Variable-Rate Deep Image Compression through Spatially-Adaptive Feature Transform

Myungseo Song   Jinyoung Choi   Bohyung Han

ECE & ASRI, Seoul National University, Korea

{micmic123, jin0.choi, bhhan}@snu.ac.kr

Abstract

We propose a versatile deep image compression network based on Spatial Feature Transform (SFT) [45], which takes a source image and a corresponding quality map as inputs and produce a compressed image with variable rates. Our model covers a wide range of compression rates using a single model, which is controlled by arbitrary pixel-wise quality maps. In addition, the proposed framework allows us to perform task-aware image compressions for various tasks, e.g., classification, by efficiently estimating optimized quality maps specific to target tasks for our encoding network. This is even possible with a pretrained network without learning separate models for individual tasks. Our algorithm achieves outstanding rate-distortion trade-off compared to the approaches based on multiple models that are optimized separately for several different target rates. At the same level of compression, the proposed approach successfully improves performance on image classification and text region quality preservation via task-aware quality map estimation without additional model training. The code is available at the project website[1].

1. Introduction

Image compression has been studied actively for decades and has recently become more critical with exploding use of mobile devices for capturing and sharing images. Lossy image compression is a particularly useful technique in this trend since it reduces required space and transmission cost significantly at the expense of quality of reconstructed images. The techniques based on deep learning [5, 6, 11, 18, 23–25, 28–31, 43, 46, 47] have started to outperform traditional codecs including JPEG [36], JPEG2000 [42] and BPG [8]. Many learning-based approaches adopt autoencoder networks for coding nonlinear transformation and optimizing rate-distortion trade-off [13]. They have arisen as new candidates for next generation image compression

---

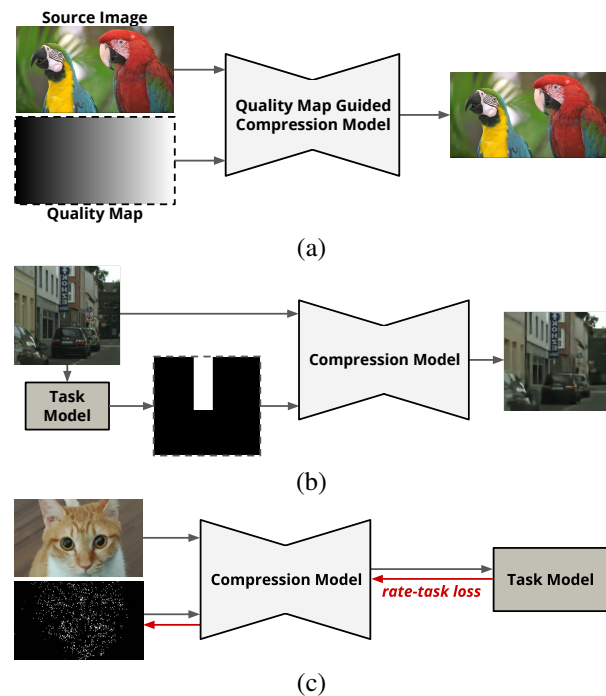[1] https://github.com/micmic123/QmapCompression

Figure 1. The proposed framework. (a) Our compression model takes a quality map as an input along with an image. The compression for the image is conditioned on the quality map, which indicates pixel-wise importance. (b) Task-aware compression at test time via non-uniform quality map. One can create a quality map manually or use an output of a pretrained task model (*e.g.*, object detection results). (c) A task-aware quality map from pretrained models can be estimated at encoding phase by minimizing rate-task loss for the quality map without fine-tuning the models.

standards due to their high performance and applicability compared to the traditional hand-engineered codecs.

The rate-distortion optimization in learned lossy image compression methods is realized by minimizing a combined loss function based on compression ratio and distortion between an original and a corresponding output image. For the objective function, most of existing algorithms rely on a uniform compression ratio over image space. However, all pixels in an image are not equally important and a spatially-
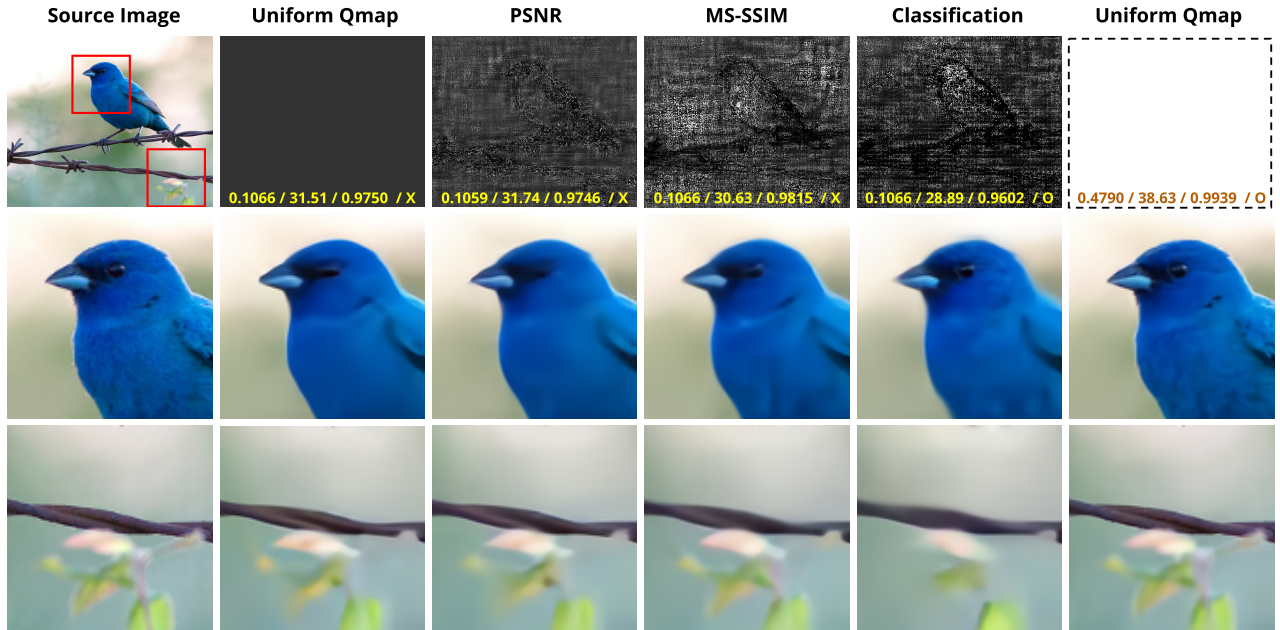
Figure 2. Compression results with various quality maps using our model. All images are obtained by the same model. Except for uniform quality maps (2nd and 6th column), the quality maps are inferred by optimizing the rate-task loss. The numbers in each quality map denote bits per pixel (bpp)/PSNR (dB)/MS-SSIM/classification result of the corresponding reconstructed image. Results by two different uniform maps show that our model adapts well at a wide range of bitrate. The rest of the results show that our model can be adjusted to desired task without additional training of the model. For example, one with classification-aware quality map (5th column) only succeeds in classification among all reconstructed images with similar bpp, while having the worst PSNR and MS-SSIM. Qualitatively, it maintains the quality of discriminative regions for bird recognition, *e.g.*, an eye, at the expense of the quality in other regions.

adaptive image compression by identifying regions of interest (ROIs) would be desirable for better performance. On the other hand, existing models are typically optimized for a single target compression rate and their extensions to multiple rates are not straightforward.

This paper introduces a variable-rate image compression network based on an importance map with spatially adaptive continuous values. Specifically, we optimize the objective function pertaining to rate-distortion trade-off, where image distortion is constrained by a 2D real-valued quality map that defines pixel-wise weight for computing mean squared errors (MSEs). The proposed approach employs spatially-adaptive affine transform modules, which perform pixel-wise feature transformations and result in compressed images guided by a quality map. Note that our model enables us to compress images with arbitrary compression rates and obtain compressed images with spatially-varying quality given by quality maps. In addition to the flexibility of image compression, we also propose a technique to automatically generate task-aware quality maps by backpropagation without retraining the models and construct compressed images optimized for target tasks.

Compared to the existing adaptive image compression techniques based on variable rates and ROIs [2, 3, 9, 11, 12, 18, 27, 43, 46], the proposed framework is much more flexi-

ble and generalizable. The variable-rate models [11, 12, 46] have shown comparable performance with single-rate counterparts. However, Yang *et al*. [46] handle only a few discrete levels of compression using an autoencoder. Choi *et al*. [11] and Cui *et al*. [12] have proposed continuously variable-rate models, but improper selection of quantization bin size leads to a degradation of rate-distortion performance [11]. More importantly, all of these methods do not consider explicit spatial importance for image compression. A variable-rate approach based on a recurrent neural network (RNN) [18] evaluates the distortion of individual patches in a source image to compute weighted distortion. However, it makes the quality of each patch roughly uniform, and realizes spatial adaptiveness by introducing a post-processing for dynamic bit allocation. Minnen *et al*. [32] allows coarse (per-patch) quality variation, but suffers from slow coding speed and low quality compared to recent methods. On the other hand, there exist ROI-based compression methods that reflect the given pixel-wise importance [2, 3, 9, 27], but they are limited to taking a binary mask and compressing images with predefined discrete levels, too.

Figure 1 illustrates the proposed framework for training and inference. The main contributions of our approach are summarized as follows:

- We propose a variable-rate image compression algorithm conditioned on a real-valued quality map, which guides an efficient bit allocations across pixels within an input image.

- We design an effective network architecture based on spatially-adaptive feature transform, which takes advantage of spatial information as a prior, for our conditioned image compression. Our model even excels the compression performance of fixed-rate models in a practical range of bitrates.

- We introduce a method to estimate task-specific quality map for image compression at test time. The inferred quality maps are efficient to acquire and effective to achieve good performance in the target task.

The rest of the paper is organized as follows. Section 2 reviews related literatures on deep image compression and Section 3 presents the proposed method in detail. We show experimental results and analysis in Section 4.

2. Related work

This section first discusses basic image compression techniques based on deep neural networks, and presents two kinds of adaptive models including variable-rate compression and ROI-guided compression. Then, we describe techniques about spatially-adaptive affine transform, which is closely related to the main component in our algorithm.

2.1. Deep image compression

Deep image compression models learn to minimize distortion between a pair of a source image and a reconstructed image while maximizing the likelihood of the quantized latent representation for low entropy coding cost (bitrate). The trade-off between the rate and the distortion is controlled by a Lagrange multiplier λ [13], but most of existing works [1, 4–6, 23–25, 28–31, 37, 44, 47] are limited to learning for a fixed value of λ and obtaining only a single point in a rate-distortion curve for each trained model. Various methods have been proposed to improve the rate-distortion trade-off. For example, [5, 23, 28, 31, 32, 37] incorporate entropy prediction of learned representations based on context during training, and [24, 25, 28] employ importance maps internally for dynamic bit allocation of latent representations. Some approaches introduce additional models for hyper-prior, which provide side information for a conditional entropy model [6, 23, 30].

2.2. Variable-rate compression

There exist a handful of image compression approaches to support variable-rates using a single model [11,12,18,43, 46]. Early methods [18,43] employ RNNs, where the number of iterations are used to control target rates. However, their processing time for encoding and decoding increases as image quality gets higher, which makes the methods impractical. More recent works [11, 12, 46] adopt multiple Lagrangian multiplier values λ to define the loss function and allow trained models to handle multiple rate-distortion trade-offs using feature transformations depending on λ's. Specifically, [46] presents an image compression network based on an autoencoder that supports multiple discrete levels of compression rate. Choi *et al.* [11] select quantization bin sizes in multiple discrete levels to approximate real-valued rates. However, it is not straightforward to determine the discrete levels and the bin sizes, which makes the quality of output images suboptimal. On the other hand, Cui *et al.* [12] modulates compression rates continuously via interpolation of the learned parameters in a pretrained discrete variable rate model. All these aforementioned approaches commonly adopt channel-wise affine transform conditioned on Lagrange multipliers and show comparable accuracy with single-rate models trained independently.

2.3. ROI-based compression

ROI-based models for image compression take a binary mask as an additional input to maintain higher reconstruction quality in the ROI while ignoring or discounting other areas [2, 3, 9, 27]. Agustsson *et al.* [2] synthesizes unimportant regions using a semantic label map using a generative adversarial network (GAN) to achieve extremely low bitrates. Another GAN-based method [27] minimizes the MSE of important regions directly while reducing the distortion of remaining parts indirectly using a feature matching loss for the entire image. Cai *et al.* [9] present a similar approach based on MS-SSIM as a distortion metric, where they train the model to predict ROI in a supervised manner. Akutsu *et al.* [3] employ weighted MS-SSIM conditioned on ROI masks with specific target distortion values. It is worth noting that, to our knowledge, all existing ROI-based models are limited to using binary masks.

2.4. Spatially-adaptive affine transform

The adaptive feature transformation applies a transformation with affine parameters dynamically generated from an external information. After the adaptive instance normalization [16] have been proposed for style transfer, [20, 33, 45] extend this idea to spatially-varying affine transform (denormalization) with element-wise distinction for super-resolution, semantic image synthesis, and denoising. Unlike [20,33], spatial feature transform (SFT) [45] modulates intermediate feature maps of a network without normalization. All these works rely on the external inputs when generating the spatially-varying parameters. On the other hand, our approach employs the original input of the network in addition to the external information for SFT.
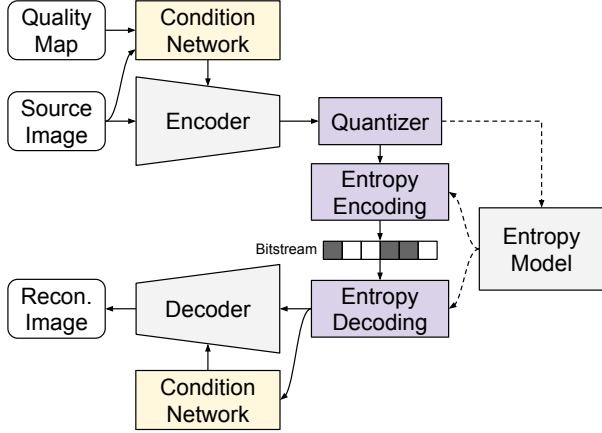
Figure 3. High-level architecture of our model. To perform effective compression conditioned on the quality map, we introduce the condition networks to the compression model.
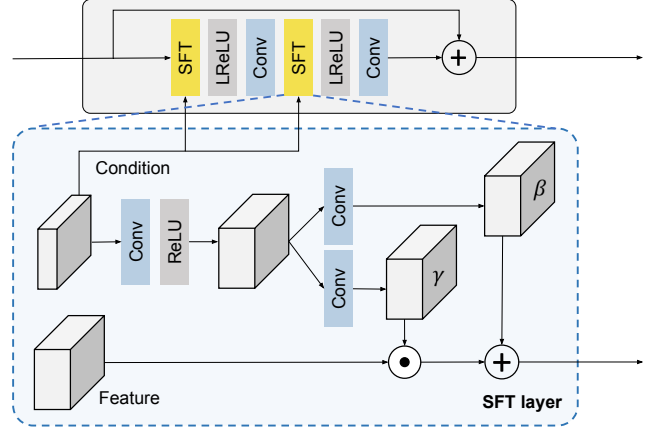

Figure 4. Illustration of Spatial Feature Transform (SFT) layer [45] and residual block we implement, which are related to [33]. The SFT layer takes an intermediate feature of previous layer and a prior condition to produce element-wise scaling and shifting parameters γ and β. As a result, the element-wise affine transformation is performed to the feature.

3. Image Compression Guided by Quality Map

This section presents our main idea of variable-rate image compression via spatially-adaptive feature transform guided by a pixel-wise quality map.

3.1. Overview

The proposed algorithm is a generalized and integrated version of ROI-based and variable-rate image compression methods. For a source image $\mathbf{x} = [x_i]_{i=1:N}$, our model takes a quality map $\mathbf{m} = [m_i]_{i=1:N}$ ($m_i \in [0, 1]$) as its side information to reflect spatial importance of $\mathbf{x}$. The quality map $\mathbf{m}$ defines pixel-wise quality levels after compression. As in the standard image compression approaches, our model consists of three primary components, an encoder, a quantizer, and a decoder. The encoder transforms $\mathbf{x}$ conditioned on $\mathbf{m}$ to a latent representation $\mathbf{y}$, which is then quantized to $\hat{\mathbf{y}}$ by the quantizer. Since rounding operation for quantized values is non-differentiable, it is relaxed to a differentiable alternative—additive uniform noise [5]—during training. After the quantization, an entropy coding, *e.g.*, arithmetic coding [38], is performed on $\hat{\mathbf{y}}$ to save it as a lossless bitstream. For reconstruction, the decoder generates a reconstructed image $\mathbf{x}'$ from $\hat{\mathbf{y}}$.

Figure 3 depicts the high-level concept of the proposed model. The unique components in our approach are the condition networks, which are designed to utilize quality maps effectively. We will make detailed discussion about the architecture of our model in Section 3.4.

3.2. Rate-distortion loss

The goal of lossy image compression is to minimize the length of the bitstream and the distortion between $\mathbf{x}$ and $\mathbf{x}'$ concurrently. This objective raises an optimization problem of minimizing $R + \lambda D$, where a Lagrange multiplier λ of a

fixed value determines the trade-off between the rate R and the distortion D. On the other hand, our model achieves the variable-rate compression by minimizing $R + \Lambda^T \mathbf{D}$, where $\Lambda = [\lambda_i]_{i=1:N} \in \mathbb{R}^N$ is now a vector of Lagrange multipliers varied by the quality map $\mathbf{m}$ instead of a constant scalar while the vectorized distortion $\mathbf{D} \in \mathbb{R}^N$ measures each pixel's distortion. Each element λ_i in Λ is determined by a corresponding m_i by a predefined and monotonically increasing function $T : [0, 1] \rightarrow \mathbb{R}^+$, *i.e.*, $\lambda_i = T(m_i)$. In other words, the higher quality level m_i is, the higher weight λ_i is for the distortion term of the corresponding pixel x_i. This framework leads to an explicit control over spatial bit allocation guided by $\mathbf{m}$.

The estimation of rate R requires to learn an entropy model P conditioned on $\mathbf{m}$, which outputs the likelihood of the given representation $\hat{\mathbf{y}}$, and replace R by an approximate entropy. By choosing squared error as the distortion metric, the training loss of the proposed spatially-adaptive variable-rate image compression model is given by

$$\mathcal{L} = -\log P(\hat{\mathbf{y}}|\mathbf{m}) + \sum_i^N \lambda_i \frac{(x_i - x_i')^2}{N}. \quad (1)$$

Note that a spatially uniform quality map regularizes the standard rate-distortion optimization with a single scalar λ.

3.3. Spatially-adaptive feature transform

Our network is characterized by Spatial Feature Transform (SFT) module [45], for which a condition network generates its appropriate inputs using an external prior. We design the specialized SFT modules to image compression, which effectively reflects a quality map for generating de-
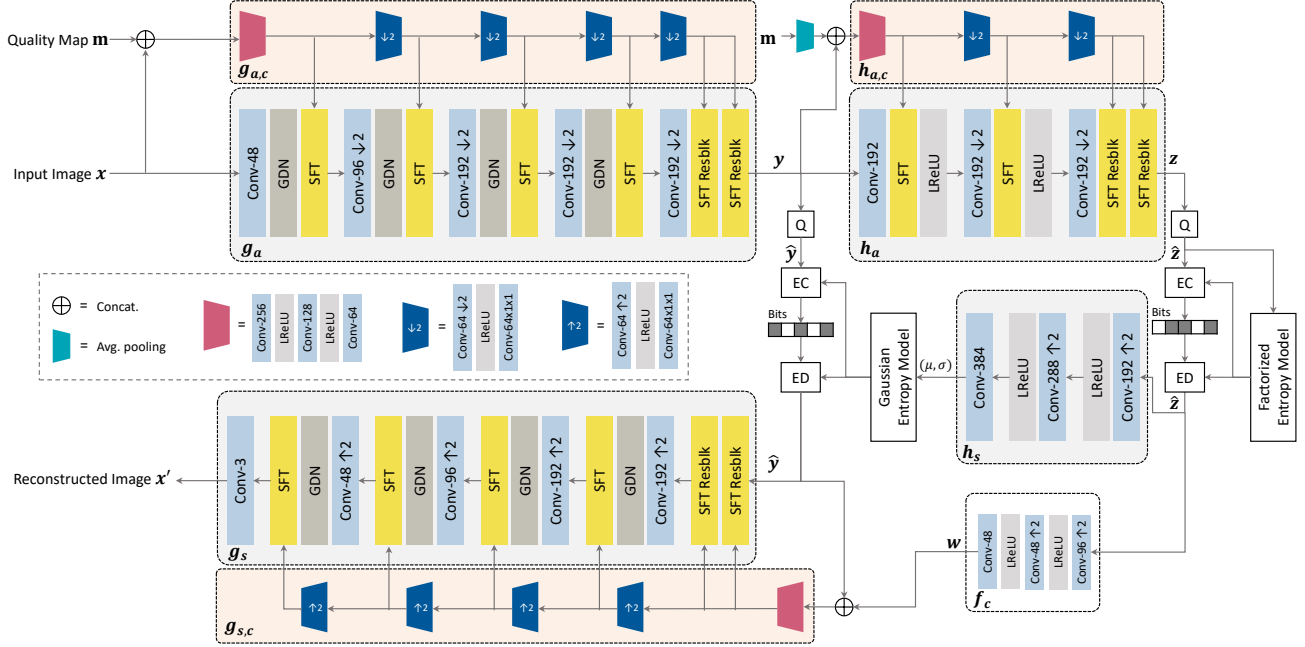
Figure 5. The network architecture of the proposed model. We insert the SFT layers into the convolutional autoencoder, and utilize several condition networks to generate prior condition features from the input image and the quality map. Each of the encoder g_a, the hyper-encoder h_a, and the decoder g_s has its own condition network $g_{a,c}$, $h_{a,c}$ and $g_{s,c}$, respectively. Note that h_s indicates the hyper-decoder. We use 3×3 kernels for the convolution layers in the condition networks and the SFT modules, and 5×5 kernels for others. We adopt a simplified version of generalized divisive normalization (GDN) and inverse GDN (IGDN) [17] in our network. EC and ED denote the entropy coding and entropy decoding, respectively.

sirable compressed representations. Figure 4 illustrates the revised structure of our SFT module. Note that the SFT module learns to generate a set of element-wise affine parameters (γ, β) for an intermediate feature map $\mathbf{f}$, depending on an external condition Ψ; the SFT layer learns a mapping function $\zeta : \Psi \mapsto (\gamma, \beta)$. Within the layer, a feature transformation is given by

$$\text{SFT}(\mathbf{f}, \Psi) = \gamma \odot \mathbf{f} + \beta, \qquad (2)$$

where $\odot$ denotes element-wise multiplication.

In the earlier work of SFT [45], all SFT layers share a condition network and employ only the external information, *i.e.*, $\mathbf{m}$ in our case, as the external prior. In our algorithm, the image compression components such as encoder and decoder adopt their own condition networks, which take the inputs of the components in addition to the external prior and generate suitable spatial importance conditions Ψ for the SFT modules. Moreover, we incorporate the hierarchical SFT layers with progressive downsampling (or upsampling) of the condition features. Also, we transform the condition features to a proper size using convolutions instead of simple methods such as average pooling [20] and nearest-neighbor downsampling [33]. These adjustments turns out to improve the capacity of our model and leads to performance gain.

3.4. Network architecture and pipeline

Figure 5 demonstrates the proposed architecture derived from the Mean & Scale (M&S) Hyperprior model [30]. Among several variations of the models introduced in [30], we choose the non-autoregressive version without a context model for our entropy model P. On top of the baseline network, our approach incorporates a condition network with multiple SFT modules for each of the three components—encoder, decoder, and hyper-encoder, where an SFT module is given a Ψ produced by the matching condition network. We present the detailed procedure of our network designed for image compression in the following.

We first encode an image $\mathbf{x}$ using an encoder $g_a(\cdot, \cdot)$ and a condition network $g_{a,c}(\cdot, \cdot)$ given a quality map $\mathbf{m}$ and generate a latent representation $\mathbf{y}$, which is given by

$$\mathbf{y} = g_a(\mathbf{x}, \Psi_1), \quad \text{where} \quad \Psi_1 = g_{a,c}(\mathbf{x}, \mathbf{m}). \qquad (3)$$

The spatially-adaptive quality information is captured in $\mathbf{y}$. A hyper-encoder $h_a(\cdot, \cdot)$ generates an image-specific side-information $\mathbf{z}$ from the latent representation $\mathbf{y}$, where another condition network $h_{a,c}(\cdot, \cdot)$ is applied to $(\mathbf{y}, \mathbf{m})$ and produces a condition variable Ψ_2 as follows:

$$\mathbf{z} = h_a(\mathbf{y}, \Psi_2), \quad \text{where} \quad \Psi_2 = h_{a,c}(\mathbf{y}, \mathbf{m}). \qquad (4)$$
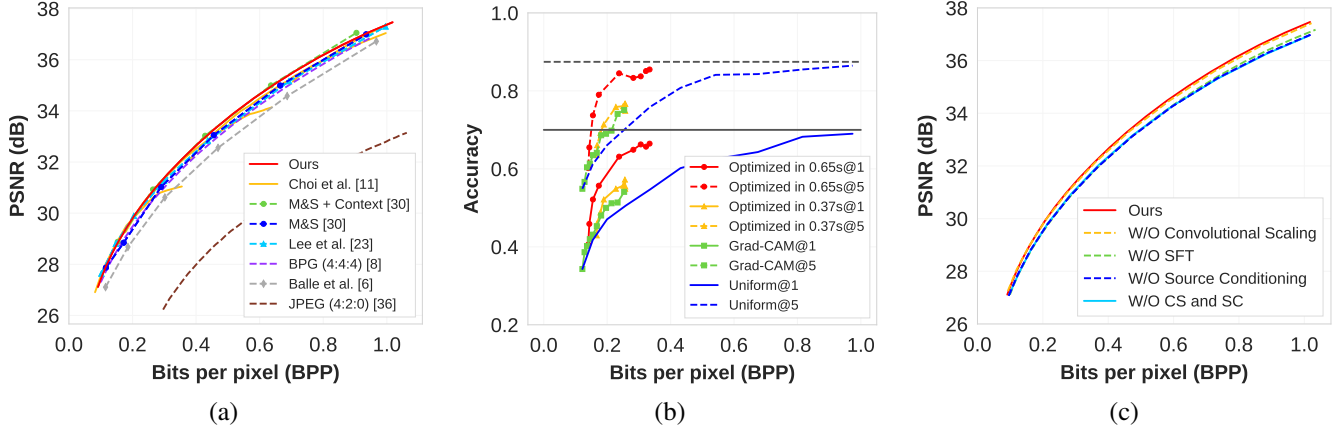
Figure 6. (a) PNSR comparison to baseline methods on the Kodak dataset. Mean & Scale (M&S) Hyperprior model without context model [30] is our counterpart single-rate model. (b) Classification accuracy comparison on the ImageNet dataset using inferred qualtiy maps and uniform quality maps. The gray lines indicate the accuracies for the original images. When we optimize the rate-classification loss three and five times with randomly initialized quality maps, the run times are 0.37 and 0.65 seconds in average, respectively. We obtain the rate-accuracy curves for the optimized quality maps by varying the weight of the cross-entropy loss. As an alternative of the optimization, we adopt Grad-CAM [40] without the ground-truth label as a quality map. (c) Ablation study results. W/O Convolutional Scaling (CS) means the replacement of the convolution layers with average pooling for scaling the intermediate features in the condition networks. In the case of W/O Source Conditioning (SC), all condition networks take only the external inputs, i.e., $\Psi_1 = g_{a,c}(\mathbf{m})$, $\Psi_2 = h_{a,c}(\mathbf{m})$, and $\Psi_3 = g_{s,c}(\mathbf{w})$. Lastly, for W/O SFT, we eliminate all the SFT layers and feed the condition features to the preceding convolution layers.

Note that $\mathbf{z}$ captures the spatial dependencies in the quantized latent representation $\hat{\mathbf{y}} = Q(\mathbf{y})$ and models the probability of $\hat{\mathbf{y}}$ conditionally. Then quantized side information $\hat{\mathbf{z}} = Q(\mathbf{z})$ is forwarded to the hyper-decoder $h_s(\cdot)$ to draw the parameters (μ, σ) of a Gaussian entropy model, which approximates the distribution of $\hat{\mathbf{y}}$.

For reconstructing the image, a decoder $g_s(\cdot, \cdot)$ operates on $\hat{\mathbf{y}}$ and the output of the condition network $g_{s,c}(\cdot, \cdot)$, which is given by

$$\mathbf{x}' = g_s(\hat{\mathbf{y}}, \Psi_3), \quad \text{where} \ \ \Psi_3 = g_{s,c}(\hat{\mathbf{y}}, \mathbf{w}). \quad (5)$$

Note that a resizing (upsampling) network $f_c(\cdot)$ is applied to $\hat{\mathbf{z}}$ to obtain a surrogate $\mathbf{w}$ of a quality map $\mathbf{m}$ since $\mathbf{m}$ may not be available in the decoder side but $\hat{\mathbf{z}}$ maintains the spatial importance information.

3.5. Task-aware image compression

The proposed algorithm provides task-aware image compression capability, which can be done by estimating a task-specific quality map using a pretrained task model at encoding time. Given a task loss function $\mathcal{L}_\text{task}$, the optimal task-specific quality map $\mathbf{m}^*$ is given by the following objective:

$$\mathbf{m}^* = \underset{\mathbf{m}}{\arg\min} \, P(\hat{\mathbf{y}}|\mathbf{m}) + \lambda \mathcal{L}_\text{task}, \quad (6)$$

where λ is a Lagrange multiplier for rate-task trade-off control. Since each element of $\mathbf{m}$ is in $[0, 1]$, the optimization is simply realized by the standard backpropagation.

The pretrained task model can be any network for arbitrary tasks even including the third-party ones. For example,

one may acquire a classification-specific quality map to improve the quality of particular semantic regions by choosing a proper pretrained classifier for $\mathcal{L}_\text{task}$ of (6). We emphasize that this additional optimization at test time encourages the candidate quality map to boost the performance in terms of the target task and it is sufficient to use a uniform quality map for decent reconstruction quality in general. Also, as a practical estimate of $\mathbf{m}^*$, one may use a saliency map from the task model or draw a location-wise information using the intermediate features or outputs of the models, e.g. ROI masks. Nonetheless, our optimized quality maps lead to higher rate-task performance than competing alternatives. Note that our framework deals with arbitrary tasks by only changing $\mathbf{m}$ at encoding time unlike other task-aware approaches that require task-specific training [10]. The supplementary document has more discussions.

4. Experiments

We now present the experimental results of our compression model in comparison to the existing ones and analyze our model in various conditions. Refer to the supplementary document for more discussions and illustrations about our results including training quality map, model complexity, qualitative results, and experimental details.

4.1. Training and evaluation

Datasets and data processing We train our model on the COCO [26] dataset with data augmentation via random cropping of 256×256 images without resizing. We evaluate
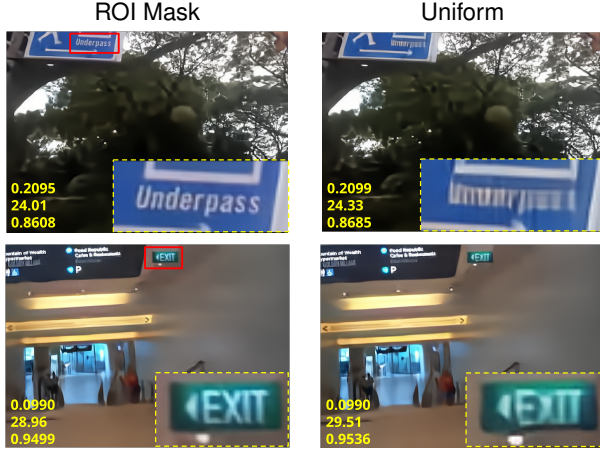
Figure 7. Text-preserving compression examples with ROI masks in our model. In the left column, ROI mask (red box) highlighting the text region is adopted as a quality map, whereas a uniform quality map is used in the right column. The yellow annotations denote the triples of bpp, PSNR (dB), and MS-SSIM.

the rate-distortion performance on the Kodak dataset [22]. For the classification-aware compression, we construct a test set using ImageNet [39] by choosing 102 categories and sampling 5 images per category randomly. We adopt the ICDAR2015 [19] dataset to obtain the text-preserving compression results.

Implementation details Our implementation relies on Pytorch [35] and an open-source compression library [7]. For the entropy coder, we adopt Range Asymmetric Numeral System [14] provided by [7]. We train our model is trained for 2M iteratrions with batch size 8. We employ the Adam [21] optimizer, where the learning rate is initially set to be 10^{-4} and decreased to 10^{-5} after 1.4M iterations. Gradient clipping [34] with threshold 1 leads to stable training according to our experience. The function $T(\cdot)$, which produces Λ from the quality map, is defined as $T(x) = 0.001e^{4.382x}$; this choice approximately results in the bpp range of $[0.1, 1.0]$ on the Kodak dataset.

4.2. Rate-distortion performance

We first evaluate the performance of our variable-rate model by feeding multiple uniform quality maps with no consideration of spatial adaptivity. For comparisons, we select recent learning-based image compression models [6, 11, 23, 30] and the classical state-of-the-art codec, BPG [8]. PSNR is employed as an evaluation metric since all these models are optimized for MSE. This experiment is conducted on the Kodak dataset.

Figure 6(a) illustrates the superior quality of our model compared to the baseline methods. Unlike other variable-rate techniques [11, 46], which achieve slightly worse or

similar performances compared to the single-rate counterparts under the equivalent architectures of the entropy model, our model outperforms the corresponding single-rate model, M&S [30]. Our model is even better or as competitive as Lee *et al*. [23] and M&S+Context [30], which adopt the time-consuming autoregressive context models. Note that the single-rate approaches require training multiple independent models to cover a wide range of rates, *e.g.*, 6 models of M&S in Figure 6(a). Our algorithm also outperforms the recent variable-rate method [11] with the autoregressive context model. We emphasize that our model does not include the context model and has potential to further improve performance with it.

4.3. Task-aware compression

Classification Figure 6(b) compares classification accuracy with three options of quality maps including inferred maps by rate-task optimization, Grad-CAM [40], and uniform quality maps. We employ the pretrained VGG16 [41] to optimize the cross-entropy loss for the rate-task optimization and extract the Grad-CAM outputs. The class with the highest prediction score in each image is selected to estimate Grad-CAM instead of the ground-truth label since the scenario is more practical. To validate the generalization performance, we use pretrained ResNet18 [15] to obtain classification accuracy. Our task-aware quality map optimization at test time significantly improves classification accuracy while incurring less than a second computational cost with a TitanXp GPU. Grad-CAM also outperforms the uniform quality maps, which demonstrates the feasibility of the task-aware compression without the ground-truth labels.

Text preserving Figure 7 demonstrates text-preserving compression results with manually given text ROI masks. The results imply that our model successfully reflects the ROI information and preserves the text regions even at harsh compression rates.

4.4. Qualitative analysis

Comparison with other methods Figure 8 shows our result with a uniform quality map in comparison with M&S Hyperprior model [30] and BPG [8] on an image from the Kodak dataset. Our method reconstructs complex textures better than others while achieving the best PSNR and MS-SSIM scores at lower or the same bitrate.

Quality map vs. bit allocation Figure 9 illustrates the reconstruction and bit allocation results for different quality maps on an image from the Kodak dataset. The bit allocation maps are given by the average negative log-likelihood of $\hat{\mathbf{y}}$ over all channels at every element.
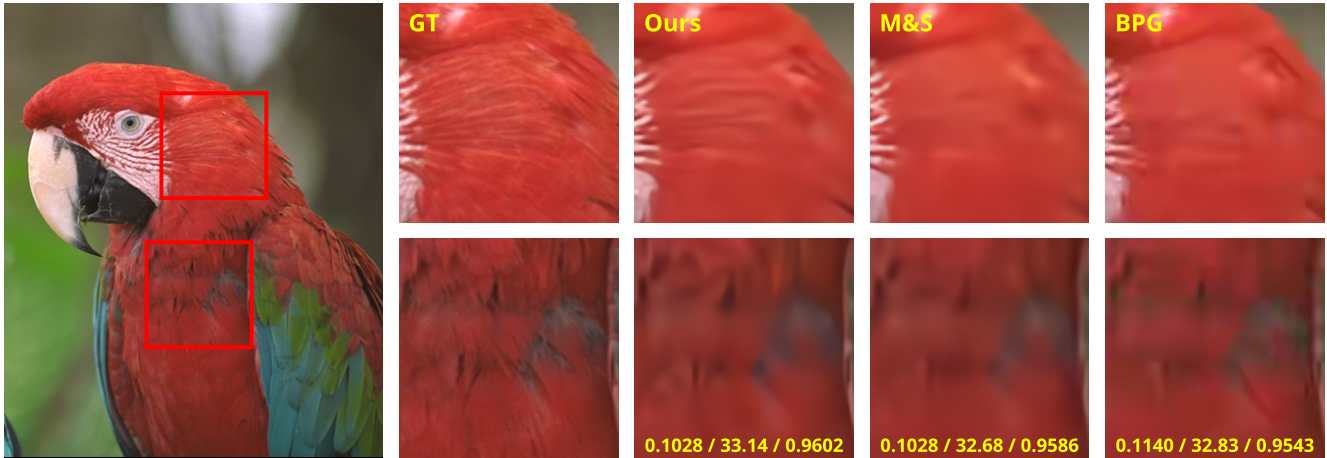
Figure 8. Comparison between our model, M&S Hyperprior model [30], and BPG [8]. The yellow annotations in the cropped images indicate bpp/PSNR (dB)/MS-SSIM of an entire image.
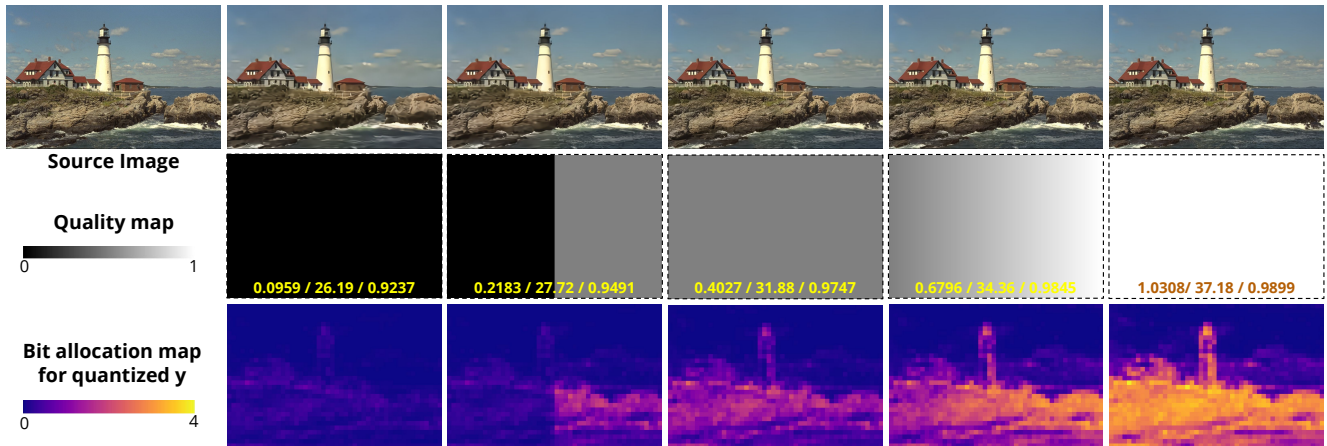

Figure 9. Our spatially-adaptive compression results for uniform and non-uniform quality maps with the corresponding bit allocation maps. The yellow numbers in the quality map denote bpp/PSNR (dB)/MS-SSIM. The reconstructed images are best views with zooming-in.

4.5. Ablation Study

Figure 6(c) presents the results from the ablation study to validate the contribution of each module on rate-distortion performance; the proposed approach adopting the SFT layers and refining the modules improves the quality of the variable-rate image compression. Most importantly, comparing our model with W/O Source Conditioning case, additional conditioning on the original input of the network as well as the external data, *e.g.*, $\mathbf{x}$ and $\mathbf{m}$ for $g_{a,c}$, contributes to the performance noticeably. This is because the adaptivity of our model is enhanced by the generation of different condition features depending on the original image given the same quality map.

5. Conclusion

We presented a novel variable-rate image compression algorithm based on an end-to-end trainable deep neural network. The proposed framework enables us to perform spatially-adaptive image compression in terms of compression rate using a single model given a real-valued pixel-wise quality map. In addition to the flexibility, our model has an additional benefit to estimate task-specific quality maps automatically at test time without additional training, which is useful to perform task-aware image compression. We design an efficient network architecture based on spatially-adaptive feature transform for the conditioned image compression. Our experiment shows that the proposed algorithm achieves outstanding performance even compared to multiple models trained independently with fixed rates and has wide applicability to various tasks in computer vision.

References

[1] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc J Van Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *NeurIPS*, 2017. 3

[2] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *ICCV*, pages 221–231, 2019. 2, 3

[3] Hiroaki Akutsu and Takahiro Naruko. End-to-end learned roi image compression. In *CVPRW*, page 0, 2019. 2, 3

[4] Sharon Ayzik and Shai Avidan. Deep image compression using decoder side information. In *ECCV*, pages 699–714. Springer, 2020. 3

[5] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *ICLR*, 2017. 1, 3, 4

[6] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *ICLR*, 2018. 1, 3, 7

[7] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. CompressAI: a PyTorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020. 7

[8] Fabrice Bellard. BPG image format, 2014. 1, 7, 8

[9] Chunlei Cai, Li Chen, Xiaoyun Zhang, and Zhiyong Gao. End-to-end optimized roi image compression. *TIP*, 29:3442–3457, 2019. 2, 3

[10] Jinyoung Choi and Bohyung Han. Task-aware quantization network for JPEG image compression. In *ECCV*, pages 309–324. Springer, 2020. 6

[11] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In *ICCV*, pages 3146–3154, 2019. 1, 2, 3, 7

[12] Ze Cui, Jing Wang, Bo Bai, Tiansheng Guo, and Yihui Feng. G-VAE: A continuously variable rate deep image compression framework. *arXiv preprint arXiv:2003.02012*, 2020. 2, 3

[13] Lee D Davisson. Rate-distortion theory and application. *Proceedings of the IEEE*, 60(7):800–808, 1972. 1, 3

[14] Jarek Duda. Asymmetric numeral systems: entropy coding combining speed of huffman coding with compression rate of arithmetic coding. *arXiv preprint arXiv:1311.2540*, 2013. 7

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 3

[17] Nick Johnston, Elad Eban, Ariel Gordon, and Johannes Ballé. Computationally efficient neural image compression. *arXiv preprint arXiv:1912.08771*, 2019. 5

[18] Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, and George Toderici. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *CVPR*, pages 4385–4393, 2018. 1, 2, 3

[19] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. ICDAR 2015 competition on robust reading. In *ICDAR*, pages 1156–1160. IEEE, 2015. 7

[20] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *CVPR*, pages 3482–3492, 2020. 3, 5

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7

[22] Eastman Kodak. Kodak lossless true color image suite (PhotoCD PCD0992). 7

[23] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. In *ICLR*, 2018. 1, 3, 7

[24] Mu Li, Wangmeng Zuo, Shuhang Gu, Jane You, and David Zhang. Learning content-weighted deep image compression. *TPAMI*, 2020. 1, 3

[25] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. In *CVPR*, pages 3214–3223, 2018. 1, 3

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 6

[27] Jonas Löhdefink, Andreas Bär, Nico M Schmidt, Fabian Hüger, Peter Schlicht, and Tim Fingscheidt. Focussing learned image compression to semantic classes for V2X applications. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1641–1648. IEEE, 2020. 2, 3

[28] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *CVPR*, pages 4394–4402, 2018. 1, 3

[29] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *NeurIPS*, 33, 2020. 1, 3

[30] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *NeurIPS*, 31:10771–10780, 2018. 1, 3, 5, 6, 7, 8

[31] David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *ICIP*, pages 3339–3343. IEEE, 2020. 1, 3

[32] David Minnen, George Toderici, Michele Covell, Troy Chinen, Nick Johnston, Joel Shor, Sung Jin Hwang, Damien Vincent, and Saurabh Singh. Spatially adaptive image compression using a tiled deep network. In *ICIP*, pages 2796–2800. IEEE, 2017. 2, 3

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019. 3, 4, 5

[34] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *ICML*, pages 1310–1318. PMLR, 2013. 7

[35] 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. *NeurIPS*, 32:8026–8037, 2019. 7

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

[37] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *ICML*, pages 2922–2930. PMLR, 2017. 3

[38] Jorma Rissanen and Glen Langdon. Universal modeling and coding. *IEEE Trans. Inf. Theory*, 27(1):12–23, 1981. 4

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 7

[40] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 6, 7

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 7

[42] Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi. The JPEG 2000 still image compression standard. *IEEE Signal processing magazine*, 18(5):36–58, 2001. 1

[43] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. In *CVPR*, pages 5306–5314, 2017. 1, 2, 3

[44] Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep generative models for distribution-preserving lossy compression. In *NeurIPS*, 2018. 3

[45] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, pages 606–615, 2018. 1, 3, 4, 5

[46] Fei Yang, Luis Herranz, Joost van de Weijer, José A Iglesias Guitián, Antonio M López, and Mikhail G Mozerov. Variable rate deep image compression with modulated autoencoder. *SPL*, 27:331–335, 2020. 1, 2, 3, 7

[47] Zhisheng Zhong, Hiroaki Akutsu, and Kiyoharu Aizawa. Channel-level variable quantization network for deep image compression. *IJCAI*, 2020. 1, 3