Review

Low-Power Ultra-Small Edge AI Accelerators for Image Recognition with Convolution Neural Networks: Analysis and Future Directions

Weison Lin *, Adewale Adetomi and Tughrul Arslan

Institute for Integrated Micro and Nano Systems, University of Edinburgh, Edinburgh EH9 3FF, UK; adewale.adetomi@ed.ac.uk (A.A.); tughrul.arslan@ed.ac.uk (T.A.)

* Correspondence: Weison.Lin@ed.ac.uk

Abstract: Edge AI accelerators have been emerging as a solution for near customers' applications in areas such as unmanned aerial vehicles (UAVs), image recognition sensors, wearable devices, robotics, and remote sensing satellites. These applications require meeting performance targets and resilience constraints due to the limited device area and hostile environments for operation. Numerous research articles have proposed the edge AI accelerator for satisfying the applications, but not all include full specifications. Most of them tend to compare the architecture with other existing CPUs, GPUs, or other reference research, which implies that the performance exposé of the articles are not comprehensive. Thus, this work lists the essential specifications of prior art edge AI accelerators and the CGRA accelerators during the past few years to define and evaluate the low power ultra-small edge AI accelerators. The actual performance, implementation, and productized examples of edge AI accelerators are released in this paper. We introduce the evaluation results showing the edge AI accelerator design trend about key performance metrics to guide designers. Last but not least, we give out the prospect of developing edge AI's existing and future directions and trends, which will involve other technologies for future challenging constraints.

Keywords: edge AI accelerator; CGRA; CNN

Citation: Lin, W.; Adetomi, A.; Arslan, T. Low-Power Ultra-Small Edge AI Accelerators for Image Recognition with Convolution Neural Networks: Analysis and Future Directions. *Electronics* **2021**, *10*, 2048. <https://doi.org/10.3390/electronics10172048>

Academic Editor: Fabian Khateb

Received: 30 June 2021

Accepted: 19 August 2021

Published: 25 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Convolution neural network (CNN), widely applied to image recognition, is a machine learning algorithm. CNN is usually adopted by software programs supported by the artificial intelligence (AI) framework, such as TensorFlow and Caffe. These programs are usually run by central processing units (CPUs) or graphics processing units (GPUs) to form the AI systems which construct the image recognition models. The models trained by massive data such as big data and infer the result by the given data have been commonly seen running on cloud-based systems.

Hardware platforms for running AI technology can be sorted into the following hierarchies: data center bound system, edge-cloud coordination system, and 'edge' AI devices. The three hierarchies of hardware platforms from the data center to edge devices require different hardware resources and are exploited by various applications according to their demands. The state-of-the-art applications for image recognition such as unmanned aerial vehicles (UAVs), image recognition sensors, wearable devices, robotics, remote sensing satellites belong to the third hierarchy and are called edge devices. Edge devices refer to the devices connecting to the internet but near the consumers or at the edge of the whole Internet of things (IoT) system. They are also called edge AI devices when they utilize AI algorithms. The targeted AI algorithm of the accelerators in this paper is CNN.

The mentioned applications contain several features, which demand a dedicated system to cover. Edge AI holds several advantages and is qualified to deal with them. The specific capabilities of edge AI technology are:

- Edge AI can improve the user experience when AI technology and data processing are near customers more and more.
- Edge AI can reduce the data transition latency, which implies real-time processing ability.
- Edge AI can run under no internet coverage to offer privacy through local processing.
- Edge AI pursues the compact size and manages power consumption to meet the mobility and limited power source.

Some edge AI systems are neither power-sensitive nor size-limited such as surveillance systems for face recognition and unmanned shops. Designed as an immobile system is a specific feature of these edge AI systems. Although these kinds of applications do not care about power consumption and size, they tend to be aware of data privacy more. As a result, they also avoid using the first and second hierarchy platforms. However, the scope of these power non-sensitive edge AI systems is not a target in this paper. This paper focuses on surveying AI accelerators designed for power-sensitive and size-limited edge AI devices, which are based on batteries or limited power sources such as systems using solar panels. The following mentioned edge AI refers to this kind of system. Typically, this kind of system is portable and comes to mind when people mention edge AI devices because GPU and CPU-based systems can easily substitute the non-power-sensitive and non-size-limited edge AI systems since they do not care about the power consumption and size. As an edge AI accelerator requires mobility, power consumption and area size are the most caring features. Greedily, when the two features meet the requirement, the computation ability will be expected to be as high as possible. High computation ability helps to solve the critical feature of this kind of edge AI device. The critical feature is the real-time computing ability for predicting or inferring the subsequent decision by pre-trained data.

CPUs and GPUs have been used extensively in the first two hierarchies of AI hardware platforms for practicing CNN algorithms. Due to the inflexibility of CPUs and the high-power consumption of GPUs, they are not suitable for power-sensitive edge AI devices. As a result, power-sensitive edge AI devices require a new customized and flexible AI hardware platform to implement arbitrary CNN algorithms for real-time computing with low power consumption.

Furthermore, as the edge devices develop into various applications such as monitoring natural hazards by UAV, detecting radiation leakage for nuclear disaster by robotics, and remote sensing in space by satellites, these applied fields are more critical than usual. These critical environments, such as radiation fields, can cause a system failure. As a result, not only power consumption and area size are the key but also the fault tolerance of edge AI devices for satisfying their compact and mobile feature with reliability. There are various research articles have been proposed targeting fault tolerance. Reference [1] introduces a clipped activation technique to block the potentially faulty activations and maps them to zero on a CPU and two GPUs. Reference [2] focuses on systolic array fault mitigation, which utilizes fault-aware pruning with/without retraining technique. The retraining feature takes 12 min to finish the retraining at least, and the worst case is 1 h for AlexNet. It is not suitable for edge AI. For permanent faults, Reference [3] proposes a fault-aware mapping technique to minus the permanent fault in MAC units. For power-efficient technology, Reference [4] proposes a computation re-use-aware neural network technique to reuse the weight by constructing a computational reuse table. Reference [5] uses an approximate computing technique and retrains the network by getting the resilient neurons. It also shows that dynamic reconfiguration is the crucial feature for the flexibility to arranging the processing engines. These articles focus on fault tolerance technology specifically. Some of them address the relationship between accuracy and power-efficient together but lack computation ability information [4,5]. Besides these listed articles, there are still many published works targeting fault tolerance in recent years, which indicates that the edge AI with fault tolerance is the trend.

In summary, edge AI devices' significant issues are power sensitivity, device size limitation, limited local-processing ability, and fault tolerance. The limitation of power

sensitivity, device size, and local-processing ability are bound by the native constraints of the edge AI devices. The constraints are generated by the limited power source such as battery and the portable feature, which causes the size of an edge AI system limited. To address these issues, examining the three key features of the prior art of the accelerators tailored for edge AI devices is necessary for providing future design directions.

From the point of view of a completed edge AI accelerator, the released specifications of the above fault tolerance articles are not comprehensive because they focus on the fault tolerance feature more than a whole edge AI accelerator. Furthermore, most related edge AI accelerator survey works tend to focus on an individual structure's specific topics and features without comparing the three features. Although each article of the individual structure is interesting to read and learn about, without direct comparison in the same standard units, it is unclear to know how good and what the generation of each structure is. This situation also makes the designers hard to compare the structure of each edge AI accelerator and determine which design is more suitable to reference. As a result, this paper focuses on evaluating the prior arts of edge AI accelerator on the three key features.

The prior arts focused on in this work are not limited to the released prior art accelerators but also edge accelerators architecture based on coarse-grained reconfigurable array (CGRA) technology because to achieve the flexibility of the hardware structure and deal with its compact size, one of the solutions for edge AI platforms is dynamic reconfiguration. The reconfigurable function realizes different types of CNN algorithms such that they can be loaded into an AI platform depending on the required edge computing. Moreover, the reconfigurable function also potentially provides fault tolerance to the system by reconfiguring the connections between processing elements (PEs). Overall, this survey will benefit those looking up the low-power ultra-small edge AI accelerators' specifications and setting up their designs. This paper helps designers choose or design a suitable architecture by indicating reasonable parameters for their low-power ultra-small edge AI accelerator. The rest of this paper is organized as follows: Section 2 introduces the hardware types adopted by AI applications. Section 3 introduces the edge AI accelerators, including prior edge AI accelerators, CGRAs accelerators, the units used in this paper for evaluating their three key features, and the suitable technologies for implementing the accelerators. Section 4 releases the analysis result and indicates the future direction. Conclusion and future works are summarized in Section 5.

2. System Platform for AI Algorithms

For achieving the performance of AI algorithms, several design trends of a complete platform for AI systems, such as cloud training and inference, edge-cloud coordination, near-memory computing, and in-memory computing, have been proposed [6]. Currently, AI algorithms rely on the cloud or edge-cloud coordinating platforms, such as Nvidia's GPU-based chipsets, Xilinx's Versal platform, MediaTek's NeuroPilot platform, and Apple's A13 CPU [7]. The advantages and disadvantages of CPU and GPU for applying on edge devices are shown in Table 1 [8,9]. As shown in Table 1, CPU and GPU are more suitable for data-center-bound platforms due to CPU's sequential processing feature and GPU's power high power consumption. They do not meet the demand of low-power edge devices, which are strictly power-limited and size sensitive [10].

Edge-cloud coordination systems belong to the second hierarchy, which cannot run in areas with no network coverage. Data transfer through the network has significant latency, not acceptable for real-time AI applications such as security and emergency response [9]. Privacy is another concern when personal data is transferred through the internet. Low-power edge AI devices require hardware to support high-performance AI computation with minimal power consumption in real-time. As a result, designing a reconfigurable AI hardware platform that allows the adoption of arbitrary CNN algorithms for low-power edge AI devices with no internet coverage is the trend.

Table 1. Pros and cons of CPU, GPU, and Edge AI accelerator.

Pros and Cons	Processor		
	CPU	GPU	Edge AI accelerator
Advantage	<ul style="list-style-type: none"> Easily implement any algorithms. 	<ul style="list-style-type: none"> Can process high throughput video data. High memory bandwidth. High parallel processing ability [11–14]. 	<ul style="list-style-type: none"> Power and computation efficient. Compact size. Customizable design for the specific application.
Disadvantage	<ul style="list-style-type: none"> The sequential processing feature does not match the characteristic of CNN, requiring massively parallel computing. 	<ul style="list-style-type: none"> requires massive power support. Restricts its application for power-sensitive edge devices. Images in a streaming video and some tracking algorithms are inputted sequentially but not parallel [15]. 	<ul style="list-style-type: none"> Customizable for the specific targeted application (inflexible for all types of computations). Computational power is limited compared to the data center CPU and GPU.
Application platform	<ul style="list-style-type: none"> More suitable for a data center. Cooperate with AI accelerator. 	<ul style="list-style-type: none"> More suitable for a data center. Cooperate with AI accelerator. 	<ul style="list-style-type: none"> Customized for specific edge devices. Can cooperate with CPU or GPU.

3. Edge AI Accelerators

Several architectures and methods have been proposed to achieve compact size, low power consumption, and computation ability of edge devices. The following Sections 3.2 and 3.3, introduce the released prior art edge AI accelerators and state-of-the-art edge accelerators based on CGRA, potentially suitable for low-power edge AI devices.

Most of the proposed review articles introduce accelerators feature-by-feature, and some miss mentioning the three key features. These articles tend to report the existing works only but not compare them. On the other hand, another edge AI accelerator articles contain all three key features. Still, the result they release is hard to understand because they only release comparison results with reference accelerators. It turns out that the used units in these articles for comparison of the architecture area, power consumption, and computation ability are not expected by the edge AI designers' community. Instead of using millimeters squared (mm^2), watt (W), operations per second (OPs), the results show how many 'times' better their reference works. As a result, we decide to compare the edge AI accelerators and CGRA architectures by the units used by most AI accelerator designers.

3.1. Specification Normalization and Evaluation

The presented unit of the computation ability in the following section is OPs (operation per second); MOPs, GOPs, and TOPs represent Mega, Giga, and Tera OPs, respectively. The arithmetic of each accelerator varies in data representation, e.g., floating-point (FP) and fixed-point (fixed). To compare the accelerators, using different arithmetic will lose impartiality. As a result, converting the units will be the following task.

The computation ability will be represented as c . If the arithmetic of the accelerator is FP, its c will be defined as c_{FP} . On the other hand, c_{Fixed} means the computation ability under Fixed arithmetic. In the computation rows of Tables 2–7, the initial c is the original data released by the reference works. It might vary in arithmetic type and precisions. Based on [16,17], the computation ability of FP (c_{FP}) can be converted to the computation ability of c_{Fixed} by scaling three times. As a result, (1) is introduced.

Converted computation ability to a fixed point is defined as follows:

$$cFixed = cFP \times 3 \quad (1)$$

However, because not all accelerators have the same data precision, *cFixed* is not convincing while comparing the accelerators' ability. Reference [18] indicates that if a structure is not being optimized for precisions like Nvidia does in their GPUs, the theoretical performance of half-precision follows the natural $2\times/4\times$ speedups against single/double precisions, respectively. As a result, accelerators' computation ability performance needs to be normalized to 16-bit as it is used in the majority, without loss of generality. After normalization, the computation ability of each accelerator can be represented as *cFixed16*. To specify the accelerators' performance fairly, (2) is introduced. The lastest computation abilities shown in the computation rows in Tables 2–7 are the computation ability in 16 bit fixed-point format.

Converted computation ability to a 16-bit fixed point is defined as follows:

$$cFixed16 = cFixed \times (precision/16) \quad (2)$$

To specify the accelerators' synergy performance, (3) is introduced to represent the accelerators' evaluation value. Since the edge devices require low power consumption and compact size, in (3), the denominator will be power consumption p (w) times chip size s (mm²), and the fraction will be computation ability *cFixed16* (GOPs).

The equation for evaluating accelerator's synergy performance:

$$Evaluation\ value\ (E) = cFixed16 / (p \times s) \quad (3)$$

3.2. Prior Art Edge AI Accelerators

The following will show the edge AI accelerators [9,19–28], which focus on the demands of edge devices and are organized into Tables 2–4 according to their precision and power consumption. Table 2 shows the accelerators with 16-bit precision and containing less than a watt power consumption. Table 3 shows the 16-bit precision accelerators have relatively high-power consumption, higher than a watt. Table 4 shows the accelerators which do not belong to 16-bit precision.

Table 2. Prior Art Edge AI Accelerators.

Three Key Features and the Evaluation Value	Edge AI Accelerators		
	Kneron 2018 [9]	Eyeriss (MIT) 2016 [19]	1.42TOPS/W 2016 [21]
Computation ability	152 GOPs	84 GOPs	64 GOPs
Precision	16-bit Fixed	16-bit Fixed	16-bit Fixed [9]
Power consumption	350 mW	278 mW	45 mW
Size	TSMC 65 nm RF 1P6M Core area 2 mm × 2.5 mm	TSMC 65 nm LP 1P9M Chip size 4.0 mm × 4.0 mm Core area 3.5 mm × 3.5 mm	TSMC 65 nm LP 1P8M Chip size 4.0 mm × 4.0 mm
Evaluation value <i>E</i>	86.86 (core)	18.88 24.66 (core)	88.88
Implementation	Implemented in TSMC 65 nm technology	Implemented in TSMC 65 nm technology	Implemented in TSMC 65 nm technology
Commercial product example			-
	Packaged in a USB stick	Packaged on a PCIe interface card (no public sales channel found)	

Table 3. Prior Art Edge AI Accelerators.

Three Key Features and the Evaluation Value	Edge AI Accelerators			
	Myriad x (Intel) 2017 [22]	NVIDIA Tegra X1 TM660M 2019 [23,24]	Rockchip RK1808 2018 [25]	Texas Instruments AM5729 2019 [28]
Computation ability	1 TFlops =3 TOPs	472 GFlops =1.42 TOPs	100 GFlops =300 GOPs	120 GOP/s
Precision	16-bit FP	16-bit FP	16-bit FP 300 GOPs@ INT16	16-bit Fixed
Power consumption	<2 Watt	5–10 Watts (module-level)	~3.3 W (module-level)	≈6.5 W (module-level)
Size	8.1 mm × 8.8 mm (package)	28-nm 23 mm × 23 mm	22 nm ≈13.9 mm × 13.9 mm	28-nm 23 mm × 23 mm
Evaluation value E	21.55 (package s)	5.34×10^{-4} (module-level p)	3.81 (module-level p)	3.5×10^{-5} (module-level p)
Implementation	Implemented in TSMC 16 nm technology	Implemented as system on chip (probably TSMC 16 nm technology)	Implemented as system on chip (22 nm technology)	Implemented as system on chip (28 nm technology)
Commercialized product example	Packaged in a USB stick	Packaged on a PCIe interface card	Packaged in a USB stick	Packaged on a developing board

Table 4. Prior Art Edge AI Accelerators.

Three Key Features and the Evaluation Value	Edge AI Accelerators		
	GTI Lightspeur SPR2801S 2019 [16]	Optimizing FPGA-based 2015 [20]	Google Edge TPU 2018 [26,27]
Computation ability	5.6 TFlops =9.45 TOPs	61.62 GFlops =61.62 GOPs [20]	4 TOPs =2 TOPs
Precision	Input activations: 9-bit FP Weights: 14-bit FP	32-bit FP	INT8
Power consumption	600 mW 2.8 TOPs@300 mW	18.61 Watt	2 W (0.5 W/TOPs)
Size	28 nm 7.0 × 7.0 mm	On Virtex7 VX485T	5.0 mm × 5.0 mm
Evaluation value E	329.14	–	40.96
Implementation	Implemented in TSMC 28 nm technology	Implemented on the VC707 board which has a Xilinx FPGA chip Virtex7 485t	Implemented in ASIC (undisclosed technology)
Commercialized product example	Packaged in a USB stick	–	Packaged in a USB stick

After calculating their evaluation value E , accelerators [9,21] show similar abilities, $E = 80$ s. On the other hand, References [19,22,26] share similar E in the 20 s. Although some of the accelerators have close evaluation value E , the $cFixed16$, p , s values of these accelerators still need to be examined because the accelerators might apply for different purposes and environments. For example, Reference [16] has the highest evaluation value E , but its size is 9.8 times [9], 3 times [21], and nearly 2 times [26]. Overall, the evaluation value E can tell us a general efficiency of an AI accelerator, which is computation ability per unit area and watt. In Tables 2–4, several accelerators [23,25,28] lack details of the specifications since they only release the module-level data. As a result, the evaluation value E of them should be treated more conservatively. On the other hand, Reference [20] is a completed system on an FPGA board and does not release its size on a single chip, so its evaluation value E is hard to measure. Nevertheless, its data are a good study material for the designers who intend to build their future projects on an FPGA board for prototyping.

Some works, such as [29], use analog components and memristors to mimic neurons for CNN computing. However, none of the commercial proposed systems uses memristors. Several developers have researched memristor technology, including HP, Knowm, Inc. (Santa Fe, NM, USA), Crossbar, SK Hynix, HRL Laboratories, and Rambus. HP built the first workable memristor in 2008, yet until now, it still has a distance from prototype to commercial application.

Knowm, Inc. sold their fabricated memristor for experimentation purposes. Again, the memristor is not intended for application in commercial products [30]. Besides, it is worth mentioning that many CPUs in smartphones contain built-in neural processing units (NPU) or AI modules; for example, MediaTek Helio P90, Apple A13 Bionic, Samsung Exynos 990, Huawei Kirin 990, etc. However, individual NPU or so-called AI modules' detailed performance in these commercial CPUs is not public. As a result, these AI modules are hard to compare with the pure AI accelerators, but it is worth keeping an eye on these commercial products to prevent losing the latest information.

3.3. Coarse-Grained Cell Array Accelerators

Dynamically reconfigurable technology is the key feature of an edge AI hardware platform for flexibility and fault tolerance. The term 'dynamic' explains that during the runtime, reconfiguring the platform is still possible. Generally, reconfigurable architectures can be grouped into two major types: fine-grained reconfigurable architecture (FGRA) and coarse-grained reconfigurable architecture (CGRA). FGRA contains a large amount of silicon to be allocated for interconnecting the logic together, which implies that FGRA impacts the rate for reconfiguring devices in real-time due to the larger bitstreams of instructions needed. As a result, CGRA is a better solution for real-time computing.

Reference [31] presents many CGRAs and categorizes them into different categories: early pioneering, modern, larges, and deep learning. The article includes plentiful information, and the authors also collect the statistics to let readers know the developing trend of CGRA comparing to GPU. However, Reference [31] does not focus on the three key features' comparison for the CGRAs. For understanding the performance between CGRAs and determine which architectures are the potential candidates for edge AI accelerators, this paper presents architectures [32–42] published in the recent few years for comparison. For uniting the units and reference standards used in each article, this paper consults the references and converts the various units to be standardized according to the revealed information of each architecture in Tables 5–7. Table 5 shows the CGRAs use 32-bit precision, while Table 6 shows the 16-bit precision CGRAs. Last but not least, Table 7 presents the CGRAs, which neither use 32-bit, nor 16-bit precision.

Table 5. Coarse-Grained Cell Array Accelerators.

Three Key Features and the Evaluation Value	Coarse-Grained Cell Array ACCELERATORS			
	ADRES 2017 [32]	VERSAT 2016 [33]	FPCA 2014 [36]	SURE Based REDEFINE 2016 [37]
Computation ability	4.4 GFlops =26.4 GOPs (A9)	1.17 GFlops =7.02 GOPs (A9)	9.1 GFlops =54.6 GOPs (A9)	450 Faces/s ≈201.6 GOPs (Reference)
Precision	32-bit FP (A9)	32-bit Fixed	32-bit FP (A9)	32-bit Fixed
Power consumption	115.6 mW	44 mW	12.6 mW	1.22 W
Size	0.64 mm ²	0.4 mm ²	Xilinx Virtex6 XC6VLX240T	5.7 mm ²
Evaluation value <i>E</i>	356.84	398.86	–	29.48
Implementation	Implemented in TSMC 28 nm technology	Implemented in UMC 130 nm technology	Implemented in Xilinx Virtex-6 FPGA XC6VLX240T	Implemented in 65 nm technology

Table 6. Coarse-Grained Cell Array Accelerators.

Three Key Features and the Evaluation Value	Coarse-Grained Cell Array Accelerators		
	TRANSPIRE 2020 [34]	DT-CGRA 2016 [38,39]	Heterogenous PULP 2018 [42]
Computation ability	136MOPs (binary8 benchmark)	95 GOPs	170 MOPs
Precision	8/16-bit Fixed	16-bit Fixed	16-bits Fixed
Power consumption	0.57 mW	1.79 W	0.44 mW
Size	N/A	3.79 mm ²	0.872 mm ²
Evaluation value <i>E</i>	–	14	443.08
Implementation	Implemented in 28 nm technology	Implemented in SMIC 55 nm process	Implemented in STMicroelectronics 28 nm technology

Table 7. Coarse-Grained Cell Array Accelerators.

Three Key Features and the Evaluation Value	Coarse-Grained Cell Array Accelerators		
	SOFTBRAIN 2017 [35]	Lopes et al. 2017 [40]	Eyeriss v2 2016 [41]
Computation ability	452 GOPs (test under 16-bit mode)	1.03 GOPs =1.545 GOPs	153.6 GOPs
Precision	64-bit Fixed (DianNao)	24-bit Fixed	Weight/iacts: 8-bit Fixed psum: 20-bit Fixed
Power consumption	954.4 mW	1.996 mW 0.45 mm ²	160 mW
Size	3.76 mm ²	(not including the buffer, memory, and control systems)	24.5 mm ² (2 times of v1 [19])
Evaluation value <i>E</i>	125.96	1720.1	39.18
Implementation	Implemented in 55 nm technology	Implemented in 90 nm technology	Implemented in a 65 nm technology.

Some of the works do not show the computation ability in OPs, the standard reference unit for the AI accelerator designers and clients. Instead, a few of them compare their computation ability with the ARM Cortex A9 processor [32,33,36]. For example, Reference [33] releases Versat's performance in operation cycles by running the benchmarks and shows the ARM Cortex A9 processor's ability on those benchmarks for comparison. However, the article does not show Versat's OPs, the standard reference unit for the AI accelerator designers and clients. According to the results, it shows that Versat is 2.4 times faster than the ARM Cortex A9 processor on average. Then, Reference [43] shows the performance of the ARM Cortex A9 processor is 500 mega floating points per second (MFlops). Af-

ter calculation, Versat's operation ability is equal to 1.17 Giga Flops (GFlops). However, GFlops is still not the preferred unit for edge AI devices' designers and clients. Based on (1), GFlops can be converted to GOPs by scaling three times. Finally, the performance of Versat is gotten and equal to 3.51 GOPs in 32-bit precision. We adopt (2) to get its cFixed16, 7.02 GOPs, as shown in Table 5, for easily comparing to other accelerators. Similar works are done for the rest of the architectures in Tables 5–7. For exception, the area size of [34] cannot be found out due to the lack of information. Reference [36] does the work on an FPGA, so its core size is unable to be evaluated.

The computation unit used by [37] is faces-per-second because it targets face recognition, making the computation unit's converting work even harder and having significant deviation when referencing the ability from other similar work. Table 5 shows the computation ability of [37] is 450 faces/second, roughly equal to 201.6 GOPs [44]. In [37], it achieves recognizing 30 faces in a frame while the frame rate is 15 per second, which amounts to 450 recognitions to be performed per second. On the other hand, the reference work [44] recognizes up to 10 objects in a frame with a 60 per second frame rate. As a result, the converted computation ability of [37] remains for reference with a certain deviation. Reference [40] has the highest evaluation value E of all the listed works. However, the revealed size of [40] is only part of the architecture, so edge AI designers should be more conservative in assessing specific architecture specifications. Reference [42] does not release its computation ability. Since [42] shares the same architecture with [45], the operation/power ratio in [45] can be the reference. Furthermore, Reference [42] contains double cores and extra heterogeneous PEs comparing to [45]. As a result, the overall operation/power ratio of [42] would be higher than being evaluated.

Overall, the evaluation results show that [37,38,41] have tens' grade evaluation values E , between 10 to 40. References [32,33,35,42] have hundreds' grade evaluation values E , between 100 to 400. According to CGRA's evaluation value E , CGRAs show the potential ability to execute edge AI applications, like the outstanding prior art edge AI accelerators in Tables 2–4.

3.4. Implementation Technology

The implementation method of each edge AI accelerator has been shown in Tables 2–7. Most of the prior art edge AI accelerators in Tables 2–4 have been commercialized and taped out as ASIC chips, such as [9,16,22,23,25,26,28], which can be easily found online for sale in different system packages. Commonly seen, the system packages adopted by the systems are developing boards or USB sticks; the examples are in Tables 2–4. On the academic side, References [19,21] are also built as ASIC, but [20] is implemented in a single FPGA chip. When it comes to CGRA accelerators in Tables 5–7, although most of them are not taped out to get the physical chip, they have been synthesized on ASIC standard library. The technology they use has been organized in Tables 5–7.

From the implementation information in Tables 2–7, it can be noticed that FPGA and application-specific integrated circuits (ASIC) are the most used approaches to implement edge AI accelerators, including CGRA based, for their customized ability and low-power consumption. The non-recurring engineering expense (NRE expense) and flexibility of ASIC are high and low, respectively, compared to FPGA. As a result, building a system by ASIC has a higher cost than FPGA when the amount of the products is small. Not only that, but also the developing time of ASIC is longer than FPGA. At the beginning of the system developing process, FPGA-based platforms are the better solution due to their high throughput, reasonable price, low power consumption, and reconfigurability [46]. Accordingly, at the prototyping stage, building future AI platform design on a suitable FPGA platform at the system level [47] is suggested.

4. Architecture Analysis and Design Direction

Figure 1 organizes the accelerators whose evaluation value E is in the grade of tens or hundreds from Tables 2–7. References [20,34,36,40] are not included in Figure 1 because

they lack the area data at the chip level. The power consumption data of [23,25,28] is only released at module-level, so to be fair, they are not listed in Figure 1, either. The yellow area in Figure 1 represents the area size of each accelerator. Every bar in Figure 1 is composed of two parts, up and down in blue and orange color, respectively. The upper part in blue represents the GOPs/mm²; the lower orange represents the mW/mm². The two lines, the upper one and the lower one in Figure 1, represent evaluation value E and the ratio of GOPs and mW, respectively. To focus on accelerators targeting ultra-small areas, the accelerators whose area size is below 10 mm² are selected, and their three key features are presented in Figure 2a. In Figure 2a, each accelerator contains two bars. The left orange one represents the power consumption in mW while the right blue one represents the computation ability, GOPs. According to Figure 2a, the accelerators can be grouped into two categories by area size as units' grade and decimal grade. In the units' grade group are [9,35,37,38]. On the other hand, Refs. [32,33,42] belong to the decimal grade group. In the units' grade group, References [9,35] share a similar ratio of GOPs/mW while [37] has a relatively lower ratio, and [38] has the lowest ratio. The below analyzes the result of the ratio. Although [37] has close computation ability about 1.3 times [9] and 0.45 times [35], it consumes too much power, near 3.5 times [9] and 1.3 times [35]. When it comes to [38], its computation ability almost reaches a hundred but with huge power consumption, even higher than [37]. As a result, References [9,35] have better computation ability performance and power consumption in the units mm² area grade. It is interesting to know that the area size and GOPs/mW ratio positively correlate in the decimal grade group. Reference [42] has a relatively large area size compared to its computation ability. For more detail, the next paragraph will introduce the analysis.

Figure 1. Power consumption and operations per area statistics.

Figure 2b shows the normalized three key features of the accelerators in Figure 2a. The three key features of the accelerators are normalized to the same grade of computation ability by scaling up [32,33,38,42] and scaling down [35,37], linearly [48]. The result shows that except for [38,42], the remaining five accelerators have a similar trend in power consumption and area size. After normalization, the result emphasizes the unsatisfactory performance of [38,42] for low-power edge AI devices. Reference [38] consumes too much power while [42] has a too big area compared to its computation ability. However, if the targeting application requires ultra-low power consumption and can accept hundred-grade MOPs, Reference [42] is a good choice. Overall, a trading-off between computation ability and power consumption can be taken once the architecture size has been chosen. Designers can set accelerator's specifications according to its targeting application. As a result, if designers want to design an architecture for an edge AI accelerator in an ultra-small area

(units' mm^2 area size), the power consumption and operation ability should be in the order of hundreds of mWs and GOPs, respectively.

Figure 2. (a) Three key features statistics (accelerators under 10 mm²) and (b) the statistics normalized to the same grade of GOPs.

Figure 3 shows the scenario of the AI applications from the original to the future. It also illustrates the model of the AI application distribution in the yellow and purple oval, mentioned as three hierarchies. Figure 4 shows the legends of Figure 3.

Figure 3. The scenario of AI applications on hardware platforms. (Figure legends can be found in Figure 4.).

Figure 4. Figure legends of Figure 3.

The history of the developing data center-based AI in the first hierarchy can trace back to the 1980s. At that time, AI systems had developed to solve several specific problems as the decision-making ability of a human expert. It is achieved by if-then rules instead through conventional procedural code and composed by the inference engine and the knowledge base units. The idea is like the AI technology we are using nowadays. Rather than the knowledge base, we use big data and deep learning. Several modern database AI achievements are well noticed; the examples are Deep Blue (a chess computer 1996) and AlphaGO (a board game Go program 2014). The significant technology gap between Deep Blue and AlphaGO is deep learning neural network that can handle a larger branching factor in GO games.

Nowadays, much different information is collected by internet browsers, databases, and sensors through big data technology. Massively collected data is a good material for data training. According to the different individual applications, the trained weights are different. Trained weights are stored in cloud systems and wait for the unjudged data to come in. A specific application-oriented AI system uses a suitable deep learning neural network algorithm to infer the result, then transmit the result back to the requiring system. Cloud-based data center AI system is powerful due to its excellent hardware such as supercomputer or high-end GPU. It is hard to reach directly by ordinary consumers.

Thanks to the well-developed internet and personal consumer electronics, AI technology can benefit the general public. As a result, edge AI is a popular topic. The current development status of edge AI can be seen in both yellow and purple oval in Figure 3. The yellow oval shows that the data center AI system is connected with several edge AI devices through internet infrastructures such as routers and cell towers. Some edge AI systems need to work with intermediary systems such as a server and a PC to process data before transmitting or receiving data from the cloud. This type of edge AI system, which requires connection to the data center, is called an edge-cloud coordination system, which is currently well-spread and famous such as virtual assistants, e.g., Siri.

In Figure 3, the face recognition door lock can send the unjudged data, i.e., face picture, to the cloud for processing and get the inference result. However, in reality, the method is challenged due to safety concerns. As a result, systems like door locks requiring high standard security usually adopt edge AI systems instead of edge-cloud coordination systems. Because of the local inference feature of edge, AI systems can provide systems processing the data without an internet connection. The local inference feature is a big move for a device to promise data privacy, security, real-time processing, and working under no

internet coverage environment. This paper has introduced several edge AI accelerators, which target this field to satisfy the demand for edge AI devices.

The latest generation telecommunication standard, e.g., 5G, can be a valuable technology in the future edge AI development when it comes to training on chips. Training on chips might require something like distributed computing, which is achieved by the Internet of things (IoT) through 5G or future communication protocols. The idea is shown as the orange oval illustrated in Figure 3.

For example, autonomous cars contain pre-trained weights during the developing time to allow their AI system to make decisions for self-driving. The car manufacturer, Tesla, does epochal milestones on this page. However, pre-trained data might have some flaws due to the environment's change with an unexpected scenario. There are several non-fatal crashes reported in [49]. The idea of training on chips is that while human drivers engage the autopilot for a system non-aware situation but human handleable, the system can share the information with other compatible systems for training on chips. Training on chips can adopt distributed computing technology to reduce the computation.

The future trend of the research is how to distribute the data for training and share the weights between compatible systems, which might demand 5G or future telecommunication protocol to implement massive data exchange. Training on Chips also encounters another challenge. As shown in Figure 3, edge AI systems along different unmanned shops may experience several significant issues about which people are caring. The significant issues are personal data utilization and a standard platform/protocol for data exchanging, etc. Although personal data utilization may sound beyond the edge AI accelerators' technical design level, it is suggested that an accelerator designer involves a data encryption algorithm with data training on chips during the design level. Another significant topic is data training exchanging protocol. Applying training on chips to universal edge AI devices or systems is demanded to maximize the benefit of training on chips. The answer might be a cross-platform protocol, which will require the coordination of the designers.

Furthermore, there are several popular research of edge AI systems targeting harsh environments mentioned in the introduction. The scenario of a UAV working in a nuclear power station and a remote sensing satellite in Figure 3 are examples. They should be aware while designing future trend edge AI accelerators.

In summary, training on chips in edge AI accelerators will be a popular research topic. The distributed computing technology, 5G or future telecommunication protocol, data encryption for training on chips, cross-platform data exchange protocol, and harsh environment tolerance will involve in the developing path of the edge AI accelerators in the future.

5. Conclusions and Future Works

This paper has presented a survey of up-to-date edge AI accelerators and CGRA accelerators that can apply to image recognition systems and introduced the evaluation value E for both edge AI accelerators and CGRAs. CGRA architectures meet the evaluation value E of the existing prior art edge AI accelerators, which implies the potential suitability of CGRA architectures for running edge AI applications. The result reveals the evaluation values E of prior art edge AI accelerators and CGRAs are between tens to four hundred, indicating that edge AI accelerators' future design trend should meet this grade. Overall, the analysis shows that the future design of ultra-small area (under 10 mm^2) accelerators' power consumption and operation ability should be in the order of hundreds of mWs and GOPs, respectively.

As the edge devices are finding their way into various applications such as monitoring natural hazards by UAVs, detecting radiation leakage for nuclear disaster by robotics, and remote sensing in space by satellites, these applied fields are more critical than usual. Many research articles are targeting the fault tolerance feature for edge AI accelerators in recent years, which indicates that the trend of the edge AI accelerators is resilient in terms of reliability and high radiation field applicability. Finally, we illustrate the current status of

edge AI applications with a future vision, which addresses the technologies involved in the future design of edge AI accelerators and their challenges.

Author Contributions: Conceptualization, W.L., A.A. and T.A.; methodology, W.L.; formal analysis, W.L.; investigation, W.L.; resources, W.L. and T.A.; data curation, W.L.; writing—original draft preparation, W.L.; writing—review and editing, W.L. and T.A.; visualization, W.L.; supervision, T.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, and further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hoang, L.-H.; Hanif, M.A.; Shafique, M. FT-ClipAct: Resilience Analysis of Deep Neural Networks and Improving their Fault Tolerance using Clipped Activation. In Proceedings of the 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 9–13 March 2020.
2. Zhang, J.J.; Gu, T.; Basu, K.; Garg, S. Analyzing and mitigating the impact of permanent faults on a systolic array based neural network accelerator. In Proceedings of the 2018 IEEE 36th VLSI Test Symposium (VTS), San Francisco, CA, USA, 22–25 April 2018.
3. Hanif, M.A.; Shafique, M. Dependable Deep Learning: Towards Cost-Efficient Resilience of Deep Neural Network Accelerators against Soft Errors and Permanent Faults. In Proceedings of the 2020 IEEE 26th International Symposium on On-Line Testing and Robust System Design (IOLTS), Napoli, Italy, 13–15 July 2020.
4. Yasoubi, A.; Hojabr, R.; Modarressi, M. Power-Efficient Accelerator Design for Neural Networks Using Computation Reuse. *IEEE Comput. Archit. Lett.* **2017**, *16*, 72–75.
5. Venkataramani, S.; Ranjan, A.; Roy, K.; Raghunathan, A. AxNN: Energy-efficient neuromorphic systems using approximate computing. In Proceedings of the 2014 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), La Jolla, CA, USA, 11–13 August 2014.
6. You, Z.; Wei, S.; Wu, H.; Deng, N.; Chang, M.-F.; Chen, A.; Chen, Y.; Cheng, K.-T.T.; Hu, X.S.; Liu, Y.; et al. *White Paper on AI Chip Technologies*; Tsinghua University and Beijing Innovation Centre for Future Chips: Beijing, China, 2008.
7. Montaqim, A. Top 25 AI Chip Companies: A Macro Step Change Inferred from the Micro Scale. Robotics and Automation News 2019. Available online: <https://roboticsandautomationnews.com/2019/05/24/top-25-ai-chip-companies-a-macro-step-change-on-the-micro-scale/22704/> (accessed on 4 May 2021).
8. Simonyan, K.; Zisserman, A. Very deep convolution networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556.
9. Du, L.; Du, Y.; Li, Y.; Su, J.; Kua, Y.-C.; Liu, C.-C.; Chang, M.C.F. A Reconfigurable Streaming Deep Convolutional Neural Network Accelerator for Internet of Things. *IEEE Trans. Circuits Syst.* **2018**, *65*, 198–208. [CrossRef]
10. Clark, C.; Logan, R. *Power Budgets for Mission Success*; Clyde Space Ltd.: Glasgow, UK, 2011; Available online: <http://mstl.atl.calpoly.edu/~{}workshop/archive/2011/Spring/Day%203/1610%20-%20Clark%20-%20Power%20Budgets%20for%20CubeSat%20Mission%20Success.pdf> (accessed on 4 May 2021).
11. Yazdanbakhsh, A.; Park, J.; Sharma, H.; Lotfi-Kamran, P.; Esmailzadeh, H. Neural acceleration for GPU throughput processors. In Proceedings of the 48th International Symposium on Microarchitecture, Waikiki, HI, USA, 5–9 December 2015.
12. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.-R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
13. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, New York, NY, USA, 3–7 November 2014.
14. Vasudevan, A.; Anderson, A.; Gregg, D. Parallel multi channel convolution using general matrix multiplication. In Proceedings of the 2017 IEEE 28th International Conference on Application-Specific Systems, Architectures and Processors (ASAP), Seattle, WA, USA, 10 July 2017.
15. Guo, K.; Sui, L.; Qiu, J.; Yu, J.; Wang, J.; Yao, S.; Han, S.; Wang, Y.; Yang, H. Angel-eye: A complete design flow for mapping CNN onto embedded FPGA. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **2018**, *37*, 35–47. [CrossRef]
16. Gyr Falcon Technology Inc. (GTI). Lightspeur® 2801S. Available online: <https://www.gyrfalcontech.ai/solutions/2801s/> (accessed on 4 May 2021).
17. Farahini, N.; Li, S.; Tajammul, M.A.; Shami, M.A.; Chen, G.; Hemani, A.; Ye, W. 39.9 GOPs/watt multi-mode CGRA accelerator for a multi-standard basestation. In Proceedings of the 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, China, 19–23 May 2013.
18. Abdelfattah, A.; Anzt, H.; Boman, E.G.; Carson, E.; Cojean, T.; Dongarra, J.; Gates, M.; Grützmacher, T.; Higham, N.J.; Li, S.; et al. A survey of numerical methods utilizing mixed precision arithmetic. *arXiv* **2020**, arXiv:2007.06674.

19. Chen, Y.-H.; Krishna, T.; Emer, J.S.; Sze, V. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE J. Solid State Circuits* **2017**, *52*, 262–263. [\[CrossRef\]](#)
20. Zhang, C.; Li, P.; Sun, G.; Guan, Y.; Xiao, B.; Cong, J. Optimizing FPGA-based accelerator design for deep convolution neural networks. In Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Monterey, CA, USA, 22–24 February 2015.
21. Sim, J.; Park, J.-S.; Kim, M.; Bae, D.; Choi, Y.; Kim, L.-S. A 1.42TOPS/W deep convolution neural network recognition processor for intelligent IoE systems. In Proceedings of the 2016 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 31 January–4 February 2016.
22. Oh, N. Intel Announces Movidius Myriad X VPU, Featuring ‘Neural Compute Engine’, AnandTech 2017. Available online: <https://www.anandtech.com/show/11771/intel-announces-movidius-myriad-x-vpu> (accessed on 4 May 2021).
23. NVIDIA. JETSON NANO. Available online: <https://developer.nvidia.com/embedded/develop/hardware> (accessed on 4 May 2021).
24. Wikipedia, T. Available online: https://en.wikipedia.org/wiki/Tegra#cite_note-103 (accessed on 4 May 2021).
25. Toybrick. TB-RK1808M0. Available online: <http://t.rock-chips.com/portal.php?mod=view&aid=33> (accessed on 5 May 2021).
26. Coral. USB Accelerator. Available online: <https://coral.ai/products/accelerator/> (accessed on 5 May 2021).
27. DIY MAKER. Google Coral edge TPU. Available online: <https://s.fanpiece.com/SmVAXcY> (accessed on 5 May 2021).
28. Texas Instruments. AM5729 Sitara Processor. Available online: <https://www.ti.com/product/AM5729> (accessed on 5 May 2021).
29. Shafiee, A.; Nag, A.; Muralimanohar, N.; Balasubramonian, R.; Paul Strachan, J.; Hu, M.; Williams, R.S.; Srikumar, V. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In Proceedings of the 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Seoul, Korea, 18–22 June 2016.
30. Arensman, R. Despite HPs Delays, Memristors Are Now Available. Electronics 360. 2016. Available online: <https://electronics360.globalspec.com/article/6389/despite-hp-s-delays-memristors-are-now-available> (accessed on 4 May 2021).
31. Podobas, A.; Sano, K.; Matsuoka, S. A Survey on Coarse-Grained Reconfigurable Architectures from a Performance Perspective. *IEEE Access* **2020**, *8*, 146719–146743. [\[CrossRef\]](#)
32. Karunaratne, M.; Mohite, A.K.; Mitra, T.; Peh, L.-S. HyCUBE: A CGRA with Reconfigurable Single-cycle Multi-hop Interconnect. In Proceedings of the 2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC), Austin, TX, USA, 18–22 June 2017.
33. Lopes, J.D.; de Sousa, J.T. Versat, a Minimal Coarse-Grain Reconfigurable Array. In Proceedings of the International Conference on Vector and Parallel Processing, Porto, Portugal, 28–30 June 2016.
34. Prasad, R.; Das, S.; Martin, K.; Tagliavini, G.; Coussy, P.; Benini, L.; Rossi, D. TRANSPIRE: An energy-efficient TRANSprecision floatingpoint Programmable archItectuRE. In Proceedings of the 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 9–13 March 2020.
35. Nowatzki, T.; Gangadhar, V.; Ardalani, N.; Sankaralingam, K. Stream-Dataflow Acceleration. In Proceedings of the 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), Toronto, ON, Canada, 24–28 June 2017.
36. Cong, J.; Huang, H.; Ma, C.; Xiao, B.; Zhou, P. A Fully Pipelined and Dynamically Composable Architecture of CGRA. In Proceedings of the 2014 IEEE 22nd Annual International Symposium on Field-Programmable Custom Computing Machines, Boston, MA, USA, 11–13 May 2014.
37. Mahale, G.; Mahale, H.; Nandy, S.K.; Narayan, R. REFRESH: REDEFINE for Face Recognition Using SURE Homogeneous Cores. *IEEE Trans. Parallel Distrib. Syst.* **2016**, *27*, 3602–3616. [\[CrossRef\]](#)
38. Fan, X.; Li, H.; Cao, W.; Wang, L. DT-CGRA: Dual-Track Coarse Grained Reconfigurable Architecture for Stream Applications. In Proceedings of the 2016 26th International Conference on Field Programmable Logic and Applications (FPL), Lausanne, Switzerland, 29 August–2 September 2016.
39. Fan, X.; Wu, D.; Cao, W.; Luk, W.; Wang, L. Stream Processing Dual-Track CGRA for Object Inference. *IEEE Trans. VLSI Syst.* **2018**, *26*, 1098–1111. [\[CrossRef\]](#)
40. Lopes, J.; Sousa, D.; Ferreira, J.C. Evaluation of CGRA architecture for real-time processing of biological signals on wearable devices. In Proceedings of the 2017 International Conference on ReConFigurable Computing and FPGAs (ReConFig), Cancun, Mexico, 4–6 December 2017.
41. Chen, Y.-H.; Yang, T.-J.; Emer, J.; Sze, V. Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices. *IEEE Trans. Emerg. Sel. Topics Circuits Syst.* **2019**, *9*, 292–308. [\[CrossRef\]](#)
42. Das, S.; Martin, K.J.; Coussy, P.; Rossi, D. A Heterogeneous Cluster with Reconfigurable Accelerator for Energy Efficient Near-Sensor Data Analytics. In Proceedings of the 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 27–30 May 2018.
43. Nikolskiy, V.; Stegailov, V. Floating-point performance of ARM cores and their efficiency in classical molecular dynamics. *J. Phys. Conf. Ser.* **2016**, *681*, 012049. [\[CrossRef\]](#)
44. Kim, J.Y.; Kim, M.; Lee, S.; Oh, J.; Kim, K.; Yoo, H. A 201.4 GOPS 496 mW Real-Time Multi-Object Recognition Processor with Bio-Inspired Neural Perception Engine. *IEEE J. Solid-State Circuits* **2010**, *45*, 32–45. [\[CrossRef\]](#)
45. Gautschi, M.; Schiavone, P.D.; Traber, A.; Loi, I.; Pullini, A.; Rossi, D.; Flamand, E.; Gürkaynak, F.K.; Benini, L. Near-Threshold RISC-V Core with DSP Extensions for Scalable IoT Endpoint Devices. *IEEE Trans. Very Large Scale Integr. Syst.* **2017**, *25*, 2700–2713. [\[CrossRef\]](#)

-
46. Shawahna, A.; Sait, S.M.; El-Maleh, A. FPGA-Based Accelerators of Deep Learning Networks for Learning and Classification: A Review. *IEEE Access* **2019**, *7*, 7823–7859. [[CrossRef](#)]
 47. Lavagno, L.; Sangiovanni-Vincentelli, A. System-level design models and implementation techniques. In Proceedings of the 1998 International Conference on Application of Concurrency to System Design, Fukushima, Japan, 23–26 March 1998.
 48. Takouna, I.; Dawoud, W.; Meinel, C. Accurate Mutlicore Processor Power Models for Power-Aware Resource Management. In Proceedings of the 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, Sydney, NSW, Australia, 12–14 December 2011.
 49. Wikipedia. Tesla Autopilot. Available online: https://en.wikipedia.org/wiki/Tesla_Autopilot#Hardware_3 (accessed on 6 August 2021).