End-to-end optimized image compression with competition of prior distributions

Benoit Brummer
intoPIX
Mont-Saint-Guibert, Belgium
b.brummer@intopix.com

Christophe De Vleeschouwer
Université catholique de Louvain
Louvain-la-Neuve, Belgium
christophe.devleeschouwer@uclouvain.be

Abstract

Convolutional autoencoders are now at the forefront of image compression research. To improve their entropy coding, encoder output is typically analyzed with a second autoencoder to generate per-variable parametrized prior probability distributions. We instead propose a compression scheme that uses a single convolutional autoencoder and multiple learned prior distributions working as a competition of experts. Trained prior distributions are stored in a static table of cumulative distribution functions. During inference, this table is used by an entropy coder as a look-up-table to determine the best prior for each spatial location. Our method offers rate-distortion performance comparable to that obtained with a predicted parametrized prior with only a fraction of its entropy coding and decoding complexity.

1. Introduction

Image compression typically consists of a transformation step (including quantization) and an entropy coding step that attempts to capture the probability distribution of a transformed context to generate a smaller compressed bitstream. Entropy coding ranges in complexity from simple non-adaptive encoders [26, 24] to complex arithmetic coders with adaptive context models [15, 23]. The entropy coding strategy has been revised to address the specificities of learned compression. More specifically, for recent works that make use of a convolutional autoencoder [12] (AE) as the all-inclusive transformation and quantization step, the entropy coder relies on a cumulative probability model (CPM) trained alongside the AE [5]. This model estimates the cumulative distribution function (CDF) of each channel coming out of the AE and passes these learned CDFs to an entropy coder such as range encoding [16].

Such a simple method outperforms traditional codecs like JPEG2000 but work is still needed to surpass complex codecs like BPG. Johannes Ballé et al. (2018) [6] proposed analyzing the output of the convolutional encoder with another AE to generate a floating-point scale parameter that differs for every variable that needs to be encoded by the entropy coder, thus for every location in every channel. This method has been widely used in subsequent works but introduces substantial complexity in the entropy coding step because a different CDF is needed to encode every variable in the latent representation of the image, whereas the single AE method by Ballé et al. (2017) [5] reused the same CDF table for every latent spatial location.

Our work uses the principle of competition of experts [22, 14] to get the best out of both worlds. Multiple prior distributions compete for the lowest bit cost on every spatial location in the quantized latent representation. During training, only the best prior distribution is updated in each spatial location, further improving the prior distributions specialization. CDF tables are fixed at the end of training. Hence, at testing, the CDF table resulting in the lowest bitcost is assigned to each spatial location of the latent representation. The rate-distortion (RD) performance obtained is comparable to that obtained with a parametrized distribution [6], yet the entropy coding process is greatly simplified since it does not require a per-variable CDF and can build on look-up-tables (LUT) rather than the computation of analytical distributions.

2. Background

Entropy coders such as range encoding [16] require ***cdf***s where, for each variable to be encoded, the probability that a smaller or equal value appears is defined for every allowable value in the latent representation space. Johannes Ballé et al.'s seminal work (2017) [5] consists of an AE, computing a latent image representation consisting in C_L channels of size $H_L \times W_L$, and a CPM, consisting of one CDF per latent output channel, which are trained conjointly. The latent representation coming out of the encoder is quantized then passed through the CPM. The CPM defines, in a parametrized and differentiable manner, a CDF per channel. At the end of training, the CPM is evaluated at every possible value[1] to generate the static CDF table. The CDF table is not differentiable, but going from a differentiable CPM to a static CDF table speeds up the encoding and decoding process. The

CDF table is used to compress latent representations with an entropy coder, the approximate bit cost of a symbol is the binary logarithm of its probability.

Ballé et al. (2018) improved the RD efficiency by replacing the unique CDF table with a Gaussian distribution parametrized with a hyperprior (HP) sub-network [6]. The HP generates a scale parameter, and in turn a different CDF, for every variable to be encoded. Thus, complexity is added by exploiting the parametrized Gaussian prior during the entropy coding process, since a different CDF is required for each variable in the channel and spatial dimensions.

Minnen et al. proposed a scheme where one of multiple probability distributions is chosen to adapt the entropy model locally [21]. However, these distributions are defined a posteriori, given the encoder trained with a global entropy model. Thus [21] does not perform as well as the HP scheme [6] per [19, Fig. 2a]. In contrast, the present method jointly optimizes the local entropy models and the AE in an end-to-end fashion that results in greater performance. Minnen et al. [19] later proposed to improve RD with the use of an autoregressive sequential context model. However, as highlighted in [13], this is obtained at the cost of increased runtime by several orders of magnitude. Subsequent works have attempted to reduce complexity of the neural network architecture [10] and to bridge the RD gap with Minnen's work [13], but entropy coding complexity has remained largely unaddressed and has instead evolved towards increased complexity [19, 7, 20] compared to [6]. The present work builds on Ballé et al. (2017) [5] and achieves the performance of Ballé et al. (2018) [6] without the complexity introduced by a per-variable parametrized probability distribution. We chose Ballé et al. (2017) as a baseline because it corresponds to the basic unit adopted as a common reference and starting point for most models proposed in the recent literature to improve compression quality [6, 19, 13, 20]. Due to its generic nature, our contribution remains relevant for the newer, often computationally more complex, incremental improvements on Ballé et al. (2017).

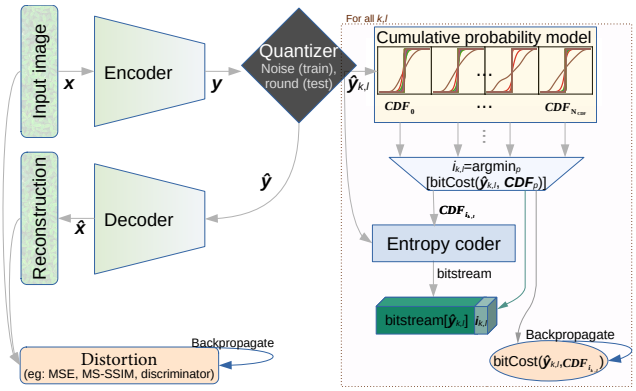

Figure 1. AE compression scheme with competition of prior distributions. The AE architecture is detailed in [6, Fig. 4]. The indices $\boldsymbol{i}$ denote the indices of CDF tables that minimizes the bitcount for each latent spatial dimension. Loss = Distortion + λ bitCost.

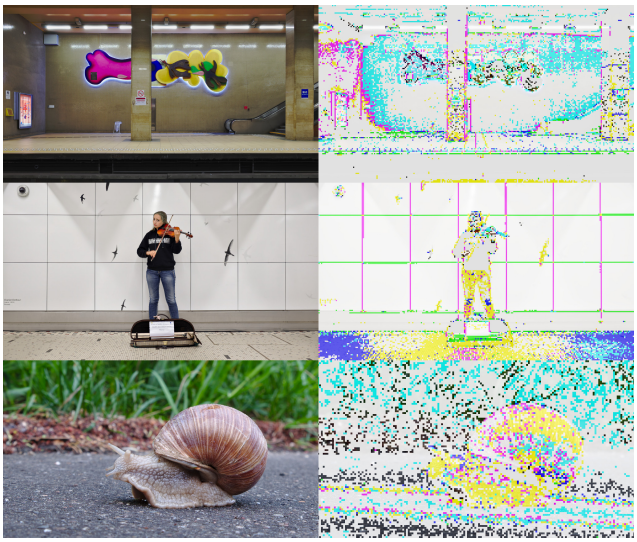

Figure 2. Segmentation of three test images [1]: each distinct color represents one of 64 CDF tables used to encode a latent spatial location (16×16 pixels patch)

3. Competition of prior distributions

Our proposed method introduces competitions of expert [22, 14] prior distributions: a single AE transforms the image and a set of prior distributions are trained to model the ***CDF*** of the latent representation in each spatial location. For each latent spatial dimension the CDF table which minimizes bit cost is selected; that prior is either further optimized on the features it won in the training mode, or its index is stored for decoding in the inference mode. This scheme is illustrated in Figure 1, and three sample images are segmented by "winning" CDF table in Figure 2.

All prior distributions are estimated in parallel by considering N_{CDF} CDF tables, and selecting, as a function of the encoded latent spatial location, the one that minimizes the entropy coder bitcount. The CDF table index is determined for each spatial location by evaluating each CDF table in inference. This can be done in a vectorized operation given sufficient memory. During training the CPM is evaluated instead of CDF tables such that the probabilities are up to date and the model is differentiable, and the bit cost is returned as it contributes to the loss function. The cost of CDF table indices has been shown to be neglectable due to the reasonably small number of priors, which in turns results from the fact that little gain in latent code entropy has been obtained by increasing the number of priors.

In all our experiments , the AE architecture follows the one in Ballé et al. (2018) [6], without the HP, since we found that the AE from [6] offers better RD than the one described in Ballé et al. (2017) [5], even with a single CDF table. A functional training loop is described in Algorithm 1.

Algorithm 1 Training loop

$\boldsymbol{y} \leftarrow$ model.Encoder($\boldsymbol{x}$)
$\hat{\boldsymbol{y}} \leftarrow$ quantize($\boldsymbol{y}$)
$\hat{\boldsymbol{x}} \leftarrow$ clip(model.Decoder($\hat{\boldsymbol{y}}$), 0, 1)
$distortion \leftarrow$ visualLossFunction($\hat{\boldsymbol{x}}, \boldsymbol{x}$)
for $0 \leq k < \mathrm{H_L}$ and $0 \leq l < \mathrm{W_L}$ **do**
  $\boldsymbol{bitCost}\,[k,l] \leftarrow \min_{i<\mathrm{N_{CDF}}}$
    $\left| -\log_2 \left(\boldsymbol{CPM}_i(\hat{\boldsymbol{y}}[k,l] + 0.5) - \boldsymbol{CPM}_i(\hat{\boldsymbol{y}}[k,l] - 0.5) \right) \right|$
end for ▷ ***CPM*** is the differentiable version of ***CDF***
$Loss \leftarrow distortion \times \lambda + |\boldsymbol{bitCost}|$
$Loss$.backward()

4. Experiments

4.1. Method

These experiments are based on the PyTorch implementation of Ballé et al. (2018) [6] published by Liu Jiaheng [9, 13]. To implement our proposed method, the HP is omitted in favor of competition of expert prior distributions. The CPM is that defined in [9] with an additional $\mathrm{N_{CDF}}$ dimension to compute all CDF tables in parallel. Theoretical results are verified using the torchac range coder [18, 17, 16]. A functional training loop is described in Algorithm 1, and source code is provided on https://github.com/trougnouf/Manypriors. To ensure that all priors get an opportunity to train, the prior distributions that have not been used for at least fifty steps are randomly assigned to spatial locations with largest bitcounts, to be forced to train. The Adam optimizer [11] is used with a starting learning rate (LR) of 0.0001 for the AE and 0.001 for the CPM. Performance is tested every 2500 steps in inference mode on the validation set, and the LR is decayed by a factor of 0.99 if the performance have not improved for two tests. Reported performance is the one of the model taht minimizes $(\text{visualLoss} \times \lambda + \text{bitCost})$ on the validation set at the end of training. Base models are trained for six million steps at $\lambda = 4096$ with the mean squared error (MSE) loss. Smaller λ values and MS-SSIM models are trained for four million steps starting from the base model with their LR and optimizer reset. All models use $\mathrm{C_H} = 192$ (hidden layers channels) and $\mathrm{C_L} = 256$ (output channels) such that a single base model is needed for each prior configuration. The training and validation dataset is made of free-license images from Wikimedia Commons [3]; mainly "Category:Featured pictures on Wikimedia Common" which consists of 13928 images of the highest quality. The images are cropped into 1024^2 pixels patches on disk to speed up further resizing, then they are resized on-the-fly by a random factor down to 256^2 pixels during training. A batch size of 4 patches is used. The kodak set [2] is used as a validation set and the CLIC professional test dataset [4] is used for testing.

The RD curve of our "multiprior" model is compared with that of the HP model [6], which is trained from scratch using Liu Jiaheng's PyTorch implementation [9, 13]. Liu Jiaheng's code differs slightly from the paper's definition [6] in that a Laplace distribution is used in place of the normal distribution to stabilize training. Complexity is measured as the number of GMac (billion multiply-accumulate operation) using the ptflops counter [25] and the number of memory lookup operations is calculated manually.

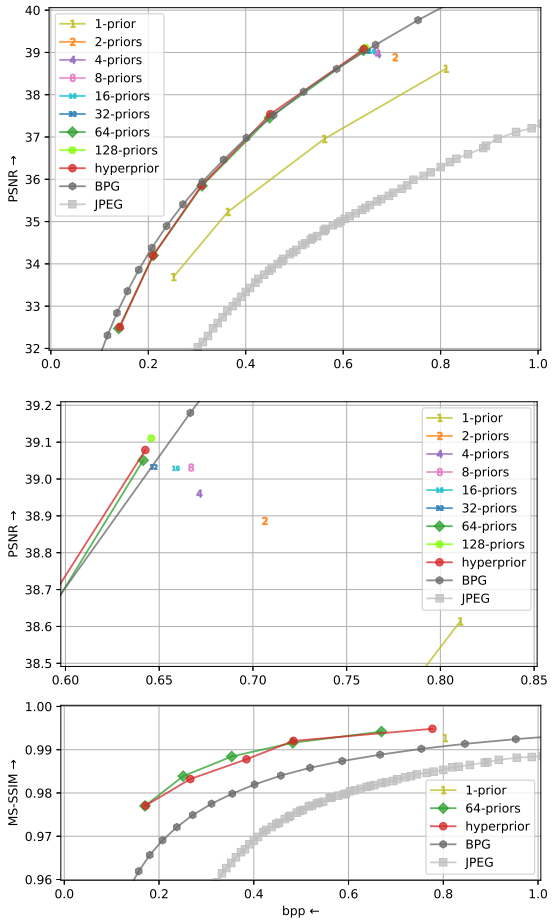

Figure 3. Top: PSNR RD curve of a 64-priors model on the CLIC pro. test set, compared with the HP model [6], and the BPG and JPEG codecs. Middle : Zoom in on models with 1, 2, 4, 8, 16, 32, and 64 priors. Bottom: MS-SSIM RD curve.

4.2. Results

The PSNR RD curve measured on the CLIC professional test set [4] is shown on top of Figure 3. The performance of a 64-priors model is in line with that of the HP model : they both perform slightly better than BPG at high bpp, and achieve significantly better RD than the single-prior model. In the middle, the RD value at $\lambda = 4096$, the highest bitrate, is shown for 1, 2, 4, 8, 16, 32, 64, and 128 prior distributions. 128-priors offer marginal gains and costs an increased training time (1.5) and encoding time. MS-SSIM performance of fine-tuned models is shown in the bottom of Figure 3; the 64-priors model still performs similarly to [6],

and both learned compression models benefit from this more perceptual metric compared with traditional codecs.

Computational complexity of our Manypriors has been compared to the one of the HP model [6]). This complexity is expressed in GMac for the neural network parts and number of memory lookup operations. The lack of a HP AE saves 3 % to 6 % GMac, depending on whether only the HP decoder (image decoding) or the whole HP codec (image encoding) is used. Decoding with the Manypriors scheme is greatly simplified compared to [6] because the CDF tables generation process takes the optimal indices stored as side-information and looks up one static CDF table per latent spatial dimension, that is C_L (typically 256) fewer lookups than with a HP. During encoding, the Manypriors scheme must lookup every latent variable with every CDF table in order to determine the most cost effective CDF tables. This results in N_{CDF} (typically 64) times more lookup operations than the HP scheme overall, although these lookup operations are relatively cheap because only two values are needed (variable$\pm$0.5), whereas each CDF table lookup in [6] returns L probabilities. Moreover, it is challenging to make an accurate CDF LUT for the HP scheme, because quantizing the distribution scale parameter reduces the accuracy of the resulting CDFs, negatively impacting the bitrate. This challenge is exacerbated when the distribution has multiple parameters [19] or a mixture of distributions [7] is used. In Figure 3, LUT are replaced by accurate but complex Laplace distribution computation for the HP scheme in order to maximize the reported RD performance.

Time complexity is measured for every step on CPU, where it can be reliably profiled due to synchroneous execution. It is summarized in Table 1 with the following distinct sub-categories: NN (neural network) is the time spent in the AE, CDF generation is the time spent building the CDF tables for a specific image, and entropy is the bitstream generation. All operations are done using the PyTorch framework in python, except for entropy encoding which makes use of the torchac range coding library [18, 17], written in C++, and the prior indices are compressed using the LZMA library [8]. The total encoding time of the 64-priors model is 0.32 time that of the HP model and the decoding time is 0.42 times that of the HP model. The timing is more significant when it is broken down by sub-category because each component has a different response time depending on the hardware (and software) architecture in place. The AE ("NN") encoding time is 0.90 that of the HP scheme and decoding time is 0.95 time as much as the HP. Both the hyper-encoder and hyper-decoder are called during encoding, thus it appears that each part of the HP sub-network costs 5 % of the AE time. The time taken to build the CDF tables for the HP model was measured both by estimating the per-variable Laplace distributions ("full-precision") and with a quantized scale parameter LUT. In any case, finding the best indices of

Table 1. Breaking down the image encoding and decoding time, in seconds. Image: 4.5 MP snail [1]. CPU: AMD Ryzen 7 2700X. Time avg. of 50 runs.

(↓)		Hyperprior (Ballé2018)	64-priors (ours)	ratio (ours÷HP)
Encoding	NN encode: main + hyperprior	3.81 + 0.41	3.79 + 0.00	0.90
	entropy encode, main + hyperprior	0.15 + 0.02	0.15 + 0.00	
	CDF: select indices + gather tables	0.00 + FP: 15.95 LUT: 5.66	1.90 + 0.81	**FP: 0.17 LUT: 0.48**
	encode (total)	FP: 20.33 LUT: 10.04	6.65	**FP: 0.32 LUT: 0.66**
Decoding	NN decode : main + hyperprior	10.66 + 0.34	10.50	0.95
	CDF : gather tables	FP: 15.95 LUT: 5.66	0.81	**FP: 0.05 LUT: 0.14**
	entropy decode : main + hyperprior	0.24 + 0.02	0.24	0.92
	decode (total)	FP: 27.21 LUT: 16.92	11.54	**FP: 0.42 LUT: 0.68**

a 64-priors model appears to be relatively inexpensive and the total CDF tables generation time is only 0.17 to 0.48 that of the HP model (depending on whether the HP model uses full-precision or LUT) for encoding. During decoding, the 64-priors model spends 0.05 to 0.14 as much time building the CDF tables as the HP model, because the optimal CDF table indices have already been determined during encoding and they are included in the bitstream.

5. Conclusion

Convolutional autoencoders trained for compression are optimized for both rate and distortion. Rate is estimated with a cumulative probability model, which in turns generates a CDF for every latent variable to be encoded. A single CDF per latent channel is not sufficient to capture the statistics at the output of the encoder, nor to allow the encoder to express a wide variety of features. To support multiple statistics, the hyperprior [6] parametrizes a standard distribution, but this introduces a great deal of complexity in the entropy coding stage because the CDF differs for every latent variable to be encoded. The proposed method uses multiple prior distributions working as a competition of experts to capture the relevant features which they specialize on. This approach is advantageous because the learned CDF tables are stored in a static LUT once training is finished, and a model trained with 64 prior distributions performs with a similar RD as one trained with a HP sub-network. Moreover, a learned CDF table includes the CDF for all channels in the latent code. Hence, accessing the CDF table for a spatial location provides the CDF for each of its channels and the number of lookups is reduced to the number of latent spatial locations. In our experiments, CDF tables generation in the encoding step takes 0.17 to 0.48 as much time with a 64-priors model as it does with the HP model (depending on the precision of the HP model). This ratio is lowered to 0.05 to 0.14 during decoding because the prior indices have already been determined during the encoding.

6. Acknowledgements

This research is funded by the Walloon Region. Computational resources were provided by CISM/UCL and CÉCI, funded by the FRS-FNRS under convention 2.5020.11.

References

[1] Commons test photographs. https://commons.wikimedia.org/wiki/Category:Commons_Test_Photographs. Accessed: 2020-10-22. 2, 4

[2] True color kodak images. http://r0k.us/graphics/kodak/. Accessed: 2020-01-20. 3

[3] Wikimedia commons. https://commons.wikimedia.org. Accessed: 2020-04-03. 3

[4] Challenge on learned image compression. http://challenge.compression.cc/tasks/, 2020. 3

[5] Johannes Ballé, Valero Laparra, and Eero Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017*, 2017. 1, 2

[6] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 4

[7] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 4

[8] Lasse Collin and Igor Pavlov. Xz utils, Jul 2009. 4

[9] Liu Jiaheng. compression. https://github.com/liujiaheng/compression, 2020. 3

[10] Nick Johnston, Elad Eban, Ariel Gordon, and Johannes Ballé. Computationally efficient neural image compression. *CoRR*, abs/1912.08771, 2019. 2

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[12] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991. 1

[13] Jiaheng Liu, Guo Lu, Zhihao Hu, and Dong Xu. A unified end-to-end framework for efficient deep image compression. *arXiv preprint arXiv:2002.03370*, 2020. 2, 3

[14] Shunta Maeda. Fast and flexible image blind denoising via competition of experts. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2239–2247, 2020. 1, 2

[15] Detlev Marpe, Heiko Schwarz, and Thomas Wiegand. Context-based adaptive binary arithmetic coding in the h.264/avc video compression standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):620–636, 2003. 1

[16] Gloria Martín. Range encoding: an algorithm for removing redundancy from a digitised message. In *Video and Data Recording Conference, Southampton, 1979*, pages 24–27, 1979. 1, 3

[17] Fabian Mentzer. torchac. https://github.com/fab-jul/L3C-PyTorch/tree/master/src/torchac, 2020. 3, 4

[18] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Practical full resolution learned lossless image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4

[19] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10771–10780. Curran Associates, Inc., 2018. 2, 4

[20] D. Minnen and S. Singh. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3339–3343, 2020. 2

[21] D. Minnen, G. Toderici, S. Singh, S. J. Hwang, and M. Covell. Image-dependent local entropy models for learned image compression. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 430–434, 2018. 2

[22] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. volume 80 of *Proceedings of Machine Learning Research*, pages 4036–4044, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 1, 2

[23] Alexander Rhatushnyak, Jan Wassenberg, Jon Sneyers, Jyrki Alakuijala, Lode Vandevenne, Luca Versari, Robert Obryk, Zoltan Szabadka, Evgenii Kliuchnikov, Iulia-Maria Comsa, Krzysztof Potempa, Martin Bruse, Moritz Firsching, Renata Khasanova, Ruud van Asseldonk, Sami Boukortt, Sebastian Gomez, and Thomas Fischbacher. Committee draft of jpeg xl image coding system, 2019. 1

[24] Thomas Richter, Joachim Keinert, Antonin Descampe, and Gael Rouvroy. Entropy coding and entropy coding improvements of jpeg xs. In *2018 Data Compression Conference*, pages 87–96, 2018. 1

[25] Vladislav Sovrasov. Flops counter for convolutional networks in pytorch framework. https://github.com/sovrasov/flops-counter.pytorch, 2020. 3

[26] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992. 1