Image Signal Processing in the Context of Deep Learning Applications

[1]Ali-Khusein and [2]Urquhart
[1,2] First Moscow State University, Russia.
[1] alikhusein60@gmail.com

Correspondence should be addressed to Ali-Khusein : alikhusein60@gmail.com.

Article Info
Journal of Computing and Natural Science (http://anapub.co.ke/journals/jcns/jcns.html)
Doi: https://doi.org/10.53759/181X/JCNS202404002
Received 05 January 2023; Revised from 25 March 2023; Accepted 25 June 2023.
Available online 05 January 2024.

Abstract – Deep learning accelerators are a specialized sort of hardware architecture designed to enhance the computational efficiency of computers engaged in deep neural networks (DNNs) training. The implementation of DNNs in embedded vision applications might potentially be facilitated by the integration of energy-effective accelerators of deep learning into sensors. The lack of recognition for their significant impact on accuracy is a notable oversight. In previous iterations of deep learning accelerators integrated inside sensors, a common approach was bypassing the image signal processor (ISP). This deviation from the traditional vision pipelines had a detrimental impact on the performance of machine learning models trained on data that had undergone post-ISP processing. In this study, we establish a set of energy-efficient techniques that allow ISP to maximize their advantages while also limiting the covariate shift between the target dataset (RAW images) and the training dataset (ISP-analyzed images). This approach enables the practical use of in-sensor accelerators. To clarify, our results do not minimize the relevance of in-sensor accelerators. Instead, we highlight deficiencies in the methodology used in prior research and propose methodologies that empower in-sensor accelerators to fully exploit their capabilities.

Keywords – Image Signal Processor, Digital Signal Processor, Image Processing Unit, Deep Neural Networks.

I. INTRODUCTION

Deep learning is now a widely discussed subject within the field of machine learning (ML). The backdrop consists of several hierarchical levels of neural networks. The deep learning technique is used to apply nonlinear model modifications and high-level model abstractions in large datasets. The latest advancements in deep learning architectures have significantly contributed to the progress of several domains within the field of artificial intelligence. The use of comparable procedures seen in machine learning models is also prevalent in deep learning models. The deep learning workflow by Zhou et al. [1] demonstrates the use of three key processing phases: data preprocessing and analysis, DL model training and building, and interpretation and validation. This approach is designed to effectively tackle real-world difficulties. In the context of feature extraction, it is worth noting that the deep learning (DL) model diverges from traditional machine learning (ML) modeling due to its fully automated nature, and its process has been illustrated in **Fig 1**. Several widely used ML approaches include support vector machines, k-nearest neighbor analysis, random forests, decision trees, linear repression, naïve Bayes, k-means clustering, and association rules.

An image processor, known as an image processing unit (IPU), image signal processor (ISP), or image processing engine (IPE), is considered a dedicated DSP (digital signal processor) or a multi-media processor [2] utilized for the processing of images. It is often found in digital cameras; however, it may also be present in other devices on occasion. The significance of the image processor's speed is growing due to the continuous rise in pixel count of image sensors. Photographers desire to avoid delays caused by the camera's image processor, allowing them to seamlessly continue shooting without perceiving any processing activity. As a result, there is a need for further refinement of image processors to effectively manage larger volumes of data within the same or perhaps reduced time frame.

The current prevailing paradigm in machine learning algorithms, often referred to as deep learning, has achieved exceptional performance in several domains of application, establishing itself as the forefront of technological advancement. According to Zhang, Cheng, Zhang, Liu, and Wei [3], the poor integration of deep neural networks (DNNs) in embedded systems might be attributed to their substantial power, processing, and memory demands. The use of vision applications allows for the simultaneous execution of basic analytical methods, hence creating an optimal setting for the attainment of energy efficiency. Several accelerators, both in-sensor and near-sensor, use the parallel readout capabilities of image sensors to effectively execute convolution operations. Within the conventional imaging pipeline, characterized by its power-intensive nature and high level of precision, an ISP does a multitude of non-linear processing activities on images prior to its subsequent analysis. Nevertheless, contemporary in-sensor accelerators evade ISP, so affecting the imaging procedure

and diminishing the quality of the captured images (refer to **Fig 2**). This study examines the implications of the dataflow architecture choice and proposes an alternative solution.

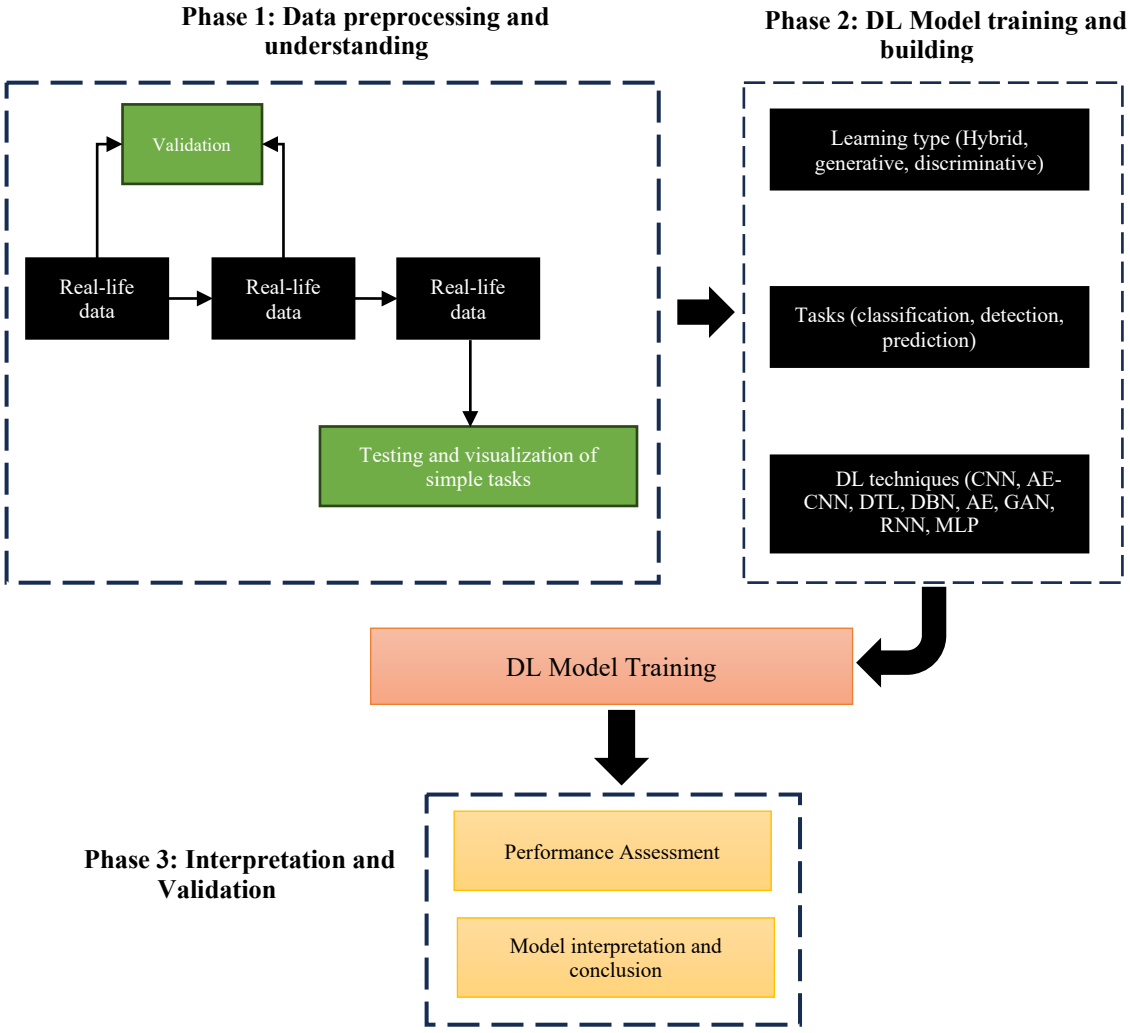


Fig 1. Process of deep learning (DL) and other machine learning (ML) techniques

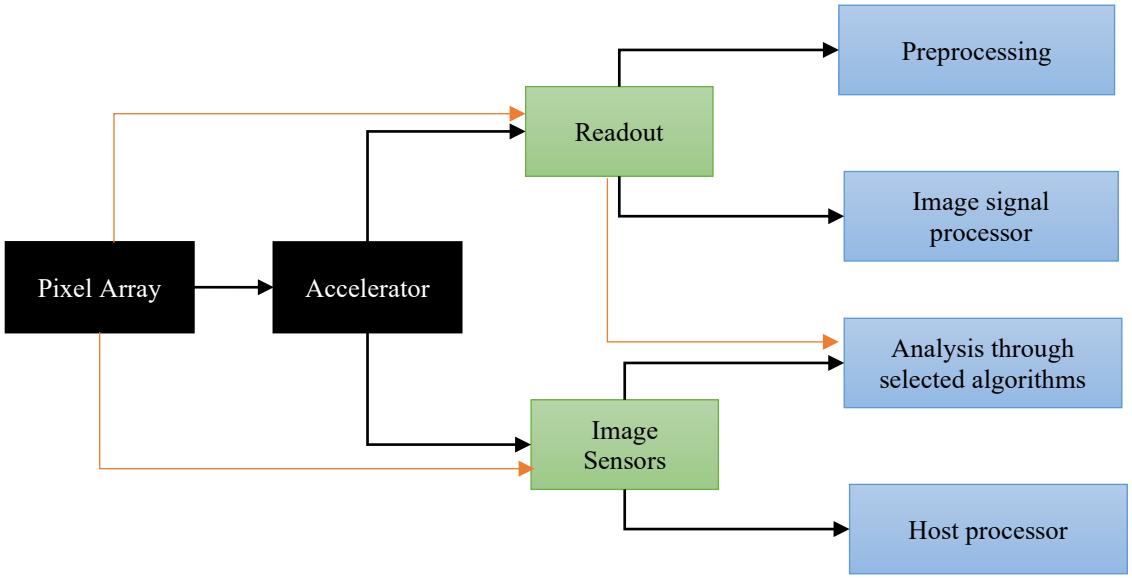


Fig 2. Flowchart contrasting the traditional visioning pipeline (shown by black arrows) with the in-sensor accelerators (shown by pick arrows).

Specifically, it has been observed that the incorporation of pre-ISP training accelerators into an image sensor presents a challenge in terms of deployment. The distributions acquired by deep neural networks (DNNs) vary from the RAW photographs produced by the image sensor because to the common practice of training DNNs on images that have previously undergone processing by the ISP. The phenomenon known as covariate shift refers to the issue arising from disparate distributions between the training data, consisting of ISP processed images, and the target data, including RAW photographs.

Our investigation revealed that this discrepancy resulted in a significant decrease of 60% in accuracy for in-sensor accelerator pipelines. One potential solution is to use pre-ISP data, namely RAW images, for the purpose of training your neural networks. Nevertheless, accomplishing this task would need either generating new extensive datasets of unprocessed photographs, which is a financially demanding operation, or compromising by using a reduced amount of training data, thereby diminishing the level of accuracy. Our proposed approach offers an alternate solution by allowing in-sensor accelerators to use commercially available deep neural networks (DNNs) trained on images processed by image signal processors (ISPs). Additionally, it allows training using large-scale contemporary datasets, while maintaining high accuracy levels.

The following paragraphs outline some key contributions of this research endeavor. The research presents a straightforward image signal pre-processing pipeline, which boosts the accuracy of the in-sensor accelerators that use deep neural networks trained on standard datasets processed by in-camera image signal processors. This is achieved by the implementation of pixel binning and gamma compression techniques on the captured image. The implementation of necessary local transformations inside an image sensor may be easily achieved, therefore obviating the need for an ISP and facilitating the integration of in-sensor accelerators. The proposed methodology demonstrates enhancements in detection accuracies ranging from 25% to 60%, reductions in response times by 34%, and energy savings of 30% when compared to systems that maintain the ISP in close proximity to the sensor. The rest of the paper is organized as follows: Section II presents a discussion of machine vision pipeline, while Section III, a literature review of the concept is critically presented. Section IV focusses on the results and discussion of identifying the required pipeline operations, and the accuracy detection of RAW, proposed, and ISP data. Lastly, Section V presents final remarks regarding the research.

II. THE MACHINE VISION PIPELINE

The traditional machine vision pipeline consists of the three steps listed in **Table 1** below.

Table 1. Steps in traditional machine visioning

Sensing	The image sensor utilizes photodiodes to convert light into electrical impulses that are similar to light intensity. The discontinuous look of RAW photographs may be attributed to the use of color filter arrays by the image sensor, which enables the separate collection of red, green, and blue intensities (refer to **Fig 3(a)**). It should be noted that the proposed procedures we have provided need only small modifications during the readout process in order to be applicable to different color filter array designs.
Preprocessing	The ISP assumes the responsibility for doing the preprocessing tasks. These methods are often limited to manufacturers and are specifically devised to augment the aesthetic appeal of the ultimate product. However, it is worth noting that there are some shared procedures inside the pipelines, one of which is demosaicing. This particular process involves the conversion of a raw image, composed of separate channels, into a cohesive color composite. Finally, the RAW images undergo a conversion process to a widely used format such as JPEG or PNG.
Inference	In the context of operations such as detection and classification, specialized algorithms designed for particular applications are executed on a central processing unit (CPU) of a host system. The pipeline is modified by using in-sensor accelerators to perform convolutional operations directly inside the image sensor.

The detection and transmission of picture data is facilitated by the use of an image sensor. Various electronic imaging tools, such as clinical imaging tools, digital cameras, camera modules, night vision instruments, such as sonar, thermal imaging devices, and radar, extensively use these sensors. Additionally, a diverse range of other electronic imaging equipment also relies on these sensors. Video camera tubes are instances of analog sensors that were used in previous times. According to Fu, Liu, Chen, Wang, and Chou [4], active pixel sensors in CMOS technology use charge-coupled device (CCD) semiconductors. Analog sensors, in contrast to their digital counterparts, are characterized by their use of vacuum tubes as opposed to flat-panel detectors. Preprocessing of images is a necessary step prior to their use in the training and inference stages of the model. Several instances of this phenomenon may be seen, such as the act of cropping, rotating, and modifying the color scheme of a picture.

The potential enhancement of model training and inference times might be achieved with the use of picture preprocessing techniques. In the domain of image processing, the application of DNNs is employed for the purpose of making predictions about the content of an image, often referred to as inference. The properties that have been created are then analyzed by the host computer. The process of feature extraction is executed in this pipeline prior to the implementation of any ISP activities. Given that this dataset is atypical for post-ISP training, the observed features exhibit noticeable distinctions. As a result, there is a discernible alteration in the distribution of covariates between the training data (images processed by the ISP) and the target data (RAW photographs), leading to a significant influence on the effectiveness of the application.

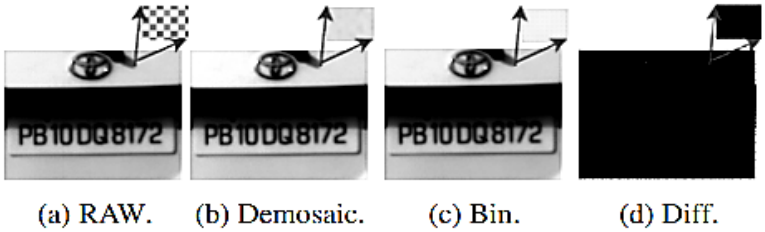


Fig 3: (a) There is a tiling pattern in the RAW image. (b) Demo-saicing creates a picture that is smooth. (c) Binning also produces a plane image. (d) The variation between binned and demosaiced images is negligible (MSE = 4.68 for the provided 400400 image).

III. LITERATURE REVIEW

According to Kumar, Zhu, Gao, Wang, Lanza, and Thakur [5], the integration of an electrical memristor array with 2D materials, facilitated by the unlimited stacking enabled by van der Waals (vdW) bonding, allows for the achievement of high-density integration in a three-dimensional (3D) configuration. The potential enhancement of energy efficiency in 2D photodetectors may be achieved by their integration with a memristor array. The memristor exhibits nonlinearity in its resistance value when subjected to an applied stimulus (as seen in **Fig 4 A**, hence classifying it as a kind of nonvolatile memory. The resistance knob may be conceptualized as the neural network weight. The device has the capability to perform matrix computing operations, which typically require many CMOS transistors, simultaneously.

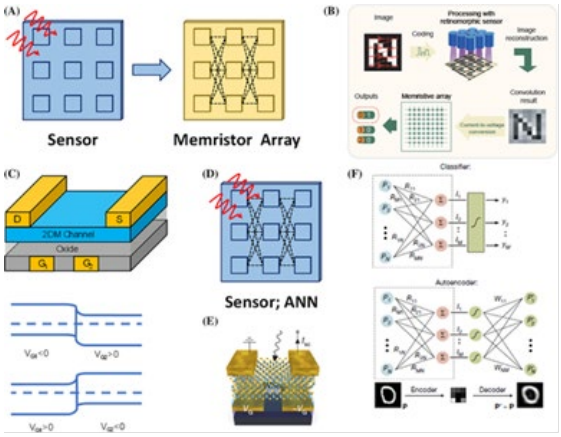
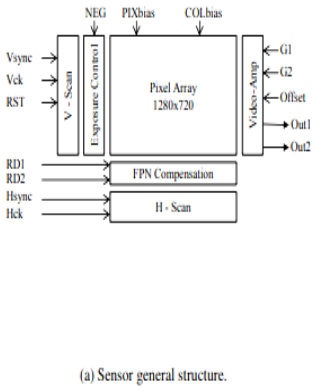


Fig 4. (A) The memristor array may be connected with 2D photodetectors to increase energy efficiency. (B) The processing and sensing of images by a memristor and detector array. (C) By modifying the 2DM band using double back gates, a 2DM photodetector with a variable rate of response may be produced. (D-F)



(a) Sensor general structure.

Optical format	1/2 inch
Pixel Size	5.6um x 5.6um
Pixel Array Size	1280x720
Fill Factor	30%
MicroLens	Available
Dynamic range	> 140*dB*
Output Format	Differential Analog Output
Maximum Frame rate	60Hz
Maximum H scanning rate	80MHz
CMOS process	Standard 0.18 1P3M
High light Log sensitivity	50mV/decade
Low light linear sensitivity	1V/lux*s @ t_{expo}=40ms
Random Noise	0.5mV over all DR
FPN	<1mV
Shutter Mode	Rolling Shutter
Power consumption	230mW

(b) Sensor Characteristics summary.

Fig 5. Sensor Specifications

Lin et al. [6] described a mode, which integrates a neural network with a sensor chip: (D) A sensor may function as a neural network to identify basic traits in addition to collecting data. WSe2 photodiode schematic diagram in (E). The bottom gate electrodes control the photoresponse and the gadget operates in a short circuit situation. (F) The system may be taught in two modes: classifier and autoencoder. B) This image was used with permission.

The neuromorphic vision system comprises of memory networks and sensors that replicate the biological functionality and hierarchical arrangement of the retina. The user's text is too short to be rewritten academically. **Fig 4B** displays a processing flowchart and a neuromorphic vision system. The sensor was constructed using a van der Waals (vdW) heterostructure including layers of WSe2, h-BN, and Al_2O_3. This design enabled the sensor to be electronically toggled between a "ON" and "OFF" state. Through the use of discrete programming of individual detectors inside a 33 bionic retinal sensor, we successfully executed many conventional kernel operations inherent to a Convolutional Neural Network (CNN) algorithm. Various convolution kernels were used to carry out a diverse array of image-processing tasks, such as eliminating extraneous information and extracting essential ones.

In contrast to the dynamic kernel function used in a conventional convolutional neural network (CNN) approach [7] posit that the kernel function's representation in the sensor could not be altered during its utilization. Following the extraction process, it was necessary to transform the optoelectronic signal into a digital signal prior to transmitting it to the memristor crossbar system. The box type was not explicitly mentioned in the paper; nonetheless, it is likely that a printed circuit board (PCB) was used instead of a surface mount package (SMP). In contrast to computers built on the von Neumann architecture, memristors do not need the conversion of analog signals into digital format for processing purposes. The neuromorphic

system demonstrated exceptional cross-tracking capabilities in an environment only composed of analog signals by the integration of retinomorphic sensors with recurrent neural network (RNN).

If the physical features of a photodetector array, such as electric memristor resistance, could be controlled by an exterior voltage, it has the potential to function as a neural network. Many scholars have made advancements in the field of image identification by developing photodiode arrays using 2D materials as an alternative to the conventional von Neumann design, which is considered antiquated. Consequently, it is possible to integrate additional calculations inside the detector (**Fig 4**D) in order to enhance its in-field capabilities. The photodiode array of the device functioned as the neural network responsible for the real-time recognition and processing of images, therefore obviating the need for laborious and ineffective individual processing stages. As a result, intelligent recognition has become faster and more efficient compared to its previous state. The responsivity of each pixel device may be independently modified by electrically doping the channel materials via the use of double-gate electrodes, as seen in **Fig 4**E.

The artificial neural network (**Fig 4**F) was used by the system to perform both supervised and unsupervised learning tasks. The determination of the weight of the neural network was based on the responsiveness shown by each individual unit. The sensitivity of the system was adjusted in real-time to enhance its accuracy in identifying objects. The system's speed was only constrained by the optical-electrical conversion process, since both image processing and image sensing were executed inside an analog signal framework. The process of detecting and identifying an image was accomplished in under 50 nanoseconds. In comparison to traditional methodologies, the system has potential capabilities to process a staggering quantity of 20 million photographs within a single second. Moreover, the biological neural network demonstrated significantly reduced energy consumption per operation as compared to conventional machine vision systems (MVSs), which use substantial power throughout each operation. The range of energy values considered in this context spans from 10^-15 to 10^-13 Joules.

According to [8], stereo vision systems find use in different fields, such as robotics (evidenced by the NASA Curiosity rover), autos (specifically in pedestrian identification, driving assistance, and accident avoidance), industrial safety, and security (in the counting of people). Contemporary computer vision methodologies heavily depend on the use of image processing techniques and calibration methods in order to address geometric obstacles, such as rectifying projection alteration and designing of stereo pairs. The application of stereo vision models for depth calculation becomes unfeasible due to the inability to extract depth information caused by the loss of contrast in oversaturated areas. In order to address these challenges, Han, Bian, Liu, Zeng, and Tian [9] developed a stereo vision system. At the core of this camera are two logarithmic sensors, namely the B&W WDR NSC1005 sensors, which are under the direction of a shared controller.

Fig 5 presents a comprehensive summary of the technical specifications of the sensors. The NSC1005 demonstrates the ability to maintain a constant level of contrast sensitivity throughout a wide range of 140 dB. The differential analog outputs of both sensors are connected to twelve-bit differential analog-to-digital converters (ADCs). In this design, the generation of sensor control indicators (Vck, Vsync, Hsync, RST, Hck, RD1, and RD2) for all sensors is accomplished by an FPGA, hence achieving synchronization at the pixel level. The Field-Programmable Gate Array (FPGA) is responsible for receiving the digitized levels of the left and right pixels. Upon transmission to the host PC, the Field-Programmable Gate Array (FPGA) will undertake the task of compressing the left and right channels. These channels are both 12 bits in size and will be combined into a single channel with a size of 24 bits.

The pipeline serves as the fundamental framework for the majority of current machine vision systems. The ISP has a substantial role in influencing both the accuracy and energy consumption of these systems. The study conducted by Hong, Siau, and Kim [10] examined the impact of various ISP practices on overall productivity. The use of a software application to convert images processed by an ISP into RAW images has led to the identification of the crucial role played by denoising, gamma compression, and demosaicing in retaining the high accuracy levels in various applications. Due to the inherent non-linear and non-invertible characteristics of several stages within the ISP pipeline, the reconstruction of RAW images can only be achieved by approximate methods. For instance, the introduction of noise is necessary in order to enhance the image signal. The effect of demosaicing and denoising techniques on task efficiency may have been overstated by Sharif, Ali Naqvi, and Biswas [11] as a result of their simplified approach to RAW picture reconstruction.

It has been observed that the implementation of a smoothing technique that effectively mitigates the tiling artifacts caused by the color filter array, while also exhibiting partial denoising capabilities, is sufficient to achieve a high level of accuracy in many applications. During an email correspondence, Buckler expressed his belief in the credibility of the subject matter. Pixel binning is a technique that produces continuous images by reducing noise and smoothing the discontinuities in RAW photographs, resulting in a level of continuity similar to that achieved with demosaicing. The parallel readout architecture of image sensors also enables pixel binning, making them highly appropriate for in-sensor learning acceleration. Given the aforementioned circumstances, we propose the implementation of a concise two-step processing pipeline, which integrates pixel binning and gamma compression.

IV. RESULTS AND DISCUSSION

Identifying the Required Pipeline Operations

Both the RAW and ISP-processed iterations of the image effectively capture the structural and geometric aspects of the scene, since they serve as interchangeable representations of the same visual content. The main differentiating factors among them are the outcomes resulting from changes in data distribution caused by localized modification techniques such as

gamma compression. To acquire knowledge about representations, machine vision algorithms undergo training using images that have been processed by an ISP. There is a potential argument in favor of training networks using RAW images; nevertheless, many problems now hinder this approach. There are several challenges in the field of RAW image processing that need to be addressed. Firstly, there is a scarcity of publicly available datasets containing RAW images. This paper aims to partially overcome this issue by introducing a dataset of RAW images for public use. Secondly, there is a lack of a universally standardized RAW image format. Lastly, the substantial file sizes of RAW images pose difficulties in their transfer, particularly when compared to more compressed formats such as PNG and JPEG.

In order to estimate the distribution of data in ISP-generated images, we suggest a critical image signal preprocessing pipeline, which influences gamma compression and pixel binning techniques. Consequently, the application of pre-existing deep neural networks (DNNs) is feasible in both near-sensor and in-sensor accelerators due to the diminished covariate divergence between RAW images (targets) and ISP-generated (training). In order to assess the impact of our recommended pipeline on the distribution of data, we investigate its influence on the distribution of pixel intensities.

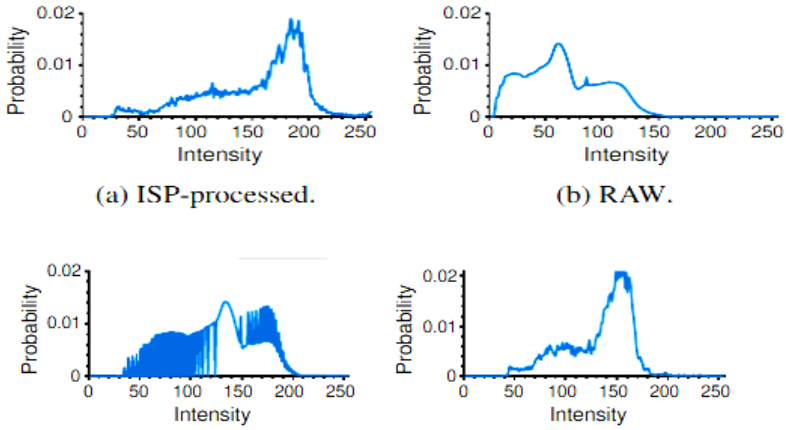


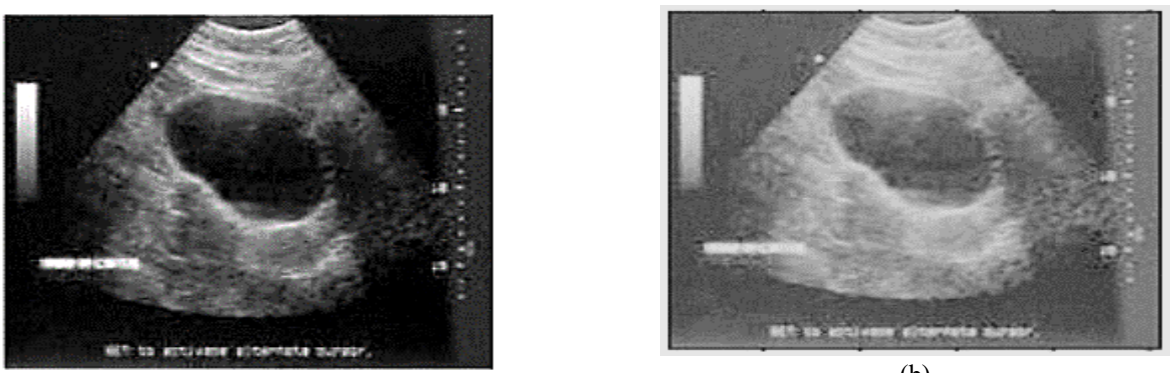
Fig 6. Differences in intensity distribution between RAW and ISP-processed images are striking. If (b) is compressed with gamma, the resulting non-linear transformation may be denoised using binning to get a rough approximation of (a).

Gamma Compression

Gamma correction is a nonlinear operation used to rectify the luminance, which refers to the picture brightness level, of individual pixels within an image. These functions are designed to map brightness levels in order to correct for the non-linear luminance impact shown by display devices. The power-law expression is shown by the following expression.

$$AX^{\Upsilon} = Y \quad (1)$$

The 'Y' output value is retrieved by increasing the 'X' input value to the gamma power. Gamma encoding or compression refers to the use of a gamma value ϒ that is less than 1. The application of a gamma value of 0.6 to the dark sections, as seen in **Fig 7** (a) and (b), will result in an increase in brightness. The term "Gamma decoding" or "Gamma expansion" refers to the process of decoding or expanding a gamma value ϒ that is greater than 1.



(a) (b)

Fig 7. Typical gammma compression of images with value 0.6

In contrast to image sensors, the human visual system exhibits a logarithmic response rather than a linear reaction to light. The prioritization of visual quality is of utmost importance for ISPs (ISPs). To substantiate our claims, we may analyze the intensity distributions of the images before and after undergoing image signal processing (ISP), as seen in **Fig 6**. The

process of gamma compression involves the application of a non-linear, localized change to each pixel, wherein the pixel value is multiplied by an exponent that is less than 1. In order to estimate the intensity distribution after ISP processing, it is observed that gamma compression may be used (refer to **Fig 6**(c)). The compression of gamma photographs will be performed in accordance with Adobe's 1998 standard.

$$P_{norm}{}^{\gamma} = P_{gamma} \tag{2}$$

Pixel Binning

The frame rate and signal-to-noise ratios of digitalized cameras may be improved with the use of pixel binning, a clocking mechanism that combines the charge accumulated by adjacent CCD pixels. The binning procedure is executed by on-chip CCD clock time circuit to determine the parallel and serial shift registers, before increasing the CCD analogue output.

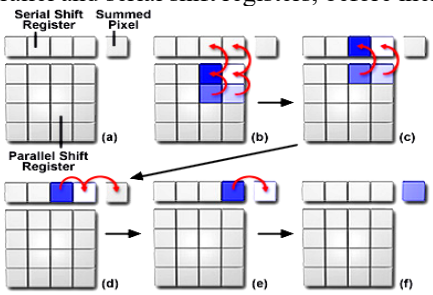


Fig 8. Illustrative example of the 2 x 2 binning technique, aimed at enhancing the understanding of the pixel binning procedure.

The schematic representation in **Fig 8**(a) illustrates a simplified arrangement consisting of a four-by-four array of parallel shift register pixels, a 4-gate serial shift registers, and a summation pixel, also known as the output nodes. When photons emitted from the light source make contact with the CCD photodiodes, electron accumulation occurs inside each pixel, as seen in **Fig 8**(b) by the 4-blue shade square found in the top right part of the parallel shift registry. The term "well depth" pertains to the electron count range per pixel inside a charge-coupled device (CCD), often falling within the range of 30,000 to 350,000. There exists a clear correlation between the depth of a well and the dynamic range of a charge-coupled device (CCD). The electrons gathered at every pixel site or photogate is determined by the intensity of incoming light and the duration of exposure for the picture. Once a single cycle of illumination has been administered to the charge-coupled device (CCD) and the picture has been acquired, the electrons are focused through the serial and parallel shift registers, output amplifiers, and definitely the A/D (analogue-to-digital) converter circuit. Binning may enhance picture sharpness and reduce exposure time, hence increasing the camera's sensitivity to slightly defocused light.

An instance of this phenomenon is shown in **Fig. 8**(b), whereby the progression of each integrated pixel inside the parallel register is illustrated as advancing by one gate, ultimately resulting in the arrangement displayed in **Fig 8** (c). In this scenario, electrons originating from two pixels are retained inside the parallel shift register, whereas electrons from the remaining two pixels are sent towards the serial shift register. Subsequently, the electrons that remain in the corresponding shift register are sent to the adjacent gate components of the serial registers in a subsequent step, as seen in **Fig 8** (c). The process of charging the summing pixel, as seen in **Fig 8** (d) and (e), necessitates the last phases of transferring charge from the serial register. In **Fig 8** (f), the summing well displays a charge equivalent to four pixels, which is poised to be sent to the output amplifier. Subsequently, the signal will undergo amplification and digitization before to its transmission to other integrated circuits. The process of reading from the array continues until all elements inside the array have been handled in their entirety. In this context, a super pixel is seen, which is a larger pixel resulting from the amalgamation of the spatial extent of four adjacent pixels. The resolution of the image has been decreased by 50%, but the signal-to-noise ratio has been increased by a 4-factor.

Pixel binning is a widely used technique for subsampling in which the data is subjected to an averaging process followed by conversion into decimal form. Similar to the process of averaging, binning is a technique that reduces the presence of Gaussian noise. **Fig 6**(d) illustrates the process of denoising by using the technique of binning the gamma compressed picture, resulting in the creation of a uniform intensity distribution. An additional benefit of pixel binning is its capacity to enhance analysis efficiency by subsampling of the image. The use of deep neural network (DNN) based techniques for object identification necessitates a substantial allocation of time and resources when processing high-resolution images, owing to the utilization of small convolutional kernels. Images with dimensions exceeding 1,000 by 1,000 are unable to be accommodated within the available RAM, even when using a 6 GB NVIDIA GTX 1060 GPU. The significance of image size has been disregarded in prior research on in-sensor accelerators due to the use of images that were too small to provide significant object recognition. The process of sub-sampling the ‘s’ images based on ‘w’, the binning window ‘w’ may be mathematically represented as follows:

$$\frac{1}{2w+1}\sum_{k=-w}^{+w} P(s \times j + k, s \times i + k) = P_{bin}(i,j) \tag{3}$$

Hardware Implementation

The aforementioned actions provide an adequate approximation of the data distributions of images processed by the ISP, as seen in **Fig 6**. In order to get a high level of accuracy and facilitate the application of pre-existing DNNs for deep learning

accelerators inside sensors, it is essential to include pixel binning and gamma compression prior to the readout of the image sensor. In order to achieve this objective, we suggest the modifications that follows to the image sensors.

Logarithmic pixels

Linear photodiodes are highly suitable for use in image sensors due to their ability to convert light energy into a voltage that is directly proportionate to the incident light. The connection between input light energy and the corresponding logarithmic output voltage in logarithmic pixels is linear. This linearity is attributed to the sub-threshold activity of the source supporter transistors in the dynamic pixel sensors. The logarithmic function is often used as an approximation for gamma compression. To facilitate the integration of gamma compression inside the sensor, logarithmic pixels might be used.

Pixel binning

Binning may be accomplished by using a traditional circuit for averaging during the reading phase. The parallel readout pipeline culminates in a series of gathering capacitors, which may be electrically connected to facilitate charge sharing and averaging. Pixel binning is a widely used technique that is compatible with the majority of contemporary image sensors. Although the inclusion of supplementary circuitry for averaging necessitates an increase in sensor power consumption, the resulting increment is deemed insignificant according to Ngampruetikorn, Schwab, and Stephens [12]. Moreover, as the length of the binning window increases, the readout time drops quadratically, resulting in a reduction in net sensing energy.

Accuracy Detection of RAW, Proposed and ISP Data

The replication of the effects of our suggested sensing system may be achieved via the use of pixel binning and gamma compression approaches to RAW images. The effect of dataflow choice on accuracy has been insufficiently acknowledged in prior research on in-sensor accelerators, mostly owing to evaluations conducted on images that have already undergone processing by an image signal processor. Nevertheless, our evaluation is based on the analysis of RAW images captured using a commercially available camera, so replicating the observations that an in-sensor accelerator would make. A test dataset including 225 RAW images was gathered, encompassing 1,215 distinct cars. The dataset is now available for public access. The images provided are digital replicas of the original analog versions, without any modifications made by the ISP. The impact on precision: The accuracy of object recognition is assessed by using the modelzoo [13] in TensorFlow as a benchmark. This benchmark comprises many cutting-edge object detection networks. **Fig 9** illustrates the detection precision of RAW images, images pre-processed by the recommended pipeline, ISP-processed images within vehicle-identification case.

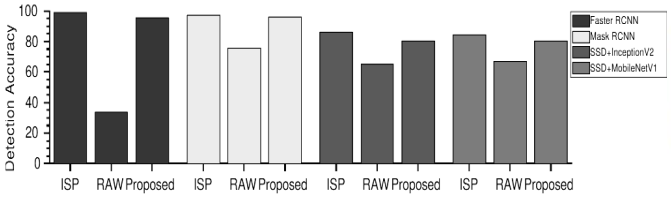


Fig 9. The accuracy of detecting objects in post-ISPs, RAW images, and images processed using the suggested pipeline.

Fig 10. Off-the-shelf DNN employed in RAW images using the suggested methodology.

When the estimated bounding box exceeds the genuine bounding box by a margin of over 40 percent, it is deemed that the object has been detected (see to **Fig 10**). The recognition accuracy often experiences a 25-60% improvement when comparing in-sensor accelerators to RAW images. **Fig 11** presents the mAP (mean average precision) value for various networks and preprocessing techniques. The application of robust feature extractors with the MobileNet and Faster RCNN implementations leads to an increased occurrence of false positive within ISP-generated images, resulting in a reduction in net mAP values.

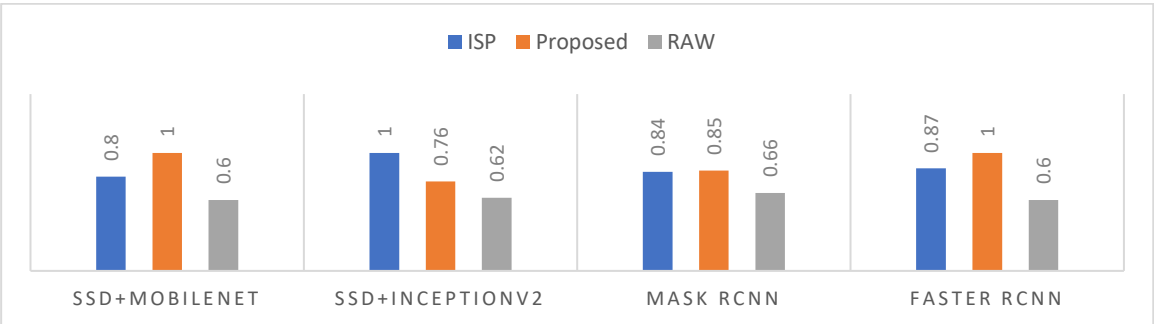


Fig 11. Mean average accuracy for various networks.

The Raspberry Pi 3 microcontroller is used to ascertain the overall duration needed by a typical ISP pipeline for the purpose of assessing the impact of the ISP on the response time and energy demands of an embedded vision system. The

Raspberry Pi is compatible with TensorFlow-lite, which is a streamlined iteration of the widely-used TensorFlow application programming interface (API). The implementation of Quantized SSD MobileNet V2 is used in order to ensure that it can be accommodated within the available memory capacity. According to the findings of Kumar, Kumar, and Saad [14], this particular model is used for the evaluation of energy efficiency. In comparison to the conventional ISP pipeline employed for the vehicle-identification challenge, our proposed method exhibits a reduction in latency by 34% and a decrease in energy usage by 30%. The results of this study indicate that the use of near-sensor accelerators, which maintain the ISP, might potentially reduce both system energy consumption and analytical latency by circumventing the need for the ISP in the proposed pipeline.

V. CONCLUSION

Machine learning approaches have found extensive use across several domains, including but not limited to pattern identification, computational learning, and natural language processing. Over the course of recent decades, machine learning has had a significant impact on several aspects of our everyday lives. This effect is shown by the advancements in efficient online search algorithms, the development of self-driving vehicles, the progress made in computer vision technology, and the improvements in optical character recognition capabilities. This research elucidates the infeasibility of using commercially available deep neural networks (DNNs) with in-/near-sensor deep learning accelerators without prior implementation of a meticulously designed preprocessing strategy. The use of binned readouts and logarithmic pixels facilitates the straightforward integration of fundamental processes, such as pixel binning and gamma compression, into an image sensor. This approximation effectively replicates the data-generating distribution of the images processed by the ISP. The proposed method enables the application of commercially available deep neural networks (DNNs) in accelerators located inside or near the sensors. This implementation leads to notable improvements, including a decrease in error rate by approximately 26% to 59%, a reduction in the consumption of the system energy by approximately 29%, and a reduction in analysis latency by 35%.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding was received to assist with the preparation of this manuscript.

Competing Interests

There are no competing interests.

References

[1]. Z. Zhou et al., “Operation of AIMS in deep learning workflow,” ASVIDE, vol. 10, pp. 137–137, 2023.
[2]. Lokesh, Chandana, “Vlsi modeling of high performance digital signal processors for wireless sensor nodes,” I-manag. S J. Digit. Signal Process., vol. 5, no. 2, p. 22, 2017.
[3]. H. Zhang, J. Cheng, J. Zhang, H. Liu, and Z. Wei, “A regularization perspective based theoretical analysis for adversarial robustness of deep spiking neural networks,” Neural Netw., vol. 165, pp. 164–174, 2023.
[4]. H.-K. Fu, Y.-L. Liu, T.-T. Chen, C.-P. Wang, and P.-T. Chou, “The study of spectral correction algorithm of charge-coupled device array spectrometer,” IEEE Trans. Electron Devices, vol. 61, no. 11, pp. 3796–3802, 2014.
[5]. P. Kumar, K. Zhu, X. Gao, S.-D. Wang, M. Lanza, and C. S. Thakur, “Hybrid architecture based on two-dimensional memristor crossbar array and CMOS integrated circuit for edge computing,” Npj 2D Mater. Appl., vol. 6, no. 1, 2022.
[6]. P. Lin et al., “Three-dimensional memristor circuits as complex neural networks,” Nat. Electron., vol. 3, no. 4, pp. 225–232, 2020.
[7]. A. Kumar, K. Abhishek, X. Liu, and A. Haldorai, “An Efficient Privacy-Preserving ID Centric Authentication in IoT Based Cloud Servers for Sustainable Smart Cities,” Wireless Personal Communications, vol. 117, no. 4, pp. 3229–3253, Nov. 2020, doi: 10.1007/s11277-020-07979-8.
[8]. K. Ashok, M. Ashraf, J. Thimmia Raja, M. Z. Hussain, D. K. Singh, and A. Haldorai, “Collaborative analysis of audio-visual speech synthesis with sensor measurements for regulating human–robot interaction,” International Journal of System Assurance Engineering and Management, Aug. 2022, doi: 10.1007/s13198-022-01709-y.
[9]. F. Han, Y. Bian, B. Liu, Q. Zeng, and Y. Tian, “Research on calibration of a binocular stereo-vision imaging system based on the artificial neural network,” J. Opt. Soc. Am. A Opt. Image Sci. Vis., vol. 40, no. 2, pp. 337–354, 2023.
[10]. S.-G. Hong, K. Siau, and J.-W. Kim, “The impact of ISP, BPR, and customization on ERP performance in manufacturing SMEs of Korea,” Asia Pac. J. Innov. Entrep., vol. 10, no. 1, pp. 39–54, 2016.
[11]. S. M. A. Sharif, R. Ali Naqvi, and M. Biswas, “Beyond joint demosaicking and denoising: An image processing pipeline for a pixel-bin image sensor,” in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021.
[12]. V. Ngampruetikorn, D. J. Schwab, and G. J. Stephens, “Energy consumption and cooperation for optimal sensing,” Nat. Commun., vol. 11, no. 1, p. 975, 2020.
[13]. A. Salari, A. Djavadifar, X. Liu, and H. Najjaran, “Object recognition datasets and challenges: A review,” Neurocomputing, vol. 495, pp. 129–152, 2022.
[14]. S. Kumar, R. Kumar, and Saad, “Real-Time Detection of Road-Based Objects using SSD MobileNet-v2 FPNlite with a new Benchmark Dataset,” in 2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2023.