Causal Contextual Prediction for Learned Image Compression

Zongyu Guo, Zhizheng Zhang, Runsen Feng and Zhibo Chen, *Senior Member, IEEE,*

***Abstract*— Over the past several years, we have witnessed impressive progress in the field of learned image compression. Recent learned image codecs are commonly based on autoencoders, that first encode an image into low-dimensional latent representations and then decode them for reconstruction purposes. To capture spatial dependencies in the latent space, prior works exploit hyperprior and spatial context model to build an entropy model, which estimates the bit-rate for end-to-end rate-distortion optimization. However, such an entropy model is suboptimal from two aspects: (1) It fails to capture global-scope spatial correlations among the latents. (2) Cross-channel relationships of the latents remain unexplored. In this paper, we propose the concept of separate entropy coding to leverage a serial decoding process for causal contextual entropy prediction in the latent space. A *causal context model* is proposed that separates the latents across channels and makes use of channel-wise relationships to generate highly informative adjacent contexts. Furthermore, we propose a *causal global prediction model* to find global reference points for accurate predictions of undecoded points. Both these two models facilitate entropy estimation without the transmission of overhead. In addition, we further adopt a new group-separated attention module to build more powerful transform networks. Experimental results demonstrate that our full image compression model outperforms standard VVC/H.266 codec on Kodak dataset in terms of both PSNR and MS-SSIM, yielding the state-of-the-art rate-distortion performance.**

***Index Terms*—Learned image compression, causal context model, causal global prediction, improved entropy model.**

I. Introduction

Lossy image compression is a fundamental technique for image transmission and storage. Since the concept of hybrid coding was proposed [1], [2], the hybrid coding framework has shown strong vitality in the field of media compression. This coding framework, combining prediction and transformation, has been improved for decades, not only to achieve better performance but also to meet the rising demand of novel multimedia applications. Since the standardization of H.264/AVC in 2003 [3], the technique of intra prediction has become an important component of image compression. Researchers have been developing increasingly complex intra prediction methods for decades, expecting more accurate predictions and thus more efficient compression. Intra prediction basically tries to capture the spatial correlations of the image to reduce the spatial redundancies in the compressed bitstream.

Over the past few years, the progress of learning-based image compression is impressive [4], [5], [6], [7]. Despite the short history of this field, end-to-end optimized compression models have shown the potential to outperform traditional manually designed codecs. Learned image compression commonly follows a pipeline consisting of transformation, quantization and lossless entropy coding, which resembles traditional transform coding [8]. Specifically, a non-linear transform is first performed on the input image, mapping it to latent representations. Then the latents are quantized, yielding discrete representations. To losslessly compress the discrete latents, an entropy model is used to estimate their discrete entropy [9]; this model is later improved as a hierarchical entropy model, *i.e.*, a hyperprior model [5]. More recently, several studies introduce the concept of context model [10], [11], which is an autoregressive model over latents. Such a context model usually employs mask convolution [12] to aggregate local contexts for efficient entropy coding.

Worthy of mention is that the non-linear transform and entropy model play different roles for compression although they both can decorrelate natural images. While the non-linear transform disentangles the image into compact latent representations, the entropy model exploits a probabilistic structure to establish an accurate estimation for bit-rate and eventually leads to the generation of smaller files [5], [10], [11]. As a part of the entropy model, the autoregressive context model [10], [11] is more like a *prediction* module, which leverages adjacent available latents to predict undecoded points. Such an autoregressive model incurs a significant computational penalty during decoding due to the serial decoding process, but it largely improves the rate-distortion performance.

However, the entropy model, which incorporates adjacent contexts and hierarchical priors, is suboptimal in terms of two aspects. On the one hand, adjacent context modeling is not allowed to capture global correlations of the latents which are beneficial for sufficient information decorrelation. Some previous works [13], [14] present spatial-adaptive encoders/decoders, the transformation of which is adaptive to global image contents. Despite this, the exploitation of global contexts within the entropy model is still ignored. In theory, a pair of image patches with similar context should be transformed into similar latent representations, even if they are spatially far away from each other (see Fig.1). Thus, going beyond the context modeling of adjacent contexts towards a global-scope context modeling is a theoretically powerful modification to deliver more accurate entropy estimation. However, a critical challenge is imposed when performing global-scope context modeling in the entropy model. That

Zongyu Guo, Zhizheng Zhang, Runsen Feng and Zhibo Chen are with the Department of Electronic Engineer and Information Science, University of Science and Technology of China, Hefei, Anhui, 230026, China. Corresponding Author: Zhibo Chen (chenzhibo@ustc.edu.cn).

This work was supported in part by NSFC under Grant U1908209, 61632001 and the National Key Research and Development Program of China 2018AAA0101400.

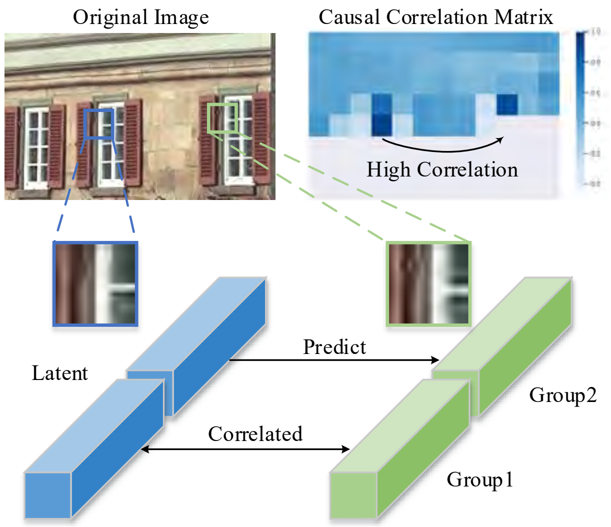
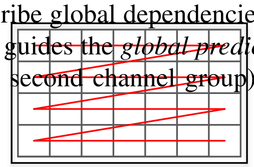
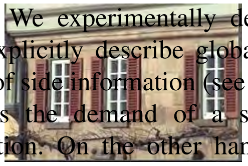


Fig. 1: An overview of our proposed causal global prediction model. A correlation matrix is calculated from the first channel group to describe global dependencies. This correlation matrix is *causal* and guides the *global prediction* for these undecoded channels (the second channel group).





is, *how can global correlation information be established at the decoder?* We experimentally demonstrate that it is not valuable to explicitly describe global references through the transmission of side information (see Section III-B for details), which implies the demand of a special design to realize global prediction. On the other hand, in addition to spatial redundancies, cross-channel relationships remain unexplored. In fact, after learning-based transformation, information located spatially is embedded into channel dimension. Therefore, cross-channel relationship modeling is meaningful as well. Compared with traditional compression, this is a special issue for neural image compression because the latent space is now 3-D. An optimal entropy model should be able to eliminate channel-wise redundancies for improved entropy estimation.

In this paper, to address the two abovementioned issues, we propose a causal context model and a causal global prediction model for more accurate entropy estimation. Experimentally, we find that the spatial correlations of the entire latents can be approximated by partial channels, indicating that there are redundancies among different channels. We thus design a causal context model with an improved mask convolution, where we separate the latents into two groups. Subsequently, we propose a novel causal global prediction model to reduce global redundancies among the latent variables. Specifically, once the first latent group is decoded (Group 1 in Fig.1), the proposed causal context model will aggregate both the spatially adjacent latents and the first half-channel elements in the current spatial location to yield highly informative adjacent contexts. In addition. the proposed causal global prediction model will leverage the decoded group to model global correlations, establishing a causal correlation matrix that guides the global predictions for the remaining undecoded channels (Group 2 in Fig.1). Different from previous hyperprior model, the proposed causal global prediction model does not require the transmission of overhead to describe global correlations.

Through causal contextual prediction within the entropy model, the contexts are embedded into two separate channel groups, where one latent group is directly inferred as usual while the other group is causally inferred upon the condition of the prior one. In addition, we adopt a new group-separated attention module to strengthen the representation ability of the transform networks. Motivated by [15], [16], this module enables independent feature-map attention across separated groups and is demonstrated to be more powerful than the attention mechanism used in previous compression method [7]. Our technical contributions can be summarized as follows:

- We propose the concept of separate entropy coding by dividing the latent representation into two channel groups for more effective context modeling. We thereby propose a causal context model that makes use of cross-channel redundancies to generate highly informative adjacent contexts.
- We pinpoint that exploiting global-scope contexts is vital for the entropy modeling, and propose a causal global prediction model to conduct prediction in a global scope. By extending the concept of separate entropy coding, this model does not require extra transmission of overhead but can still establish global reference information.
- We adopt group-separated attention module to strengthen the non-linear transform networks, which is demonstrated to be more powerful than previous attention designs. We take this point as a side contribution in this paper.

We integrate the proposed causal context model and causal global prediction model, along with the group-separated attention module, building our learned image compression network. We verify the effectiveness of each component with sufficient ablation studies. Experimental results demonstrate that our approach outperforms previous learned image compression schemes [10], [7] in terms of both PSNR and MS-SSIM [17]. Moreover, compared with traditional codec VTM 8.0[1], our method achieves 5.1% BD rate savings on the Kodak dataset[2] in terms of PSNR.

Our work described in this paper is an extension of our pioneering short conference paper [18]. In [18], we introduce the causal context model[3], which is the first instantiation of the concept of separate entropy coding. The differences between this paper and [18] are described as follows. Firstly, here we extend the concept of separate entropy coding to propose a novel causal global prediction model that successfully reduces global redundancies of the latents and achieves more efficient entropy estimation. Secondly, we provide detailed analyses to explain the design of separate entropy coding, as illustrated in Section IV-D. Thirdly, an improved attention module is employed to enhance the non-linear transform networks. Fourthly, we conduct extensive ablation studies that verify the effectiveness of our proposed new techniques. We also evaluate the coding time of our method.

[1] https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-8.0

[2] http://r0k.us/graphics/kodak/

[3] In [18], it is termed 3-D context entropy model. We here term it as causal context model to avoid ambiguity with 3-D mask convolution.

The remainder of this paper is organized as follows. Section II briefly overviews some relevant approaches. Section III illustrates the motivation of global prediction in the entropy model. Section IV introduces our proposed causal context model and causal global prediction model along with the proposed group-separated attention layer. Section V is the experimental section. We conclude this paper in the last section.

II. Related Work

A. Traditional Image Compression

Traditional image compression standards, such as JPEG [19], JPEG2000 [20] and BPG (HEVC intra) [4], are widely used in practice. They heavily rely on manually designed modules but always follow the concept of transform coding [8]. Since the standardization of H.264/AVC [3], intra prediction has become an important component of image compression. In the case of H.264, there are a total of 9 intra prediction modes. With regard to H.265/HEVC [21], 35 optional intra prediction modes are adopted for different sizes of prediction units (PUs). The ongoing H.266/VVC [22] further applies 65 intra modes. Intra prediction plays an important role in reducing the spatial redundancies of images in hybrid compression framework.

Additionally, it is worthwhile to note the intra block prediction (IBC) technique, which was first standardized in the screen content coding extensions [23] of HEVC. Intra block compensation is observed to be effective for screen contents. Therefore, IBC utilizes a bit-consuming displacement vector (referred as block vector or BV) to represent the relative displacement between the current prediction unit and the reference block. Such classical technique can conduct global searching for more efficient intra prediction and it is similar to the global prediction idea in this paper. However, our solution for global prediction of the latents does not require the transmission of prediction vectors.

B. Learned Image Compression

1) Framework: In the field of learned image compression, researchers initially study the problem of non-differentiable quantization in artificial neural network [24], [25]. The first learned compression method to outperform JPEG is an RNN-based model [4]. This framework also supports coding scalability [26], [27] but is only optimized for distortion. The work of [9] attempts to estimate the *rate* as discrete entropy. These works make it feasible to train an end-to-end compression network optimized for the rate-distortion trade-off. In [5], a hyperprior model is proposed that transmits extra side information to model the spatial structure of the latents. This work also parameterizes the distribution of the latents to a zero-mean Gaussian scale model (GSM). After that, the parametric form of latent distribution is improved from Single Gaussian Model (SGM) [10] to Gaussian Mixture Model (GMM) [7]. In addition, motivated by PixelCNN [12], the concept of context model is introduced for efficient entropy estimation [10], [11]. However, this type of model only covers local regions but does not pay attention to global scope.

2) Global Context: An early work [14] uses a spatially adaptive post-process to dynamically adjust bit rate according to a target reconstruction quality. Later, both [13] and [28] use an importance map to for more adaptive bit rate allocation. All of those works [14], [13], [28] incorporate image content to affect compression, but they do not have explicit entropy model. Recently, Lee et al. [29] also explore to utilize global context within the entropy model. However, they are unable to establish accurate global references at the decoder side. Aggregating possible global contexts is hard to determine accurate global reference information, thereby leading to the suboptimal performance of their entropy model.

3) Cross-Channel Relationship: Previous works [28], [30], [31] also consider the relationships between latent channels. One early work [28] does not have a parametric entropy model which limits their performance. Unlike separate entropy coding in our paper, Chen et al. [30] adopt a channel-autoregressive 3-D mask convolution that slides over the latent representations across channels. In their method, every latent element is conditioned on adjacent decoded elements that are spatial-channel neighbors. Therefore, when applying a $5\times5\times5$ mask convolution, they could only learn correlations from adjacent two channels. Li et al. [31] propose a similar idea to ours but their entropy model is also autoregressive along channels, which complicates the practical entropy coding and thus increases the time complexity. Instead of channel-autoregressive entropy coding, here we demonstrate that separating the latents into two groups can help network adaptively learn appropriate context, which is simple but effective.

III. Problem Definition

A. Learned Compression Framework

In the framework of learned image compression, a natural image, denoted by $\boldsymbol{x}$, is first encoded as latent representations $\boldsymbol{y}$ through an analysis transform $g_a(\boldsymbol{x}|\boldsymbol{\phi})$. Then the latents $\boldsymbol{y}$ are quantized to discrete values $\hat{\boldsymbol{y}}$, which will be losslessly coded by algorithms such as arithmetic coding. At the decoder side, a synthesis transform $g_s(\hat{\boldsymbol{y}}|\boldsymbol{\theta})$ recovers $\hat{\boldsymbol{y}}$ to reconstruct an image $\hat{\boldsymbol{x}}$.

$$\begin{aligned} \boldsymbol{y} &= g_a(\boldsymbol{x}|\boldsymbol{\phi}), \\ \hat{\boldsymbol{y}} &= Q(\boldsymbol{y}), \\ \hat{\boldsymbol{x}} &= g_s(\hat{\boldsymbol{y}}|\boldsymbol{\theta}). \end{aligned} \tag{1}$$

If taking entropy model into account, the above neural compression schemes can be intepreted as variational autoencoders (VAEs) [32]. In particular, the rate-distortion objective of end-to-end compression has a strong relationship with the loss function of β-VAE [33], where there is a hyperparameter balancing the latent channel capacity with reconstruction quality:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}[-\log_2 p_{\hat{\boldsymbol{y}}}(\hat{\boldsymbol{y}})] + \lambda \cdot \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}[d(\boldsymbol{x}, \hat{\boldsymbol{x}})]. \tag{2}$$

The first term is the rate that corresponds to the cross entropy between the natural (marginal) distribution and the learned entropy model. The second term measures the reconstruction quality according to the given distortion metric d (e.g., PSNR or MS-SSIM). Imposing a variant constraint by adjusting λ influences the disentangling effect of the non-lienar transform

[4] https://bellard.org/bpg/

[33] and determines the model rate. In the work of [5], to reduce the spatial redundancies among latent variables, a hyperprior model is proposed, which assigns a few extra bits as side information to transmit some spatial structure information. Such a hyperprior model helps to learn an accurate entropy model and thereby achieves a better estimation of $p_{\hat{\boldsymbol{y}}}(\hat{\boldsymbol{y}})$. This hyper model can be roughly divided into a hyper analysis transform $h_a(\boldsymbol{y}|\boldsymbol{\phi_h})$ and a synthesis transform $h_s(\hat{\boldsymbol{z}}|\boldsymbol{\theta_h})$ as

$$\begin{aligned} \boldsymbol{z} &= h_a(\boldsymbol{y}|\boldsymbol{\phi_h}), \\ \hat{\boldsymbol{z}} &= Q(\boldsymbol{z}), \\ P_{\hat{y}|\hat{z}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}) &\leftarrow h_s(\hat{\boldsymbol{z}}|\boldsymbol{\theta_h}). \end{aligned} \tag{3}$$

However, such spatial structure information is inaccurate because the side information is transmitted after several downsampling layers, leading to geometric information loss. To capture the adjacent correlations of $\hat{\boldsymbol{y}}$, a context model f_c [10], [11] is equipped as a component of hyperprior model. It is usually in the form of mask convolution with parameter $\boldsymbol{\theta_c}$. At the expense of decoding speed, we now decode all points serially, *e.g.*, in raster scan order. Then, the hyperprior entropy model can be depicted as follows:

$$\begin{aligned} \boldsymbol{c_n} &= f_c(\hat{\boldsymbol{y}}_{\boldsymbol{i_1}}, \hat{\boldsymbol{y}}_{\boldsymbol{i_2}}, ..., \hat{\boldsymbol{y}}_{\boldsymbol{i_k}}|\boldsymbol{\theta_c}), \\ P_{\hat{y}|\hat{z}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}) &\leftarrow h_s(\hat{\boldsymbol{z}}, \boldsymbol{c_n}|\boldsymbol{\theta_h}), \end{aligned} \tag{4}$$

where $\boldsymbol{i_1, i_2, ..., i_k}$ indicate the indices of the subset of decoded points for aggregating adjacent context $\boldsymbol{c_n}$. Since at any point the decoder can only access the already decoded latents, the local context model only touches the left and top points which are adjacent to the current decoding point.

B. Global Prediction

Basically, the idea of global prediction aims at further leveraging all previously decoded latents to predict the current decoding point, and this can be formulated as

$$\begin{aligned} \boldsymbol{c_n} &= f_c(\hat{\boldsymbol{y}}_{\boldsymbol{1}}, \hat{\boldsymbol{y}}_{\boldsymbol{2}}, ..., \hat{\boldsymbol{y}}_{\boldsymbol{n-1}}|\boldsymbol{\theta_c}), \\ P_{\hat{y}|\hat{z}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}) &\leftarrow h_s(\hat{\boldsymbol{z}}, \boldsymbol{c_n}|\boldsymbol{\theta_h}), \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{1}}, \hat{\boldsymbol{y}}_{\boldsymbol{2}}, ..., \hat{\boldsymbol{y}}_{\boldsymbol{n-1}}$ represent all points that are already decoded. According to Fig.1, we observe that global spatial redundancy remains in the latents. We herein conduct an exploratory experiment to verify that such global redundancies heavily influence the performance of the entropy model.

A common way to describe global correlation is to calculate the similarity between all points. However, in the task of compression where decoding is serial, we are only interested in the similarities[5] between all previously decoded points and the current decoding point. Therefore, we calculate *causal* correlation matrices (see Fig.1) that help us search global reference points. We should keep in mind that the correlation matrix is not directly available at the decoder side. The decoder cannot compute the similarity between other points and the current decoding point before it decodes out the specific feature vector in the current decoding location.

Model	Anchor	Case 1	Case 2
Rate (bpp)	0.1588	0.1435	0.1799
PSNR (dB)	28.53	28.54	28.56

TABLE I: Making use of global reference information is advantageous for entropy estimation. In case 1, the transmission cost of global reference information is not considered and thus case 1 is not a practical codec. In case 2 we assign additional bits to transmit the coordinates of reference points in advance.

We start with an ablation experiment to study the effect of global context on entropy model, where the baseline network already adopts hierarchical priors and local context model [10]. As shown in Table I, in both case 1 and case 2, the decoder is assumed to be able to make use of the coordinates of global reference points for more efficient entropy estimation. However, in case 1, we do not calculate the transmission cost of the global reference information, while in case 2, we assign additional bits to transmit the coordinates of reference points in advance[6]. During the training stage in case 2, we modify the loss function to include the transmission overhead. Comparing case 1 and case 2 with the baseline model, we can conclude that incorporating global references into entropy model significantly improves entropy estimation. Specifically, the required bit-rate decreases from 0.1588 to 0.1435 bit per pixel (bpp) while the reconstruction quality is maintained as we that can observe from case 1.

The above experiments demonstrate the advantages of incorporating accurate global context into the entropy model. However, it also reveals a critical problem: *how to efficiently establish accurate global correlation information at the decoder side?* In this paper, we propose a causal global prediction model that does not require transmission of overhead but can still conduct global prediction. It is based on the causal context model, which aims at generating highly informative adjacent context via latent separation.

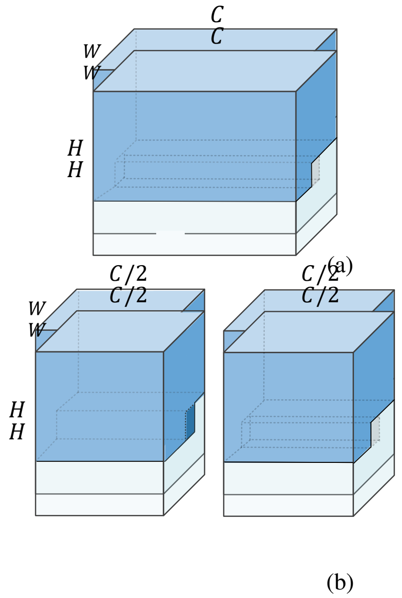
IV. Proposed Method

A. Causal Context Model

Inspired by the conditional RGB prediction of PixelCNN++ [34], we propose a causal context model to improve entropy estimation, as described in our pioneer conference paper [18]. The causal context model separates the latents $\hat{\boldsymbol{y}}_{\boldsymbol{n}}$ in spatial location n into two groups across channels. When it comes to decode $\hat{\boldsymbol{y}}_{\boldsymbol{n}}$, the distribution of the first channel group $\hat{\boldsymbol{y}}_{\boldsymbol{n,1}}$ is predicted as usual, which adopts a mask convolution $f_{c,1}$ (as in Fig.2a) to generate adjacent context $\boldsymbol{c_{n,1}}$ and combines with hyperpriors $\hat{\boldsymbol{z}}$. Thanks to the serial decoding process, once the first group $\hat{\boldsymbol{y}}_{\boldsymbol{n,1}}$ is decoded, we can take this known group to estimate the second group $\hat{\boldsymbol{y}}_{\boldsymbol{n,2}}$. By modifying the conventional mask convolution, the improved mask convolution $f_{c,2}$, which is shown in Fig.2b, is able to aggregate both the adjacent decoded latents and the first half latents in current spatial location, generating more informative adjacent contexts $\boldsymbol{c_{n,2}}$. The improved mask convolution can be regarded as a mutation between mask convolution and causal

[5] Here, we compute the cosine distance to measure similarity.

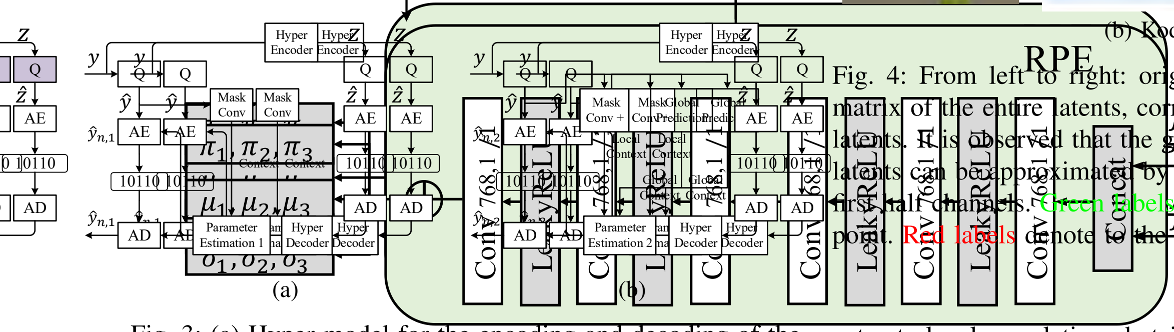
[6] Since coordinates are integers as well, we use another hierarchical entropy model to estimate the rate of reference coordinates.



(a)

(b)

Fig. 2: (a) The structure of conventional mask convolution [12]. (b) The improved mask convolution in our proposed causal context model (denoted by Mask Conv +).



(a)

(b)

Fig. 3: (a) Hyper model for the encoding and decoding of the first group $\hat{\boldsymbol{y}}_{\boldsymbol{n,1}}$. (b) Incorporation of the adjacent and global context for the second group $\hat{\boldsymbol{y}}_{\boldsymbol{n,2}}$.

convolution [35]. It leverages the serial decoding process to generate causal contexts. The whole process can be formulated as follows:

$$\begin{aligned} \boldsymbol{c_{n,1}} &= f_{c,1}(\hat{\boldsymbol{y}}_{\boldsymbol{i_1}}, \hat{\boldsymbol{y}}_{\boldsymbol{i_2}}, ..., \hat{\boldsymbol{y}}_{\boldsymbol{i_k}}), \\ P_{\hat{\boldsymbol{y}}_{\boldsymbol{n,1}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}_{\boldsymbol{n,1}}|\hat{\boldsymbol{z}}) &\leftarrow h_{s,1}(\hat{\boldsymbol{z}}, \boldsymbol{c_{n,1}}|\boldsymbol{\theta_h}), \\ \boldsymbol{c_{n,2}} &= f_{c,2}(\hat{\boldsymbol{y}}_{\boldsymbol{i_1}}, \hat{\boldsymbol{y}}_{\boldsymbol{i_2}}, ..., \hat{\boldsymbol{y}}_{\boldsymbol{i_k}}, \hat{\boldsymbol{y}}_{\boldsymbol{n,1}}), \\ P_{\hat{\boldsymbol{y}}_{\boldsymbol{n,2}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}_{\boldsymbol{n,2}}|\hat{\boldsymbol{z}}) &\leftarrow h_{s,2}(\hat{\boldsymbol{z}}, \boldsymbol{c_{n,2}}|\boldsymbol{\theta_h}). \end{aligned} \quad (6)$$

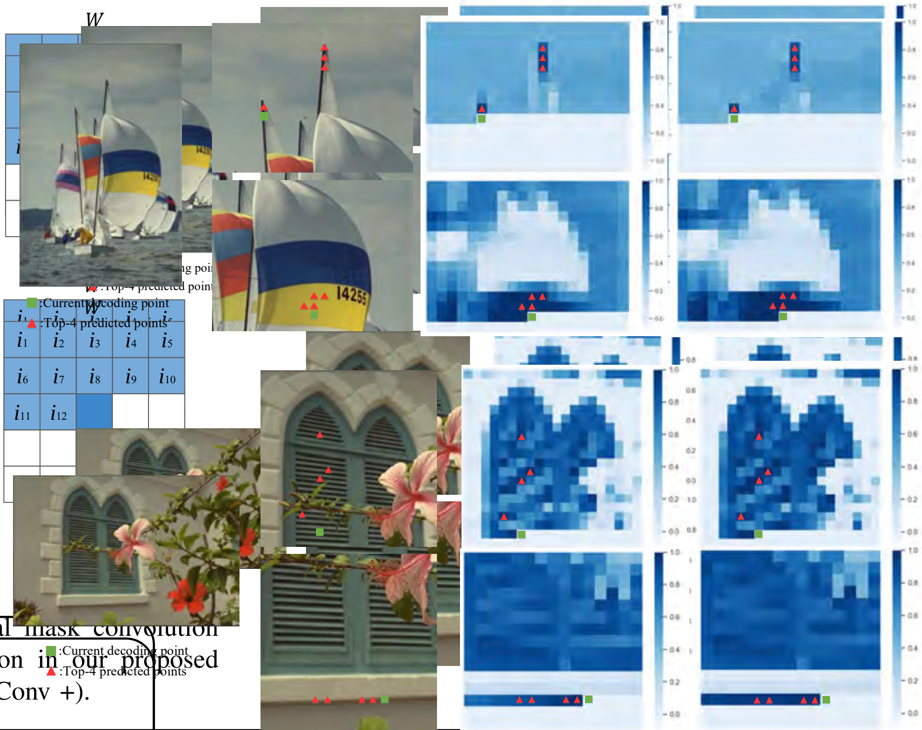
This causal context model can effectively extract the cross-channel relationship to facilitate entropy estimation of the second channel group. Note that the concept of separate entropy coding actually divides the decoding of $\hat{y}_n$ into two stages. At different stages, the network parameters are independent. As a result, the following parameter estimation modules $h_{s,1}$ and $h_{s,2}$ do not share weights.

B. Causal Global Prediction Model

The above causal context model is helpful for capturing both the spatial and channel redundancies. However, it only extracts local correlations but ignores global correlations. In this section, we improve it and propose a *causal global prediction model* to utilize the long-range correlations of the latents. The proposed causal global prediction model separates the latents $\hat{\boldsymbol{y}}$ into two groups as well. The compression process of the first channel group $\hat{\boldsymbol{y}}_{\boldsymbol{n,1}}$ is unchanged, only utilizing hyperprior and local context model, as illustrated in Fig.3a. However, to estimate the distribution of the second latent group, the entropy model now incorporates both the improved adjacent context $\boldsymbol{c_{n,2}}$ and the global context $\boldsymbol{c_{n,3}}$ together as

$$\begin{aligned} \boldsymbol{c_{n,1}} &= f_{c,1}(\hat{\boldsymbol{y}}_{\boldsymbol{i_1}}, \hat{\boldsymbol{y}}_{\boldsymbol{i_2}}, ..., \hat{\boldsymbol{y}}_{\boldsymbol{i_k}}), \\ P_{\hat{\boldsymbol{y}}_{\boldsymbol{n,1}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}_{\boldsymbol{n,1}}|\hat{\boldsymbol{z}}) &\leftarrow h_{s,1}(\hat{\boldsymbol{z}}, \boldsymbol{c_{n,1}}|\boldsymbol{\theta_h}), \\ \boldsymbol{c_{n,2}} &= f_{c,2}(\hat{\boldsymbol{y}}_{\boldsymbol{i_1}}, \hat{\boldsymbol{y}}_{\boldsymbol{i_2}}, ..., \hat{\boldsymbol{y}}_{\boldsymbol{i_k}}, \hat{\boldsymbol{y}}_{\boldsymbol{n,1}}), \\ \boldsymbol{c_{n,3}} &= f_{c,3}(\hat{\boldsymbol{y}}_{\boldsymbol{1}}, \hat{\boldsymbol{y}}_{\boldsymbol{2}}, ..., \hat{\boldsymbol{y}}_{\boldsymbol{n-1}}, \hat{\boldsymbol{y}}_{\boldsymbol{n,1}}), \\ P_{\hat{\boldsymbol{y}}_{\boldsymbol{n,2}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}_{\boldsymbol{n,2}}|\hat{\boldsymbol{z}}) &\leftarrow h_{s,2}(\hat{\boldsymbol{z}}, \boldsymbol{c_{n,2}}, \boldsymbol{c_{n,3}}|\boldsymbol{\theta_h}), \end{aligned} \quad (7)$$

where the improved adjacent context $\boldsymbol{c_{n,2}}$ is generated by the causal context model $f_{c,2}$ and global context $\boldsymbol{c_{n,3}}$ is generated by the causal global prediction model $f_{c,3}$. The causal global prediction model enables the decoder to establish accurate global reference information via latent separation, which will be introduced in the next section. Fig.3b presents the diagram of the entropy coding process for the second group.



(b) Kodim07

Fig. 4: From left to right: original image patch, correlation matrix of the entire latents, correlation matrix of the first half latents. It is observed that the global correlations of the entire latents can be approximated by the correlation matrices of the first half channels. Green labels represent the curent decoding point. Red labels denote to the four selected reference points.

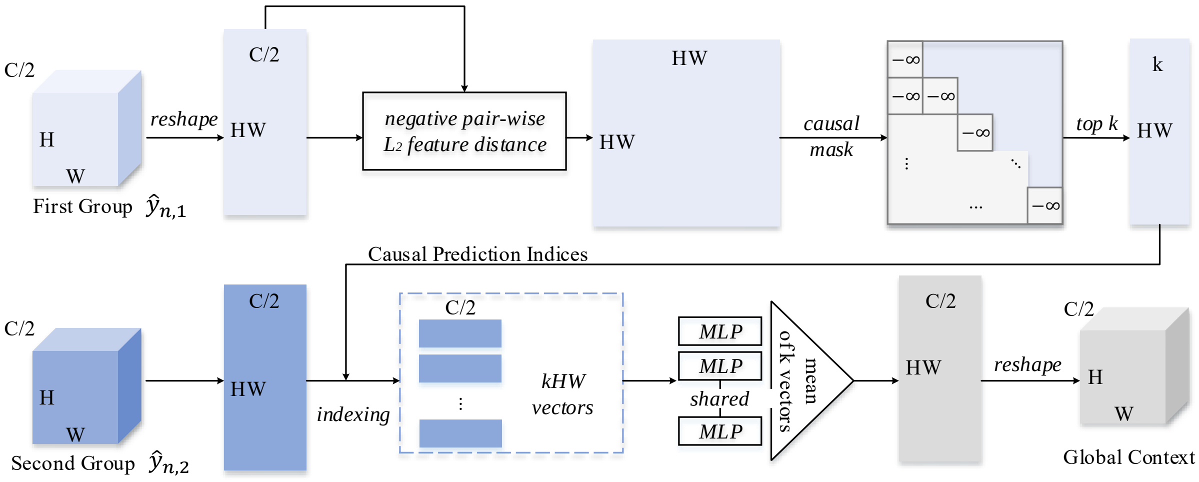
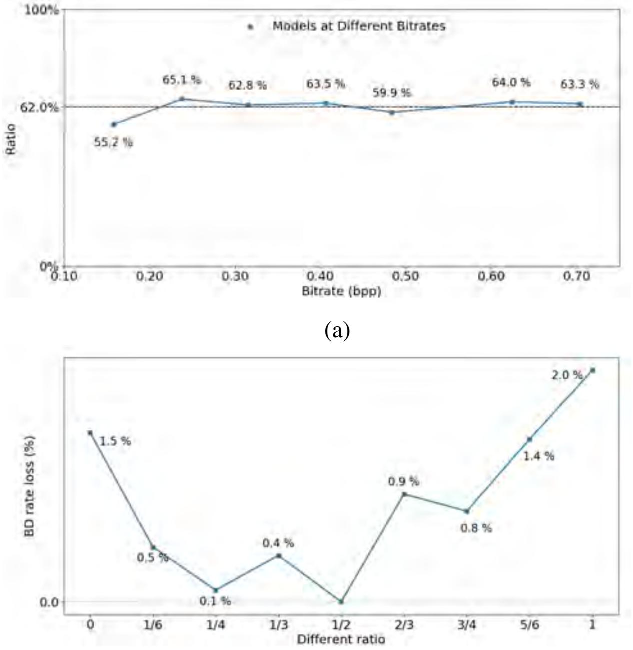


Fig. 5: Global prediction via latent separation.

C. Global Prediction via Latent Separation

The previous discussions in Section III-B show the potential of global context for entropy estimation under the premise that the decoder can establish accurate global reference information. However, it is very bit-consuming to explicitly transmit the coordinates of relevant points. As shown in Fig. 4, we find that the similarities among the half latent vectors match the similarities among the entire latent vectors. It implies that the spatial correlations of the entire latent variables can be approximated by the correlation matrices that are calculated from the first half channels. This observation motivates us to generate bit-free global correlations at the decoder side by making use of the concept of latent separation. The latent vector in each spatial location can roughly represent the context in a 16×16 area of original image because the encoding transform network downsamples natural image for four times with stride=2. Therefore, there is a close relationship between the latent correlation matrix and the image context.

Fig.5 illustrates the process of separate global prediction. After decoding the first latent group $\hat{\boldsymbol{y}}_{n,1}$, we then calculate the negative pair-wise $\mathcal{L}_2$ distances among all half vectors, yielding a $HW \times HW$ correlation matrix. Since the decoding process is *causal*, the correlation matrix is masked to be an upper triangular matrix. To avoid occasional inaccurate predictions, we preserve only the top-k correlated points and gather their indices to generate the causal prediction indices. These indices, denoting the k most similar points regarding the current decoding point, are used for indexing and selecting appropriate features in the second group. After indexing, we have k vectors for each point, which will be utilized to predict current decoding point. The dimension of each prediction vector is $1 \times 1 \times c/2$. Considering that the latents have a total of HW points, we now have kHW vectors as predicted information. Since current prediction vectors are directly gathered from previously decoded points, we then send these vectors into a shared multi-layer perceptron (MLP), which is a common process for learning effective representations [36]. We finally average the output of MLP to generate global context. In our experiments, we use top-4 predictions to generate global context which will be investigated in the following analysis. We also visualize some examples of top-4 predictions in Fig.4.



(a)

(b)

Fig. 6: (a) If separating the latents into two groups on average, the first group is representative that occupies a higher percentage in bitstream. (b) Different separation ratios affect the BD rate performance. We evaluate different methods on the whole Kodak dataset. The baseline here is the simplified model [10] deployed with our proposed separate entropy coding modules.

D. Analysis

1) Why separating the latents into two groups on average: The network is expected to adaptively arrange appropriate channels in different groups during training. As shown in

Fig.6a, when the first and second groups have the same number of channels, the first group accounts for around 62% of the bitstream at different bitrates. This statistic demonstrates that to accurately model global correlations, the proposed latent-separated models are more likely to assign those representative and bit-consuming channels to the first group. We also explore to adjust the ratio of channel numbers in different groups. The results are shown in Fig.6b. In this figure, the left-most point represents the case of a small ratio value where we assign two channels in the first group to calculate correlation matrix. These two channels in the first group are informative and representative of image contexts to guide the prediction of the second channel group. And ratio=1 is another extreme case that makes the whole context model degrade to conventional 2-D context model [10] without channel separation. It is observed that the model achieves good results when the first channel group occupies 1/6∼1/2 channels. Too small or too large ratio degrades compression performance since the concept of separate entropy coding requires some channels in the first group to ease the compression of the rest channels in the second group. In particular, ratio≈0 performs better than ratio=1 slightly since the former can still utilize coarse correlation information to achieve more effective compression.

In addition, it is reasonable to simply separate the latent channels into two groups. On the one hand, the parameter estimation modules in different groups do not share weights. On the other hand, the current two-group separation already divides the decoding period into two stages and thus increases time complexity. Dividing the latent channels into more groups would further complicate the decoding in terms of both space complexity and time complexity.

2) Difference between global context and local context: The global context should be distinguished from the local context. Local context [10], [11] is extracted by a mask convolution layer, *aggregating* possible context without performing explicit prediction. But here the causal global prediction model searches the set of all decoded points, selecting the top-k correlated points. These selected points will then take their second channel group as references to *predict* the undecoded channel group in current decoding point.

3) Difference between global context and hyperprior model: The generated global context is also different from hierarchical priors. Both of them take advantage of the global structures of the latents to improve entropy estimation. However, the hyperprior model transmits additional overhead, i.e., side information, to model rough global structures with downsampling and upsampling layers. Our causal global prediction model does not transmit overhead but still enables effective global prediction with all accessible points. In terms of compression performance, global context is complementary to hierarchical priors, even if they only work for the second channel group. However, similar to local context model [10], latent-separated global prediction results in a causality problem for decoding that increases decoding complexity.

4) The influence of the value of k: The proposed causal global prediction model selects the k most relevant points as references to predict the entropy of undecoded channel group. It is important to determine an appropriate number of reference points because it directly influences the prediction performance. Actually, in traditional compression standards such as H.264 [3], different prediction modes also involve different numbers of reference points. For example, vertical prediction mode and horizontal prediction mode directly copy one adjacent pixel. Some diagonal prediction modes would involve two relevant adjacent pixels. For DC prediction mode and planar mode [37], they operate on more adjacent pixels. Here, we explore the effects of the number of reference points on our learned image compression method.

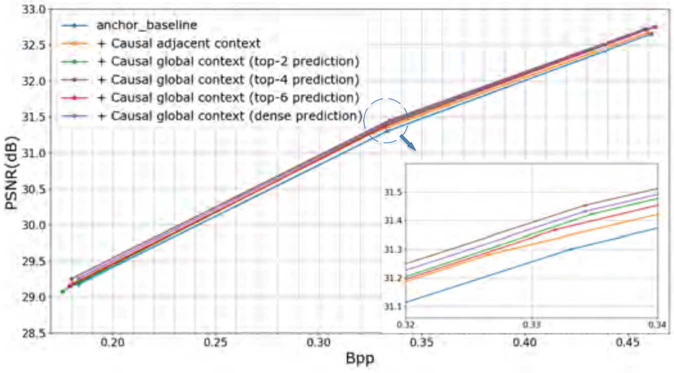
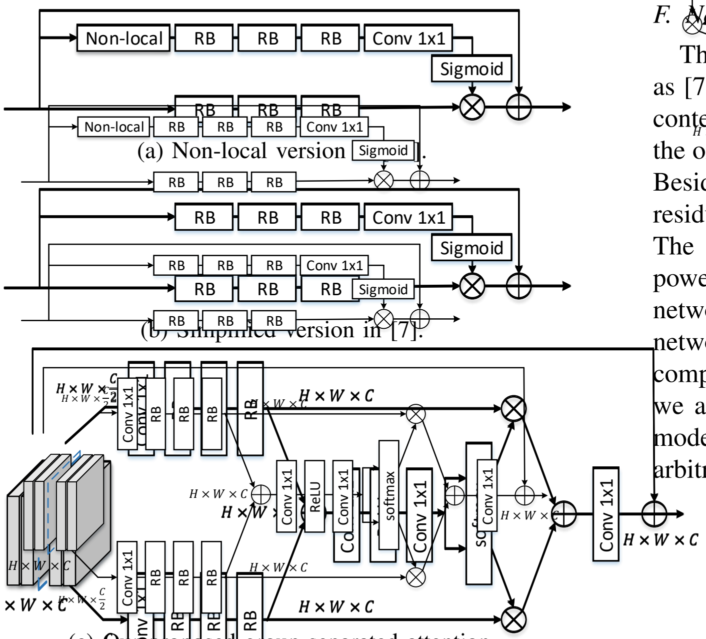


Fig. 7: Exploring the effects of different k values in our proposed causal global prediction model.

As shown in Fig.7, we compare different k values, where baseline is [10]. In this figure, dense prediction refers to using all decoded points to predict the current point while top-$2, 4, 6$ prediction only involves using several relevant points for prediction. It is found that $k = 4$ achieves the best performance, which is slightly better than dense prediction. This can be explained by the fact that some global references are inaccurate because the global correlation matrix is approximated by half channels. We emphasize that the separate global prediction model is not a simple non-local attention [38]. In the task of compression, the decoder cannot explicitly calculate attention map at the decoder side. Furthermore, the proposed causal global prediction model is also different from the causal non-local attention [39] used for image distribution modeling. Our method determines accurate reference points, while the causal non-local attention simply aggregates all previous points and cannot establish spatial correlations.

5) Analysis of cross-channel redundancy: We should note that the cross-channel redundancy still exists after non-linear transform and it is utilized with spatial correlation together. On the one hand, the transform network is not efficient enough to remove the cross-channel redundancy. In the non-linear transform network, one critical component is Generalized Divisive Normalization (GDN) layer that helps to decorrelate information across channels [9]. This GDN layer achieves a small mutual information between transformed components as illustrated in [40]. However, as shown in the figure 1 of [40], the mutual information after GDN transform is not completely zero, which implies that residual correlation exists after such non-linear transform. On the other hand, the cross-channel relationship is utilized together with spatial correlation. As-



(a) Non-local version in [30].

(b) Simplified version in [7].

(c) Our proposed group-separated attention.

Fig. 8: Comparisons of different versions of attention. RB represents residual block, which contains three 3×3 convolution layers (stride=1) and three ReLU activation layers.

sume that we have two same latent vectors in different spatial location. Now we want to use the first latent vector to predict the other undecoded latent vector. Even if the cross-channel redundancy was removed entirely, we can still determine the similarity between these two latent vectors by examining their first half vectors because their first half channels are exactly the same. We thus can conduct causal global prediction for the second channel group.

E. Group-Separated Attention Module in Transforms

In the previous sections, to reduce the cross-channel redundancies and global redundancies among the latents, we introduce the novel causal context model and causal global prediction model. Both of them facilitate entropy estimation for the second latent group. In addition to the improved entropy model, we further propose a new group-separated attention layer to enhance the non-linear transform networks.

There are several previous works with respect to learned image compression using attention modules to enhance the encoder and decoder. Chen et al. [30] suggests employing a residual non-local module for compression. Later, Cheng et al. simplify this attention layer [7] by removing the non-local block with comparable performance. In this paper, inspired by [16], [15], we empirically modify the attention structure and adopt an improved group-separated attention module, which enables separate feature-map attention in two groups. The channel splitting process is expected to gather those channels with similar characteristics. Detailed comparisons between these three types of attention can be found in Fig.8. Experimental results indicate that this group-separated attention module is more powerful than other attention mechanisms (shown later in the experimental section).

F. Network Architecture

The architecture of our main compression network is similar as [7]. Our main contributions are to employ the novel causal context model and causal global prediction model and replace the original attention module with a separate attention module. Besides, we enhance the parameter estimation module using residual blocks [41] as described in our pioneer work [18]. The main network structure is shown in Fig.9. To build a powerful compression model, we integrate a postprocessing network GRDN [42] as suggested in [29]. This postprocessing network is lightweight and achieves a good balance between complexity and performance. To estimate the entropy of $\hat{y}$, we assume the distribution of y subject to Gaussian mixture model (GMM) and this is a generalized form to approximate arbitrary probability distribution. Mathematically,

$$\begin{aligned} p(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}) &\sim \sum_{i=1}^{I} \pi_i \mathcal{N}(\mu_i, \sigma_i^2), \\ \pi_i, \mu_i, \sigma_i^2 &\leftarrow h_s(\hat{\boldsymbol{z}}, (\boldsymbol{c_{n,1}}, \boldsymbol{c_{n,2}}, \boldsymbol{c_{n,3}})|\boldsymbol{\theta_h}) \end{aligned} \tag{8}$$

where $\pi_i, \mu_i, \sigma_i, i \in \{1, ..., I\}$ are the estimated parameters of the entropy model (we choose $I = 3$ following [7], [18]). Eq.8 models a continuous probability density regarding y. To estimate the discrete probability of $\hat{y}$, the continuous function $p(y|\hat{z})$ is convolved with a unit uniform distribution, which is usually implemented in the form of additive uniform noise,

$$P_{\hat{y}|\hat{z}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}) = (\sum_{i=1}^{I} \pi_i \mathcal{N}(\mu_i, \sigma_i^2) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2}))(\hat{\boldsymbol{y}}). \tag{9}$$

As explained in the work of [9], such additive uniform noise helps to solve the problem of non-differential quantization. Note that we sent the hard-quantized latent variable $\hat{y}$ into the synthesis decoder during training, which is similar to [43] and is found to achieve better performance. We describe the experimental results in the following section.

V. Experiments

A. Implementation Details

Our compression model is trained on the whole ImageNet training set [44]. The original images are cropped to 256×256 patches. Minibatches of 8 of these patches are used to update network parameters. We divide the training period into three stages. First, we train the main compression network with rate-distortion constraint (as in Eq.2). This training stage lasts for 1.5 million iterations. Second, we fix the main compression network and optimize the postprocessing GRDN to minmize distortion, which lasts for 400,000 iterations. Third, both the main network and the postprocessing network are optimized jointly to achieve the optimal rate-distortion performance. The third training stage lasts for 600,000 iterations. For all three training stages, we use the Adam optimizer [45] with an initial learning rate of $5e-5$. The learning rate decays to $1e-5$ after 300,000 iterations.

Similar to previous works, the distortion term is measured by two quality metrics, PSNR and MS-SSIM [17]. During training, a Lagrange multiplier λ balances the rate-distortion trade-off. When optimizing for PSNR, we train different

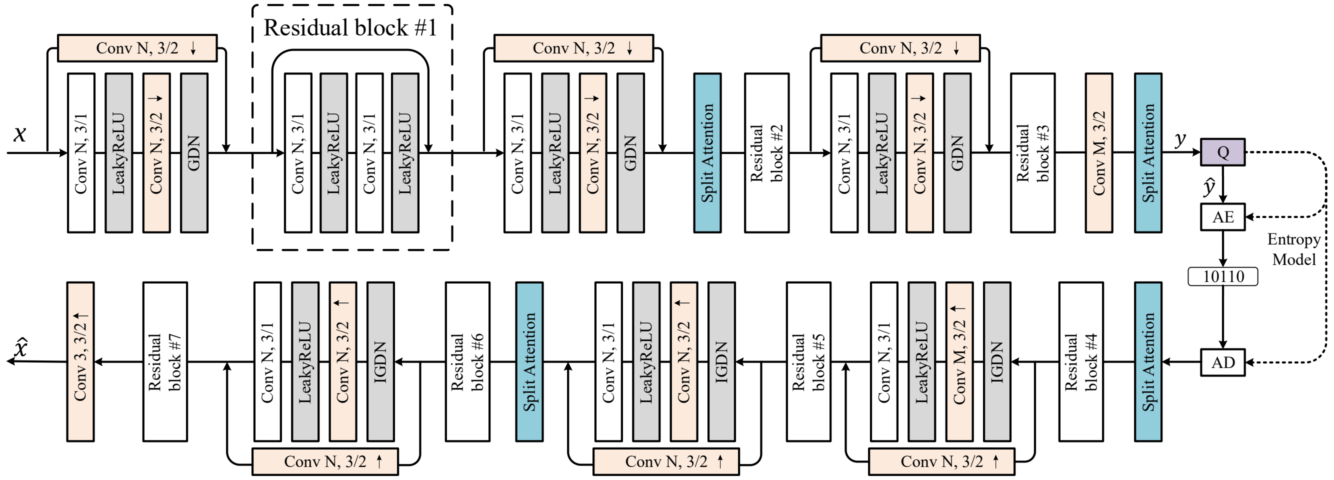


Fig. 9: Architecture of our main compression network. We choose N=192, M=192 when bitrates are less than 0.50 bpp. For higher bitrate, we set N=192, M=320 to enhance the model capacity following the setting in [5].

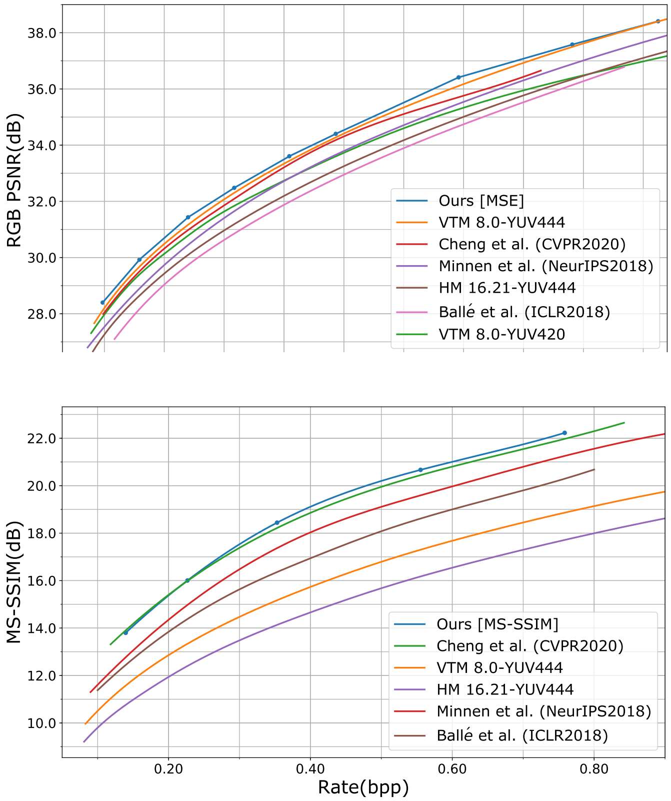


Fig. 10: Comparisons of RD curves with other methods.

models with values of λ ranging from 100 to 3072, where $D = MSE(\hat{x}, x)$. In terms of MS-SSIM, we train five models with $\lambda \in \{6, 16, 40, 100, 180\}$, where $D = 1-\text{MSSSIM}(\hat{x}, x)$.

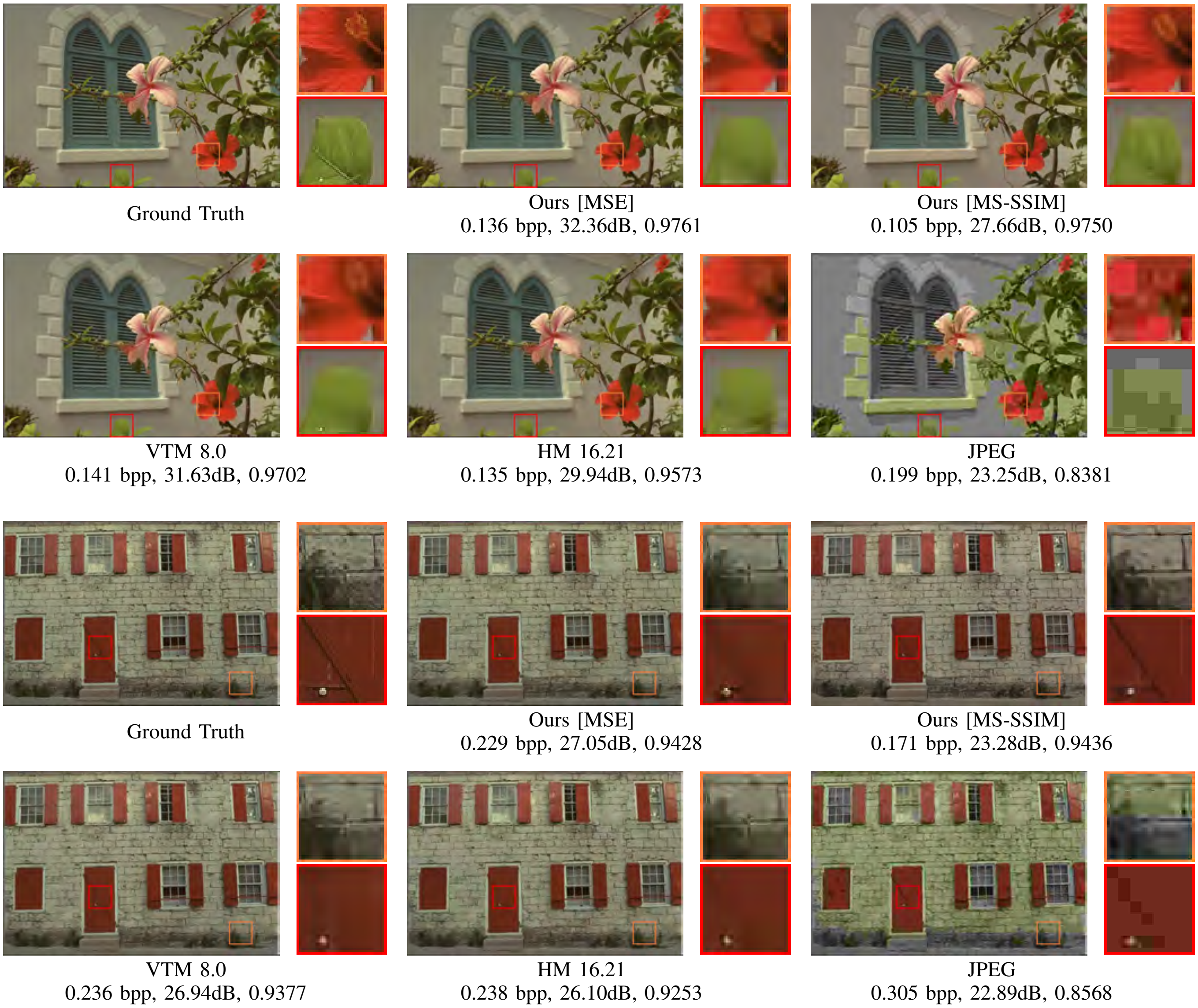
B. Performance

For comparison, we evaluate different methods on the Kodak dataset which contains 24 uncompressed images. The rate distortion curves are shown in Fig.10. We compare the performance of our method with other existing approaches including [7], [10], [5] and traditional codecs, such as VVC intra and HEVC intra. Specifically, the results of Ballé et al. (ICLR2018) [5] are obtained from our reproduced model, which performs closely as their report. The results of Minnen et al. (NeurIPS2018) [10] and Cheng et al. (CVPR2020) [7] are directly taken from their papers because our reproduced statistics are slightly lower than their report. The reproduction gap may be caused by differences in the training set. While we directly use the whole ImageNet training set, [10] does not mention the training set and [7] uses a subset of ImageNet to avoid negative samples. For VVC and HEVC, we use the official test model VTM 8.0 and HM 16.21[7], with the YUV444 configuration in all-intra mode.

Fig.10 quantitatively shows that our method achieves the state-of-the-art rate-distortion performance. Specifically, it outperforms VTM 8.0 at all bitrates in terms of PSNR and performs better than the previous state-of-the-art method [7] at most bitrates in terms of MS-SSIM. Our method achieves 5.1% BD rate savings against VTM 8.0 (covering 0.15, 0.4, 0.7 and 1.0 bpp). There is an irregular point in our PSNR-rate curve (at 0.68 bpp), which is the turning point that indicates the change of model capacity. In Fig.11, we provide some examples that exhibit pleasant qualitative results obtained by our method, especially the model optimized for MS-SSIM.

Furthermore, our proposed method is expected to present better results on images with a lot of repeat patterns such as screen-captured images. Here, we evaluate the performance of our method on some screen-captured images from HEVC standard test sequences of screen content. We notice that there are some screen-captured images with very simple contexts, which require only few bits ($<$ 0.01 bpp) but would easily deliver a high-quality reconstruction (PSNR $>$ 45dB). Therefore, this kind of images would have significant influence on the average rate-distortion performance and makes the average statistics unreliable. We instead provide individual RD curves of two

[7] https://vcgit.hhi.fraunhofer.de/jct-vc/HM/-/releases/HM-16.21


Ground Truth

Ours [MSE]
0.136 bpp, 32.36dB, 0.9761

Ours [MS-SSIM]
0.105 bpp, 27.66dB, 0.9750

VTM 8.0
0.141 bpp, 31.63dB, 0.9702

HM 16.21
0.135 bpp, 29.94dB, 0.9573

JPEG
0.199 bpp, 23.25dB, 0.8381

Ground Truth

Ours [MSE]
0.229 bpp, 27.05dB, 0.9428

Ours [MS-SSIM]
0.171 bpp, 23.28dB, 0.9436

VTM 8.0
0.236 bpp, 26.94dB, 0.9377

HM 16.21
0.238 bpp, 26.10dB, 0.9253

JPEG
0.305 bpp, 22.89dB, 0.8568

Fig. 11: Visual comparisons.

specific screen-captured images to compare our method with VVC. As shown in Fig. 12, we use our full compression model that is still trained on the ImageNet dataset for comparison. It can be observed that our method achieves better performance when the bitrate is low. It verifies the effectiveness of our proposed causal prediction method. However, VVC gradually performs better than our method when the bitrate and PSNR value increase. This result can be explained from the perspective of autoencoder (AE) limit of neural compression model [46]. Our neural compression model is a VAE-based model, where the encoder transform and the decoder transform are non-invertible. It is different from the linear invertible transform such as DCT used in traditional codecs. The non-invertible transform network introduces errors and information loss, thereby sets a lower bound of distortion. In contrast, invertible transform used in traditional codecs does not bring in information loss. When the reconstruction quality is very high, the learning-based compression model would reach the AE limit and performs worse than traditional codecs. Since screen-captured images have a lot of repeated patterns and some even have very simple contexts, the reconstruction quality of screen-captured image is naturally very high. Therefore, when using learning-based codecs to compress screen-captured images, it will easily encounter the AE-limit issue due to the high-quality reconstructions. As a result, in this case of compression of screen-captured images, our method performs better than VVC when the bitrate is low but is not excellent as VVC when the PSNR value is high..

C. Ablation Study

In Section IV-D, we conduct an exploratory experiment to investigate the effects of different prediction modes where the baseline is a relative simple model [10]. To further verify the effectiveness of different modules, we conduct ablation studies under the same experimental settings where the baseline is now our main compression network.

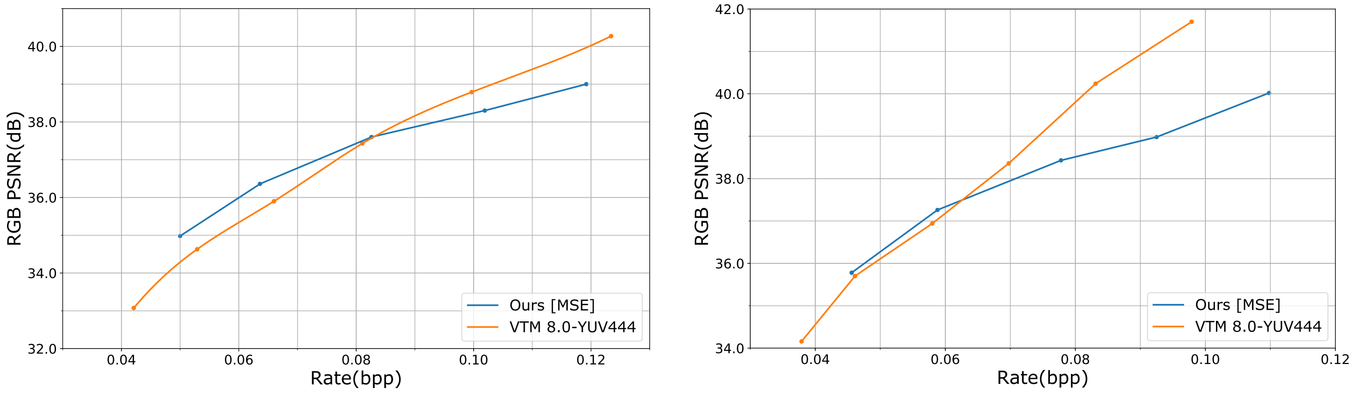


Image 1

Image 2

Fig. 12: Individual RD curves of two screen-captured images.

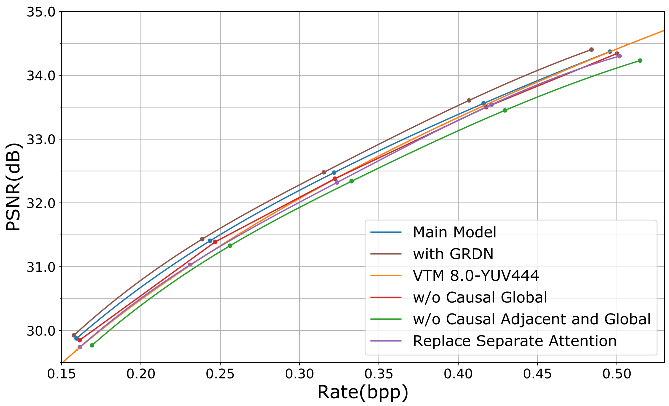


Fig. 13: Ablation study where baseline is our main compression model.

As shown in Fig.13, *Main Model* represents our main compression network without postprocessing. Additional postprocessing GRDN boosts performance by approximately 0.1 dB (the brown line) and this corresponds to the optimal performance of our method in Fig.10. Compared with the main compression network, both the causal adjacent context model and causal global prediction model are important to achieve improved rate-distortion performance (as shown by the red line and the green line, respectively). Note that *w/o causal global* refers to removing the causal global prediction model while preserving causal context model. Considering that our baseline network is powerful enough, such improvements are encouraging. In addition, replacing the proposed separate attention module by the attention layer used in [7] also results in a performance drop. In conclusion, our proposed three elements, including causal context model, causal global prediction model and separate attention layer, contribute to the state-of-the-art performance of our method.

D. Coding time

To evaluate the running time of the models fairly, we limit the accessible resources to one 2080 Ti GPU and two cores of an Intel E5-2699 v4 CPU. We modify our previous implementation submitted to CLIC 2020 to evaluate the coding time of three different entropy models. Here, *Mask Context* is the conventional context model with 2-D mask convolution [10] that is taken as the base model. *Causal Context* corresponds to our submission to CLIC competition [18] which adopts the adjacent causal context model. As shown in Table II, since the serial decoding separates the decoding process across channels, the decoding time increases from average 6.7 seconds to average 7.9 seconds. Note that we test all the 24 Kodak images of which resolutions are 768×512. *Causal Context + Causal Global Prediction* is the most powerful entropy model proposed in this paper. However, despite the state-of-the-art rate-distortion performance, this complex entropy model requires 37.7 seconds for decoding on average. It is unsurprising because the global searching of references points consumes much more time. There remains much room for optimization in terms of both software and hardware.

	Encoding Time (s)	Decoding Time (s)
Masked Context [10]	3.4	6.7
Causal Context	3.5	7.9
Causal Context + Causal Global Prediction	3.5	38.7

TABLE II: Encoding and decoding time on Kodak dataset. The statistics are averaged on the whole dataset. Those three methods have the similar encoding time because encoding can be computed in parallel.

VI. Conclusion

In this paper, we explore reducing the global redundancies and cross-channel redundancies among the latent variables in an entropy model. It is observed that separating the latents is advantageous for improving the entropy model. To this end, we first propose a causal context model to generate highly informative adjacent context. We then extend it to a causal global prediction model to conduct global prediction with accurate global references. Both models make use of channel-wise redundancies to facilitate entropy estimation for a specific latent group. While the separate entropy model suffers from an increased decoding time, it significantly improves the rate-distortion performance. In addition, we adopt a new group-separated attention module that enables independent feature-map attention in separate groups, thereby enhancing the transform networks. Experimental results indicate that our method achieves the state-of-the-art performance in terms of both PSNR and MS-SSIM.

REFERENCES

[1] A. Habibi, "Hybrid coding of pictorial data," *IEEE Transactions on Communications*, vol. 22, no. 5, pp. 614–624, 1974.

[2] R. Forchheimer, "Differential transform coding: A new hybrid coding scheme," in *1981 Picture Coding Symposium (PCS)*, 1981.

[3] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h. 264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[4] G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, and M. Covell, "Full resolution image compression with recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5306–5314.

[5] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[6] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 221–231.

[7] Z. Cheng, H. Sun, M. Takeuchi, and J. 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*, 2020, pp. 7939–7948.

[8] V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 9–21, 2001.

[9] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[10] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 771–10 780.

[11] J. Lee, S. Cho, and S.-K. Beack, "Context-adaptive entropy model for end-to-end optimized image compression," in *the 7th International Conference on Learning Representations, ICLR 2019*, 2019.

[12] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1747–1756.

[13] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3214–3223.

[14] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S. Jin Hwang, J. Shor, and G. Toderici, "Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4385–4393.

[15] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 407–10 416.

[16] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] Z. Guo, Y. Wu, R. Feng, Z. Zhang, and Z. Chen, "3-d context entropy model for improved practical image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 116–117.

[19] G. K. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[20] M. Rabbani, "Jpeg2000: Image compression fundamentals, standards and practice," *Journal of Electronic Imaging*, vol. 11, no. 2, p. 286, 2002.

[21] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[22] G. J. Sullivan and J.-R. Ohm, "Versatile video coding towards the next generation of video compression," in *2018 Picture Coding Symposium (PCS)*, 2018.

[23] J. Xu, R. Joshi, and R. A. Cohen, "Overview of the emerging hevc screen content coding extension," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 50–62, 2015.

[24] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," *arXiv preprint arXiv:1703.00395*, 2017.

[25] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Advances in Neural Information Processing Systems*, 2017, pp. 1141–1151.

[26] W. Li, "Overview of fine granularity scalability in mpeg-4 video standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 301–317, 2001.

[27] Z. Zhang, Z. Chen, J. Lin, and W. Li, "Learned scalable image compression with bidirectional context disentanglement network," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1438–1443.

[28] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Conditional probability models for deep image compression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4394–4402.

[29] J. Lee, S. Cho, and M. Kim, "A hybrid architecture of jointly learning image compression and quality enhancement with improved entropy minimization," *arXiv preprint arXiv:1912.12817*, 2019.

[30] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, and Y. Wang, "Neural image compression via non-local attention optimization and improved context modeling," *arXiv preprint arXiv:1910.06244*, 2019.

[31] M. Li, K. Zhang, W. Zuo, R. Timofte, and D. Zhang, "Learning context-based non-local entropy modeling for image compression," *arXiv preprint arXiv:2005.04661*, 2020.

[32] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[33] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[34] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[35] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[36] X. Liu, J.-Y. Lee, and H. Jin, "Learning video representations from correspondence proposals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4273–4281.

[37] T. Wiegand, "Draft itu-t recommendation and final draft international standard of joint video specification (itu-t rec. h. 264— iso/iec 14496-10 avc)," *JVT-G050*, 2003.

[38] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[39] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "Pixelsnail: An improved autoregressive generative model," in *International Conference on Machine Learning*. PMLR, 2018, pp. 864–872.

[40] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," in *4th International Conference on Learning Representations, ICLR 2016*, 2016.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] D.-W. Kim, J. Ryun Chung, and S.-W. Jung, "Grdn: Grouped residual dense network for real image denoising and gan-based real-world noise modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

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

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[45] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[46] L. Helminger, A. Djelouah, M. Gross, and C. Schroers, “Lossy image compression with normalizing flows,” *arXiv preprint arXiv:2008.10486*, 2020.

Zongyu Guo received the B.S. degree from University of Science and Technology of China in 2019. He is currently a Ph.D. student in the Department of Electronic Engineering and Information Science, in University of Science and Technology of China, advised by Prof. Zhibo Chen. His research interests include image/video compression, image inpainting, and generative modelling of data distribution.

Zhizheng Zhang received the B.S. degree in electronic information engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2016. He is currently pursuing the Ph.D. degree with the University of Science and Technology of China, Hefei, China. His current research interests include image/video compression, intelligent media understandings, and reinforcement learning.

Runsen Feng received the B.S. degree in electronic information engineering from the University of Science and Technology of China, Hefei, China, in 2018. He is currently a Ph.D. student at the University of Science and Technology of China. His research interests include learning-based image coding and video coding.

Zhibo Chen (M’01-SM’11) received the B. Sc., and Ph.D. degree from Department of Electrical Engineering Tsinghua University in 1998 and 2003, respectively. He is now a professor in University of Science and Technology of China. His research interests include image and video compression, visual quality of experience assessment, immersive media computing and intelligent media computing. He has more than 150 publications and more than 50 granted EU and US patent applications. He is IEEE senior member, Secretary/Chair-Elect of IEEE Visual Signal Processing and Communications Committee, and member of IEEE Multimedia System and Applications Committee. He was TPC chair of IEEE PCS 2019 and organization committee member of ICIP 2017 and ICME 2013, served as TPC member in IEEE ISCAS and IEEE VCIP.