JPEG COMPLIANT COMPRESSION FOR DNN VISION

Kaixiang Zheng, Ahmed H. Salamah*, Linfeng Ye*, and En-Hui Yang, Fellow, IEEE*

{k56zheng,ahamsalamah,l44ye,ehyang}@uwaterloo.ca

2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10221982

ABSTRACT

Conventional image compression techniques are mostly developed for the human visual system. However, with the extensive use of deep neural networks (DNNs), more and more images will be consumed by DNN-based intelligent machines, which makes it crucial to develop image compression techniques customized for DNN vision while being JPEG compliant. In this paper, we first propose a new distortion measure, dubbed the sensitivity weighted error (SWE). Then, we develop OptS, a DNN-oriented compression algorithm with full JPEG compatibility, which designs optimal quantization tables for DNN models based on SWE. To test the performance of our algorithm, experiments of image classification are conducted on the ImageNet dataset for two prevailing DNN models. Results demonstrate that our algorithm achieves better rate-accuracy (R-A) performance than the default JPEG. For some DNN model, the compression ratio of our algorithm can reach 8.3×[1], reducing the compression rate (bits per pixel, bpp) of the default JPEG by 57.4% with no accuracy loss. Our source code is available at https://github.com/zkxufo/OptS.git.

***Index Terms*—** Image compression, deep learning, JPEG, quantization table, distortion measure.

1. INTRODUCTION

Image compression is one of the fundamental domains in image processing and computer vision (CV), which has been well developed in the past decades. Among all image compression approaches, JPEG [1] has become the *defacto* one, considering its extensive application. JPEG is a lossy compression technique developed for the human visual system, dubbed human-oriented compression (HOC) in this paper. HOC algorithms can aggressively compress an image as they allow information loss during compression which is imperceptible for humans.

With the success of deep learning, a dramatic change is now happening in terms of how images are consumed. In the last decade, numerous deep neural networks (DNNs) have been proposed and used for CV tasks, such as image classification [2, 3], semantic segmentation [4, 5], object detection [6, 7], image generation [8, 9], etc. More and more images have been and will continue to be viewed by DNN-based intelligent machines, where humans would step in only when these intelligent machines fail their vision tasks. However, DNN perception differs in general from human perception [10]. The information loss incurred in the process of HOC may not be imperceptible for DNNs anymore, resulting in degraded DNN performance [11].

To aggressively compress images without sacrificing DNN's performance, DNN-oriented compression (DOC) techniques are of urgent need. So far, there are only a few of research works addressing the DOC problem [12], [10], [13]. These existing approaches, however, either do not provide a better rate-accuracy (R-A) tradeoff compared to the default JPEG, or are computationally expensive.

In this paper, we first propose a new distortion measure dubbed the sensitivity weighted error (SWE). Given a DNN, the SWE distortion measure is customized for the vision of that DNN, based on the sensitivity of the training loss function of that DNN with respect to perturbations in the discrete cosine transform (DCT) domain. With the SWE distortion measure, any DCT-based HOC algorithm including JPEG can be converted to a DOC algorithm without additional complexity. Based on SWE, we then develop OptS, a DOC algorithm with full JPEG compatibility, which designs optimal quantization tables for JPEG in conjunction with SWE. It is shown experimentally that when evaluated on the ImageNet validation set [14], our proposed DOC algorithm can improve the classification accuracy of tested popular DNNs by as much as 0.93% over the default JPEG at the same compression rate in bits per pixel (bpp), or reduce the compression rate of the default JPEG by as much as 57.4% at the same accuracy. Moreover, if we tolerate some accuracy loss up to 0.47%, then our compression ratio can even reach 13.3×, reducing the compression rate of the default JPEG by 73.5%.

The rest of the paper is organized as follows. We formulate the problem in Section 2 and propose our method in Section 3, with experimental results shown in Section 4.

2. PROBLEM FORMULATION

In JPEG, an encoder first partitions an image x into B non-overlapping 8×8 blocks, and then applies DCT to each of

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN203035-22, and by the Canada Research Chairs Program. All authors are affiliated to the Department of Electrical and Computer Engineering, University of Waterloo.

*These authors contributed equally to this work.

[1]The compression ratio equals 24 divided by the compression rate.

these non-overlapping blocks to obtain DCT coefficients C. After flattening the DCT coefficients in each block in the zigzag order, we get $M = 64$ sequences $\{C_i\}_{i=1}^{M}$ of DCT coefficients, each with length B, where each sequence $C_i = \{C_{i,j}\}_{j=1}^{B}$ corresponds to a distinct DCT frequency position $1 \le i \le M$. We further model these sequences as independent sources.

After DCT, the next step of JPEG is to perform quantization. Define a quantization table $Q = \{q_1, q_2, \ldots, q_M\}$, where q_i is the quantization step size for C_i. Given Q, we can quantize DCT coefficients $C_{i,j}$ into DCT indices $K_{i,j}$ by $K_{i,j} = \lfloor C_{i,j}/q_i \rceil$. This step is also referred to as hard decision quantization (HDQ). Note that $K_i = \{K_{i,j}\}_{j=1}^{B}$ is the sequence of DCT indices corresponding to the ith frequency position. As is well known, quantization is the step where distortion is introduced, which makes JPEG a lossy compression technique. Finally, a lossless coding method is utilized to encode the DCT indices.

With the adoption of HDQ, JPEG compression can be formulated as the following optimization problem

$$\inf_{Q} \; R(Q), \;\; \text{s.t.} \;\; D(Q) \le D_T, \tag{1}$$

where $R(Q)$ denotes the number of bits per block resulting from quantization and lossless coding, $D(Q)$ is the distortion per block introduced by the quantization process, and D_T is the prescribed block-wise distortion budget. In this problem, we are optimizing over Q, because both R and D are solely dependent on Q given a fixed lossless coding method. Therefore, the JPEG compliant image compression problem can be formulated as a quantization table designing problem.

Decompose $D(Q)$ over different sources into $D(Q) = \sum_{i=1}^{M} D(C_i, q_i)$, where $D(C_i, q_i)$ is the average distortion of C_i with quantization step size q_i. Then, we can rewrite (1) into an equivalent problem

$$\inf_{\substack{Q: D(C_i,q_i) \le D_i, \forall i \\ \sum_{i=1}^{M} D_i = D_T}} R(Q), \tag{2}$$

where D_i is the portion of distortion budget allocated to the i^{th} source. Consider the inequality constraints in (2): $D(C_i, q_i) \le D_i, 1 \le i \le M$. Given an image x and a specific distortion measure, C_i and $D(\cdot,\cdot)$ are fixed, so the selection of q_i is determined by D_i. Therefore, designing an optimal quantization table Q^* is equivalent to finding an optimal distortion allocation $\{D_i^*\}_{i=1}^{M}$.

For HOC, Problem (2) has already been solved in [15] by Yang *et al.* In this paper, we focus on solving (2) for DOC, by first proposing a new distortion measure $D(\cdot,\cdot)$ customized for DNN vision.

3. METHODOLOGY

3.1. DNN's Sensitivity to DCT Perturbations

Following the previous formulation, we can see that distortion is of pivot significance in our problem. Therefore, a natural question is how to measure the distortion in the case of DOC. Conventionally, in the case of HOC, mean squared error (MSE) is widely used to measure the distortion between the original DCT coefficients $C_{i,j}$ and quantized DCT coefficients $q_i K_{i,j}$: $D(C_i, q_i) = \frac{1}{B}\sum_{j=1}^{B}(C_{i,j} - q_i K_{i,j})^2$ for all $1 \le i \le M$. Note that in MSE, all DCT frequencies are treated equally. For DNN vision, however, it was observed in [10] that DNN's sensitivity varies across DCT frequencies. This motivates us to differentiate the distortions resulting from the quantization of different sources C_i. To this end, let us formally define DNN's sensitivity to DCT perturbations.

Given a DNN, we denote its loss function used in its training stage as $L(\cdot)$. Suppose ΔC amount of perturbation is added to the DCT coefficients C of an image. We want to understand how sensitive the loss function $L(\cdot)$ of the DNN is with respect to the perturbation ΔC. To simplify our discussion, the derivation below is limited to a single channel of an image, but can be easily extended to the multiple channels of the image. By Taylor expansion, we have

$$L(C + \Delta C) = L(C) + [\nabla L(C)]^T \Delta C + o(||\Delta C||) \tag{3}$$

when ΔC is small enough, where $\nabla L(C)$ is the gradient vector of $L(\cdot)$ with respect to C, and $||\Delta C||$ denotes the l_2 norm of ΔC. Therefore, the loss increase caused by the perturbation is upper bounded by

$$\begin{aligned}
|\Delta L| &= |L(C + \Delta C) - L(C)| \le |[\nabla L(C)]^T \Delta C| + |o(||\Delta C||)| \\
&= \left| \sum_{i=1}^{M} \frac{\mathrm{d}L}{\mathrm{d}C_i^T} \Delta C_i \right| + |o(||\Delta C||)| \\
&\le \sum_{i=1}^{M} \left| \frac{\mathrm{d}L}{\mathrm{d}C_i^T} \Delta C_i \right| + |o(||\Delta C||)| \\
&\le \sum_{i=1}^{M} \sqrt{\sum_{j=1}^{B} \left(\frac{\partial L}{\partial C_{i,j}} \right)^2 \cdot \sum_{j=1}^{B} \Delta C_{i,j}^2} + |o(||\Delta C||)|
\end{aligned} \tag{4}$$

where the last inequality follows from the Cauchy–Schwarz inequality. Squaring both sides of (4), we have

$$\begin{aligned}
\Delta L^2 &\le \left(\sum_{i=1}^{M} \sqrt{\sum_{j=1}^{B} \left(\frac{\partial L}{\partial C_{i,j}} \right)^2 \cdot \sum_{j=1}^{B} \Delta C_{i,j}^2} \right)^2 + |o(||\Delta C||^2)| \\
&\le M \sum_{i=1}^{M} \sum_{j=1}^{B} \left(\frac{\partial L}{\partial C_{i,j}} \right)^2 \cdot \sum_{j=1}^{B} \Delta C_{i,j}^2 + o(||\Delta C||^2)
\end{aligned} \tag{5}$$

Note that (5) is valid for any small perturbation ΔC. When the perturbation ΔC is limited only to the DCT frequency i, (5) then becomes

$$\Delta L^2 \le \sum_{j=1}^{B} \left(\frac{\partial L}{\partial C_{i,j}} \right)^2 \cdot ||\Delta C_i||^2 + o(||\Delta C_i||^2) \tag{6}$$

where $||\Delta C_i||^2 = \sum_{j=1}^{B} \Delta C_{i,j}^2$. Therefore, the squared rate of change of $L(\cdot)$ with respect to C_i is upper bounded by

$$\frac{\Delta L^2}{||\Delta C_i||^2} \le \sum_{j=1}^{B} \left(\frac{\partial L}{\partial C_{i,j}} \right)^2 + o(1) \tag{7}$$

In view of (7), we now define the given DNN's sensitivity to the perturbation at the i^{th} frequency position as

$$s_i = \sum_{j=1}^{B} \left(\frac{\partial L}{\partial C_{i,j}} \right)^2 = \left|\left| \frac{\mathrm{d}L}{\mathrm{d}C_i} \right|\right|^2, \ 1 \le i \le M \tag{8}$$

Across all frequencies, DNN's sensitivity can be characterized by the set $S = \{s_1, s_2, \ldots, s_M\}$. Note that the computation of such defined DNN's sensitivity $S = \{s_1, s_2, \ldots, s_M\}$ depends on the input image C (in the DCT domain), the DNN, and the ground truth label of the input image C, which may not be available at the encoder.

3.2. Offline Estimation of the Sensitivity

To remove the dependency on the availability of the DNN and the ground truth label of the input image C at the time of encoding, we estimate the sensitivity S for each target DNN offline.

Specifically, for each target DNN, we first randomly select N image samples, and then estimate S by taking the sample mean of the sensitivity over the N image samples:

$$s_i = \frac{1}{N} \sum_{k=1}^{N} \sum_{j=1}^{B} \left(\frac{\partial L}{\partial C_{i,j}^k} \right)^2, \ 1 \le i \le M \tag{9}$$

where C^k is the DCT coefficients of the k^{th} input image x^k. The partial derivatives in (9) are obtained through backpropagation.

3.3. Sensitivity Weighted Error (SWE)

Go back to (5). Note that $M = 64$ is fixed. With the sensitivity S estimated via (9), we can now measure the distortions between the original DCT coefficients $C_{i,j}$ and quantized DCT coefficients $q_i K_{i,j}$ for each source C_i and for the whole image respectively as follows:

$$D(C_i, q_i) = \frac{1}{B} \sum_{j=1}^{B} s_i (C_{i,j} - q_i K_{i,j})^2, \ 1 \le i \le M \tag{10}$$

$$D(Q) = \frac{1}{B} \sum_{i=1}^{M} \sum_{j=1}^{B} s_i (C_{i,j} - q_i K_{i,j})^2. \tag{11}$$

The above distortion measure is called the sensitivity weighted error (SWE). In view of our derivation in Subsection 3.1, the SWE for an image is a good estimation of the upper bound of the squared DNN loss increase caused by quantization. Minimizing the SWE subject to a rate constraint will in turn reduce the squared DNN loss increase.

3.4. Optimal Allocation of SWE (OptS)

Based on the proposed SWE, we now turn back to solve the distortion allocation problem formulated in Section 2. Actually, to solve problem (2) for DOC, one can simply embed SWE into any existing algorithm for HOC as a substitute for the MSE, since SWE is a generic distortion measure independent of the algorithm. In this paper, we select the algorithm proposed in [15], named OptD, as the HOC baseline; and build the DOC counterpart, named OptS, upon it.

Algorithm 1 OptS for the Luminance Channel

Input: x, $S = \{s_1, s_2, \ldots, s_M\}$, d, q_{max}
Output: $Q = \{q_1, q_2, \ldots, q_M\}$
  if $s_i \sigma_i^2 < d$ **then**
    set $q_i = q_{max}$
  else
    if $i = 1$ **then**
      set $q_i = min\left\{ \left\lfloor \sqrt{\frac{12d}{s_i}} \right\rfloor, q_{max} \right\}$
    else
      set $q_i = max\{q_i \in \mathcal{Q} : D_{Lap}(\lambda_i, q_i) \le \frac{d}{s_i}\}$
    end if
  end if

In OptD, authors first model DC and AC coefficients with uniform and Laplacian distributions, so that $D(C_1, q_1) = q_1^2/12$ and $D(C_i, q_i)$, $2 \le i \le M$, can be approximated by $D_{Lap}(\lambda_i, q_i)$ calculated in (14). Then, a parameter d named water level is determined given the distortion budget D_T. With d and variances of sources $\{\sigma_i^2\}_{i=1}^M$, one can obtain the optimal distortion allocation $\{D_i^*\}_{i=1}^M$ which leads to the optimal quantization table Q^*.

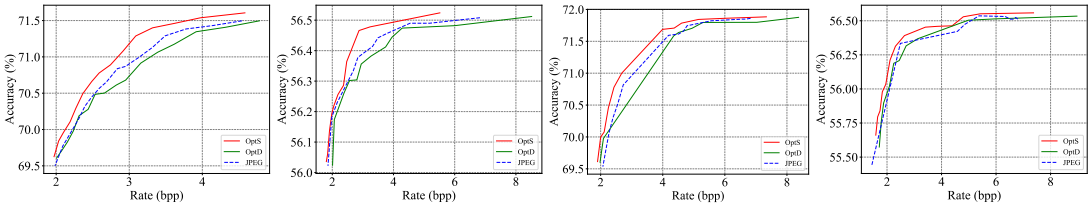
$$\lambda_i = \frac{1}{B} \sum_{j=1}^{B} |C_{i,j}|, \ 2 \le i \le M \tag{12}$$

$$z_i = q_i - \lambda_i + \frac{q_i}{e^{q_i/\lambda_i} - 1} \tag{13}$$

$$D_{Lap}(\lambda_i, q_i) = 2\lambda_i^2 - \frac{2q_i(\lambda_i + z_i - 0.5q_i)}{e^{z_i/\lambda_i}(1 - e^{-q_i/\lambda_i})} \tag{14}$$

However, OptD in [15] is proposed for grayscale images only, while most DNN models accept color images as inputs. Therefore, we have to extend OptD to accommodate color images before updating it to OptS. Conventionally, color images are converted to the YCbCr format before compression. In fact, the compression of the luminance channel (Y) can be directly handled by the original OptD. However, one cannot apply the original OptD to chrominance channels (Cb and Cr) respectively to get two quantization tables, because chrominance channels share the same quantization table in the JPEG framework. Following the same principle of designing the single-channel OptD, we manage to develop a two-channel OptD for chrominance channels. Note that this extension is highly nontrivial because of the aforementioned restriction.

Now that we have the complete version of OptD, we can update it to OptS by replacing MSE with SWE. The resulting algorithms for luminance and chrominance channels are demonstrated in Alg. 1 and Alg. 2, where $\mathcal{Q} = \{1, 2, \ldots, q_{max}\}$, and q_{max} is a predetermined maximum quantization step size. Note that the sensitivity used in Alg. 2 has $2M$ entries as it includes the sensitivity of both Cb and Cr channels. Also, some notations are overloaded in two algorithms for simplicity.



(a) MobileNetV2 (b) AlexNet (c) MobileNetV2 (d) AlexNet

Fig. 1: Evaluation of the R-A performance. Fig. 1a and 1b correspond to Setting 1; Fig. 1c and 1d correspond to Setting 2. It is worth mentioning that for MobileNetV2 and AlexNet, DeepN-JPEG yields 68.43% and 52.07% of accuracy with the rate of 7.9 bpp (its accuracy is low). Therefore, to avoid distortion, these points are removed from the figures.

Algorithm 2 OptS for Chrominance Channels

Input: x, $S = \{s_1, \ldots, s_M, s_{M+1}, \ldots, s_{2M}\}$, d, q_{max}
Output: $Q = \{q_1, q_2, \ldots, q_M\}$
  if $max\{s_i\sigma_i^2, s_{i+M}\sigma_{i+M}^2\} < d$ **then**
    set $q_i = q_{max}$
  else if $s_i\sigma_i^2 < d \leq s_{i+M}\sigma_{i+M}^2$ **then**
    if $i = 1$ **then**
      set $q_i = min\Big\{\Big\lfloor\sqrt{\frac{12d}{s_{i+M}}}\Big\rfloor, q_{max}\Big\}$
    else
      set $q_i = max\{q_i \in \mathcal{Q} : D_{Lap}(\lambda_{i+M}, q_i) \leq \frac{d}{s_{i+M}}\}$
    end if
  else if $s_{i+M}\sigma_{i+M}^2 < d \leq s_i\sigma_i^2$ **then**
    if $i = 1$ **then**
      set $q_i = min\Big\{\Big\lfloor\sqrt{\frac{12d}{s_i}}\Big\rfloor, q_{max}\Big\}$
    else
      set $q_i = max\{q_i \in \mathcal{Q} : D_{Lap}(\lambda_i, q_i) \leq \frac{d}{s_i}\}$
    end if
  else if $d \leq min\{s_i\sigma_i^2, s_{i+M}\sigma_{i+M}^2\}$ **then**
    if $i = 1$ **then**
      set $q_i = min\Big\{\Big\lfloor\sqrt{\frac{12d}{max\{s_i, s_{i+M}\}}}\Big\rfloor, q_{max}\Big\}$
    else
      set $q_i = max\Big\{q_i \in \mathcal{Q} : D_{Lap}(\lambda_i, q_i) \leq \frac{d}{s_i}, D_{Lap}(\lambda_{i+M}, q_i) \leq \frac{d}{s_{i+M}}\Big\}$
    end if
  end if

4. EXPERIMENTAL RESULTS

To evaluate the performance of OptS, we select image classification as our CV task. Accordingly, the loss function L represents the cross entropy loss. To estimate the sensitivity of tested DNN models following (9), we sample 10K images from the ImageNet ILSVRC 2012 training set [14] covering all classes. Note that all images are scaled to the size of 224 $\times$ 224 following the standard preprocessing.

Instead of the rate-distortion (R-D) performance in the HOC case, we evaluate the rate-accuracy (R-A) performance based on the ImageNet validation set [14]. Here, rate is measured by the average bpp of an image, and accuracy is the top-1 validation accuracy. For comparison, we select the default JPEG as our benchmark. Also, the performance of OptD is presented as an ablation study, in order to show the contribution of SWE alone. Two prevailing models, MobileNetV2 [16] and AlexNet [17], are adopted as our target DNNs. In all experiments, q_{max} is selected to be 100.

For JPEG, the quality of a compressed image is controlled by the quality factor (QF), while the counterpart in OptS or OptD is the water level d. A straightforward experimental design is to use fix QF and d's for three algorithms over the whole dataset. However, in this design, we have no control over the image-wise distortion, and a lot of tuning is inevitably involved. So, we design our experiments in a more sophisticated way to foster fair comparison without tuning.

Setting 1: Use a fixed QF for JPEG to compress all the images in the dataset, and record all the resulting SWEs for them. Then, for each image, adjust the image-adaptive d's in OptS and OptD to roughly match the recorded SWE using binary search.

Setting 2: Use a fixed d for OptS to compress all the images in the dataset, and record all the resulting SWEs for them. Then, for each image, adjust the image-adaptive QF in JPEG and d in OptD to do the distortion matching again.

For Setting 1, results are shown in Fig. 1a and 1b, with QF $\in [70, 98]$; for Setting 2, results are shown in Fig. 1c and 1d, with $d \in [0.005, 1]$. Some of these results are also shown in Table 1. As expected, OptS always performs better than OptD and JPEG in both settings. In addition, we also compare with an existing DOC algorithm called DeepN-JPEG [12], whose performance is mentioned in the caption of Fig. 1, yet we don't compare with another DOC algorithm dubbed GRACE [10] as it's not JPEG compliant.

Model	JPEG		OptS	
	Rate (bpp)	Acc (%)	Rate (bpp)	Acc (%)
MobileNetV2	2.1	69.53	2.2	70.46
AlexNet	6.8	56.51	2.9	56.47
AlexNet	6.8	56.51	1.8	56.04

Table 1: Comparison between JPEG and OptS.

Interestingly, our experiments also show that for small d, compression with OptS actually yields the accuracy slightly better than the inference accuracy of the raw dataset, confirming the prediction made in [18] in practice. Specifically, compared to the raw dataset without compression, OptS improves the original accuracy of MobileNetV2 (71.878%) by 0.054% with 4.2$\times$ compression ratio, and that of AlexNet (56.522%) by 0.028% with 4.0$\times$ compression ratio.

5. REFERENCES

[1] William B Pennebaker and Joan L Mitchell, *JPEG: Still image data compression standard*, Springer Science & Business Media, 1992.

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

[3] 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*, 2016, pp. 770–778.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[7] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[8] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[10] 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*, 2019, pp. 1–16.

[11] Neelanjan Bhowmik, Jack W Barker, Yona Falinie A Gaus, and Toby P Breckon, "Lost in compression: the impact of lossy image compression on variable size object detection within infrared imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 369–378.

[12] 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*, 2018, pp. 1–6.

[13] Hongshan Li, Yu Guo, Zhi Wang, Shutao Xia, and Wenwu Zhu, "Adacompress: Adaptive compression for online computer vision services," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2440–2448.

[14] 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*. Ieee, 2009, pp. 248–255.

[15] En-Hui Yang, Chang Sun, and Jin Meng, "Quantization table design revisited for image/video coding," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4799–4811, 2014.

[16] 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*, 2018, pp. 4510–4520.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] En-hui Yang, Hossan Amer, and Yanbing Jiang, "Compression helps deep learning in image classification," *Entropy (https://doi.org/10.3390/e23070881)*, 2021.