IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 July 2023, accepted 23 July 2023, date of publication 1 August 2023, date of current version 7 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3300379

RESEARCH ARTICLE

Real-Time Multi-Task ADAS Implementation on Reconfigurable Heterogeneous MPSoC Architecture

GUNER TATAR[1,2], (Member, IEEE), AND SALIH BAYAR[2], (Member, IEEE)
[1]Department of Electrical Electronic Engineering, Fatih Sultan Mehmet Vakif University, 34445 Istanbul, Turkey
[2]Department of Electrical and Electronic Engineering, Marmara University, 34840 Istanbul, Turkey

Corresponding author: Salih Bayar (salih.bayar@marmara.edu.tr)

ABSTRACT The rapid adoption of Advanced Driver Assistance Systems (ADAS) in modern vehicles, aiming to elevate driving safety and experience, necessitates the real-time processing of high-definition video data. This requirement brings about considerable computational complexity and memory demands, highlighting a critical research void for a design integrating high FPS throughput with optimal Mean Average Precision (mAP) and Mean Intersection over Union (mIoU). Performance improvement at lower costs, multi-tasking ability on a single hardware platform, and flawless incorporation into memory-constrained devices are also essential for boosting ADAS performance. Addressing these challenges, this study proposes an ADAS multi-task learning hardware-software co-design approach underpinned by the Kria KV260 Multi-Processor System-on-Chip Field Programmable Gate Array (MPSoC-FPGA) platform. The approach facilitates efficient real-time execution of deep learning algorithms specific to ADAS applications. Utilizing the BDD100K+Waymo, KITTI, and CityScapes datasets, our ADAS multi-task learning system endeavours to provide accurate and efficient multi-object detection, segmentation, and lane and drivable area detection in road images. The system deploys a segmentation-based object detection strategy, using a ResNet-18 backbone encoder and a Single Shot Detector architecture, coupled with quantization-aware training to augment inference performance without compromising accuracy. The ADAS multi-task learning offers customization options for various ADAS applications and can be further optimized for increased precision and reduced memory usage. Experimental results showcase the system's capability to perform real-time multi-class object detection, segmentation, line detection, and drivable area detection on road images at approximately 25.4 FPS using a 1920 × 1080p Full HD camera. Impressively, the quantized model has demonstrated a 51% mAP for object detection, 56.62% mIoU for image segmentation, 43.86% mIoU for line detection, and 81.56% IoU for drivable area identification, reinforcing its high efficacy and precision. The findings underscore that the proposed ADAS multi-task learning system is a practical, reliable, and effective solution for real-world applications.

INDEX TERMS ADAS, deep learning, deep processing unit, memory allocation, multi-task learning, MPSoC-FPGA architecture, Vitis-AI, quantization aware training.

I. INTRODUCTION

The emergence of deep learning (DL) and vision-based technology has significantly impacted the discourse surrounding autonomous driving. In this new era, autonomous driving systems are complicated constructs with multiple sensors and modules designed for specific functions. A crucial aspect of a robust autonomous driving system is its ability to perceive and respond to various environmental elements, such as surrounding vehicles, pedestrians, and traffic signs. DL has played a transformative role in advancing these processes beyond the capabilities of conventional algorithms. By applying DL methods to a wide range of machine vision tasks, autonomous driving systems have experienced

The associate editor coordinating the review of this manuscript and approving it for publication was Ludovico Minati.

improvements in accuracy, speed, and reliability. The integration of DL into this domain has opened up exciting possibilities for the future of autonomous vehicles. Although DL finds applicability across multiple domains, including healthcare, finance, and entertainment, its prominence is particularly evident in Advanced Driver Assistance Systems (ADAS) tasks. For instance, architectures such as VGG-Net [1], GoogleNet [2], DenseNet and ResNet [3] have been proposed for image classification and have successfully executed all tasks. In object recognition, multi-stage detectors, such as R-CNN [4] and Faster R-CNN [4], and single-stage object detection approaches, such as SSD [5] and YOLO [6], have been effectively deployed in contemporary applications. Similarly, models such as DeconvNet [7] and SegNet [8] are recognized for their proficiency in semantic segmentation, while LaneNet [9] and VPGNet [10] have demonstrated success in lane detection tasks.

An in-depth examination of each model reveals that their complexities can be attributed to the abundance of layers and parameters. This complexity stems from designers' ambition to enhance the performance of Deep Neural Networks (DNNs) by creating more extensive architectures. However, this increased complexity poses challenges in training, extracting insights, and implementing deeper architectures, which escalate alongside the growing requirements. Furthermore, using each specified DL model for a specific task makes employing them simultaneously in multi-tasking applications more feasible. This allows for more efficient utilization of the models across different tasks. Two distinct strategies emerge as potential solutions to address this complexity. The first strategy involves operating each task-specific model on separate hardware platforms, even though this method unavoidably has financial implications.

In contrast, the second strategy focuses on cultivating and expanding multi-task learning models. As the name suggests, multi-task learning operates on the premise that running numerous interconnected tasks on a single hardware platform using the same infrastructure can increase efficiency. While this approach may slightly increase the model's complexity for the designer, it simultaneously alleviates the financial duty of running distinct learning models on separate platforms. As a result, the execution of multi-task learning on constrained hardware through multi-tasking is rapidly emerging as a promising and practical strategy within ADAS.

This emerging sub-field within artificial intelligence is characterized by a groundbreaking approach: running multiple learning tasks simultaneously using a single model while leveraging the differences and similarities across these tasks. This novel paradigm holds immense promise and potential across various domains. The overall computational complexity can be significantly reduced by sharing backbone encoder computations. Moreover, multi-task learning takes advantage of various tasks' inherent and interconnected relationships, improving learning efficiency and predictive accuracy.

It becomes crucial to optimize both software and hardware architectures to optimize the processing speed and minimize inference time. Therefore, hardware-software co-design is vital for enhancing ADAS within a Multiprocessor System-on-Chip Field Programmable Gate Array (MPSoC-FPGA). This comprehensive approach spans both software and hardware dimensions, facilitating the maximum utilization of embedded resources and ultimately achieving high energy efficiency. By synergistically integrating hardware and software design, ADAS systems can unlock their full potential in various real-world applications.

In light of the discussed complexities, this paper proposes an innovative architecture tailored to address financial constraints while enabling the execution of multiple tasks simultaneously on a designated hardware platform. The novelty of this study lies in the simultaneous improvements made to the hardware accelerator platform and software enhancements. In the forthcoming chapters, we will thoroughly elucidate the unique software enhancements, the collaborative methodology involving hardware-software co-design, and the specialized memory allocation and pipeline structures implemented on the hardware accelerator side. This comprehensive examination aims to understand better the various components and innovative approaches employed in the proposed architecture.

A. MOTIVATION AND BACKGROUND

DL plays a crucial role in ADAS by enhancing autonomous driving capabilities and mitigating potential risks to life and property. The adoption of DL in ADAS is motivated by its exceptional proficiency in computer vision tasks, which are vital for the proper ADAS functionalities. DL models, including Convolutional Neural Networks (CNNs) and DNNs, demonstrate remarkable effectiveness in processing and extracting information from image and video data captured by sensors. These models can autonomously learn meaningful features, eliminating the need for manual feature engineering. Moreover, DL excels in complex pattern recognition, enabling end-to-end learning and simplifying the overall system architecture.

The scalability and adaptability of DL models allow them to generalize well across diverse driving scenarios. While other learning algorithms may find utility in specific components of ADAS, the unique strengths of DL in computer vision render it highly applicable in various ADAS applications. Consequently, our research efforts have been focused on the development and acceleration of DL-based ADAS algorithms, aiming to capitalize on the transformative potential of DL in advancing autonomous driving systems.

Selecting appropriate hardware is crucial to ensure that ADAS tasks using specified DL models run at desired performance levels without encountering resource consumption problems. Achieving this goal necessitates software optimizations on hardware accelerator platforms. Custom circuit blocks and chip architectures are leveraged to design hardware accelerators effectively. Similarly, operational cores are employed to balance performance and functional flexibility.

In embedded systems and applications, constraints such as speed, image size, power consumption, flexibility, accuracy, and memory play significant roles. Understanding the benefits of employing these sophisticated products requires a comprehensive grasp of DL models and hardware acceleration structures. DL algorithms are fundamentally based on neural networks, so a profound understanding of their architecture and configuration is critical for successful implementation in ADAS and other related domains.

At their core, DL algorithms rely on neural networks, emphasizing the importance of a profound understanding of their architecture and configuration for successful implementation in ADAS and related domains. Cloud-based DL applications are prevalent, utilizing open-source public clouds from companies like Google, Microsoft, and Amazon [14]. These companies' DL networks analyze vast amounts of data using the CPUs of thousands of servers. However, in scenarios where CPUs are insufficient for processing large datasets, they are combined with other hardware, such as FPGA, GPU, or ASIC-based accelerators, to utilize DL networks more efficiently. For example, while many companies employ computers with multi-core GPUs, Google utilizes its Tensor processor unit, and Microsoft relies on an FPGA system [14], [15].

Recent literature and research indicate a growing popularity of application-specific platforms, such as FPGAs and ASICs, due to their inherent parallelism capabilities [16]. The inherent parallelism of these platforms benefits DL model training by reducing execution time and enhancing program accuracy. Consequently, designers can focus on the DL model's core aspects by alleviating the parallelization workload. Effective hardware-software co-design is imperative to execute DL models on MPSoC architectures involving ARM and FPGA components together. Designers must allocate computationally intensive portions to be executed on the Deep Processing Unit (DPU) side while the ARM processor side handles the remaining tasks. This coordinated approach ensures efficient resource utilization and maximizes DL-based ADAS applications' performance.

Considering the specific requirements of the ADAS application under study, the KRIA KV260 MPSoC FPGA selection was carefully considered. The KRIA KV260 is a suitable choice, mainly due to its alignment with the computational demands, flexibility, and power efficiency needed for our ADAS application. Notably, it offers specific features, such as high-speed interfaces and dedicated hardware accelerators, which are well-aligned with the needs of our ADAS system. Moreover, the KRIA KV260 boasts robust support and a mature ecosystem for deep learning development, simplifying the implementation and optimization of our DL algorithms.

Furthermore, the versatility of the KRIA KV260 is evident in its frequent utilization across various domains, including ADAS, robotics, and industrial automation. These applications often necessitate striking a balance between processing power and FPGA flexibility, making the KRIA KV260 a compelling choice for our work in ADAS.

B. RELATED WORK

This section provides a concise overview of advanced previous works on multi-object detection, semantic segmentation, and multi-tasking ADAS. The literature on ADAS is vast, making it challenging to compare all studies directly. Therefore, we have focused on studies conducted in recent years that are closely related to our proposed research. We have identified common aspects regarding tasks, datasets, software-hardware co-design, hardware platforms used, FPS values, and inference accuracy. Our approach begins with concise summaries of state-of-the-art studies, categorized from single-task to multi-task studies. Subsequently, we present an encapsulating Table 1 that provides a comprehensive and accessible comparison between various studies.

Table 1 highlights a notable trend in the latest state-of-the-art studies, where GPU-based hardware architectures are predominantly utilized. The primary reason for this preference is the exceptional acceleration potential of GPUs, with hundreds of cores capable of processing thousands of threads in parallel, leading to significant performance gains compared to other architectures. As modern computational problems often exhibit parallel structures, GPUs are well-suited for tasks like video processing, image analysis, signal processing, and DL essential for running self-driving vehicles in real-time. Moreover, GPUs are programmable with advanced software packages like CUDA, TensorFlow, and PyTorch, enhancing their usability.

However, the energy consumption and high cost associated with GPUs pose constraints, particularly for battery-powered devices. As a result, designers seek architectures that offer GPU-like features while being more energy-efficient and cost-effective. In this regard, application-specific platforms, such as FPGAs and ASICs, are gaining popularity due to their inherent parallelism capability, offering advantages in program execution time and accuracy [16], [17], [18], [19], [20].

To conclude, GPU-based hardware architectures remain prevalent in state-of-the-art ADAS studies due to their powerful acceleration capabilities and user-friendly programming interfaces. Nonetheless, the adoption of application-specific platforms, especially FPGAs, is on the rise, driven by their inherent parallelism, low energy consumption, and cost-effectiveness. Developing high-level software frameworks for FPGA programming has further contributed to their increased use in DL applications.

Based on the given information, we highlighted the strengths and weaknesses of the studies that showed the closest similarity to our research from the references in Table 1.

For example, in their insightful work, Ghorbel et al., [27], proposed a method that utilizes GPU acceleration to parallelize the eyes detection algorithm based on Viola and Jones, a development aimed at designing an innovative smart

TABLE 1. ADAS multi-tasks implementation in the literature.

Ref.	Drivable area & Line detection	Semantic segmentation	Multi object det.	Data-sets	HW-SW co-design	Platform	FPS	Evaluation (%)
[27]	Y	N	N	FDDB	Y	C-based GPU Omap4460	55ms	87.33 (Accu.)
[22]	Y	N	N	COCO	N	NVIDIA Jetson Nano	20	60 (Accu.)
[24]	Y	Y	N	Cityscapes KITTI	N	NVIDIA RTX TITAN	N	72.32 (mIoU)
[25]	Y	Y	N	tuSimple Cityscapes BDD100K	N	NVIDIA RTX 2080Ti GPU	N	95.73 (Accu.) 70.9 (F1) 57.03 (Accu.)
[26]	Y	Y	N	Their	N	GTX 1080Ti Xavier NX	50 15	94.51 (F1)
[23]	Y	Y	N	Cityscapes	N	NVIDIA RTX 2080Ti & ROS	N	N
[33]	Y	Y	N	KITTI	N	NVIDIA TITAN GPU	N	94.21 (AP)
[28]	N	N	Y	VOT VBT	Y	Zynq UltraScale+ MPSoC ZCU3EG	123 53.3	66.25 (mIoU)
[29]	N	N	Y	Their	N	NVIDIA GTX 1060	16.68	0.986 (mAP)
[30]	N	N	Y	GTSRB	N	NVIDIA Jetson Xavier AGX & Xavier Nx	43.59 23.17	0.621 (mAP)
[32]	N	N	Y	CVC-09 FLIR ADAS OSU, KAIST	N	NVIDIA Jetson Nano	3	84.1 (mAP)
[34]	Y	N	Y	Air learning database	Y	Xavier NX Jetson TX2	6	91 (Accu.)
[31]	N	Y	N	Cityscapes	N	GTX TITAN X Maxwell (GPU)	30	70 (mIoU)
[35]	Y	Y	Y	N	N	NVIDIA Jetson Xavier & TI TDA2x	10 15	39.78 (mAP)
ADAS Multi task- learning (Proposed study)	Y	Y	Y	BDD100K Waymo Cityscapes KITTI	Y	KRIA KV260 MPSoC FPGA	25.4	Object det.: 51 (mAP) Segment.: 56.62 (mIoU) Drivable: 81.56 (mIoU) Line det.: 43.86 (IoU)

* Y/N: Yes/No, IoU: Intersection over Union, mIoU: mean Intersection over Union, mAP: mean Average Precision, FPS: Frame per Second, Accu: Accuracy

wheelchair. The authors subsequently apply this research to craft a human-machine interface to govern intelligent wheelchair control. Notably, their work incorporates a significant element of software-hardware co-design. However, their study appears to lack comprehensive coverage of multi-task learning, which warrants further exploration to augment the design and performance of intelligent wheelchairs. In the concluding remarks of their study, Ghorbel et al. explicitly acknowledge the persistent challenge posed by GPUs in the context of energy consumption and efficiency. This issue is particularly pronounced in electronic systems powered by batteries, potentially limiting the viability and practical applicability of their research in real-world settings.

Moreover, the study employs the Omap4 4460 platform, which may be prohibitive considering the price-performance ratio. It is crucial to examine whether more cost-effective alternatives could achieve comparable, if not superior, results in designing intelligent wheelchairs and their respective human-machine interfaces. This would improve the accessibility and affordability of the technology for a broader user base.

In their study, the authors [28] undertook a comparative analysis of three disparate deep convolutional neural network hardware accelerator implementation methods. These encompassed coarse-grained, fine-grained, and sequential Vitis-AI strategies. Two bespoke DNN architectures were developed within the System Verilog and FINN frameworks, displaying the flexibility and applicability of these models. Notably, despite achieving high performance in terms of FPS rate, the authors did not explore multi-task implementation in their work, which leaves room for further investigation into improving computational efficiency. Another critical observation to be made about the study is their choice of the high-performance MPSoC ZCU3EG for the exclusive task of object detection. Given the processing capabilities of this particular system, it could potentially be better leveraged by distributing the computational load across multiple tasks, thereby optimizing resource use. This single-task

focus leaves the potential for more comprehensive, multi-task approaches largely unexplored. Such methodologies could prove beneficial for enhancing processing efficiency and performance in future research endeavours.

An additional noteworthy study is the one conducted by Cho et al., [33], titled "Multi-task Self-supervised Visual Representation Learning for Monocular Road Segmentation," which attained a commendable 94.23% Average Precision (AP) performance. Within their research, the authors employed a multi-task framework for segmentation-based roadway identification. Their study utilized the KITTI dataset and relied on the considerably costly NVIDIA TITAN GPU as their hardware. In a distinctive shift from traditional literature, the authors examined the utility of unsupervised stereo-based indicators to acquire high-level semantic knowledge for monocular route detection. Their experimental outcomes indicated an above-average performance, albeit not significantly superior compared to our research. Our study boasts several advantages over Cho et al.'s research, such as a broader array of utilized datasets (KITTI, BDD100K, Waymo and CityScapes), the adopted methodology, the concurrent execution of four distinct tasks, and the incorporation of a backbone encoder.

In the seminal work by Krishnan et al., [34], an in-depth exploration was conducted on the automation of domain-specific SoC design intended for autonomous vehicles, with particular emphasis on Unmanned Aerial Vehicles (UAVs). The scope of their research primarily encompassed nano, micro, and mini-UAVs, for which they utilized the capabilities of Xavier NX and Jetson TX2 to fashion domain-specific SoCs accelerators. Their research findings confirmed an acceleration factor of approximately 2.25x, 1.62x, and 1.43x, respectively, across the range of UAVs studied.

It is crucial to note that the approach employed by the authors, though commendable in its results, diverges from ours in a significant aspect, namely the software-hardware co-design, which is the cardinal characteristic differentiating our study. In addition to this distinguishing factor, our research incorporates semantic segmentation, which remarkably enhances the FPS value. This constitutes a considerable advantage when the two studies are juxtaposed.

Furthermore, the original study conducted by Krishnan et al. exhibits a preference for GPU-based accelerator platforms. Despite their numerous benefits, it is widely recognized that such platforms are mainly inefficient in energy consumption, owing to their reliance on hundreds or even thousands of CUDA cores. This aspect further underscores our research's distinctiveness and potential advantages, which circumvents this significant energy inefficiency challenge.

In their substantive research, Lai et al. [35] introduced a Multi-Task Semantic Attention Network (MTSAN) designed to amalgamate segmentation and object detection functionalities for real-time applications in ADAS. Although their approach substantially reduced false alarm frequency, it did so at the cost of increased computational resources. The researchers reported an FPS rate of 10 at a resolution of 512×256 on NVIDIA's Jetson Xavier and a slightly improved 15 FPS at the exact resolution on Texas Instruments' TDA2x platform. However, given the considerable investment associated with NVIDIA's Jetson platforms, these FPS rates could be more impressive when juxtaposed with our model's superior performance.

Moreover, the authors characterized their hardware as low-power, a claim inconsistent with the commonly recognized high energy consumption intrinsic to GPU-based architectures. Their study further reveals a subtle under emphasis on the multi-task functionality of their model, creating an opportunity for a more thorough model evaluation and comparison. Contrasting the performance metrics of our model with those of the authors provides a clearer picture of our superiority. Enhancements on numerous fronts, including FPS, power consumption, and memory resource utilization, are observed in the presented model. It showcases superior real-time performance, exhibiting an improved FPS rate compared to [34] and [35]. Furthermore, our model's mAP and mIoU results exceed those of the MTSAN model, which underscores our superior performance in object detection and segmentation tasks.

Regarding power consumption, our model operates on an MPSoC architecture, renowned for delivering robust computational performance while consuming significantly less power than GPU-based architectures. This issue summarises our model being more energy-efficient while providing full ADAS functionalities. Regarding memory resources, we leverage the Xilinx Kria KV260 platform, renowned for optimal memory resource utilization, which results in a more memory-efficient solution.

In summary, our model presents a more balanced solution for ADAS applications by providing robust performance coupled with energy efficiency and cost-effectiveness. This situation culminates in our model outperforming those proposed by [34] and [35] across multiple evaluation parameters, asserting its superiority in this field.

As Table 1 shows, there are many methods and studies for ADAS development. Here, we have tried to present famous state-of-the-art outcomes closest to our work. In addition, studies involving software optimization and the use of hardware accelerators in designing ADAS applications have become popular. Hardware accelerators can take the form of CPUs, ASICs, GPUs, and FPGAs. In such event, CPUs are combined with other hardware for utilizing the DL networks. When hardware accelerators are inadequate, one or more of the platforms noted above, like CPUs, GPUs, FPGAs, and ASICs, are used together for the training inference of DL algorithms [15], [36].

Recently, application-specific platforms (e.g. FPGAs, ASICs) are becoming more popular [37] due to their structural parallelism capability, favouring DL algorithm training in program execution time and accuracy. It is to be highlighted that there is a gap in the literature on the need for a design encompassing high FPS throughput and mAP-mIoU value,

better performance at a lower cost, multi-tasking on a single hardware, and integration into memory-constrained devices to enhance the performance of ADAS tasks. Therefore, undertaking an integrated hardware-software design incorporating the above-mentioned features is crucial to enable the widespread adoption of autonomous vehicle technologies.

In light of this information, we preferred the FPGA-based MPSoC architecture because of its ability to perform parallel processing, its price/performance compatibility and its more flexible structure. We have realized an efficient hardware-software co-design on the MPSoC structure. We created our own DPU architecture on the programmable logic (PL) side so that the ADAS multi-task learning yields better results than the existing studies. We arranged the computationally demanding parts of the DL model to be on the DPU and the video pre-processing, task allocations and inference parts on the ARM. For DPU-ARM data interaction, we used the advanced extensible interface - direct memory access (AXI-DMA) and pipeline architecture.

C. CONTRIBUTIONS

We have given the study's main contributions to the literature in a list as follows;

1) We developed ADAS multi-task learning system, which can perform several tasks, including semantic segmentation, multi-object detection, line detection, and derivable area detection, on a single piece of hardware. This approach can lead to efficient and effective development of embedded systems, particularly for ADAS applications.
2) We enhanced the efficiency of our model for resource-limited embedded devices by examining its backbone. To reduce its memory footprint on constrained platforms, we quantized the model using int-8 bits.
3) We constructed effective optimizations to the proposed model to improve its performance without affecting the accuracy of the inference.
4) We assembled the programmable DPU reserved for the convolutional neural network.
5) Through hardware-software co-design, our proposed approach offers high performance and low energy consumption when compared to other hardware architectures for similar tasks. We conducted a feasibility study by deploying the proposed model on low-power embedded devices and demonstrated real-time processing using a prototype design. Specifically, we investigated the ability to program traditional programming languages, such as Python and C++, in heterogeneous MPSoC architecture with hardware-software co-design.

The remaining paper scenario is as follows. Section II contains a discussion on the design methodology of the proposed model. Initially, an overlay is designed utilizing AMD-Xilinx Vivado 2022.1, which is then imported to the MPSoC-FPGA-based development board. The performance of software design and enhancements within the Python environment follows this. Notably, the focus here is also on the model's training, quantization, and compilation. Lastly, the experimental setup and execution of the *.xmodel* file quantized and compiled using *vai_q_pytorch* are deliberated. In Section III, a rigorous comparative analysis of our findings with analogous studies from the existing literature is undertaken, with primary emphasis on parameters such as power consumption, reliability, longevity, accuracy, and overall efficiency in order to optimize the quality of the design. Concluding the paper, Section IV encapsulates the summative observations drawn from the entirety of the article and clarifies prospective avenues for future research.

II. METHODOLOGY

This section raises a broad methodological strategy adopted for our research. It commences with a discussion on the structure of the multi-task learning network and its pertinent subtopics. Following this, an exploration of the hardware design section and its associated subtopics is presented, developed utilizing the Vitis unified software platform and the Vivado 2022.1 integrated development environment. Ultimately, the QAT for the model incorporating a ResNet-18 shared backbone encoder with SSD is addressed in this work.

A. INTRODUCTION

DNNs have become a fundamental tool extensively employed in ADAS, prompting an increasing focus on enhancing and accelerating DL-based ADAS algorithms. Achieving such improvements and accelerations has relied mainly on software optimizations. These optimizations typically involve techniques to speed up the algorithm's execution, parallelise it, or integrate libraries like PThreads, OpenCL, and MPI into the algorithms. The ultimate goal of software optimizations is to address potential algorithm bottlenecks using the specified methods. However, the growing complexity of DNN models, which may comprise millions or even billions of parameters, has sometimes rendered software-based optimizations inadequate.

The advancements in semiconductor process technology have opened up new possibilities for using hardware as an accelerator and synergising it with software optimizations. This hardware and software integration has emerged as a prominent trend in recent state-of-the-art studies. In line with these developments, our research incorporates a DPU-based hardware accelerator complemented by software optimisation utilising high-level languages such as C/C++ and Python.

The contributions of this study to the existing scholarly discourse are elucidated in Section I-C, where a comprehensive overview of our research's key findings is presented. Subsequently, this section delves into the specifics of our contributions. The discussion sequence commences with a detailed exploration of the system's infrastructure, an in-depth analysis of the hardware design, and an investigation

into the software design and the associated optimisation methodologies. This study enhances DL-based ADAS algorithms by combining hardware acceleration with software optimizations. The synergistic integration of hardware and software offers the potential to address the challenges posed by increasingly complex DNN models and optimise the performance of ADAS systems effectively. The subsequent sections will delve into the technical aspects of this integration and its implications for ADAS applications.

B. MULTI-TASK LEARNING NETWORK

In the current landscape of deep learning networks, the prevailing approach often involves addressing a specific task. However, practical applications frequently demand higher efficiency achieved through integrating multiple individual algorithms into a single comprehensive learning framework. This integration is made possible by multi-task learning networks, which consolidate various tasks into a cohesive unit. The efficacy of this approach hinges on exploiting the relationships between these diverse tasks and optimizing them for real-time applications.

Multi-task learning networks combine different functions into a well-suited task, capitalizing on the interrelationships between these tasks. This study proposes a unified learning network that combines four distinct learning algorithms for ADAS. We enhance efficiency in real-time applications by integrating algorithms that perform multiple individual tasks into a single framework. Consequently, the network gains a more accurate representation of functions by sharing features among tasks, leading to improved learning efficiency and enhanced prediction accuracy. Additionally, by sharing backbone layers, multi-task learning can reduce the overall network size (memory footprint) and computational complexity. This advantage is particularly beneficial for resource-constrained platforms and fulfils quick inference requirements.

Our proposed ADAS multi-task learning approach comprises a shared backbone encoder, a segmentation subnet, and a detection subnet, as depicted in Fig. 1. The specific details of these subnet models have been thoroughly discussed to provide a comprehensive understanding of our proposed methodology. The shared backbone encoder is a typical feature extractor, efficiently utilizing extracted features across different tasks. The segmentation subnet focuses on semantic segmentation, while the detection subnet concentrates on object detection.

1) SHARED BACKBONE ENCODER

Various computer vision tasks, including object detection, are addressed using complex CNN architectures. The construction of object detection or segmentation architectures is based on CNN models initially trained for image classification, thanks to the principle of transfer learning. In this context, CNN serves as the feature extractor and forms the backbone of the object detection model. This backbone processes input data, focusing on specific features to extract detailed conceptual elements.

The study's model is rooted in the ResNet18 architecture rather than using deeper structures such as DenseNet, ResNet101, or GoogleNet. These have numerous hyper-parameters and high computational densities. This choice enables the operation of the model on open-source and resource-limited embedded devices. The model is constructed through transfer learning, with initial training targeting object detection, segmentation, and classification. Performance optimization of the model is achieved by employing a shared backbone. This approach facilitates efficient multi-task learning for ADAS and reduces the computational load, positioning it as an effective solution for real-world applications. The strategic design of the model was undertaken with these considerations in mind.

2) DETECTION SUBNET

Our research adopted a single-stage SSD [5] as a detection decoder to fulfil real-time application needs and enable quick inference. The SSD employs a multi-scale feature map for swift object detection. However, as the CNN structure progressively downsizes the spatial dimensions, it also lowers the resolution of the feature map. Hence, the SSD uses lower-resolution layers for detecting and positioning larger-scale objects. As such, our research opted for a 4×4 feature map for identifying large objects. After VGG16, SSD introduces six extra convolution layers, with five dedicated to object detection. This layer configuration prompted us to generate six predictions on three rather than the conventional four, leading to approximately 8732 predictions across these layers. This strategic choice contributed to our model's efficiency and robustness, enhancing its ability to make precise object detection under various conditions.

3) SEGMENTATION SUBNET

The segmentation subnet's architectural design incorporates several learnable up-sampling layers. This subnet comprises a 3×3 convolution layer, layers of batch normalization, and multiple layers with activation functions. The output tensor size of one convolution block layer and the input tensor size of another layer remain consistent, with only up-sampling altering the tensor size.

A convolution block is initially applied at the subnet's bottom, facilitating the extraction of meaningful semantic features. Instead of using pooling to extract low-resolution features, three convolution blocks incorporating a dilated layer are employed. Given that the pooling process can result in a loss of detail, applying dilated convolution provides more relevant results for extracting deep features than standard convolution and subsequent pooling processes. After the deep feature extraction, up-sampling is performed to restore the spatial resolution. In the processing methodology, hints about objects provided by the encoder properties are integrated into the decoder side to delineate the boundaries with more

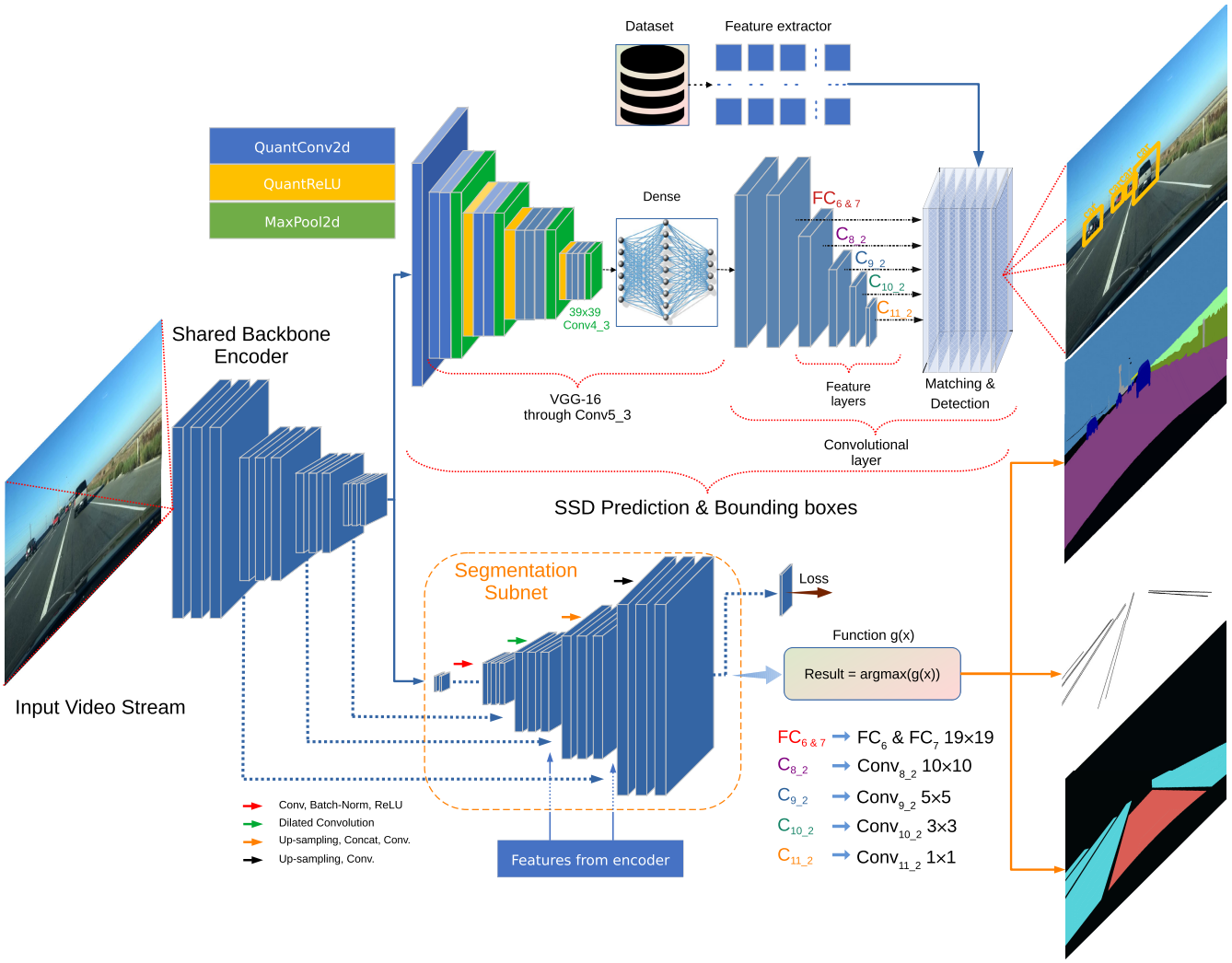


FIGURE 1. General overview of the ADAS multi-task learning.

precision. Feature summation is used in place of element accumulation to enhance inference accuracy.

4) SOFTWARE OPTIMIZATION

FPGAs and SoCs implement domain-specific architectures to optimize CNN in applications, including inference rate, latency, and hardware utilization. Our ADAS multi-task learning model runs on ZYNQ Ultrascale+ MPSoC architecture as a co-design where DPU and ARM are used together. The DPU accelerates the computing workloads of DL inference algorithms commonly used in various computer vision applications. Therefore, the DPU performs the CNN computations here, while the ARM performs the pre-and post-processing of the input signal. Furthermore, depth-wise separable convolution (DWSC) is adapted to the on-chip pipeline method to process efficiently in parallel, thereby reducing off-chip memory access. Thus, the DPU core significantly improves run-time scheduling efficiency during the computation of layers. DWSC decreases computational intensity [38] approximated to normal convolution operations. DWSC has the fundamental principle of separating two procedures dept-wise convolution (dwc) and point-wise convolution (pwc). Pwc has a 1×1 standard structure that follows deep convolution (see Fig. 2). It collects feature information from different channels in the exact spatial location, thus reducing the computational cost and memory footprint of separable convolutional networks. In an effort to optimize memory utilization on the ARM (Quad-Core Cortex A-53) processing engines (PEs), careful data allocation was enacted to prevent unnecessary re-reading. This facilitated the creation of an adaptable infrastructure for diverse neural network models. Similarly, multi-threading and a pipeline architecture were implemented to leverage the DPU and Quad-Core PEs' full potential on the Kria KV260 development board. As a result, high efficiency was attained by mitigating delays during data communication between the PL and the PS. The processing duration of the PL, observable during data transactions via AXI-DMA, underscores this improvement.

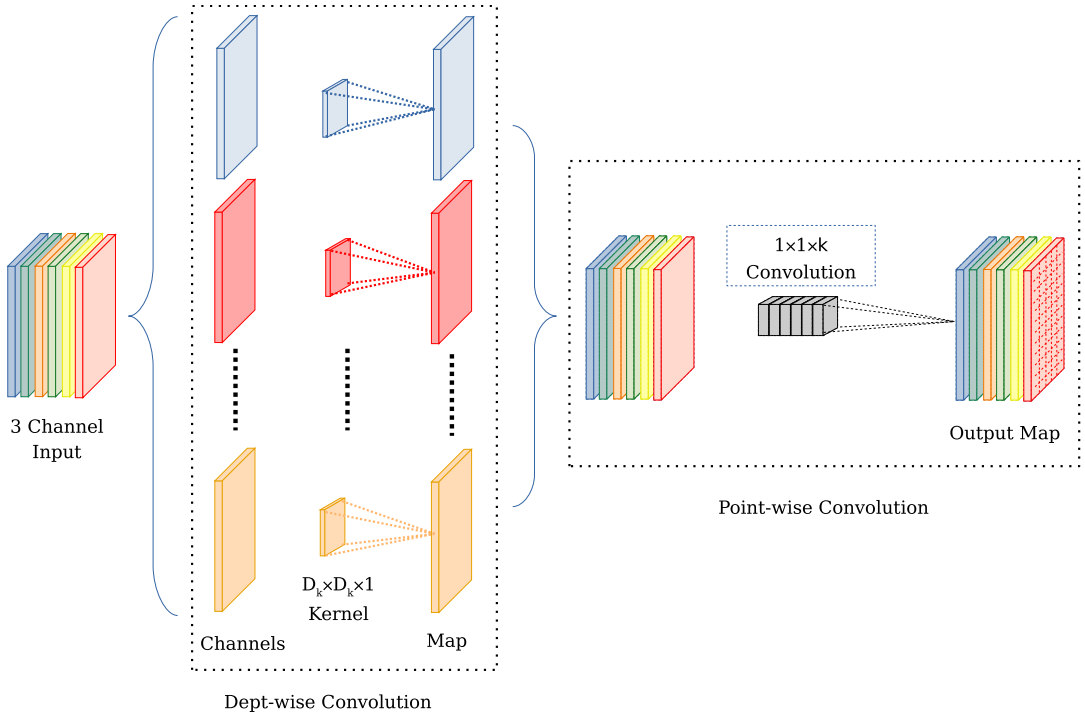


FIGURE 2. Separable dept-wise convolution architecture.

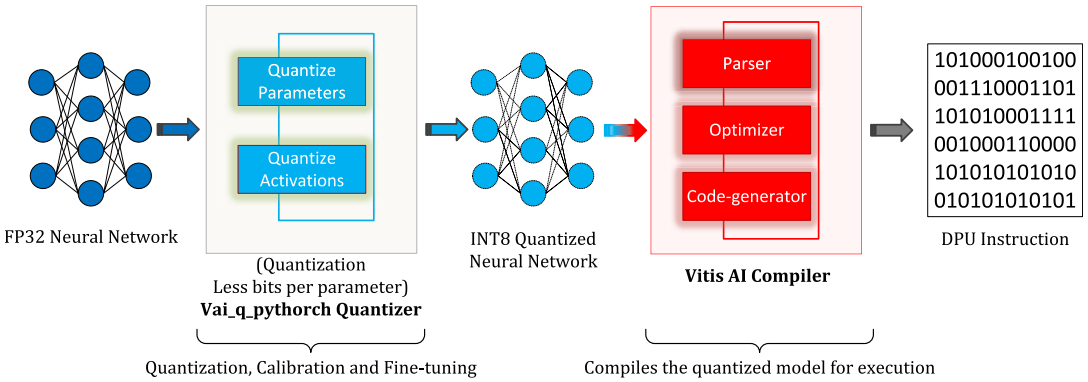


FIGURE 3. An overview of quantization and compilation of FP32 bits NN model.

A significant software-side enhancement was the quantization of the model, as depicted in Fig. 3. This modification resulted in a model with a memory footprint well-suited to resource-constrained devices, thereby considerably reducing computational density. Although model quantization may slightly impair inference accuracy, the benefits are sufficiently significant to overlook this minor setback. A detailed account of the improvements and inference results obtained are comprehensively discussed in Section III.

C. HARDWARE (OVERLAY) DESIGN

The ZYNQ architecture comprises a customizable MPSoC incorporating a quad-core ARM Cortex-A53, dual-core ARM Cortex RF53, ARM Mali 400MP and a conventional PL integrated circuit (IC). In addition, MPSoC is equipped with fast and efficient connections and supports the AXI standard. The system's design phase requires tasks to be allocated between the processor and the FPGA sections, known as PS and PL, based on the system requirements. This allocation is a critical step as the overall speed and functionality of the program depend on how tasks are distributed between PS and PL sections. The entire system's performance is influenced by the tasks assigned to each team. Our research proposal allocated the high-speed and computationally intensive parts to PL while assigning the remaining roles to PS.

Identifying the functional blocks in the hardware design, we integrated them as IPs to establish the necessary AXI-DMA interfaces between PL and PS, as depicted in Fig. 4. Digital hardware development was carried out using the Vivado 2022.1 integrated development environment,

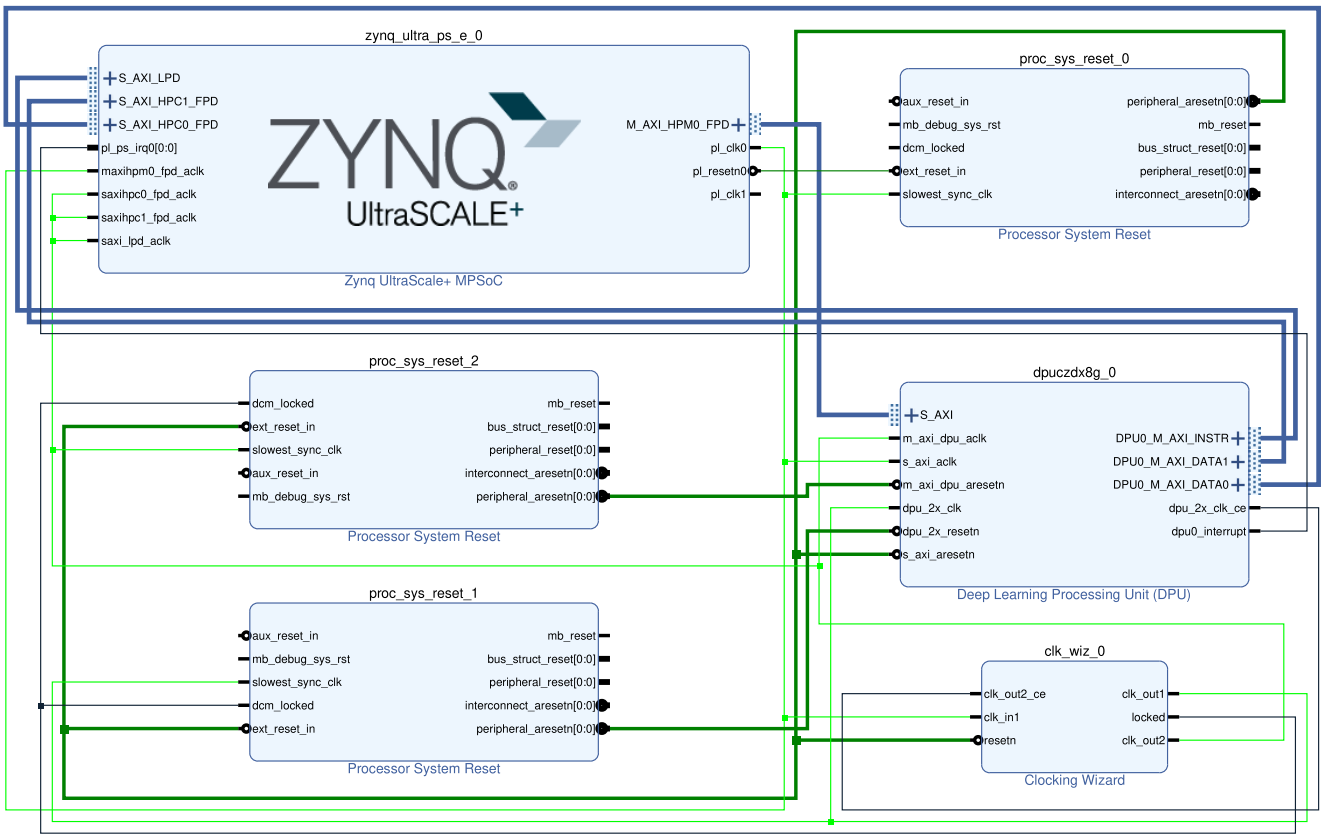


FIGURE 4. Overlay design of ADAS multi-task learning hardware.

while high-level synthesis integrated system design was conducted in PL. The hardware accelerator design was prepared in the PL environment and comprised the overlay comprising IP blocks.

The ARM is utilized for pre-processing and post-processing tasks in this study, exploiting the capabilities of high-level languages such as Python and C++ and the OpenCV library. As part of the methodology, the decode thread is programmed to carry out resize functions on images of 1920 × 1080 and 320 × 512 resolutions. During the ML inference phase, scale and mean value subtraction operations are performed on the ARM, leveraging the capabilities of the Vitis-AI library. The remainder of the task is executed within the PL, specifically in the DPU. Due to the convolution process's intricate nature and heavy processing demands, the high-performance computational abilities of the DPU are employed to carry out this process successfully.

The execution of tasks in this study specifically involved the use of the AXI4 stream interface, a decision influenced by the superior functionalities of the Direct Memory Access (DMA) feature, which facilitated a notably rapid transfer of image pixel values. Strategic adjustments to the primary port were carried out to facilitate seamless interaction between the PS and the PL. These modifications included setting the bit width for the AXI HPM0 FPD at 32, the AXI HPM1 FPD at 128, and the AXI HPM0 LPD at 32. Further configurations were made to the data width of the secondary interface, with the AXI HPC0 and HPC1 explicitly set at 128 and the AXI LPD at 32. These configurations proved essential in maintaining the AXI4 protocol for the DPU architecture, a B4096 model. Such adaptations, aimed at enhancing the system's design, improved the interaction between hardware and software components while optimizing task execution within the system.

1) HARDWARE OPTIMIZATION

DL models exist in various types, and within resource-constrained environments, it is unlikely to create a single hardware parameter that caters to all models. Typically, designs with fewer hardware resources lead to improved FPGA timing, higher clock frequency, increased throughput, and reduced power consumption. For applications similar to ADAS, the presented design provides a viable compromise with the opportunity for customization through fine-tuned parameters. The DPU retrieves instructions from off-chip memory to guide the computing engine's procedure, with the Vitis-AI compiler generating these instructions, which include layer-fusion and optimizations.

This setup uses on-chip memory for input activation, feature maps, and output metadata buffering to maximize

efficiency. Further, software optimization is employed to reuse data as much as possible, minimizing external memory bandwidth requirements. The DPU architecture's computing engine leverages a deep pipeline, and the PEs fully utilize fine-grained building blocks such as multipliers, adders, and accumulators. This methodology maximizes the benefits of the hardware accelerator structure.

It is possible to configure the convolution architectures available within the DPU IP architecture to align with the parallelism of the convolution unit. Architectures vary in PL resource consumption. Larger architectures consume more resources and thus show higher performance. On the other hand, we prefer smaller architectures for our resource-constrained device. When we use minor architecture, we experience a decrease in our device performance. To avoid dropping the performance, we used a Double Data Rate (DDR) approach to improve the performance we get with the DPU. In this formatting, we have specified 1x clock input for general logic and 2x for Digital Signal Processing (DSP) slices.

The cascade length's usage presents a crucial aspect of leveraging DSP. A trade-off between resource usage and timing performance always exists when determining the DSP cascade's size. For instance, deploying a more significant cascade length reduces resource consumption but delivers subpar timing performance. Conversely, opting for a shorter cascade length reduces resource usage and significantly improves timing performance. Hence, in smaller devices with limited logic resources, it is advisable to employ more extensive cascade lengths.

In light of resource usage and timing performance, the study deduced that the DSP's ideal maximum cascade length is four. The DPU IP core's use of DSP elements forms the basis for whether DSP usage is high or low. In instances where DSP element usage is low, multiplication only is performed, whereas high DSP element usage involves both multiplication and accumulation. By setting DSP usage as high in the hardware design, a reduction in the computational density of the quantized DWSC model led to increased timing performance. Moreover, enabling UltraRAM memory in the hardware design allowed for the use of the more significant DPU architecture, given the device's lack of sufficient BRAM, effectively reducing the resource constraint of the development board. Fine-tuning performed on DSP, BRAM, and UltraRAM culminated in a more optimized PL process, yielding timing requirement values at an optimal level. As a consequence, the timing summary revealed values of 1.135ns for *Worst Negative Slack*, 0.001ns for *Worst Hold Slack*, 2.000ns for *Worst Pulse Width Slack* and finally, 0.009ns for *Total Negative Slack*.

D. QUANTIZATION AWARE TRAINING (QAT)

Designers have been tailoring more comprehensive architectures to enhance the performance of CNN models. Such adaptations, including broader and more profound CNN architecture, have successfully reduced the classification error rate for specific problems. Authors of a particular study [39], utilizing various CNN models, exemplified the relationship between computational density and memory requirements in ImageNet classification. They observed a decrease in the ImageNet classification error rate from 17% to 2.9%. Consequently, expanding the network model incurs an increase in computational complexity. This escalation, in turn, leads to a considerable surge in memory requirements. Additionally, bandwidth concerns arise due to the millions of parameters found in CNN models.

Techniques such as model pruning, weight quantization, and activation function quantization [19], [40] aid in reducing computational complexity. Misuse of the quantization method can diminish inference accuracy, while the pruning method can prevent network over-fitting during training. Aiming for a goal-oriented model, the designer needs to balance these trade-offs. A potential pitfall of quantizing CNN model weights and activation functions is data loss, attributed to the inability to restore the floating point after quantization and de-quantization fully. To articulate this issue in mathematical terms;

$$x = f_d(f_q(x, s_x, z_x), s_x, z_x) + \Delta_x \tag{1}$$

where;

f_d and f_q are de-quantization and quantization functions, respectively. Δ_x is an undetermined small value. Suppose $\Delta_x = 0$, the quantized integer models' inference accuracies are the same as those of the floating point models. Unfortunately, this is not the case. The model performs well after training when the model parameters are in FP32 (32 bits floating-point arithmetic). However, setting the precision to int-8 (8 bits integer) or lower can lead to standard inference even if the network is well-trained. In contrast, the quantized network has a much lower memory requirement than the floating point counterpart, resulting in less energy consumption by the system. As a result, the quantized model is more suitable for battery-powered embedded devices.

This study delves into the implementation of real-time ADAS for an FPGA-based MPSoC hardware accelerator by quantizing the ResNet18 model with SSD assets. The weights and activation functions of the model were quantified as int-8 bit low precision integers and a performance comparison of the network was carried out. The PyTorch framework was utilized to construct the model, and the *vai_q_pytorch* library was used to quantize the weights and activation functions. It is noteworthy that *vai_q_pytorch* is a Vitis AI quantizer-supported library that operates on the PyTorch framework.

The Vitis AI [21] quantizer takes in the floating point model, conducts pre-processing, and then quantizes the weights and activation/biases at the specified bit-width. The pre-processing performed by Vitis AI folds the batch normalization and eliminates nodes from the model that are not necessary for inference. Thanks to batch normalization, simultaneous learning is possible across layers in the network. Without batch normalization, the use of a high learning

rate could lead to the issue of disappearing gradients. However, with batch norms, a higher learning rate can be used since alterations in one layer do not impact the others.

Only the initial value is set as QAT is employed since the learning rate value will undergo automatic updates during training. The authors' extensive investigation and explanation of QAT are presented in their work [42], examining its various components and mechanisms. This article is recommended for those desiring a more complete and in-depth understanding of QAT. The detailed information in this work can provide further illumination and enrichment to the reader's comprehension of this specific area of study.

The QAT technique in neural networks strives to minimize the effect of data loss during training, with the inference accuracy of the model experiencing only minimal impact. Given that the weight and activation tensors change during neural network training, a quantization and de-quantization layer can be added for each varying tensor in QAT [41]. Differing from (1), (2) and (3) can be defined in the following manner;

$$\hat{x} = f_d(f_q(x, s_x, z_x), s_x, z_x) \tag{2}$$

$$\hat{x} = s_x(clip(round(\frac{1}{s_x}x + z_x), \alpha_q, \beta_q) - z_x) \tag{3}$$

Data types for quantized tensors are still floating-point tensors. Therefore, we need to train as if there were no quantization layers. In addition, the main problem with QAT is that such quantization layers cannot be differentiated [41]. On the other hand, the straight-through estimation (STE) [43] derivative strategy excels when used for QAT. The identity function in the clipping range $[\alpha, \beta]$ and the constant function outside of the clipping range $[\alpha, \beta]$ are how STE handles the quantization and de-quantization functions. Thus, the resulting derivatives are 1 if $[\alpha, \beta]$ is in the clipping range and 0 if outside of the field. We can define symmetric quantization mathematically as in 4.

$$\frac{\partial \hat{x}}{\partial x} = \begin{cases} 1 & if\ \alpha \le x \le \beta \\ 0 & else \end{cases} \tag{4}$$

Scaling factors can be discovered during QAT thanks to STE. For instance, the Learned Step-Size Quantization (LSQ) [44] is obtained from the scaling elements' gradient quantization function. Starting from 1, we can get 5 to 8 as follows;

$$\begin{aligned} \frac{\partial \hat{x}}{\partial s_x} &= \frac{\partial s_x}{\partial s_x}\left(clip(round\left(\frac{1}{s_x}x, \alpha_q, \beta_q\right)\right) \\ &+ s_x \frac{\partial\left(clip\left(round\left(\frac{1}{s_x}x\right), \alpha_q, \beta_q\right)\right)}{\partial s_x} \end{aligned} \tag{5}$$

$$\begin{aligned} &= clip\left(round\left(\frac{1}{s_x}x\right), \alpha_q, \beta_q\right) \\ &+ s_x \frac{\partial\left(clip\left(round\left(\frac{1}{s_x}x\right), \alpha_q, \beta_q\right)\right)}{\partial s_x} \end{aligned} \tag{6}$$

If we define the numerator part of (5) as any variable (θ) in order not to rewrite it at length; $\theta =$ $clip\left(round\left(\frac{1}{s_x}x\right), \alpha_q, \beta_q\right)$

$$\cong \begin{cases} \theta + s_x \dfrac{\partial\left(\frac{1}{s_x}\right)x}{\partial s_x} & if\ \alpha \le x \le \beta \\ a_q + s_x \dfrac{\partial\left(b_q\right)}{\partial s_x} & if\ x < \alpha \\ b_q + s_x \dfrac{\partial\left(b_q\right)}{\partial s_x} & ifx > \beta \end{cases} \tag{7}$$

$$= \begin{cases} round\left(\dfrac{x}{s_x}\right) - \dfrac{x}{s_x} & if\ \alpha \le x \le \beta \\ a_q & if\ x < \alpha \\ b_q & if\ x > \beta \end{cases} \tag{8}$$

Here, it is possible to learn or adaptively select different bit widths for each layer in a model or a uniform bit width for the entire model in this case. The *vai_q_pytorch* library may quantize the activation functions of the model according to the following equations.

$$QuantReLU_{(x,z_x,y_x,k)} = \begin{cases} z_y & if\ x < z_x \\ z_y + k(x - z_x) & if\ x \ge z_x \end{cases} \tag{9}$$

When $z_x = 0,\ z_y = 0$ and $k = 1$, the generally utilised ReLU in DL models is a particular case of the above description.

$$ReLU_{(x,0,0,1)} = \begin{cases} 0 & if\ x < z_x \\ 1 & if\ x \ge z_x \end{cases} \tag{10}$$

Here, we have given the Mathematically analysis steps of the QuantReLU function.

$$y = ReLU(x, 0, 0, 1) \tag{11}$$

$$= \begin{cases} 0 & if\ x < z_x \\ 1 & if\ x \ge z_x \end{cases} \tag{12}$$

$$= s_y(y_q - z_y) \tag{13}$$

$$\begin{aligned} &= ReLU(s_x(x_q - z_x), 0, 0, 1) \\ &= \begin{cases} 0 & if\ s_x(x_q - z_x) < 0 \\ (s_x(x_q - z_x) & if\ s_x(x_q - z_x) \ge 0 \end{cases} \\ &= \begin{cases} 0 & if\ x_q < z_x \\ s_x(x_q - z_x) & if\ x_q \ge z_x \end{cases} \end{aligned} \tag{14}$$

Consequently;

$$\begin{aligned} y_q &= \begin{cases} z_y & if\ (x_q < z_x \\ z_y + \dfrac{s_x}{s_y}(x_q - z_x) & if\ x_q \ge z_x \end{cases} \\ &= ReLU(x_q, z_x, z_y, \frac{s_x}{s_y}) \end{aligned} \tag{15}$$

Hereby, to achieve the QuantReLU corresponding to the floating-point $y_q = ReLU(x, 0, 0, 1)$, we require to serve;

$$y_q = ReLU(x_q, z_x, z_y, \frac{s_x}{s_y}) \tag{16}$$

where; z_x and z_y are zero points, s is a positive floating-point scale element and x_q is quantized matrices,

TABLE 2. Use of datasets.

Datasets	Train	Validation	Test
BDD100K+Waymo	70000	10000	20000
KITTI	7480	-	7517
CityScapes	2975	500	1525

E. EXPERIMENTAL SETUP

This section details the experimental setup required for the real-time execution of ADAS multi-task learning. The discussion initiates with an explanation of the fundamental operation of the model and the significance of pipeline design. Following this, insights about the datasets used for this study will be shared. The next segment will dive into the process of model training. Concluding the section, an overview of the hardware-software co-design involved in this study will provide a comprehensive understanding of the overall process.

1) SYSTEM SETUP

Our research leverages the capabilities of the AMD Xilinx Zynq UltraScale+ MPSoCs, uniting an FPGA with PL and an ARM processor inside a PS into a cohesive entity. The chosen experimental setup utilizes the Zynq UltraScale+ MPSoC Kria KV260 Vision AI development board from AMD Xilinx, offering 4GB DDR memory due to its compatibility and efficiency in executing the investigated algorithm.

The focus is primarily on the ARM Cortex A53 (PS) and PL. The PS coordinates multiple operations, encompassing monitor connections, pre-processing and post-processing task management, the interface functions oversight, USB interface regulation, and operating system activity control. Simultaneously, PL develops optimized on-chip and off-chip memory access techniques, formulates pipeline strategies, and supports hardware acceleration functions. As depicted in Figure 5, multiple threads are established as pipelines and operated in parallel to maximize efficiency. This pipeline design strategy yields considerable benefits, contributing to a roughly 50% throughput increase, and diminishing design complexity and resource usage, as illustrated in Fig. 6. With each implemented FPGA kernel embodying a single thread, the inherent parallelism within this thread can be fully exploited.

2) DATASETS

We are concentrating our research on improving autonomous driving in driver-operated and driverless vehicles by combining object detection and segmentation in a multi-task learning approach. We trained our model on four publicly available datasets: BDD100K, Waymo, KITTI, and CityScapes. We mostly used the BDD100K+Waymo dataset, known for its various autonomous driving scenes, and KITTI, offering object detection in three separate classes for both object detection and segmentation tasks. For semantic segmentation across 19 categories, we turned to the CityScapes dataset.

Table 2 outlines the dataset distribution we used for training and inference. We merged datasets into common categories to further enhance inference accuracy and efficiency. In particular, we merged the CityScapes and BDD100K datasets for segmentation tasks, while the BDD100K and Waymo datasets were utilized for object detection. The data was portioned for model training, testing, and validation, resulting in highly promising outcomes for multi-task learning. The resulting data subsets have demonstrated highly favourable outcomes for multi-task learning. Evaluating the performance of our model, we employed standard mAP (17) and mIoU (18) criteria. The mAP was used to measure object recognition, while the mIoU was used to evaluate segmentation. The results suggest that consolidating datasets into shared categories significantly improved the inference's accuracy and efficiency in multi-task learning.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \tag{17}$$

here, AP_k is the AP of class k, n is the number of classes,

$$mIoU = \frac{\left(\frac{1}{n_{cl}}\right)\sum_i n_{ii}}{\left(t_i + \sum_j n_{ji} - n_{ii}\right)} \tag{18}$$

where n_{cl} represents the number of classes, t_i is the total number of pixel in class i, n_{ii} represents true positives, n_{ji} false negatives.

3) MODEL TRAINING

The construction of the sophisticated multi-task model necessitated a meticulous selection of loss functions. Furthermore, the extensive capacity of the model and the management of sizable datasets required the use of a high-performance GPU. The model was segmented into several subnets subjected to independent training to alleviate the computational burdens associated with end-to-end training.

The initialization of the shared backbone subnet was carried out during the pre-training phase, which capitalized on the extensive versatility of the ImageNet dataset, recognized for its proficiency in large-scale image classification assignments. This crucial step provided the backbone with meaningful representations for both tasks. The training protocol encompassed several stages. Initially, the semantic segmentation and backbone encoder subnets were rendered passive, followed by the training of the multi-object detection subnet. Afterwards, training was initiated for the semantic segmentation and backbone encoder subnets, achieved by temporarily turning off the object detection subnet. Each subnet was subjected to 100,000 training iterations to ensure thorough learning.

The commencement of the training phase was centred on setting up the general contextual information within the images, designated as weights. This stage was sustained until the loss function could indicate convergence towards a global minimum value. During the training phase, the quantization of weight and activation functions was expedited by applying QAT, aiming to curtail unnecessary quantization tasks. Additional steps included pre-processing measures, such as

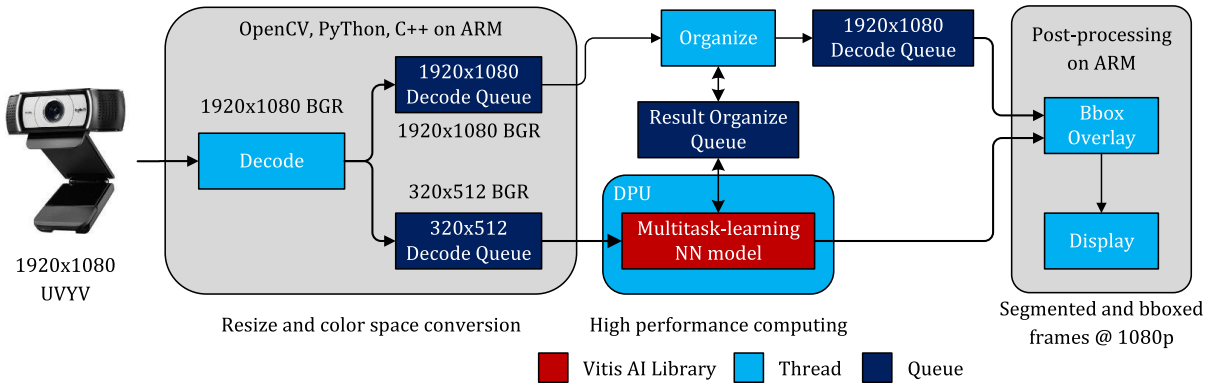


FIGURE 5. Fundamental execution of the ADAS multi-task learning.

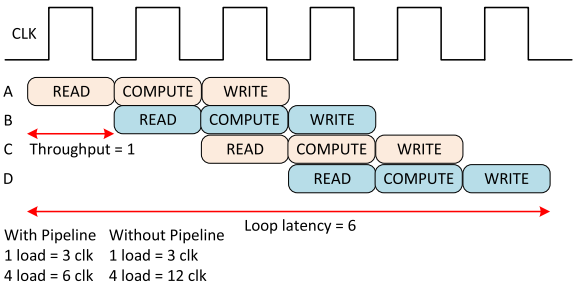


FIGURE 6. Kernel pipeline design.

TABLE 3. Specifications of the computer used in evaluation.

Parameters	Value	Unit
CPU Manufacturer	INTEL	-
CPU Variant	i7-7700HQ	-
CPU Clock Frequency	3.8	GHz
CPU Core Size	8	-
Cache Size	6	MB
RAM Size	32	GB
GPU Manufacturer	NVIDIA	-
GPU Chipset	GTX1080	-
GPU RAM Size	6	GB

adapting the size of the input image, required explicitly for multi-object detection, to align directly with the mesh input size.

The accomplishment of the pre-training phase led to the fine-tuning of the entire multi-tasking model, with task-relevant labelled data incorporated into the process. The parameters of the shared backbone network were updated in parallel with those of the task-specific subnets. The optimization algorithms previously described were utilized to minimize the weighted sum of the loss functions. A series of tests on various hyperparameters, encompassing learning rate, weight decay, and batch size, were conducted to achieve the model's most effective configuration. A validation set was employed to monitor the model's performance.

The objective classification function for the object detection subnet was defined as $focal_loss2d$ and a soft L_1 loss was employed for bounding box regression, thereby tailoring the model specifically for object detection tasks. Notably, $focal_loss2d$ effectively countered the influence of class imbalance during the training phase, whereas the soft L_1 loss function played a role in diminishing the effect of outliers in bounding box regression. Stochastic gradient descent (SGD) was subsequently applied to refine the model, setting a learning rate 1e-5 and a momentum value of 0.9.

In the SSD multi-box configuration, a batch-size ratio of 16 was set, and binary cross-entropy (BCELoss) was utilized as the loss function, guaranteeing a proficient training process. Encoder weights were initialized via a pre-trained ImageNet model for the segmentation subnet. The choice of *LovaszSoftmaxLoss* [45] facilitated pixel-level classification and semantic segmentation tasks. The model's optimization relied on the SGD optimizer, assigning a learning rate of 1-e2. Notably, during training, the batch size for the segmentation subnet was designated as 2.

After training and fine-tuning procedures, the ADAS multi-task learning network was evaluated on the test set, employing relevant metrics for each task. Specifically, mAP was used for object detection, while IoU served the segmentation task. Such assessments aided in approximating the model's effectiveness and efficiency, subsequently offering critical insights for potential enhancements and refinements.

SGD and LovaszSoftmaxLoss can be specified mathematically as in 19, 20 and 21, respectively.

$$\theta = \theta - \eta.\nabla_\theta J(\theta; x^{(i)}; y^{(i)}) \tag{19}$$

$$\nabla = \frac{\partial}{\partial x}\hat{\imath} + \frac{\partial}{\partial y}\hat{\jmath} + \frac{\partial}{\partial k}\hat{z} \tag{20}$$

here, ∇ is the gradient operator, η is the learning rate, $x^{(i)}$ is the training sample, and $y^{(i)}$ denotes the label, respectively.

$$loss(f) = \frac{1}{|C|}\sum_{c\in C}\overline{\Delta_{J_c}}(m(c)) \tag{21}$$

where Δ_{J_c} is Jaccard loss extension, $m(c)$ is the vector of errors and C represents class.

You can find the features of the workstation we use for training in Table 3.

4) HARDWARE - SOFTWARE CO-DESIGN

This research emphasizes the creation of a versatile system input compatible with data from camera or sensor fusion.

Such flexibility supports modifications in input modes via adjustments to the kernel or root file format of the real-time operating system. This requirement is crucial given the pivotal role that machine vision systems serve in the automotive industry, particularly in object detection and semantic segmentation. Consequently, optimizing hardware and software co-designs is critical to meet the demands of throughput and power consumption.

The adopted approach in this study entails a meticulous evaluation of the algorithm, arranging its components based on the time consumption profiles for each process. The formation of this hierarchy was directed by the criteria established within this research. This process facilitated the identification of specific segments of the algorithm where single-instruction multiple-data (SIMD) operations were prevalent. An observed increase in power consumption in certain algorithmic sections was attributed to an escalation in algorithmic latency and the frequency of memory access. Given that, each memory access operation requires energy, an escalation in these operations' frequency directly affects the system's energy consumption. Consequently, power consumption elevates with an increase in the frequency of these accesses, as demonstrated in certain parts of our algorithm. This insight proved vital in identifying the algorithm's energy-intensive areas, optimising our overall system design for enhanced energy efficiency.

Modifications were made to several settings to circumvent potential resource constraints and amplify overall performance. ALU parallelism was set to 8, RAM usage to High, channel augmentation was enabled, and the DSP cascade length was extended to 4. Channel augmentation is optional to boost DPU efficiency, especially when the number of input channels is considerably less than the available channel parallelism. This scenario is frequently seen in numerous CNNs where the input channel of the first layer typically comprises three, failing to utilize the hardware channels optimally. However, even when the number of input channels surpasses channel parallelism, channel augmentation can be advantageous, albeit requiring more logic resources. Despite the related costs, this feature could enhance the efficiency of most CNNs. These optimizations' importance lies in achieving a highly efficient system design.

Table 4 outlines two unique DPU configurations and their respective utilization methods. Both configurations pose credible options for inference tasks, with this research choosing the configuration presented in case 1. Each variable emphasized in the table affects inference and memory consumption in distinctive ways. For instance, activating channel augmentation can improve the overall efficiency for numerous CNNs, although it may result in elevated LUTs consumption, thereby creating obstacles for devices with restricted memory. Moreover, it is critical to acknowledge that LUTs consumption may also vary among different DPU architectures, such as B1152, B3136, and B4096. Performance and memory prerequisites are further influenced by the types of ReLU used in convolution and ALU. A suggested setting for ALU parallelism is 4 for devices with memory restrictions. Nevertheless, this study chose a setting of 8 to cater to performance requirements. This modification, while leading to extra consumption of LUTs, FFs BlockRAMs, and DSPs resources, is often considered negligible when prioritizing performance.

TABLE 4. Performance comparison of two distinct DPU configurations for hardware-software co-design.

DPU Architecture (B4096)	Case 1	Case 2
Channel Augmentation	Enable	Disable
ALU Parallel	8	4
DSP48 Maximal Cascade Length	4	4
RAM & DSP48 Usage	High	High
Convolution ReLU Type	ReLU+ReLU6	ReLU+LeakyReLU+ReLU6
ALU ReLU Type	ReLU+ReLU6	ReLU+ReLU6
Memory Consumption (Average)	Higher	Lower

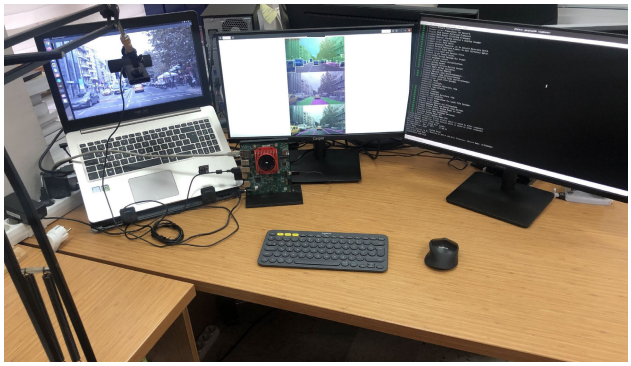


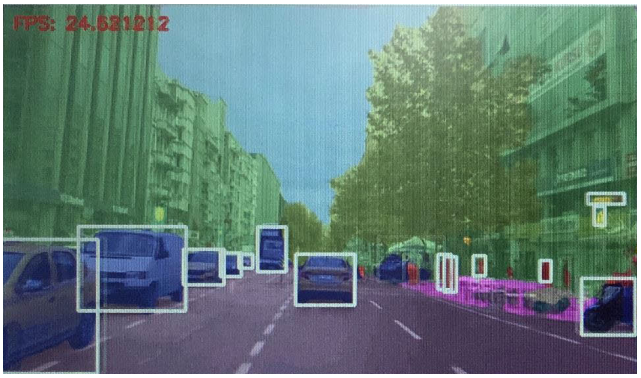
FIGURE 7. A Comprehensive examination of the real-time implementation of ADAS multi-task learning.

TABLE 5. Performance comparison of the studies in terms of specified metrics.

Ref.	Methods	Params.	Power (W)	mAP (%)	mIoU / IoU (%)	GFLOPs
[13]	Multi-task learning	N	12.1	N	57.59	13.6
[28]	Multiple object det.	0.115M	3.118	N	67	N
[48]	Multi-task learning	N	14.3	62.8	57.6	9
[49]	Autonomous Driving	65.2M	9.8	N	N	N
[50]	Multi-task learning	65.2M	7.34	N	N	N
ADAS Multi-task Learning	Multi-task learning	11.4M	7.19	51	Segm.: 56.62 Drivable: 81.56 Line: 43.86	25

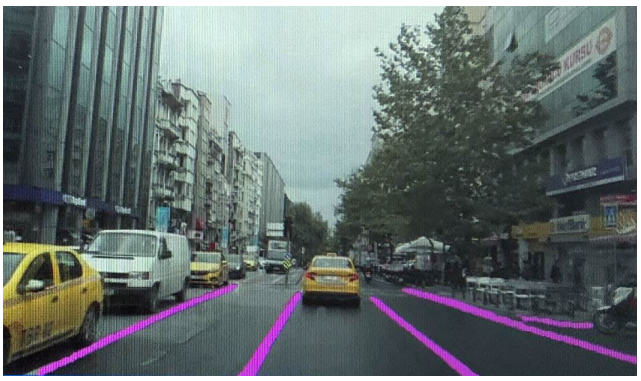
III. RESULTS AND DISCUSSION

This research involves the Kria KV260 development board and the Logitech c930e camera, as illustrated in Fig. 7. The development board obtains a real-time video stream via USB for processing on ARM. DPU enables the PL environment to conduct complex computation and convolution operations,



(a) Multiple object detection

(b) Drivable area detection



(c) Line detection

FIGURE 8. Real-time implementation of ADAS multi-task learning.

which are conveyed to the monitor through an HDMI connection. Performance evaluation was conducted via various ADAS use cases, including object detection, segmentation, line detection, and detection in drivable areas. The real-time results of these specified ADAS tasks are demonstrated in Fig. 8, showing proficient input video data processing.

The platform's accuracy was gauged through the mAP value for object detection and the mIoU value for segmentation and line detection. In contrast, the platform's throughput was evaluated using the FPS value. Analogous investigations employing mAP and mIoU metrics have been presented for reader comprehension. A literature review identified a demand for multi-task ADAS research deploying MPSoC FPGA. The methodologies discussed presently might appear designated for diverse applications; however, their future

TABLE 6. Performance comparison of the study in INT8 and FP32 model types.

ADAS Multi-task Learning	Floating Point (FP32)	Quantization Aware Training (INT8)
Segmentation mIoU (%)	57.95	56.62
Multiple object detection mAP (%)	51.29	51
Drivable Area Detection mIoU (%)	82.63	81.56
Line Detection IoU (%)	43.65	43.86
GFLOPs	6.36	25

incorporation in numerous sectors, notably those involving autonomous vehicle technology and ADAS, is virtually inevitable. Notably, within the MPSoC-FPGA environment, employing hardware-software co-design is poised to yield enhanced outcomes, especially when software-accelerated techniques are reinforced by hardware. Integrating hardware and software in such a harmonized approach can significantly leverage system performance, fostering advancements in the forthcoming era of autonomous and assisted driving technologies.

In Table 5, we have reviewed the literature encompassing our specified criteria. As discernible from the Table 5, the power consumption, the number of tasks, and the performance (GOPs) we recommend surpass those of the other studies. Notably, although operation at [28] appears optimal regarding power consumption, it only carries out the multi-object detection task. The study closest to ours was conducted at [47], where the researchers undertook multiple object detection and segmentation tasks. Our investigation indicates that their study's mAP and mIoU values are satisfactory. However, their power consumption exceeds ours, and they perform fewer tasks. Thus, as demonstrated in the table, our ADAS multi-task learning stands out in terms of both the number of functions and the evaluation outcomes.

The performance evaluation results for object detection using the specified dataset showed 51% mAP, indicating that the Kria KV260 Vision AI platform can accurately detect objects in real time. For segmentation, the platform achieved 56.62% mIoU, demonstrating its ability to segment objects accurately in complex scenarios. In line detection, the platform reached 43.86% IoU, indicating its ability to detect lines in the environment accurately. In seeing the derivable area, the platform also achieved 81.56% mIoU, demonstrating its ability to detect the derivable location accurately. Furthermore, the platform reached a throughput of 25.4 FPS at the optimized + pipeline design, indicating its real-time ability to process multiple ADAS use cases.

To assess the performance of ADAS multi-task learning on the Kria KV260 Vision AI Starter Kit Board, we conducted a comparative study between two precision, FP32 and QAT-INT8, as illustrated in Table 6. This comparison aimed to gauge how precision influences model performance concerning computational metrics (FLOPs) and task-specific outcomes.

TABLE 7. Performance comparison of DPU with and without optimized.

Thread size	Optimized + Pipeline			None-Optimized		
	DPU run-time (ms)	ARM (PEs) average run-time (ms)	FPS	DPU run-time (ms)	ARM (PEs) average run-time (ms)	FPS
t - 1	33.551	146.874	10.5	33.901	153.784	8.6
t - 2	34.095	156.342	19	34.194	165.864	15
t - 3	34.788	176.1	22	34.491	186.78	18.86
t - 4	34.839	220.5	25.4	34.766	209.687	21
Memory footprint	62.86 MB			69 MB		

TABLE 8. Resource consumption of hardware architectures used in similar studies.

ADAS multi-task learning				[48]			[49]		
Resource	Available	Utilization	Utilization (%)	Available	Utilization	Utilization (%)	Available	Utilization	Utilization (%)
LUTs	117120	63445	54.17	70560	39772	56.37	274080	75000	27.3
LUTRAM	57600	7180	12.47	28800	3650	12.67	144000	N	N
FFs	234240	112272	47.93	141120	59045	41.84	548160	146000	26.7
BRAM	144	135	93.75	216	123	56.94	912	280	30.7
URAM	64	48	75	N			0	N	
DSP	1248	774	62.02	360	211	58.61	2520	N	
BUFG	352	6	1.7	196	3	1.53	NA	N	
Platform	XCK26			XCZU3EG			ZU9		
DPU Architecture	1 x B4096			1 x B1152F			2 x B4096		
Frequency (MHz)	300			525			600		

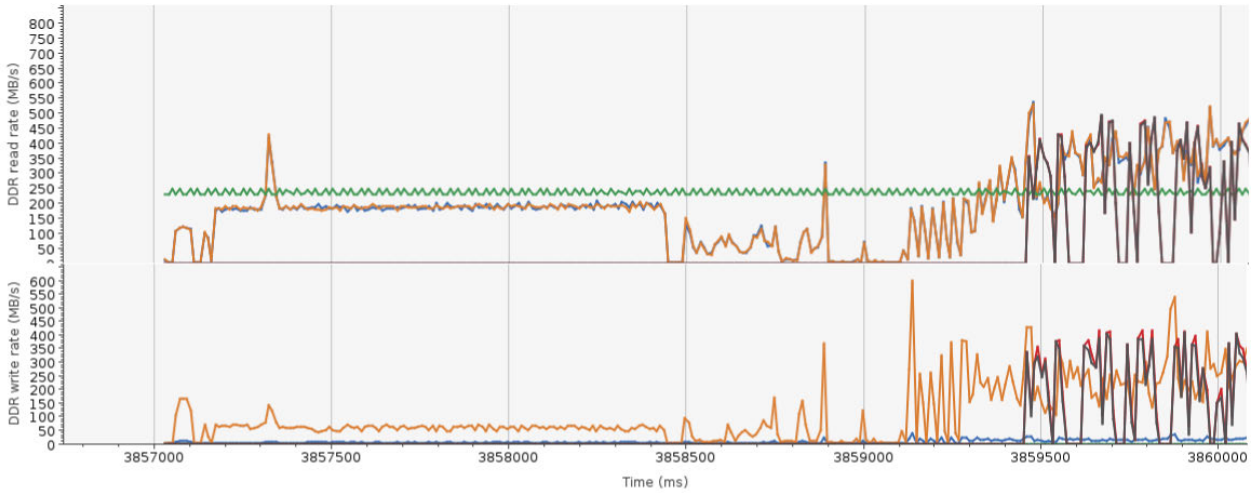


FIGURE 9. Performance evaluation of DDR ports read/write in Vitis AI analyzer.

The FP32 model delivered 6.36 GFLOPs for the computational metrics, while the INT8 model attained nearly 25 GFLOPs. This boost in computational speed for the INT8 model is anticipated due to the reduced numerical precision, which leads to a lesser computational complexity and memory usage. Consequently, it enables more operations to be performed every second, resulting in a higher FLOPs value for the INT8 model. Alongside this, a minor trade-off between precision and task-specific performance outcomes is noticed, with the FP32 model showing slightly improved performance in segmentation and derivable area detection tasks, owing to its higher numerical precision.

Conversely, the INT8 model showed comparable or slightly superior performance in object and line detection tasks. Thus, the selection between FP32 and INT8 hinges on the application's specific requirements. FP32 may be preferable when the highest accuracy is a priority and computational resources are not a limiting factor. However, the INT8 model is a viable alternative for scenarios prioritizing computational efficiency and speed, still delivering competitive performance. The optimal balance depends on the specific constraints and requirements of the application.

As can be seen from Table 6, we preferred QAT-INT8 and FP32 for comparison. There are methods for quantization, including post-training quantization (PTQ) (also called direct-quantization) and QAT. The memory footprint after quantization is similar in both methods. The main difference lies in the performance of the model after quantization. Our

model that we trained with QAT was typically more resilient to the effects of quantization and outperformed a model we quantized with PTQ (i.e., it produced more accurate predictions) given the same amount of memory. QAT aims to reduce the accuracy disruption caused by the quantization process, but the process is more time-consuming and computationally expensive than PTQ.

We utilized multi-threading to improve the study's throughput and observed a significant performance improvement as in Table 7. We followed the performance variation of the DPU across varying thread sizes. It is well-known that the DPU we utilized has a maximum thread limit of 4. While employing a vast number of threads enhances the performance, it also significantly increases the DPU run-time value. Thus, we can elaborate on a trade-off between the thread size and run time. We also incorporated a DPU pipeline to enhance the performance of CNNs processing on the FPGA fabric, resulting in further progress in the platform's throughput. We use the Vitis AI analyzer tool to measure the platform's performance inference for all allocation threads and tasks, as shown in Fig. 9. In this context, every color denotes the speeds of read and write operations in megabytes per second (MB/s) for five distinct DDR ports.

Additionally, the platform had a low memory footprint, indicating its efficiency in memory utilization. As a result, the Kria KV260 Vision AI platform delivers high accuracy and throughput while maintaining low power consumption and memory footprint. Furthermore, the platform's multi-task implementation and multi-class object detection capabilities allow it to process complex ADAS use cases. Our ADAS multi-task learning successfully integrated a complex and large model into a development board thanks to the hardware and software optimizations. Table 8 depicts the resource usage of similar analyses. We offered a glimmer of hope for resource-constrained devices by multi-tasking on a single development board. Our study incorporated the B4096 DPU architecture provided by AMD Xilinx, resulting in maximum efficiency at low frequencies and with limited resource usage. Our inference success and resource usage are commendable compared to the other two studies. Our hardware and software optimizations allowed for optimal utilization of the development board, creating sufficient resource space to include various ADAS tasks. Overall, our study is a cost-effective and efficient ADAS research solution that can be deployed in real-world applications with minimal modifications.

It is imperative to note that future investigations will not remain restricted to CNNs and DNNs. Despite the extensive applicability of these networks, there is an evolving trajectory towards more simplified structures in artificial intelligence, which could yield superior results compared to the highly effective structures comprising convolutional layers, such as DNNs and CNNs.

Indeed, relentless technological advancement has enabled current models to offer plausible solutions to contemporary problems. However, the escalating complexity of these problems necessitates formulating novel, diverse structures. This concept is well-exemplified in the research indicated in [50]. Contrary to the traditional CNN methodology, this study introduces the Physics Informed DL model. This paradigm can be characterized as a physics-knowledge-informed deep learning framework, wherein physics-based domain knowledge is assimilated into the data-driven model as soft constraints. These constraints serve to guide and adjust the data-driven model.

In parallel, the research in [51] elucidates Extreme Learning Machines (ELMs) as an alternative to traditional deep learning methodologies, which typically encompass Deep Belief Networks and Constrained Boltzmann Machines. This approach streamlines the training process, a phase typically protracted by the intricate fine-tuning of numerous parameters and the complexity of the hierarchical structure. ELMs achieve this through a non-iterative, rapid training process facilitated by a random feature-matching mechanism.

IV. CONCLUSION AND FUTURE WORK

This research proposes a highly effective and efficient approach for executing multi-task ADAS on an MPSoC-FPGA. The solution focuses on detecting multiple objects, lane identification, drivable area detection, and semantic segmentation while maintaining minimal power consumption. The method encompasses a range of software and hardware enhancements, including integrating multiple models and a unified learning algorithm, resulting in a remarkable 9% reduction in memory usage. Hardware optimizations like parallelization and pipeline architecture were leveraged to enhance efficiency, compatibility, and speed. The holistic optimizations improved accuracy, superior performance, and reduced energy consumption, offering a promising pathway for developing embedded systems.

Notably, our research approach employed a single B4096 DPU, a measure that substantially reduced resource consumption compared to prior research endeavours. This strategy culminated in our system achieving an energy consumption rate of 7.19w, an FPS value of 25.4, a memory footprint of nearly 62.86MB, a multi-object detection rate of 51% mAP, a segmentation rate of 56.62% mIoU, a drivable area detection rate of 81.56% mIoU, and a line detection rate of 43.86% IoU. Given the data and results, ADAS multi-task learning offers an effective, efficient, sustainable, and precise system design for real-time ADAS applications. Furthermore, with its low power consumption, cost-effectiveness, and compact design, this system presents a compelling solution for real-world applications, as the experimental results demonstrate the proposed method's feasibility for conducting real-time processing in low-power embedded devices for on-road testing.

ADAS plays a vital role in modern automotive safety and convenience by addressing imminent challenges by integrating sensors, complex algorithms, and advanced computing resources. The research highlights the significant potential of utilizing MPSoC FPGAs, with their parallel computing and reconfigurable logic capabilities, in achieving successful

ADAS deployment. MPSoC FPGAs offer critical elements like parallel computing and pipeline architectures that have the potential to enhance the performance of ADAS applications significantly. By leveraging PL and PEs, custom hardware accelerators such as DPUs can be created, enabling fast and low-latency image and video data processing. Integrating advanced software frameworks and libraries like OpenCV, Python, and C++ further simplifies the deployment of complex algorithms required for successful ADAS applications. Another avenue to improve ADAS efficacy is through multi-task learning, where a single model is trained to simultaneously perform various tasks, including object and lane detection. However, this approach introduces challenges related to memory allocation and resource management. To address memory and resource constraints, quantized aware training can be employed, producing compact and efficient models with minimal performance degradation. Nonetheless, optimizing quantized models requires careful consideration of trade-offs between accuracy, performance, and memory consumption.

Despite significant progress in developing ADAS systems with MPSoC FPGAs and other computing resources, specific challenges still need to be addressed. Integrating ADAS with advanced methods like autonomous driving necessitates heightened reliability, safety, and security. Moreover, ADAS systems operating in harsh environments require specialized hardware and software architectures. While MPSoC FPGAs, parallel computing architectures, and software frameworks like OpenCV and Python promise efficient and high-performance ADAS systems, addressing these challenges is crucial for widespread adoption and successful deployment across diverse applications.

Our future research aims to explore how innovative technologies, such as advanced LIDAR or RADAR systems, can be integrated into the multi-task learning approach of ADAS. We also plan to incorporate refined computer vision and ML algorithms to enhance object detection and tracking accuracy. Additionally, we envision advancing the decision-making process in ADAS by leveraging Vehicle-to-Everything (V2X) data to coordinate activities across multiple vehicles, thereby enhancing road safety. Furthermore, we are exploring the potential impact and integrative possibilities of emerging technologies like quantum computing and neuromorphic computing, which may revolutionize ADAS. These innovative concepts form the basis of our future research plans as we explore their potential and transformative effects in ADAS systems.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[7] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1520–1528, doi: 10.1109/ICCV.2015.178.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[9] D. Neven, B. D. Brabandere, S. Georgoulis, M. Proesmans, and L. V. Gool, "Towards end-to-end lane detection: An instance segmentation approach," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Changshu, China, Jun. 2018, pp. 286–291, doi: 10.1109/IVS.2018.8500547.

[10] S. Lee, J. Kim, J. S. Yoon, S. Shin, O. Bailo, N. Kim, T.-H. Lee, H. S. Hong, S.-H. Han, and I. S. Kweon, "VPGNet: Vanishing point guided network for lane and road marking detection and recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1965–1973, doi: 10.1109/ICCV.2017.215.

[11] Ò. Lorente, I. Riera, and A. Rana, "Scene understanding for autonomous driving," 2021, *arXiv:2105.04905*.

[12] M. Teichmann, M. Weber, M. Zöllner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Changshu, China, Jun. 2018, pp. 1013–1020, doi: 10.1109/IVS.2018.8500504.

[13] J. Peng, L. Tian, X. Jia, H. Guo, Y. Xu, D. Xie, H. Luo, Y. Shan, and Y. Wang, "Multi-task ADAS system on FPGA," in *Proc. IEEE Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Hsinchu, Taiwan, Mar. 2019, pp. 171–174, doi: 10.1109/AICAS.2019.8771615.

[14] B. P. Sanjay. (2022). *Deep Learning Part 3/4, Medium*. Accessed: Jul. 25, 2022. [Online]. Available: https://medium.com/my-aiml/deep-learning-part-3-4-5c1392ecbc17

[15] P. Jawandhiya, "Hardware design for machine learning," *Int. J. Artif. Intell. Appl.*, vol. 9, no. 1, pp. 63–84, Jan. 2018, doi: 10.5121/ijaia.2018.9105.

[16] J. Borrego-Carazo, D. Castells-Rufas, E. Biempica, and J. Carrabina, "Resource-constrained machine learning for ADAS: A systematic review," *IEEE Access*, vol. 8, pp. 40573–40598, 2020, doi: 10.1109/ACCESS.2020.2976513.

[17] K. S. Zaman, M. B. I. Reaz, S. H. Md Ali, A. A. A. Bakar, and M. E. H. Chowdhury, "Custom hardware architectures for deep learning on portable devices: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6068–6088, Nov. 2022, doi: 10.1109/TNNLS.2021.3082304.

[18] M. Lebedev and P. Belecky, "A survey of open-source tools for FPGA-based inference of artificial neural networks," in *Proc. Ivannikov Memorial Workshop (IVMEM)*, Sep. 2021, pp. 50–56, doi: 10.1109/IVMEM53963.2021.00015.

[19] G. Tatar, S. Bayar, and I. Cicek, "Hardware acceleration of FIR filter implementation on Zynq SoC," in *Proc. IEEE 16th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Washington, DC, USA, Oct. 2022, pp. 1–6, doi: 10.1109/AICT55583.2022.10013522.

[20] G. Tatar, S. Bayar, and I. Cicek, "Performance evaluation of low-precision quantized LeNet and ConvNet neural networks," in *Proc. Int. Conf. Innov. Intell. Syst. Appl. (INISTA)*, Biarritz, France, Aug. 2022, pp. 1–6, doi: 10.1109/INISTA55318.2022.9894261.

[21] Vitis AI. *Xilinx*. Accessed: Jul. 15, 2022. [Online]. Available: https://www.xilinx.com/products/design-tools/vitis/vitis-ai.html

[22] M. R. Rani, M. Z. C. Mustafar, N. H. F. Ismail, M. S. F. Mansor, and Z. Zainuddin, "Road peculiarities detection using deep learning for vehicle vision system," *IOP Conf., Mater. Sci. Eng.*, vol. 1068, no. 1, 2021, Art. no. 012001, Accessed: Jul. 25, 2023. [Online]. Available: https://www.academia.edu/50181633/Road_Peculiarities_Detection_using_Deep_Learning_for_Vehicle_Vision_System

[23] T. Almeida, B. Lourenço, and V. Santos, "Road detection based on simultaneous deep learning approaches," *Robot. Auton. Syst.*, vol. 133, Nov. 2020, Art. no. 103605.
[24] A. Hernández, S. Woo, H. Corrales, I. Parra, E. Kim, D. F. Llorca, and M. A. Sotelo, "3D-DEEP: 3-dimensional deep-learning based on elevation patterns for road scene interpretation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Las Vegas, NV, USA, Oct. 2020, pp. 892–898, doi: 10.1109/IV47402.2020.9304601.
[25] M.-A. Andrei, C.-A. Boiangiu, N. Tarbă, and M.-L. Voncilă, "Robust lane detection and tracking algorithm for steering assist systems," *Machines*, vol. 10, no. 1, p. 10, Dec. 2021, doi: 10.3390/machines10010010.
[26] Y. Chen, Z. Xiang, and W. Du, "Improving lane detection with adaptive homography prediction," *Vis. Comput.*, vol. 39, no. 2, pp. 581–595, Jan. 2022, doi: 10.1007/s00371-021-02358-1.
[27] A. Ghorbel, N. B. Amor, and M. Abid, "GPGPU-based parallel computing of Viola and Jones eyes detection algorithm to drive an intelligent wheelchair," *J. Signal Process. Syst.*, vol. 94, no. 12, pp. 1365–1379, Jul. 2022, doi: 10.1007/s11265-022-01783-2.
[28] M. Machura, M. Danilowicz, and T. Kryjak, "Embedded object detection with custom LittleNet, FINN and vitis AI DCNN accelerators," *J. Low Power Electron. Appl.*, vol. 12, no. 2, p. 30, May 2022, doi: 10.3390/jlpea12020030.
[29] N. Sharma and R. D. Garg, "Cost reduction for advanced driver assistance systems through hardware downscaling and deep learning," *Syst. Eng.*, vol. 25, no. 2, pp. 133–143, Nov. 2021, doi: 10.1002/sys.21606.
[30] E. Güney, C. Bayilmis, and B. Çakan, "An implementation of real-time traffic signs and road objects detection based on mobile GPU platforms," *IEEE Access*, vol. 10, pp. 86191–86203, 2022, doi: 10.1109/ACCESS.2022.3198954.
[31] H.-Y. Han, Y.-C. Chen, P.-Y. Hsiao, and L.-C. Fu, "Using channel-wise attention for deep CNN based real-time semantic segmentation with class-aware edge information," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1041–1051, Feb. 2021, doi: 10.1109/TITS.2019.2962094.
[32] M. A. Farooq, P. Corcoran, C. Rotariu, and W. Shariff, "Object detection in thermal spectrum for advanced driver-assistance systems (ADAS)," *IEEE Access*, vol. 9, pp. 156465–156481, 2021, doi: 10.1109/ACCESS.2021.3129150.
[33] J. Cho, Y. Kim, H. Jung, C. Oh, J. Youn, and K. Sohn, "Multi-task self-supervised visual representation learning for monocular road segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Diego, CA, USA, Jul. 2018, pp. 1–6, doi: 10.1109/ICME.2018.8486472.
[34] S. Krishnan, Z. Wan, K. Bhardwaj, P. Whatmough, A. Faust, S. Neuman, G.-Y. Wei, D. Brooks, and V. J. Reddi, "Automatic domain-specific SoC design for autonomous unmanned aerial vehicles," in *Proc. 55th IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Chicago, IL, USA, Oct. 2022, pp. 300–317, doi: 10.1109/MICRO56248.2022.00033.
[35] C.-Y. Lai, B.-X. Wu, V. M. Shivanna, and J.-I. Guo, "MTSAN: Multi-task semantic attention network for ADAS applications," *IEEE Access*, vol. 9, pp. 50700–50714, 2021, doi: 10.1109/ACCESS.2021.3068991.
[36] (2017). *Hardware options for Machine/Deep Learning | MS&E 238 Blog*. Accessed: Jul. 25, 2023. [Online]. Available: https://mse238blog.stanford.edu/2017/07/gnakhare/hardware-options-for-machinedeep-learning/
[37] R. Chen, T. Wu, Y. Zheng, and M. Ling, "MLoF: Machine learning accelerators for the low-cost FPGA platforms," *Appl. Sci.*, vol. 12, no. 1, p. 89, Dec. 2021, doi: 10.3390/app12010089.
[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.
[39] C. Wu, M. Wang, X. Chu, K. Wang, and L. He, "Low-precision floating-point arithmetic for high-performance FPGA-based CNN acceleration," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 15, no. 1, pp. 1–21, Mar. 2022, doi: 10.1145/3474597.
[40] M. P. Véstias, R. P. Duarte, J. T. de Sousa, and H. C. Neto, "A fast and scalable architecture to run convolutional neural networks in low density FPGAs," *Microprocessors Microsyst.*, vol. 77, Sep. 2020, Art. no. 103136, doi: 10.1016/j.micpro.2020.103136.
[41] L. Mao. *Quantization for Neural Networks*. Accessed: Jul. 25, 2023. [Online]. Available: https://leimao.github.io/article/Neural-Networks-Quantization/
[42] P.-E. Novac, G. B. Hacene, A. Pegatoquet, B. Miramond, and V. Gripon, "Quantization and deployment of deep neural networks on microcontrollers," *Sensors*, vol. 21, no. 9, p. 2984, Apr. 2021, doi: 10.3390/s21092984.
[43] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.
[44] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," 2019, *arXiv:1902.08153*.
[45] M. Berman, A. R. Triki, and M. B. Blaschko, "The Lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4413–4421, doi: 10.1109/CVPR.2018.00464.
[46] J. Wang and S. Gu, "FPGA implementation of object detection accelerator based on vitis-AI," in *Proc. 11th Int. Conf. Inf. Sci. Technol. (ICIST)*, Chengdu, China, May 2021, pp. 571–577, doi: 10.1109/ICIST52614.2021.9440554.
[47] S. Fang, L. Tian, J. Wang, S. Liang, D. Xie, Z. Chen, L. Sui, Q. Yu, X. Sun, Y. Shan, and Y. Wang, "Real-time object detection and semantic segmentation hardware system with deep learning networks," in *Proc. Int. Conf. Field-Program. Technol. (FPT)*, Naha, Japan, Dec. 2018, pp. 389–392, doi: 10.1109/FPT.2018.00081.
[48] A. Kojima and Y. Osawa, "Design and implementation of autonomous driving robot car using SoC FPGA," in *Proc. Int. Conf. Field-Program. Technol. (ICFPT)*, Tianjin, China, Dec. 2019, pp. 441–444, doi: 10.1109/ICFPT47387.2019.00088.
[49] S. Kalapothas, G. Flamis, and P. Kitsos, "Efficient edge-AI application deployment for FPGAs," *Information*, vol. 13, no. 6, p. 279, May 2022, doi: 10.3390/info13060279.
[50] J. Zhang, Y. Zhao, F. Shone, Z. Li, A. F. Frangi, S. Q. Xie, and Z.-Q. Zhang, "Physics-informed deep learning for musculoskeletal modeling: Predicting muscle forces and joint kinematics from surface EMG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 484–493, 2023, doi: 10.1109/TNSRE.2022.3226860.
[51] J. Zhang, Y. Li, W. Xiao, and Z. Zhang, "Non-iterative and fast deep learning: Multilayer extreme learning machines," *J. Franklin Inst.*, vol. 357, no. 13, pp. 8925–8955, Sep. 2020, doi: 10.1016/j.jfranklin.2020.04.033.

GUNER TATAR (Member, IEEE) was born in Kahramanmaras, Elbistan. He received the degree, in 2007, the B.S. degree from the Electronics and Communication Department, Marmara University, in 2014, and the master's degree from Marmara University, in 2017, where he is currently pursuing the Ph.D. degree in EEE (English). Following graduation, he worked for a year as a scholarship student in developing biomedical imaging and diagnostic systems infrastructure, financially supported by the Ministry of Development, in 2017. Since then, he has been a Research Assistant with the Department of EEE, Fatih Sultan Mehmet Vakif University. His research interests include reconfigurable computing, dynamic and partial reconfiguration of AMD Xilinx FPGA, multiprocessors, embedded multicore architecture, deep learning, and driver assistant systems.

SALIH BAYAR (Member, IEEE) received the B.S. degree in electronics and communication engineering from Yıldız Technical University, Istanbul, Turkey, in 2003, the M.S. degree in electrical engineering and information technology specialization in systems engineering from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2007, and the Ph.D. degree from the Department of Computer Engineering, Boğaziçi University, Istanbul. He was a Research Assistant with the Department of Computer Engineering, Boğaziçi University, between 2007 and 2013. He was a research and development engineer and the manager, from 2013 to 2017, in a leading software company in Istanbul. Since 2017, he has been an Assistant Professor with the Electrical and Electronics Department, Marmara University, Istanbul. His main research interests include parallel computing, machine learning, image processing, FPGAs, multi-processor, and embedded multi-core architectures.