

γ -QUANT: Towards Learnable Quantization for Low-bit Pattern Recognition

Mishal Fatima^{1*}, Shashank Agnihotri^{1*}, Marius Bock^{2*}, Kanchana Vaishnavi Gandikota², Kristof Van Laerhoven², Michael Moeller², and Margret Keuper^{1,3}

¹ University of Mannheim

² University of Siegen

³ MPI for Informatics, Saarland Informatics Campus

Abstract. Most pattern recognition models are developed on pre-processed data. In computer vision, for instance, RGB images processed through image signal processing (ISP) pipelines designed to cater to human perception are the most frequent input to image analysis networks. However, many modern vision tasks operate without a human in the loop, raising the question of whether such pre-processing is optimal for automated analysis. Similarly, human activity recognition (HAR) on body-worn sensor data commonly takes normalized floating-point data arising from a high-bit analog-to-digital converter (ADC) as an input, despite such an approach being highly inefficient in terms of data transmission, significantly affecting the battery life of wearable devices. In this work, we target low-bandwidth and energy-constrained settings where sensors are limited to low-bit-depth capture. We propose γ -QUANT, i.e. the task-specific *learning* of a non-linear quantization for pattern recognition. We exemplify our approach on raw-image object detection as well as HAR of wearable data, and demonstrate that raw data with a learnable quantization using as few as 4-bits can perform on par with the use of raw 12-bit data. All code to reproduce our experiments is publicly available via github.com/Mishalfatima/Gamma-Quant.

1 Introduction

Deep learning techniques have revolutionized the performance of numerous pattern recognition tasks in the last decade by training on large-scale image datasets. Yet, comparably little attention has been paid to the type and quantization of the input data. In computer vision, for instance, most pipelines consider pre-processed sRGB images with a standard bit depth of 8 bits. In the imaging process, cameras usually capture visual information in a higher-bit-depth RAW format which is converted to the standard format by an image signal processor (ISP) using a series of operations including black light subtraction, demosaicking, denoising, white balancing, gamma correction, color manipulation, and tone-mapping to finally obtain a visually pleasing 8-bit sRGB image. As photography-oriented ISP pipelines may not be optimal for vision tasks, recent works [30, 13,

*Equal contribution.

38, 33] have also successfully optimized ISP pipelines together with the downstream vision task. Yet, the idea to explicitly optimize the quantization for a given (automated) machine learning task has not been exploited so far.

Similarly, most studies in human activity recognition (HAR) from body-worn data simply use linearly quantized high-bit (e.g. 12-bit) information from the sensors. While settings of lower bit quantizations have been investigated (see [3, 10]), the idea to *learn* an optimal quantization has not been studied.

In both settings, computer vision and HAR, the quantization of the analog data into a digital signal plays a critical role in balancing data quality, memory requirements, and energy consumption at the analog-to-digital converter (ADC), see e.g. [23]. Moreover, significant bandwidth can be saved if the recorded data is sent to the cloud in a low-bit format for further analysis.

In this paper, we demonstrate that a tailored *learned* quantization has significant advantages over a naïve linear quantization. Specifically, we study a learnable non-linear quantization via

$$\mathcal{Q}(\mathcal{X}, \gamma, \mu) = Q_{\hat{N}}(\text{sign}(\mathcal{X} - \mu) \cdot |\mathcal{X} - \mu|^\gamma) \quad (1)$$

for (normalized) analog input values \mathcal{X} , a linear quantizer $Q_{\hat{N}}$ to a target bit-depth of \hat{N} bit, and learnable parameters γ as well as an offset μ . This learnable non-linear quantization is optimized together with neural networks for specific tasks. We refer to our approach (1) as γ -QUANT.

We demonstrate that γ -QUANT improves the performance of object detection on raw data on diverse vision datasets like the PASCAL-RAW dataset [32] and the RAOD dataset [46] by simulating γ -QUANT on analog signals by using raw (high bit depths) images of the respective datasets. We show generality of the approach by conducting a similar study for a completely different modality, i.e., inertial, body-worn sensor data, using different datasets commonly used in HAR.

In summary, the contributions of this work are as follows:

- We show that naïve low-bit quantization of accelerometer data as well as of images harms model performance.
- We propose γ -QUANT, a learnable non-linear quantization (parameterized similar to a gamma correction), as a solution.
- We demonstrate that our proposed method allows reducing the recorded data to up to 4-bit for object detection and even 2-bit for human activity recognition without significant performance drops in comparison to high-bit data. Moreover, we demonstrate that learning the quantization via γ -QUANT yields systematic improvements over a classical (linear) quantization.

2 Related Work

Work in the direction of jointly optimizing sensor and neural network parameters is limited, though some substantial contributions have been made recently. We summarize them in our two application domains, computer vision and wearable sensor data analysis, separately.

2.1 Quantization of Wearable Sensor Data

Energy efficiency is crucial for wearable human activity recognition on edge devices with limited battery capacity. Transmission of HAR data accounts for significant energy consumption at the wearable device [22] which can be minimized by reducing expensive communications, for instance, by aggregating or compressing data, and doing on-device feature extraction and classification to avoid sending raw signals. Recent work has focused on efficient inference of neural networks for on-device HAR using pruning [24], adaptive inference [35] and quantization [12, 11, 49]. Unlike the quantization of HAR data, network quantization has been widely studied. Exemplary approaches include sub-byte and mixed-precision quantization with adaptive inference in 1D CNNs [11], full-integer quantization of DeepConv LSTM [49], and binary quantization of weights and activations in neural networks [12]. Orthogonal to these techniques, power savings can also be achieved by turning off the sensors when inactive or lowering their sampling rates [31, 48] or adapting sampling rates per activity [47, 9]. Further techniques have also been proposed to handle such HAR data captured at variable sampling rates, including modifications to neural network architectures [28], and data augmentation [16]. Unlike these techniques, we address the often-overlooked challenge of reducing energy use during data capture by applying low-bit quantization directly at the inertial sensor, which can provide task-aware data compression immediately at acquisition, complementing the existing energy-saving techniques.

2.2 Codesigning Imaging System and Vision Models

Instead of the traditional approach where imaging hardware and perception models are developed independently, a recent trend in efficient machine learning is to code-sign imaging systems and computer vision models [20], creating tightly integrated solutions that maximize performance while reducing the hardware or computing requirements. [21] formulate imaging building blocks as context-free grammar whose parameters can be optimized through a reinforcement learning framework. [14, 7, 6] jointly optimize for the downstream computer vision model along with the optical layer to exploit its potential computation capability. [40] proposes a differentiable approach to jointly optimize the size and distribution of pixels on the imaging sensor along with the downstream computer vision model. However, none of these works deal with quantization at the sensor, which is the focus of our work.

RAW images contain more information than standard RGB images (sRGB), which could potentially be beneficial for higher-level vision tasks such as object detection [27, 46, 45]. In this context, a hardware-in-the-loop method is introduced in [30] to optimize hardware ISP for end-to-end task-specific networks by using zero-order optimization. Diamond *et al.* [13] use Anscombe networks as neural ISPs which are jointly trained with task-specific neural networks. [29] train a minimal neural ISP pipeline for object detection that improves generalization

to unseen camera sensors. Robidoux *et al.*[38] optimize HDR ISP hyperparameters together with detector network, whereas [33] train a neural network for automatic exposure selection that is trained jointly with ISP pipeline and a network for HDR object detection. Instead of using the whole ISP, [3, 46] identifies the key components of the ISP for downstream vision tasks. Prior works [27, 46, 26] learn specific parameterized ISP functions such as demosaicking and gamma correction, color correction in an end-to-end fashion along with the object detector. While some of these works [27, 46] learn a gamma correction, this is a part of the ISP pipeline after a digital image has been obtained, while we propose to learn this function before quantization at the sensor for the analog input. While [3, 10] consider the effect of low bit quantization on vision tasks, they do not optimize this process. Instead of learning fixed ISP parameters, [44] introduces a scene-adaptive ISP to automatically generate an optimal ISP pipeline and the corresponding ISP parameters to maximize the detection performance. Yet, the learned computational steps are applied to the digital (= already quantized) image. To the best of our knowledge the *learning* of a quantization to be applied within the analog-to-digital converter of a camera, has not been considered before.

3 Our Approach γ -QUANT

3.1 Preliminaries

Many sensors use a physical effect to induce a voltage which is converted to a digital measurement by an analog-to-digital converter (ADC). For instance, variable capacitance Micro-Electro-Mechanical Systems (MEMS) accelerometers that are present in many mobile and wearable systems, sense tiny mass' distance changes between two capacitor plates. An in-chip ADC converts the resulting amplified voltages to quantized digital values that are handled in a local processing unit. Similarly, in imaging, a CMOS sensor uses photodiodes (and the photo-electric effect) to induce a charge, which is transferred to a capacitor. The charge is converted to a voltage, amplified and passed to the ADC. For both modalities the ADC typically uses a linear quantization with at least 8, more commonly 12 or even 16 bits per value. Yet, both the energy consumption as well as the readout speed are crucially influenced by the bit-depth, such that a reduction can have significant benefits: According to [43], simply reducing the image bit-depth from 16-bit to 12-bit has shown speedups by a factor of two in industry. It was shown in [5] that – for energy-constrained sensors – ADCs alone contribute up to 50% of the energy consumption in an image sensor. Similarly, in wearable sensorics, transferring the recorded accelerometer data wirelessly has shown to be the by far most energy-consuming operation, such that a reduced bit-depth has an immediate and significant effect on the wearable's battery life.

3.2 A Learnable Quantization Approach

The most straight-forward ADCs produce lower bit-depths via linear quantization, i.e., converting each analog value \mathcal{X} (assumed to be normalized to $[0, 1]$) to

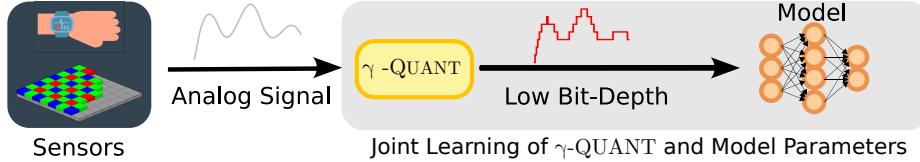


Fig. 1. γ -QUANT learns the quantization of an ADC to convert the analog signal of a sensor to a low bit depth digital signal, which is subsequently sent to the neural network for performing the downstream task. The parameters of the quantization are trained jointly with the parameters of the neural network in a task-specific fashion. γ -QUANT can reduce the energy consumption of the sensor significantly with a minimal loss in performance.

a digital value \mathcal{X}_Q via

$$Q_{\hat{N}}(\mathcal{X}) = \left\lfloor \mathcal{X} \cdot (2^{\hat{N}} - 1) \right\rfloor, \quad (2)$$

where \hat{N} is the desired bit depth, and $\lfloor \cdot \rfloor$ is the floor operation.

In imaging, the Image Signal Processor (ISP) converts the raw digital values to standard RGB (sRGB) images through a sequence of pre-processing steps including demosaicking, denoising, white-balancing, color conversion, and tonemapping. Yet, low bit-depth raw-images often lead to poor visual quality. As the human visual system rather scales logarithmically, [3, 1] proposed to scale the analog signals before quantization via

$$\mathcal{X}_{\log} = \log(\mathcal{X} + \epsilon), \quad (3)$$

where ϵ is required to bound the input to the logarithm from below. [3] use $\epsilon=1$ and quantize the resulting signal via

$$\mathcal{X}_Q = \left\lfloor \frac{\mathcal{X}_{\log} - \min(\mathcal{X}_{\log})}{\max(\mathcal{X}_{\log}) - \min(\mathcal{X}_{\log})} \cdot (2^{\hat{N}} - 1) \right\rfloor. \quad (4)$$

Yet, the quantization is ad-hoc, independent of the target bit-depth \hat{N} , and the effect of ϵ heavily depends on the dynamic range of the analog signal, such that it might need domain- and downstream-task specific adaption.

For body-worn accelerometer data, values are typically in $[-1, 1]$ (up to a scaling) as accelerations for each axis can happen in two (opposite) directions, reflected by different signs. Thus, a logarithmic quantization is not straightforward. Moreover, accelerometers might constantly yield a non-zero signal in the gravitational field of the earth such that zero might not be a natural candidate for the finest quantization anymore.

For these reasons, we propose γ -QUANT, i.e., an automated approach for *learning* the quantization of sensor signals in an automated and *task-specific* way, see Fig. 1. More precisely, we propose to learn a non-linear low-bit ADC parameterized via (1) along with a neural network for a specific task by optimizing

$$\min_{\theta, \gamma, \mu} \mathbb{E}_{(\mathcal{X}, \mathbf{y})} [\mathcal{L}(\eta(Q(\mathcal{X}, \gamma, \mu); \theta), \mathbf{y})], \quad (5)$$

where η is the task-specific neural network with parameters θ , γ and μ are the parameters of the learnable ADC (i.e., the quantizer \mathcal{Q}), and \mathcal{L} is a suitable loss function comparing the network output and the ground truth prediction y .

To optimize (5), we propose to *simulate* the learnable ADC $\mathcal{Q}(\mathcal{X}, \gamma, \mu)$ by using readily available high-bit data $\hat{\mathcal{X}}$ as an approximation to the analog signal \mathcal{X} . Since the quantization operation stops the flow of gradients in the backward pass, we use a straight-through estimator to allow gradient-based optimization of γ and μ .

While a simulation enables the training of (5) with standard first-order methods on large data sets, the resulting learned (non-linear) ADCs can be realized in hardware: For body-worn sensors such as 3D accelerometer MEMS, the in-chip Digital Signal Processor can be re-configured to adjust ADC settings such as the resolution or sensitivity range through internal registers. These subsequently alter the step size or quantization level of the digital output. Custom transfer functions or lookup tables can be implemented in the sensor to apply non-linear scaling. For CMOS image sensors non-linear quantization in the ADC can also be realized as demonstrated in [15, 4]. Thus, our learnable quantization framework targets specific application (such as smart watches for human activity recognition or object detection cameras in an autonomous driving setting) where dedicated hardware is built/programmed using the task-specific learned quantization.

3.3 Specifics of γ -QUANT for HAR and Object Detection

When applying γ -QUANT to imaging applications, we built upon findings of prior works (e.g. [3]) that low intensity values are important to resolve in a more fine-grained manner. Therefore fix the offset $\mu = 0$, and simplify γ -QUANT to

$$\mathcal{Q}(\mathcal{X}, \gamma) = Q_{\hat{N}}(\mathcal{X}^\gamma) = \left\lfloor \mathcal{X}^\gamma \cdot (2^{\hat{N}-1}) \right\rfloor, \quad (6)$$

where we assume \mathcal{X} to be normalized to $[0, 1]$. More specifically, when simulating \mathcal{X} from a high-bit digital signal, we divide the digital signal by $2^N - 1$ for an N -bit signal. For a faithful simulation of an analog signal, N needs to be significantly larger than our target bit depth \hat{N} . The resulting γ -QUANT approach can learn to resemble a simple linear quantization ($\gamma = 1$), compress large values more strongly ($\gamma < 1$), or compress small values more strongly ($\gamma > 1$), see Fig. 2, right.

When applying γ -QUANT to body-worn sensors for HAR, we use the full approach (1) as data is typically normalized to $[-1, 1]$ and offsets can be important. Explicitly writing out the linear quantizer for $[-1, 1]$ our approach becomes

$$\mathcal{Q}(\mathcal{X}, \gamma, \mu) = \text{round} \left(((\text{sign}(\mathcal{X} - \mu) \cdot |\mathcal{X} - \mu|^\gamma + 1) / 2 \cdot (2^N - 1)) \right), \quad (7)$$

resulting in exemplary curves illustrated in Fig. 2 on the left. As the function (7) is non-differentiable at $\mathcal{X} = \mu$ for $\gamma < 1$, we found the replacement of $|\mathcal{X} - \mu|^\gamma$ by $(|\mathcal{X} - \mu| + \epsilon)^\gamma$ for a small ϵ , e.g. $\epsilon = 10^{-3}$ to stabilize the training process.

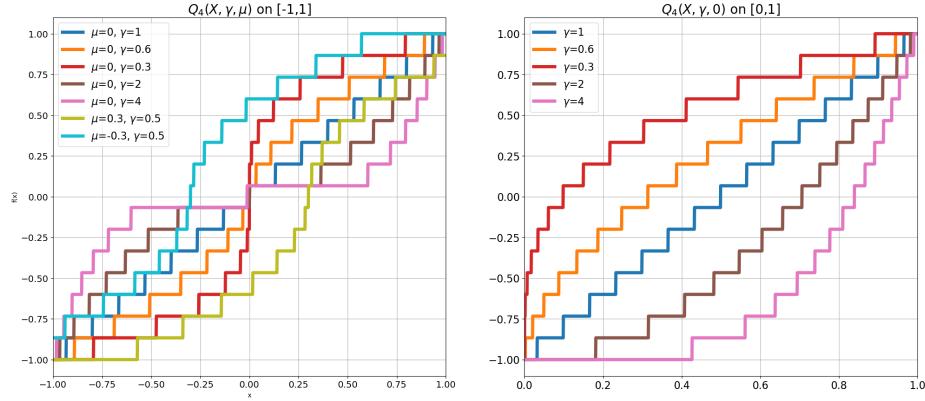


Fig. 2. Exemplifying different γ -QUANT quantization for HAR on $[-1, 1]$ data (left) and object detection on $[0, 1]$ data (right).

4 Experiments on Learnable Quantizations for Accelerometer-based Human Activity Recognition

As a first use case of our γ -QUANT we focus on inertial sensors, specifically accelerometers used in the context of Human Activity Recognition. The following will outline datasets used as well as experiments conducted. As accelerations measured by wearable sensors can go in both directions, value ranges of the investigated datasets are between $[-1, 1]$. We thus apply the version of γ -QUANT for wearable accelerometer data as shown in (7).

4.1 Datasets

In total, we investigate five wearable accelerometer datasets (see Table 1). The WEAR dataset [2] records participants outdoors, while performing a set of workout-related activities such as running, stretching and strength-based activities. Similarly, the Hang-Time [18] dataset records a team of basketball players during their practice session consisting of a warm-up, drill and game session. Recorded in a biology wet lab, the Wetlab dataset [39] records recurring activities, such as pipetting, occurring during a DNA extraction experiment. Lastly, the RWHAR [42] and SBHAR dataset [37] has participants consists of various of locomotion activities such as walking stairs, with the SBHAR dataset providing annotations of additional transitional activity periods that mark the transition from one to another activity. Note that all but the RWHAR dataset provide continuous recording data, thus providing an additional *NULL*-class which represents times during the recording participants did not perform any of the activities of relevancy.

Table 1. Investigated HAR datasets. Table provides: participant count, activity count (classes), sensor axes count and overall recording scenario. Each sensor axes provides accelerometer data sampled at 50Hz.

Dataset	participants	classes	axes	scenario
WEAR [2]	18	19	12	body-weight workout
Wetlab [39]	22	9	3	laboratory
Hang-Time [18]	24	6	3	basketball
RWHAR [42]	15	8	21	locomotion
SBHAR [37]	30	13	3	locomotion + transitional

4.2 Experimental Setup and Implementation

During experiments, we follow a Leave-One-Subject-Out (LOSO) cross-validation, where each participant in the dataset is used as the validation set exactly once, while all other participants are used for training. We report the class-averaged macro F1-score averaged across all validation splits (i.e., participants). We further seed-average our experiments repeating each experiment using a set of three different random seeds. For all experiments we employ a sliding window of one second with a 50% overlap, normalize all accelerometer signals between $[-1, 1]$ using min-max normalization applied across the full dataset, a weighted cross-entropy loss and the Adam optimizer, with a learning rate of $1e^{-4}$, weight decay of $1e^{-6}$. We train for 30 epochs with a batch size of 100, using a learning rate schedule that multiplies the learning rate by a factor of 0.9 every 10 epochs. As a model architecture of choice we use the DeepConvLSTM, a widely-adopted HAR model [34]. For each dataset we compare results using linear quantized and γ -QUANT quantized accelerometer signals as input, differing bit depths to be either 2 or 4. We further provide results using the raw accelerometer signal as provided by the datasets. We further compare a dataset-wide versus sensor-axis-specific learnable quantization using γ -QUANT, i.e., learning a (γ, μ) -pair for each sensor axis in the dataset. In all experiments using γ -QUANT we initialize γ with 0.4, as we generally estimate differences accelerations close to 0, i.e., fine-grained movements, to contain more information than the differences in large accelerations, i.e., strong movements.

4.3 Evaluation Results

Table 2 presents the per-dataset results of the experiments described above. It is evident that with 4-bit linear quantization, performance on the WEAR and Wetlab dataset drops significantly compared to using raw data. With 2-bit linear quantization, all five investigated HAR datasets witness a substantial decline in performance. However, across all datasets, γ -QUANT consistently outperforms models trained on linearly quantized data. In particular, for the WEAR and Wetlab datasets, sensor-axis-specific learnable quantizations lead to notable improvements, achieving prediction F1-scores comparable to raw data, even with 2-bit input.

Table 2. Per-dataset HAR results comparing training using raw data with training using linear, dataset-wide γ -QUANT or per-axis γ -QUANT quantized accelerometer signals. We provide results for bit depths $\hat{N} = 2$ and $\hat{N} = 4$.

		WEAR	Wetlab	Hang-Time	RWHAR	SBHAR
	<i>Raw data</i>	71.01 ± 0.18	25.61 ± 0.07	33.88 ± 0.09	71.21 ± 0.90	54.58 ± 0.22
$\hat{N}=4$	linear	63.74 ± 0.22	21.72 ± 0.18	33.69 ± 0.22	69.78 ± 0.63	52.14 ± 0.26
$\hat{N}=4$	γ -QUANT	70.14 ± 0.51	23.00 ± 0.24	33.63 ± 0.17	69.28 ± 0.46	52.58 ± 0.20
$\hat{N}=4$	γ -QUANT(per-axis)	70.17 ± 0.39	23.63 ± 0.20	33.67 ± 0.17	71.16 ± 1.26	52.52 ± 0.21
$\hat{N}=2$	linear	58.91 ± 0.13	12.62 ± 0.34	28.01 ± 0.14	64.82 ± 0.62	33.02 ± 0.39
$\hat{N}=2$	γ -QUANT	60.45 ± 0.43	12.94 ± 0.18	30.67 ± 0.12	71.41 ± 0.51	40.18 ± 0.32
$\hat{N}=2$	γ -QUANT(per-axis)	63.95 ± 0.14	16.85 ± 0.54	30.60 ± 0.22	69.21 ± 0.32	41.88 ± 0.32

Figure 3 illustrates representative quantization functions learned during HAR experiments. The left subplot shows that different datasets yield distinct global scaling factors γ , while the learnable offset μ remains close to zero across datasets. However, as demonstrated in the right subplot for the Wetlab dataset, if γ -QUANT is learned individually on a per-axis level, both a negative offset ($\mu < 0$) for the x-axis as well as a positive offset ($\mu > 0$) for the z-axis of the sensor are learned. Supported by the overall pattern of improved prediction results of the per-sensor application of γ -QUANT, we hypothesize that the learned offsets allow a better compensation of the static gravitational components captured by fixed accelerometers, where certain axes exhibit constant acceleration depending on their orientation relative to the ground. This compensation works especially well for sensors whose orientation is rather static, e.g., because they are placed at body parts that do not exhibit strong rotations, e.g., the chest or waist of a participant, or capture mostly activities, which do not exhibit strong movements such as sitting. Conversely, one can only expect smaller gains from per-axis quantization on sensors which exhibit frequent changes in orientation due to fast movements of e.g. sport-related activities such as present in the Hang-Time dataset, as gravitational components do not remain static.

5 Experiments on Learnable Quantizations for Object Detection on CMOS Sensor Data

5.1 Datasets

We use the PASCAL RAW dataset [32] having a bit-depth of 12, with three classes, namely ‘Car’, ‘Bicycle’, and ‘Person’. The training data consists of 2129 images, and the test data is 2130 images. Experiments on a second data set, RAOD [46], can be found in the appendix.

5.2 Experimental Setup and Implementation

We use the mmdetection [8] framework from openmmlab for conducting experiments. Following previous works [46, 32] that perform object detection on

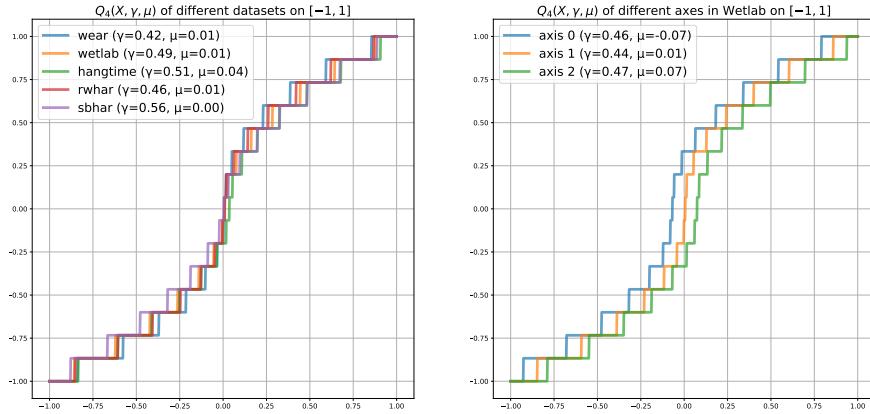


Fig. 3. Visualization of exemplary learned quantization function using γ -QUANT. The left plot shows the per-dataset learned quantization functions. The right plot shows the per-axis quantization functions when predicting the Wetlab dataset. γ and μ values in both plots are averaged across validation splits.

RAW images, we used Faster-RCNN [36], and PAA [19] models with pre-trained ResNet50 [17] backbone. We conduct experiments by quantizing the Bayer patterned higher bit-depth RAW images to lower bit depths of 4, 6, 8, and 12 using γ -QUANT and compare against Linear quantization (Linear Quant) as well as the logarithmic quantization (4). The Bayer pattern RGGB image is viewed as a four-channel image, subsampled, and converted to a three-channel RGB image by averaging the two green channels. All the images are scaled to a range of $[0, 255]$ before being used as input to the neural network. We report the following quantitative evaluation metrics: mean average precision (mAP) is calculated by averaging the Average Precision (AP) values across the intersection over union (IoU) thresholds ranging from $[0.5, 0.95]$ with a step size of 0.05, resulting in 10 threshold values. We also report performance with IoU-thresholds of 0.5 and 0.75 (AP50 and AP75) for each case.

Following the multi-scale training setup commonly used [36, 25], during training, the shorter image side is scaled to one of the sides randomly selected from a set of sides: $[480, 512, 544, 576, 608, 640, 672, 704, 736, 768, 800]$ using nearest neighbor interpolation and the longer side is scaled to maintain the aspect ratio. At test time, the shorter edge length is kept at 800. We use a batch size of 16 and train the FasterRCNN model for 140 epochs and PAA for 70 epochs. For FasterRCNN, we apply warm-up for the first 1000 iterations, linearly increasing the lr from $1e^{-4}$ to $2.5e^{-3}$ followed by a multi-step learning rate scheduler. We use SGD with Nesterov Momentum [41] as the optimizer with a weight decay of $1e^{-4}$. For PAA, we apply a warmup strategy for the first 4000 iterations, followed by cosine annealing for the learning rate schedule. The base learning rate is set to $1e^{-3}$, with a weight decay of $1e^{-3}$.

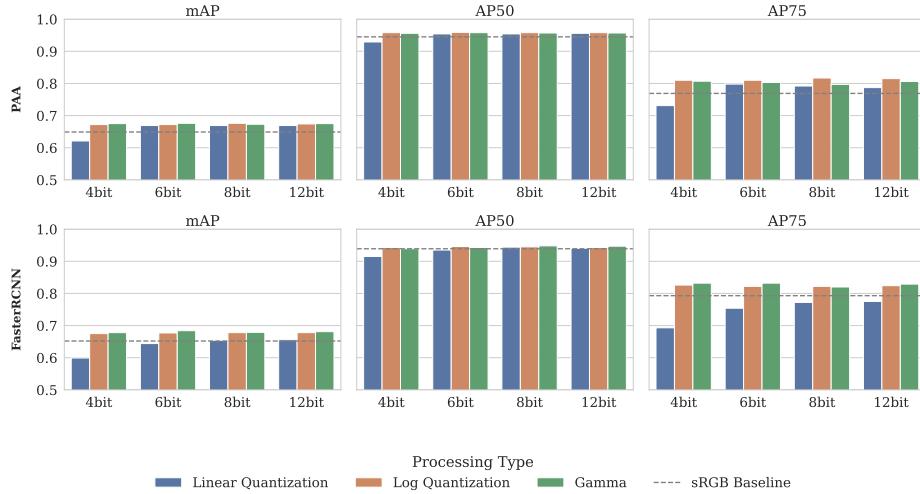


Fig. 4. Results with γ -QUANT, log and linear quantization on PAA and FasterRCNN for different bit depths. The blue dashed lines represented the performance of the model on sRGB input images.

5.3 Evaluation Results

To demonstrate the effectiveness of the proposed γ -QUANT, we design a set of experiments across 4 different quantization levels, i.e., 4, 6, 8, 12 bits. For every quantization level, we run three quantization methods, namely, linear quantization, logarithmic quantization as shown in (4), and γ -QUANT. The quantitative results obtained for different bit depths using the three models on PASCAL RAW are presented in Fig. 4.

We observe that performance with *4-bit linear quantization is the worst* across all architectures. Low-bit linear quantization is especially hurting details in lower luminance regions, which account for a majority of the intensity values and therefore result in the lowest performance. Results with γ -QUANT show systematic improvement across both architectures. In particular, there is surprisingly little difference between the different bit depths, indicating that - although a 4-bit image might not be visually pleasing - it is a sufficient bit depth for faithful object detection. Moreover, it is highly encouraging that the performance on standard (ISP processed) RGB images is met or even surpassed: γ -QUANT learns to scale the distribution of pixel values such that low-intensity pixels are amplified, thus providing more contrast in the input. This allows features and details to be better visible, enabling the model to learn more informed feature representations.

Our experiments demonstrate that the log-quantization (4) performs on par with our proposed γ -QUANT. While this might be discouraging at first glance, such results are based on a well chosen value of ϵ , i.e., $\epsilon = 1$ when simulating analog signals with digital 12-bit raw data or, correspondingly, $\epsilon \approx 0.00024414$

for analog signals scaled to $[0, 1]$. More concretely, for an analog signal, there is no natural ϵ , such that it becomes a hyperparameter to be tuned. Our framework can be seen as an automatic (differentiable) way of learning such a hyperparameter. In particular, ϵ in log-quantization could also be learned in the same framework. While this (and further much more flexible) parametrizations of the quantization are an interesting direction for future research, we decided to study γ -QUANT for the sake of simplicity: The range of the output in the log-quantization depends on ϵ such that a rescaling to $[0, 1]$ needs to be included in the learning. Our γ -QUANT automatically preserves the $[0, 1]$ range. We thus consider the fact that γ -QUANT reaches the performance of a logarithmic quantization to be encouraging. In particular, for FasterRCNN, the learned values of γ are 0.294, 0.338, 0.354, 0.359 for 4, 6, 8, and 12 bits, respectively. Thus, the logarithmic shape of the curve is learned to be optimal and does not need to be derived from prior assumptions that might be violated in significantly different application scenarios.

6 Conclusion

Our proposed task-specific quantization framework, γ -QUANT offers a significant advancement in optimizing the non-linear quantization of ADCs for pattern recognition with different modalities. By moving beyond traditional high-bit-depth linear quantization and manual choice of a non-linear quantization (e.g., based on human perception), we develop an automatic task-aware quantization framework. This helps achieve substantial improvements in object detection and human activity recognition from body-worn sensors with sensor hardware constraints such as energy consumption and memory usage compared to traditional data processing pipelines. Our results highlight the potential of γ -QUANT to maintain the performance of high-bit data in different tasks while minimizing energy consumption, memory usage, and data transmission costs, ultimately contributing to more efficient and sustainable machine learning workflows. While our current work is focused on hardware constraints at the sensor, in the future, we would expand on this framework to combine efficient network architectures and network quantization for inference, offering further improvements in computational efficiency.

Limitations: It is impossible to have a dataset with truly analog signals. Thus, in our work, high-bit depth RAW images serve as a proxy for analog signals. While this is a valid assumption that stems its roots in signal processing theory, there is still some loss of information between true analog and high-bit depth RAW input. Ideally, we would like to test our proposed γ -QUANT on sensors directly to work on analog signals, with the potential benefit of further gains in accuracy.

Acknowledgments

This work was supported by the DFG research project WASDEO (grant number 506589320) and by the DFG Research Unit 5336 – Learning to Sense (L2S). We further gratefully acknowledge the University of Siegen’s OMNI cluster and the high-performance computer HoreKa at the National High-Performance Computing Center at the Karlsruhe Institute of Technology (NHR@KIT) for providing computational resources.

References

1. Bermak, A., Kitchen, A.: A novel adaptive logarithmic digital pixel sensor. *IEEE Photonics Technology Letters* **18**(20), 2147–2149 (2006). <https://doi.org/10.1109/LPT.2006.883893>
2. Bock, M., Kuehne, H., Van Laerhoven, K., Moeller, M.: WEAR: An Outdoor Sports Dataset for Wearable and Egocentric Activity Recognition. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **8**(4) (2024). <https://doi.org/10.1145/3699776>
3. Buckler, M., Jayasuriya, S., Sampson, A.: Reconfiguring the imaging pipeline for computer vision. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
4. Cao, Y., Pan, X., Zhao, X., Wu, H.: An analog gamma correction scheme for high dynamic range cmos logarithmic image sensors. *Sensors* **14**(12), 24132–24145 (2014). <https://doi.org/10.3390/s141224132>, <https://www.mdpi.com/1424-8220/14/12/24132>
5. Chae, Y., Cheon, J., Lim, S., Kwon, M., Yoo, K., Jung, W., Lee, D.H., Ham, S., Han, G.: A 2.1 m pixels, 120 frame/s cmos image sensor with column-parallel $\delta\sigma$ adc architecture. *IEEE Journal of Solid-State Circuits* **46**(1), 236–247 (2010)
6. Chang, J., Sitzmann, V., Dun, X., Heidrich, W., Wetzstein, G.: Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Scientific reports* **8**(1), 1–10 (2018)
7. Chang, J., Wetzstein, G.: Deep optics for monocular depth estimation and 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10193–10202 (2019)
8. Chen, K., Wang, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark (2019)
9. Cheng, W., Erfani, S., Zhang, R., Kotagiri, R.: Learning datum-wise sampling frequency for energy-efficient human activity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018), 1
10. Christie, O., Rego, J., Jayasuriya, S.: Analyzing sensor quantization of raw images for visual slam. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 246–250 (2020). <https://doi.org/10.1109/ICIP40778.2020.9191352>
11. Daghero, F., Burrello, A., Xie, C., Castellano, M., Gandolfi, L., Calimera, A., Macii, E., Poncino, M., Pagliari, D.J.: Human activity recognition on microcontrollers with quantized and adaptive deep neural networks. *ACM Transactions on Embedded Computing Systems (TECS)* **21**(4), 1–28 (2022)
12. Daghero, F., Xie, C., Pagliari, D.J., Burrello, A., Castellano, M., Gandolfi, L., Calimera, A., Macii, E., Poncino, M.: Ultra-compact binary neural networks for human activity recognition on risc-v processors. In: Proceedings of the 18th ACM International Conference on Computing Frontiers. pp. 3–11 (2021)
13. Diamond, S., Sitzmann, V., Julca-Aguilar, F., Boyd, S., Wetzstein, G., Heide, F.: Dirty pixels: Towards end-to-end image processing and perception. *ACM Transactions on Graphics (TOG)* **40**(3), 1–15 (2021)
14. Fu, Y., Zhang, Y., Wang, Y., Lu, Z., Boominathan, V., Veeraraghavan, A., Lin, Y.: Sacod: Sensor algorithm co-design towards efficient cnn-powered intelligent phlatcam. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5168–5177 (2021)
15. Ham, S., Lee, Y., Jung, W., Lim, S., Yoo, K., Chae, Y., Cho, J., Lee, D., Han, G.: Cmos image sensor with analog gamma correction using nonlinear single-slope

- adc. In: 2006 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 4 pp.–3581 (2006). <https://doi.org/10.1109/ISCAS.2006.1693400>
- 16. Hasegawa, T.: Smartphone sensor-based human activity recognition robust to different sampling rates. *IEEE Sensors Journal* **21**(5), 6930–6941 (2021). <https://doi.org/10.1109/JSEN.2020.3038281>
 - 17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
 - 18. Hoelzemann, A., Romero, J.L., Bock, M., Laerhoven, K.V., Lv, Q.: Hang-Time HAR: A Benchmark Dataset for Basketball Activity Recognition Using Wrist-Worn Inertial Sensors. *Sensors* **23**(13) (2023). <https://doi.org/10.3390/s23135879>
 - 19. Kim, K., Lee, H.S.: Probabilistic anchor assignment with iou prediction for object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 355–371. Springer (2020)
 - 20. Klinghoffer, T., Somasundaram, S., Tiwary, K., Raskar, R.: Physics vs. learned priors: Rethinking camera and algorithm design for task-specific imaging. In: 2022 IEEE International Conference on Computational Photography (ICCP). pp. 1–12. IEEE (2022)
 - 21. Klinghoffer, T., Tiwary, K., Behari, N., Agrawalla, B., Raskar, R.: Diser: Designing imaging systems with reinforcement learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23632–23642 (2023)
 - 22. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials* **15**(3), 1192–1209 (2012)
 - 23. Leñero-Bardallo, J.A., Fernández-Berni, J., Rodríguez-Vázquez, Á.: Review of adcs for imaging. In: Image Sensors and Imaging Systems 2014. vol. 9022, pp. 145–150. SPIE (2014)
 - 24. Liberis, E., Lane, N.D.: Differentiable neural network pruning to enable smart applications on microcontrollers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **6**(4), 1–19 (2023)
 - 25. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
 - 26. Liu, A., Mu, S., Xu, S.: A learnable color correction matrix for raw reconstruction. In: British Machine Vision Conference (2024)
 - 27. Ljungbergh, W., Johnander, J., Petersson, C., Felsberg, M.: Raw or cooked? object detection on raw images. In: Scandinavian Conference on Image Analysis. pp. 374–385. Springer (2023)
 - 28. Malekzadeh, M., Clegg, R., Cavallaro, A., Haddadi, H.: Dana: Dimension-adaptive neural architecture for multivariate sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **5**(3), 1–27 (2021)
 - 29. Morawski, I., Chen, Y.A., Lin, Y.S., Dangi, S., He, K., Hsu, W.H.: Genisp: Neural isp for low-light machine cognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 630–639 (2022)
 - 30. Mosleh, A., Sharma, A., Onzon, E., Mannan, F., Robidoux, N., Heide, F.: Hardware-in-the-loop end-to-end optimization of camera image processing pipelines. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7529–7538 (2020)
 - 31. Nakajima, Y., Murao, K., Terada, T., Tsukamoto, M.: A method for energy saving on context-aware system by sampling control and data complement. In: International Symposium on Wearable Computers (ISWC) 2010. pp. 1–4 (2010). <https://doi.org/10.1109/ISWC.2010.5665860>

32. Omid-Zohoor, A., Ta, D., Murmann, B.: Pascalraw: raw image database for object detection (2014)
33. Onzon, E., Mannan, F., Heide, F.: Neural auto-exposure for high-dynamic range object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7710–7720 (June 2021)
34. Ordóñez, F.J., Roggen, D.: Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. MDPI Sensors **16**(1) (2016). <https://doi.org/10.3390/s16010115>
35. Rashid, N., Demirel, B.U., Abdullah Al Faruque, M.: Ahar: Adaptive cnn for energy-efficient human activity recognition in low-power edge devices. IEEE Internet of Things Journal **9**(15), 13041–13051 (2022). <https://doi.org/10.1109/JIOT.2022.3140465>
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
37. Reyes-Ortiz, J.L., Oneto, L., Samà, A., Parra, X., Anguita, D.: Transition-Aware Human Activity Recognition Using Smartphones. Neurocomputing **171** (2016). <https://doi.org/10.1016/j.neucom.2015.07.085>
38. Robidoux, N., Capel, L.E.G., Seo, D.e., Sharma, A., Ariza, F., Heide, F.: End-to-end high dynamic range camera pipeline optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6297–6307 (2021)
39. Scholl, P.M., Wille, M., Van Laerhoven, K.: Wearables in the Wet Lab: A Laboratory System for Capturing and Guiding Experiments. In: International Joint Conference on Pervasive and Ubiquitous Computing. ACM (2015). <https://doi.org/10.1145/2750858.2807547>
40. Sommerhoff, H., Agnihotri, S., Saleh, M., Moeller, M., Keuper, M., Kolb, A.: Differentiable sensor layouts for end-to-end learning of task-specific camera parameters. arXiv preprint arXiv:2304.14736 (2023)
41. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 1139–1147. PMLR, Atlanta, Georgia, USA (17–19 Jun 2013), <https://proceedings.mlr.press/v28/sutskever13.html>, 3
42. Sztyler, T., Stuckenschmidt, H.: On-Body Localization of Wearable Devices: An Investigation of Position-Aware Activity Recognition. In: International Conference on Pervasive Computing and Communications. IEEE (2016). <https://doi.org/10.1109/PERCOM.2016.7456521>
43. Teledyne Vision Solutions: Bit depth, full well, and dynamic range (2024), <https://www.teledynevisionsolutions.com/en-in/learn/learning-center/imaging-fundamentals/bit-depth-full-well-and-dynamic-range/>, accessed: 2025-04-09
44. Wang, Y., Xu, T., Fan, Z., Xue, T., Gu, J.: Adaptiveisp: Learning an adaptive image signal processor for object detection. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 112598–112623. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/cc596a803bedc7a03a87e98c77a22efe-Paper-Conference.pdf
45. Wu, Z., Wang, X., Jia, M., Liu, M., Sun, C., Wu, C., Wang, J.: Dense object detection methods in raw uav imagery based on yolov8. Scientific reports **14**(1), 18019 (2024)

46. Xu, R., Chen, C., Peng, J., Li, C., Huang, Y., Song, F., Yan, Y., Xiong, Z.: Toward raw object detection: A new benchmark and a new model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13384–13393 (2023)
47. Yan, Z., Subbaraju, V., Chakraborty, D., Misra, A., Aberer, K.: Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In: 2012 16th International Symposium on Wearable Computers. pp. 17–24 (2012). <https://doi.org/10.1109/ISWC.2012.23>
48. Zheng, L., Wu, D., Ruan, X., Weng, S., Peng, A., Tang, B., Lu, H., Shi, H., Zheng, H.: A novel energy-efficient approach for human activity recognition. Sensors **17**(9), 2064 (2017)
49. Zhou, H., Zhang, X., Feng, Y., Zhang, T., Xiong, L.: Efficient human activity recognition on edge devices using deepconv lstm architectures. Scientific Reports **15**(1), 13830 (2025)

Table Of Content

The supplementary material covers the following information:

- Section A: Visualizations
- Section B: Additional Results: RAOD Dataset.
- Section C: Additional Results: Gamma Initialization (HAR).

A Visualizations

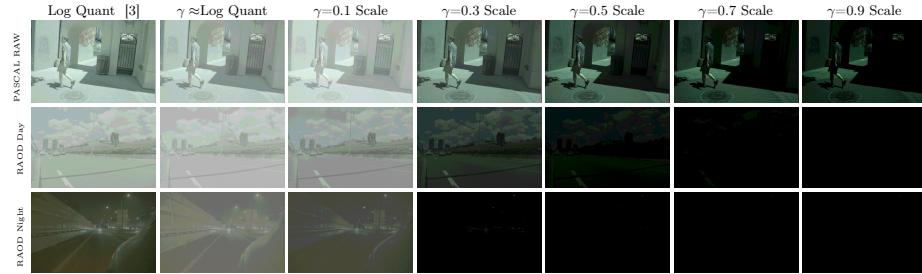


Fig. 5. We visualize the effect of different values of gamma on a **4-bit** quantized image. Top row shows images from PASCAL RAW and bottom two rows with RAOD day and night scenarios. The first column applies logarithmic quantization with $\epsilon=1$, whereas the second column approximates log values using appropriate gamma values. The γ value for PASCAL RAW is chosen to be 0.136 and 6.395e-2 for RAOD. The next five columns show gamma scaling with different chosen gammas.

B Additional Results: RAOD Dataset

We provide results with RAOD dataset for linear quantization and γ -QUANT. RAOD is a 24 bit HDR RAW data introduced by [46]. γ -QUANT performs best in all cases. Since, RAOD dataset includes day and night scenarios, we also experiment with conditioning gamma on the time of the day. We see that there is marginal difference in the results of the two experiments.

C Additional Results: Gamma Initialization (HAR)

In Table 3 we provide results of comparing a initialization of $\gamma = 1$, i.e. linear quantization, with the initialization provided in the paper of $\gamma = 0.4$. One can

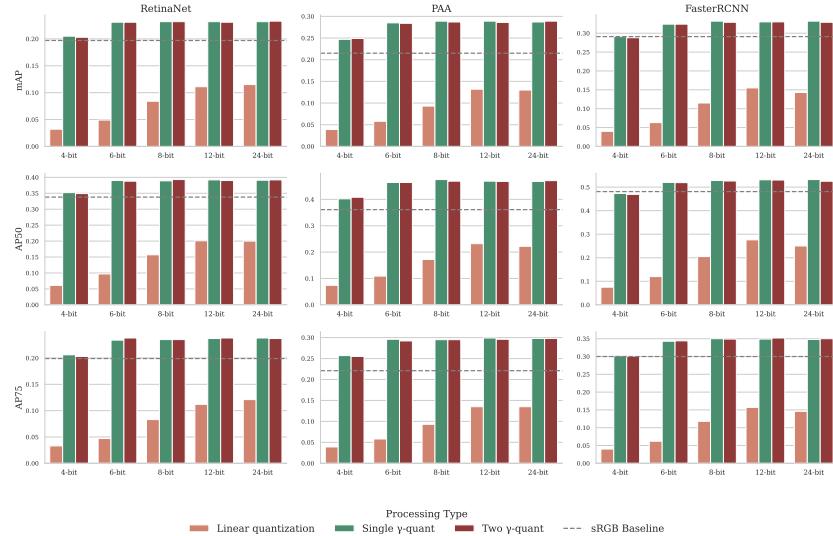


Fig. 6. Detection results on the RAOD dataset using γ -QUANT and linear quantization with PAA, Faster R-CNN, and RetinaNet across different bit depths. Blue dashed lines represent performance using standard sRGB input images. γ -QUANT consistently outperforms standard baselines across all architectures and quantization levels. mAP reflects average performance across 10 IoU thresholds; AP50 and AP75 represent results at 0.5 and 0.75 IoU thresholds, respectively.

see that results only marginally differ across the two different initializations of γ , with our chosen initialization of $\gamma = 0.4$ slightly performing better, which further supports our assumption made in the main paper that differences in smaller accelerations contain more information than differences in large accelerations.

Table 3. Per-dataset HAR results comparing training the chosen initializations of $\gamma = 0.4$ with a initialization $\gamma = 1.0$, i.e. linear quantization. We provide results for bit depths $\hat{N} = 2$ and $\hat{N} = 4$. One can see that results only marginally differ with $\gamma = 0.4$ performing slightly better, supporting our assumptions made in the main paper.

		WEAR	Wetlab	Hang-Time	RWHAR	SBHAR
$\hat{N} = 4$	Global ($\gamma = 1.0$)	70.17 ± 0.13	23.47 ± 0.29	33.89 ± 0.04	70.28 ± 0.10	52.92 ± 0.28
	Global ($\gamma = 0.4$)	70.14 ± 0.51	23.00 ± 0.24	33.63 ± 0.17	69.28 ± 0.46	52.58 ± 0.20
	Per-axis ($\gamma = 1.0$)	68.93 ± 0.41	23.45 ± 0.28	33.82 ± 0.09	70.09 ± 0.91	53.58 ± 0.15
	Per-axis ($\gamma = 0.4$)	70.17 ± 0.39	23.63 ± 0.20	33.67 ± 0.17	71.16 ± 1.26	52.52 ± 0.21
$\hat{N} = 2$	Global ($\gamma = 1.0$)	60.55 ± 0.65	12.98 ± 0.12	29.63 ± 0.12	70.67 ± 0.20	35.75 ± 0.55
	Global ($\gamma = 0.4$)	60.45 ± 0.43	12.94 ± 0.18	30.67 ± 0.12	71.41 ± 0.51	40.18 ± 0.32
	Per-axis ($\gamma = 1.0$)	62.76 ± 0.24	14.35 ± 0.14	28.60 ± 0.26	67.30 ± 0.36	36.56 ± 0.33
	Per-axis ($\gamma = 0.4$)	63.95 ± 0.14	16.85 ± 0.54	30.60 ± 0.22	69.21 ± 0.32	41.88 ± 0.32