Advanced Technologies in Display Manufacturing and AI: A Survey on Mura, Defect Recognition, Autofocus Leveling, Micro-Display Technology, Edge AI, and TPU Optimization

www.surveyx.cn

Abstract

This survey paper explores the integration of advanced technologies in display manufacturing and artificial intelligence, focusing on mura detection, defect recognition, autofocus leveling, micro-display technology, edge AI, and TPU optimization. These technologies enhance efficiency, accuracy, and scalability across various applications. Edge AI significantly improves real-time data processing, essential for dynamic environments like the Metaverse. Micro-LED technology, noted for its brightness and energy efficiency, shows promise in augmented reality and wearable devices, though challenges like high production costs persist. The survey emphasizes the development of scalable edge infrastructures and the integration of federated learning to enhance model quality in decentralized environments. Future research should focus on optimizing resource allocation, enhancing cooperative perception, and addressing privacy concerns in multi-access edge computing. The exploration of hybrid architectures, such as combining spiking and artificial neural networks, offers potential for energy-efficient AI deployments. Additionally, partitioning strategies are vital for optimizing resource utilization in IoT-Edge-AI applications. The study concludes that integrating MRAM into XR applications can lead to significant energy savings, highlighting its potential for future hardware design. Overall, the survey underscores the transformative impact of these technologies and the need for ongoing research to address emerging challenges and support innovative applications.

1 Introduction

1.1 Significance of Keywords

The integration of advanced technologies such as mura detection, defect recognition, autofocus leveling, micro-display technology, edge AI, and TPU optimization is crucial in modern display manufacturing and artificial intelligence. Mura detection ensures high-quality displays by identifying and correcting visual inconsistencies that detract from user experience, while real-time defect recognition enhances manufacturing efficiency and product reliability, particularly in high-speed environments. The importance of AI systems is further highlighted in public health applications, such as mask detection during the COVID-19 pandemic [1].

Autofocus leveling is vital for imaging systems, where precise focus adjustments are necessary for optimal image clarity, especially in medical imaging applications. Micro-display technology has transformed the development of compact, high-resolution screens essential for consumer electronics and advanced virtual reality systems, significantly improving user experiences in emerging digital environments like the Metaverse [2].

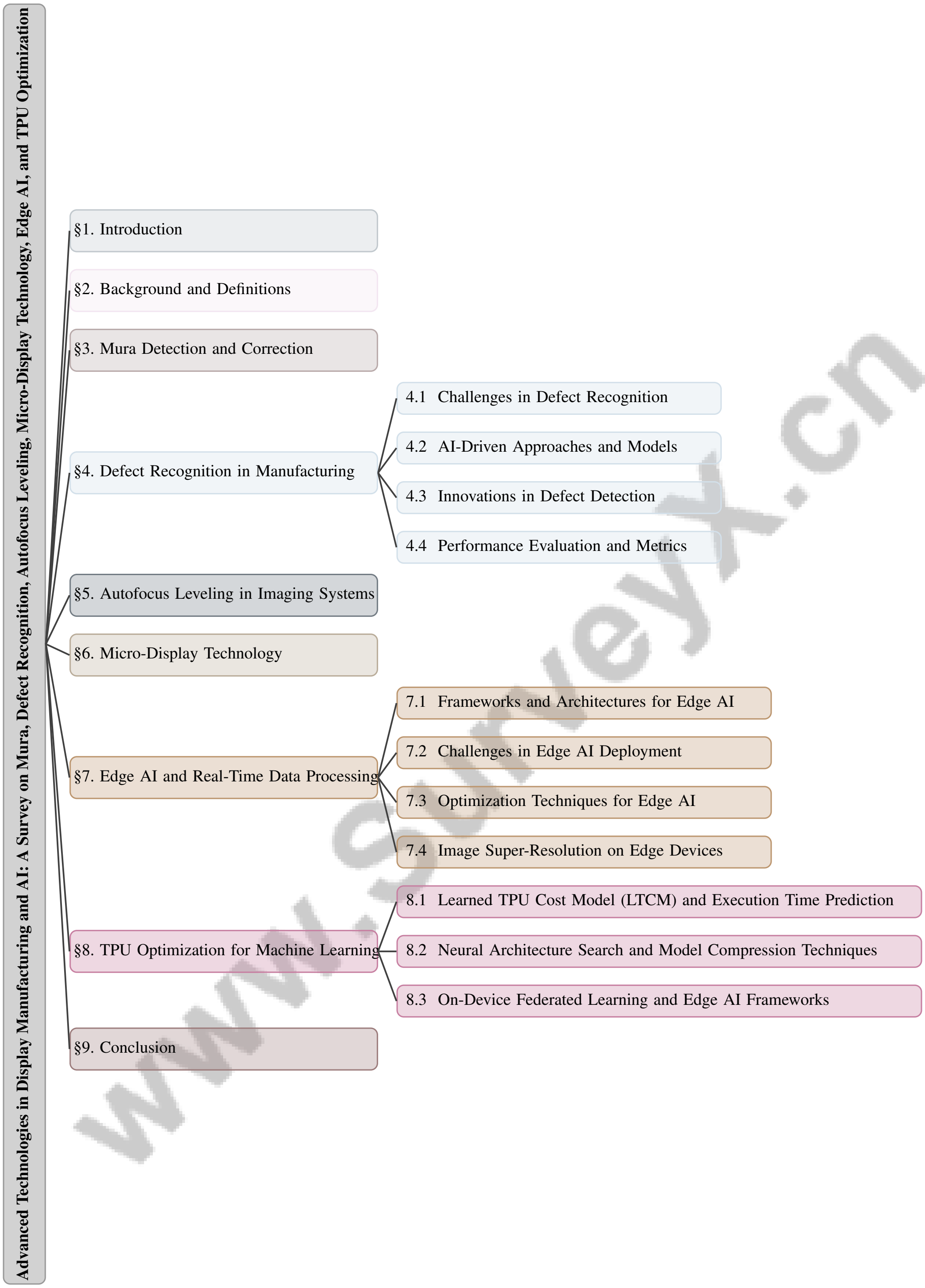
Advanced Technologies in Display Manufacturing and AI: A Survey on Mura, Defect Recognition, Autofocus Leveling, Micro-Display Technology, Edge AI, and TPU Optimization
§1. Introduction
§2. Background and Definitions
§3. Mura Detection and Correction
§4. Defect Recognition in Manufacturing
4.1 Challenges in Defect Recognition
4.2 AI-Driven Approaches and Models
4.3 Innovations in Defect Detection
4.4 Performance Evaluation and Metrics
§5. Autofocus Leveling in Imaging Systems
§6. Micro-Display Technology
§7. Edge AI and Real-Time Data Processing
7.1 Frameworks and Architectures for Edge AI
7.2 Challenges in Edge AI Deployment
7.3 Optimization Techniques for Edge AI
7.4 Image Super-Resolution on Edge Devices
§8. TPU Optimization for Machine Learning
8.1 Learned TPU Cost Model (LTCM) and Execution Time Prediction
8.2 Neural Architecture Search and Model Compression Techniques
8.3 On-Device Federated Learning and Edge AI Frameworks
§9. Conclusion

Figure 1: chapter structure

Edge AI, which deploys AI algorithms at the network's edge, enhances performance and reduces latency, particularly in time-sensitive applications requiring rapid data processing. The efficiency gains in resource management for edge computing underscore the role of AI/ML in this context [3]. TPU optimization further improves machine learning task efficiency, promoting faster and more energy-efficient computations, which is increasingly important given the environmental impact of AI technologies and the demand for sustainable computing solutions.

AI-generated, for reference only.

Each technology contributes uniquely to advancing display manufacturing and AI, addressing current limitations and paving the way for future innovations. The convergence of edge computing and AI not only enhances performance and operational efficiency but also addresses critical challenges such as data privacy, security, and sustainability. By processing vast amounts of data generated by mobile and IoT devices at the network edge, these technologies facilitate real-time analytics and decision-making, minimizing latency and reducing the environmental footprint associated with centralized cloud data transmission [4, 5, 6, 7].

1.2 Objectives of the Survey

This survey aims to examine the integration and optimization of advanced technologies in display manufacturing and artificial intelligence, focusing on mura detection, defect recognition, autofocus leveling, micro-display technology, edge AI, and TPU optimization. A primary objective is to address the challenges and limitations of existing technologies in realizing the comprehensive vision of the Metaverse, emphasizing the need for enhanced performance and reduced latency in AI-driven environments [8]. The survey also explores the convergence of edge computing and deep learning, highlighting the necessity for low-latency, efficient AI services to overcome cloud computing limitations [9].

Additionally, the survey investigates the sustainability of AI technologies, providing insights into energy efficiency and carbon emissions through large-scale energy datasets for edge AI [10]. It examines the implications of integrating edge AI in display manufacturing, particularly for defect recognition and autofocus leveling, thereby enhancing system efficiency and reliability [11]. The survey proposes strategies to enhance performance, reduce latency, and improve scalability through partitioning techniques in IoT-Edge-AI environments [12]. Furthermore, it addresses security and privacy issues within the MEC environment, focusing on the roles of Software Defined Networking (SDN) and Network Function Virtualization (NFV) in mitigating these concerns [13].

Moreover, the survey discusses the lack of hardware accelerators supporting both dynamic precision and diverse activation functions, which limits AI workload performance [14]. It also provides a comprehensive overview of various computing paradigms such as Grid, Cloud, Fog, Edge, Serverless, and Quantum Computing, deliberately excluding specific implementations to maintain a broad perspective [7]. The survey includes benchmarks for evaluating different hardware architectures for XR applications, particularly in hand detection and eye segmentation tasks [2].

The survey aims to analyze the current state and future prospects of edge computing and AI technologies, emphasizing their applications in display manufacturing and AI-driven systems. By examining the integration of deep learning and edge computing, it seeks to illuminate how these technologies can enhance operational efficiency, security, and sustainability in manufacturing processes. Furthermore, it highlights the necessity for innovative approaches to manage the challenges of deploying AI at the network edge, contributing to the evolution of more intelligent and resource-efficient systems across various industries [9, 15, 4, 16, 7].

1.3 Structure of the Survey

This survey is meticulously structured to guide the reader through the intricate landscape of advanced technologies in display manufacturing and artificial intelligence. It begins with an **Introduction**, highlighting the significance of key technologies such as mura detection, defect recognition, autofocus leveling, micro-display technology, edge AI, and TPU optimization, and outlining the objectives of the survey. Following this, the **Background and Definitions** section provides foundational understanding, defining essential terms and discussing their relevance and applications in the industry.

The survey thoroughly investigates various technologies, starting with . This section emphasizes methodologies and challenges associated with identifying and rectifying visual inconsistencies in display technologies like liquid crystal displays (LCDs) and other flat-panel displays (FPDs). It highlights the role of advanced machine vision techniques, such as deep convolutional neural networks, in enhancing defect recognition and classification while addressing complexities from environmental factors and image quality variations affecting inspection accuracy [17, 18, 15, 19, 20]. Following this is the section on **Defect Recognition in Manufacturing**, which examines AI-driven approaches and recent innovations in defect detection, along with performance evaluation metrics.

AI-generated, for reference only.

Next, the survey explores **Autofocus Leveling in Imaging Systems**, discussing advancements and specific techniques such as the Infrared Microscopy Automated Location System (IMAL) to enhance focus accuracy. The section on **Micro-Display Technology** highlights the development of compact, high-resolution screens, comparing them with traditional technologies and exploring future directions.

The discussion then shifts to **Edge AI and Real-Time Data Processing**, examining the deployment of AI at the network edge, focusing on frameworks, challenges, optimization techniques, and applications like image super-resolution. The section titled delves into advanced techniques aimed at improving Tensor Processing Units (TPUs). It highlights the implementation of the Learned TPU Cost Model (LTCM), which utilizes a neural network to predict execution times based on graph embeddings, enabling more efficient decision-making in compiler optimizations and autotuning. Additionally, it discusses integrating neural architecture search methodologies to enhance model accuracy and inference speed, illustrating how these strategies contribute to significant performance improvements in machine learning applications on TPUs [21, 22].

Finally, the **Conclusion** synthesizes key findings and insights from the survey, discussing future directions and potential advancements in display manufacturing and AI technologies. Throughout the survey, comprehensive datasets such as those introduced by the DeepEn2023 benchmark provide critical insights into energy consumption at various levels, underscoring the importance of sustainability in AI technologies [10].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Core Concepts in Display Manufacturing

The integration of advanced technologies in display manufacturing enhances the quality and efficiency of display systems. Edge AI technology plays a crucial role in real-time data processing for applications like augmented reality and interactive displays, requiring immediate responsiveness [11]. The Nonlinear Joint Transform Correlator (NL-JTC) exemplifies advancements in optical systems, enabling full amplitude-and-phase convolution with minimal latency, essential for real-time display adjustments [23]. Efficient resource management in fog and edge computing environments is critical due to resource limitations and dynamic workloads, ensuring seamless operation in resource-constrained settings [3].

Micro-LED technology, while promising for high-resolution displays, faces significant technical challenges that impede its commercialization [24]. The evolution of computing technologies is vital for addressing diverse social needs and advancing capabilities in display manufacturing, driven by the demand for efficient, scalable, and adaptable solutions [7]. High efficiency in data processing and reduced latency are foundational in display manufacturing, especially for edge AI applications, ensuring the performance and reliability of modern display systems [25]. Developing low-power, efficient hardware is imperative for extended reality (XR) applications, which require intensive computational resources for tasks such as hand detection and eye segmentation [2].

2.2 Definitions and Applications

Understanding the definitions and applications of fundamental technologies in display manufacturing and AI is essential for driving innovation. Edge Computing (EC) processes data at the network's edge, significantly reducing latency and bandwidth usage, crucial for environments like the Metaverse [9]. The Internet of Things (IoT) extends this by interconnecting devices that collect and exchange data, often in resource-constrained settings, introducing unique security challenges [13].

AI, particularly through Deep Neural Networks (DNN), automates tasks traditionally requiring human cognition, with AI-driven defect recognition being critical in manufacturing despite limitations like those of fast CMOS sensors [20]. Edge AI optimizes operational processes, enhancing data security and reliability in display manufacturing [11]. Micro-LED technology advances, such as in GaN-based high-brightness green micro-LEDs, improve display performance by addressing brightness and color conversion efficiency challenges [24].

Energy efficiency in AI applications is optimized through datasets like DeepEn2023, providing insights into energy consumption across configurations, crucial for enhancing energy usage in edge AI applications [10]. Wake Vision advances TinyML applications with a dataset of over 6

AI-generated, for reference only.

million images for person detection, ensuring high label accuracy [26]. The convergence of edge computing and deep learning facilitates scalable AI solutions, addressing resource provisioning and task offloading challenges [3]. Partitioning techniques in IoT-Edge-AI environments enhance performance, reduce latency, and improve scalability, meeting modern application demands [12].

The definitions and applications of these technologies underscore their significance in advancing display manufacturing and AI, highlighting the necessity for continued innovation to overcome challenges and enhance system performance. Integrating real-world datasets, such as FPHAB for hand detection and OpenEDS for eye segmentation, supports the development of AI models tailored for specific applications, contributing to the evolution of memory-oriented design space exploration in edge AI [2]. Researchers continue to navigate security risks and resource management complexities across diverse computing paradigms, emphasizing the importance of adaptive strategies in response to performance variability [7].

3 Mura Detection and Correction

Category	Feature	Method
Challenges in Traditional Vision-Based AOI Systems	Analog and Precision Processing	Flex-PE[14], VIP[27]
	Edge and Efficiency	ISEA[28], VDOT[25]
	Robustness Enhancement	APGA[29]
Innovations in Mura Detection	Advanced Circuit Technologies	APDFD[30]

Table 1: This table presents a summary of the challenges associated with traditional vision-based Automated Optical Inspection (AOI) systems and the recent innovations in mura detection technologies. It categorizes the challenges into analog and precision processing, edge and efficiency, and robustness enhancement, while highlighting advanced circuit technologies as a key innovation in mura detection. The methods listed illustrate the ongoing efforts to improve display manufacturing through enhanced detection methodologies.

Category	Feature	Method
Challenges in Traditional Vision-Based AOI Systems	Analog and Precision Processing	Flex-PE[14], VIP[27]
	Edge and Efficiency	ISEA[28], VDOT[25]
	Robustness Enhancement	APGA[29]
Innovations in Mura Detection	Advanced Circuit Technologies	APDFD[30]

Table 2: This table presents a summary of the challenges associated with traditional vision-based Automated Optical Inspection (AOI) systems and the recent innovations in mura detection technologies. It categorizes the challenges into analog and precision processing, edge and efficiency, and robustness enhancement, while highlighting advanced circuit technologies as a key innovation in mura detection. The methods listed illustrate the ongoing efforts to improve display manufacturing through enhanced detection methodologies.

Ensuring high-quality display manufacturing necessitates precise mura detection and correction, a challenge unmet by conventional vision-based Automated Optical Inspection (AOI) systems. These systems struggle with the demands of modern production, prompting the need for innovative detection methodologies. Table 4 provides a detailed comparison of traditional vision-based AOI systems, memristor-controlled circuits, and deep learning algorithms, emphasizing the innovations in mura detection technologies that address the challenges faced by conventional systems. Table 2 provides a detailed summary of the challenges faced by traditional vision-based AOI systems and the innovations in mura detection technologies that address these issues.

3.1 Challenges in Traditional Vision-Based AOI Systems

Traditional AOI systems face significant hurdles in detecting mura, primarily due to inefficiencies in data processing linked to Analog-to-Digital Converters (ADCs), which introduce delays detrimental to the rapid inspection required in high-speed production [30]. Scalability is another issue; AOI systems often need specialized programming to handle diverse product lines, complicating adaptability and inspection criteria [27]. Communication bottlenecks from transmitting high-dimensional features further slow processing and data integration, reducing inspection efficacy [28].

Method Name	Data Processing	Real-Time Processing	Hardware Limitations
APDFD[30]	Adc Delays	Analog Domain Processing	Memristor Characteristics
VIP[27]	Communication Bottlenecks	Real-time Inspections	Precision And Adaptability
ISEA[28]	Communication Bottleneck	Communication Challenges	-
APGA[29]	-	-	-
Flex-PE[14]	Communication Bottlenecks	Large Datasets	Precision Adaptability
VDOT[25]	Data Processing Inefficiencies	Real-time Inefficiencies	Hardware Resource Constraints

Table 3: This table presents a comparative analysis of various methods addressing challenges in traditional vision-based Automated Optical Inspection (AOI) systems. It categorizes the methods based on their approaches to data processing, real-time processing capabilities, and hardware limitations, highlighting specific issues such as ADC delays, communication bottlenecks, and hardware constraints.

As illustrated in Figure 2, these challenges can be categorized into three primary areas: data processing inefficiencies, real-time processing issues, and hardware limitations. Each category highlights specific obstacles such as ADC delays, scalability issues, communication bottlenecks, and the need for advanced algorithms and adaptable solutions to improve mura detection in manufacturing. Table 3 provides a comprehensive overview of the methods employed to tackle these challenges faced by traditional vision-based AOI systems, emphasizing the areas of data processing, real-time processing, and hardware limitations.

Traditional systems also struggle with real-time processing of dynamic edge data, exacerbated by the large, heterogeneous datasets from IoT devices. Limited computational resources and low-latency demands complicate accurate mura detection. Effective edge AI applications, like those using Spiking Neural Networks (SNNs), require multiple time steps, challenging deployment in latency-sensitive contexts. Optimizing strategies such as hybrid quantization and early exit techniques is crucial for balancing inference time and accuracy [31, 32, 33, 34].

Additionally, obtaining dense segmentations and pre-identifying critical regions for image augmentation present significant challenges in mura detection [29]. Hardware limitations, particularly in adapting to varying precision and activation functions, further restrict traditional AOI systems [14]. These challenges underscore the need for advanced algorithms and adaptable solutions to improve mura detection in contemporary manufacturing [25].

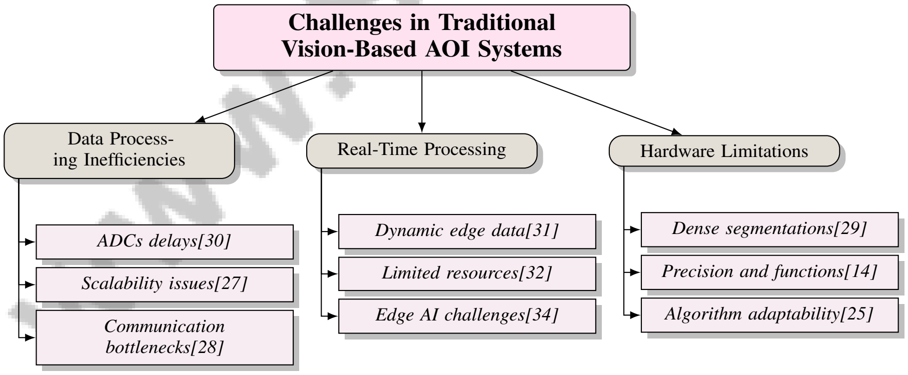

Figure 2: This figure illustrates the primary challenges faced by traditional vision-based Automated Optical Inspection (AOI) systems, categorized into data processing inefficiencies, real-time processing issues, and hardware limitations. Each category highlights specific obstacles such as ADC delays, scalability issues, communication bottlenecks, and the need for advanced algorithms and adaptable solutions to improve mura detection in manufacturing.

3.2 Innovations in Mura Detection

Recent innovations have significantly enhanced the accuracy and efficiency of mura detection in display manufacturing. Memristor-controlled circuits, for example, enable real-time pixel differencing and background subtraction, improving detection speed and precision by dynamically adapting to varying display conditions [30].

Advanced algorithms utilizing deep learning have transformed mura detection. These algorithms, leveraging convolutional neural networks and semi-supervised methods, adeptly classify image patches as defective or non-defective, streamlining analysis of complex defects at the atomic scale. This reduces reliance on human expertise, which can be inconsistent, and ensures robustness even with limited datasets, making them ideal for challenging environments [17, 35, 36].

Edge AI technologies have further revolutionized mura detection by enabling real-time data processing at the network's edge, crucial for handling the vast data from high-resolution displays [37, 38, 39, 19]. These technologies support lightweight models that operate under resource constraints, enhancing inspection system scalability and responsiveness.

The integration of Internet of Things (IoT) devices in display manufacturing has improved operational efficiency by facilitating real-time monitoring and analysis of production lines. This advancement enables immediate defect detection using sophisticated machine vision technologies, such as deep convolutional neural networks, which enhance inspection accuracy and reduce environmental dependency. The use of edge computing minimizes latency and enhances quality control responsiveness by processing data closer to the source [4, 40, 15]. IoT devices enable early detection of potential mura defects, supporting predictive maintenance and enhancing production quality.

In recent years, the field of manufacturing has witnessed significant advancements in defect recognition, driven by the integration of artificial intelligence (AI) and innovative detection methodologies. The challenges associated with traditional defect recognition systems have prompted researchers to explore new approaches that enhance both accuracy and efficiency. As illustrated in Figure 3, this figure presents a hierarchical classification of defect recognition, detailing the multifaceted challenges encountered in the domain. It emphasizes the complexity and interrelated aspects of defect recognition systems, while also showcasing the AI-driven approaches that have emerged to tackle these issues. Furthermore, the figure highlights the innovations in detection techniques and the performance evaluation metrics that are crucial for assessing the effectiveness of these advancements. This comprehensive overview not only underscores the significant progress made in the field but also sets the stage for a deeper analysis of the methodologies that have contributed to overcoming traditional challenges.

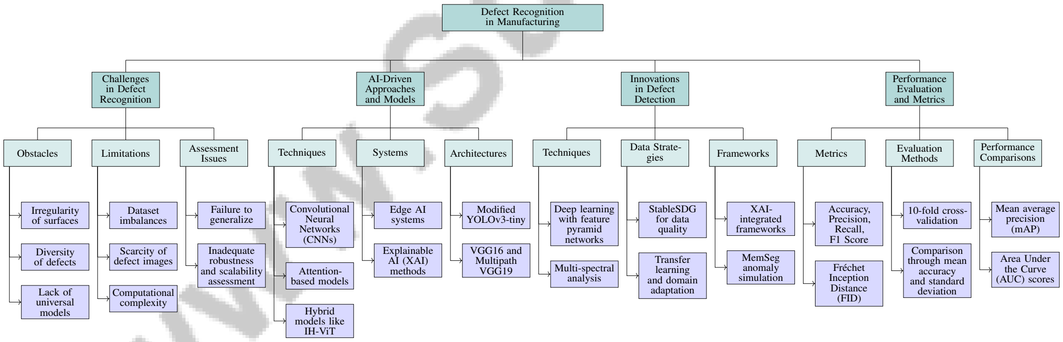
Defect Recognition in Manufacturing
Challenges in Defect Recognition
AI-Driven Approaches and Models
Innovations in Defect Detection
Performance Evaluation and Metrics
Obstacles
Limitations
Assessment Issues
Techniques
Systems
Architectures
Techniques
Data Strategies
Frameworks
Metrics
Evaluation Methods
Performance Comparisons
Irregularity of surfaces
Diversity of defects
Lack of universal models
Dataset imbalances
Scarcity of defect images
Computational complexity
Failure to generalize
Inadequate robustness and scalability assessment
Convolutional Neural Networks (CNNs)
Attention-based models
Hybrid models like IH-ViT
Edge AI systems
Explainable AI (XAI) methods
Modified YOLOv3-tiny
VGG16 and Multipath VGG19
Deep learning with feature pyramid networks
Multi-spectral analysis
StableSDG for data quality
Transfer learning and domain adaptation
XAI-integrated frameworks
MemSeg anomaly simulation
Accuracy, Precision, Recall, F1 Score
Fréchet Inception Distance (FID)
10-fold cross-validation
Comparison through mean accuracy and standard deviation
Mean average precision (mAP)
Area Under the Curve (AUC) scores

Figure 3: This figure illustrates the hierarchical classification of defect recognition in manufacturing, detailing the challenges faced, AI-driven approaches, innovations in detection, and performance evaluation metrics. The structure emphasizes the complexity and interrelated aspects of defect recognition systems, highlighting significant advancements and methodologies in addressing traditional challenges.

Feature	Traditional AOI Systems	Memristor-Controlled Circuits	Deep Learning Algorithms
Detection Technique	Vision-based Aoi	Pixel Differencing	Cnn Classification
Real-time Processing	Limited BY Adcs	Improved Speed	Streamlined Analysis
Hardware Adaptability	Low Adaptability	Dynamic Adaptation	Robust With Constraints

Table 4: Comparison of detection techniques, real-time processing capabilities, and hardware adaptability across traditional vision-based AOI systems, memristor-controlled circuits, and deep learning algorithms. This table highlights the advancements in mura detection technologies, addressing the limitations of conventional systems in display manufacturing.

4 Defect Recognition in Manufacturing

4.1 Challenges in Defect Recognition

Defect recognition in manufacturing faces considerable obstacles due to the irregularity of surfaces and the diversity of defects, which complicate automated detection and classification [41]. The lack of universal models capable of recognizing industrial objects in complex images further exacerbates this issue [42]. High accuracy demands, alongside dataset imbalances, pose significant challenges for developing robust recognition systems [43]. Methodologies often falter in generalizing across varied backgrounds and defect types, leading to suboptimal detection outcomes [44]. The scarcity of defect images for training further limits model accuracy in diverse scenarios [45]. Additionally, the computational complexity of deep learning models and the opaque nature of some explainable AI (XAI) methods restrict their application in real-time [46]. Benchmarks often fail to meet specific XR workload requirements, leading to inflated performance metrics that do not reflect real-world scenarios [2].

Figure 4 illustrates the primary challenges in defect recognition within manufacturing, categorized into surface irregularity, model limitations, and computational complexity. This figure highlights key obstacles such as irregular surfaces, lack of universal models, and the demands of deep learning methodologies. The robustness and scalability of AI/ML methods are often inadequately assessed, emphasizing the need for innovative solutions to enhance defect recognition systems' accuracy, efficiency, and adaptability [3].

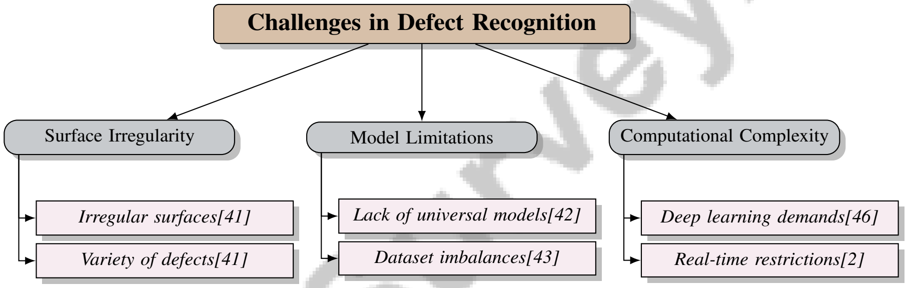

Figure 4: This figure illustrates the primary challenges in defect recognition within manufacturing, categorized into surface irregularity, model limitations, and computational complexity, highlighting key obstacles such as irregular surfaces, lack of universal models, and deep learning demands.

4.2 AI-Driven Approaches and Models

AI methodologies have significantly advanced defect recognition in manufacturing by enhancing detection accuracy and efficiency. Convolutional Neural Networks (CNNs) have been pivotal, with applications like VGG16 and Multipath VGG19 improving recognition accuracy, even with limited data [47, 42]. Techniques such as MemSeg streamline detection by identifying abnormal regions end-to-end, while attention-based models like MA-Q2L enhance localization and recognition [36, 48]. Edge AI systems, using flexible SIMD multi-precision elements like Flex-PE, further support real-time detection [14]. Explainable AI (XAI) methods enhance transparency in visual inspections, making systems more user-friendly [46]. Hybrid models like IH-ViT, combining CNN and Vision Transformer architectures, improve generalization by capturing local and global features, beneficial in low-data scenarios [43]. The modified YOLOv3-tiny architecture demonstrates AI's adaptability across diverse applications, crucial for varying manufacturing environments [49]. These AI-driven advancements continue to transform defect recognition, addressing traditional challenges and paving the way for efficient inspection processes [25].

4.3 Innovations in Defect Detection

Innovations in defect detection have notably enhanced manufacturing capabilities. The integration of deep learning with feature pyramid networks, such as the CNN RetinaNet model, effectively handles various defect scales [50]. The GC10-DET dataset provides a benchmark for evaluating detection models [17]. Multi-spectral analysis alongside deep learning enhances detection by utilizing multiple spectral channels [51]. StableSDG improves training data quality by generating images from perturbed samples [45]. Transfer learning and domain adaptation allow models to perform with minimal labeled data [52]. Edge AI technologies facilitate real-time analysis, enhancing responsiveness in manufacturing [11]. XAI-integrated frameworks improve model performance and interpretability, suitable for industrial applications [46]. MemSeg's anomaly simulation and memory modules boost detection accuracy, while self-attention mechanisms address class imbalance [36, 48]. Flex-PE architecture optimizes resource utilization for real-time detection [14]. The MVGG19 model outperforms traditional models, proving effective for industrial recognition [42]. These innovations advance defect detection by introducing methodologies like EDDN and Multipath VGG19, enhancing accuracy and efficiency in identifying industrial defects. Large-scale datasets and transfer learning strategies enable robust performance, addressing data imbalance and variability, thus improving quality control and operational efficiency [17, 42, 52, 15].

4.4 Performance Evaluation and Metrics

Benchmark	Size	Domain	Task Format	Metric
CP[32]	14,112	Object Detection	Detection	mAP, Latency
DeepEn2023[10]	1,000,000	Energy Efficiency	Energy Consumption Measurement	Energy Consumption, Carbon Emissions
GC10-DET[17]	3,570	Metallic Surface Defect Detection	Defect Classification	Recall, Average Precision
DNN-Edge[53]	7,481	Object Detection	2D And 3D Object Detection	mAP, AP3D
FER-Bench[54]	5,000	Facial Expression Recognition	Classification	Accuracy, Latency
WV[26]	6,000,000	Person Detection	Binary Image Classification	F1-score, Accuracy
EdgeAI-Bench[34]	500	Computer Vision	Image Classification	Inference Latency, Accuracy
XR-AI[2]	112,759	Computer Vision	Image Segmentation	Energy Delay Product

Table 5: This table provides a comprehensive overview of various benchmarks utilized in the evaluation of defect recognition systems and related tasks. It details the size, domain, task format, and metrics for each benchmark, offering insights into the diverse applications and performance metrics employed in contemporary research.

Assessing defect recognition systems' performance is crucial for reliability in manufacturing. Metrics like Accuracy, Precision, Recall, and F1 Score evaluate models such as MemoryMamba, which outperform baseline methods [55]. Fréchet Inception Distance (FID) assesses image quality and model accuracy on augmented datasets, highlighting robust data augmentation's importance [45]. Classification accuracy is compared through mean accuracy and standard deviation [52]. Surface defect recognition achieved 99.1

5 Autofocus Leveling in Imaging Systems

5.1 Advancements in Imaging Systems

Recent advancements in imaging systems have significantly enhanced autofocus leveling, crucial for achieving precise image clarity. Innovations in circuit design have reduced complexity and power consumption, leading to efficient motion detection with minimal digital processing [30]. This optimization is particularly beneficial for real-time applications. Deep learning models have revolutionized defect detection and quality assurance by automating focus adjustments and pattern recognition, which are essential for precise integrated circuit characterization [27, 56]. Automation of these processes ensures consistent focus, thereby improving system reliability.

The integration of advanced algorithms and real-time processing capabilities has enhanced the responsiveness of imaging systems in dynamic environments. By leveraging edge computing and adaptive precision training, these systems optimize resource utilization and performance, making them ideal for applications in the Internet of Things (IoT) and artificial intelligence (AI) domains [4, 57, 7]. This automation significantly boosts imaging performance, especially in high-precision applications.

These advancements underscore the importance of integrating cutting-edge technologies such as deep learning and edge computing to enhance imaging precision and operational efficiency across applications like automated optical inspection and smart surveillance. Such innovations not only

AI-generated, for reference only.

improve image quality but also address challenges related to scalability, adaptability, and real-time processing in constrained devices, promoting broader adoption in industrial and consumer settings [17, 39, 27, 4, 19]. These developments significantly contribute to the field of imaging technology, paving the way for future innovations.

5.2 Infrared Microscopy Automated Location System (IMAL)

The Infrared Microscopy Automated Location System (IMAL) represents a significant advancement in autofocus leveling, particularly for integrated circuit characterization. By combining an autofocus mechanism with sophisticated graph-matching methods, IMAL enables precise location and characterization of structures within integrated circuits [56]. This integration is crucial for accurate alignment and focusing in microscopy, necessary for detailed examination of complex circuit patterns.

IMAL addresses the limitations of traditional autofocus techniques, which often struggle to maintain precision in dynamic environments. Advanced graph-matching algorithms allow IMAL to efficiently analyze complex integrated circuit layouts, improving focus adjustment accuracy and speed. The infrared microscopy approach, paired with a motorized optical system, enables precise detection of conductive tracks beneath silicon layers, overcoming challenges related to data redundancy and noise [58, 56, 18]. This capability is essential for high-resolution imaging and detailed structural analysis, particularly in semiconductor manufacturing.

Additionally, IMAL's automation reduces the need for manual intervention, enhancing efficiency and reproducibility in imaging processes. The system autonomously adjusts focus based on real-time data, ensuring consistent imaging quality and minimizing human error. This automation is particularly advantageous in high-throughput environments like aerospace, automotive, and nuclear industries, where rapid and precise imaging is crucial for maintaining production efficiency and quality standards. Advanced systems such as Automated Defect Recognition (ADR) using Convolutional Neural Networks (CNN) further reduce analysis time and subjective interpretation of defects, achieving high accuracy rates, such as 94.2

The integration of IMAL into imaging systems marks a significant advancement in autofocus leveling. By combining advanced autofocus mechanisms with graph-matching techniques, IMAL enhances imaging precision and efficiency, facilitating automation in microscopy and imaging technologies. This innovation streamlines processes, reduces manual intervention, and increases result reliability in applications such as integrated circuit inspection and defect recognition in crystalline materials [39, 27, 56, 35, 20].

6 Micro-Display Technology

6.1 Advancements in Micro-LED Technology

Recent innovations in micro-LED technology have markedly improved high-resolution display performance, primarily through a streamlined fabrication process using a double-layer thin-film structure. This advancement enhances pixel density and color performance, offering significant benefits over traditional methods [59]. Addressing challenges such as complex manufacturing and high costs is vital for integrating micro-LEDs into mainstream technologies [60]. Efforts to simplify production could unlock new market opportunities and drive further innovation.

High-brightness green micro-LED displays have been developed, achieving up to 250,000 cd/m^2 brightness with a 2.8 V turn-on voltage [24]. These features highlight micro-LEDs' potential for superior brightness and energy efficiency, ideal for next-generation applications. Their ability to deliver high-brightness displays with lower power consumption is advantageous for portable devices needing extended battery life.

Micro-LEDs, composed of micrometer-sized LED devices, offer self-luminous displays that surpass LCD and OLED technologies in luminance, speed, and longevity. As research continues to improve manufacturing and materials, the micro-LED market is expected to grow rapidly, with applications in smartphones, televisions, and wearables. This evolution promises superior visual experiences and aims to facilitate high-performance display commercialization across various industries [15, 24, 61, 59, 60].

6.2 Comparison with Traditional Display Technologies

Micro-LED technology significantly advances over traditional displays, offering superior luminance, which far exceeds conventional LCD and OLED displays, making it suitable for VR, AR, and wearables [24]. Micro-LEDs also provide faster response times, essential for applications requiring rapid image updates and minimal motion blur, alongside better energy efficiency, which extends battery life and promotes sustainability [60].

Micro-LEDs exhibit superior longevity, outlasting both LCD and OLED technologies due to the stability of their materials, which resist degradation over time. Enhanced color stability ensures consistent performance across various conditions [24]. These attributes position micro-LEDs as a transformative force in the display market, with expected applications in smartphones, televisions, and wearables [24, 59, 60, 15].

6.3 Applications and Future Directions

Micro-LED technology shows immense potential across various applications due to its superior brightness, energy efficiency, and compact form factor. It is particularly promising for AR and MR technologies, where high brightness and wide color gamut enhance visual experiences [59]. Micro-LEDs are also being explored for wearables, where low power consumption and high efficiency extend battery life, crucial for consumer acceptance. Their compact size allows integration into small devices like smartphones and wearables, maintaining high display quality [59, 60].

Future research will focus on improving manufacturing processes to enhance yield and reduce costs, facilitating broader adoption [24]. Developing automated systems and novel materials will address production challenges, ensuring reliability and scalability [60]. Protective encapsulation techniques are also being developed to enhance longevity and durability, increasing resilience against environmental factors and extending operational lifespan [24, 59, 60].

These advancements are set to revolutionize micro-display applications, leveraging high brightness, energy efficiency, and rapid response times. As research progresses, improvements in manufacturing, materials, and driving technologies will accelerate micro-LED commercialization, enhancing integration into diverse devices and establishing micro-LEDs as a foundational technology in future display systems, particularly for AR and MR applications [24, 59, 60].

7 Edge AI and Real-Time Data Processing

Edge AI's real-time data processing is fundamentally linked to its supporting frameworks and architectures. As the need for responsive systems escalates, it's crucial to explore methodologies that enhance edge AI deployment, focusing on frameworks that optimize performance in distributed computing environments. Understanding these foundational structures is vital for appreciating their role in advancing edge AI applications.

7.1 Frameworks and Architectures for Edge AI

Deploying edge AI effectively necessitates frameworks that optimize distributed computing performance. The Device-Edge Co-Inference Framework (DECI) exemplifies this by partitioning deep neural networks between on-device and server components to enhance edge inference and balance computational loads [62]. Hierarchical methodologies optimize convolutional neural network models for edge AI accelerators, balancing accuracy, latency, and power consumption, which is essential in resource-constrained settings [63].

MobileInst enhances segmentation and tracking capabilities using a mobile vision transformer and a dual-transformer instance decoder, maintaining low latency [64]. AirComp's task-oriented over-the-air computation aggregates noisy local feature vectors from multiple devices, optimizing for maximum inference accuracy [65]. Research divides edge AI into algorithm-level and system-level designs, emphasizing tailored approaches for edge environments [66].

The HierTrain framework deploys DNN training tasks over Multi-access Edge Computing (MEC) architecture, using hybrid parallelism to enhance scalability and efficiency [67]. In the Internet of Vehicles context, integrating edge caching, computing, and AI is crucial for real-time processing, highlighting the need for Edge Intelligence Systems (EIS) to support immediate data processing and decision-making [68].

Resource-efficient parallel split learning approaches on platforms like Nvidia Jetson Xavier AGX demonstrate enhanced computational efficiency and scalability [69]. The advancement of sophisticated frameworks is crucial for effectively deploying edge AI, enabling vast data processing from interconnected devices while addressing resource constraints and security vulnerabilities [4, 5, 6, 70]. These frameworks optimize resource management and facilitate seamless integration, playing a critical role in edge AI applications' success.

7.2 Challenges in Edge AI Deployment

Deploying AI at the network edge faces significant challenges due to limited computational and storage resources on edge devices. These constraints hinder support for advanced neural networks during inference, especially with complex deep learning models requiring substantial computational power and energy [66]. Fast data processing and privacy concerns complicate deployment, particularly regarding data transmission between devices and servers [57].

Resource allocation and communication latency exacerbate deployment challenges. High latency in data transmission between edge devices and cloud centers slows down training, complicating real-time data processing and decision-making vital for edge AI performance [67]. This issue is critical in scenarios requiring rapid responses, such as vehicular operations, where safety and efficiency depend on real-time processing [68].

The trade-off between on-device computational costs and communication overhead can compromise accuracy or induce excessive latency, challenging edge AI performance [66]. Existing methods often rely on smaller, less accurate models or necessitate high-resource edge devices, creating an unfavorable balance between cost, complexity, and accuracy.

Optimization is complicated by the non-convex nature of problems involving joint transmit precoding and receive beamforming, essential for task-oriented over-the-air computation, limiting traditional methods' applicability in rapid processing scenarios [57]. Determining the optimal combination of deployment operators and tiers to minimize latency while maintaining acceptable accuracy in AI model inference remains unresolved, emphasizing the need for innovative strategies to tackle edge AI implementation challenges [9, 5, 70, 4, 6].

Addressing these challenges requires innovative solutions enhancing edge AI systems' efficiency, security, and scalability. This includes designing energy-efficient computing architectures for edge environments, optimizing data acquisition techniques for real-time processing, and establishing standardized interfaces for deploying AI systems at the network edge, leveraging vast data generated by mobile and IoT devices [4, 6].

7.3 Optimization Techniques for Edge AI

Optimizing AI performance at the edge involves deploying various techniques to address inherent constraints, such as limited computational resources and the need for real-time processing. EdgeSplit's effectiveness lies in its adaptive model partitioning and bandwidth allocation strategies, minimizing training time while leveraging each edge device's capabilities [69]. Complexity-driven algorithms for neural network pruning selectively remove filters contributing minimally to overall network complexity, enhancing efficiency by reducing trainable parameters while maintaining accuracy [71]. Integrating hardware/software co-design frameworks optimized for Bit-line Computing (BC) architectures further enhances performance and energy efficiency for edge AI applications [72].

The MOSAIC framework achieves significant improvements in the throughput-accuracy tradeoff, achieving up to 4.75x higher throughput with negligible accuracy loss compared to traditional methods [73]. Holistic designs that integrate communication, computation, and learning ensure adaptability and efficiency, addressing high computational and memory demands often exceeding mobile device capabilities [74].

Resource allocation techniques are pivotal in optimizing edge AI performance. The ISCC approach determines optimal resource allocation for sensing power, communication time, and quantization bits, maximizing discriminant gain and improving overall efficiency [75]. Adaptive precision training

(APT) dynamically adjusts layer precision during training, optimizing energy consumption and memory usage while maintaining learning effectiveness [57].

Future research should focus on developing efficient edge computing architectures, exploring cooperative vehicular networks, and enhancing the robustness of edge AI frameworks [68]. Exploring new architectures for edge intelligence and enhancing model adaptability while addressing privacy and security concerns remains crucial. The optimization of AI at the edge necessitates a multifaceted approach, including complexity-driven pruning algorithms, holistic system design, efficient resource allocation, and innovative hardware solutions, collectively enhancing edge AI capabilities to meet modern application demands while maintaining efficiency, scalability, and sustainability.

7.4 Image Super-Resolution on Edge Devices

Implementing image super-resolution (SISR) techniques on edge devices marks a significant advancement in edge AI, enabling high-quality image processing in resource-constrained environments. The edge-SR (eSR) method exemplifies this, utilizing one-layer architectures designed for edge devices employing convolutional layers to efficiently upscale images, addressing challenges of limited computational power and storage [19].

CycleGANs effectively transform out-of-distribution images, enhancing recognition capabilities on edge devices by adapting to diverse input data [39]. This adaptability is crucial for maintaining high performance across varying datasets, as demonstrated by the DELTA model, validated against real-world datasets for its effectiveness in image super-resolution tasks [18].

Integrating the ISCC approach into SISR systems improves inference accuracy and resource utilization while remaining adaptable to varying conditions in edge AI environments [75]. The Sequential Concept Drift Detection (SCDD) method allows for real-time monitoring of data distribution changes, ensuring SISR techniques remain effective despite dynamic conditions [76].

Frameworks like Deep-Edge effectively reduce model update times, ensuring latency-sensitive applications meet deadlines [77]. MOSAIC demonstrates that processing input frames from multiple cameras can achieve significant throughput with minimal accuracy loss, highlighting the potential for high-performance SISR on edge devices [73].

Future research should enhance collaborative learning methods, such as federated learning, which aligns with integrating federated learning within edge AI frameworks [78]. The successful implementation of SISR techniques on edge devices is significantly bolstered by advancements in model architecture, optimized communication protocols, and tailored hardware solutions, collectively addressing unique challenges posed by edge computing environments and the need for efficient data processing and analysis [4, 6]. These advancements ensure that edge AI systems deliver high-resolution image outputs efficiently, enhancing application capabilities across diverse domains, from mobile devices to industrial monitoring systems.

8 TPU Optimization for Machine Learning

Optimizing Tensor Processing Units (TPUs) for machine learning involves methodologies that enhance performance and efficiency. This section explores significant advancements in TPU optimization, focusing on the Learned TPU Cost Model (LTCM), which improves execution time prediction. By analyzing LTCM's capabilities and its integration with adaptive methodologies, insights into its role in resource management and system efficiency for TPUs are gained.

8.1 Learned TPU Cost Model (LTCM) and Execution Time Prediction

The Learned TPU Cost Model (LTCM) enhances TPUs' execution times by using advanced machine learning techniques for accurate execution time prediction. This model optimizes resource allocation, minimizing latency and improving overall system efficiency. LTCM represents TPU kernels as undirected graphs, allowing neural networks to estimate execution times effectively, thus ensuring peak performance and energy efficiency [25]. Integrating LTCM with adaptive methodologies like EdgeSplit, which segments models to align with each edge device's computational capabilities, further optimizes TPU operations and resource utilization [69]. Techniques such as the 'nanhu-vdot' method target the acceleration of large language models on TPUs, improving execution time and processing efficiency [25].

Experiments using CNN benchmarks, including LeNet-5, AlexNet, and VGG16 on datasets like CIFAR-10 and CIFAR-100, demonstrate LTCM's robustness under constrained edge AI conditions, highlighting its adaptability across various scenarios [72]. Deployments of models like YOLOv7 on platforms such as the Xilinx ZCU102 FPGA illustrate significant performance and energy efficiency improvements, reinforcing LTCM's impact on edge AI applications [79]. Structured search methodologies in facial expression recognition tasks showcase LTCM's cost-saving benefits, reducing computational expenses while maintaining high accuracy [63].

LTCM and its associated methodologies represent a substantial advancement in TPU optimization, contributing to sustainable and efficient AI deployments at the edge. By addressing challenges in execution efficiency and resource management, LTCM enhances TPU performance, paving the way for more responsive and energy-efficient AI systems. Future research could extend these optimization strategies to include additional activation functions and further reduce precision operations while maintaining accuracy [80].

8.2 Neural Architecture Search and Model Compression Techniques

Neural architecture search (NAS) and model compression techniques, including pruning and quantization, are vital for enhancing TPUs' performance by optimizing machine learning models' accuracy and efficiency, facilitating deployment in resource-constrained environments like mobile devices and edge AI applications. These methods aid in developing cost models that predict execution times and guide configuration tuning, thereby improving inference speed and reducing computational overhead [71, 81, 78, 22, 31].

NAS employs advanced search algorithms to automatically identify optimal neural network architectures, enhancing model performance and efficiency. Utilizing graph embeddings and feedforward neural networks in NAS captures complex relationships within computation graphs, significantly improving prediction accuracy relative to traditional models [22]. This ensures selected architectures are high-performing and well-suited for TPUs' computational capabilities.

Model compression techniques, including pruning, factorization, quantization, and compact model design, are essential for optimizing TPUs [78]. These methods reduce neural networks' size and complexity, enabling faster processing and lower energy consumption. Pruning, in particular, eliminates redundant parameters, decreasing computational overhead without compromising performance [71]. The versatility of pruning modes allows for competitive performance without retraining, saving time and resources.

Quantization enhances model efficiency by lowering model parameters' precision, benefiting TPUs adept at low-precision computations. The quantization and retraining of models like the KWT model demonstrate significant size reductions while maintaining acceptable accuracy, facilitated by custom RISC-V instructions for hardware acceleration [82]. Adaptive precision training (APT) offers further optimization by dynamically adjusting each neural network layer's precision during training. This approach uses the same precision for both forward and backward propagation, optimizing memory usage and energy consumption [57]. By tailoring precision levels to each layer's specific requirements, APT ensures TPUs operate efficiently across diverse workloads.

Future research aims to expand NAS and model compression methodologies to other application areas, optimize configurations for a broader range of output categories, and develop frameworks for automatic extraction of optimal configurations [83]. These advancements will enhance TPUs' scalability and adaptability, ensuring their continued relevance in the evolving machine learning and artificial intelligence landscape.

8.3 On-Device Federated Learning and Edge AI Frameworks

Integrating federated learning within edge AI frameworks signifies a paradigm shift in distributed AI deployment, enhancing operational efficiency and data privacy. Federated learning aggregates locally trained models from multiple edge devices, achieving performance metrics comparable to centralized AI systems while significantly reducing communication overhead and preserving user privacy. This decentralized approach keeps sensitive data on the device, mitigating privacy risks and aligning with regulatory compliance requirements [84].

Edge AI frameworks, like In-Edge AI, optimize real-time AI task execution by intelligently scheduling tasks across edge nodes and mobile devices, enhancing AI applications' flexibility and efficiency. This collaborative approach effectively utilizes computational resources, minimizing latency and improving system responsiveness [67]. Integrating federated learning with these frameworks facilitates seamless data processing and model training, enabling edge devices to contribute to a collective intelligence network while operating autonomously.

Future research in this domain focuses on refining benchmarks to encompass diverse workloads and exploring additional metrics that capture edge AI performance nuances [85]. Expanding these benchmarks will deepen understanding of edge computing capabilities and foster the development of more robust and adaptable AI systems. Additionally, investigating hybrid models that merge classic and deep learning strengths presents a promising avenue for enhancing TPU optimization in edge AI applications [19].

Exploring wireless techniques and optimizing TPU performance within edge AI frameworks represent additional growth areas, enhancing AI deployments' scalability and efficiency [86]. By concentrating on these areas, researchers can devise innovative solutions addressing resource constraints and dynamic environments inherent in edge computing [69]. Future inquiries will also examine integrating model compression techniques and supporting multi-device scenarios to further enhance edge AI frameworks' efficiency and applicability [87].

The convergence of federated learning and edge AI frameworks holds significant promise for advancing AI systems' capabilities, enabling more efficient, secure, and scalable deployments across diverse application scenarios. This integration enhances edge devices' performance while contributing to a more intelligent and responsive computing ecosystem. Future research could optimize collaborative frameworks between edge and cloud learning, refine communication protocols for high-mobility scenarios, and investigate innovative methods for leveraging noise beneficially in learning processes [88].

9 Conclusion

This survey delves into the integration of advanced technologies within display manufacturing and artificial intelligence, highlighting pivotal areas such as mura detection, defect recognition, autofocus leveling, micro-display technology, edge AI, and TPU optimization. These innovations collectively enhance efficiency, accuracy, and scalability across various sectors. The deployment of edge AI markedly improves real-time data processing, offering superior speed and efficiency, which is critical in dynamic settings like the Metaverse. Micro-LED technology emerges as a standout in display manufacturing, known for its superior brightness and energy efficiency, especially in augmented reality and wearable devices, although challenges like high production costs and complex manufacturing processes persist.

The survey underscores the imperative of developing scalable and robust edge infrastructures to support the proliferation of edge AI applications. Future research should focus on optimizing resource allocation, enhancing cooperative perception among vehicles, and addressing privacy and security issues in multi-access edge computing environments. The integration of federated learning within edge AI frameworks marks a significant advancement, enhancing model quality and scalability in decentralized learning environments.

Exploring hybrid architectures, such as those combining spiking neural networks and artificial neural networks, offers potential for reducing energy consumption while maintaining accuracy, contributing to more sustainable AI deployments. Partitioning is identified as a critical strategy for optimizing resource utilization, ensuring timely processing in IoT-Edge-AI applications. Future research directions include developing green, ethical, and trusted AI systems, exploring quantum AI, and fostering collaboration between software and hardware design to meet the evolving demands of edge AI applications.

The survey highlights the transformative impact of computing technologies across diverse sectors and the necessity for ongoing research to address emerging challenges. It concludes that integrating MRAM into the memory hierarchy of XR applications can achieve substantial energy savings and

AI-generated, for reference only.

area reductions, presenting a promising approach for future hardware design. The pivotal role of edge AI in enhancing real-time data processing capabilities is emphasized, alongside addressing existing challenges and proposing future research directions. As these technologies evolve, the focus will likely shift towards developing energy-efficient solutions and scalable architectures to meet the increasing demands of contemporary applications, thereby improving the performance and reliability of display manufacturing and AI technologies while fostering innovative applications in emerging domains.

AI-generated, for reference only.

References

[1] Kinshuk Sengupta and Praveen Ranjan Srivastava. Hrnet: Ai on edge for mask detection and social distancing, 2022.

[2] Vivek Parmar, Syed Shakib Sarwar, Ziyun Li, Hsien-Hsin S. Lee, Barbara De Salvo, and Manan Suri. Memory-oriented design-space exploration of edge-ai hardware for xr applications, 2023.

[3] Ai-based fog and edge computing: A systematic review taxonomy and future directions.

[4] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang. Edge intelligence: Paving the last mile of artificial intelligence with edge computing, 2019.

[5] Shuiguang Deng, Hailiang Zhao, Weijia Fang, Jianwei Yin, Schahram Dustdar, and Albert Y Zomaya. Edge intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet of Things Journal*, 7(8):7457–7469, 2020.

[6] Dianlei Xu, Tong Li, Yong Li, Xiang Su, Sasu Tarkoma, Tao Jiang, Jon Crowcroft, and Pan Hui. Edge intelligence: Empowering intelligence to the edge of network. *Proceedings of the IEEE*, 109(11):1778–1837, 2021.

[7] Muhammed Golec and Sukhpal Singh Gill. Computing: Looking back and moving forward, 2024.

[8] Luyi Chang, Zhe Zhang, Pei Li, Shan Xi, Wei Guo, Yukang Shen, Zehui Xiong, Jiawen Kang, Dusit Niyato, Xiuquan Qiao, and Yi Wu. 6g-enabled edge ai for metaverse: Challenges, methods, and future research directions, 2022.

[9] Xiaofei Wang, Yiwen Han, Victor CM Leung, Dusit Niyato, Xueqiang Yan, and Xu Chen. Convergence of edge computing and deep learning: A comprehensive survey. *IEEE communications surveys & tutorials*, 22(2):869–904, 2020.

[10] Xiaolong Tu, Anik Mallik, Haoxin Wang, and Jiang Xie. Deepen2023: Energy datasets for edge artificial intelligence, 2023.

[11] Zimu Zheng. Kubeedge-sedna v0.3: Towards next-generation automatically customized ai engineering scheme, 2023.

[12] Guoxing Yao and Lav Gupta. A survey on the use of partitioning in iot-edge-ai applications, 2024.

[13] Cheng Wang, Zenghui Yuan, Pan Zhou, Zichuan Xu, Ruixuan Li, and Dapeng Oliver Wu. The security and privacy of mobile-edge computing: An artificial intelligence perspective. *IEEE Internet of Things Journal*, 10(24):22008–22032, 2023.

[14] Mukul Lokhande, Gopal Raut, and Santosh Kumar Vishvakarma. Flex-pe: Flexible and simd multi-precision processing element for ai workloads, 2024.

[15] Yuanyuan Ding, Junchi Yan, Guoqiang Hu, and Jun Zhu. Cognitive visual inspection service for lcd manufacturing industry, 2021.

[16] Yuyi Mao, Xianghao Yu, Kaibin Huang, Ying-Jun Angela Zhang, and Jun Zhang. Green edge ai: A contemporary survey, 2024.

[17] Xiaoming Lv, Fajie Duan, Jia-jia Jiang, Xiao Fu, and Lin Gan. Deep metallic surface defect detection: The new benchmark and detection network. *Sensors*, 20(6):1562, 2020.

[18] Zhenyu Wang and Shahriar Nirjon. Characterizing disparity between edge models and high-accuracy base models for vision tasks, 2024.

[19] Pablo Navarrete Michelini, Yunhua Lu, and Xingqun Jiang. edge-sr: Super-resolution for the masses, 2021.

[20] Mathis Hoffmann, Thomas Köhler, Bernd Doll, Frank Schebesch, Florian Talkenberg, Ian Marius Peters, Christoph J. Brabec, Andreas Maier, and Vincent Christlein. Joint super-resolution and rectification for solar cell inspection, 2021.

[21] Humberto Carvalho, Pavel Zaykov, and Asim Ukaye. Leveraging the hw/sw optimizations and ecosystems that drive the ai revolution, 2022.

[22] Samuel Kaufman, Phitchaya Mangpo Phothilimthana, and Mike Burrows. Learned tpu cost model for xla tensor programs. In *Proc. Workshop ML Syst. NeurIPS*, pages 1–6, 2019.

[23] Nonlinear optical joint transform correlator for low latency convolution operations.

[24] ZHANG Jie, WANG Guanghua, DENG Feng, YANG Wenyun, GAO Sibo, LU Chaoyu, MENG Zeyang, GAO Shuxiong, CHANG Cheng, CAO Kunyu, et al. Fabrication of gan-based micro-led green micro-display with high brightness. *Infrared Technology*, 46(10):1186–1191, 2024.

[25] Xu-Hao Chen, Si-Peng Hu, Hong-Chao Liu, Bo-Ran Liu, Dan Tang, and Di Zhao. Research on llm acceleration using the high-performance risc-v processor "xiangshan" (nanhu version) based on the open-source matrix instruction set extension (vector dot product), 2024.

[26] Colby Banbury, Emil Njor, Andrea Mattia Garavagno, Matthew Stewart, Pete Warden, Manjunath Kudlur, Nat Jeffries, Xenofon Fafoutis, and Vijay Janapa Reddi. Wake vision: A tailored dataset and benchmark suite for tinyml computer vision applications, 2024.

[27] Faraz Waseem, Sanjit Menon, Haotian Xu, and Debashis Mondal. Vizinspect pro – automated optical inspection (aoi) solution, 2022.

[28] Xu Chen, Khaled B. Letaief, and Kaibin Huang. On the view-and-channel aggregation gain in integrated sensing and edge ai, 2024.

[29] Kaiyang Cheng, Claudia Iriondo, Francesco Calivá, Justin Krogue, Sharmila Majumdar, and Valentina Pedoia. Adversarial policy gradient for deep learning image augmentation, 2019.

[30] Olga Krestinskaya and Alex Pappachen James. Real-time analog pixel-to-pixel dynamic frame differencing with memristive sensing circuits, 2018.

[31] Rakshith Jayanth, Neelesh Gupta, and Viktor Prasanna. Benchmarking edge ai platforms for high-performance ml inference, 2024.

[32] Chinthaka Gamanayake, Lahiru Jayasinghe, Benny Ng, and Chau Yuen. Cluster pruning: An efficient filter pruning method for edge ai vision applications, 2020.

[33] Nemin Qiu and Chuang Zhu. Low latency of object detection for spikng neural network, 2023.

[34] Jaskirat Singh, Bram Adams, and Ahmed E. Hassan. On the impact of black-box deployment strategies for edge ai on latency and model performance, 2024.

[35] Nik Dennler, Antonio Foncubierta-Rodriguez, Titus Neupert, and Marilyne Sousa. Learning-based defect recognition for quasi-periodic microscope images, 2020.

[36] Minghui Yang, Peng Wu, and Hui Feng. Memseg: A semi-supervised method for image surface defect detection using differences and commonalities. *Engineering Applications of Artificial Intelligence*, 119:105835, 2023.

[37] Binxiao Huang, Jason Chun Lok Li, Jie Ran, Boyu Li, Jiajun Zhou, Dahai Yu, and Ngai Wong. Hundred-kilobyte lookup tables for efficient single-image super-resolution, 2024.

[38] Sudhakar Sah, Ravish Kumar, Honnesh Rohmetra, and Ehsan Saboori. Token pruning using a lightweight background aware vision transformer, 2024.

[39] August Lidfelt, Daniel Isaksson, Ludwig Hedlund, Simon Åberg, Markus Borg, and Erik Larsson. Enabling image recognition on constrained devices using neural network pruning and a cyclegan, 2020.

[40] Hamza Ali Imran, Usama Mujahid, Saad Wazir, Usama Latif, and Kiran Mehmood. Embedded development boards for edge-ai: A comprehensive report, 2020.

[41] Marc Wenskat. First attempts in automated defect recognition in superconducting radio-frequency cavities, 2019.

[42] Ioannis D. Apostolopoulos and Mpesiana Tzani. Industrial object, machine part and defect recognition towards fully automated industrial monitoring employing deep learning. the case of multilevel vgg19, 2020.

[43] Xiaoibin Wang, Shuang Gao, Yuntao Zou, Jianlan Guo, and Chu Wang. Ih-vit: Vision transformer-based integrated circuit appear-ance defect detection, 2023.

[44] Weixi Wang, Xichen Zhong, Xin Li, Sizhe Li, and Xun Ma. Overhead line defect recognition based on unsupervised semantic segmentation, 2023.

[45] Yichun Tai, Kun Yang, Tao Peng, Zhenzhen Huang, and Zhijiang Zhang. Defect image sample generation with diffusion prior for steel surface defect recognition, 2024.

[46] Truong Thanh Hung Nguyen, Phuc Truong Loc Nguyen, and Hung Cao. Xedgeai: A human-centered industrial inspection framework with data-centric explainable edge ai approach, 2024.

[47] Jingwen Fu, Xiaoyan Zhu, and Yingbin Li. Recognition of surface defects on steel sheet using transfer learning, 2019.

[48] Xin Zuo, Yu Sheng, Jifeng Shen, and Yongwei Shan. Multi-label sewer pipe defect recognition with mask attention feature enhancement and label correlation learning, 2024.

[49] Vittorio Mazzia, Francesco Salvetti, Aleem Khaliq, and Marcello Chiaberge. Real-time apple detection system using embedded systems with hardware accelerators: An edge ai application, 2020.

[50] Alberto García-Pérez, María José Gómez-Silva, and Arturo de la Escalera. Automated defect recognition of castings defects using neural networks, 2022.

[51] Haiyong Chen, Yue Pang, Qidi Hu, and Kun Liu. Solar cell surface defect inspection based on multispectral convolutional neural network, 2018.

[52] Hongyu Li, Peng Jiang, and Tiejun Wang. Deep learning based intelligent coin-tap test for defect recognition, 2022.

[53] Lukas Stäcker, Juncong Fei, Philipp Heidenreich, Frank Bonarens, Jason Rambach, Didier Stricker, and Christoph Stiller. Deployment of deep neural networks for object detection on edge ai devices with runtime optimization, 2021.

[54] Heath Smith, James Seekings, Mohammadreza Mohammadi, and Ramtin Zand. Realtime facial expression recognition: Neuromorphic hardware vs. edge ai accelerators, 2024.

[55] Qianning Wang, He Hu, and Yucheng Zhou. Memorymamba: Memory-augmented state space model for defect recognition, 2024.

[56] Raphaël Abelé, Jean-Luc Damoiseaux, Redouane El Moubtahij, Jean-Marc Boi, Daniele Fronte, Pierre-Yvan Liardet, and Djamal Merad. Spatial location in integrated circuits through infrared microscopy. *Sensors*, 21(6):2175, 2021.

[57] Tian Huang, Tao Luo, and Joey Tianyi Zhou. Adaptive precision training for resource constrained devices, 2020.

[58] Aaron Yi Ding, Ella Peltonen, Tobias Meuser, Atakan Aral, Christian Becker, Schahram Dustdar, Thomas Hiessl, Dieter Kranzlmuller, Madhusanka Liyanage, Setareh Magshudi, Nitinder Mohan, Joerg Ott, Jan S. Rellermeyer, Stefan Schulte, Henning Schulzrinne, Gurkan Solmaz, Sasu Tarkoma, Blesson Varghese, and Lars Wolf. Roadmap for edge ai: A dagstuhl perspective, 2021.

[59] Longheng Qi, Peian Li, Xu Zhang, Ka Ming Wong, and Kei May Lau. Monolithic full-color active-matrix micro-led micro-display using ingan/algainp heterogeneous integration. *Light: Science & Applications*, 12(1):258, 2023.

[60] Moojin Kim. Introduction and research trends on micro led technology. *Advanced Industrial SCIence*, 3(3):14–19, 2024.

[61] Stanislava Soro. Tinyml for ubiquitous edge ai, 2021.

[62] Jiawei Shao and Jun Zhang. Communication-computation trade-off in resource-constrained edge inference, 2020.

[63] Mohammadreza Mohammadi, Heath Smith, Lareb Khan, and Ramtin Zand. Facial expression recognition at the edge: Cpu vs gpu vs vpu vs tpu, 2023.

[64] Renhong Zhang, Tianheng Cheng, Shusheng Yang, Haoyi Jiang, Shuai Zhang, Jiancheng Lyu, Xin Li, Xiaowen Ying, Dashan Gao, Wenyu Liu, and Xinggang Wang. Mobileinst: Video instance segmentation on the mobile, 2023.

[65] Dingzhu Wen, Xiang Jiao, Peixi Liu, Guangxu Zhu, Yuanming Shi, and Kaibin Huang. Task-oriented over-the-air computation for multi-device edge ai, 2022.

[66] Yuanming Shi, Kai Yang, Tao Jiang, Jun Zhang, and Khaled B. Letaief. Communication-efficient edge ai: Algorithms and systems, 2020.

[67] Deyin Liu, Xu Chen, Zhi Zhou, and Qing Ling. Hiertrain: Fast hierarchical edge ai learning with hybrid parallelism in mobile-edge-cloud computing, 2020.

[68] Jun Zhang and Khaled B Letaief. Mobile edge intelligence and computing for the internet of vehicles. *Proceedings of the IEEE*, 108(2):246–261, 2019.

[69] Mingjin Zhang, Jiannong Cao, Yuvraj Sahni, Xiangchun Chen, and Shan Jiang. Resource-efficient parallel split learning in heterogeneous edge computing, 2024.

[70] Sukhpal Singh Gill, Muhammed Golec, Jianmin Hu, Minxian Xu, Junhui Du, Huaming Wu, Guneet Kaur Walia, Subramaniam Subramanian Murugesan, Babar Ali, Mohit Kumar, Kejiang Ye, Prabal Verma, Surendra Kumar, Felix Cuadrado, and Steve Uhlig. Edge ai: A taxonomy, systematic review and future directions, 2024.

[71] Muhammad Zawish, Steven Davy, and Lizy Abraham. Complexity-driven cnn compression for resource-constrained edge ai, 2022.

[72] Marco Rios, Flavio Ponzina, Alexandre Levisse, Giovanni Ansaloni, and David Atienza. Bit-line computing for cnn accelerators co-design in edge ai inference, 2022.

[73] Ila Gokarn, Hemanth Sabella, Yigong Hu, Tarek Abdelzaher, and Archan Misra. Mosaic: Spatially-multiplexed edge ai optimization over multiple concurrent video sensing streams, 2023.

[74] Khaled B. Letaief, Yuanming Shi, Jianmin Lu, and Jianhua Lu. Edge artificial intelligence for 6g: Vision, enabling technologies, and applications, 2021.

[75] Dingzhu Wen, Peixi Liu, Guangxu Zhu, Yuanming Shi, Jie Xu, Yonina C. Eldar, and Shuguang Cui. Task-oriented sensing, computation, and communication integration for multi-device edge ai, 2022.

[76] Takeya Yamada and Hiroki Matsutani. A sequential concept drift detection method for on-device learning on low-end edge devices, 2023.

[77] Anirban Bhattacharjee, Ajay Dev Chhokra, Hongyang Sun, Shashank Shekhar, Aniruddha Gokhale, Gabor Karsai, and Abhishek Dubey. Deep-edge: An efficient framework for deep learning model update on heterogeneous edge, 2020.

[78] Han Cai, Ji Lin, Yujun Lin, Zhijian Liu, Haotian Tang, Hanrui Wang, Ligeng Zhu, and Song Han. Enable deep learning on mobile devices: Methods, systems, and applications, 2022.

[79] Federico Nicolas Peccia, Svetlana Pavlitska, Tobias Fleck, and Oliver Bringmann. Efficient edge ai: Deploying convolutional neural networks on fpga with the gemmini accelerator, 2024.

[80] Zishen Wan, Ashwin Sanjay Lele, and Arijit Raychowdhury. Circuit and system technologies for energy-efficient edge robotics, 2022.

AI-generated, for reference only.

[81] Yiwei Zhao, Ziyun Li, Win-San Khwa, Xiaoyu Sun, Sai Qian Zhang, Syed Shakib Sarwar, Kleber Hugo Stangherlin, Yi-Lun Lu, Jorge Tomas Gomez, Jae-Sun Seo, Phillip B. Gibbons, Barbara De Salvo, and Chiao Liu. Neural architecture search of hybrid models for npu-cim heterogeneous ar/vr devices, 2024.

[82] Aness Al-Qawlaq, Ajay Kumar M, and Deepu John. Kwt-tiny: Risc-v accelerated, embedded keyword spotting transformer, 2024.

[83] Zichao Shen, Neil Howard, and Jose Nunez-Yanez. Big-little adaptive neural networks on low-power near-subthreshold processors, 2023.

[84] Muhammad Shafique, Alberto Marchisio, Rachmad Vidya Wicaksana Putra, and Muhammad Abdullah Hanif. Towards energy-efficient and secure edge ai: A cross-layer framework, 2021.

[85] Wenbin Li, Hakim Hacid, Ebtesam Almazrouei, and Merouane Debbah. A comprehensive review and a taxonomy of edge machine learning: Requirements, paradigms, and techniques, 2023.

[86] Weier Wan, Rajkumar Kubendran, Clemens Schaefer, S. Burc Eryilmaz, Wenqiang Zhang, Dabin Wu, Stephen Deiss, Priyanka Raina, He Qian, Bin Gao, Siddharth Joshi, Huaqiang Wu, H. S. Philip Wong, and Gert Cauwenberghs. Edge ai without compromise: Efficient, versatile and accurate neurocomputing in resistive random-access memory, 2021.

[87] Shubha R. Kharel, Prashansa Mukim, Piotr Maj, Grzegorz W. Deptuch, Shinjae Yoo, Yihui Ren, and Soumyajit Mandal. Automated and holistic co-design of neural networks and asics for enabling in-pixel intelligence, 2024.

[88] Dwith Chenna. Evolution of convolutional neural network (cnn): Compute vs memory bandwidth for edge ai, 2023.

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.