IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 March 2023, accepted 24 April 2023, date of publication 28 April 2023, date of current version 3 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3270710

RESEARCH ARTICLE

Exploring the Viability of Bypassing the Image Signal Processor for CNN-Based Object Detection in Autonomous Vehicles

JORDAN CAHILL[1], ASHKAN PARSI[2], DARRAGH MULLINS[1], JONATHAN HORGAN[3], ENDA WARD[3], CIARÁN EISING[4], (Member, IEEE), PATRICK DENNY[4], (Member, IEEE), BRIAN DEEGAN[1], (Member, IEEE), MARTIN GLAVIN[1], (Member, IEEE), AND EDWARD JONES[1], (Senior Member, IEEE)

[1]Department of Electrical and Electronic Engineering, University of Galway, Galway 091, H91 TK33 Ireland
[2]Xperi, Parkmore East, Galway, H91 V0TX Ireland
[3]Valeo Vision Systems, Tuam, Galway, H54 Y276 Ireland
[4]Department of Electronic and Computer Engineering, University of Limerick, Limerick 061, V94 T9PX Ireland

Corresponding author: Jordan Cahill (j.cahill5@universityofgalway.ie)

This work was supported in part by the Science Foundation Ireland under Grant 13/RC/2094, and in part by the European Regional Development Fund through the Southern and Eastern Regional Operational Program to Lero–the Science Foundation Ireland Research Centre for Software (www.lero.ie).

ABSTRACT In the field of autonomous driving, cameras are crucial sensors for providing information about a vehicle's environment. Image quality refers to a camera system's ability to capture, process, and display signals to form an image. Historically, ''good quality'' in this context refers to images that have been processed by an Image Signal Processor (ISP) designed with the goal of providing the optimal experience for human consumption. However, image quality perceived by humans may not always result in optimal conditions for computer vision. In the context of human consumption, image quality is well documented and understood. Image quality for computer vision applications, such as those in the autonomous vehicle industry, requires more research. Fully autonomous vehicles inevitably encounter constraints concerning data storage, transmission speed, and energy consumption. This is a result of enormous amounts of data being generated by the vehicle from suites made up of multiple different sensors. We propose a potential optimization along the computer vision pipeline, by completely bypassing the ISP block for a class of applications. We demonstrate that doing so has a negligible impact on the performance of Convolutional Neural Network (CNN) object detectors. The results also highlight the benefits of using raw pre-ISP data, in the context of computation and energy savings achieved by removing the ISP.

INDEX TERMS Object detection, image signal processor, autonomous vehicles, neural networks, raw data, Bayer filter.

I. INTRODUCTION

Object detection is a widely studied field concerned with extracting appropriate information from an image and identifying subjects from the extracted data. Object detection is used in a multitude of different applications, such as autonomous vehicles [1], [2], security [3], robotics [4], and medical [5]. Almost all existing object detection algorithms are designed for use with images processed by a typical Image Signal Processor (ISP) in a standard camera system. The ISP is responsible for taking the raw data captured by a camera sensor, and converting it to a suitable image to be viewed by a human. However, a typical ISP associated with modern cameras, including camera sensors in the automotive industry, primarily produces images designed for human viewing in mind [6], [7]. What defines ''good quality'' is a widely debated topic and no definitive answer exists. Traditionally, humans were considered to be the primary consumer of camera images, and as such ISPs are commonly designed to display images that are intended to meet a

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

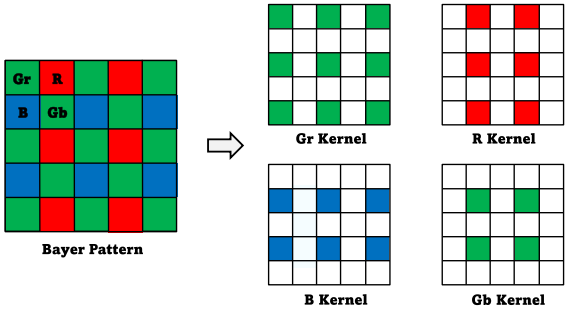


FIGURE 1. Bayer format and filtering.

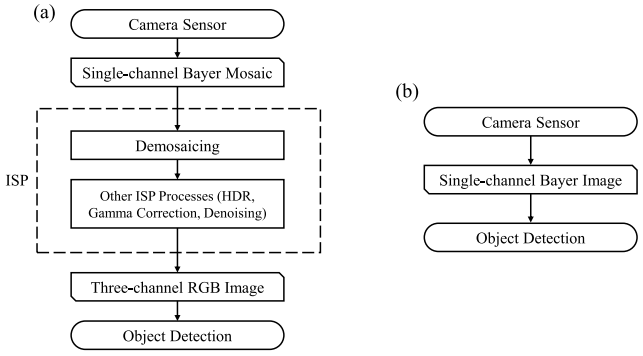


FIGURE 2. *(a)* Traditional image acquisition pipeline; *(b)* Proposed image acquisition pipeline bypassing the image signal processor.

human's expectation of what the underlying data represents. This representation is usually not optimized for the needs of computer vision applications, and the ideal parameters for human vision and computer vision may even be mutually antagonistic.

To produce a typical color image, the camera combines images from three channels, i.e., red (R), green (G), and blue (B), combined as an RGB image. Photosensors laid out in an array capture color information, with a single color being captured per pixel. These arrays are known as color filter arrays (CFAs). A Bayer filter is by far the most commonly used CFA format consisting of a series of 2×2 grids each containing two green pixels, one blue pixel, and one red pixel [8], as shown in Figure 1. The Bayer filter dedicates twice as many green pixels as red and blue pixels due to the human eye being more sensitive to green [9]. From here, the ISP converts this pattern into an image designed to be suitable for human viewing through several discrete processes, including noise filtering and gamma correction, as well as the conversion of the Bayer image to standard RGB format, known as demosaicing. In image processing, "*raw*" can be used either to refer to an image that has not been processed by an ISP or to refer to an image that has not undergone image compression. In this study, we use the term "*raw*" to refer to a mosaic image that has not been processed by an ISP.

This paper examines the viability of using pre-ISP images for object detection in an autonomous vehicle application, using Convolutional Neural Networks (CNNs). To do so we evaluate the performance of two different CNN-based object detectors when trained and evaluated on conventionally processed RGB images, against the performance when trained and evaluated on raw Bayer images, with no ISP processing. In [10], Bucker et al. claim that the demosaic and gamma correction steps are the two most critical processes for computer vision task performance. Thus we train and evaluate a third model on an image set whereby only gamma correction is applied to the original Bayer images. The images evaluated in this research were captured using linear range cameras, however, it is worth noting that many autonomous cameras are High Dynamic Range (HDR) cameras, and thus would be a possible consideration for future work.

We also explore the potential savings in the context of computation time and power consumption as a result of removing the ISP from a traditional image acquisition pipeline. Note that this analysis focuses on the ISP and is independent of the specific algorithms used for object detection. Figure 2 shows the traditional image acquisition pipeline, as well as our proposed acquisition pipeline. Additionally, this work provides a brief commentary on the lack of comprehensive datasets consisting of pre-ISP Bayer images available within the public domain and their importance for computer vision.

The specific contributions of this study include:

- A detailed comparison of object detection performance between two CNN-based models trained and evaluated on sets of images both before and after exposure to image signal processing.
- A comparison of relevant public datasets containing pre-ISP images, and a brief commentary on the lack of availability of said datasets, particularly for the use case of object detection.
- A brief description of the discrete processes in an ISP, highlighting where the ISP may negatively affect computer vision performance, where relevant.
- An analysis of potential savings to be made in terms of energy consumption and computation time, through bypassing the ISP.

The remainder of section I provides a background on related work in the field. Section II provides a discussion on the modular nature of an autonomous vehicle system architecture. Section III provides a brief overview of ISP processing blocks, and an insight into the impact tuning an ISP towards human vision has on computer vision performance. Section IV covers CNN-based object detection and the models implemented in this work, as well as a commentary on available datasets. Section V evaluates the results of object detection using the YOLOv5 and Faster R-CNN models on Bayer and RGB images. Section VI provides a discussion on the system-level performance benefits of bypassing the ISP for computer vision. Section VII concludes the paper.

A. RELATED WORK

Studies have shown benefits in bypassing the demosaicing step and using CFA images such as Bayer filter images, for feature extraction computer vision applications. In [11], Zhou et al. explore the impact of demosaicing on gradient-based feature extraction. Zhou et al. demonstrate both theoretically and experimentally that the raw Bayer pattern images are sufficient for gradient-based feature extraction algorithms with a negligible drop in performance. Trifan and Neves corroborate this in [12] using CFA images for Scale Invariant Feature Transform (SIFT) [13] and Speeded-Up Robust Features (SURF) [14] feature description with minimal drop in performance.

In a conventional computer vision system, the ISP consumes a significant amount of computational resources, processing time and energy consumption [10]. In an attempt to reduce computation, Buckler et al. [10] examined the role of an ISP in both classical and CNN-based computer vision, by disabling stages of a traditional ISP. They claim that the majority of ISP processes are unnecessary for computer vision applications, however, they found gamma correction, denoising, and demosaicing to have significant effects on computer vision. They use a software tool they refer to as the *Configurable & Reversible Imaging Pipeline* (CRIP) to convert ISP-processed images to raw images.

However, since several ISP processes are non-linear and non-invertible, the resultant images are only approximate reconstructions, which may amplify the effects of individual processes. This claim is backed by Lubana et al. who surmise that Buckler et al. may have overestimated the effects of these ISP operations due to the images examined being of a pseudo-Bayer format [15].

Furthermore, due to the nature of the interpolation process, the demosaicing step may lead to errors in the form of erroneously estimated colors, or an artificial jagged pattern known as the zipper effect [16], [17].

A common trend with a lot of the related work is restrictions due to a lack of pre-ISP images available to the public domain. Chandra et al. use subsampled RGB images in a format they refer to as pseudo-Bayer images [18] for gesture recognition. In their work, they mention the benefit a comprehensive collection of publicly available Bayer images would have on their research. The research presented builds from the related work mentioned above on using Bayer images for computer vision, by extending the work to an object detection use case in an autonomous vehicle context.

II. AUTONOMOUS VEHICLE SYSTEM ARCHITECTURE

The typical architecture of a software system within an autonomous vehicle is different from that of a traditional PC system. Autonomous vehicle architectures typically include several layers of software, usually consisting of many ECUs (Electronic Computing Units) each responsible for an individual task such as vision, area mapping, steering, braking, and infotainment. The exact architecture may vary, depending on the implementation. A typical autonomous vehicle software architecture may contain the following [19]:

1) Perception Layer: Responsible for collecting sensor data from cameras, LiDARs, radar, and other sensors, and processing the data to generate a 3D map of the vehicle's surroundings. The data can then be used to identify objects such as other vehicles and pedestrians, as well as their locations.
2) Prediction and Planning Layer: Responsible for using the data from the perception layer and predicting the behavior of surrounding objects in order to plan a safe path for the vehicle.
3) Control layer: Responsible for controlling the vehicle's acceleration, braking, and steering to navigate the planned path from the prediction and planning layer
4) User Interface Layer: Responsible for providing the user with an interface to interact with the autonomous vehicle. The user interface layer includes displays and controls for setting destinations, infotainment systems, and other features.

The modularity of a typical autonomous vehicle software architecture makes it feasible to adjust one subsystem without a drastic effect on the overall system. Furthermore, modularity allows for the potential for varying degrees of image processing depending on the use case. For use cases such as Advanced Driving Assistance Systems (ADAS) and infotainment systems, where human-viewable images are required, an ISP may be used in order to produce an image that is appropriate for human vision. However, in cases where the user is not required to view the images, such as automated collision avoidance through object detection, the raw Bayer images may prove sufficient.

III. IMAGE SIGNAL PROCESSOR

The Image Signal Processor (ISP) is a processing block involved in the image acquisition process responsible for converting the data captured by a camera sensor into a usable image for a given application, be it human consumption or computer vision. The ISP is made up of many discrete processes, some of which may be destructive or irreversible. Thus, as an image passes through each block, the quantity of information decreases. Without the reference data captured directly at the sensor, it is impossible to re-capture the data lost as a result of some of these processes. Figure 3 shows an example image processing pipeline. Note that an individual ISP may not be limited to the processes in Figure 3 and may have additional processes, depending on the individual ISP.

Lens shading correction—is applied to improve the spacial uniformity of the illumination and color captured by a sensor. These non-uniformities are more common around the image periphery due to lens distortion. Lens shading correction is particularly critical for automotive cameras that employ fish-eye lenses.

Auto white balance (AWB)—is responsible for correcting the color temperature of the final image, such that white

TABLE 1. An overview of relevant available datasets.

Dataset	Year	No. of Images	Pre-ISP Images	Labelled	Automotive Targets
PASCALRAW	2014	4,259	✓	✓	✓
Oxford RobotCar	2016	19,556,490	✓	✗	✓
INTEL-TAU	2019	7,022	✓	✗	✗
ZRR	2021	20,000	✓	✗	✗
G-MIND	2023	65,000	✓	✓	✓

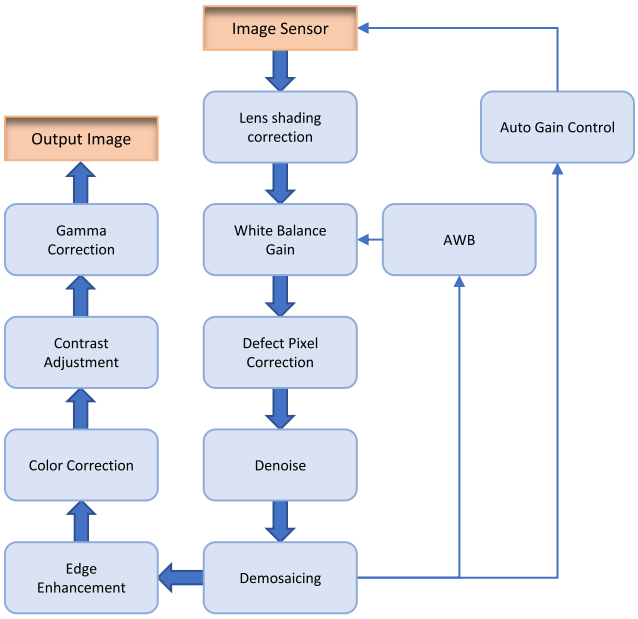


FIGURE 3. Example ISP pipeline. In practice, ISPs are not limited to the blocks provided in the Figure and conversely may not contain all processes shown.

objects appear a desired white. AWB values are calculated using the input image and these values are used to apply digital white balance gains to the red, green, and blue channels appropriately.

Auto gain control (AGC) / Auto exposure control (AEC)—is a feedback loop within the ISP, responsible for managing the gain and exposure for the next frame, which is calculated using a weighted average of the current exposure. The gain in a digital imaging device is a way of increasing the sensitivity of the sensor to light and represents a relationship between the sensor and the number of electrons captured.

Defect pixel correction—corrects defective pixels captured by an image sensor. A defective or dead pixel refers to a pixel in which the pixel, at a sensor level, is turned off, creating a black spot within the image. Dead pixels increase over an image sensor's lifetime and may be the result of malfunctioning transistors or photodiodes.

Denoise—is the process of filtering an image to remove the appearance of noise in an image. When removing noise, it is very difficult to preserve texture within an image, thus losing high-frequency information. The loss of this high-frequency data can be a negative factor when detecting image gradients, a key component in feature detectors. Denoising is particularly irreversible without the reference image, due to the stochastic nature of noise.

Demosaicing—or color interpolation, as discussed is the process of converting a sensor's raw data, typically in the form of a Bayer CFA, into a color RGB image. While demosaicing is certainly a critical operation in terms of producing an image designed for human viewing, it can also introduce imaging artifacts. Some of these artifacts include zippering or staircase artifacts on edges, as well as false colors within an image. The former of these artifacts can have a negative impact on edge detection performance due to repeating edges. These effects can be mitigated by more complex demosaicing filters, however, this comes at a computational cost. As a demosaiced image is an interpolation of the raw sensor data to generate an RGB image, it is again impossible to accurately reverse without estimation.

Edge enhancement—is responsible for enhancing edges within an image, to make it appear sharper. Excessive edge enhancement can be detrimental to computer vision applications, however, particularly gradient-based algorithms, by the artificial creation of duplicate edges.

Color correction—is responsible for correcting any color error generated as a result of the spectral characteristics of the optical lens or color filters used in an image sensor, or in the lighting of the scene. The color correction matrix also corrects any cross-talk between pixels where color information from one pixel contaminates an adjacent pixel. In most ISPs, color correction is applied for better color accuracy, however excessive color correction may accentuate noise, negatively impacting computer vision.

Contrast Adjustment—is responsible for enhancing image contrast, thus improving separation between different grey levels. A clear partition between grey levels can be beneficial for gradient-based algorithms and can thus improve machine vision performance. However, contrast tuning towards human vision needs may differ from computer vision needs, and in some cases have a negative impact on performance.

Gamma correction—is responsible for adjusting the gamma of an image, defining the relationship between a pixel's numerical value, and the represented luminance. If twice the number of photons are captured by an image

sensor, the system will receive a signal with double the amplitude. However, the human eye perceives light non-linearly, in that twice the amount of photons would only seem a fraction brighter. Gamma correction uses the information encoded by the image sensor to adjust the contrast in the final image appropriately. While gamma correction is crucial for human viewing, its impact on machine vision is less documented.

IV. OBJECT DETECTION

A. DATASETS

The primary roadblock to working with pre-ISP images, particularly in the context of deep learning, is the lack of a comprehensive dataset consisting of raw Bayer images available in the public domain. One approach is the use of pseudo-Bayer images where some studies attempt to generate a Bayer image from an already processed RGB image [10], [18]. The problem with this approach is that many of the ISP processes are destructive and non-linear, so the raw images can only be approximately reconstructed [15].

For training and evaluation, we used two different raw datasets. The first dataset we used is the PASCALRAW dataset published by researchers from Stanford University containing pre-ISP images [20]. The dataset consists of 4,259 12-bit images containing three annotated classes: 1,765 cars, 4,077 persons, and 708 bicycles. The images are unprocessed Bayer images captured using a Nikon D3200 DSLR camera.

The second dataset is a dataset constructed internally within the University of Galway, Ireland, known as the Galway Multimodal Infrastructure Node Dataset (G-MIND). The dataset contains data captured from a suite of sensors, including sensors typically used in automotive applications, for the purpose of sensor benchmarking. A subsection of the dataset contains 65,000 annotated pre-ISP images with automotive targets. As such, it is a suitable second dataset for use in this study on the effect of bypassing the ISP on object detection.

For each dataset, we obtained three sets of images, each set having a different amount of image processing carried out, but containing the same information. The three sets of images consisted of the raw Bayer data with no processing, the raw Bayer data with *only* gamma correction, and the fully processed RGB images. The gamma-corrected, and ISP-processed images were obtained by processing the raw Bayer images using Fast-openISP, an open-source software implementation of a conventional ISP. We examine the effects of training and evaluate the performance of CNN-based object detection on each of these different image sets containing the same real-world information, with varying levels of image processing. Both YOLOv5 and Faster R-CNN use OpenCV's *imread()* to upscale grayscale/single-channel images to three-channel images. As stated previously, Buckler et al. claim in [10] that the demosaic and gamma correction steps are the most important ISP processes for computer vision performance, thus we choose our datasets as such to investigate this for our use case. Figure 4 displays the visual difference between each set of images. For the gamma-corrected image set, a gamma correction coefficient of $\frac{1}{2.2}$ was applied to each Bayer image, to match the target monitor gamma in the sRGB color space [21].

B. YOLO: YOU ONLY LOOK ONCE

YOLO (You Only Look Once) is a widely-used Convolutional Neural Network (CNN) for object detection [22], [23], [24]. The first model YOLOv1, was proposed in 2016 [25], and since then several iterations have led to the latest —at the time of writing —improved model YOLOv5, released in 2020 [26]. YOLO is commonly used for its high performance while maintaining both high training and detection speed [27].

The YOLOv5 architecture consists of three primary parts, the backbone, the neck, and the head [28]. The backbone is responsible for extracting the important features within an image. YOLOv5 incorporates CSPNets (Cross Stage Partial Networks) into Darknet as its backbone, known as CSPDarknet. The neck generates feature grids from the features extracted by the backbone through the use of a path aggregation network (PAN). Finally, the head is responsible for generating output vectors with class probabilities and bounding boxes for the final detections.

While YOLO architectures provide impressive detection speeds, early iterations of YOLO struggled with the detection of small targets within an image [29]. This is because without detailed grid division there may tend to be several targets within the same feature grid. YOLOv5 improves on this by transmitting each batch of training data through a data loader, capable of image enhancements such as scaling, color space adjustment, and mosaic enhancement. Moreover, it provides the YOLOv5 algorithm with high adaptability to different sizes of images.

C. FASTER R-CNN

Faster R-CNN is a state-of-the-art Region-Based Convolutional Neural Network (R-CNN) for object detection [30]. It consists of two parts: (1) a Region Proposal Network (RPN) for generating a list of Regions of Interest (RoIs), in which each region is likely to contain objects; and (2) it employs a Fast R-CNN network [31] for classifying an image into objects, including the background as an object, and refining the boundaries of the region proposals in (1).

The significance of traditional R-CNN networks is that they bring the high accuracy of CNNs when it comes to classification tasks to the problem of object detection. Traditionally, the primary drawback with R-CNN networks was the computational load [32]. The R-CNN requires a pass through the network layers for *each* object proposal in order to extract features. To reduce this computational burden, Faster R-CNN uses the RPN, a fully convolutional network, to generate object proposals to be fed into the second module. The second module, Fast R-CNN, is responsible for refining

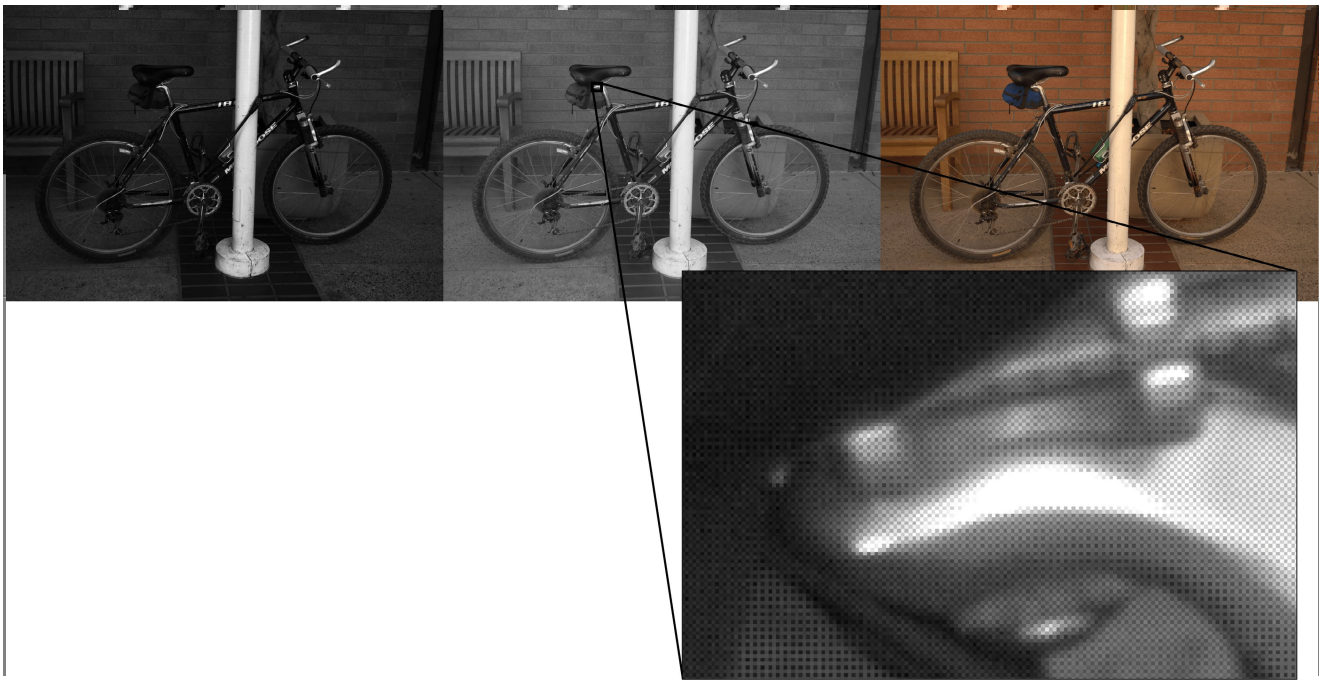


FIGURE 4. Figure showing three images, each in different formats - Bayer, Bayer with gamma correction, and RGB. The image on the left is a Bayer image with no ISP processing. The center image is a Bayer image with only gamma correction, zoomed in to showcase the Bayer filter pattern. The image on the right is an ISP-processed RGB image. This particular image is taken from the PASCALRAW dataset.

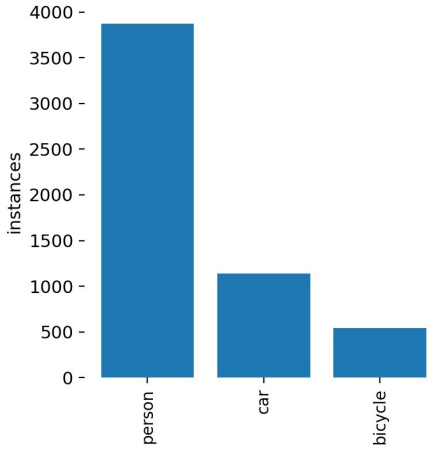


FIGURE 5. Class breakdown of the PASCALRAW dataset training set.

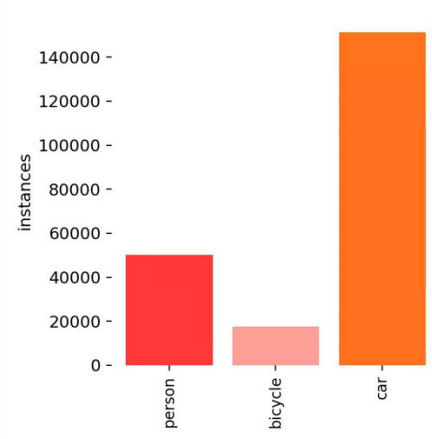


FIGURE 6. Class breakdown of the G-MIND dataset training set.

the proposals. The key technique is the sharing of the same convolutional layers for the RPN and Fast R-CNN detector up to their own fully connected layers. Now the network is only required to pass through the layers once to produce and refine object proposals.

D. MODEL TRAINING

The class balance for training for both the PASCALRAW, and the G-MIND datasets are shown in Figure 5, and Figure 6, respectively. The primary variable in this research when comparing model performance is whether each image has been processed by an ISP, thus model hyperparameters were kept as default wherever possible. Each model was trained with default hyperparameters for 50 epochs, the point where the Mean Average Precision (mAP) reaches a plateau, for a fair comparison. It is worth noting that by tuning individual hyperparameters, higher performance may potentially be achieved for a specific image set, however, that is not the purpose of this research.

Figure 7 shows the training Precision-Recall curves for the PASCALRAW dataset, while Figure 8 shows the training Precision-Recall curve for the G-MIND dataset.

E. EVALUATION METRICS

The performance of each model was evaluated using mAP@0.5, and mAP@0.5:0.95, using the definitions in [33].

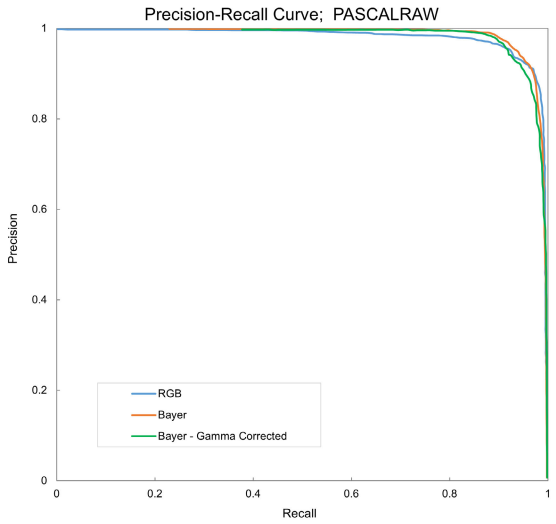


FIGURE 7. Precision-Recall curve for the PASCALRAW dataset, for the YOLO object detection model.

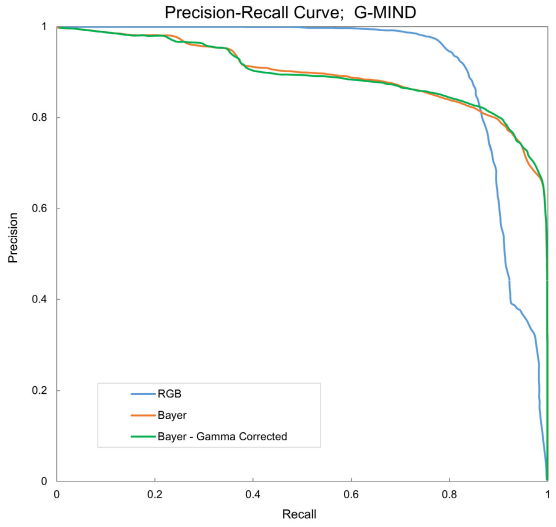


FIGURE 8. Precision-Recall curve for the G-MIND dataset, for the YOLO object detection model.

To determine whether each detection is a true positive or a false positive, most object detection challenges use both a confidence score and the Intersection of the bounding boxes over the Union of the bounding boxes (IoU) [34]. A confidence score ranging between 0 and 1 is provided by the model per detection, representing how likely the classification of an object is to be correct. Precision is defined as the number of true positives divided by the sum of true positives and false positives. In this case, the denominator is equal to the total number of detections made by the model, whether correct or not. Similarly, Recall is defined as the number of true positives divided by the sum of true positives and false negatives. This denominator is equal to the number of ground truth instances.

Mean Average Precision (mAP) is calculated by observing the graph of precision versus recall for each class. Average Precision (AP) is defined as the precision averaged across all unique recall levels, or in essence, the area under the P-R curve for a single class. mAP is defined as the mean of AP across all classes, and a high mAP is a good indicator of how well a model performs when compared to other models applied to the same test dataset.

The value mAP@0.5 refers to a mean Average Precision calculated using an IoU threshold of 0.5. As this value increases, the model is required to make a prediction closer to the ground truth. Furthermore, mAP@0.5:0.95 is a stricter metric that tests the model's performance at different IoU threshold increments (in this case, from 0.5 to 0.95 in steps of 0.05), and takes the average at each increment. This inherently lends a greater amount of importance to localization when compared with mAP@0.5. In this study, we use mAP and mAP0.5:0.95 as our key performance indicators for model accuracy.

V. MODEL PERFORMANCE COMPARISON

Table 2 presents object detection results where training and evaluation are both performed on the same image type for the PASCALRAW dataset. For example, row 2 presents the results for a YOLOv5 model trained on an RGB training set, and subsequently evaluated on a separate RGB test set. Similarly, row 4 presents the results for a YOLOv5 model trained on a Bayer image training set, with only gamma correction applied, and subsequently evaluated using a Bayer image test set, with only gamma correction applied. When observing the average precision (AP) of these models when applying an IoU of 0.5, the performance of the model varies only slightly by ∼1-2%. The variance in performance does increase by a further ∼1% when using a tighter IoU restriction taking the average at intervals between 0.5 and 0.95.

Table 3 shows the same results for the G-MIND dataset. Similar to the PASCALRAW dataset, a negligible drop in $mAP_{0.5}$ can be observed across all three image types. The Table does, however, show a decrease in $mAP_{0.5}$ performance when compared against the PASCALRAW set. This can be attributed to G-MIND being a more difficult image set, in that it contains much smaller objects. The localization ability for each model is higher when trained on the G-MIND dataset, particularly for the Faster R-CNN model, as observed by the difference between the $mAP_{0.5}$ and the $mAP_{0.5:0.95}$ for each. This is due to the G-MIND annotations being much tighter around the objects, leaving less negative space for the models to learn.

Table 4 presents object detection results for the PASCAL-RAW dataset whereby the images are trained on different image types and subsequently evaluated on an image set consisting of raw Bayer images with no ISP processing. Table 5 shows the same results for the G-MIND dataset. As one might expect, the performance is worse in instances

(a) (b) (c)

FIGURE 9. Figure showing three corresponding detections from each image type from the PASCALRAW dataset using YOLO object detection. *(a)* Bayer images with no ISP processing; *(b)* Bayer images with only gamma correction; *(c)* Full ISP processed RGB images.

TABLE 2. PASCALRAW: Object detection results from models trained on the listed training set, and subsequently evaluated using a test set consisting of images in the same listed format; GC = Gamma Correction.

	Training Set	$mAP_{0.5}$	$mAP_{0.5:0.95}$
YOLOv5	RGB	0.971	0.684
	Bayer (No GC)	0.949	0.648
	Bayer (w/ GC)	0.954	0.651
Faster R-CNN	RGB	0.961	0.697
	Bayer (No GC)	0.952	0.659
	Bayer (w/ GC)	0.959	0.673

TABLE 3. G-MIND: Object detection results from models trained on the listed training set, and subsequently evaluated using a test set consisting of images in the same listed format; GC = Gamma Correction.

	Training Set	$mAP_{0.5}$	$mAP_{0.5:0.95}$
YOLOv5	RGB	0.885	0.626
	Bayer (No GC)	0.862	0.601
	Bayer (w/ GC)	0.864	0.605
Faster R-CNN	RGB	0.905	0.767
	Bayer (No GC)	0.896	0.759
	Bayer (w/ GC)	0.885	0.714

where the training set is different from the evaluation set. The drop in performance is only slight, particularly for the easier PASCALRAW, potentially indicating how an existing model may perform if applied directly to pre-ISP images, thus eliminating the need to retrain existing models.

TABLE 4. PASCALRAW: Object detection results using models trained with the listed training set, and subsequently evaluated using a test set consisting of pre-ISP Bayer images; GC = Gamma Correction.

	Training Set	$mAP_{0.5}$	$mAP_{0.5:0.95}$
YOLOv5	RGB	0.955	0.66
	Bayer (No GC)	0.949	0.648
	Bayer (w/ GC)	0.95	0.651
Faster R-CNN	RGB	0.957	0.663
	Bayer (No GC)	0.952	0.659
	Bayer (w/ GC)	0.944	0.646

TABLE 5. G-MIND: Object detection results using models trained with the listed training set, and subsequently evaluated using a test set consisting of pre-ISP Bayer images; GC = Gamma Correction.

	Training Set	$mAP_{0.5}$	$mAP_{0.5:0.95}$
YOLOv5	RGB	0.775	0.493
	Bayer (No GC)	0.862	0.601
	Bayer (w/ GC)	0.845	0.557
Faster R-CNN	RGB	0.792	0.592
	Bayer (No GC)	0.896	0.759
	Bayer (w/ GC)	0.878	0.727

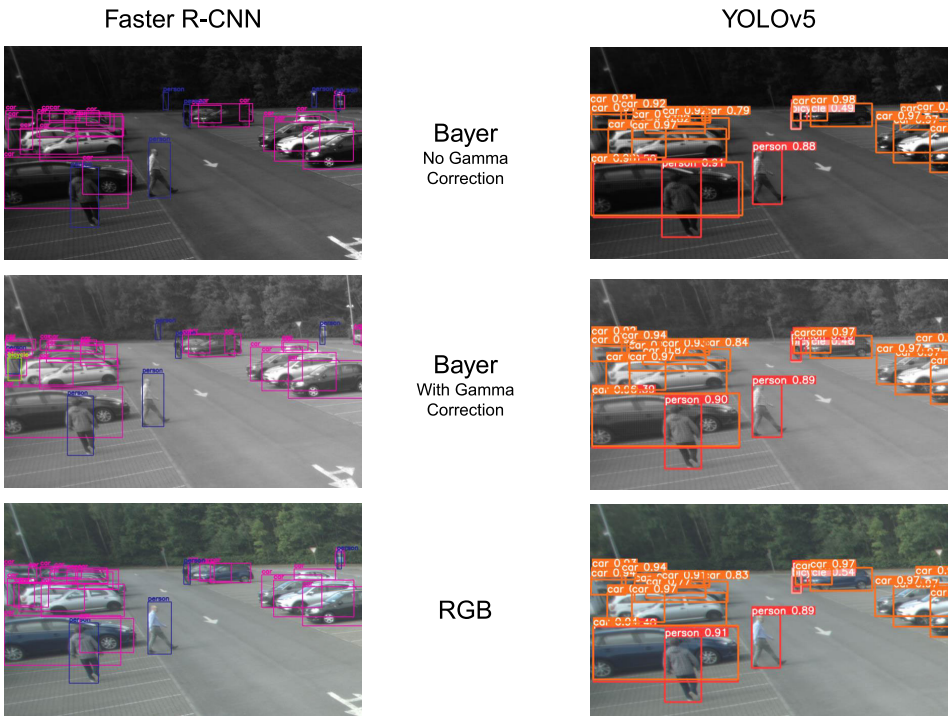


FIGURE 10. Figure showing three corresponding detections from each image type from the G-MIND dataset using both the Faster R-CNN and YOLOv5 object detection models. *Top:* Bayer images with no ISP processing; *Middle:* Bayer images with only gamma correction; *Bottom:* Full ISP processed RGB images.

Figure 9 displays sample detections for each image set from the PASCALRAW dataset. In these samples, YOLOv5 was used for detection. Figure 9(a) contains detections for raw Bayer images with no ISP processing. Figure 9(b) contains detections for Bayer images with only gamma correction, thus the images are slightly brighter than those of 9(a). Figure 9(c) contains full ISP-processed RGB images. Note that despite the images in Figure 9(a) appearing darker, and one might say less pleasing to the human eye than the other image sets, the performance is consistent across all three. Similarly, Figure 10 displays sample detections from the G-MIND dataset, using both YOLOv5 and Faster R-CNN for detection. For the sake of brevity, we do not include sample images for every permutation of image set and detection model.

VI. SYSTEM LEVEL PERFORMANCE BENEFITS

A. POWER CONSUMPTION

Power costs in a typical camera can be divided into two different components: the camera sensor and the signal processor. A study by Likamwa et al. looks at the energy consumption of a number of camera sensors across different industry use cases [35]. Typical usage ranged from 137.1 mW for a security camera, to 338.6 mW for a mobile camera, with an automotive camera listed as 189.5 mW. More modern commercial sensors range from 261 mW to 743 mW [36], [37].

Image signal processing is generally implemented in a dedicated RISC (Reduced Instruction Set Computer) or hardware-accelerated block. However, an ISP may also be implemented in software, or in a hybrid software/RISC architecture. It is therefore difficult to provide a single exact number for power saving, as it will vary from system to system. Furthermore, the power usage may vary depending on the number of cameras and other factors such as desired resolution and frame rate. For these reasons, we do not attempt to provide a single exact number for the savings made through our approach, only a theoretical overview.

OmniVision's OAX4010 ISP has a power consumption of 522 mW when measured with two OX02A10 camera sensors at 1824 × 940 @ 60 fps [38]. Each OX02A10 sensor requires an active power consumption of 450 mW [39]. Hence, eliminating the ISP in this instance provides a theoretical power saving of ∼35%.

B. COMPUTATION TIME

Computation speed is critical for safety in autonomous vehicle applications. For most use cases, the latency introduced by an ISP would not be much of an issue, but for time-critical uses such as the automotive industry, every millisecond is crucial. The amount of latency introduced by an ISP is usually not readily available outside of industry. GEO SemiConductor's GW4100 is listed as having a low latency mode, with the latency typically being $1/6^{th}$ of a frame. Arm writes that it should take no longer than 150 ms to go from an image sensor to the image being processed to the image being displayed on-screen [40]. Furthermore, Arm also claims that in a machine vision application, a vehicle

should not travel more than 250mm between a camera image being acquired and the image being presented to the decision-making process. Anything longer means the machine vision system may be too slow to react in driving situations where accurate and timely decisions are critical. In [40], Arm do not provide any data to justify these statements on latency requirements. However, based on the use cases described, these time constraints appear reasonable. To put it into context, a vehicle traveling at 50 km per hour covers 250 mm in 18 ms, with 50 km/h being a typical speed for an urban driving scenario.

VII. CONCLUSION

In this paper, we explore the potential benefits of bypassing an Image Signal Processor (ISP) when training a Convolutional Neural Network (CNN) for object detection. Typical ISPs are designed to produce an image deemed suitable for human viewing, with the optimal parameters for computer vision often being mutually antagonistic. We train two CNN-based object detection algorithms on two different pre-ISP datasets. A negligible drop in overall object detection performance can be observed when evaluating using the same datasets, with similar performance being exhibited for the two datasets considered. This indicates a potential for computer vision algorithms such as object detection to operate without the need for an ISP block, without suffering substantial performance loss. The absence of an image processor in the middle of the system also allows for the image sensor to interface directly with the processing system, allowing for computer vision to be implemented and contained entirely in the camera if desired.

We discuss the savings to be made in terms of computation time and power consumption. Furthermore, when evaluating the performance of a model pre-trained on ISP-processed RGB images, the model performs at a similar level when evaluated with a test set containing Bayer images, as with a test set containing post-ISP images. This allows for the use of already existing models to be employed on pre-ISP Bayer images.

While this paper has explored the impact and benefits of bypassing the ISP in computer vision applications with a preliminary investigation in object detection, the research presented in this paper will be expanded in future work, by observing the effects of bypassing an ISP on other computer vision applications such as semantic segmentation, and traffic sign recognition, as well as more challenging conditions such as applications in low light or adverse weather.

REFERENCES

[1] B. Wu, A. Wan, F. Iandola, P. H. Jin, and K. Keutzer, "SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 129–137.

[2] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1907–1915.

[3] Dhiraj and D. K. Jain, "An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery," *Pattern Recognit. Lett.*, vol. 120, pp. 112–119, Apr. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786551930011X

[4] A. Coates and A. Y. Ng, "Multi-camera object detection for robotics," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 412–419.

[5] D. Sarikaya, J. J. Corso, and K. A. Guru, "Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1542–1549, Jul. 2017.

[6] K.-H. Thung and P. Raveendran, "A survey of image quality measures," in *Proc. Int. Conf. Tech. Postgraduates (TECHPOS)*, Dec. 2009, pp. 1–4.

[7] L. Yahiaoui, J. Horgan, B. Deegan, S. Yogamani, C. Hughes, and P. Denny, "Overview and empirical analysis of ISP parameter tuning for visual perception in autonomous driving," *J. Imag.*, vol. 5, no. 10, p. 78, Sep. 2019. [Online]. Available: https://www.mdpi.com/2313-433X/5/10/78

[8] B. E. Bayer, "Color imaging array," U.S. Patent 3 971 065, Jul. 20, 1976.

[9] M. S. Safna Asiq and W. S. Emmanuel, "Efficient colour filter array demosaicking with prior error reduction," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1191–1199, Apr. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157819301673

[10] M. Buckler, S. Jayasuriya, and A. Sampson, "Reconfiguring the imaging pipeline for computer vision," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 975–984.

[11] W. Zhou, L. Zhang, S. Gao, and X. Lou, "Gradient-based feature extraction from raw Bayer pattern images," *IEEE Trans. Image Process.*, vol. 30, pp. 5122–5137, 2021.

[12] A. Trifan and A. J. R. Neves, "On the use of feature descriptors on raw image data," in *Proc. 5th Int. Conf. Pattern Recognit. Appl. Methods*, Feb. 2016, pp. 655–662.

[13] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[14] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314207001555

[15] E. S. Lubana, R. P. Dick, V. Aggarwal, and P. M. Pradhan, "Minimalistic image signal processing for deep learning applications," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4165–4169.

[16] M. A. Eskelinen and J. Hamalainen, "Dangers of demosaicing: Confusion from correlation," in *Proc. 9th Workshop Hyperspectral Image Signal Processing, Evol. Remote Sens. (WHISPERS)*, Sep. 2018, pp. 1–5.

[17] S. Mihoubi, P.-J. Lapray, and L. Bigué, "Survey of demosaicking methods for polarization filter array images," *Sensors*, vol. 18, no. 11, p. 3688, Oct. 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/11/3688

[18] M. Chandra and B. Lall, "A novel method for CNN training using existing color datasets for classifying hand postures in Bayer images," *Social Netw. Comput. Sci.*, vol. 2, no. 2, pp. 1–10, Apr. 2021.

[19] M. N. Ahangar, Q. Z. Ahmed, F. A. Khan, and M. Hafeez, "A survey of autonomous vehicles: Enabling communication technologies and challenges," *Sensors*, vol. 21, no. 3, p. 706, Jan. 2021.

[20] A. Omid-Zohoor, D. Ta, and B. Murmann. (2015). *PASCALRAW: Raw Image Database for Object Detection. Stanford Digital Repository*. [Online]. Available: http://purl.stanford.edu/hq050zr7488

[21] E. Bilissi, R. Jacobson, and G. Attridge, "Perceptibility and acceptability of gamma differences of displayed SRGB images," in *Proc. PICS*. Springfield, VA, USA: IS&T, 2003, pp. 120–125.

[22] A. Malta, M. Mendes, and T. Farinha, "Augmented reality maintenance assistant using YOLOv5," *Appl. Sci.*, vol. 11, no. 11, p. 4758, May 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/11/4758

[23] Y. Chen, C. Zhang, T. Qiao, J. Xiong, and B. Liu, "Ship detection in optical sensing images based on YOLOv5," in *Proc. 12th Int. Conf. Graph. Image Process. (ICGIP)*, Jan. 2021, pp. 102–106.

[24] 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)*, 2016, pp. 779–788.

[25] G. Jocher, L. Changyu, A. Hogan, L. Yu, changyu98, P. Rai, and T. Sullivan, "Ultralytics/YOLOV5: Initial release," Tech. Rep., Jun. 2020, doi: 10.5281/zenodo.3908560.

[26] G. Jocher, "Ultralytics/YOLOV5: V5.0–YOLOv5-P6 1280 models, AWS, supervise.ly and YouTube integrations," Tech. Rep., Apr. 2021, doi: 10.5281/zenodo.4679653.

[27] B. Yan, P. Fan, X. Lei, Z. Liu, and F. Yang, "A real-time apple targets detection method for picking robot based on improved YOLOv5," *Remote Sens.*, vol. 13, no. 9, p. 1619, Apr. 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/9/1619
[28] J. Zhu, X. Li, P. Jin, Q. Xu, Z. Sun, and X. Song, "MME-YOLO: Multi-sensor multi-level enhanced Yolo for robust vehicle detection in traffic surveillance," *Sensors*, vol. 21, no. 1, p. 27, Dec. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/21/1/27
[29] W. Wu, H. Liu, L. Li, Y. Long, X. Wang, Z. Wang, J. Li, and Y. Chang, "Application of local fully convolutional neural network combined with YOLO V5 algorithm in small target detection of remote sensing image," *PLoS ONE*, vol. 16, no. 10, pp. 1–15, Oct. 2021, doi: 10.1371/journal.pone.0259283.
[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28. Red Hook, NY, USA: Curran Associates, 2015, pp. 1–14.
[31] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[32] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 650–657.
[33] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 740–755.
[34] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.
[35] R. LiKamWa, B. Priyantha, M. Philipose, L. Zhong, and P. Bahl, "Energy characterization and optimization of image sensing toward continuous mobile vision," in *Proc. 11th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2013, pp. 69–82.
[36] OmniVision. *OV04690-B68Y-1C*. OVT.com. Accessed: Mar. 9, 2023. [Online]. Available: https://www.ovt.com/products/ov04690-b68y-1c/
[37] OmniVision. *OX08B40-B68Y-001D-Z*. OVT.com. Accessed: Mar. 9, 2023. [Online]. Available: https://www.ovt.com/products/ox08b40-b86y-001d-z/
[38] OmniVision. *OAX4010-B69G-1B-Z*. OVT.com. Accessed: Mar. 9, 2023. [Online]. Available: https://www.ovt.com/products/oax4010-b69g-1b-z/
[39] OmniVision. *OX02A10-E85Y-LD*. OVT.com. Accessed: Mar. 9, 2023. [Online]. Available: https://www.ovt.com/products/ox02a10-e85y-ld/
[40] C. Babla. *Arm Introduces New Automotive Image Signal Processor to Advance Adoption of Driver Assistance and Automation Technologies*. arm.com. Accessed: Mar. 9, 2023. [Online]. Available: https://www.arm.com/company/news/2022/02/arm-adds-new-isp-to-advance-adoption-of-adas-and-automation-technologies

JORDAN CAHILL received the B.Eng. degree (Hons.) from the University of Galway, in 2018, where he is currently pursuing the Ph.D. degree. He is a member of the Connaught Automotive Research (CAR) Group, under the supervision of Prof. Edward Jones and Prof. Martin Glavin. His research interests include exploring novel methods of optimizing sensor data within an autonomous vehicle context, primarily along the computer vision pipeline.

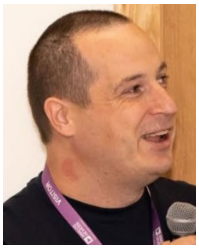
ASHKAN PARSI received the B.E. degree in computer software and the M.Sc. degree in artificial intelligence from the Shahrood University of Technology, Shahrood, Iran, in 2010 and 2013, respectively, and the Ph.D. degree in electrical and electronic engineering from the University of Galway, Ireland, in 2021. From 2012 to 2016, he was a Software Developer, a Senior Researcher, and a Technical Project Manager on several national projects with the Iran Telecommunication Research Center (ITRC). He is currently leading and conducting research in the development of signal and image processing for advanced sensors and machine vision technologies in a range of problems in connected and autonomous vehicles with Xperi Company. He has awarded the Government of Ireland Postgraduate Research Scholarship (IRC), for his Ph.D. study. His research interests include signal processing, machine learning, and algorithm design.

DARRAGH MULLINS received the B.E. degree in energy systems engineering and the Ph.D. degree in electronic engineering from the University of Galway, in 2013 and 2018, respectively. His Ph.D. research topic involved the application of imaging sensors and signal processing to wastewater treatment plant performance sensing. He was a Postdoctoral Research Fellow with the University of Galway, from 2018 to 2022, where he managed a research program and co-supervised six Ph.D. student projects, involving sensors and V2X communication systems for pedestrian and vehicle monitoring from both vehicle and fixed infrastructure point-of-view. He is currently a Senior Technical Officer and an Adjunct Lecturer with the School of Engineering, University of Galway. He was awarded the Lero Director's Prize for Education and Public Engagement, in 2020.

JONATHAN HORGAN is a Computer Vision and Deep Learning Architecture Manager and a Senior Expert with Valeo Vision Systems. He has worked in the field of computer vision for over 16 years, with a focus over the last ten years on automotive computer vision for advanced driver assistance systems (ADAS), automated parking, and automated driving. He is currently working on next-generation advanced computer vision and deep learning, with the ultimate goal of achieving fully autonomous driving and parking. He has 25 publications in peer-reviewed journals and conferences and over 100 patents published in the field of automotive computer vision.

ENDA WARD received the B.E. degree in electronic engineering from the University of Galway, in 1999, and the M.Eng.Sc. degree in electronic engineering, with a focus on biomedical electronics, in 2002.

He is responsible for defining the camera product roadmap for surround and automated driving applications with Valeo. He has worked with key technology experts across the supply chain and within OEMs to define optimal system architectures. He was lecturing for a number of years in the areas of electronics and computing systems with Atlantic Technological University, Ireland. Later, he moved into industry, working in the biomedical space and has spent the last 16 years in automotive ADAS design. He holds several patents in the area of automotive vision.



CIARÁN (HUGHES) EISING (Member, IEEE) received the degree in electronic and computer engineering and the Ph.D. degree from the University of Galway, in 2003 and 2010, respectively. From 2009 to 2020, he was a Computer Vision Team Lead and an Architect with Valeo Vision Systems, where he also held the title of a Senior Expert. In 2016, he was awarded the position of an Adjunct Lecturer with the University of Galway. In 2020, he joined the University of Limerick.

PATRICK DENNY (Member, IEEE) received the B.Sc. degree in experimental physics and mathematics from NUI Maynooth, Ireland, in 1993, the M.Sc. degree in mathematics from the University of Galway, Ireland, in 1994, and the Ph.D. degree in physics from the University of Galway, in 2000, while researching electromagnetic planetary physics with GFZ Potsdam, Germany.

From 1999 to 2001, he was an RF Engineer with AVM GmbH, Germany, where he has designed and developed the hardware for the first integrated GSM/ISDN/USB modem. After working in supercomputing development at Compaq/Hewlett Packard, from 2001 to 2002, he joined Connaught Electronics Ltd. (subsequently Valeo), Tuam, Ireland, as a Team Leader of radiofrequency design. For over 20 years, he was a Senior Expert with Valeo, designing and developing novel radiofrequency and imaging systems, including the first mass production high dynamic range automotive camera for leading car companies, including BMW, Daimler, and Land Rover. In 2010, he became an Adjunct Professor of automotive electronics with the University of Galway. In 2022, he became a Lecturer of artificial intelligence with the Department of Electronic and Computer Engineering, University of Limerick, Ireland. His research interests include automotive imaging technology and its extension into other domains, algorithmic designs, artificial intelligence, applied mathematics, and the industrialization of advanced technologies.

BRIAN DEEGAN (Member, IEEE) received the bachelor's degree in computer engineering and the M.Sc. degree in biomedical engineering from the University of Limerick, in 2004 and 2005, respectively, and the Ph.D. degree in biomedical engineering from the University of Galway, in 2011. The focus of his research was the relationship between blood pressure and cerebral blood flow in humans. From 2011 to 2022, he was with Valeo Vision Systems, as a Vision Research Engineer, focusing on image quality. In 2022, he joined the Department of Electrical and Electronic Engineering, University of Galway, as a Lecturer and a Researcher. His research focus is on high dynamic range imaging, LED flicker, top-view harmonization algorithms, and the relationship between image quality and machine vision.

MARTIN GLAVIN (Member, IEEE) received the B.E. degree in electronic engineering and the Ph.D. degree in algorithms and architectures for high-speed data communication systems from the University of Galway, Ireland, in 1997 and 2004, respectively, and the Higher Diploma degree in third level education, in 2007. He was a Lecturer (fixed term contract), from September 1999 to December 2003, and became a permanent member of the Academic Staff, in January 2004. He is currently the Joint Director of the Connaught Automotive Research (CAR) Group, University of Galway. He is also a Funded Investigator in Lero, the Irish Software Research Centre. He currently has a number of Ph.D. students and postdoctoral researchers, working in collaboration with industry in the areas of signal processing and embedded systems for automotive and agricultural applications.

EDWARD JONES (Senior Member, IEEE) received the B.E. and Ph.D. degrees in electronic engineering from the University of Galway, Ireland. His Ph.D. research was on the development of computational auditory models for speech processing. He is currently a Professor of electrical and electronic engineering with the School of Engineering, University of Galway. From 2009 to 2010, he was a Visiting Researcher with the Department of Electrical Engineering, Columbia University, New York, NY, USA. He was a Visiting Fellow with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia. He also has a number of years of industrial experience in senior positions, in both start-ups and multinational companies, including Toucan Technology Ltd., PMC-Sierra Inc., Innovada Ltd., and Duolog Technologies Ltd. He also represented Toucan Technology and PMC-Sierra on international standardization groups ANSI T1E1.4 and ETSI TM6. His current research interests include DSP algorithm development and embedded implementation for applications in biomedical engineering, speech and audio processing, and image processing. He is a Chartered Engineer and a fellow of the Institution of Engineers of Ireland.