JPEG INSPIRED DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

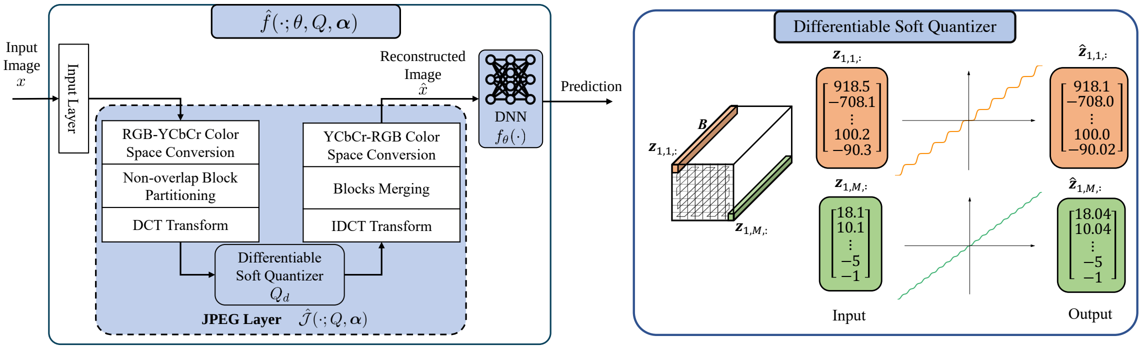
Although it is traditionally believed that lossy image compression, such as JPEG compression, has a negative impact on the performance of deep neural networks (DNNs), it is shown by recent works that well-crafted JPEG compression can actually improve the performance of deep learning (DL). Inspired by this, we propose JPEG-DL, a novel DL framework that prepends any underlying DNN architecture with a trainable JPEG compression layer. To make the quantization operation in JPEG compression trainable, a new differentiable soft quantizer is employed at the JPEG layer, and then the quantization operation and underlying DNN are jointly trained. Extensive experiments show that in comparison with the standard DL, JPEG-DL delivers significant accuracy improvements across various datasets and model architectures while enhancing robustness against adversarial attacks. Particularly, on some fine-grained image classification datasets, JPEG-DL can increase prediction accuracy by as much as 20.9%. Our code is available on `https://github.com/JpegInspiredDl/JPEG-Inspired-DL.git`.

1 INTRODUCTION

JPEG compression (Pennebaker & Mitchell, 1992) is the *defacto* lossy image compression technique with ubiquitous presence in real-world applications. With the development of deep learning, more and more images, potentially compressed by JPEG, are consumed by deep neural networks (DNNs). Naturally, it's of interest to study how JPEG compression will impact DNN performance for computer vision tasks, and extensive research has been conducted along this line such as Dodge & Karam (2016); Liu et al. (2018) and Xie & Kim (2019). These initial explorations establish a widely accepted view that the information loss caused by JPEG obscures important features in the input image, thereby negatively impacting DNN performance.

However, it was shown by Yang et al. (2021) that the above conventional wisdom does not hold anymore if JPEG compression is applied intelligently and adaptively on a per image basis. Indeed, Yang et al. (2021) showed that if applied appropriately, JPEG compression can actually improve DNN performance at least in theory. They further proposed to train a DNN based on images with various JPEG quality levels. While they managed to demonstrate improved performance with a specially designed DNN topology, the model is too cumbersome to be fully trained, leading to suboptimal performance. Moreover, the adherence to the default JPEG quantization also limits the effectiveness of this method. On the other hand, given a fixed DNN model, Zheng et al. (2023) and Salamah et al. (2024b) found that its performance could be slightly improved if the input images got compressed by JPEG with optimized quantization parameters. However, the performance gain is not significant due to the frozen DNN model. Although these existing works provide promising insights to improve DNN performance with JPEG compression, a solid solution that can fully unleash the potential of this idea has yet to be found.

To address the above issue, in this paper, we propose jointly optimizing both JPEG quantization operation and a DNN to achieve greater effectiveness. To this end, we first introduce a trainable JPEG compression


$\hat{f}(\cdot;\theta,Q,\boldsymbol{\alpha})$
Input Image x
Input Layer
Reconstructed Image $\hat{x}$
DNN $f_\theta(\cdot)$
Prediction
RGB-YCbCr Color Space Conversion
Non-overlap Block Partitioning
DCT Transform
YCbCr-RGB Color Space Conversion
Blocks Merging
IDCT Transform
Differentiable Soft Quantizer Q_d
JPEG Layer $\hat{\mathcal{J}}(\cdot;Q,\boldsymbol{\alpha})$
Differentiable Soft Quantizer
B
$\boldsymbol{z}_{1,1,:}$
$\boldsymbol{z}_{1,M,:}$
$\hat{\boldsymbol{z}}_{1,1,:}$
$\hat{\boldsymbol{z}}_{1,M,:}$
918.5 −708.1 ⋮ 100.2 −90.3
918.1 −708.0 ⋮ 100.0 −90.02
18.1 10.1 ⋮ −5 −1
18.04 10.04 ⋮ −5 −1
Input
Output

(a) (b)

Figure 1: (a) The JPEG-DL framework is illustrated, with both standard DNN's operations and the forward/inverse processes of the standard JPEG pipeline are shown in white boxes as non-trainable components. The JPEG pipeline receives a conventionally preprocessed input image $x \in \mathbb{R}^{3\times W\times H}$ sampled from the underlying task's dataset. The differentiable soft quantizer (Q_d) and the underlying DNN architecture (f_θ), shown in blue, formed a unified new architecture $\hat{f}(x;\theta,Q,\boldsymbol{\alpha})$ parametrized with these being the only trainable components. (b) As an example, we show $\boldsymbol{z}_{1,:,:}$, i.e., the DCT representation of the Y channel of an image, by a tensor consisting of B blocks of DCT coefficients. Each 8×8 block contains $M = 64$ DCT frequencies, ordered from low to high in a zigzag manner. Then, we show how $\boldsymbol{z}_{1,1,:}$ and $\boldsymbol{z}_{1,M,:}$ are quantized by $Q_d(\cdot\,; q=1, \alpha=10)$ and $Q_d(\cdot\,; q=0.5, \alpha=16)$, respectively.

layer, structure of which is illustrated in Fig. 1a. This layer follows the standard JPEG pipeline to convert a preprocessed input image, after passing through after the input layer, into blockwise Discrete Cosine Transform (DCT) coefficients in the YCbCr color space. A novel differentiable soft quantizer (Q_d) is then applied to quantize the DCT coefficients at each frequency position, followed by the standard JPEG inverse process to reconstruct the RGB image. We then present JPEG-DL, a novel deep learning (DL) framework that that introduces a new DNN architecture by inserting a JPEG layer directly after the original input layer of any underlying DNN architecture. This layer can be considered an integral part of any underlying DNN architecture, forming a new unified DNN architecture, whose parameters get optimized jointly with DNN model weights during training. The core of JPEG-DL is Q_d as it substitutes the non-differentiable, hard quantization in JPEG with a differentiable, soft quantization operation defined by a neat analytical formula, which not only facilitates gradient-based optimization for quantization parameters but also introduces additional trainable non-linearity to the overall image understanding pipeline. To validate the effectiveness of JPEG-DL, we conducted extensive experiments for image classification on six datasets including four fine-grained classification datasets (Wah et al., 2011; Khosla et al., 2011; Nilsback & Zisserman, 2008; Parkhi et al., 2012), CIFAR-100 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009). Results show that JPEG-DL significantly and consistently outperforms the standard DL across various DNN architectures, with a negligible increase in model complexity. Specifically, JPEG-DL improves classification accuracy by up to 20.9% on some fine-grained classification dataset, while adding only 128 trainable parameters to the DL pipeline. Moreover, the superiority of JPEG-DL over the standard DL is further demonstrated by the enhanced adversarial robustness of the learned unified architecture.

The main contributions of this paper can be summarized as follows:

- We introduce a novel trainable JPEG layer leveraging differentiable soft quantizers with nice analytical formulas.

- Based on the new JPEG layer, we propose a new DL framework dubbed JPEG-DL, which jointly optimizes the JPEG layer and the DNN model during training.
- The outstanding performance of JPEG-DL over the standard DL is verified by comprehensive experimental results on various image classification datasets across multiple DNN architectures and for a variety of tasks including adversarial defense.

2 BACKGROUND AND RELATED WORK

JPEG applications. JPEG was originally developed as a lossy image compression technique based on transform coding, which reduces image file sizes by generating their compact representations. Beyond its traditional use, JPEG has found numerous applications in deep learning: (1) it has been utilized as a data augmentation technique to improve robustness of DNNs against image compression (Benbarrad et al., 2022); (2) it has been employed as an empirical defense method against adversarial attacks, effectively reducing adversarial perturbations and enhancing the adversarial robustness of DNNs (Dziugaite et al., 2016; Das et al., 2017; Guo et al., 2018); and (3) it has been integrated into the knowledge distillation (Hinton et al., 2015) framework by Salamah et al. (2024a), where it helps the teacher model transfer knowledge to the student model in a more effective way. In contrast, this paper focuses on leveraging JPEG to enhance the natural performance, instead of the robust performance, of DNNs without relying on any teacher model.

Optimizing JPEG Compression for DNN vision. As a lossy image compression technique, JPEG is developed specifically for the human visual system. As a result, while the information loss introduced by JPEG is often imperceptible to humans, it can significantly degrade the performance of DNNs. This issue gives rise to a line of research which optimizes JPEG compression based on DNN perception. For instance, given a pretrained DNN, Xie & Kim (2019), Zheng et al. (2023) and Salamah et al. (2024b) first derive its sensitivity to different DCT frequencies, based on which they customize JPEG quantization tables for this DNN to reach the optimal rate-accuracy tradeoff. Another popular direction is to make JPEG trainable, which integrates the JPEG encoder and DNN into an end-to-end differentiable training framework (Luo et al., 2020; Xie et al., 2022). These methods create differentiable proxies for image quantization and bitrate calculation, so that JPEG compression can be optimized via backpropagation in order to minimize a total loss considering both DNN performance and bitrate. However, all the above works focus on mitigating DNN performance degradation in the presence of JPEG compressed images, but offer little to no improvement on DNN performance when raw images are given as input. On the contrary, this paper leverages JPEG purely as a tool to improve DNN performance on raw images, regardless of its compression capability.

Improving DNN performance with JPEG. Due to the aforementioned reason, JPEG compression generally hurts DNN performance. However, it's recently demonstrated by Yang et al. (2021) that, with an oracle guiding the compression process, one can select an optimal quality level to compress each image, enabling the DNN to make its best possible prediction for it. By applying this adaptive JPEG compression across a set of images, one can actually improve the DNN prediction accuracy considerably. This phenomenon, termed "compression helps" in the original paper, is justified by the fact that compression can remove noise and disturbing background features, thereby highlighting the main object in an image, which helps DNNs make better prediction. Due to the need of ground truth labels in the compression stage, the above adaptive compression scheme is not realizable in most real world applications; however, the discovery of "compression helps" at least show the potential of using JPEG to improve DNN performance. Thus motivated, the authors in turn proposed an implementable way to improve DNN performance with JPEG. They built a new DNN topology that incorporates 11 parallel branches of an underlying pretrained model, with each branch obtaining as input either the raw image or its compressed version at a varying quality level. The penultimate layer representations from all branches are concatenated and fed into a classification head, which is then trained together with those 11 model backbones as a unified DNN structure. This new DNN topology is clearly too complex to train, so the authors opted for partial training to mitigate the training complexity, therefore resulting in a suboptimal performance. In contrast, our proposed method introduces only 128 additional trainable parameters to the existing DL pipeline, causing negligible complexity increase. Moreover, it's noteworthy that our method is orthogonal to theirs, as our trainable JPEG quantization tables can be embedded into their framework to replace the default JPEG quantization tables, thereby further improving the performance.

Trainable Activation Functions. Relentless efforts have been made to search for activation functions that can improve DNN performance. Among all the directions, trainable activation functions have gained particular interest for their flexibility, expressive power, and adaptability during training. Chen & Chang (1996) propose the adjustable generalized hyperbolic tangent function, which extends the classic hyperbolic tangent function by introducing parameters to control the saturation level and slope of the function. He et al. (2015) introduce the parametric ReLU (PReLU), a variant of ReLU with a trainable parameter that adjusts the negative part of ReLU. More recently, Kolmogorov-Arnold Networks (KANs) (Liu et al., 2024) employ trainable activation functions on edges, with nodes simply summing all the incoming activations. Interestingly, the differentiable soft quantizers in our JPEG layer can be interpreted as trainable activation functions whose input is the DCT representations of images. These trainable soft quantizers effectively introduce additional nonlinearity to DNN models, thus improving their expressive power.

3 JPEG-DL: JPEG INSPIRED DL

3.1 PROBLEM FORMULATION

In the JPEG pipeline, an RGB image $x \in \mathbb{R}^{3\times W\times H}$ is first converted to the YCbCr color space and then partitioned into B non-overlapping 8×8 blocks, where DCT is applied to each of these blocks to obtain the corresponding DCT coefficients. For each color channel, the DCT coefficients in each block are then flattened following the zigzag order, resulting in $M = 64$ frequency positions ordered from low frequency to high frequency. Therefore, we denote the DCT coefficients of the image x as $\boldsymbol{z} = [z_{l,m,n}]$, where $l = 1, 2, 3$ corresponds to the color channel Y, Cb and Cr respectively, $1 \leq m \leq M$ is the index for frequency position, and $1 \leq n \leq B$ is the index for block. Up to this point, all operations involved are differentiable. Next, quantization tables $Q_Y = [q_1, q_2, \dots, q_M]$ and $Q_C = [q_{M+1}, q_{M+2}, \dots, q_{2M}]$ are used for the luminance (Y) and chrominance (CbCr) channels respectively to quantize their DCT coefficients. Following uniform quantization, we obtain quantized DCT coefficients $\hat{z}_{l,m,n} = \lfloor z_{l,m,n}/q_m \rceil \cdot q_m$ for $l = 1$, and $\hat{z}_{l,m,n} = \lfloor z_{l,m,n}/q_{M+m} \rceil \cdot q_{M+m}$ for $l = 2, 3$. Note that quantization is non-differentiable due to the use of the rounding operation. Finally, inverse operations including blockwise inverse DCT (IDCT) transform, blocks merging and YCbCr-to-RGB color space conversion are conducted over $\hat{\boldsymbol{z}}$ sequentially to obtain the reconstructed RGB image $\hat{x}$. Similar to their forward counterparts, all these inverse operations are differentiable. Denoting the composition of all the above operations as $\mathcal{J}$, we then have $\hat{x} = \mathcal{J}(x; Q)$, where $Q = (Q_Y, Q_C)$. Hereafter, the mapping $\mathcal{J}$ stands for the JPEG encoding-decoding operation.

In supervised learning, each $x \in \mathcal{X}$ corresponds to a ground truth label $y \in \mathcal{Y}$. Let f_θ represent a DNN model with trainable weights θ, and let $\mathcal{L}$ denote the loss function used to train this DNN. In standard DL, the primary objective is to solve the following minimization problem:

$$\min_{\theta} \mathbb{E}[\mathcal{L}(f_\theta(x), y)]. \tag{1}$$

In contrast, JPEG-DL tries to improve the performance of DNN by jointly training it with the JPEG operation. As a result, the formulation should be instead:

$$\min_{\theta, Q} \mathbb{E}[\mathcal{L}(f_\theta(\mathcal{J}(x; Q)), y)]. \tag{2}$$

However, in order to solve (2) with gradient descent, the key challenge is caused by the non-differentiable quantization operation, which makes the gradients w.r.t. Q almost zero everywhere. To address this issue, we will introduce a *differentiable soft quantizer* (Q_d) in the next subsection, replacing the uniform quantizer (Q_u) used in $\mathcal{J}$.

3.2 DIFFERENTIABLE SOFT QUANTIZER

Denote the index set of uniform quantization as

$$\mathcal{A} = \{-L, -L+1, \dots, 0, \dots, L-1, L\}. \tag{3}$$

For convenience, $\mathcal{A}$ is also regarded as a vector of length $2L+1$. Multiplying $\mathcal{A}$ with a quantization step size q, we get the corresponding reconstruction space

$$\hat{\mathcal{A}} = q \times [-L, -L+1, \dots, 0, \dots, L-1, L]. \tag{4}$$

Again, we will regard $\hat{\mathcal{A}}$ as both a vector and a set.

To randomly quantize a DCT coefficient z to an element in $\hat{\mathcal{A}}$, we invoke from Yang & Hamidi (2024) a trainable conditional probability mass function (CPMF) $P_\alpha(\cdot|z)$ over the reconstruction space $\hat{\mathcal{A}}$ or equivalently the index set $\mathcal{A}$ given z, where $\alpha > 0$ is a trainable parameter:

$$P_\alpha(iq|z) = \frac{e^{-\alpha(z-iq)^2}}{\sum_{j\in\mathcal{A}} e^{-\alpha(z-jq)^2}}, \ \forall i \in \mathcal{A}. \tag{5}$$

Extend z to a vector of length $2L+1$, i.e., $[z]_{2L+1} = [\ \overbrace{z, \dots, z}^{2L+1 \text{ times}}\ ]$. Then, the CPMF $P_\alpha(\cdot|z)$, regarded as a vector of length $2L+1$, can be easily computed via the softmax operation $\sigma(\cdot)$:

$$\left[P_\alpha(\cdot|z)\right]_{2L+1} = \sigma\Big(-\alpha \times \Big(\left[z\right]_{2L+1} - \hat{\mathcal{A}}\Big)^2\Big). \tag{6}$$

With the CPMF $P_\alpha(\cdot|z)$, z is now quantized to each $iq \in \hat{\mathcal{A}}$ with probability $P_\alpha(iq|z)$. Note that as $\alpha \to \infty$, $P_\alpha(\cdot|z)$ approaches an one-hot vector with probability 1 at the nearest point to z in $\hat{\mathcal{A}}$ and 0 elsewhere. Therefore, the resulting random quantizer effectively functions as the deterministic uniform quantizer $Q_u(z) = \lfloor z/q \rceil \cdot q$.

Based on the CPMF $P_\alpha(\cdot|z)$, we can now define a differentiable soft quantizer Q_d as the conditional expectation of iq given z, i.e.,

$$Q_d(z) = \mathbb{E}[iq|z] = \sum_{i\in\mathcal{A}} P_\alpha(iq|z) \cdot iq. \tag{7}$$

Similarly, as $\alpha \to \infty$, Q_d also goes to Q_u. Fig. 2 shows how the shape of Q_d varies w.r.t α, given a fixed q.

This soft quantizer Q_d serves as an analytical proxy for Q_u. It's differentiable everywhere, allowing gradients to flow through it smoothly. More importantly, compared to Q_u, Q_d involves a trainable parameter

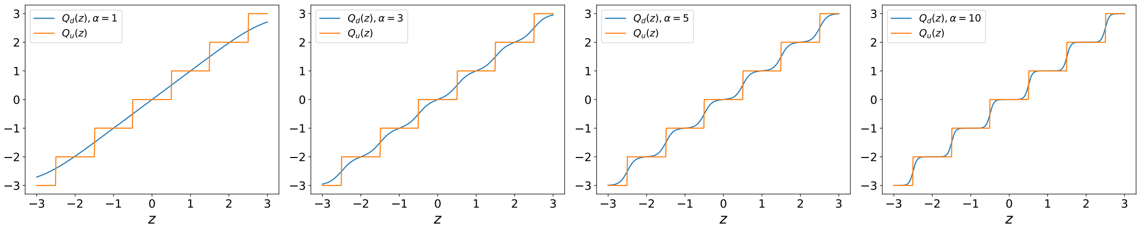


Figure 2: Illustration of Q_u vs. Q_d with $\alpha = 1, 3, 5, 10$, where L and q are set to 3 and 1, respectively.

α which can adjust the softness of the quantizer, thereby introducing more flexibility. In view of these nice properties, Q_d is the ideal candidate in place of Q_u used in $\mathcal{J}$.

As a side note, for the reader who is not familiar with quantization, but familiar with the attention operation used in transformer models (Vaswani, 2017), Q_d can be regarded as an attention operation in broad sense, with the query being z, the key and value being $\hat{\mathcal{A}}$, and the similarity metric between the query and key being negative squared distance instead of dot product.

3.3 OVERALL FRAMEWORK OF JPEG-DL

Substituting Q_u in $\mathcal{J}$ with Q_d, we get a differentiable JPEG layer $\hat{\mathcal{J}}$ parameterized by Q and $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_Y, \boldsymbol{\alpha}_C)$. $\boldsymbol{\alpha}_Y = [\alpha_1, \alpha_2, \dots, \alpha_M]$ and $\boldsymbol{\alpha}_C = [\alpha_{M+1}, \alpha_{M+2}, \dots, \alpha_{2M}]$ are α tables for the luminance and chrominance channels respectively, used in conjunction with Q_Y and Q_C to quantize DCT coefficients. Following the proposed soft quantization, we obtain quantized DCT coefficients $\hat{z}_{l,m,n} = Q_d(z_{l,m,n}; q_m, \alpha_m)$ for $l = 1$, and $\hat{z}_{l,m,n} = Q_d(z_{l,m,n}; q_{M+m}, \alpha_{M+m})$ for $l = 2, 3$, where $Q_d(z; q, \alpha)$ denotes a differentiable soft quantizer defined in (7) parameterized by a quantization step q and a scaling factor α. Overall, for an input image x, we have $\hat{x} = \hat{\mathcal{J}}(x; Q, \boldsymbol{\alpha})$. Therefore, we can rewrite (2), the JPEG-DL formulation, as

$$\min_{\theta, Q, \boldsymbol{\alpha}} \mathbb{E}[\mathcal{L}(f_\theta(\hat{\mathcal{J}}(x; Q, \boldsymbol{\alpha})), y)], \tag{8}$$

where the expectation can be approximated by the empirical mean over a mini-batch in actual training. Thanks to the use of Q_d, (8) can now be solved by gradient descent with ease (see Appendix A.1 for the derivative analytical formulas).

After training with JPEG-DL, we get the optimized parameters θ^*, Q^* and $\boldsymbol{\alpha}^*$. Then, we consider the composition of the JPEG layer and the underlying DNN as a unified DNN model $\hat{f}(x; \theta^*, Q^*, \boldsymbol{\alpha}^*) = f_{\theta^*}(\hat{\mathcal{J}}(x; Q^*, \boldsymbol{\alpha}^*))$ to do validation. Concretely, any raw input x should be fed into $\hat{\mathcal{J}}$, instead of directly to f_{θ^*}, thus allowing the reconstructed image $\hat{x}$ to be fed into the underlying DNN f_{θ^*}. In other words, the JPEG layer is prepended as the first layer of any underlying DNN.

To conclude this section, we refer readers to Fig. 1a and Fig. 1b for an illustration of the inner workings of the JPEG-DL framework.

4 EXPERIMENTS

CIFAR-100. We evaluate our proposed method using a transformer-based architecture and four state-of-the-art convolutional neural networks (CNNs): EfficientFormer-L1 (Li et al., 2022), ResNet (He et al., 2016), VGG (Simonyan & Zisserman, 2014), MobileNet (Sandler et al., 2018), and ShuffleNet (Ma et al., 2018). For ResNet, we employ CIFAR-specific versions: ResNet32, ResNet56, and ResNet110. For VGG, we utilize VGG8 and VGG13. All CNN architectures follow the training recipe from CRD (Tian et al., 2020) (see Appendix A.3), while for the transformer-based architecture, EfficientFormer-L1, we adhere to the setup proposed by Xu et al. (2023) (see Appendix A.5).

Fine-grained Tasks. These tasks involve visually similar classes and typically feature fewer training samples per class compared to conventional classification tasks. We evaluate our method on four datasets: CUB-200-2011 (Wah et al., 2011), Stanford Dogs (Khosla et al., 2011), Flowers (Nilsback & Zisserman, 2008), and Pets (Parkhi et al., 2012). For CNN architectures, we employ PreAct ResNet-18 (He et al., 2016) and DenseNet-BC (Huang et al., 2017), following the experimental setup and architecture modifications of Zhang et al. (2017). For the transformer-based architecture, we use EfficientFormer-L1, adopting the setup outlined by Xu et al. (2023) (see Appendices A.4 and A.5).

JPEG-layer settings for CIFAR-100 and fine-grained tasks. To train our JPEG layer, we first study the gradient nature of Q_d with respect to q and α in (5) by changing one parameter and fixing the other. Both q and α are sharable parameters among different numbers of blocks per image. During the calculation of the gradient of Q_d with respect to α, our analysis shows that when α is sufficiently large, such that Q_d is not too far from Q_u, the gradient magnitude for α is almost zero, as shown in Appendix A.1. This indicates that $\boldsymbol{\alpha}$ will not be updated effectively if initialized within a reasonable range. Experiments also confirmed that making $\boldsymbol{\alpha}$ trainable does not significantly impact model performance. As a result, we choose not to train $\boldsymbol{\alpha}$ in our framework. However, when we explore the calculation of the gradient of Q_d with respect to q, the accumulated gradient received from different blocks per image during backpropagation results in unstable gradient magnitudes at a fixed value of α, as shown in our analysis in Appendix A.1. To address this, we use the ADAM optimizer, which adapts the learning rate for each trainable parameter q in our JPEG layer, enabling more efficient training and better convergence. For CIFAR-100, we set the learning rate to 0.003 across all tested models. For the fine-grained datasets, we set the JPEG learning rate to 0.005. Across these datasets, we fix $\alpha_m = 5$ for all $1 \le m \le 2M$, and set $L = 2^{b-1}$ in (3), where b is a tunable hyperparameter set to 8.

ImageNet-1K. For all experiments on this dataset, we utilize the standard training recipes shown by Paszke et al. (2019) without any modifications. We use SqueezeNet (Iandola, 2016), ResNet-18 and ResNet-34 as our testing underlying models.

JPEG-layer settings for ImageNet-1K. We will utilize specific settings to control the gradient magnitude to ensure more stable updates for Q. We define the Gradient Scaling Constants $\hbar_m = \alpha_m q_m^2$, $1 \le m \le 2M$, which allows us to control the magnitude of gradients w.r.t Q. Specifically, we fix $\hbar_m = 0.7$ for all $1 \le m \le 2M$. During training, we update the value of α_m to be $\hbar_m / q_m^2$ before calculating the gradients w.r.t Q. As a result of controlling the maximum gradient magnitude, we can optimize Q using an SGD optimizer with 0.5 learning rate, instead of an ADAM optimizer. This approach ensures stable and efficient training for quantization table updates (see Appendix A.2 for more details). For this dataset, b is equal to 11 across all tested models.

Quantization Table Initialization. For CIFAR-100 and fine-grained tasks, we initialize the quantization table Q based on the reciprocal of sensitivity for each DCT frequency, given a pre-trained model of the underlying DNN architecture, following the approach described by Zheng et al. (2023); Salamah et al. (2024b). Broadly speaking, the sensitivity of a frequency indicates the rate of change of the loss function w.r.t. the perturbation on this frequency, so we favor a smaller quantization step for a more sensitive frequency to limit the distortion amount on it. For ImageNet-1K, we adopt the strategy from Esser et al. (2019), where the initialization of q_m is based on the average of absolute values of DCT coefficients across all blocks in all training images that will be quantized by q_m. Specifically, $q_m = 2\sum_{k=1}^{N}\sum_{n=1}^{B}|z_{1,m,n}^{(k)}|/(NB\sqrt{2^{b-1}})$ for $1 \le m \le M$, i.e. Q_Y, and $q_m = \sum_{k=1}^{N}\sum_{l=2}^{3}\sum_{n=1}^{B}|z_{l,m,n}^{(k)}|/(NB\sqrt{2^{b-1}})$ for $M+1 \le m \le 2M$, i.e., Q_C, where k is the index for image and N is the number of training images.

Table 1: Top-1 validation accuracy (%) for Baseline and JPEG-DL on CIFAR-100. The Baseline results are from Tian et al. (2020). For JPEG-DL, we report the mean and standard deviation of experimental results over three runs.

Method	Res32	Res56	Res110	VGG8	VGG13	MobileNetV2	ShuffleNetV2
Baseline	71.14	72.34	73.79	70.36	73.77	64.6	71.82
JPEG-DL	**71.92**$_{\pm 0.31}$ (+0.78)	**73.39**$_{\pm 0.19}$ (+1.05)	**74.46**$_{\pm +0.11}$ (+0.67)	**71.10**$_{\pm +0.41}$ (+0.74)	**75.32**$_{\pm 0.10}$ (+1.55)	**65.91**$_{\pm 0.11}$ (+1.31)	**73.04**$_{\pm 0.16}$ (+1.22)

CIFAR-100 and Fine-grained tasks Results. The performance of JPEG-DL is shown in Tables 1 and 2. Across all seven tested models for CIFAR-100, JPEG-DL consistently provides improvements, with gains of up to 1.53% in top-1 accuracy. In the fine-grained tasks, JPEG-DL offers a substantial performance increase,

Table 2: Top-1 validation accuracy (%) on various fine-grained image classification tasks and model architectures. We report the mean and standard deviation of experimental results over three runs.

Model	Method	CUB-200	Dogs	Flowers	Pets
ResNet-18	Baseline	54.00±1.43	63.71±0.32	57.13±1.28	70.37±0.84
	JPEG-DL	**58.81**±0.12 (+4.81)	**65.57**±0.37 (+1.86)	**68.76**±0.57 (+11.63)	**74.84**±0.66 (+4.47)
DenseNet-121	Baseline	57.70±0.44	66.61±0.17	51.32±0.57	70.26±0.79
	JPEG-DL	**61.32**±0.43 (+3.62)	**69.67**±0.58 (+3.06)	**72.22**±1.05 (+20.90)	**75.90**±0.68 (+5.64)

Table 3: Top-1 validation accuracy (%) on ImageNet with different model architectures.

Method	SqueezeNetV1.1	Resnet18	Resnet34
Baseline	57.95	69.75	73.31
JPEG-DL	**58.26** (+0.31)	**70.13** (+0.38)	**73.54** (+0.23)

with improvements of up to 20.90% across all datasets using two different models. Additional results for CIFAR-100 and fine-grained tasks using a transformer-based model are shown in Appendices A.5. On fine-grained tasks, we have extended our results to include a comparison with more baselines that address the non-differentiability and zero derivative problems of JPEG quantization. Additionally, we provide a comprehensive comparison with non-learnable and learnable preprocessing methods. For a detailed analysis of these comparisons, please refer to Appendices A.6 and A.7.

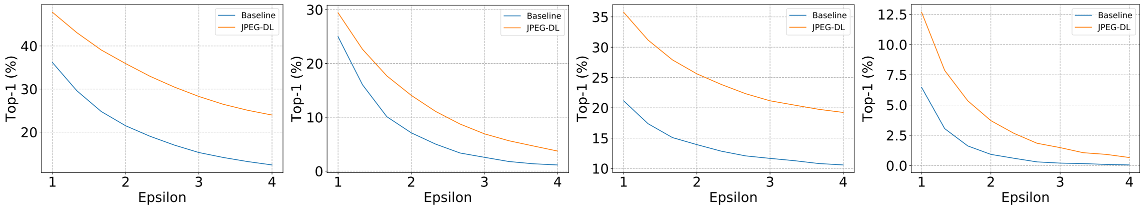
ImageNet-1K Results. The performance of JPEG-DL is shown in Table 3. With a trivial increase in complexity (adding 128 parameters), JPEG-DL achieves a gain of 0.31% in top-1 accuracy for SqueezeNetV1.1 compared to the baseline using a single round of Q_d quantization operation. By increasing the number of quantization rounds to five, we observe an additional improvement of 0.20%, leading to a total gain of 0.51% over the baseline. The best results are indicated in bold, and values in parentheses indicate relative accuracy gains over the baseline.

5 ANALYSIS AND DISCUSSION

Robustness. JPEG has generally been used as an empirical defense mechanism against adversarial attacks by mitigating adversarial perturbations and enhancing the robustness of DNNs, as discussed in Section 2. To evaluate the adversarial robustness of JPEG-DL models in comparison to standard DNN, we conduct experiments using two attack methods, FGSM and PGD, on CIFAR-100 with two different models from Table 1. The perturbation budget, Epsilon, ranged from 1 to 4 for both attack methods. For PGD, we applied 5 steps with a perturbation step size of $(2.5 \times \text{Epsilon})/\text{steps}$, following the setup used by Madry (2017). As shown in Fig. 3, the JPEG-DL models significantly improve the adversarial robustness compared to the standard DNN models, with improvements of up to 15% for FGSM and 6% for PGD. We will show some examples of the quantization tables used in this study in the next subsection.

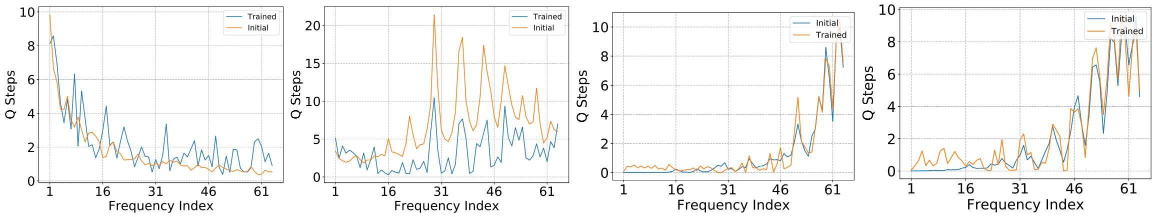
Designed Quantization Tables. Fig. 4 presents the Y and CbCr quantization tables, both at initialization and after convergence, for VGG13 trained on CIFAR-100 and ResNet-18 trained on the CUB200 dataset. The estimated sensitivity values used to initialize these quantization tables are provided in Appendix A.9.

Feature maps visualization. Fig. 5 presents the feature maps extracted after the first dense block in DenseNet-121 for both the JPEG-DL model and the baseline model, trained on the CUB200 dataset using the model from Table 2. The output of the feature maps at this stage is of size 56×56, and both sets are shown in the same sequence using the same original image. The shown example was incorrectly classified by the baseline model, while the JPEG-DL model correctly classified it. In this figure, it is evident



(a) FGSM for VGG13 (b) PGD for VGG13 (c) FGSM for Res56 (d) PGD for Res56

Figure 3: Evaluate the adversarial robustness of JPEG-DL models in comparison to standard DNN on VGG13 and Res56 for CIFAR-100 against FGSM and PGD attacks.



(a) VGG13 (Y Channel) (b) VGG13 (CbCr Channel) (c) ResNet-18(Y Channel) (d) ResNet-18 (CbCr Channel)

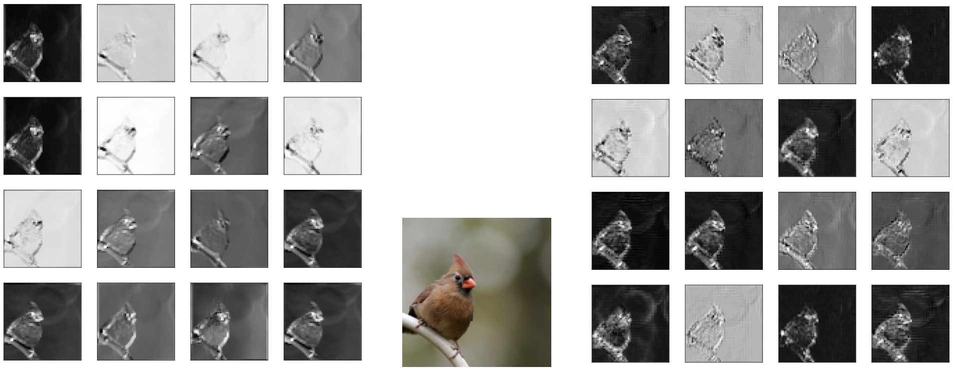
Figure 4: Initial and final quantization tables for VGG13 trained on CIFAR-100 and ResNet18 trained on CUB200, with frequency indices arranged in the default zigzag order.

that the feature maps from the JPEG-DL model show significantly better contrast between the foreground information (the bird) and the background compared to the feature maps generated by the baseline model. Specifically, the foreground object in the JPEG-DL feature maps is enclosed within a well-defined contour, making it visually distinguishable from the background. In contrast, the baseline model's feature maps show a more blended structure, where the foreground contains higher energy in low frequencies, causing it to blend more smoothly with the background. Additionally, another example in Appendix A.10 shows a similar phenomenon in addition to background information being more effectively removed, following the same setup as the first example. These discrepancies in feature map clarity and contrast are propagated through subsequent blocks of DenseNet-121, eventually contributing to the misclassification problem observed in the baseline model.

Interpretability using CAM. Fig. 6 illustrates GradCAM++ visualizations (Chattopadhay et al., 2018) for two examples from the CUB-200 dataset, comparing the baseline model with JPEG-DL using ResNet-18 as our underlying model. In both instances, the baseline model incorrectly classifies the input, while the JPEG-DL model correctly classifies it. The visualizations highlight how JPEG-DL focuses more precisely on the main object in the image, demonstrating the model's improved attention to key regions that contribute to correct classification. This highlights the effectiveness of JPEG-DL in enhancing model interpretability and performance, and it also supports the case for background removal discussed in the previous subsection.

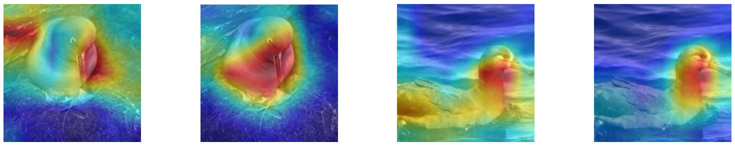
6 CONCLUSION

In contrast to the conventional understanding that JPEG compression negatively impacts the DL performance, this paper introduces a novel trainable JPEG compression layer into the DL pipeline to improve DL performance, termed JPEG-DL. This layer enables gradient-based optimization for quantization pa-



(a) Baseline (b) Input (c) JPEG-DL

Figure 5: Feature maps of size 56×56 are shown after the first dense block in DenseNet-121 for both JPEG-DL and baseline models Figs. in 5a and 5c, respectively, using an original input shown in Fig.5b. The JPEG-DL model highlights the foreground (bird) more distinctly, while the baseline model shows less contrast, contributing to its misclassification.



(a) Baseline (b) JPEG-DL (c) Baseline (d) JPEG-DL

Figure 6: GradCAM++ visualization for baseline and JPEG-DL models on CUB200 using two examples, where the baseline model incorrectly classified them and the JPEG-DL model correctly classified them.

rameters and can be integrated seamlessly with any underlying DNN to be trained jointly. To validate the effectiveness of JPEG-DL, we conduct extensive experiments on six image classification datasets and show significant gains over the standard DL. Additionally, we demonstrate that JPEG-DL improves the adversarial robustness of the learned model.

REFERENCES

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

Tajeddine Benbarrad, Salaheddine Kably, Mounir Arioua, and Nabih Alaoui. Compression-based data augmentation for cnn generalization. In *International Conference on Cybersecurity, Cybercrimes, and Smart Emerging Technologies*, pp. 235–244. Springer, 2022.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Chyi-Tsong Chen and Wei-Der Chang. A feedforward neural network with function shape autotuning. *Neural networks*, 9(4):627–641, 1996.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *ArXiv*, abs/1705.02900, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6. IEEE, 2016.

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of JPG compression on adversarial images. *CoRR*, abs/1608.00853, 2016.

Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Forrest N Iandola. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer, 2011.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yanyu Li, Geng Yuan, Yang Wen, Eric Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *arXiv preprint arXiv:2206.01191*, 2022.

Zihao Liu, Tao Liu, Wujie Wen, Lei Jiang, Jie Xu, Yanzhi Wang, and Gang Quan. DeepN-JPEG: A deep neural network favorable JPEG-based image compression framework. In *Proceedings of the 55th Annual Design Automation Conference*, pp. 18. ACM, 2018.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

Xiyang Luo, Hossein Talebi, Feng Yang, Michael Elad, and Peyman Milanfar. The rate-distortion-accuracy tradeoff: Jpeg case study. *arXiv preprint arXiv:2008.00605*, 2020.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

Aleksander Madry. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

William B. Pennebaker and Joan L. Mitchell. *JPEG Still Image Data Compression Standard*. Kluwer Academic Publishers, Norwell, MA, United States, 1992. ISBN 0442012721.

Ahmed H Salamah, Shayan Mohajer Hamidi, and En-Hui Yang. A coded knowledge distillation framework for image classification based on adaptive jpeg encoding. *Pattern Recognition*, pp. 110966, 2024a.

Ahmed H Salamah, Kaixiang Zheng, Linfeng Ye, and En-Hui Yang. Jpeg compliant compression for dnn vision. *IEEE Journal on Selected Areas in Information Theory*, 2024b.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Richard Shin and Dawn Song. Jpeg-resistant adversarial images. In *NIPS 2017 workshop on machine learning and computer security*, volume 1, pp. 8, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Zhengzhong Tu, Peyman Milanfar, and Hossein Talebi. Muller: Multilayer laplacian resizer for vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6877–6887, 2023.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Xiufeng Xie and Kyu-Han Kim. Source compression with bounded dnn perception loss for iot edge computer vision. In *The 25th Annual International Conference on Mobile Computing and Networking*, pp. 1–16, 2019.

Xiufeng Xie, Ning Zhou, Wentao Zhu, and Ji Liu. Bandwidth-aware adaptive codec for dnn inference offloading in iot. In *European Conference on Computer Vision*, pp. 88–104. Springer, 2022.

Zhiqiu Xu, Yanjie Chen, Kirill Vishniakov, Yida Yin, Zhiqiang Shen, Trevor Darrell, Lingjie Liu, and Zhuang Liu. Initializing models with larger ones. In *The Twelfth International Conference on Learning Representations*, 2023.

En-Hui Yang and Shayan Mohajer Hamidi. Systems and methods for training deep learning models. 2024.

En-Hui Yang, Hossam Amer, and Yanbing Jiang. Compression helps deep learning in image classification. *Entropy*, 23(7):881, 2021.

Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13876–13885, 2020.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Kaixiang Zheng, Ahmed H. Salamah, Linfeng Ye, and En-Hui Yang. Jpeg compliant compression for dnn vision. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 1875–1879, 2023. doi: 10.1109/ICIP49359.2023.10221982.

A APPENDIX

A.1 PARTIAL DERIVATIVES OF Q_d

With the CPMF $P_\alpha(\cdot|z)$ shown in (6), z is now quantized to each $\hat{z} \in \hat{\mathcal{A}}$ with probability $P_\alpha(\hat{z}|z)$. Denote this random mapping by

$$\hat{z} = Q_{\mathrm{p}}(z). \tag{9}$$

Note that given z, $\hat{z}$ is a random variable taking values in $\hat{\mathcal{A}}$ with distribution $P_\alpha(\cdot|z)$.

A.1.1 QUANTIZATION STEP AND INPUT

The partial derivatives of $Q_d(z)$ w.r.t. q and z are obtained as

$$\frac{\partial Q_d(z)}{\partial z} = 2\alpha \mathrm{Var}\big\{Q_p(z)\big\}, \tag{10a}$$

$$\begin{aligned}\frac{\partial Q_d(z)}{\partial q} &= \frac{1}{q}\Big(\mathbb{E}\big\{Q_p(z)\big\} + (2\alpha z)\mathrm{Var}\big\{Q_p(z)\big\} \\ &\qquad - (2\alpha)\mathrm{Skew}_{\mathrm{u}}\big\{Q_p(z)\big\}\Big),\end{aligned} \tag{10b}$$

where for any random variable V,

$$\mathrm{Skew}_{\mathrm{u}}(V) \triangleq \sum_v v^3 \mathbb{P}_V(v) - \Big(\sum_v v\mathbb{P}_V(v)\Big)\Big(\sum_v v^2\mathbb{P}_V(v)\Big).$$

A.1.2 SCALING FACTOR α

The partial derivatives of $Q_d(Z)$ w.r.t. α obtained as

$$\begin{aligned}\frac{\partial Q_d(z)}{\partial \alpha} &= \frac{\partial \sum_{i\in\mathcal{A}} iqP_\alpha(iq|z)}{\partial \alpha} \\ &= \sum_{i\in\mathcal{A}} iq\frac{\partial P_\alpha(iq|z)}{\partial \alpha}\end{aligned} \tag{11}$$

$$\begin{aligned}\frac{\partial P_\alpha(iq|z)}{\partial \alpha} &= \frac{-(z-iq)^2 e^{-\alpha(z-iq)^2}}{\sum_{j\in\mathcal{A}} e^{-\alpha(z-jq)^2}} + \frac{e^{-\alpha(z-iq)^2}\sum_{j\in\mathcal{A}}(z-jq)^2 e^{-\alpha(z-jq)^2}}{(\sum_{j\in\mathcal{A}} e^{-\alpha(z-jq)^2})^2} \\ &= P_\alpha(iq|z)\sum_{j\in\mathcal{A}}(z-jq)^2 P_\alpha(jq|z) - (z-iq)^2 P_\alpha(iq|z)\end{aligned} \tag{12}$$

Plugging (2) in (1) yields

$$\begin{aligned}\frac{\partial Q_d(z)}{\partial \alpha} &= \sum_{i\in\mathcal{A}} iqP_\alpha(iq|z)\sum_{j\in\mathcal{A}}(z-jq)^2 P_\alpha(jq|z) - \sum_{i\in\mathcal{A}} iq(z-iq)^2 P_\alpha(iq|z) \\ &= \mathbb{E}\{Q_p(z)\}\mathbb{E}\{(z-Q_p(z))^2\} - \mathbb{E}\{Q_p(z)(z-Q_p(z))^2\} \\ &= -Cov\{Q_p(z),(z-Q_p(z))^2\}\end{aligned} \tag{13}$$

To gain a better understanding, we fix q and b, and analyzed how $\frac{\partial Q_d(z)}{\partial \alpha}$ behaves for different values of α. From Fig. 7, it is evident that the gradient magnitude decreases as α increases. This indicates that Q_d approaches the shape of the uniform quantizer Q_u, as shown previously in Fig. 2, leading to a reduction in the gradient magnitude for α. As a result, when α becomes sufficiently large, the gradient approaches zero, effectively preventing further updates to α, making it act as a non-trainable parameter at a certain point of the training process. Based on this observation and verified experimental results, we opt not to train α for simple image classification tasks like CIFAR-100 and fine-grained datasets, as it does not impact the overall performance.

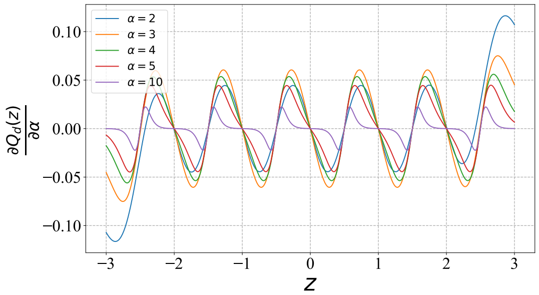
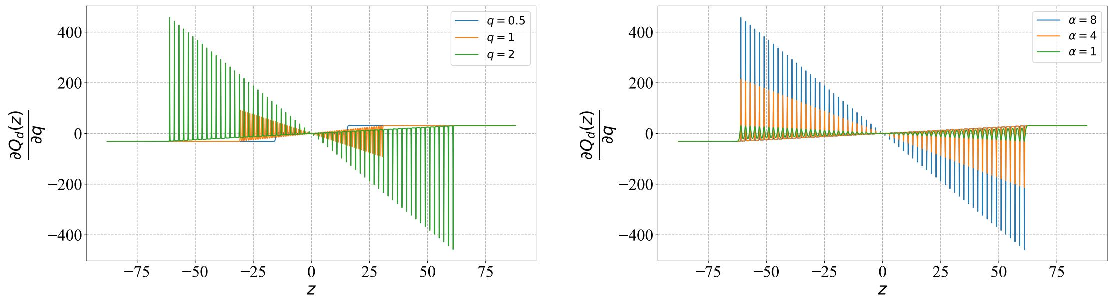


Figure 7: This Figure illustrates the partial derivatives of $Q_d(z)$ w.r.t. the scaling factor α, where $b = 3$ and $q = 1$.

A.2 GRADIENT SCALING

Figures 8a and 8b demonstrate the impact of varying α and q on the gradient magnitude of $Q_d(z)$ with respect to q. As observed, decreasing the value of q while keeping α constant leads to a decrease in the gradient magnitude. Similarly, decreasing the value of α while maintaining a fixed q also results in a reduction of the gradient magnitude. To address the instability in updating the trainable parameters Q, we propose utilizing α to regulate the magnitude of $\frac{\partial Q_d(z)}{\partial q}$, as illustrated in Figure 8b. By leveraging this control mechanism of adjusting α, we can stabilize the magnitude of the gradients that update q. This behavior suggests a relationship between q and α in controlling the gradient magnitude of $Q_d(z)$.

To explore the relationship between α and q, we refer to the exponent in (5), $\alpha(z - iq)^2$, which can be rewritten by expressing $z = cq$, resulting in the form $\alpha q^2(c-i)^2$. From this, we define a new term, $\hbar = \alpha q^2$, referred to as the *Gradient Scaling Constant*. To further illustrate this relationship, in Fig. 9, we set $\hbar = 2$, and by selecting different pairs of α and q values, we demonstrate that the maximum magnitude of $\frac{\partial Q_d(z)}{\partial q}$ remains invariant. This confirms that the gradient magnitude can be effectively controlled by adjusting α based on the last updated value of q, according to the specified $\hbar$ value. This adjustment allows for controlled quantization updates, reducing the potential instability during training. Moreover, by controlling the gradient magnitude, we simplify the optimization process, enabling the use of a single learning rate for all q values by using the SGD optimizer, instead of the ADAM optimizer. This gradient scaling mechanism is analogous to the ADAM optimizer, which adapts different learning rates for individual trainable parameters based on momentum and recent gradient magnitudes.



(a)

(b)

Figure 8: Figures 8a and 8b illustrate the partial derivatives of $Q_d(z)$ w.r.t. the parameter q for cases where $\alpha = 8$ and q varies, and α varies and $q = 2$, respectively. For both figures, we set $b = 6$.

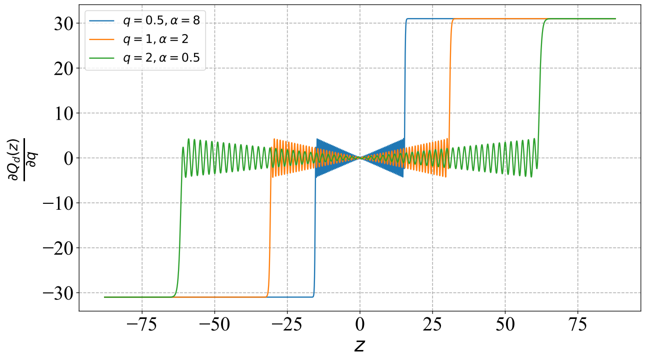


Figure 9: Demonstrating impact of $\hbar$ on gradient magnitude for various combinations of q and α. We fixed $\hbar$=2 and b=6.

A.3 CNN-BASED ARCHITECTURES SETTING

For CIFAR100, we deploy a stochastic gradient descent (SGD) optimizer with a momentum of 0.9, a weight decay of 0.0005, and a batch size of 64. We initialize the learning rate as 0.05, and decay it by 0.1 every 30 epochs after the first 150 epochs until the last 240 epoch. For MobileNetV2, ShuffleNetV1 and ShuffleNetV2, we use a learning rate of 0.01 as this learning rate is optimal for these models in a grid search, while 0.05 is optimal for other models.

For fine-grained classification tasks, all networks are trained from scratch and optimized by SGD with a momentum of 0.9, weight decay of 0.0001, and an initial learning rate of 0.1,. The learning rate is divided by 10 after epochs 100 and 150 for all datasets, and the total epochs are 200. We set batch size 32 for these fine-grained classification tasks. We use the standard data augmentation technique for ImageNet (Deng et al., 2009), *i.e.*, flipping and random cropping. This experimental setup is also outlined by Yun et al. (2020).

A.4 FINE-GRAINED MODEL ARCHITECTURES

We use standard ResNet-18 with 64 filters and DenseNet-121 with a growth rate of 32 for image size 224 × 224. For fine-grained classification tasks, we use PreAct ResNet-18 (He et al., 2016), which modifies the first convolutional layer with kernel size 3×3, strides 1 and padding 1, instead of the kernel size 7×7, strides 2 and padding 3, for image size 32×32 by following Zhang et al. (2017). We use DenseNet-BC structure (Huang et al., 2017), and the first convolution layer of the network is also modified in the same way as in PreAct ResNet-18 for image size 32×32.

A.5 TRANSFORMER-BASED SETTINGS AND RESULTS

In this section, we compare the performance of JPEG-DL compared to its baseline using the EfficientFormer-L1 model (Li et al., 2022) on CIFAR-100, as well as two fine-grained datasets, Flowers and Pets. We followed the experimental setup described by Xu et al. (2023), adhering to the same configurations mentioned in Table 4. The learning rate was set to 0.003 for CIFAR-100 and 0.005 for the fine-grained tasks, aligning with our standard settings in Section 4.

Interestingly, we found that the best performance for transformer-based architectures was achieved when the quantization tables were initialized with all ones, effectively representing the highest quality factor quantization table for JPEG. The top-1 validation accuracy performance is presented in Table 5, demonstrating the significant improvements achieved by JPEG-DL over the baseline.

Table 4: Hyper-parameter setting on EfficientFormer-L1.

Settings	CIFAR-100	Pets	Flowers
batch size	512	512	512
warmup epochs	50	100	100
training epochs	300	600	600

Table 5: Top-1 accuracy (%) on CIFAR-100 and two fine-grained image classification tasks using EfficientFormer-L1. We report the mean and standard deviation of the experimental results over three runs with different random seeds. The best results are indicated in bold, and values in parentheses indicate relative accuracy gains over the baseline.

Model	Method	Accuracy (%)
CIFAR-100	Baseline	80.27$_{\pm 0.33}$
	JPEG-DL	**80.49**$_{\pm 0.23}$ (+0.22)
Flowers	Baseline	69.78$_{\pm 0.33}$
	JPEG-DL	**73.05**$_{\pm 0.12}$ (+3.27)
Pets	Baseline	65.52$_{\pm 1.23}$
	JPEG-DL	**69.28**$_{\pm 0.42}$ (+3.76)

A.6 COMPARISON WITH MORE BASELINES

In Table 6, we have included comparisons against other baselines that employed JPEG quantization in the pipeline, but with heuristic approaches to handle the non-differentiability and zero derivative problems of JPEG quantization. For example, Ballé et al. (2016) allow optimization via stochastic gradient descent by replacing the quantizer with an additive i.i.d. uniform noise, which has the same width as the quantization bins, where $\hat{c} = c + q * U(-0.5, 0.5)$. However, during validation, they employed Q_u using the trained quantization tables. Shin & Song (2017) employ a third-order polynomial approximation of the rounding function to make JPEG differentiable. The gradient approximation through the round function is a key aspect of certain neural network approaches. Esser et al. (2019)[1] employ a straight-through estimator, originally proposed by Bengio et al. (2013), to achieve this approximation. This method treats the round function as a pass-through operation during backpropagation, allowing for effective gradient estimation. Our experimental results demonstrate that we consistently outperform all tested baselines. Even though these methods find a way to make the quantization differentiable, they still hurt the performance. This is consistent with the well-known wisdom in the literature, which also shows that they did not take advantage of the higher level of non-linearity introduced into DNN architectures, as proposed by our JPEG-DL framework.

Table 6: Top-1 validation accuracy (%) on various fine-grained image classification tasks and model architectures. We report the mean and standard deviation of experimental results over three runs.

Model	Method	CUB-200	Dogs	Flowers	Pets
ResNet-18	Baseline	54.00±1.43	63.71±0.32	57.13±1.28	70.37±0.84
	Ballé et al. (2016)	50.78±2.21 (-3.22)	53.47±7.37 (-10.24)	55.46±0.59 (-1.67)	56.14±17.16 (-14.23)
	Shin & Song (2017)	55.34±0.14 (+1.34)	63.03±0.56 (-0.68)	55.78±1.44 (-1.35)	71.45±1.01 (+1.08)
	Esser et al. (2019)	51.58±0.18 (-2.42)	60.45±0.23 (-3.26)	58.04±0.58 (+0.91)	68.81±0.55 (-1.56)
	JPEG-DL	**58.81**±0.12 (+4.81)	**65.57**±0.37 (+1.86)	**68.76**±0.57 (+11.63)	**74.84**±0.66 (+4.47)
DenseNet-121	Baseline	57.70±0.44	66.61±0.17	51.32±0.57	70.26±0.79
	Ballé et al. (2016)	52.00±1.41 (-5.70)	60.07±6.41 (-6.54)	46.60±2.87 (-4.72)	61.91±1.88 (-8.35)
	Shin & Song (2017)	57.19±0.78 (-0.51)	66.90±0.13 (+0.29)	51.04±0.87 (-0.28)	69.95±1.21 (-0.31)
	Esser et al. (2019)	56.46±0.30 (-1.24)	64.89±0.12 (-1.72)	55.98±0.24 (+4.60)	69.58±0.59 (-0.68)
	JPEG-DL	**61.32**±0.43 (+3.62)	**69.67**±0.58 (+3.06)	**72.22**±1.05 (+20.90)	**75.90**±0.68 (+5.64)

A.7 Comparison with JPEG-based Data Augmentation and Other Preprocessing Methods

JPEG-based Data Augmentation. We have compared and implemented JPEG-based data augmentation across three different sets, each with varying ranges of quantity factor (QF). For each tested range, we randomly select a QF for each image within the mini-batch. These sets have been tested on fine-grained datasets for all the models evaluated. Table 7 shows corresponding results for these tested sets. Consistent with the common knowledge in the literature, JPEG-based data augmentation in general degrades the accuracy performance.

Non-Learnable and Learnable Preprocessing Methods. We have tested the performance of non-learnable and learnable preprocessing methods during training and validation across all tested models on all tested fine-grained datasets, as demonstrated in Table 7. For non-learnable preprocessing, we have considered applying denoising using a Gaussian kernel and histogram equalization for each image within a mini-batch. As for the learnable one, we have compared it with Tu et al. (2023) as a learnable resize module, which has a bandpass nature in that it learns to boost details in certain frequency subbands that benefit the downstream recognition models. We have applied the same setup mentioned in their paper, in which they fine-tuned the model to achieve some improvement [2]

---

[1] We followed the experimental setup defined by Esser et al. (2019) using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, adhering to the same learning rate decay schedule as the underlying model. Additionally, we initialized Q by setting it to $2|Q|/\sqrt{2^b - 1}$ and implemented gradient scaling during training. We also maintained consistency by using the same number of bits used by JPEG-DL.

[2] We have verified the correctness of our PyTorch implementation with their TensorFlow implementation shown `https://colab.research.google.com/github/google-research/google-research/blob/master/muller/muller_demo.ipynb`.

Table 7: JPEG-DL is compared to JPEG-based augmentation (orange), non-learnable (yellow), and learnable preprocessing (blue) methods. JPEG-DL employs three different chrominance subsampling schemes shown in green color. Top-1 validation accuracy (%) on various fine-grained image classification tasks and model architectures. We report the mean and standard deviation of experimental results over three runs. The best results are indicated in bold, and values in parentheses indicate relative accuracy gains over the baseline.

Dataset	Method	ResNet-18	DenseNet-121
CUB-200	Baseline	54.00±1.43	57.70±0.44
	Rand. QF [1:50]	52.86±0.69 (-1.14)	55.72±0.72 (-1.98)
	Rand. QF [50:100]	54.53±0.41 (+0.53)	56.98±0.90 (-0.72)
	Rand. QF [1:100]	53.91±0.43 (-0.09)	57.20±0.33 (-0.50)
	Denoising	54.82±0.40 (+0.82)	56.02±1.00 (-1.68)
	Equalization	49.00±0.31 (-5.00)	54.32±0.57 (-3.38)
	Learnable Resize	54.98±0.19 (+0.98)	57.35±0.22 (-0.35)
	JPEG-DL (4:2:0)	58.00±0.12 (+4.00)	60.55±0.71 (+2.85)
	JPEG-DL (4:2:2)	58.11±0.10 (+4.11)	**61.51**±0.41 (+3.81)
	JPEG-DL (4:4:4)	**58.81**±0.12 (+4.81)	61.32±0.43 (+3.62)
Dogs	Baseline	63.71±0.32	66.61±0.17
	Rand. QF [1:50]	60.61±0.17 (-3.10)	64.93±0.17 (-1.68)
	Rand. QF [50:100]	63.14±0.30 (-0.57)	66.97±0.18 (+0.36)
	Rand. QF [1:100]	62.12±0.16 (-1.59)	65.59±0.54 (-1.02)
	Denoising	62.52±0.41 (-1.19)	66.36±0.34 (-0.25)
	Equalization	62.38±0.29 (-1.33)	67.12±0.19 (+0.51)
	Learnable Resize	63.10±0.43 (-0.61)	67.90±0.32 (+1.29)
	JPEG-DL (4:2:0)	65.57±0.31 (+1.86)	68.90±0.08 (+2.29)
	JPEG-DL (4:2:2)	**65.64**±0.16 (+1.93)	**69.85**±0.78 (+3.24)
	JPEG-DL (4:4:4)	65.57±0.37 (+1.86)	69.67±0.58 (+3.06)
Flowers	Baseline	57.13±1.28	51.32±0.57
	Rand. QF [1:50]	58.33±0.48 (+1.20)	51.44±0.79 (+0.12)
	Rand. QF [50:100]	55.98±0.89 (-1.15)	51.67±1.10 (+0.35)
	Rand. QF [1:100]	57.55±0.66 (+0.42)	52.32±0.73 (+1.00)
	Denoising	57.28±1.06 (+0.15)	50.37±0.79 (-0.95)
	Equalization	60.56±0.98 (+3.43)	61.72±0.21 (+10.40)
	Learnable Resize	57.41±0.48 (+0.28)	51.21±0.82 (-0.11)
	JPEG-DL (4:2:0)	67.58±1.50 (+10.45)	68.01±1.17 (+16.69)
	JPEG-DL (4:2:2)	67.75±1.19 (+10.62)	68.79±1.08 (+17.47)
	JPEG-DL (4:4:4)	**68.76**±0.57 (+11.63)	**72.22**±1.05 (+20.90)
Pets	Baseline	70.37±0.84	70.26±0.79
	Rand. QF [1:50]	69.71±0.54 (-0.66)	68.93±0.50 (-1.33)
	Rand. QF [50:100]	70.00±0.57 (-0.37)	68.63±0.82 (-1.63)
	Rand. QF [1:100]	69.52±0.62 (-0.85)	70.60±1.05 (+0.34)
	Denoising	69.15±1.07 (-1.22)	69.03±1.13 (-1.23)
	Equalization	73.00±0.79 (+2.63)	75.04±1.30 (+4.78)
	Learnable Resize	70.04±0.72 (-0.33)	68.83±1.85 (-1.43)
	JPEG-DL (4:2:0)	74.64±1.34 (+4.27)	**76.21**±0.35 (+5.95)
	JPEG-DL (4:2:2)	74.81±0.51 (+4.44)	76.17±1.83 (+5.90)
	JPEG-DL (4:4:4)	**74.84**±0.66 (+4.47)	75.90±0.68 (+5.64)

Impact of Chromance subsampling on JPEG-DL. In Table 7, we evaluate various subsampling schemes for all tested models across all tested fine-grained datasets. Notably, although there is no clear winner among different chroma subsampling methods, DNNs indeed respond to color information differently from human; color information is more important to DNNs than human since chroma subsampling formats 4:4:4 and 4:2:2, in general, give rise to better accuracy performance than the 4:2:0 subsampling format which is adopted predominantly in the image and video coding.

A.8 Fixing Hyperparamters across different datasets

In Table 8, we evaluate the difference between the hyperparameters used for CIFAR-100 and the fine-grained task, where the only difference lies in the learning rate. In the following setup, we present the results when a learning rate of 0.003 is applied to the fine-grained dataset. We observe that the performance difference is marginal, and in some cases, we can even achieve higher gains in model performance. This suggests that the choice of learning rate can have a significant impact on the effectiveness of the model.

Table 8: Top-1 validation accuracy (%) on various fine-grained image classification tasks and model architectures. We report the mean and standard deviation of experimental results over three runs.

Model	Learning Rate	CUB-200	Dogs	Flowers	Pets
ResNet-18	0.005	**58.81**±0.12 (+4.81)	**65.57**±0.37 (+1.86)	68.76±0.57 (+11.63)	74.84±0.66 (+4.47)
	0.003	58.62±0.50 (+4.61)	65.45±0.17 (+1.74)	**69.61**±1.16 (+12.48)	**74.90**±0.82 (+4.53)
DenseNet-121	0.005	**61.32**±0.43 (+3.62)	**69.67**±0.58 (+3.06)	72.22±1.05 (+20.90)	**75.90**±0.68 (+5.64)
	0.003	60.92±0.50 (+3.22)	69.53±0.49 (+2.92)	**72.45**±0.92 (+21.13)	75.83±0.64 (+5.56)

A.9 Sensitivity and Q Table Initialization

Figure 10 presents the estimated sensitivity values for all models considered in the CUB200 dataset and a subset of models used for CIFAR-100, as proposed by Salamah et al. (2024b); Zheng et al. (2023). These estimated sensitivity values are used to initialize the quantization table for JPEG-DL. Figure 11 illustrates the initial and converged values of the quantization table at the end of training.

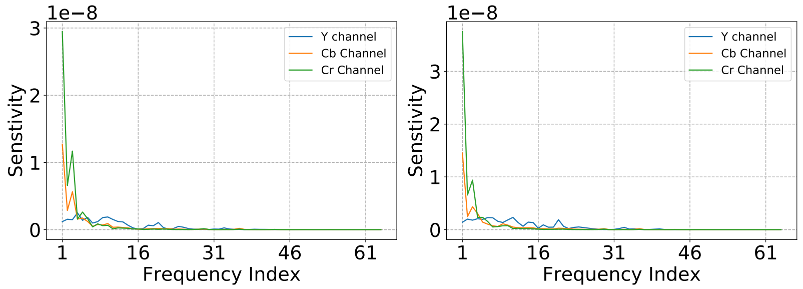
A.10 Feature maps visualization

Figure 12 presents the feature maps extracted after the first dense block in DenseNet-121 for both the JPEG-DL and the baseline models, trained on the Flowers dataset using the model from Table 2. The shown example was incorrectly classified by the baseline model, while the JPEG-DL model correctly classified it.

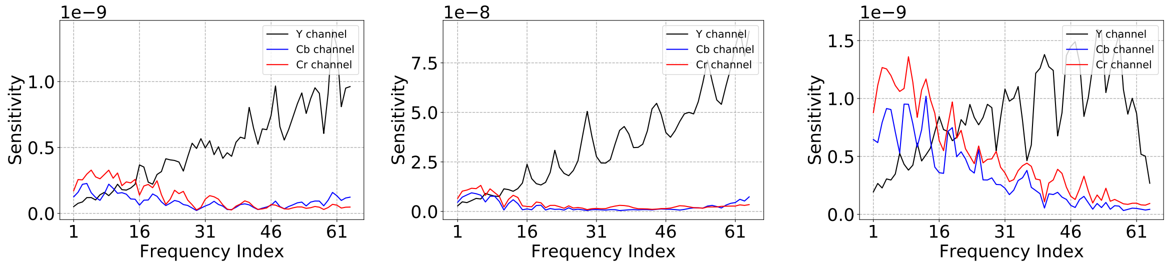
A.11 Mitigating the Inference Overhead of Differentiable Soft Quantizers

The reconstruction space $\hat{\mathcal{A}}$ is configured with a specific size, determined by the parameter L, which is set to 2^{b-1}, as mentioned in Section 4. For CIFAR-100 and Fine-grained datasets, b is set to 8 bits, resulting in a reconstruction space length of 513. For ImageNet-1K, b is set to 11 bits, resulting in a length of 2047. This high dimensionality of the reconstruction space poses a computational challenge during inference when calculating the conditional probability mass function (CPMF) $P_\alpha(\cdot|z)$. To address this, the support of each CPMF is restricted to the five closest points in $\hat{\mathcal{A}}$ to the coefficient z being quantized. This simplification, referred to as the masked CPMF, significantly reduces computational complexity without compromising the effectiveness of the quantization scheme $\mathcal{Q}_d$.

Figure 13 demonstrates the rapid decay of probability mass in CPMFs as the reconstruction level $\hat{z}$ deviates from the quantized value the uniform quantizer $\mathcal{Q}_u(z)$. The set $\mathcal{M}$, consisting of five reconstruction levels



(a) ResNet18 (b) DenseNet-121



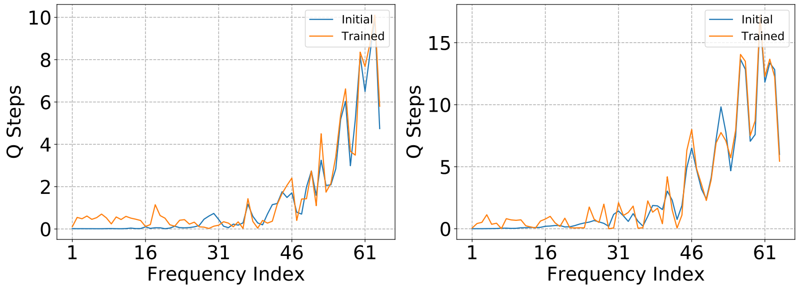
(c) VGG13 (d) Res56 (e) ShuffleNetV2

Figure 10: The estimated sensitivity is shown for two pre-trained models on the CUB200 (first row) and three pre-trained models on the CIFAR-100 (second row), with sensitivity indices arranged in the default zigzag order.

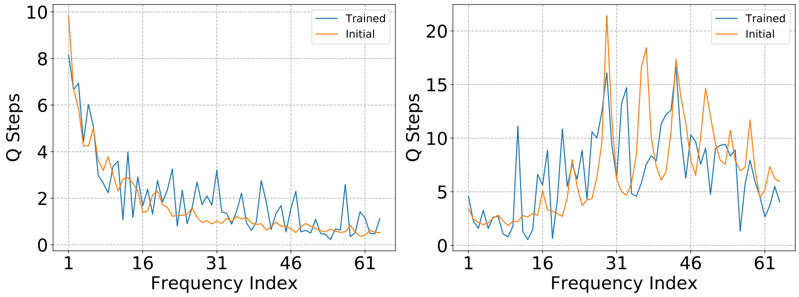
α	$\mathcal{Q}_d(z = 0.5)$		
	Full Space	Masked	Δ
1	0.5	0.5027	0.0027
3	0.5	0.5	5.96e-08
5	0.5	0.5	5.96e-08
10	0.5	0.5	0

Table 9: $\mathcal{Q}_d(z = 0.5; q = 1, \alpha)$ values resulting from the full-space CPMF and the masked CPMF at different α, alongside the absolute difference Δ between the full-space and masked cases.

centered around $\mathcal{Q}_u(z)$, captures a significant portion of the total probability mass, even in extreme cases where the quantization step q and the parameter α are very small. This observation suggests that restricting the CPMF support to $\mathcal{M}$ (masked CPMF) is a valid simplification. Instead of computing the full-space CPMF with length $2L + 1$, we can efficiently calculate the masked CPMF with length 5 using softmax on $\mathcal{M}$. The impact of this simplification on the quantization scheme $\mathcal{Q}_d$ is negligible, particularly for high values of α, as shown in Table 9. Therefore, using the masked CPMF in our implementation is a justified simplification that significantly reduces computational complexity without compromising accuracy.



(a) DenseNet-121 (Y Channel) (b) DenseNet-121 (CbCr Channel)



(c) Res56 (Y Channel) (d) Res56 (CbCr Channel)

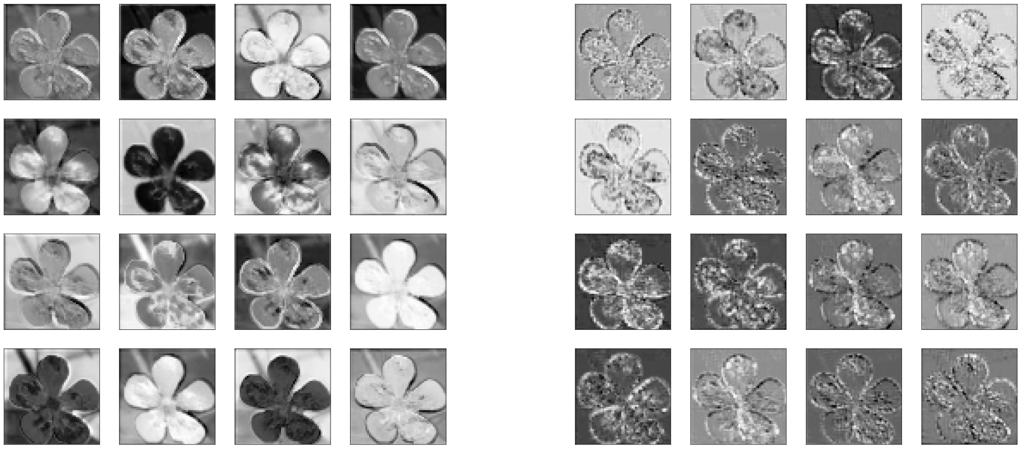
Figure 11: Initial and final quantization tables for Res56 trained on CIFAR-100 and DenseNet-121 trained on CUB200.

Specifically, Fig. 13 illustrates how fast the probability mass of a reconstruction level $\hat{z}$ decays in CPMFs as $\hat{z}$ deviates from the uniform quantizer $\mathcal{Q}_u(z)$. It's clear that the set $\mathcal{M} = \{\mathcal{Q}_u(z - 2q, \mathcal{Q}_u(z) - q, \mathcal{Q}_u(z), \mathcal{Q}_u(z) + q, \mathcal{Q}_u(z) + 2q\}$ of reconstruction levels retain a total probability mass close to 1, even in an adversarially chosen scenario where both q and α are extremely small[3] (i.e., 1) and the coefficient C being quantized lies exactly at a quantization threshold. Therefore, instead of computing the full-space CPMF with length $2L + 1$ following (6), we only conduct softmax on $\mathcal{M}$ to get the masked CPMF with length 5. The error in $\mathcal{Q}_d$ caused by this masking is negligible as shown in Table 9, especially in the high α case. Therefore, it's totally justified to simplify the full-space CPMF to the masked CPMF in our implementation.

Table 10 presents a comparison of inference time and throughput between the standard model and our unified model, which incorporates the JPEG layer with the underlying model. The inference times for both the full-space CPMF and the masked CPMF are measured. The reported inference time is averaged across the entire validation set of ImageNet using Resnet18 from Table 3. These results demonstrate the effectiveness of the proposed approach in addressing the computational complexity of the JPEG-DL framework.

---

[3] As q and α increase, the CPMF becomes even sharper, and as a result, $\mathcal{M}$ will capture increasingly higher probability mass.



(a) Baseline

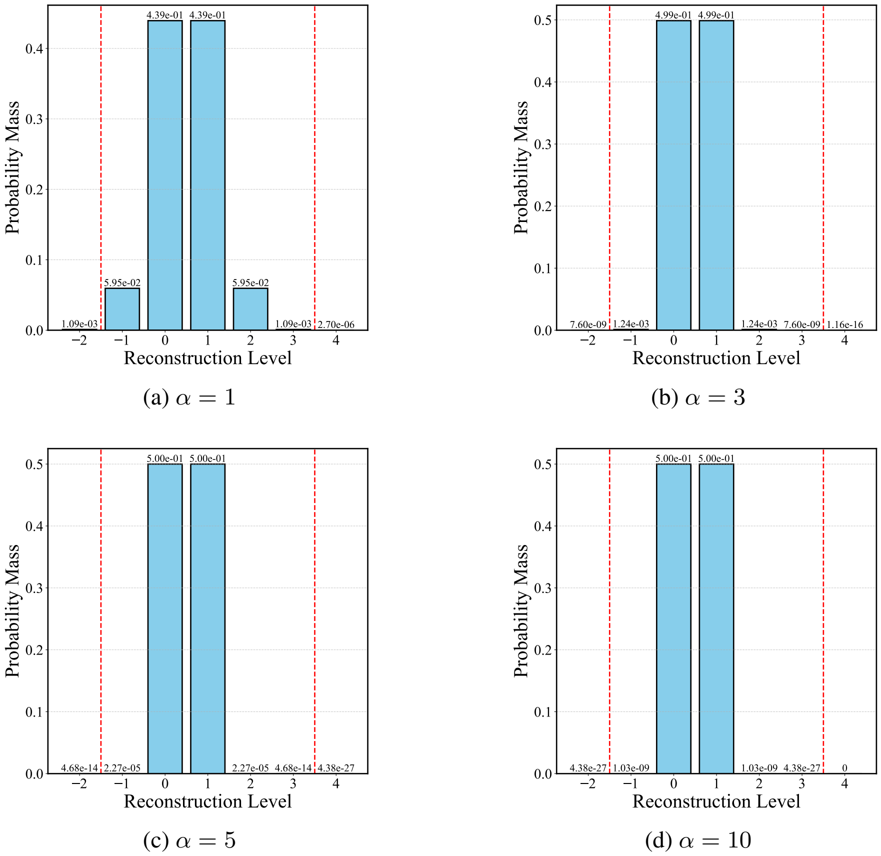
(b) JPEG-DL

(c) Input

Figure 12: Feature maps of size 56×56 are shown after the first dense block in DenseNet-121 for both JPEG-DL and baseline models in Figs. 12a and 12b, respectively, using an original input shown in Fig. 12c. The JPEG-DL model highlights the foreground (flower) more distinctly, while the baseline model shows less contrast, contributing to its misclassification.

Inference Settings	Inference time (milliseconds)	Inference Throughput (img/sec)
Baseline	5.46	180.65
JPEG-DL (Masked CPMFs)	7.43	133.79
JPEG-DL (Full Space)	18.86	52.37

Table 10: Comparison of inference time between different inference settings measured for Resnet18 on ImageNet, following the experimental setup mentioned in Section 4. Evaluation is conducted on NVIDIA RTX A5000 GPUs.



(a) $\alpha = 1$

(b) $\alpha = 3$

(c) $\alpha = 5$

(d) $\alpha = 10$

Figure 13: Partial visualization of CPMFs $P_\alpha(\cdot|z = 0.5)$ computed using (6), with $L = 1023$, $q = 1$, and α values of (a) 1, (b) 3, (c) 5, and (d) 10. The reconstruction levels between two red dashed lines represent $\mathcal{M}$.