







# Contextformer: A Transformer with Spatio-Channel Attention for Context Modeling in Learned Image Compression

A. Burakhan Koyuncu<sup>1,2</sup>, Han Gao<sup>4</sup>, Atanas Boev<sup>2</sup>, Georgii Gaikov<sup>3</sup>,  
Elena Alshina<sup>2</sup>, and Eckehard Steinbach<sup>1</sup>

<sup>1</sup> Technical University of Munich, Munich, Germany  
[burakhan.koyuncu@tum.de](mailto:burakhan.koyuncu@tum.de)

<sup>2</sup> Huawei Munich Research Center, Munich, Germany

<sup>3</sup> Huawei Moscow Research Center, Moscow, Russia

<sup>4</sup> Tencent America, Palo Alto, USA

**Abstract.** Entropy modeling is a key component for high-performance image compression algorithms. Recent developments in autoregressive context modeling helped learning-based methods to surpass their classical counterparts. However, the performance of those models can be further improved due to the underexploited spatio-channel dependencies in latent space, and the suboptimal implementation of context adaptivity. Inspired by the adaptive characteristics of the transformers, we propose a transformer-based context model, named Contextformer, which generalizes the de facto standard attention mechanism to spatio-channel attention. We replace the context model of a modern compression framework with the Contextformer and test it on the widely used Kodak, CLIC2020, and Tecnick image datasets. Our experimental results show that the proposed model provides up to 11% rate savings compared to the standard Versatile Video Coding (VVC) Test Model (VTM) 16.2, and outperforms various learning-based models in terms of PSNR and MS-SSIM.

**Keywords:** Context Model, Learned Image Compression, Transformers

## 1 Introduction

Recent works in learned image compression outperform hand-engineered classical algorithms such as JPEG [46] and BPG [8], and even reach the rate-distortion performance of recent versions of video coding standards, such as VVC [1]. The most successful learning-based methods use an autoencoder based on [6,33], where the entropy of the latent elements is modeled and minimized jointly with an image distortion metric. The entropy modeling relies on two principles – backward and forward adaptation [3]. The former employs a hyperprior estimator utilizing a signaled information source. The latter implements a context model, where previously decoded symbols are used for entropy estimation without a need for signaling. Due to its efficiency, a wide variety of context model architectures were explored in the recent literature [20,32,33,34,38,39,52]. We categorize those

architectures into the following groups w.r.t. their targets: (1) increased support for spatial dependencies; (2) exploitation of cross-channel dependencies; (3) increased context-adaptivity in the entropy estimation. For instance, we consider the methods such as [13,14,26,52] in the first category since those methods aim to capture long distant relations in the latent space. The 3D context [11,31,32] and channel-wise autoregressive context model [34] fall in the second category. In those works, entropy estimation of each latent element can also use information from the spatially co-located elements of previously coded channels. In [34] the authors show that entropy estimation, which mainly relies on cross-channel dependencies, outperforms their previous spatial-only model [33]. Most often, the context models use non-adaptive masked convolutions [29]. Those are location-agnostic [29], i.e., the same kernel is applied to each latent position, which potentially reduces the model performance. Even for a larger kernel size, the performance return is marginal, as only a small set of spatial relations between symbols can be utilized. [20,39] propose an adaptive context model, where the selection of latent elements to be used is based on pair-wise similarities between previously decoded elements. Furthermore, [38] uses a transformer-based context model to achieve context adaptivity for the spatial dimensions. However, those models have limited context adaptivity, as they are partially or not applying adaptive modeling for the cross-channel elements. For instance, in [20], although the primary channel carries on average 60% of the information, the context model does not employ any adaptive mechanism for modeling it.

Attention is a widely used deep learning technique that allows the network to focus on relevant parts of the input and suppress the unrelated ones [36]. In contrast to convolutional networks, attention-based models such as transformers [45] provide a large degree of input adaptivity due to their dynamic receptive field [35]. This makes them a promising candidate for a high-performance context model. Following the success of transformers in various computer vision tasks [10,16,17,22,38], we propose a transformer-based context model, *Contextformer*. Our contribution is threefold: (1) We propose a variant of the Contextformer, which adaptively exploits long-distance spatial relations in the latent tensor; (2) We extend the Contextformer towards a generalized context model, which can also capture cross-channel relations; (3) We present algorithmic methods to reduce the runtime of our model without requiring additional training.

In terms of PSNR, our model outperforms a variety of learning-based models, as well as VTM 16.2 [43] by a significant margin of 6.9%–10.5% in average bits saving on the Kodak [18], CLIC2020 [44] and Tecnick [2] image datasets. We also show that our model provides better performance than the previous works in a perceptual quality-based metric MS-SSIM [47].

## 2 Related Work

### 2.1 Learned Image Compression

Presently, the state-of-the-art in lossy image compression frameworks is fundamentally a combination of variational autoencoders and transform coding [19],

where the classical linear transformations are replaced with learned non-linear transformation blocks, e.g., convolutional neural networks [3]. The encoder applies an analysis transform  $g_a(\mathbf{x}; \phi)$  mapping the input image  $\mathbf{x}$  to its latent representation  $\mathbf{y}$ . This transform serves as dimensionality reduction. The latent representation  $\mathbf{y}$  is quantized by  $Q(\cdot)$  and is encoded into the bitstream. In order to obtain the reconstructed image  $\hat{\mathbf{x}}$ , the decoder reads the quantized latent  $\hat{\mathbf{y}}$  from the bitstream and applies the synthesis transform  $g_s(\hat{\mathbf{y}}; \theta)$ , which is an approximate inverse of  $g_a(\cdot)$ .

Aiming to reduce the remaining coding redundancy in latent space, Ballé et al. [5] introduced the factorized density model, which estimates the symbol distribution by using local histograms. During training, a joint optimization is applied to minimize both the symbol entropy and the distortion between the original and the reconstructed image. Knowledge of the probability distribution and coding methods such as arithmetic coding [40] allows for efficient lossless compression of  $\hat{\mathbf{y}}$ . Later, Ballé et al. [6] proposed using a hyperprior, which employs additional analysis and synthesis transforms  $h_{a/s}(\cdot)$  and helps with modeling of the distribution  $p_{\hat{\mathbf{y}}}(\hat{\mathbf{y}}|\hat{\mathbf{z}})$  conditioned on the side information  $\hat{\mathbf{z}}$ . The side information is modeled with a factorized density model, whereas  $p_{\hat{\mathbf{y}}}(\hat{\mathbf{y}}|\hat{\mathbf{z}})$  is modeled as a Gaussian distribution. Their proposed framework can be formulated as

$$\hat{\mathbf{y}} = Q(g_a(\mathbf{x}; \phi)), \quad (1)$$

$$\hat{\mathbf{x}} = g_s(\hat{\mathbf{y}}; \theta), \quad (2)$$

$$\hat{\mathbf{z}} = Q(h_a(\hat{\mathbf{y}}; \phi_h)), \quad (3)$$

$$p_{\hat{\mathbf{y}}}(\hat{\mathbf{y}}|\hat{\mathbf{z}}) \leftarrow h_s(\hat{\mathbf{z}}; \theta_h), \quad (4)$$

and the loss function  $\mathcal{L}$  of end-to-end training is

$$\mathcal{L}(\phi, \theta, \phi_h, \theta_h, \psi) = \mathbf{R}(\hat{\mathbf{y}}) + \mathbf{R}(\hat{\mathbf{z}}) + \lambda \cdot \mathbf{D}(\mathbf{x}, \hat{\mathbf{x}}) \quad (5)$$

$$= \mathbb{E}[\log_2(p_{\hat{\mathbf{y}}}(\hat{\mathbf{y}}|\hat{\mathbf{z}}))] + \mathbb{E}[\log_2(p_{\hat{\mathbf{z}}}(\hat{\mathbf{z}}|\psi))] + \lambda \cdot \mathbf{D}(\mathbf{x}, \hat{\mathbf{x}}), \quad (6)$$

where  $\phi$ ,  $\theta$ ,  $\phi_h$  and  $\theta_h$  are the optimization parameters and  $\psi$  denotes the parameters of the factorized density model  $p_{\hat{\mathbf{z}}}(\hat{\mathbf{z}}|\psi)$ .  $\lambda$  is the Lagrange multiplier regulating the trade-off between rate  $\mathbf{R}(\cdot)$  and distortion  $\mathbf{D}(\cdot)$ .

## 2.2 Context Model

Higher compression performance requires more accurate entropy models, which would need an increased amount of side information [33]. To overcome this limitation, Minnen et al. [33] proposed a context model, which estimates the entropy of current latent element  $\hat{\mathbf{y}}_i$  using the previously coded elements. Their approach extends Eq. (4) to

$$p_{\hat{\mathbf{y}}_i}(\hat{\mathbf{y}}_i|\hat{\mathbf{z}}) \leftarrow g_{ep}(g_{cm}(\hat{\mathbf{y}}_{< i}; \theta_{cm}), h_s(\hat{\mathbf{z}}; \theta_h); \theta_{ep}), \quad (7)$$

where the context model  $g_{cm}(\cdot)$  is implemented as a 2D masked convolution.  $g_{ep}(\cdot)$  computes the entropy parameters and  $\hat{\mathbf{y}}_{< i}$  denotes the previously coded

local neighbors of current latent element  $\hat{\mathbf{y}}_i$ . Their proposed 2D context model requires 8.4% fewer bits than BPG [8].

Further improvements of the context model have been proposed (see Figs. 1a to 1e). In [13,14,52] a multi-scale context model was implemented, which employs multiple masked convolutions with different kernel sizes in order to learn various spatial dependencies simultaneously. [11,31,32] employ 3D masked convolutions in order to exploit cross-channel correlations jointly with the spatial ones.

Minnen and Singh proposed a channel-wise autoregressive context model [34]. It splits the channels of the latent tensor into segments and codes each segment sequentially with the help of a previously coded segment. This reduces the sequential steps and outperforms the 2D context model of [33]. However, this approach uses only cross-channel correlations and omits the spatial ones.

Qian et al. [39] proposed a context model, which combines 2D masked convolutions with template matching to increase the receptive field and provide context adaptivity. They search for a similar patch in the previously coded positions and use the best match as a global reference for the entropy model.

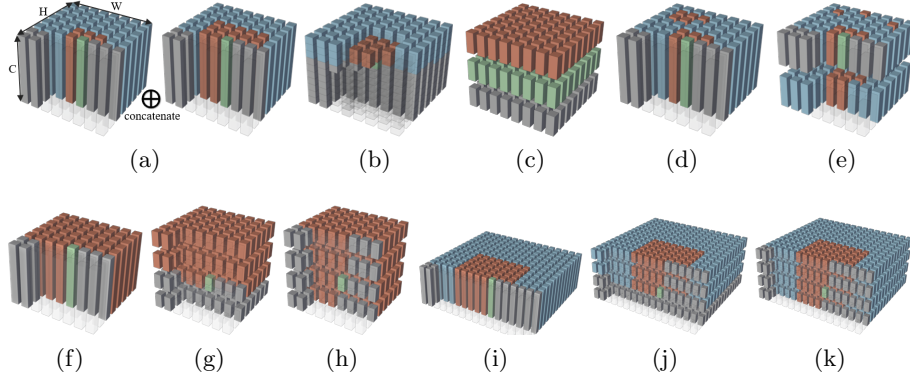
Guo et al. [20] proposed a context model, which can be seen as an extension of [39]. In their approach, the channels of the latent tensor are split into two segments. The first segment is coded with a 2D masked convolution, similar to [34]. Coding of the second segment is done using two different mechanisms: MaskConv+, an “improved” version of the 2D masked convolutions, and a global prediction. Additional to the local neighbors, MaskConv+ uses the spatially co-located elements from the first segment. The global prediction is made by calculating the similarity between all elements from the first segment. The indices of the top  $k$  similar elements (from the corresponding position in the first segment) are used to select elements in the second segment and include those in the entropy model. They reported average bits savings of 5.1% over VTM 8.0 [43].

Qian et al. [38] replaced the CNN-based hyperprior and context model with a transformer-based one which increased the adaptivity of the entropy model. They proposed two architectures for their context model – serial and parallel model. The serial model processes the latent tensor sequentially similar to [33]. The parallel one uses the checkboard like grouping proposed in [21] to increase the decoding speed. They achieved competitive performance with some of the CNN-based methods such as [12].

### 2.3 Transformers

Self-attention is an attention mechanism originally proposed for natural language processing [45]. Later, it was adopted in various computer vision tasks, where it outperformed its convolutional counterparts [10,16,17,22].

The general concept of a transformer is as follows. First, an input representation with  $S$  sequential elements is mapped into three different representations, query  $\mathbf{Q} \in \mathbb{R}^{S \times d_q}$ , key  $\mathbf{K} \in \mathbb{R}^{S \times d_k}$  and value  $\mathbf{V} \in \mathbb{R}^{S \times d_v}$  with separate linear transformations. Then, the attention module dynamically routes the queries with the key-value pairs by applying a scaled dot-product. The attention module is



**Fig. 1.** Illustration of the latent elements used by the context model (■) to estimate the entropy of the current latent (■) in (a–e) for the prior-arts and (f–k) our proposed context model. Previously coded and yet to be coded elements are displayed as (■) and (■), respectively. The displayed prior-art models are (a) multi-scale 2D context [13,14,52], (b) 3D context [11,31,32], (c) channel-wise autoregressive context [34], (d) 2D context with global reference [39], and (e) context with advanced global reference [20]. Note that in (c) each (■) is coded simultaneously by using only a part of the elements presented as (■), and in (e), the primary channel segment is shown at the bottom for better visibility. Our models with different configurations are shown in (f) Contextformer( $N_{cs}=1$ ), (g) Contextformer( $N_{cs}>1, sfo$ ), (h) Contextformer( $N_{cs}>1, cfo$ ). Note that the (serial) transformer-based context model of [38] employs similar mechanism as (f). (i–k) show the versions of our models (f–h) using the sliding window attention.

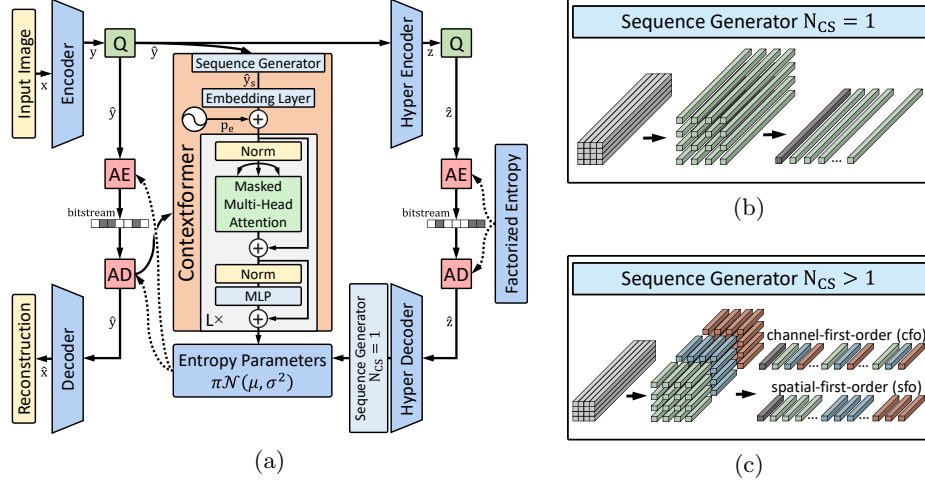
followed by a point-wise multi-layer perceptron (MLP). Additionally, the multi-head attention splits the queries, keys, and values into  $h$  sub-representations (so-called heads), and for each head, the attention is calculated separately. The final attention is computed by combining each sub-attention with a learned transformation ( $\mathbf{W}$ ). The multi-head attention enables parallelization and each intermediate representation to build multiple relationships.

To preserve coding causality in autoregressive tasks, the attention mechanism has to be limited by a mask for subsequent elements not coded yet. The masked multi-head attention can be formulated as:

$$Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}, \quad (8)$$

$$\text{head}_i(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{softmax} \left( \frac{\mathbf{Q}_i \mathbf{K}_i^T}{d_k} \odot \mathbf{M} \right) \mathbf{V}_i, \quad (9)$$

where  $\mathbf{Q}_i \in \mathbb{R}^{S \times \frac{d_q}{h}}$ ,  $\mathbf{K}_i \in \mathbb{R}^{S \times \frac{d_k}{h}}$  and  $\mathbf{V}_i \in \mathbb{R}^{S \times \frac{d_v}{h}}$  are the sub-representations, the mask  $\mathbf{M} \in \mathbb{R}^{S \times S}$  has ones in its lower triangle and the rest of its values are minus infinity, and  $\odot$  stands for the Hadamard product.



**Fig. 2.** Illustration of (a) our proposed model with the Contextformer, (b) sequence generator for the Contextformer with spatial attention, Contextformer( $N_{cs}=1$ ), and (c) sequence generator for the Contextformer with spatio-channel attention, Contextformer( $N_{cs}>1$ ). The prepended start token is shown in dark gray in (b-c). Inspired by [27], we use channel-wise local hyperprior neighbors to increase performance; thus, regardless of the selected  $N_{cs}$ , we apply the sequence generator depicted in (b) to the output of the hyperdecoder.

### 3 Our Approach

In this section, we introduce our transformer-based context model, *Contextformer*, which provides context adaptivity and utilizes distant spatial relations in the latent tensor. We present two versions of the model: a simple version, which uses spatial attention; and a more advanced version, which employs attention both in the spatial and channel dimensions for entropy modeling.

#### 3.1 Contextformer with Spatial Attention

The proposed Contextformer builds on top of the architecture introduced in [14]. In the encoder, this model employs  $3 \times 3$  convolution layers with GDN activation function [4] and residual non-local attention modules (RNAB) [50]. The structure of the decoder is very similar to the one of the encoder, with the exception that residual blocks (ResBlock) [11] are used in the first layer to enlarge the receptive field. Additionally, the model adopts a hyperprior network, the multi-scale context model [52] and the universal quantization [53]. This model estimates the distribution  $p_{\hat{y}}(\hat{y}|\hat{z})$  with a single Gaussian distribution. In our approach, we use a Gaussian mixture model [12] with 3 mixture components  $k_m$ , which is known to increase the accuracy of the entropy model.

In contrast to Cui et al. [14], we use a Contextformer instead of their multi-scale context model, as shown in Fig. 2a. First, the latent  $\hat{\mathbf{y}} \in \mathbb{R}^{H \times W \times C}$  is rearranged into a sequence of spatial patches  $\hat{\mathbf{y}}_s \in \mathbb{R}^{\frac{HW}{p_h p_w} \times (p_h p_w C)}$ . Here,  $H$ ,  $W$  and  $C$  stand for the height, width and number of channels;  $(p_h, p_w)$  corresponds to the shape of each patch. Usually, patch-wise processing reduces complexity, especially for large images [16]. However, the latent  $\hat{\mathbf{y}}$  already has a 16 times lower resolution than the input image, which makes learning an efficient context model harder and leads to a performance drop. To remedy this issue, we set the patch size to  $1 \times 1$ , so each sequential element corresponds to one pixel in the latent tensor (see Fig. 2b).

The Contextformer has  $L$  transformer layers with a similar architecture to that of ViT [16]. Each layer requires an intermediate tensor with an embedding size of  $d_e$ . Therefore, we apply a learnable linear transformation  $\mathbb{R}^{HW \times C} \rightarrow \mathbb{R}^{HW \times d_e}$  (embedding layer). In order to introduce permutation-variance, we add a learned position encoding similar to the one in [16,17], but we apply it to the first layer only. We prepend the latent sequence  $\hat{\mathbf{y}}_s$  with a zero-valued start token to ensure the causality of coding. We use masking in the attention as described in Eq. (9), and multi-head attention with 12 heads. Multi-head allows our model to independently focus on different channel segments of  $\hat{\mathbf{y}}$ .

### 3.2 Contextformer with Spatio-Channel Attention

Although multi-head attention is computationally efficient in handling cross-channel dependencies, it can explore relationships in a single channel only partially. For example, consider a Contextformer with a single transformer layer. Given a latent tensor  $\hat{\mathbf{y}} \in \mathbb{R}^{H \times W \times C}$  and its sequential representation  $\hat{\mathbf{y}}_s \in \mathbb{R}^{S \times C}$ ; the  $n$ -th sub-representation of the sequence  $\hat{\mathbf{y}}_s(n, h_i) \in \mathbb{R}^{1 \times \frac{C}{k}}$  can only attend to the previous representations  $\hat{\mathbf{y}}_s(<n, h_i)$  with the same head index  $h_i$ . This means that the attention between different channel segments is not considered. Another limitation arises from the way the model behaves w.r.t.  $\hat{\mathbf{y}}$ . For modeling the entropy of latent element  $\hat{\mathbf{y}}(i, j, c)$  (with spatial coordinates  $(i, j)$  and channel index  $c$ ), the context model cannot access information from the spatially co-located elements from other channels  $\hat{\mathbf{y}}(i, j, \neq c)$ . This limits exploiting the cross-channel dependencies in  $\hat{\mathbf{y}}$ , and, therefore, the performance of the model.

To remedy this issue, we generate spatio-channel patches in the latent space  $\hat{\mathbf{y}}_s \in \mathbb{R}^{\frac{HWC}{p_h p_w p_c} \times (p_h p_w p_c)}$ , where  $p_c$  corresponds to the size of each channel segment, and total number of channel segments is  $N_{cs} = \frac{C}{p_c}$ . In a special case of  $(p_h, p_w) = (H, W)$ , our patch generation method is similar to the slicing method in channel-wise context modeling [34], but our model has a multi-head attention added. In this work, we set  $p_h$  and  $p_w$  to 1 as already discussed in Section 3.1. Splitting the latent tensor into multiple channel segments enables two different coding methods, spatial-first-order (*sfo*) and channel-first-order (*cfo*). The first method prioritizes the spatial dimensions and codes all latent elements from a channel segment sequentially before starting with the next segment. The second method prioritizes the channel dimension, and codes all channel segments with

the same spatial coordinate sequentially, before coding elements from the next spatial coordinate.

Spatio-channel sequence generation allows the standard transformer to use channel attention, from which a generalized context model can be obtained. To illustrate this, we compare how dependencies in the latent space are handled by Contextformer for various  $N_{cs}$ , and how those dependencies are handled by context models in the prior-art.

**In case of  $N_{cs} = 1$ .** The attention is limited to a spatial one (see Section 3.1). However, such a model provides faster encoding and decoding due to the smaller number of required autoregressive steps and still has better performance than some models in the prior-arts. As illustrated in Fig. 1f, Contextformer ( $N_{cs} = 1$ ) has a larger receptive field than [13,14,52], and also employs learned context adaptivity. Additionally, in [39] the receptive field is limited to a single reference and its neighboring latent elements, whereas Contextformer ( $N_{cs} = 1$ )’s receptive field is dynamic and theoretically unlimited. Notably, the serial model of [38] (best performing one) uses a similar context model as Contextformer ( $N_{cs} = 1$ ). However, their model has only a sparse-attention mechanism, whereas Contextformer employs the full attention mechanism.

**In case of  $C \geq N_{cs} > 1$ .** We achieve a context model that can exploit both spatial and cross-channel dependencies. As shown in Figs. 1g and 1h, both Contextformers ( $N_{cs} > 1$ ) with different coding order handle spatio-channel relationships. Moreover, in [20] only non-primary channel segments could be selected for entropy estimation, while the receptive field of the Contextformer’s receptive field can adapt to cover every channel segment. Compared to [38], the Contextformer ( $N_{cs} > 1$ ) provides a more adaptive context model due to the spatio-channel attention. In the extreme case  $N_{cs} = C$ , the model employs the spatio-channel attention to its full extend by computing the attention between every single latent element. Other implementations can be seen as a simplification of the extreme case for balancing the trade-off between performance and complexity. The Contextformer ( $N_{cs} = C$ ) can be regarded as a 3D context model with a large and adaptive receptive field.

### 3.3 Handling High-Resolution Images

Although the receptive field size of the Contextformer is theoretically unlimited, computing attention for long input sequences, e.g., high resolution images, is not feasible due to the quadratic increase of memory requirement and computational complexity with increasing the length of the input sequence. Therefore, we have to limit the receptive field of our model and use sliding-window attention as described in [17,37]. Inspired by 3D convolutions, we implemented a 3D sliding-window to traverse the spatio-channel array. Unlike 3D convolutions, the receptive field only slides across spatial dimensions and expands to encompass all elements in the channel dimension. In Figs. 1i to 1k one can see



the sliding-window attention mechanism for various Contextformer variants. For computational efficiency, during training, we used fixed-sized image patches and omitted the sliding-window operations. During inference, we set the size of the receptive field according to the sequence length ( $HW N_{cs}$ ) used for training.

### 3.4 Runtime optimization

Generally, autoregressive processes cannot be efficiently implemented on a GPU due to their serialized nature [21,26,34]. One commonly used approach [15,45] is to pad a set of sequences to have a fixed length, thus enabling the processing of multiple sequences in parallel during training (we refer to this method as Pad&Batch). A straightforward implementation of the sliding-window attention uses dynamic sequence lengths for every position of the window. We call this *dynamic sequence* processing (DS). The padding technique can be combined with the sliding-window by applying masking to the attention mechanism. However, one still needs to calculate attention for each padded element, which creates a bottleneck for the batched processing.

We propose a more efficient algorithm to parallelize the sliding-window attention. The first step of the algorithm calculates the processing order (or *priority*) for every position of the sliding window and then groups the positions with the same processing order for batch processing. One possible processing order is to follow the number of elements in each window and process them from the smallest to the largest number of elements. We refer to this method as Batched Dynamic Sequence (BDS). Note that transformers are sequence-to-sequence models; they simultaneously compute an output for each element in a sliding window and preserve causality of the sequence due to the masking. Therefore, we can skip computation of intermediate channel segments and calculate the output of the last channel segment for each spatial coordinate of the sliding window, which we refer to as skipping intermediate channel segments (SCS). It is worth mentioning that both the BDS and SCS methods can only be applied in the encoder, where all elements of the latent tensor are simultaneously accessible. For the decoder-side runtime optimization, we adopted the wavefront coding described in [30], which is similar to the partitioning slices used in VVC [42]. We use the same processing priority to the independent sliding windows along the same diagonal, which allows for simultaneous decoding of those windows. More information about the proposed runtime optimization algorithms can be found in the supplementary materials.

## 4 Implementation Details

We present a few variants of the Contextformer by changing its parameters  $\{L, d_e, N_{cs}, co\}$ , where  $L$ ,  $d_e$  and  $N_{cs}$  correspond to number of layers, embedding size and number of channel segments, and  $co$  corresponds to the coding order – either spatial-first (*sfo*) or channel-first (*cfo*). For all models, we used

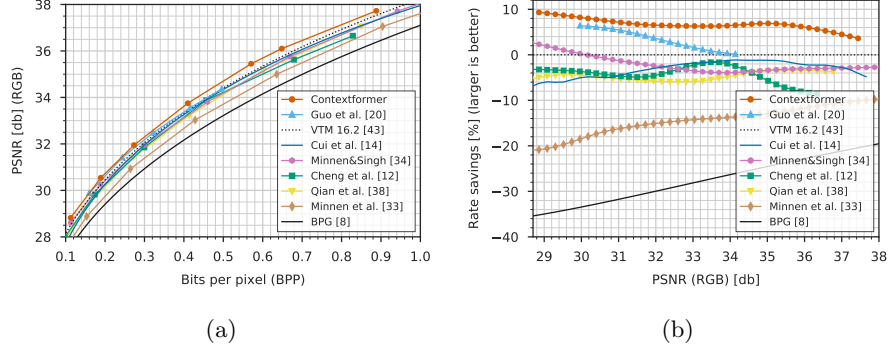
the same number of heads  $h$  and MLP size of  $d_{mlp}$ . We selected the base configuration of the Contextformer as  $\{L=8, d_e=384, d_{mlp}=4d_e, h=12, N_{cs}=4, co=cfo\}$ . More information about the architectural details can be found in the supplementary materials.

For training of all variants, we used random  $256 \times 256$  image crops from the Vimeo-90K dataset [49], batch size of 16, and ADAM optimizer [25] with the default settings. We trained our models for 120 epochs ( $\sim 1.2M$  iterations). Following [7], we used the initial learning rate of  $10^{-4}$  and reduced it by half every time the validation loss is nearly constant for 20 epochs. For this purpose, we used the validation set of Vimeo-90K. We selected mean-squared-error (MSE) as the distortion metric  $\mathbf{D}(\cdot)$  and trained a set of models with  $\lambda \in \{0.002, 0.004, 0.007, 0.014, 0.026, 0.034, 0.070\}$  to cover various bitrates. We also obtained models optimized for MS-SSIM [48] by finetuning the models trained for MSE, in the same fashion for  $\sim 500K$  iterations. By default, we selected both intermediate layer size  $N$  of the encoder and decoder, and bottleneck size  $M$  as 192. To increase the model capacity at the high target bitrates ( $\lambda_{5,6,7}$ ), we increased  $M$  from 192 to 312 by following the common practice [12,33,34].

## 5 Experimental Results

Unless specified otherwise, we used the base configuration (see Section 4) and tested its performance on the Kodak image dataset [18]. We set the spatial receptive field size to  $16 \times 16$  for the sliding-window attention. We compared the performance with the following models: the 2D context models by Minnen et al. [33] and Cheng et al. [12], the multi-scale 2D context model by Cui et al. [14], the channel-wise autoregressive context model by Minnen and Singh [34], the context model with an advanced global reference by Guo et al. [20], and the transformer-based context model by Qian et al. [38]. When the source was available, we ran the inference algorithms of those methods; in other cases, we took the results from the corresponding publications. For a fair comparison, we used the model from [20] without the GRDN post-filter [24]. Similarly, we used the serial model from [38], since it performs better and is more related to our approach. We also compared with the results achieved by classical compression algorithms such as BPG [8] and VTM 16.2 [43]. In order to test the generalization capability of our model, we also tested its performance on CLIC2020 [44], and Tecnick [2] datasets. Additionally, we present the impact of different configurations of our model on its complexity. More details about the performance comparison, and examples of compression artifacts can be found in the supplementary materials.

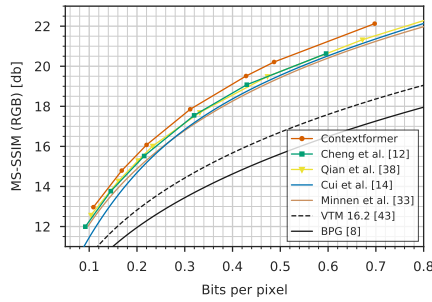
**Performance.** In Fig. 3a we show the rate-distortion performance of the Contextformer with a spatio-channel attention mechanism and the comparative performance of prior methods on the Kodak image dataset. Our method qualitatively surpasses the rest in terms of PSNR for all rate points under test. According to the Bjøntegaard Delta rate (BD-Rate) [9], our method achieves average saving of 6.9% over VTM 16.2, while the model in [20] provides 3.7% saving over



**Fig. 3.** Illustration of (a) the rate-distortion performance and (b) the rate savings relative to VTM 16.2 as a function of PSNR on the Kodak dataset showing the performance of our model compared to various learning-based and classical codecs.

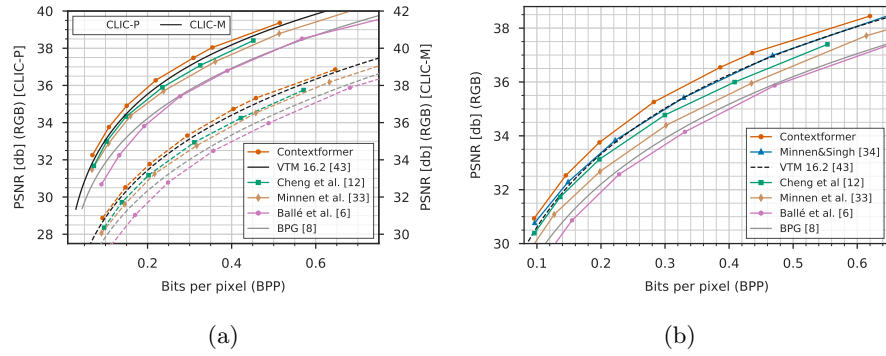
the same baseline. On average, our model saves 10% more bits compared to the multi-scale 2D context model in [14]. Notably, the only difference between our model and the one in [14] is the context and entropy modeling, and both methods have similar model sizes. The performance of our method and the prior in terms of the generalized BD-Rate metric [34] is shown in Fig. 3b. Our model achieves state-of-the-art performance by reaching 9.3% rate savings for low bitrate and 4% rate savings at the highest quality over VTM 16.2.

We also evaluated our model optimized for MS-SSIM [48]. Fig. 4 shows that our model also outperforms previous methods for this perceptual quality metric. On average, our models saves 8.7% more bits than Cheng et al. [12].



**Fig. 4.** Illustration of the rate-distortion performance in terms of MS-SSIM on Kodak dataset showing the performance of our model compared to various learning-based and classical codecs. All learned methods were optimized for MS-SSIM.

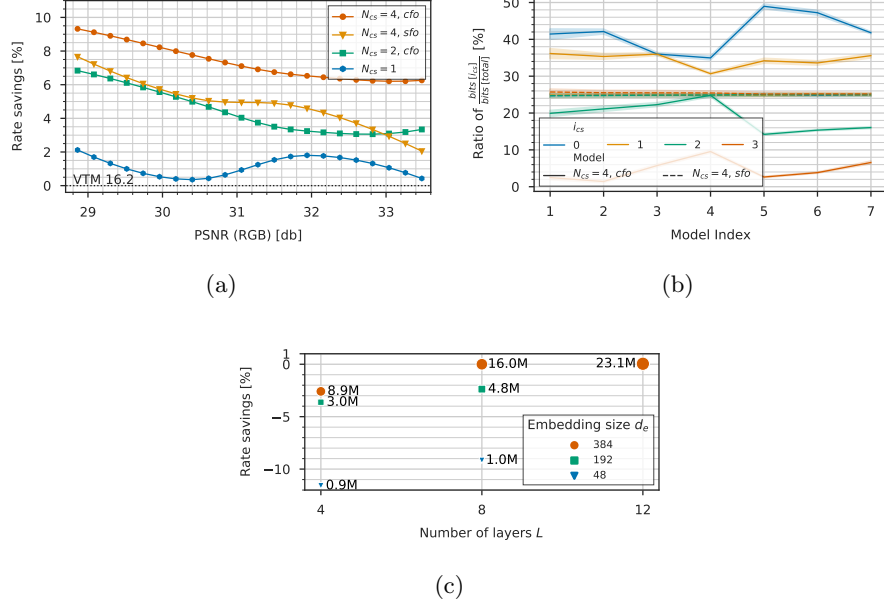
**Generalization Capability.** In order to show the generalization capability, we also evaluated our Contextformer on CLIC2020-Professional and -Mobile [44], and Tecnick [2] datasets. In terms of BD-Rate [9], our model achieves average savings of 9.8%, 5.8%, and 10.5% over VTM 16.2 on those datasets, respectively (see Fig. 6). Evaluating on the generalized BD-Rate metric [34] reveals that our method provides up to 11.9% and 6.6% relative bit savings over VTM 16.2 on CLIC2020-Professional and -Mobile datasets. On Tecnick dataset, the Contextformer saves up to 12.4% more bits over VTM 16.2.



**Fig. 5.** Comparison of the rate-distortion performance (a) on CLIC2020-Professional (solid line, left vertical axis) and CLIC2020-Mobile (dashed line, right vertical axis) datasets, and (b) on Tecnick dataset.

**Contextformer Variants.** In Fig. 6a we show the performance of our model for a varying number of channel segments  $N_{cs}$  and coding order. From the figure, one can see that increasing the number of channel segments increases the performance of the model since having more channel segments allows the models to explore more of the cross-channel dependencies. However, there is a trade-off between the number of segments and the complexity – the computational cost increases quadratically with raising  $N_{cs}$ . According to our observations, training the Contextformer with more than four segments increases training complexity too much to justify the minor performance gains.

On average, the Contextformer (*cfo*) outperforms the same model in a spatial-first-order (*sfo*) configuration, which highlights the greater importance of cross-channel dependencies than the spatial ones. For instance, in the Contextformer (*sfo*), the primary channel segment can only adopt spatial attention due to the coding order. In Fig. 6b we show the distribution of information in each channel segment. The Contextformer (*cfo*) stores more than 70% of all information in the first two channel segments, while in Contextformer(*sfo*) the information is almost equally distributed along with the segments. We observed that the



**Fig. 6.** Various ablation studies conducted with the Contextformer on Kodak dataset. (a) illustrates the rate savings relative to VTM 16.2 for the Contextformer with different number of channel segments  $N_{cs}$  and coding order. In (b), the percentile bit occupation per channel segment is shown for models with different coding orders. Each model index depicts a model trained for a specific  $\lambda$ . Notably, increasing the model capacity allocates more bits in the first two segments for  $cfo$  variant. (c) is the illustration of the average BD-Rate performance of various model sizes relative to base model. The annotations indicate the total number of entropy and context model parameters.

spatial-first coding provides a marginal gain in low target bitrates ( $bpp < 0.3$ ) and images with a uniformly distributed texture such as “kodim02” (in Kodak image dataset). This suggests that spatial dependencies become more pronounced in smoother images.

**Model Size.** Fig. 6c shows the performance of the Contextformers for different model sizes compared to the base configuration. Change in the network depth  $L$  and embedding size  $d_e$  have similar effects on the performance, whereas best performance can be achieved when both are increased. However, we observed that the return diminishes after a network depth of 8 layers. Since the base model already achieves state-of-the-art performance and further upscaling of models increases the complexity, we did not experiment with larger models. Note that the proposed network of [14], which our model is based on, has approximately the same total number of entropy and context model parameters (17M) as our

base model, whereas our model shows additionally 10.1% BD-Rate coding gain on Kodak dataset.

**Runtime Complexity.** Table 1 shows the encoding and decoding complexity of our model, some of the learning based- prior arts and VTM [43]. We tested the learning-based methods on a single NVIDIA Titan RTX, and ran the VTM [43] on Intel Core i9-10980XE Intel Core i9-10980XE. In our model, we used proposed BDS and SCS optimizations in the encoder and wavefront coding in the decoder. For low resolution images, our methods have close performance to the one of the 3D context. For 4K images, we observed even bigger benefits by the parallelization. The relative encoding time increases only 3x w.r.t. the one on the Kodak dataset, while the increase in number of pixels is 20-fold. Such speed-up shows that encoder methods with online rate-distortion optimization such as [51] have unexplored potential. Moreover, we achieve 9x faster decoding compared to a 3D context model with the proposed wavefront coding.

**Table 1.** Encoding and decoding time of different compression frameworks.

Method	Enc. Time [s]		Dec. Time [s]	
	Kodak	4K	Kodak	4K
DS	56	1240	62	1440
BDS (ours)	32	600	–	–
BDS&SCS (ours)	8	120	–	–
Wavefront (ours)	40	760	44	820
3D context [11]	4	28	316	7486
2D context [12]	2	54	6	140
VTM 16.2 [43]	420	950	0.8	2.5

## 6 Conclusion

In this work, we explored learned image compression architectures using a transformer-based context model. We proposed a context model that utilizes a multi-head attention mechanism and uses spatial dependencies in the latent space to model the entropy. Additionally, we also proposed a more generalized attention mechanism, spatio-channel attention, which can constitute a powerful context model. We showed that a compression architecture that employs the spatio-channel attention model achieves state-of-the-art rate-distortion performance.

While using an entropy model with spatio-channel attention brings noticeable gain, it also increases the runtime complexity. We addressed this issue by proposing an algorithm for efficient modeling while keeping the architecture unchanged. Future work will further investigate efficient attention mechanisms (e.g., [23,28,41]) aiming to bridge the gap to a real-time operation.

## References

1. Versatile Video Coding. Standard, Rec. ITU-T H.266 and ISO/IEC 23090-3 (Aug 2020)
2. Asuni, N., Giachetti, A.: Testimages: a large-scale archive for testing visual devices and basic image processing algorithms. In: STAG. pp. 63–70 (2014)
3. Ballé, J., Chou, P.A., Minnen, D., Singh, S., Johnston, N., Agustsson, E., Hwang, S.J., Toderici, G.: Nonlinear transform coding. *IEEE Journal of Selected Topics in Signal Processing* **15**(2), 339–353 (2020)
4. Ballé, J., Laparra, V., Simoncelli, E.P.: Density modeling of images using a generalized normalization transformation. In: 4th International Conference on Learning Representations, ICLR 2016 (2016)
5. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. In: 5th International Conference on Learning Representations, ICLR 2017 (2017)
6. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: International Conference on Learning Representations (2018)
7. Bégin, J., Racapé, F., Feltman, S., Pushparaja, A.: Compressai: a pytorch library and evaluation platform for end-to-end compression research. arXiv preprint arXiv:2011.03029 (2020)
8. Bellard, F.: Bpg image format (2015), accessed: 2022-06-01. URL <https://bellard.org/bpg>
9. Bjontegaard, G.: Calculation of average psnr differences between rd-curves. VCEG-M33 (2001)
10. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
11. Chen, T., Liu, H., Ma, Z., Shen, Q., Cao, X., Wang, Y.: End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Transactions on Image Processing* **30**, 3179–3191 (2021). <https://doi.org/10.1109/TIP.2021.3058615>
12. Cheng, Z., Sun, H., Takeuchi, M., Katto, J.: Learned image compression with discretized gaussian mixture likelihoods and attention modules. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7939–7948 (2020)
13. Cui, Z., Wang, J., Bai, B., Guo, T., Feng, Y.: G-vae: A continuously variable rate deep image compression framework. arXiv preprint arXiv:2003.02012 (2020)
14. Cui, Z., Wang, J., Gao, S., Guo, T., Feng, Y., Bai, B.: Asymmetric gained deep image compression with continuous rate adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10532–10541 (2021)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). <https://doi.org/10.18653/v1/N19-1423>, <https://aclanthology.org/N19-1423>
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth

- 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
17. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)
  18. Franzen, R.: Kodak lossless true color image suite (1999)
  19. Goyal, V.K.: Theoretical foundations of transform coding. *IEEE Signal Processing Magazine* **18**(5), 9–21 (2001)
  20. Guo, Z., Zhang, Z., Feng, R., Chen, Z.: Causal contextual prediction for learned image compression. *IEEE Transactions on Circuits and Systems for Video Technology* (2021)
  21. He, D., Zheng, Y., Sun, B., Wang, Y., Qin, H.: Checkerboard context model for efficient learned image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14771–14780 (2021)
  22. Jiang, Y., Chang, S., Wang, Z.: Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074* (2021)
  23. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are RNNs: Fast autoregressive transformers with linear attention. In: III, H.D., Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 119, pp. 5156–5165. PMLR (13–18 Jul 2020), <https://proceedings.mlr.press/v119/katharopoulos20a.html>
  24. Kim, D.W., Ryun Chung, J., Jung, S.W.: Grdn: Grouped residual dense network for real image denoising and gan-based real-world noise modeling. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 0–0 (2019)
  25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
  26. Koyuncu, A.B., Cui, K., Boev, A., Steinbach, E.: Parallelized context modeling for faster image coding. In: *2021 International Conference on Visual Communications and Image Processing (VCIP)*. pp. 1–5. IEEE (2021)
  27. Lee, J., Cho, S., Beack, S.K.: Context-adaptive entropy model for end-to-end optimized image compression. In: *6th International Conference on Learning Representations, ICLR 2018* (2018)
  28. Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S.: Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824* (2021)
  29. Li, D., Hu, J., Wang, C., Li, X., She, Q., Zhu, L., Zhang, T., Chen, Q.: Involution: Inverting the inheritance of convolution for visual recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12321–12330 (2021)
  30. Li, M., Ma, K., You, J., Zhang, D., Zuo, W.: Efficient and effective context-based convolutional entropy modeling for image compression. *IEEE Transactions on Image Processing* **29**, 5900–5911 (2020). <https://doi.org/10.1109/TIP.2020.2985225>
  31. Liu, H., Chen, T., Shen, Q., Ma, Z.: Practical stacked non-local attention modules for image compression. In: *CVPR Workshops*. p. 0 (2019)
  32. Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., Van Gool, L.: Conditional probability models for deep image compression. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4394–4402 (2018)
  33. Minnen, D., Ballé, J., Toderici, G.: Joint autoregressive and hierarchical priors for learned image compression. In: *NeurIPS* (2018)



34. Minnen, D., Singh, S.: Channel-wise autoregressive entropy models for learned image compression. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 3339–3343. IEEE (2020)
35. Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Intriguing properties of vision transformers. arXiv preprint arXiv:2105.10497 (2021)
36. Niu, Z., Zhong, G., Yu, H.: A review on the attention mechanism of deep learning. *Neurocomputing* **452**, 48–62 (2021). <https://doi.org/https://doi.org/10.1016/j.neucom.2021.03.091>, <https://www.sciencedirect.com/science/article/pii/S092523122100477X>
37. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: International Conference on Machine Learning. pp. 4055–4064. PMLR (2018)
38. Qian, Y., Sun, X., Lin, M., Tan, Z., Jin, R.: Entroformer: A transformer-based entropy model for learned image compression. In: International Conference on Learning Representations (2021)
39. Qian, Y., Tan, Z., Sun, X., Lin, M., Li, D., Sun, Z., Hao, L., Jin, R.: Learning accurate entropy model with global reference for image compression. In: International Conference on Learning Representations (2020)
40. Rissanen, J., Langdon, G.G.: Arithmetic coding. *IBM Journal of research and development* **23**(2), 149–162 (1979)
41. Roy, A., Saffar, M., Vaswani, A., Grangier, D.: Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics* **9**, 53–68 (2021)
42. Sullivan, G., Ohm, J., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding standard **22**, 1648–1667 (12 2012)
43. Team, J.V.E.: Versatile video coding (vvc) reference software: Vvc test model (vtm) (2022), accessed: 2022-06-01. URL [https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware\\_VTM](https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM)
44. Toderici, G., Theis, L., Johnston, N., Agustsson, E., Mentzer, F., Ballé, J., Shi, W., Timofte, R.: Workshop and challenge on learned image compression (clic2020)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
46. Wallace, G.K.: The jpeg still picture compression standard. *IEEE transactions on consumer electronics* **38**(1), xviii–xxxiv (1992)
47. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
48. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
49. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)* **127**(8), 1106–1125 (2019)
50. Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y.: Residual non-local attention networks for image restoration. In: International Conference on Learning Representations (2019), <https://openreview.net/forum?id=HkeGhoA5FX>
51. Zhao, J., Li, B., Li, J., Xiong, R., Lu, Y.: A universal encoder rate distortion optimization framework for learned compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1880–1884 (2021)

- 52. Zhou, J., Wen, S., Nakagawa, A., Kazui, K., Tan, Z.: Multi-scale and context-adaptive entropy model for image compression. arXiv preprint arXiv:1910.07844 (2019)
- 53. Ziv, J.: On universal quantization. IEEE Transactions on Information Theory **31**(3), 344–347 (1985)