Driven Innovations in Semiconductor Manufacturing

AI technologies are reshaping semiconductor manufacturing by enhancing efficiency, accuracy, and scalability. The utilization of Large Language Models (LLMs) in the design and verification of semiconductor components exemplifies this transformation, as evidenced by a three-phase PWM generator that successfully passed design verification, streamlining the design process [77]. Integrating iterative feedback from Electronic Design Automation (EDA) tools into LLM-generated testbenches further optimizes design processes, improving test coverage and bug detection, thereby enhancing device performance [76].

The Shisha framework highlights AI's capability in efficiently scheduling CNN pipelines on heterogeneous computing platforms, achieving comparable results to exhaustive searches while significantly reducing exploration time [80]. AI's role extends to memory technologies, with the In-Memory Classifier offering substantial energy savings and classification accuracy akin to discrete systems, promoting energy-efficient solutions in manufacturing [31]. Automated dataset generation, coupled with

EDA tool feedback, significantly enhances model performance, addressing data scarcity challenges [32].

AI's contribution to resource optimization is exemplified by the FeFET-CiM annealer, which achieves superior energy efficiency and scalability for large combinatorial optimization problems through in-memory computations [33]. In neuroanatomical analysis, AI workflows facilitate high-throughput processing of large datasets, enhancing automation and efficiency in data-intensive tasks [2]. Multi-task network pruning on embedded systems exemplifies efficient resource utilization and real-time operation capabilities, underscoring AI's impact on maintaining competitive performance across multiple tasks [3].

Distributed training methods for CNNs on mobile and edge devices demonstrate significant speedup and reduced memory usage without sacrificing accuracy, showcasing AI's role in optimizing computational processes for semiconductor manufacturing [36]. The S4oC framework's dynamic adaptation to unknown security threats and real-time resource optimization further emphasizes AI's transformative effect on manycore systems [48].

Innovations such as dynamic knowledge graphs and explainable search engines illustrate AI's potential in semiconductor manufacturing by enhancing information retrieval and visibility through interlinking production-related data, enabling design engineers to learn from past failures. Consequently, semiconductor manufacturing processes can become more efficient and cost-effective while achieving higher performance outcomes [22, 8]. The semiconductor industry continues to leverage AI to push the boundaries of technological capabilities.

6.2 AI-Driven Optimization Techniques

AI-driven optimization techniques are revolutionizing hardware systems by employing sophisticated methodologies to enhance performance and efficiency. Reinforcement learning in chip placement exemplifies AI's potential to refine design processes through experiential learning, optimizing new placements and improving layout efficiency [55]. Taguchi's design of experiments optimizes CNN parameters, systematically identifying configurations that enhance defect detection accuracy [81].

The MSN approach in deep neural networks explores the search space by maintaining multiple candidate solutions, achieving superior optimization outcomes through solution diversity [82]. The MPNA framework enhances performance by leveraging parallelism in heterogeneous computing arrays, optimizing dataflows to significantly reduce memory access [41]. Collaborative intelligence through federated learning, as illustrated by FLNet, allows multiple clients to improve model accuracy while preserving data privacy [54].

In CNN optimization, feature enhancement-collection blocks estimate performance metrics accurately with minimal execution time, integrating AI to refine model performance while minimizing computational overhead [45]. The iterative feedback loop in LLM-aided testbench generation exemplifies AI's role in enhancing verification processes through continuous learning from EDA tool outputs, improving test accuracy and comprehensiveness [76].

Optimization techniques also address security challenges in hardware systems, as conventional methods often face vulnerabilities to advanced attacks and entail significant overhead in power, performance, and area (PPA) [23]. AI-driven strategies are essential for mitigating these vulnerabilities and enhancing hardware robustness.

RMDL's ensemble learning approach improves accuracy and robustness by handling diverse data types and reducing overfitting through dropout techniques [42]. These advancements underscore the transformative impact of AI-driven optimization techniques on hardware systems, offering innovative solutions for design and implementation efficiency. By harnessing AI, the hardware industry continues to advance performance and efficiency, meeting the demands of modern technology.

7 Challenges and Future Directions

The integration of AI technologies in integrated circuit (IC) design encounters significant challenges that must be addressed to realize its full potential. This section explores the obstacles in IC design, including resource limitations, complexities of hardware description languages (HDLs), and security vulnerabilities arising from AI methodologies, providing a comprehensive overview of the barriers to

Feature	AI-Driven Innovations in Semiconductor Manufacturing	AI-Driven Optimization Techniques
Efficiency Enhancement	Energy Savings	Performance Improvement
Optimization Technique	Iterative Feedback	Reinforcement Learning
Application Area	Semiconductor Design	Chip Placement

Table 2: This table presents a comparative analysis of AI-driven innovations and optimization techniques in semiconductor manufacturing. It highlights key features such as efficiency enhancement, optimization techniques, and application areas, demonstrating the diverse roles AI plays in advancing semiconductor processes and design methodologies.

innovation in AI-driven IC design while paving the way for potential solutions and future research directions.

7.1 Challenges in Integrated Circuit (IC) Design

AI integration into IC design faces numerous challenges that hinder its effective implementation. A primary issue is the limited resources for training Large Language Models (LLMs) on HDL code, which restricts their ability to accurately interpret and repair HDL syntax and functionality [49]. This limitation is compounded by syntax errors and semantic inaccuracies in HDL generated by existing LLMs, adversely affecting the reliability and efficiency of automated design processes [77].

Fixed-function imaging pipelines optimized for traditional photography result in inefficient processing power and energy use in computer vision applications, highlighting the need for adaptable imaging solutions [50]. The complexity of unstructured pruning methods further complicates acceleration compared to structured approaches, particularly in resource-constrained environments.

Security poses another significant challenge, particularly in detecting Trojans in compiled circuits, complicating AI integration in quantum circuit design and introducing substantial risks. Existing security methods often lack the adaptability to address emerging threats in real-time, leaving systems vulnerable to sophisticated adversarial attacks, such as those targeting convolutional neural networks (CNNs) [11, 83, 84, 8].

The variability in performance of ensemble methods, such as Random Multimodel Deep Learning (RMDL), underscores the need for deterministic modeling approaches to ensure consistent outcomes across diverse datasets, especially as data complexity increases in fields like image and text classification [83, 85, 42, 86, 11]. Moreover, the computational complexity introduced by additional non-linear parameters in methods such as Volterra-based convolution necessitates more efficient strategies in IC design.

The challenges of writing and maintaining parsers for technical drawings, which require expert knowledge, further complicate AI integration in IC design. Other significant barriers include the need for extensive datasets for effective model training, improved interpretability of deep learning models to enhance user trust, and substantial computational resources for training advanced models, which limit accessibility and innovation in the field [79, 5, 12].

These multifaceted challenges necessitate ongoing research and innovation focused on developing advanced methodologies that enhance IC capabilities while addressing the unique demands posed by AI algorithms and heterogeneous computing architectures. Leveraging approaches such as knowledge graph composition for improved data retrieval and employing deep learning techniques can facilitate adaptation to the evolving landscape of chip design and meet the increasing computational demands of AI applications [7, 14, 5, 13, 8].

7.2 Future Research Directions

The integration of AI in chip design opens numerous avenues for future research aimed at enhancing performance, efficiency, and adaptability. A key area involves optimizing model architectures and hyperparameters to improve resource efficiency and operational performance, particularly within neural architecture search frameworks. Continued refinement of architecture design guidelines is essential for advancing AI-driven methodologies in chip design [87].

Future research should also focus on scaling models like HENet for larger networks and datasets, particularly in applications such as object detection and instance segmentation [38]. Additionally, refining Power Normalization functions across various neural network architectures could yield significant performance improvements [37].

Enhancing the resilience of intellectual property (IP) protection techniques is vital, including the exploration of novel materials and methods to address limitations in current approaches [23]. Research should also target the refinement of out-of-distribution (OOD) detection mechanisms and the adaptability of reinforcement learning agents to bolster the robustness of AI-driven methodologies in chip design [88].

Further exploration into advanced learning techniques for automating the verification of novel designs and identifying constraints will be crucial for AI integration [4]. Enhancing the iterative debugging capabilities of LLMs and investigating multi-LLM architectures could expand the range of HDL debugging tasks addressed [49].

In medical imaging, research should prioritize developing specialized architectures tailored to specific tasks and exploring novel modalities for enhanced diagnostic capabilities [57]. Improving training datasets for LLMs and integrating visual data processing capabilities will also enhance HDL code generation [77].

The exploration of advanced methods for counteracting outlier sensitivity in causal reasoning (CR) generation and integrating automated complicating variable identification techniques can enhance chip design frameworks [17]. Additionally, future research should aim to enhance learning algorithms to better predict and counteract emerging security threats while efficiently managing complex interactions within multi-layer graph models [48]. Addressing these opportunities will drive significant advancements in AI and chip design, leading to more intelligent, efficient, and versatile hardware systems.

7.3 Data and Training Challenges

The integration of AI into chip design faces substantial challenges related to data availability and training, critical for developing and optimizing AI models. A significant hurdle is the scarcity of high-quality, labeled training data necessary for accurate and reliable model training. The Deep Chaos Synchronization (DCS) method mitigates this issue by reducing dependency on labeled data, addressing challenges associated with data availability in AI-integrated chip design [89].

The effectiveness of AI models is contingent on the quality and diversity of training data. Current data augmentation techniques, particularly in the context of porous media, highlight the challenges of generating diverse and representative datasets that enhance model robustness and generalization [90]. Additionally, reliance on confidential design data from companies presents barriers to effective machine learning application in chip design due to privacy concerns [54].

Noise introduced during image encoding, particularly from zero-padding, complicates data preparation and may affect the fidelity and utility of training datasets [91]. Moreover, the dependency on initial Discrete Cosine Transform (DCT) coefficients in semantic communication methods underscores AI models' sensitivity to data preprocessing techniques, impacting inference accuracy if critical components are lost [40].

Developing effective defenses against cyberattacks targeting neuronal behavior presents additional challenges in AI-integrated chip design, emphasizing the need for robust data security and integrity [92]. Enhancing AI models' adaptability to various circuit types and improving efficiency in handling larger-scale designs also remain critical areas for ongoing research [93].

These challenges illustrate a multifaceted landscape of data management and training requirements, underscoring the need for innovative solutions addressing issues such as effective information retrieval from extensive documentation, integration of heterogeneous chiplet architectures, and establishing standards for interconnecting diverse computing resources. Such advancements are essential for overcoming design process obstacles and propelling the field forward [5, 14, 7, 8].

7.4 Architectural and Computational Constraints

The integration of AI technologies in chip design presents significant architectural and computational constraints that must be addressed to optimize performance and efficiency. A key challenge is the rapid evaluation of numerous configurations without relying on time-consuming compiler runs, as demonstrated by methods predicting memory compiler performance [94]. This capability is crucial for addressing architectural constraints and enabling quicker iterations and design optimizations.

The complexity introduced by layer grouping and normalization adjustments in Mini-Batch Serialization (MBS) highlights computational constraints in AI integration, necessitating careful resource management to ensure efficient training of Convolutional Neural Networks (CNNs) [95]. This complexity is compounded by the need to account for adversarial attack transferability across different network architectures, a factor often overlooked in existing benchmarks [84].

Data heterogeneity among clients poses additional challenges in model convergence, complicating AI integration in chip design. This variability affects machine learning model performance, necessitating strategies to manage data diversity and ensure robust model training [54]. Furthermore, scaling challenges arise in architectures like Chain-NN, particularly when dealing with extremely large networks or highly variable input sizes that deviate from the designed architecture [43].

The Configurable Imaging Pipeline (CIP) method exemplifies an approach to mitigate computational constraints by approximating minimal processing directly in the sensor, thereby reducing extensive processing power requirements [50]. This strategy reflects broader efforts to streamline processing requirements in vision tasks, emphasizing the importance of architectural innovations in overcoming computational limitations.

The constraints identified in the literature underscore the urgent need for innovative solutions and methodologies to enhance AI integration in chip design. Such integration is crucial for developing efficient and scalable hardware systems capable of addressing the increasing computational demands of modern applications, particularly concerning heterogeneous chiplet architectures that promise improved cost efficiency and reduced design complexity. Additionally, leveraging dynamic knowledge graphs can facilitate better visibility and retrieval of critical information from extensive documentation, supporting informed decision-making during the chip design process and ultimately leading to more robust and adaptable computing systems [14, 8].

7.5 Security and Ethical Considerations

Integrating AI into chip design necessitates a thorough examination of security and ethical considerations to ensure the responsible deployment of AI technologies. A critical security concern is the vulnerability of IC designs to unauthorized access and manipulation during fabrication and supply chain processes, which can lead to intellectual property (IP) theft and the insertion of hardware Trojans. This issue highlights the need for robust security measures to protect IC designs from exploitation [23].

In AI-powered chip design, the use of LLMs introduces additional security challenges, particularly regarding the trustworthiness of datasets used for training these models. Mitigating risks associated with untrusted datasets is essential to prevent the propagation of biases and vulnerabilities in LLM-powered chip design processes [96]. Furthermore, verifying CNNs without altering their architecture, as enabled by fuzzy logic, presents a promising approach to managing output uncertainty, thereby enhancing AI model reliability and security [97].

Adversarial transferability poses significant challenges in AI integration, as adversarial attacks can exploit vulnerabilities across network architectures. Strategies to mitigate these effects are crucial for ensuring robust model training and evaluation, thus enhancing the security of AI-driven chip designs [84]. While innovative methods like reconfiguring imaging pipelines can optimize processing efficiency, they may not be suitable for all vision applications, particularly those requiring higher fidelity images or specific image signal processing (ISP) stages, necessitating careful consideration of application-specific requirements [50].

Ethical considerations are pivotal in the AI and chip design integration process. Ensuring that AI technologies are developed and deployed with respect for privacy, fairness, and transparency is essential for fostering trust and acceptance of AI-integrated systems [21].

Collectively, these security and ethical considerations emphasize the importance of developing comprehensive frameworks and strategies to address the complex challenges of AI integration in chip design. By prioritizing security measures and adhering to ethical principles throughout the semiconductor design process, the industry can effectively mitigate risks associated with IP theft, hardware Trojans, and other vulnerabilities, fostering the safe and responsible development of

AI-driven technologies that are increasingly integral to our professional, social, and private lives [8, 23, 21].

8 Conclusion

The incorporation of artificial intelligence (AI) into chip design marks a transformative leap in semiconductor technology, offering substantial improvements in efficiency, performance, and functionality across diverse applications. This evolution is exemplified by the strategic implementation of 3D integration in neuromorphic computing, which not only enhances performance but also reduces costs, thereby demonstrating the capability of AI-driven methodologies to simplify intricate design processes. The deployment of CNN-based workflows in neuroanatomical mapping further exemplifies AI's transformative influence, achieving heightened accuracy and efficiency in processing extensive datasets.

In the realm of quantum circuit security, the efficacy of TrojanNet in detecting Trojans with high accuracy highlights AI's pivotal role in fortifying the security of sophisticated computing systems. Additionally, the development of robust watermarking frameworks is crucial to counteract emerging threats, ensuring the ethical application of generative AI tools and preserving the integrity of AI-integrated systems.

The proposed CNN framework effectively balances performance with computational complexity, proving its utility as a real-time performance estimation tool indispensable for optimizing hardware systems. These advancements collectively highlight AI's integral role in chip design, ushering in a new era of innovation and efficiency within semiconductor technology.

As AI-driven methodologies continue to evolve, they are set to revolutionize the industry by enabling the development of more intelligent, efficient, and adaptable hardware systems. By harnessing AI's potential, the semiconductor industry stands to achieve significant progress in design automation, knowledge management, and process optimization, propelling the next wave of technological advancements.

AI-generated, for reference only.

References

[1] Eren Kurshan, Hai Li, Mingoo Seok, and Yuan Xie. A case for 3d integrated system design for neuromorphic computing ai applications, 2021.

[2] Christian Schiffer, Hannah Spitzer, Kai Kiwitz, Nina Unger, Konrad Wagstyl, Alan C. Evans, Stefan Harmeling, Katrin Amunts, and Timo Dickscheid. Convolutional neural networks for cytoarchitectonic brain mapping at large scale, 2020.

[3] Flora Dellinger, Thomas Boulay, Diego Mendoza Barrenechea, Said El-Hachimi, Isabelle Leang, and Fabian Bürger. Multi-task network pruning and embedded optimization for real-time deployment in adas, 2021.

[4] Dries Van Daele, Nicholas Decleyre, Herman Dubois, and Wannes Meert. An automated engineering assistant: Learning parsers for technical drawings, 2019.

[5] Jeffrey Dean. The deep learning revolution and its implications for computer architecture and chip design, 2019.

[6] M. Guthaus, C. Batten, E. Brunvand, P. E. Gaillardon, D. harris, R. Manohar, P. Mazumder, L. Pileggi, and J. Stine. Nsf integrated circuit research, education and workforce development workshop final report, 2023.

[7] Anna Goldie, Azalia Mirhoseini, and Jeff Dean. That chip has sailed: A critique of unfounded skepticism around ai for chip design, 2024.

[8] H. Abu-Rasheed, C. Weber, J. Zenkert, P. Czerner, R. Krumm, and M. Fathi. A text extraction-based smart knowledge graph composition for integrating lessons learned during the microchip design, 2021.

[9] Mengming Li, Wenji Fang, Qijun Zhang, and Zhiyao Xie. Specllm: Exploring generation and review of vlsi design specification with large language model, 2024.

[10] Ruizhe Zhong, Xingbo Du, Shixiong Kai, Zhentao Tang, Siyuan Xu, Hui-Ling Zhen, Jianye Hao, Qiang Xu, Mingxuan Yuan, and Junchi Yan. Llm4eda: Emerging progress in large language models for electronic design automation, 2023.

[11] Abdallah Moubayed, MohammadNoor Injadat, Nouh Alhindawi, Ghassan Samara, Sara Abuasal, and Raed Alazaidah. A deep learning approach towards student performance prediction in online courses: Challenges based on a global perspective, 2024.

[12] Nn e xplainer: Learning convolut.

[13] Tamay Besiroglu, Nicholas Emery-Xu, and Neil Thompson. Economic impacts of ai-augmented rd, 2023.

[14] Zhuoping Yang, Shixin Ji, Xingzhen Chen, Jinming Zhuang, Weifeng Zhang, Dharmesh Jani, and Peipei Zhou. Challenges and opportunities to enable large-scale computing via heterogeneous chiplets, 2024.

[15] Víctor Mayoral-Vilches, Juan Manuel Reina-Muñoz, Martiño Crespo-Álvarez, and David Mayoral-Vilches. Ros 2 on a chip, achieving brain-like speeds and efficiency in robotic networking, 2024.

[16] Animesh Basak Chowdhury, Marco Romanelli, Benjamin Tan, Ramesh Karri, and Siddharth Garg. Retrieval-guided reinforcement learning for boolean circuit minimization, 2024.

[17] Niki Triantafyllou and Maria M. Papathanasiou. Deep learning enhanced mixed integer optimization: Learning to reduce model dimensionality, 2024.

[18] P. J. Zhou, Q. Yu, M. Chen, Y. C. Wang, L. W. Meng, Y. Zuo, N. Ning, Y. Liu, S. G. Hu, and G. C. Qiao. A 0.96pj/sop, 30.23k-neuron/mm$^2 heterogeneousneuromorphicchipwithfullerene-likeinterconnectiontopologyforedge-aicomputing$, 2024.

[19] Xingquan Li, Simin Tao, Zengrong Huang, Shijian Chen, Zhisheng Zeng, Liwei Ni, Zhipeng Huang, Chunan Zhuang, Hongxi Wu, Weiguo Li1, Xueyan Zhao, He Liu, Shuaiying Long, Wei He, Bojun Liu, Sifeng Gan, Zihao Yu, Tong Liu, Yuchi Miao, Zhiyuan Yan, Hao Wang, Jie Zhao, Yifan Li, Ruizhi Liu, Xiaoze Lin, Bo Yang, Zhen Xue, Fuxing Huang, Zonglin Yang, Zhenggang Wu, Jiangkao Li, Yuezuo Liu, Ming Peng, Yihang Qiu, Wenrui Wu, Zheqing Shao, Kai Mo, Jikang Liu, Yuyao Liang, Mingzhe Zhang, Zhuang Ma, Xiang Cong, Daxiang Huang, Guojie Luo, Huawei Li, Haihua Shen, Mingyu Chen, Dongbo Bu, Wenxing Zhu, Ye Cai, Xiaoming Xiong, Ying Jiang, Yi Heng, Peng Zhang, Biwei Xie, and Yungang Bao. ieda: An open-source intelligent physical implementation toolkit and library, 2023.

[20] Ian McDougall, Shayne Wadle, Harish Batchu, Michael Davies, and Karthikeyan Sankaralingam. Bryt: Data rich analytics based computer architecture for a new paradigm of chip design, 2024.

[21] Sudeep Pasricha and Marilyn Wolf. Ethical design of computers: From semiconductors to iot and artificial intelligence, 2022.

[22] Hasan Abu-Rasheed, Christian Weber, Johannes Zenkert, Roland Krumm, and Madjid Fathi. Explainable graph-based search for lessons-learned documents in the semiconductor industry, 2021.

[23] Johann Knechtel, Satwik Patnaik, and Ozgur Sinanoglu. Protect your chip design intellectual property: An overview, 2019.

[24] Guojin Chen, Haoyu Yang, Bei Yu, and Haoxing Ren. Intelligent opc engineer assistant for semiconductor manufacturing, 2024.

[25] Yao Lai, Yao Mu, and Ping Luo. Maskplace: Fast chip placement via reinforced visual representation learning, 2022.

[26] Tao Yu, Peng Gao, Fei Wang, and Ru-Yue Yuan. Non-overlapping placement of macro cells based on reinforcement learning in chip design, 2024.

[27] Ruoyu Cheng and Junchi Yan. On joint learning for solving placement and routing in chip design, 2021.

[28] Mingjie Liu, Haoyu Yang, Zongyi Li, Kumara Sastry, Saumyadip Mukhopadhyay, Selim Dogru, Anima Anandkumar, David Z. Pan, Brucek Khailany, and Haoxing Ren. An adversarial active sampling-based data augmentation framework for manufacturable chip design, 2022.

[29] Tzyy-Juin Kao and Wolfgang Fink. Pareto-optimization framework for automated network-on-chip design, 2018.

[30] Hailiang Li, Yan Huo, Yan Wang, Xu Yang, Miaohui Hao, and Xiao Wang. A lightweight inception boosted u-net neural network for routability prediction, 2024.

[31] Jintao Zhang, Zhuo Wang, and Naveen Verma. In-memory computation of a machine-learning classifier in a standard 6t sram array. *IEEE Journal of Solid-State Circuits*, 52(4):915–924, 2017.

[32] Kaiyan Chang, Kun Wang, Nan Yang, Ying Wang, Dantong Jin, Wenlong Zhu, Zhirong Chen, Cangyuan Li, Hao Yan, Yunhao Zhou, Zhuoliang Zhao, Yuan Cheng, Yudong Pan, Yiqi Liu, Mengdi Wang, Shengwen Liang, Yinhe Han, Huawei Li, and Xiaowei Li. Data is all you need: Finetuning llms for chip design via an automated design-data augmentation framework, 2024.

[33] Xunzhao Yin, Yu Qian, Alptekin Vardar, Marcel Gunther, Franz Muller, Nellie Laleni, Zijian Zhao, Zhouhang Jiang, Zhiguo Shi, Yiyu Shi, Xiao Gong, Cheng Zhuo, Thomas Kampfe, and Kai Ni. A ferroelectric compute-in-memory annealer for combinatorial optimization problems, 2023.

[34] Michal Pinos, Lukas Sekanina, and Vojtech Mrazek. Approxdarts: Differentiable neural architecture search with approximate multipliers, 2024.

[35] Kamel Abdelouahab, Cedric Bourrasset, Maxime Pelcat, François Berry, Jean-Charles Quinton, and Jocelyn Serot. A holistic approach for optimizing dsp block utilization of a cnn implementation on fpga, 2017.

[36] Pranav Rama, Madison Threadgill, and Andreas Gerstlauer. Distributed convolutional neural network training on mobile and edge clusters, 2024.

[37] Piotr Koniusz, Hongguang Zhang, and Fatih Porikli. A deeper look at power normalizations, 2018.

[38] Qiuyu Zhu and Ruixin Zhang. Henet:a highly efficient convolutional neural networks optimized for accuracy, speed and storage, 2018.

[39] Zhishang Luo, Truong Son Hy, Puoya Tabaghi, Donghyeon Koh, Michael Defferrard, Elahe Rezaei, Ryan Carey, Rhett Davis, Rajeev Jain, and Yusu Wang. De-hnn: An effective neural model for circuit netlist representation, 2024.

[40] Andrea Cavagna, Nan Li, Alexandros Iosifidis, and Qi Zhang. Semantic communication enabling robust edge intelligence for time-critical iot applications, 2022.

[41] Muhammad Abdullah Hanif, Rachmad Vidya Wicaksana Putra, Muhammad Tanvir, Rehan Hafiz, Semeen Rehman, and Muhammad Shafique. Mpna: A massively-parallel neural array accelerator with dataflow optimization for convolutional neural networks, 2018.

[42] Mojtaba Heidarysafa, Kamran Kowsari, Donald E. Brown, Kiana Jafari Meimandi, and Laura E. Barnes. An improvement of data classification using random multimodel deep learning (rmdl), 2018.

[43] Shihao Wang, Dajiang Zhou, Xushen Han, and Takeshi Yoshimura. Chain-nn: An energy-efficient 1d chain architecture for accelerating deep convolutional neural networks, 2017.

[44] Srivatsan Krishnan, Natasha Jaques, Shayegan Omidshafiei, Dan Zhang, Izzeddin Gur, Vijay Janapa Reddi, and Aleksandra Faust. Multi-agent reinforcement learning for microprocessor design space exploration, 2022.

[45] Toan-Van Nguyen, Thien Huynh-The, Van-Dinh Nguyen, Daniel Benevides da Costa, Rose Qingyang Hu, and Beongku An. An efficient deep cnn design for eh short-packet communications in multihop cognitive iot networks, 2022.

[46] Jingyi Wang and Shengchen Li. Keyword spotting system and evaluation of pruning and quantization methods on low-power edge microcontrollers, 2022.

[47] Subrata Das and Swaroop Ghosh. Trojannet: Detecting trojans in quantum circuits using machine learning, 2023.

[48] Shahin Nazarian and Paul Bogdan. S4oc: A self-optimizing, self-adapting secure system-on-chip design framework to tackle unknown threats – a network theoretic, learning approach, 2020.

[49] Xufeng Yao, Haoyang Li, Tsz Ho Chan, Wenyi Xiao, Mingxuan Yuan, Yu Huang, Lei Chen, and Bei Yu. Hdldebugger: Streamlining hdl debugging with large language models, 2024.

[50] Mark Buckler, Suren Jayasuriya, and Adrian Sampson. Reconfiguring the imaging pipeline for computer vision, 2017.

[51] Haoyu Yang and Haoxing Ren. Ililt: Implicit learning of inverse lithography technologies, 2024.

[52] Mingjie Liu, Teodor-Dumitru Ene, Robert Kirby, Chris Cheng, Nathaniel Pinckney, Rongjian Liang, Jonah Alben, Himyanshu Anand, Sanmitra Banerjee, Ismet Bayraktaroglu, Bonita Bhaskaran, Bryan Catanzaro, Arjun Chaudhuri, Sharon Clay, Bill Dally, Laura Dang, Parikshit Deshpande, Siddhanth Dhodhi, Sameer Halepete, Eric Hill, Jiashang Hu, Sumit Jain, Ankit Jindal, Brucek Khailany, George Kokai, Kishor Kunal, Xiaowei Li, Charley Lind, Hao Liu, Stuart Oberman, Sujeet Omar, Ghasem Pasandi, Sreedhar Pratty, Jonathan Raiman, Ambar Sarkar, Zhengjiang Shao, Hanfei Sun, Pratik P Suthar, Varun Tej, Walker Turner, Kaizhe Xu, and Haoxing Ren. Chipnemo: Domain-adapted llms for chip design, 2024.

[53] Anna Goldie and Azalia Mirhoseini. Placement optimization with deep reinforcement learning, 2020.

[54] Jingyu Pan, Chen-Chia Chang, Zhiyao Xie, Ang Li, Minxue Tang, Tunhou Zhang, Jiang Hu, and Yiran Chen. Towards collaborative intelligence: Routability estimation based on decentralized private data, 2022.

[55] Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Sungmin Bae, Azade Nazi, Jiwoo Pak, Andy Tong, Kavya Srinivasa, William Hang, Emre Tuncer, Anand Babu, Quoc V. Le, James Laudon, Richard Ho, Roger Carpenter, and Jeff Dean. Chip placement with deep reinforcement learning, 2020.

[56] Zhiyao Xie. Intelligent circuit design and implementation with machine learning, 2022.

[57] Ahmad Waleed Salehi, Shakir Khan, Gaurav Gupta, Bayan Ibrahimm Alabduallah, Abrar Almjally, Hadeel Alsolai, Tamanna Siddiqui, and Adel Mellit. A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability*, 15(7):5930, 2023.

[58] Georgios Zoumpourlis, Alexandros Doumanoglou, Nicholas Vretos, and Petros Daras. Non-linear convolution filters for cnn-based learning, 2017.

[59] Reem Abdel-Salam. Relating cnns with brain: Challenges and findings, 2021.

[60] Ihsan Ullah and Alfredo Petrosino. About pyramid structure in convolutional neural networks, 2016.

[61] Noemi Montobbio, Laurent Bonnasse-Gahot, Giovanna Citti, and Alessandro Sarti. Kercnns: biologically inspired lateral connections for classification of corrupted images, 2019.

[62] Mats L. Richter, Julius Schöning, Anna Wiedenroth, and Ulf Krumnack. Should you go deeper? optimizing convolutional neural network architectures without training by receptive field analysis, 2021.

[63] Alexandros Kouris, Stylianos I. Venieris, and Christos-Savvas Bouganis. Cascadecnn: Pushing the performance limits of quantisation, 2018.

[64] Théo Benoit-Cattin, Delia Velasco-Montero, and Jorge Fernández-Berni. Impact of thermal throttling on long-term visual inference in a cpu-based edge device, 2020.

[65] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Deep convolutional decision jungle for image classification, 2018.

[66] Jack Kosaian and Amar Phanishayee. A study on the intersection of gpu utilization and cnn inference, 2022.

[67] Leyuan Wang, Zhi Chen, Yizhi Liu, Yao Wang, Lianmin Zheng, Mu Li, and Yida Wang. A unified optimization approach for cnn model inference on integrated gpus, 2019.

[68] Thomas Elsken, Jan-Hendrik Metzen, and Frank Hutter. Simple and efficient architecture search for convolutional neural networks, 2017.

[69] Yizhi Liu, Yao Wang, Ruofei Yu, Mu Li, Vin Sharma, and Yida Wang. Optimizing cnn model inference on cpus, 2019.

[70] Man Yao, Jiakui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips, 2024.

[71] Yi-Lun Liao, Sertac Karaman, and Vivienne Sze. Searching for efficient multi-stage vision transformers, 2021.

[72] Yanan Sun, Bing Xue, Mengjie Zhang, and Gary G Yen. Completely automated cnn architecture design based on blocks. *IEEE transactions on neural networks and learning systems*, 31(4):1242–1254, 2019.

[73] Cheng Fu, Shilin Zhu, Hao Su, Ching-En Lee, and Jishen Zhao. Towards fast and energy-efficient binarized neural network inference on fpga, 2018.

[74] Jingchen Zhu, Chenhao Xue, Yiqi Chen, Zhao Wang, Chen Zhang, Yu Shen, Yifan Chen, Zekang Cheng, Yu Jiang, Tianqi Wang, Yibo Lin, Wei Hu, Bin Cui, Runsheng Wang, Yun Liang, and Guangyu Sun. Theseus: Exploring efficient wafer-scale chip design for large language models, 2024.

[75] Amit Sharma, Teodor-Dumitru Ene, Kishor Kunal, Mingjie Liu, Zafar Hasan, and Haoxing Ren. Assessing economic viability: A comparative analysis of total cost of ownership for domain-adapted large language models versus state-of-the-art counterparts in chip design coding assistance, 2024.

[76] Jitendra Bhandari, Johann Knechtel, Ramesh Narayanaswamy, Siddharth Garg, and Ramesh Karri. Llm-aided testbench generation and bug detection for finite-state machines, 2024.

[77] Maoyang Xiang, Emil Goh, and T. Hui Teo. Digital asic design with ongoing llms: Strategies and prospects, 2024.

[78] Yueling Zeng and Li-C. Wang. Domain knowledge graph construction via a simple checker, 2023.

[79] Paolo Faraboschi, Ellis Giles, Justin Hotard, Konstanty Owczarek, and Andrew Wheeler. Reducing the barriers to entry for foundation model training, 2024.

[80] Shisha: Online scheduling of cnn pipelines on heterogeneous architectures.

[81] Manjeet Kaur, Krishan Kumar Chauhan, Tanya Aggarwal, Pushkar Bharadwaj, Renu Vig, Isibor Kennedy Ihianle, Garima Joshi, and Kayode Owa. Taguchi based design of sequential convolution neural network for classification of defective fasteners, 2022.

[82] Ahmed Aly, David Weikersdorfer, and Claire Delaunay. Optimizing deep neural networks with multiple search neuroevolution, 2019.

[83] Ruisi Zhang and Farinaz Koushanfar. Watermarking large language models and the generated content: Opportunities and challenges, 2024.

[84] Ehsan Nowroozi, Yassine Mekdad, Mohammad Hajian Berenjestanaki, Mauro Conti, and Abdeslam EL Fergougui. Demystifying the transferability of adversarial attacks in computer networks, 2022.

[85] Bing-Yue Wu, Utsav Sharma, Sai Rahul Dhanvi Kankipati, Ajay Yadav, Bintu Kappil George, Sai Ritish Guntupalli, Austin Rovinski, and Vidya A. Chhabria. Eda corpus: A large language model dataset for enhanced interaction with openroad, 2024.

[86] Junhao Xu, Jianwei Yu, Shoukang Hu, Xunying Liu, and Helen Meng. Mixed precision low-bit quantization of neural network language models for speech recognition, 2021.

[87] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[88] Animesh Basak Chowdhury, Marco Romanelli, Benjamin Tan, Ramesh Karri, and Siddharth Garg. Invictus: Optimizing boolean logic circuit synthesis via synergistic learning and search, 2023.

[89] Majid Mobini and Georges Kaddoum. Deep chaos synchronization, 2021.

[90] C. H. Wong, S. M. Ng, C. W. Leung, and A. F. Zatsepin. The effectiveness of data augmentation in porous substrate, nanowire, fiber and tip images at the level of deep learning intelligence, 2021.

[91] Dan Wang, Tianrui Wang, and Ionuţ Florescu. Is image encoding beneficial for deep learning in finance? an analysis of image encoding methods for the application of convolutional neural networks in finance, 2020.

[92] Sergio López Bernal, Alberto Huertas Celdrán, and Gregorio Martínez Pérez. Neuronal jamming cyberattack over invasive bci affecting the resolution of tasks requiring visual capabilities, 2021.

[93] Ruizhe Zhong, Junjie Ye, Zhentao Tang, Shixiong Kai, Mingxuan Yuan, Jianye Hao, and Junchi Yan. Preroutgnn for timing prediction with order preserving partition: Global circuit pre-training, local delay learning and attentional cell modeling, 2024.

[94] Felix Last, Max Haeberlein, and Ulf Schlichtmann. Predicting memory compiler performance outputs using feed-forward neural networks, 2020.

[95] Sangkug Lym, Armand Behroozi, Wei Wen, Ge Li, Yongkee Kwon, and Mattan Erez. Mini-batch serialization: Cnn training with inter-layer data reuse, 2019.

[96] Zeng Wang, Lilas Alrahis, Likhitha Mankali, Johann Knechtel, and Ozgur Sinanoglu. Llms and the future of chip design: Unveiling security risks and building trust, 2024.

[97] Gesina Schwalbe, Christian Wirth, and Ute Schmid. Enabling verification of deep neural networks in perception tasks using fuzzy logic and concept embeddings, 2022.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

AI-generated, for reference only.