Coarse-to-Fine Hyper-Prior Modeling for Learned Image Compression

Yueyu Hu, Wenhan Yang, Jiaying Liu *
Wangxuan Institute of Computer Technology, Peking University, Beijing, China
{huyy, yangwenhan, liujiaying}@pku.edu.cn

Abstract

Approaches to image compression with machine learning now achieve superior performance on the compression rate compared to existing hybrid codecs. The conventional learning-based methods for image compression exploits hyper-prior and spatial context model to facilitate probability estimations. Such models have limitations in modeling long-term dependency and do not fully squeeze out the spatial redundancy in images. In this paper, we propose a coarse-to-fine framework with hierarchical layers of hyper-priors to conduct comprehensive analysis of the image and more effectively reduce spatial redundancy, which improves the rate-distortion performance of image compression significantly. *Signal Preserving Hyper Transforms* are designed to achieve an in-depth analysis of the latent representation and the *Information Aggregation Reconstruction* sub-network is proposed to maximally utilize side-information for reconstruction. Experimental results show the effectiveness of the proposed network to efficiently reduce the redundancies in images and improve the rate-distortion performance, especially for high-resolution images. Our project is publicly available at https://huzi96.github.io/coarse-to-fine-compression.html.

Introduction

Image compression is one of the most fundamental technology in media sharing and storage. With the booming of high-resolution visual applications like 8K streaming and Virtual Reality (VR), it is a critical issue to obtain superior efficiently compressed and high-quality image/video adapted to the limited hardware resources. Current image compression standards, *e.g.* JPEG (Wallace 1992) and BPG (Bellard. 2014) based on HEVC (Sullivan et al. 2012), still follow the transform hybrid coding framework, which is composed of the transform, quantization and entropy coding. Recent advances in machine learning methods accelerate the development of high efficient image compression methods. The state-of-the-art learning-based image compression methods now achieve better performance than hybrid codecs (*i.e.* BPG).

Transform coding methods exploit manually designed transforms, *e.g.* Discrete Cosine Transform (DCT) (Wallace 1992) to make energy compact and decorrelate pixels. The most important portion of the information in the to-be-encoded image is aggregated on lower frequencies after the transform. Those lower-frequency components are quantized with less information loss to maximally keep the quality. Besides, the coefficient of each frequency component is expected to be independently distributed after the transform, namely the redundancies have been fully squeezed out. After that, the coefficients can be efficiently encoded with entropy coding.

In recent years, end-to-end optimized image compression methods with neural networks are proposed to exploit the powerful modeling capacities of deep learning to facilitate the development of image compression methods. The methods usually utilize an end-to-end trainable model to jointly optimize the rate and distortion. The state-of-the-art networks for image compression (Minnen, Ballé, and Toderici 2018; Lee, Cho, and Beack 2018) include the Generalized Divisive Normalization (GDN) (Ballé 2018) or design non-local models (Liu et al. 2019) in the analysis and synthesis transform to reduce spatial redundancy in images. A conditional entropy model with hyper-prior and context model is included for entropy coding. The implemented learned image compression methods out-perform existing hybrid frameworks like BPG (Bellard. 2014) in both PSNR-bit-rate and MS-SSIM-bit-rate metrics.

In summary, lossy image compression methods are improved mainly along two routes.

- To reduce bit-rates with less quality loss, the information neglected by the encoder thus not stored in the bit-stream should contribute least to the reconstruction quality. Thus, the compressor needs to select the most important portion of information during encoding.
- The quantized representation of the image is encoded by its entropy model to reduce the actual number of bits to be used. This requirement is usually approached with an accurate estimation of the joint distribution of the to-be-encoded representation of the images.

*Corresponding Author. This work is supported by National Natural Science Foundation of China under contract No.61772043, Beijing Natural Science Foundation under contract No.L182002.

Though transform coding schemes are designed to meet these two requirements, they face limitations in real applications. For example, the JPEG codec partitions images into 8×8 blocks and then apply DCT. Therefore, the inter-block redundancies still exist in the bit-stream. In HEVC and BPG, intra prediction is introduced to reduce such redundancy, and context adaptive entropy codec (Marpe, Schwarz, and Wiegand 2003) is also exploited to further reduce contextual redundancy in the bit-stream. These techniques greatly reduce the bit-rate and thus improve overall rate-distortion performance. However, as the pipeline of the hybrid compressor is totally hand-crafted and already sophisticated, future improvement is more and more difficult.

Existing learning-based methods with the context model and hyper-prior still neglect some issues. First, the estimation of the probability for a to-be-encoded element in the latent representation depends on a local block of previously decoded elements, which limits the accuracy of long-term conditioned probability estimation. Also, it can hardly be accelerated by current large-scale parallel computing devices, *e.g.* GPU. Second, due to the limitation of shallow networks, existing analysis and synthesis transforms are not able to maximally reduce spatial redundancies while maintaining reconstruction quality. Third, the information transmitted using the hyper-prior is not regularized and utilized. While this part of information is encoded in the bit-stream, it is not used for the reconstruction of the image. Besides, though the combination of the hyper-prior and the context model improves overall rate-distortion performance, the way they cooperate and interact with each other has not been well studied.

This paper aims to address the limitations and explore to construct an effective and efficient framework that integrates both the hyper-level priors and the latent representations of images for efficient compression. According to our analysis, the contextual dependency captured by the context model can also be modeled as the higher-level latent representation, while the latter can be fully parallelized and include a wider range of context. The expansion of context provides a better capability to reduce underlying spatial redundancy in the latent representation. We stack multiple layers of latent representations. Each layer is the further abstract of the previous layer and is trained to approximate a factorized conditional entropy model. In this way, the total bit-rate can be further reduced. Therefore, the rate-distortion performance for the codec is improved. We design an information aggregation reconstruction sub-network to fully utilize the latent representations at different layers. The contributions of this paper are summarized as follows,

- We design a multi-layer hyper-representation framework for learned image compression. Without the additional element-wise context model, the proposed framework can model the spatial dependencies in the latent representation. It can also capture spatial redundancy in a wider range of context compared with the original context model, resulting in a more accurate conditional entropy model for the encoding of the latent representations.
- To effectively reduce spatial redundancy in the latent representations, we propose the Signal-Preserving Hyper Transform for the analysis and synthesis transforms of the hyper representations. The newly designed structure is shown to better preserve information while it reduces spatial redundancies. It enables to build a multi-layer coarse-to-fine hyper-prior model to guide the learning of latent representations for better image compression.
- We propose the information aggregation sub-network to fully utilize the encoded information in the features of different layers. The hyper latent representations serve as the side-information for the reconstruction of the image. Experimental results show significant improvement of reconstruction quality with the proposed sub-network.

Related Works

Two fundamental problems in end-to-end learned image compression are first designing decomposing transform to decorrelate the signal and second building probability model for effective entropy coding. Many deep learning-based image compression methods have been proposed in the past few years to investigate the solutions to the problems and therefore improve the compression performance with learned methods.

From the perspective of the network architecture design, these methods can be roughly divided into two categories: *fully convolutional* and *recurrent models*. *Fully convolutional networks* (Ballé, Laparra, and Simoncelli 2017; Ballé et al. 2018; Theis et al. 2017; Lee, Cho, and Beack 2018; Agustsson et al. 2017; Li et al. 2018; Mentzer et al. 2018) are trained with the rate-distortion constraint to trade-off between the bit-rate and reconstruction quality. Each trained model corresponds to one Lagrange coefficient λ that controls that balance. Thus, multiple models need to be trained to satisfy the need of different rate-distortion trade-offs. Instead, *Recurrent methods* encode images in a progressive manner. In each recurrence, a fixed dimension code is generated to encode the image or the residue signal that fails to be captured in previous steps (Toderici et al. 2017; Johnston et al. 2018; Toderici et al. 2016; Baig, Koltun, and Torresani 2017).

Recurrent models apply constraints on the dimensionality of the representations of each step. In such low dimensionality the elements are close to independently distributed. Besides, in (Toderici et al. 2017) a neural network-based binarizer is designed to model the contextual dependency in the sequence of encoded symbols to further squeeze out redundancies and facilitate entropy coding. Predictions from previous steps (Baig, Koltun, and Torresani 2017) to the to-be-encoded area further enhance the ability of de-correlation. However, the recurrent structures have issues in complexity, especially for higher ranges of bit-rate. Thus, more recent works focus on the fully convolutional structure.

A typical fully convolutional framework usually utilizes GDN (Ballé 2018) for local redundancy removal. GDN is applied as the activation function for multiple layers in the analysis transform (Ballé et al. 2018; Theis et al. 2017). This normalization consistently reduces spatial redundancy for the activation of each layer, as well as the to-be-encoded

latent representation. Different from recurrent models, the bit-rate to encode an image is minimized by evaluating the joint entropy of the quantized code and making it a term in the loss function. For probability estimation, in addition to the basic factorized model, Ballé *et al.* (Ballé et al. 2018) proposed a trainable probability density model to measure the distribution of the encoded information for entropy calculation and arithmetic coding. It is further assumed that the quantized code follows a Gaussian Distribution. The scale of the model is encoded by a hyper-prior encoder, which further reduces the spatial correlations. Successive works (Minnen, Ballé, and Toderici 2018; Lee, Cho, and Beack 2018) follow the work and generalize the hyper parameters of the probability model from both the quantized codes and the corresponding decoded contexts. The two works are the first to achieve superior performance to BPG 4:2:0 and BPG 4:4:4 (Bellard. 2014) with deep neural network architectures for image compression.

We adopt the basic structure of the fully convolutional framework, while we use no context-model for entropy coding, as it also faces complexity issues in parallel devices. Instead, we propose the coarse-to-fine architecture for conditional entropy modeling with the information aggregation sub-network and it shows improved performance.

Proposed Method

Motivation

The basic framework of variational auto-encoder based image compression is mainly composed of an analysis transform $\mathcal{G}_a$ and a synthesis transform $\mathcal{G}_s$. $\mathcal{G}_a$ maps the image to a latent representation and $\mathcal{G}_s$ decodes the image from that representation. Though the analysis transform with GDN (Ballé et al. 2018; Ballé, Laparra, and Simoncelli 2017) can largely reduce spatial redundancy among pixels to enable a factorized probability modeling of the latent representations, it might neglect some correlations due to its limited capacities to percept contexts. Later works (Minnen, Ballé, and Toderici 2018; Lee, Cho, and Beack 2018) propose to further model contextual dependency by approximating the joint distribution $P(\mathbf{X})$ of the latent representation $\mathbf{X} = \{X_i\}$ in a context-adaptive way as follows,

$$P(\mathbf{X}) \doteq \prod_i P(X_i|X_{i-1}, X_{i-2}, ..., X_{i-m}), \qquad (1)$$

where the conditioning elements of random variables $X_{i-1}, X_{i-2}, ..., X_{i-m}$ are previously encoded and decoded elements which are available when the bit representations of the current symbol are calculated. Eq. (1) is an approximation of the real joint distribution $P(\mathbf{X}) = P(X_1)P(X_2|X_1)\cdots P(X_i|X_{i-1}, ..., X_1)$ when only considering the influence of the nearest m elements. Several issues are met in this approximation. First, this approximation might fail if there exist the long-term dependencies among the variables. Second, when we actually compress an image, there exists a trade-off in selecting a proper m. A large m results in a superior modeling capacity as well as high complexity, which sets obstacles to real applications. Third, the element-wise dependency requires each probability distribution to be estimated solely, making it difficult to accelerate the encoding and decoding process with large-scale parallel computing devices.

In order to get rid of the above-mentioned issues, we propose a context model-free multi-layer approach as an alternative. The encoding and decoding modules are jointly optimized in an end-to-end manner and achieve superior time-efficiency and modeling capacity.

Coarse-to-Fine Hyper-Prior Modeling

Earlier works on learned image compression model the latent representation of images using factorized entropy models (Theis et al. 2017; Ballé, Laparra, and Simoncelli 2017), where each element is considered independently distributed. To make the assumptions of independence hold, a more complex transform for better redundancy removal and reconstruction is needed. Such analysis and synthesis transforms contain too many parameters and are less time-efficient. Therefore, we choose to allow some spatial redundancies in the latent representations. Instead, an additional layer is designed to further explore such redundancy and produce the representation $\mathbf{Y} = \{Y_i\}$ such that, conditioned on $\mathbf{Y}$, the probability distribution $P(\mathbf{X}|\mathbf{Y})$ can be approximately factorized as follows,

$$P(\mathbf{X}) = P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{Y})P(\mathbf{X}|\mathbf{Y}). \qquad (2)$$

The first equality in Eq. (2) holds as $\mathbf{Y}$ is generated from $\mathbf{X}$ in a deterministic manner. When introducing the additional layer, we reduce the difficulty of approximating the joint distribution of $P(\mathbf{X})$ by separately estimating $P(\mathbf{Y})$ and $P(\mathbf{X}|\mathbf{Y})$. First, the elements in the second layer representation $\mathbf{Y}$ are easier to model, as during the analysis transform from $\mathbf{X}$ to $\mathbf{Y}$, the dimensionality of the latent representation is further reduced and more redundancy is squeezed out. Thus the joint distribution of the latent representation $\mathbf{Y}^*$ at the most inner layer can be approximately factorized, as

$$P(\mathbf{Y}^*) = P(Y_1^*, Y_2^*, \cdots, Y_n^*) \approx \prod_i P(Y_i^*). \qquad (3)$$

Second, by minimizing the entropy of the latent representation, the network is tuned to accurately estimate $P(\mathbf{X}|\mathbf{Y})$, where $\mathbf{Y}$ denotes higher level hyper representation generated from $\mathbf{X}$. Existing works (Van den Oord et al. 2016; Mirza and Osindero 2014) show that neural networks are capable of modeling conditional probability distributions. We also train the hyper representation $\mathbf{Y}$ to embed the main component of the to-be-compressed images. Therefore the joint distribution can also be approximately factorized, as,

$$P(\mathbf{X}|\mathbf{Y}) = P(X_1, X_2, \cdots, X_n|\mathbf{Y}) \approx \prod_i P(X_i|\mathbf{Y}), \quad (4)$$

where all the elements in $\mathbf{Y}$ can be utilized as the conditions to estimate X_i, unlike existing block-conditioning context models. With wider condition the coarse-to-fine structure can explore long term correlations in images.

The overall structure of the multi-layer framework is shown in Fig. 1. The input image is transformed into the latent representation $\mathbf{X}$ with the analysis transform network.

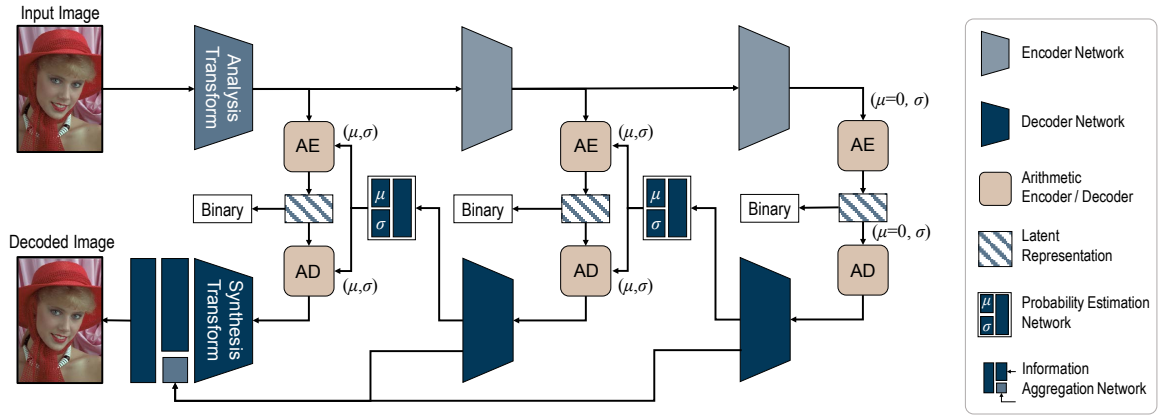


Figure 1: Overall architecture of the multi-layer image compression framework. The probability distribution of the most inner layer of hyper-prior is approaximated with a zero-mean Gaussian distribution, where the scale values σ are channel-wise independent and spatially shared trainable parameters.

It aims to squeeze out pixel-wise redundancy as far as possible. GDN is used as the activation for in the analysis transform and inverse GDN in the synthesis transform, correspondingly. To further explore the correlations, the latent representation is then further analyzed and the higher-level representation is extracted and encoded into the bit-stream. With the symmetric hyper synthesis transform network, the latent representation is decoded to form the prior $\mathbf{Y}$. With the decoded prior, we estimate the probability distribution of $\mathbf{X}$ with the prediction model. The distribution of each element in $\mathbf{X}$ is assumed to be a Gaussian distribution with mean μ and scale σ as its parameters, following previous works (Ballé et al. 2018; Minnen, Ballé, and Toderici 2018; Lee, Cho, and Beack 2018). Thus, the prediction model takes $\mathbf{Y}$ as its input and predicts μ and σ to represent the estimated distribution. According to the information theory, the minimum bit-rate required to encode $\mathbf{X}$ with the estimated distribution equals to the cross entropy of the real distribution $P(\mathbf{X}|\mathbf{Y})$ and the estimated distribution $\hat{P}(\mathbf{X}|\mathbf{Y}) \sim N(\mu, \sigma)$, denoted as,

$$R = H(\hat{P}) + D_{KL}(P||\hat{P}) = E_{P_{\mathbf{X}|\mathbf{Y}}}[-log(\hat{P})]. \quad (5)$$

In order to accelerate the convergence during the training of the multi-layer network, an additional information-fidelity loss is introduced. This loss term encourages the hyper representation $\mathbf{Y}$ to maintain the critical information in $\mathbf{X}$ during training, formulated as,

$$\min_{\mathbf{Y},\theta} \mathcal{L}_{if} = ||\mathcal{F}(\mathbf{Y}, \theta) - \mathbf{X}||_2. \quad (6)$$

In practice, the function $\mathcal{F}$ with trainable parameter θ is one convolutional layer with no non-linear activation. The information-fidelity loss takes the form of the least square error, to make the prediction of the μ and σ more accurate.

Signal Preserving Hyper Transform

To conduct coarse-to-fine modeling of images, it is important to preserve the information while performing more complex analysis and synthesis transforms in the succeeding hyper layers. Therefore, the Signal Preserving Hyper Transform is proposed to build the framework with multiple layers. We observe that elements in the latent representation produced by the main analysis transform are much less correlated compared with the pixels in natural images, as the spatial redundancy has been largely reduced. Therefore, local correlations in the feature maps are weak, while convolutions with large kernels rely on such local correlations for effective modeling. Besides, previous transform network consists of stride convolutions with the ReLU activation. Stride convolutions down sample the feature maps while activation functions like ReLU intuitively disable some of the filter neurons that produce negative values. As the dimensionality of these convolution layers needs to be limited to ensure the gradual factorization of the latent representation, the original hyper transform drops much information during the processing.

In summary, the issues of the original analysis transform in the proposed architecture lie in the two aspects: 1) Original analysis transforms fix the number of channels and down-sample the feature maps, reducing the dimensionality. 2) Combining large convolution kernels with ReLU at the beginning of the analysis transform or the end of the synthesis transform will drop some information that has not been transformed, limiting the capacity.

The Signal Preserving Hyper Transform is designed to facilitate the multi-layer structure by preserving information for coarse-to-fine analysis. The structure of the analysis and synthesis transform network is illustrated in Table 1. Instead of using large kernels in the filters, we employ a relatively small filter in the first layer with no non-linear activation and we conduct 1×1 convolutions in the rest layers. The first layer in the network expands the dimensionality of the original representation. Combined with succeeding non-linear layers, the expansion of dimensionality preserves the information of the original representation while providing the ability for non-linear modeling. We exploit a *space-to-*

Table 1: Structure of the signal preserving hyper transform.

(a) Hyper analysis transform.

Name	Operation	Output Shape	Activation
Input	/	(b, h, w, c)	/
#1 E	Conv. (3×3)	$(b, h, w, 2c)$	Linear
Down	Space-to-Depth	$(b, h/2, w/2, 8c)$	/
#2 E	Conv. (1×1)	$(b, h/2, w/2, 4c)$	ReLU
#3 E	Conv. (1×1)	$(b, h/2, w/2, 4c)$	ReLU
#4 E	Conv. (1×1)	$(b, h/2, w/2, c')$	Linear

(b) Hyper synthesis transform.

Name	Operation	Output Shape	Activation
Input	/	$(b, h/2, w/2, c')$	/
#1 D	Deconv. (1×1)	$(b, h/2, w/2, 4c)$	Linear
Up	Depth-to-Space	(b, h, w, c)	/
#2 D	Deconv. (1×1)	$(b, h, w, 4c)$	ReLU
#3 D	Deconv. (1×1)	$(b, h, w, 4c)$	ReLU
#4 D	Deconv. (3×3)	(b, h, w, c)	Linear

depth operation to reshape the tensor of the representations, making spatially adjacent elements scatter in the same location but different channels. In this way the succeeding 1×1 convolutions are able to explore spatial redundancy in a non-linear manner. At the final layer of the network, we conduct a dimensionality reduction on the tensor to make the representation compact. We symmetrically design the hyper synthesis transform to produce what is denoted as $\mathbf{Y}$ in Eq. (2) as the condition prior for outer layer and side-information for the reconstruction.

Information Aggregation for Reconstruction

In the decoding process, the synthesis transform maps latent representations back to pixels. To best reconstruct the image, the decoder needs to fully utilize the provided information in the bit-stream. The practical image and video compression usually exploits side-information to improve quality. With this idea in mind, we take hyper latent representations as side-information and aggregate information from all the layers of the hyper latent representations to reconstruct the decoded image in the proposed framework. The architecture of the information aggregation decoding network is shown in Fig. 2. Both the main latent representation and the higher order representations of smaller scales are upsampled by the decoding network to half the size of the original image. A fusion is conducted with a concatenation of the two representations. The fused representation is then processed by a residue block and then upsampled to the scale of the decoded image.

By fusing the main representation and the hyper representations, information of different scales contributes to the reconstruction of the decoded image, where the higher order representations provide the global information and the others preserve details in the image. The fusion process is conducted in smaller spatial resolutions to avoid high computational complexity. After the fusion of features, we employ a single residue block with peripheral convolution layers to map the feature maps back to pixels.

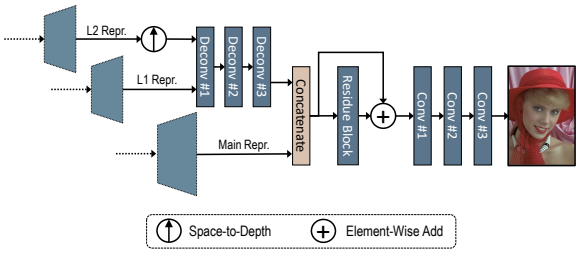


Figure 2: Information aggregation subnetwork for the reconstruction of the decoded image. The main latent representation (Main Repr.) and the two layers of hyper representations (L1 Repr. and L2 Repr.) are aggregated for the reconstruction.

Implementation Details

Probability Estimation Sub-Network As shown in Fig. 1, to conduct arithmetic coding, the entropy codec requires a probability estimation for the to-be-encoded symbol quantized from the latent representation. For the most inner layer with the smallest scale, we simply employ a zero-mean Gaussian assumption for the entropy model, with learned scale value for each channel to estimate such distribution. For the rest layers, we estimate the distribution of each element with a Gaussian distribution of predicted mean and scale. We follow (Lee, Cho, and Beack 2018) and design a prediction sub-network with a more appropriate sampling area in our proposed model to fully utilize the hyper-priors. The structure of the sub-network is illustrated in Fig. 3. With the decoded hyper representation as the input, an area of size 5×5 is sampled, centered at the to-be-predicted positions. Each sampled patch is processed using a multi-layer convolutional sub-network to estimate the probability. At the last layer of the sub-network, the patch is flattened to a vector and a fully-connected layer is used to map the feature vector to the vector of mean and scale for the current position. The local convolutions share the kernel along all the spatial positions. Therefore, the hyper latent representations are densely sampled to complete the estimation, without the potential information loss during convolutional operations.

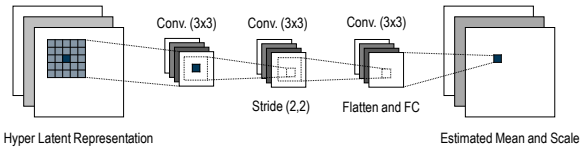
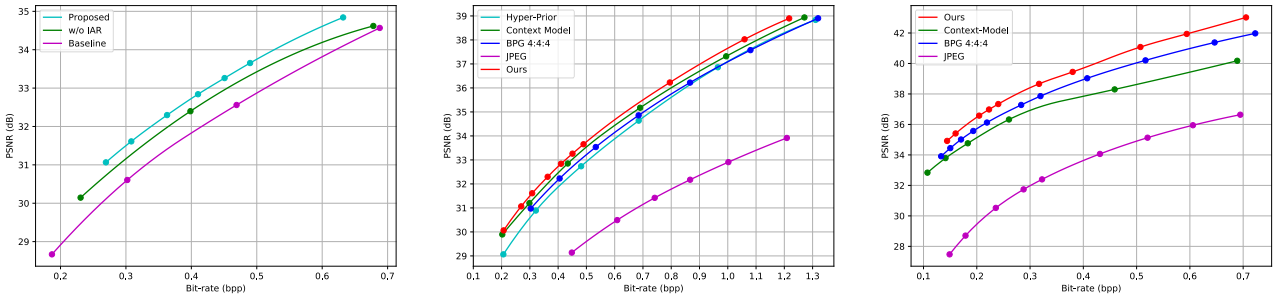


Figure 3: Estimation of the mean and scale value from decoded hyper latent representations with probability estimation subnetwork.

Model Training The network is trained jointly with rate-distortion optimization, controlled by the Lagrange parame-



(a) Ablation study on the proposed method. (b) Comparison on Kodak image set. (c) Comparison on Tecnick image set.

Figure 4: Quantitative evaluation results. (a) Ablation study on the proposed method. We set the hyper-prior model in (Ballé et al. 2018) as the baseline. The setting *w/o IAR* refers to the model with a three-layer structure but the Information Aggregation Reconstruction (IAR) sub-network is removed. (b, c) Comparison in R-D performance with learning-based methods, namely *Context-Model* (Lee, Cho, and Beack 2018) and *Hyper-Prior*, and image coding standards, namely *BPG* and *JPEG* on Kodak image set and Tecnick image set. Note that for the higher bit-rate range, the number of channels in our network is doubled.

ter λ to balance bit-rate and quality. The joint loss function is denoted as follows,

$$\mathcal{L}_{RD} = R + \lambda D, \tag{7}$$

where the distortion term can be any differentiable metric. The rate term is the sum of the cross entropy of each layer, as defined in Eq. (5). In the experiments, we train different models with different λ to evaluate the rate-distortion performance for various ranges of bit-rate.

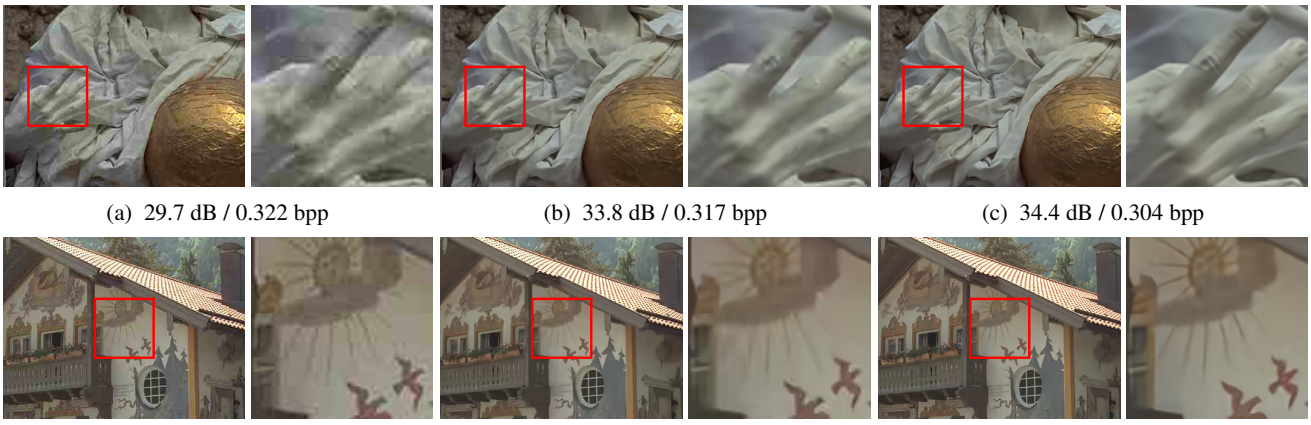
We train the network using the DIV2K dataset (Agustsson and Timofte 2017), which contains high-resolution images that have not been lossily compressed. We augment the dataset for training by additionally down-sampling the images with a scale factor 0.5. The network is trained in three stages. In the first stage, we pretrain the main analysis and synthesis transforms to achieve a good reconstruction of the images. The main analysis and synthesis transforms are pretrained with embedded context model. In the second stage, we randomly initialize the hyper transform network and train the whole model without context model in an end-to-end manner. Finally, the last layer of the original synthesis transform is replaced by the information aggregation network and the whole network is trained jointly.

We directly conduct rounding in the forward propagation, which simulates the same condition in the inference phase. As the rounding function has a gradient of zero almost everywhere, in the back-propagation, we override the derivatives of the rounding operation to be identical to the identity function $y = x$. When estimating the entropy of the representation in training, we add a uniform noise $U(-\frac{1}{2}, \frac{1}{2})$ to the original latent representation, inspired by (Ballé et al. 2018). We train the network using Adam (Kingma and Ba 2015) as the optimizer, with the initial learning rate set to 10^{-4}. In the final stage of training, we reduce the learning rate by 0.5 after every 100,000 iterations.

Experimental Results

We conduct experiments to compare the proposed method with existing learning-based and hybrid image compression methods. The experimental results include the Rate-Distortion performance and parallel acceleration analysis. We evaluate the Rate-Distortion (R-D) performance on the publicly available Kodak image set (Kodak 1993) and Tecnick SAMPLING image set (Asuni and Giachetti 2014). The Kodak image set consists of 24 lossless images of resolution 512×768. The Tecnick image set contains 40 images, with a fixed resolution 1200×1200, with which we are able to evaluate the compression performance on higher-resolution images. PSNR is used as the quality metric. We calculate bit-per-pixel (bpp) for each image and the corresponding bit-stream to show bit-rate. We reach different ranges of bit-rates by compressing images with different models, trained using different λ. The value of average PSNR and bpp are calculated across the equal value of λ or QP.

We first show an ablation study on the proposed framework in Fig. 4a. In this experimental setting, we investigate the effectiveness of the proposed Signal Preserving Hyper Transform and the Information Aggregation Reconstruction sub-network. We set the original two-layer Hyper-prior (Ballé et al. 2018) model as the baseline method. It is first observed that it is difficult to construct the multi-layer hyper-prior if the original synthesis and analysis transforms are employed. The additional hyper layer in this condition captures nearly no additional useful side-information to model the probability distribution of the upper layers. This might be due to the information loss during the forward propagation, caused by the ReLU activation and large convolutional kernels. With the proposed Signal Preserving Hyper Transform and the corresponding additional layer of hyper-prior, the network is shown to achieve a notable improvement in R-D performance, illustrated in Fig. 4a. The multi-layer structure is able to conduct in-depth analysis to better squeeze out redundancy which is neglected by the baseline model. A further improvement in performance is achieved when combining the multi-layer model with the Information Aggregation sub-network. It exploits the provided hyper representations to better reconstruct images and thus improves the quality of the decoded images.



(a) 29.7 dB / 0.322 bpp (b) 33.8 dB / 0.317 bpp (c) 34.4 dB / 0.304 bpp

(d) 27.7 dB / 0.654 bpp (e) 30.9 dB / 0.610 bpp (f) 31.3 dB / 0.610 bpp

JPEG BPG 4:4:4 Our Method

Figure 5: Comparison on visual quality with existing image compression standards (*i.e.* BPG and JPEG).

We also compare the proposed model with existing learning-based methods as well as hybrid codecs, illustrated in Fig. 4b and 4c. For traditional image codecs, *i.e* BPG and JPEG, we densely choose multiple quality parameters for the plots. BPG is the state-of-the-art hybrid image codec with intra prediction and Context-based Adaptive Binary Arithmetic Coding (CABAC). The evaluation of PSNR is conducted in RGB color space. As shown in Fig. 4b and 4c, the proposed method achieves better performance compared with BPG 4:4:4 on both Kodak and Tecnick dataset. Our model shows better performance over the state-of-the-art end-to-end method with context model and hyper-prior for entropy modeling, especially for images of higher resolution. Such images contain more long term spatial correlations, which can be well handled by the proposed coarse-to-fine structure.

In Fig. 5, we show visual results in the testing dataset to evaluate the visual quality of reconstructed images. It is observed that transform-based coding methods bring in aliasing and ringing effects, due to the loss of the high-frequency information, which harms visual quality a lot. Besides, due to the partitioning mechanism in the coding standards, blocking effects generally exist in the decoded images. The artifacts are annoying especially when the bit-rate is relatively low. The proposed end-to-end method, however, does not suffer from such effects as no partitioning and frequency related quantization is conducted in our proposed framework. We can see from Fig. 5 that JPEG produces severe blocking effects. BPG largely reduces such effects but artifacts still exist as it fails to naturally reconstruct some texture (Fig. 5a - 5c) and multi-directional edges (Fig. 5d - 5f). The high-frequency information loss on some blocks of the image degrades visual quality much, while the decoded images by the proposed method are more visually pleasing, even when achieving comparable quantitative results.

Additionally, we present the speedup *w.r.t.* GPU and CPU of the proposed network. We evaluate the network execution time on images of different sizes, *i.e.* median (1200×1200) and large (1512×2040) and calculate the acceleration ratio, as shown in Table 2. Note that due to the difference in handling floating point operations *w.r.t.* CPU and GPU, the evaluation of context-model (Lee, Cho, and Beack 2018) in the decoding time on GPU cannot be achieved. Nevertheless, the encoder end acceleration performance has demonstrated our claim and superiority.

Table 2: Acceleration ratio *w.r.t* CPU and GPU. We compare our method to the state-of-the-art context-model based method, on both the encoder (Enc) and decoder (Dec) on median (m) and large (l) images.

Method	Enc (m)	Dec (m)	Enc (l)	Dec (l)
Ours	130.5%	205.5%	163.3%	284.0%
Context-Model	117.4%	N/A	128.1%	N/A

Conclusion

In this paper, we introduce a coarse-to-fine hyper-prior guided auto-encoder for image compression. The framework is designed to decompose images into latent representations, where the elements in the latent representations can be more efficiently encoded (for the most inner layer) or conditionally modeled (for other supported layers). In this way, we better approximate the joint distribution of pixels in the to-be-encode image and allocate appropriate bits to represent the image in the bit-stream.

We propose the Signal Preserving Hyper Transform to construct the coarse-to-fine framework. By replacing the original tranforms with the proposed structure, we reduce information loss and it allows us to stack more layers of hyper representations, which therefore enhances the ability to better squeeze out spatial redundancy. It also facilitates the Information Aggregation reconstruction sub-network to fully exploit the bit-stream from all the layers to improve reconstruction quality. We achieve superior rate-distortion performance over BPG and existing learning-based methods.

References

Agustsson, E., and Timofte, R. 2017. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Agustsson, E.; Mentzer, F.; Tschannen, M.; Cavigelli, L.; Timofte, R.; Benini, L.; and Gool, L. V. 2017. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *Proc. Advances in Neural Information Processing Systems*.

Asuni, N., and Giachetti, A. 2014. TESTIMAGES: a large-scale archive for testing visual devices and basic image processing algorithms. In *Eurographics Italian Chapter Conference*, volume 1, 3.

Baig, M. H.; Koltun, V.; and Torresani, L. 2017. Learning to inpaint for image compression. In *Proc. Advances in Neural Information Processing Systems*.

Ballé, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018. Variational image compression with a scale hyperprior. In *Proc. International Conference on Learning Representations*.

Ballé, J.; Laparra, V.; and Simoncelli, E. P. 2017. End-to-end optimized image compression.

Ballé, J. 2018. Efficient nonlinear transforms for lossy image compression. In *Proc. Picture Coding Symposium*.

Bellard., F. 2014. BPG image format (http://bellard.org/bpg/). Accessed: 2019-11-18.

Johnston, N.; Vincent, D.; Minnen, D.; Covell, M.; Singh, S.; Chinen, T.; Jin Hwang, S.; Shor, J.; and Toderici, G. 2018. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. *Proc. of International Conference for Learning Representations*.

Kodak, E. 1993. Kodak lossless true color image suite (PhotoCD PCD0992). [online]. available: http://r0k.us/graphics/kodak/.

Lee, J.; Cho, S.; and Beack, S.-K. 2018. Context-adaptive entropy model for end-to-end optimized image compression. *arXiv preprint arXiv:1809.10452*.

Li, M.; Zuo, W.; Gu, S.; Zhao, D.; and Zhang, D. 2018. Learning convolutional networks for content-weighted image compression. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, H.; Chen, T.; Guo, P.; Shen, Q.; Cao, X.; Wang, Y.; and Ma, Z. 2019. Non-local attention optimized deep image compression. *arXiv preprint arXiv:1904.09757*.

Marpe, D.; Schwarz, H.; and Wiegand, T. 2003. 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.

Mentzer, F.; Agustsson, E.; Tschannen, M.; Timofte, R.; and Van Gool, L. 2018. Conditional probability models for deep image compression. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Minnen, D.; Ballé, J.; and Toderici, G. D. 2018. Joint autoregressive and hierarchical priors for learned image compression. In *Proc. Advances in Neural Information Processing Systems*.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Sullivan, G. J.; Ohm, J.-R.; Han, W.-J.; and Wiegand, T. 2012. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology* 22(12):1649–1668.

Theis, L.; Shi, W.; Cunningham, A.; and Huszár, F. 2017. Lossy image compression with compressive autoencoders.

Toderici, G.; O'Malley, S. M.; Hwang, S. J.; Vincent, D.; Minnen, D.; Baluja, S.; Covell, M.; and Sukthankar, R. 2016. Variable rate image compression with recurrent neural networks. In *Proc. International Conference on Learning Representations*.

Toderici, G.; Vincent, D.; Johnston, N.; Jin Hwang, S.; Minnen, D.; Shor, J.; and Covell, M. 2017. Full resolution image compression with recurrent neural networks. In *Proc, IEEE Conference on Computer Vision and Pattern Recognition*.

Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.; et al. 2016. Conditional image generation with pixelcnn decoders. In *Proc. Advances in neural information processing systems*.

Wallace, G. K. 1992. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics* 38(1):xviii–xxxiv.