Evaluating Privacy-Oriented Image Encoding Methods for Embedded Vision

Tyler Lash
Worcester Polytechnic Institute
100 Institute Rd, Worcester, MA 01609
tlash@wpi.edu

Abstract

Multiple methods of image encoding are investigated to test the constraints of a simple CNN pipeline where privacy and efficiency are major concerns. Using identical models and training procedures, raw images are compared against downsampling, quantization, Gaussian blur, additive noise, and Sobel edge extraction across three datasets of increasing complexity: MNIST, CIFAR-10, and a small, heavily constrained hand-gesture dataset captured on an ESP32-S3 Sense device. Results show that while digit recognition is minimally impacted by each method of information reduction, natural image classification is weak where fidelity is significantly reduced. For embedded hand-gesture recognition, it is found that shape-based encodings consistently outperform raw images despite discarding large amounts of texture and color information. Findings confirm that encoding choice alone significantly impacts learning and generalization, while also suggesting that lower-fidelity representations may actually be preferable for structured, task-specific embedded vision problems depending on the dataset and preprocessing involved.

1. Introduction

Embedded vision systems operate under tight constraints on compute, memory, power, and data bandwidth, necessitating the use of compact, efficient pipelines. Prior work has explored quantized and gradient-based input representations to improve the energy efficiency and robustness of tinyML vision models on resource-constrained platforms [5] while recent surveys of embedded AI vision systems emphasize the importance of early processing in real-time deployment [2]. This work evaluates common image encoding strategies (including downsampling, Gaussian blur, quantization, additive noise, and Sobel edge extraction) and the effect they have on classification performance across datasets of increasing complexity (MNIST, CIFAR-10, and a custom hand-gesture dataset captured on an ESP32-S3 Sense). The impact of the encoding itself on learning dynamics and final accuracy is seen by training identical models on raw and encoded images. The goal of this paper is ultimately to determine an encoding method or pipeline that can be deployed effectively in real time on low-cost embedded hardware without sacrificing model prediction accuracy, while also offering potential benefits in efficiency and privacy.

2. Related Works

Recent work has shown that modifying input representations – rather than only optimizing model architectures – can significantly impact efficiency and robustness in such settings [5].

Many of the encoding strategies considered in this work are rooted in classical image processing and early vision. Operations such as spatial downsampling, intensity quantization, and Gaussian smoothing are standard tools for reducing data fidelity while preserving coarse structural information [3]. Use of Gaussian blur comes from scale-space theory, a computer vision framework which formalizes how smoothing suppresses fine-scale detail while retaining larger-scale structures relevant to recognition tasks [4]. Edge and gradient-based representations have also been widely used as compact alternatives to raw pixels, most notably in histogram-of-oriented-gradient (or HOG) descriptors, demonstrating that boundary and shape information alone can be highly discriminative for certain tasks [1].

Similarly, privacy-preserving vision research has explored reducing visual fidelity at the input level. Prior works demonstrate that feature recognition remains feasible even with extreme spatial resolution constraints, ergo partial privacy preservation via aggressive downsampling [8]. More recent studies explicitly model the trade-offs between privacy and recognition accuracy as a function of image resolution, which highlights the need for an empirical evaluation of low-fidelity representations [9].

It is important to note results which reveal how common obfuscation techniques such as blur or pixelation can be reversed by learning-based models caution against assuming

privacy from degradation alone [6]. Related approaches such as P3 explore privacy-preserving image sharing via encoding transformations at the compression level, though these works have different system goals and assumptions than considered here [7].

Unlike prior works which typically focus on a single representation or task, this paper provides a controlled, cross-dataset evaluation of common low-fidelity encodings under identical training pipelines, including a real embedded deployment. By including both standard benchmarks and a real embedded hand-gesture dataset captured on an ESP32-S3 Sense, the effects of input encoding on learning dynamics and classification performance under realistic embedded constraints are isolated.

3. Methods

For each dataset, images are first transformed with a fixed encoding method, then used to train an identical classification model under consistent training and evaluation settings. By holding the model architecture, optimization parameters, and data splits constant across experiments, performance differences are attributed to the encoding strategy rather than variability in model or training.

3.1. Experimental Pipeline

Encodings are applied offline prior to training and evaluation. Each image is processed using exactly one encoding method, which is applied to both training and test splits. In cases where an encoding reduces spatial resolution (e.g., downsampling), images are resized back to the original input dimensions (using bilinear interpolation) in order to maintain a fixed input size.

A single CNN architecture with no encoding-specific layers or modifications is used across all experiments. All models are trained from scratch with the same optimizer, learning rate, batch size, loss function, and number of epochs (per dataset). No data augmentation or encoding-specific hyperparameter tuning is applied in each case.

Datasets are split into training and test sets using a fixed protocol. Model performance is evaluated using classification accuracy and cross-entropy loss. For the hand-gesture dataset, experiments are repeated across multiple random seeds with results reported as mean and standard deviation from those sets.

3.2. Datasets

To evaluate the effect of image encodings across varying levels of visual complexity, experiments were conducted on two standard benchmark datasets and one custom dataset collected on embedded hardware.

3.2.1 MNIST

MNIST consists of grayscale images of handwritten digits with low spatial resolution and limited visual complexity. It is used to represent a best-case scenario for aggressive image encodings, thereby allowing evaluation of how much information can be discarded ideally while still maintaining high classification accuracy.

3.2.2 CIFAR-10

CIFAR-10 is a color image dataset containing natural images across ten object classes with significantly higher visual variability than MNIST. This dataset is used to assess how these encoding strategies may scale to more complex (eg, texture or color-dependent) classification tasks.

3.2.3 Gestures

A custom hand-gesture dataset was collected using a Seeed Studio XIAO ESP32-S3 Sense equipped with 8MB onboard flash, 8MB PSRAM, and an OV3660 camera module. Raw frames (320×240 JPEG output) were buffered in PSRAM and written directly to an onboard 1GB microSD card with no further processing. The dataset consists of multiple hand poses captured under relatively consistent imaging conditions: six hand gestures with varying lighting against a blank background, yielding a set of 300 images evenly split across each of six classes. The intentionally limited dataset should reasonably reflect the constraints of real embedded vision systems, including factors like limited resolution, sensor noise, and small dataset size. This dataset is intended to provide a practical evaluation of encoding strategies in a realistic, on-device classification setting.

3.3. Encodings

Several encoding methods were applied independently to each dataset prior to training. These encodings were selected to reflect common trade-offs in embedded vision systems regarding spatial resolution, noise, precision, and feature extraction.

3.3.1 Downsampling

This encoding evaluates the extent to which coarse spatial structure alone is sufficient for classification. Spatial downsampling reduces image resolution to lower memory usage and computational cost. Images were downsampled to a fixed low resolution and then resized back to the original input dimensions using bilinear interpolation prior to model input. For MNIST and CIFAR-10, images were downsampled to 8×8 pixels, which represents an aggressive reduction in spatial detail. For the hand-gesture dataset, a slightly higher resolution (24×24) was used to preserve sufficient

hand shape information while still substantially reducing spatial complexity.

3.3.2 Quantization

This transformation reflects constraints imposed by storage, bandwidth, and low-precision arithmetic on embedded hardware, and isolates the effect of reduced intensity resolution on learning dynamics. Quantization reduces the numerical precision of image intensities by limiting bit depth. For MNIST and CIFAR-10, pixel intensities were uniformly quantized to 8 discrete levels, corresponding to 3-bit precision. For the hand-gesture dataset, a 16-level quantization was used to better accommodate sensor noise and illumination variation while still reducing representational precision.

3.3.3 Gaussian Blur

This encoding examines the importance of fine-grained detail compared to coarse spatial structure with increasing visual complexity. Gaussian blurring suppresses high-frequency image content, but preserves low-frequency structure. Images were blurred using an isotropic Gaussian kernel with standard deviation $\sigma = 1.5$ for MNIST and $\sigma = 1.0$ for both CIFAR-10 and the gesture dataset. These values were selected because they noticeably reduce fine texture and edge sharpness without completely destroying global object shape.

3.3.4 Gaussian Noise

This encoding simulates sensor noise and quantization artifacts commonly encountered in embedded imaging systems. Additive Gaussian noise is applied by adding zero-mean noise with fixed variance to each pixel intensity. Noise was added with standard deviation $\sigma = 0.15$ for MNIST, $\sigma = 0.08$ for CIFAR-10, and $\sigma = 0.05$ for the hand-gesture dataset, reflecting progressively tighter noise tolerances as dataset complexity increases.

3.3.5 Sobel Edges

Sobel edge encoding replaces raw pixel intensities with gradient magnitude information derived from first-order spatial derivatives in the horizontal and vertical directions. The resulting edge magnitude images are normalized and used as single-channel inputs to the classifier. This representation emphasizes object boundaries and structural features while discarding texture and color information; this makes it well suited for tasks where class identity is largely shape-driven. Sobel encoding is also computationally lightweight, meaning it can be efficiently implemented on embedded hardware as part of a near-sensor preprocessing pipeline.

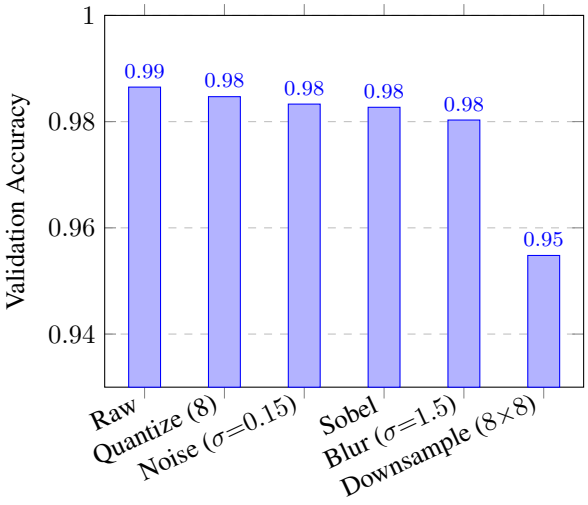

Figure 1. MNIST validation accuracy

Table 1. MNIST validation performance

Encoding	Val. Acc.	Val. Loss
Raw	0.9865	0.0460
Quantize (8 levels)	0.9847	0.0550
Noise ($\sigma = 0.15$)	0.9833	0.0518
Sobel edges	0.9827	0.0539
Gaussian blur ($\sigma = 1.5$)	0.9803	0.0636
Downsample (8×8)	0.9548	0.1439

4. Results and Discussion

MNIST and CIFAR-10 use fixed splits and report validation accuracy, whereas test accuracy for the gesture dataset is averaged over multiple random seeds. Relative trends across encodings are seen in figures and tables with results organized by dataset to emphasize effectiveness of each encoder.

4.1. MNIST

Figure 1 shows overall performance remains high for encodings that preserve digit shape with only minor degradation relative to raw images. In contrast, aggressive spatial information removal via downsampling results in a much larger reduction in accuracy.

Table 1 summarizes the corresponding numerical validation performance after training using the MNIST dataset.

4.2. CIFAR-10

CIFAR-10 presents a substantially more challenging classification task than MNIST because it consists of color images with complex backgrounds, texture, and significant variation within classes. Unlike MNIST, successful classification on CIFAR-10 relies not only on object shape, but

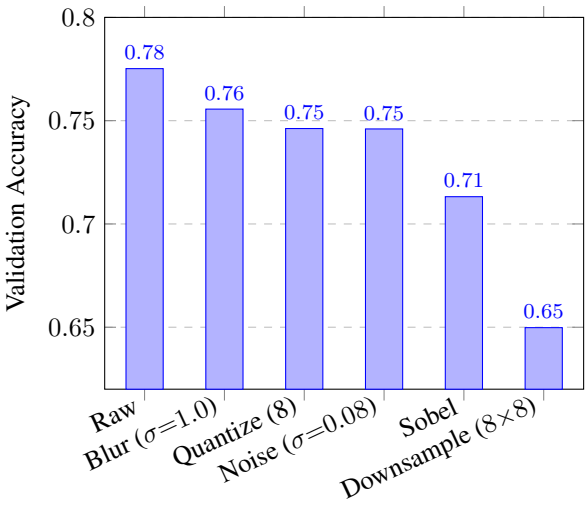

Figure 2. CIFAR-10 validation accuracy

Table 2. CIFAR-10 validation performance

Encoding	Val. Acc.	Val. Loss
Raw	0.7752	0.8384
Gaussian blur ($\sigma = 1.0$)	0.7556	0.8557
Quantize (8 levels)	0.7462	1.0120
Noise ($\sigma = 0.08$)	0.7460	0.7802
Sobel edges	0.7132	1.1098
Downsample (8×8)	0.6498	1.0899

also on fine-grained texture and color cues. This makes it well-suited for examining the limits of privacy-preserving encoding.

Figure 2 reveals that all encodings used with CIFAR-10 lead to a noticeable degradation in performance as opposed to MNIST. While the model trained on raw images achieves a validation accuracy of approximately 77.5%, even mild transformations such as Gaussian blur and color quantization reduce accuracy by 2–3%. Additive noise produces comparable degradation, whereas Sobel edge encoding and spatial downsampling result in substantially larger performance drops.

Table 2 summarizes the final validation accuracy and loss values for each encoding.

Together, these results demonstrate how privacy-preserving encodings will have a substantially larger impact on natural image classification compared to simple digit recognition. While coarse object structure is still somewhat informative, the suppressing fine-grained detail leads to reduced utility with a dataset such as CIFAR-10.

4.3. Training vs. Validation Loss

Where MNIST exhibits rapid convergence and minimal generalization gap, CIFAR-10 presents a significantly more challenging setting in which reduced representations amplify overfitting effects.

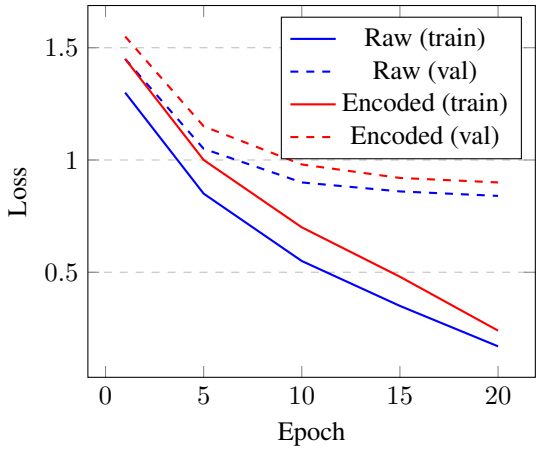

Figure 3. Training and validation loss versus epoch on CIFAR-10

Figure 3 shows training loss decreases steadily throughout training for raw inputs, while validation loss decreases initially, then plateauing early and remaining substantially higher. This divergence indicates that the model is continuining to fit training data without achieving meaningful improvements on unseen samples. The same pattern is seen with privacy-encoded inputs, although both training and validation losses converge to higher values overall.

The persistence of a large gap between training and validation loss across both raw and encoded representations suggests that performance degradation on CIFAR-10 is driven primarily by limited generalization rather than optimization failure. While privacy-preserving encodings reduce the absolute amount of information available to the model, they do not prevent memorization of remaining features, particularly in the absence of data augmentation or strong inductive biases.

4.4. Sobel Curve

While Sobel edges preserve object boundaries and coarse structure, they discard texture and color information, resulting in a fundamentally different input distribution than raw images.

As shown in Figure 4, Sobel-encoded inputs exhibit increased instability during training compared to other encodings. Validation loss in particular has sharp spikes during mid-training despite steady decreases in training loss. This suggests that while the model continues to fit the training data, it is struggling to generalize from gradient-only representations, hence unstable validation performanec.

Despite this instability, Sobel encoding manages to remain competitive on simpler, more structured tasks (e.g., MNIST and hand gestures), in which cases object boundaries are strongly correlated with class identity. However,

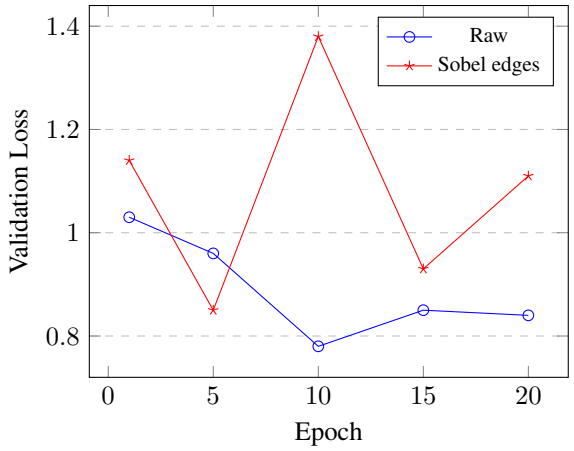

Figure 4. Validation loss versus epoch for raw and Sobel-encoded CIFAR-10 inputs

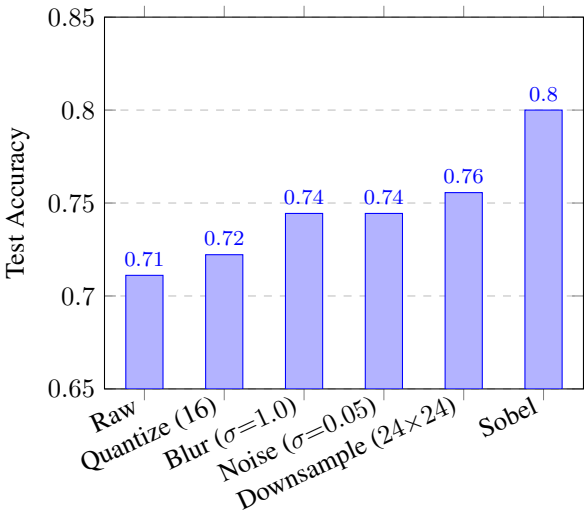

Figure 5. Hand-gesture classification accuracy

the absence of texture and color cues using a dataset with as much complexity as CIFAR-10 unfortunately limits discriminative capacity. Sobel-based representations are task-dependent: they can be extremely effective when shape dominates, but poorly suited for natural image classification where fine-grained visual detail is critical.

4.5. Hand Gestures

Classification is also demonstrated on a custom hand-gesture dataset captured using an ESP32-S3 Sense (equipped with an OV3660 camera). Compared to MNIST and CIFAR-10, this dataset is small and task-specific. Class identity is primarily determined by hand shape and finger configuration rather than texture or color.

Figure 5 summarizes the mean test accuracy across encoding methods averaged over multiple random seeds with corresponding results shown in Table 3. Sobel edge encoding achieves the highest classification accuracy, confirming that shape-based representations preserve the most discriminative information with gesture recognition. Downsampling even performs quite well, suggesting that coarse spatial structure is more than sufficient for this task.

Table 3. Hand-gesture test performance (mean ± std over random seeds)

Encoding	**Test Acc.**	**Test Loss**
Sobel edges	0.800 ± 0.067	1.064 ± 0.589
Downsample (24×24)	0.756 ± 0.084	1.250 ± 0.669
Noise (σ=0.05)	0.744 ± 0.084	0.923 ± 0.228
Gaussian blur (σ=1.0)	0.744 ± 0.102	1.164 ± 0.615
Quantize (16 levels)	0.722 ± 0.084	1.339 ± 0.626
Raw	0.711 ± 0.084	0.852 ± 0.523

Encodings that retain more appearance-based information (including raw images, quantization, Gaussian blur, and additive noise) achieve lower accuracy on average. Raw images perform worst, likely due to high background variability and limited training data, both of which hinder generalization in this setting. While this may seem counterintuitive, these results align with prior works showing how suppression of irrelevant features will naturally impose a strong inductive bias toward geometric structure. This is an interesting contrast with regard to natural image classification: while edge-based encodings degrade performance on CIFAR-10, they can be very well-suited for structured, shape-driven tasks with constrained regimes.

5. Conclusion

While MNIST classification can tolerate substantial information removal, natural image classification on CIFAR-10 is highly sensitive to transformations that suppress texture and color information. In contrast, results on the real-world embedded hand-gesture dataset demonstrate that shape-based encodings can outperform raw images when task structure is dominated by object boundaries. Encoding choice affects not only final accuracy, but also optimization stability and generalization behavior. Results emphasize that no single encoding is universally optimal, instead demonstrating the importance of matching image representations to specific task characteristics while keeping deployment-related constraints in mind. For embedded vision systems operating under strict resource and privacy requirements, carefully selected encodings can enable efficient, real-time inference without sacrificing (or, in some cases, even improving) classification performance.

References

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005. 2

[2] Alexandre Fabre et al. From near-sensor to in-sensor: A state-of-the-art review of embedded ai vision systems. *Sensors*, 24(16):5446, 2024. 1

[3] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson, 4 edition, 2018. 2

[4] Tony Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):225–270, 1994. 2

[5] Qianyun Lu and Boris Murmann. Enhancing the energy efficiency and robustness of tinyml computer vision using coarsely-quantized log-gradient input images. *ACM Transactions on Embedded Computing Systems*, 22(5):1–23, 2023. 1, 2

[6] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *arXiv*, 2016. 2

[7] Moo-Ryong Ra, Ramesh Govindan, and Antonio Ortega. P3: Toward privacy-preserving photo sharing. In *Proceedings of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2013. 2

[8] Michael S. Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 4255–4262, 2017. 2

[9] Yuntao Wang, Zirui Cheng, Xin Yi, Yan Kong, Xueyang Wang, Xuhai Xu, Yukang Yan, Chun Yu, Shwetak Patel, and Yuanchun Shi. Modeling the trade-off of privacy preservation and activity recognition on low-resolution images. *arXiv*, 2023. 2