Multi-Context Dual Hyper-Prior Neural Image Compression

Atefeh Khoshkhahtinat†, Ali Zafari†, Piyush M. Mehta‡,
Mohammad Akyash†, Hossein Kashiani†, Nasser M. Nasrabadi†
†Dept. of Computer Science & Electrical Engineering, West Virginia University, WV USA
‡Dept. of Mechanical & Aerospace Engineering, West Virginia University, WV USA
{ak00043,az00004}@mix.wvu.edu,piyush.mehta@mail.wvu.edu,
{ma00098,hk00014}@mix.wvu.edu,nasser.nasrabadi@mail.wvu.edu

***Abstract*—Transform and entropy models are the two core components in deep image compression neural networks. Most existing learning-based image compression methods utilize convolutional-based transform, which lacks the ability to model long-range dependencies, primarily due to the limited receptive field of the convolution operation. To address this limitation, we propose a Transformer-based nonlinear transform. This transform has the remarkable ability to efficiently capture both local and global information from the input image, leading to a more decorrelated latent representation. In addition, we introduce a novel entropy model that incorporates two different hyperpriors to model cross-channel and spatial dependencies of the latent representation. To further improve the entropy model, we add a global context that leverages distant relationships to predict the current latent more accurately. This global context employs a causal attention mechanism to extract long-range information in a content-dependent manner. Our experiments show that our proposed framework performs better than the state-of-the-art methods in terms of rate-distortion performance.**

***Index Terms*—Neural image compression, Transformer, Hyperprior, Global context, Entropy model, Attention**

I. Introduction

Neural image compression which is crucial to reduce the storage or transmission capacity has gained significant popularity in the computer vision field. Recently, learning-based image compression methods have achieved notable progress compared with hand-engineered codecs such as JPEG[1] and JPEG2000 [2]. Most of the image compression networks are based on the variational autoencoder (VAE) [3] which is comprised of the two key sub-networks. The first network is the main autoencoder, which is employed to map the image into a compact latent representation and transform the representation back into the original image. The second network is the entropy model which is required to estimate the entropy of the latent representation. It is apparent that improving these two sub-networks will lead to boosting compression performance. Obtaining more decorrelated and compact representation depends on the main autoencoder capability for extracting dependencies. In addition, building a powerful and accurate entropy model results in deriving less bit-rate.

A majority of learned image compression models employ a convolutional encoder to exploit image spatial correlations to achieve compressible latent representation; however, convolutional neural networks (CNNs) have some drawbacks. First, the convolution operation is only capable of leveraging local spatial dependencies within the local receptive field. Second, for various inputs, the weights of convolution filters are fixed after training. To address these issues, several solutions have been considered. Motivated by the success of attention in computer vision tasks, some works combined non-local attention modules with convolutional layers to extract long-range correlation for better compression efficiency [4], [5], [6]. However, such attention modules still do not affect the inherent local-aware property of the CNN architecture. With the rapid development of vision Transformers, Transformer-based autoencoders have been applied in image compression tasks [7]. In [7], Swin Transformer [8] is used to build the nonlinear transform which is further improved in terms of rate-distortion performance by enforcing frequency decomposition on a feature level [9]. Swin Transformer is designed upon the window-based attention and shifted-window-based attention to tackle the high computational complexity issue of ViT [10]. Despite the progress, the receptive field of the Swin Transformer is not wide enough to capture global information.

Since generating an optimum compressed bitstream relies on the entropy model, learning an accurate entropy model is vital. To this end, various works have been proposed. Ballé *et al.* [11] introduce a hyperprior that captures the spatial dependencies of the latent representation. The entropy model is approximated as a Gaussian scale mixture (GSM) where the scale parameters are predicted by the decoded hyperprior. Inspired by the PixelCNN [12], previous works [13], [14] enjoy the autoregressive component in the entropy model, which utilizes adjacent local latent representations that have already been decoded to predict the distribution of the uncoded latent. Qian *et al.*[15] present a global reference model to explore long-term connections. This algorithm searches throughout the previously decoded elements to find the most similar latent for predicting the current latent. While these approaches con-

This research is based upon work supported by the National Aeronautics and Space Administration (NASA), via award number 80NSSC21M0322 under the title of *Adaptive and Scalable Data Compression for Deep Space Data Transfer Applications using Deep Learning*.

1946-0759/23/$31.00 ©2023 IEEE
DOI 10.1109/ICMLA58977.2023.00091

tribute to estimating precisely the probability distribution of the latent representation, they still have limitations in modeling correlations. First, the hyperprior aims solely to exploit spatial dependencies of the latent space. Second, the proposed context models cannot leverage information from all the decoded elements to approximate the distribution parameters of the current element.

To address the aforementioned challenges, we propose a novel learned image compression network. The major contributions of this work can be summarized as follows:

1) The main autoencoder is built upon the Transformer termed Aggregated-window Transformer (AGWinT), which benefits from both local and aggregated-window Transformer blocks. Without increasing the computational complexity, the global information is extracted by using an aggregated-window Transformer block to achieve a more decorrelated latent representation.
2) The global context is incorporated into the entropy model to leverage long-range correlations in the latent space. This type of entropy model provides accurate probability estimation of the latent representation.
3) Two hyperpriors are introduced, one of them attempts to model inter-channel dependencies and the other one is utilized to extract inter-spatial dependencies among the elements of the latent representation via the attention mechanism.

II. Related Work

A. Learned Image Compression

Nowadays, the cutting-edge approach for lossy image compression primarily involves a combination of variational autoencoders (VAEs) and transform coding techniques [16]. This innovative approach goes beyond traditional linear transformations and incorporates learned non-linear transformation blocks. In the learned image compression framework, at the encoder side, the input image is passed through the analysis transform $g_a(\boldsymbol{x};\boldsymbol{\phi})$ to generate its compact latent representation $\boldsymbol{y}$. Then, to reduce the number of bits required to store or transmit the image, the continuous latent embeddings $\boldsymbol{y}$ are quantized to discrete symbols $\hat{\boldsymbol{y}}$, which will be subjected to lossless coding algorithms to obtain bitstream files. To reconstruct the compressed image $\hat{\boldsymbol{x}}$, the decoder retrieves the quantized latent values $\hat{\boldsymbol{y}}$ from the encoded bitstream. It then employs the synthesis transform $g_s(\hat{\boldsymbol{y}};\boldsymbol{\theta})$, which is designed to closely approximate the inverse of the analysis transform [17]. The process can be summarized as follows:

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

Ballé *et al.* [18] proposed the factorized density model, which is shared between the encoder and decoder, to effectively exploit the remaining coding redundancy within the latent space. This model is designed to estimate the latent distribution by employing local histograms. In a subsequent study, Ballé *et al.* [11] introduced hidden latent variables, represented as $\hat{\boldsymbol{z}}$, to serve as side information for capturing spatial dependencies among the elements of the latent representation. This model employs additional analysis and synthesis transforms, represented as $h_a(\hat{\boldsymbol{z}};\boldsymbol{\phi_h})$ and $h_s(\hat{\boldsymbol{z}};\boldsymbol{\theta_h})$ respectively, to conditionally model the distribution with respect to the side information. The distribution of $\hat{\boldsymbol{z}}$ is approximated using a factorized density model and $p_{\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}})$ is modeled as a zero-mean Gaussian distribution. The entropy model mechanism of this proposed network can be summarized as:

$$\begin{aligned} \boldsymbol{z} &= h_a(\boldsymbol{y};\boldsymbol{\phi}_h) \\ \hat{\boldsymbol{z}} &= Q(\boldsymbol{z}), \\ P_{\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}) &\sim \mathcal{N}(0,\, h_s(\hat{\boldsymbol{z}};\boldsymbol{\theta}_h)). \end{aligned} \tag{2}$$

B. Autoregressive-based Entropy Model

As the entropy model approaches a closer approximation to the true distribution of the latent representation, the resulting compressed file exhibits a lower bit rate. Minnen *et al.* [13] leveraged the concepts from PixelCNN [12] to expand the Gaussian scale mixture-based (GSM) entropy model into a Gaussian mixture model (GMM). This extension involves integrating a context block, which adopts an autoregressive model, along with a hyperprior network. Within this entropy model, the mean and scale parameters of the latent representation distribution are dependent on the hyperprior and causal context associated with each latent $\hat{y}_i$. Therefore, the predicted Gaussian parameters for the distribution of each latent element $\hat{y}_i$ can be expressed as follows:

$$(\mu_i, \sigma_i) = g_{ep}(g_{cm}(\hat{\boldsymbol{y}}_{<\boldsymbol{i}};\boldsymbol{\theta_{cm}}), h_s(\hat{\boldsymbol{z}};\boldsymbol{\theta_h}), \boldsymbol{\theta_{ep}}), \tag{3}$$

The context model $g_{cm}(.)$ is designed by employing a 2D masked convolution. $g_{ep}(.)$ shows the entropy parameter function and $\hat{\boldsymbol{y}}_{<\boldsymbol{i}}$ denotes the causal decoded neighbors of current latent element $\hat{y}_i$.

In [19], [20], [21], researchers proposed a multi-scale context model that utilized multiple masked convolutions with varying kernel sizes. This allowed the model to learn diverse spatial dependencies simultaneously. On the other hand, [22], [23], [24] used 3D masked convolutions to leverage both cross-channel correlations and spatial dependencies together. Qian *et al.* [15] introduced a global reference context model that aims to exploit long-range dependencies. This component examines previously decoded elements to identify the most similar latent element which is then utilized to estimate the distribution parameters of the current latent. He developed a checkerboard context model which utilizes information from closer neighboring latents to predict the uncoded latent.

C. Transformers

The Transformer [25] was originally introduced in the field of natural language processing (NLP) and had a profound impact on the NLP domain. Its remarkable success in NLP applications inspired researchers to extend the Transformer architecture to computer vision (CV) tasks, including object detection[26], image classification [27], semantic segmentation

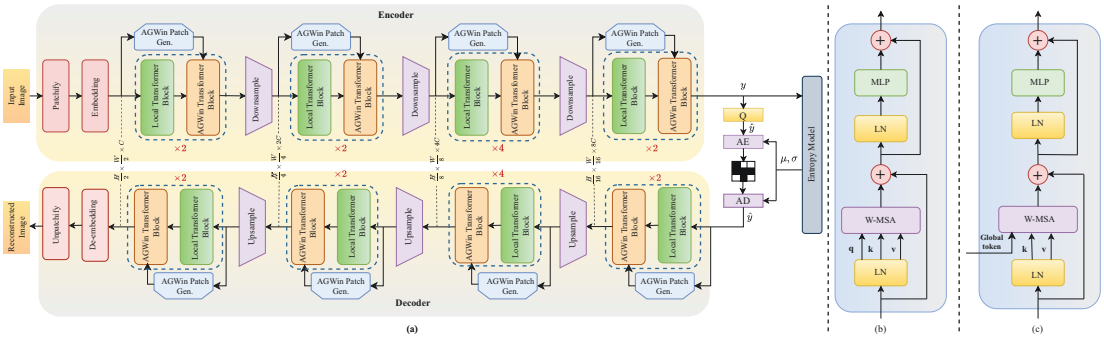

Fig. 1. Overview of our proposed AGWinT-based architecture. (b) Local Transformer block. (c) Aggregated-window Transformer block.

[28], data compression [29], and time-series analysis [30]. ViT [31] is the first pure Transformer architecture that has demonstrated superior performance compared to convolutional neural network (CNN) models like ResNet [32], [33] and attain state-of-the-art performance on multiple image classification benchmarks [34], [35], [36], [37]. Vit framework begins by partitioning the input image into non-overlapping patches. These patches are then passed through a Transformer architecture, which leverages self-attention mechanisms [38] to learn the uniform short and long-range relationships within the image more effectively. However, ViT is effective at capturing contextual information, its monolithic architecture and the quadratic computational complexity associated with self-attention pose significant obstacles in promptly deploying the model for vision tasks, where processes high-resolution images. Several works [8], [39], [40], [41], particularly the Swin Transformer [8], have been proposed to focus on achieving a balance between short and long-range spatial dependencies through using hierarchical architectures. The Swin Transformer architecture introduces a solution where self-attention is computed within local windows, and a shifted-window self-attention mechanism is employed to model interactions across different windows. Despite significant progress, the self-attention mechanism struggles to capture long-range information due to the constrained receptive field of local windows. Additionally, window-shifted attention only consider small nearby regions around each window when computing interactions. To overcome this limitation, Focal Transformer [42] is introduced by implementing more intricate self-attention modules, but this improvement comes at the cost of increased computational complexity.

III. Proposed Method

Fig.1 shows the overall architecture of our proposed model. Inspired by the GC ViT [43], we construct a main autoencoder of the compression network based on an Aggregated-Window Transformer (AGWinT). The suggested entropy model is employed to efficiently encode the quantized latent representations into a stream of ones and zeros.

A. Encoder-Decoder

*1) **Encoder**:* A hierarchical AGWinT is adopted as the encoder for obtaining feature representations at multiple resolutions by reducing spatial dimensions while increasing embedding dimensions by factors of two at 4 stages. Firstly, The input image $x \in R^{H\times W\times 3}$ is fed into the patchify layer. which consists of a 3×3 convolutional layer with a stride of 2 and the padding operation, to generate the overlapping patches. Then, the resulting patches are mapped into an embedding space with dimension C by using another 3×3 convolutional layer. This layer is named embedding. To extract semantic features, each AGWinT stage is composed of multiple local and aggregated-window Transformer blocks. After stages 1,2 and 3 a downsampling block is applied to halve the spatial resolution of the feature map and double the channel number of the feature map.

*2) **Decoder**:* An AGWinT decoder is the mirror of the encoder. We design the AGWinT decoder by replacing the patchify block with a unpatchify block, the embedding layer with a de-embedding layer, and the downsampling block with an upsampling block.

*3) **Downsampling block**:* The role of this block is to generate hierarchical representation. The input feature map is passed through the Fused-MBConv [44] to inject inductive bias into the network and model inter-channel correlations. Then, the convolution layer with a kernel size of 3 and stride of 2 is utilized to downsample the spatial feature resolution by 2, whereas the number of channels is doubled. The Fused-MBConv process can be written as:

$$\begin{aligned} \hat{x} &= DephWiseConv_{3*3}(x), \\ \hat{x} &= GELU(\hat{x}), \\ \hat{x} &= SE(\hat{x}), \\ x &= Conv_{1*1}(\hat{x}) + x. \end{aligned} \tag{4}$$

Where SE [45] refers to the Squeeze and Excitation block and GELU represents the Gaussian Error Linear Unit function.

*4) **Local Transformer Block**:* Like the Swin Transformer block [8], the local Transformer block splits the input into local windows, then a window-based multi-head self-attention (W-MSA) computes the self-attention individually for each

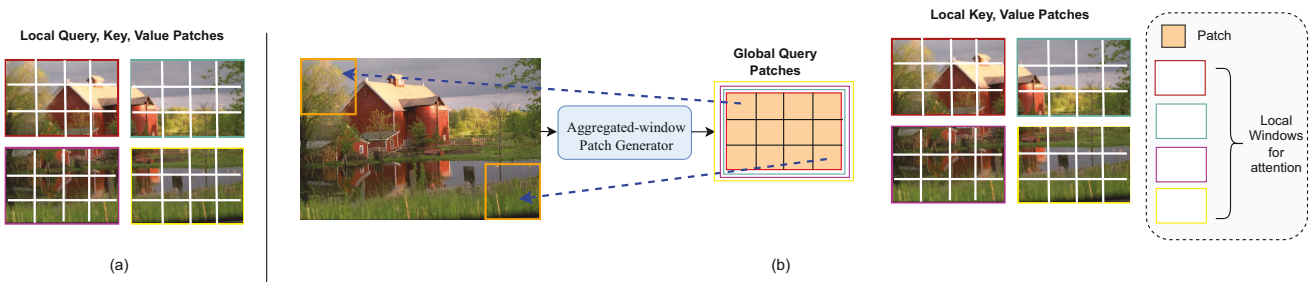

Fig. 2. An illustration of the local and global attention mechanisms. (a) The local attention mechanism operates independently within each window, where each window possesses its specific query, key, and value patches. (b) The global attention mechanism extracts query patches from the entire input feature map, aggregating information from all windows. The global queries interact with local key and value patches, facilitating the capture of long-range information through cross-region interaction.

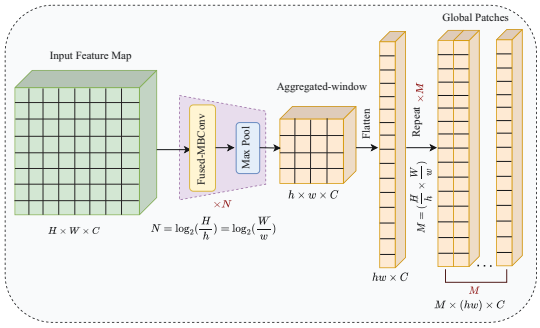

Fig. 3. The diagram of the aggregated-window patch generator. The input feature map with dimensions H and W is fed into the module, which consists of a stack of N identical layers. Each layer is composed of a Fused-MBConv block followed by a max pooling layer, which generates an aggregated window features. The number of layers N is chosen such that the extracted global feature dimension matches the size of the local window $h \times w$. Subsequently, the global window is flattened and duplicated M times, where $M = (\frac{H}{h} \times \frac{W}{w})$ corresponds to the number of local windows, to create the global query patches.

window. Computing local self-attention leads to extracting short-range information.

*5) **Aggregated-window Transformer Block**:* As shown in Fig. 2, in contrast to the local Transformer block, where each local window has its specific query patches, the aggregated-window Transformer block employs global query patches to share across all windows to interact with local keys and values. Global query patches include information from the whole input feature map which is produced by an aggregated window patch generator once at every stage.

*6) **Aggregated-window Patch Generator**:* This module consists of a stack of N identical layers where each layer is composed of Fused-MBConv followed by a max pooling layer to extract global features, as displayed in Fig. 3. The number of layers is set to N so that the output feature dimension matches the size of the local window. Then, the output is flattened and repeated to the number of windows to form the global query patches.

B. *Proposed Entropy model*

As depicted in Fig. 4, the proposed entropy model is comprised of hyperprior-networks, local context, global context, and parameter network. The outputs of hyper-networks, local context, and global context are combined together and fed into the parameter network, comprising of 1×1 convolution layers, to obtain the distribution parameters, i.e., μ,σ. The process of deriving distribution parameters can be written as:

$$\begin{aligned}(\mu, \sigma) &= pm(\phi_{sp}, \phi_{ch}, \psi_l, \psi_g),\\ \phi_{sp} = sh(\boldsymbol{y}), \phi_{ch} &= ch(\boldsymbol{y}), \psi_l = lc(\hat{\boldsymbol{y}}_{<\boldsymbol{i}}), \psi_g = gc(\hat{\boldsymbol{y}}_{<\boldsymbol{i}}),\end{aligned} \tag{5}$$

where $pm(.)$, $sh(.)$, $ch(.)$, $lc(.)$, and $gc(.)$ correspond to parameter network function, spatial-aware hyperprior model, channel-aware hyperprior model, local context, and global context, respectively. The local context is implemented using masked convolution with a kernel size of 5×5, as inspired by previous work [13]. Detailed information about the other blocks will be presented in the following sections.

*1) **Global Context Block**:* This block is introduced to effectively exploit the global context information for obtaining a more accurate entropy model. It is based on an autoregressive model that uses all the previously decoded latents to predict the current decoding latent. As shown in Fig. 5(a), masked attention is used to model the global causal correlation.

Due to the fact that the latent representations are serially decoded, the latents are unfolded into patches and then masked, represented as $p \in [H \times W, k \times k \times C]$ (where H, W, k, and C correspond to height, width, unfold kernel size, and channels dimension, respectively) to calculate the masked attention. The normalized masked patches are considered as query tokens and key tokens (neglecting the magnitude effect of patches in calculating attention score), while regular masked patches are taken into account as value tokens. The formulation for the masked-multi-head attention can be expressed as follows:

$$\begin{aligned}Attention(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) &= Concat(head_1, ..., head_m)\boldsymbol{W},\\ head_i(\boldsymbol{q_i}, \boldsymbol{k_i}, \boldsymbol{v_i}) &= softmax(\boldsymbol{q_i}\boldsymbol{k_i}^{\boldsymbol{T}} \odot \boldsymbol{M})\boldsymbol{v_i},\end{aligned} \tag{6}$$

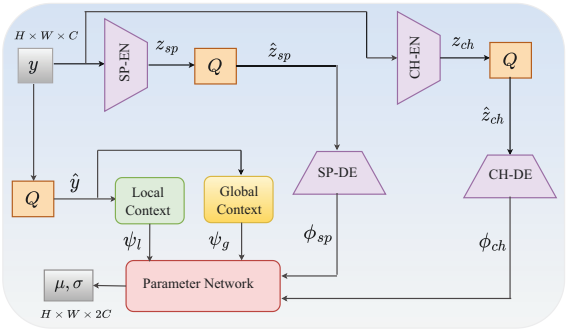

Fig. 4. Diagram of the proposed entropy model. Entropy coding operations for z_{sp} and z_{sp} are excluded for simplicity.

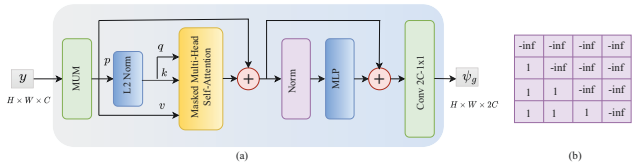

Fig. 5. Global Context Block: (a) The Transformer-based architecture, where MUM stands for the masked unfolding module. (b) An example of causal mask M.

where $q_i, k_i, v_i \in R^{HW \times \frac{K^2C}{h}}$ are the queries, keys, and values for the i-th head, respectively. $M \in R^{HW \times HW}$ represents a causal mask whose lower-triangular elements are one and the remaining elements are minus infinity.

*2) **Hperprior**:* The newly proposed hyperprior is split into two distinct groups: the Channel-aware hyperprior $\hat{\boldsymbol{z}}_{\boldsymbol{ch}}$ and the spatial-aware hyperprior $\hat{\boldsymbol{z}}_{\boldsymbol{sp}}$. Each of these hyperpriors plays a crucial role in modeling different aspects of the dependencies within the latent representation $\boldsymbol{y}$. The Channel-aware hyperprior $\hat{\boldsymbol{z}}_{\boldsymbol{ch}}$ is specifically designed to capture the inter-channel relationships at each spatial location of the latent representation, while the spatial-aware hyperprior $\hat{\boldsymbol{z}}_{\boldsymbol{sp}}$ aims to model spatial dependencies within each channel of the latent representation. Both types of hyperprior are computed in parallel and complement each other, working together to jointly extract dependencies within the latent representation $\boldsymbol{y}$.

Channel-aware Hyperprior: We have developed the Channel-aware hyperprior model, which effectively captures inter-channel dependencies within the latent representation. As illustrated in Fig. 6(a), this model is constructed based on an autoencoder architecture and utilizes channel-attention modules [46] and stacks of 1×1 convolution layers to achieve its objective. The Encoder part generates the channel-aware hyperprior $\boldsymbol{z}_{\boldsymbol{ch}} \in R^{H \times W \times \frac{C}{16}}$ from the input $\boldsymbol{y} \in R^{H \times W \times C}$, reducing the number of channels in the latent space while preserving spatial resolution. On the Decoder side, it takes the channel-aware hyperprior $\boldsymbol{z}_{\boldsymbol{ch}}$ as input and generates the desired output $\boldsymbol{\phi}_{\boldsymbol{ch}} \in R^{H \times W \times 2C}$.

Spatial-aware Hyperprior: Spatial-aware hyperprior

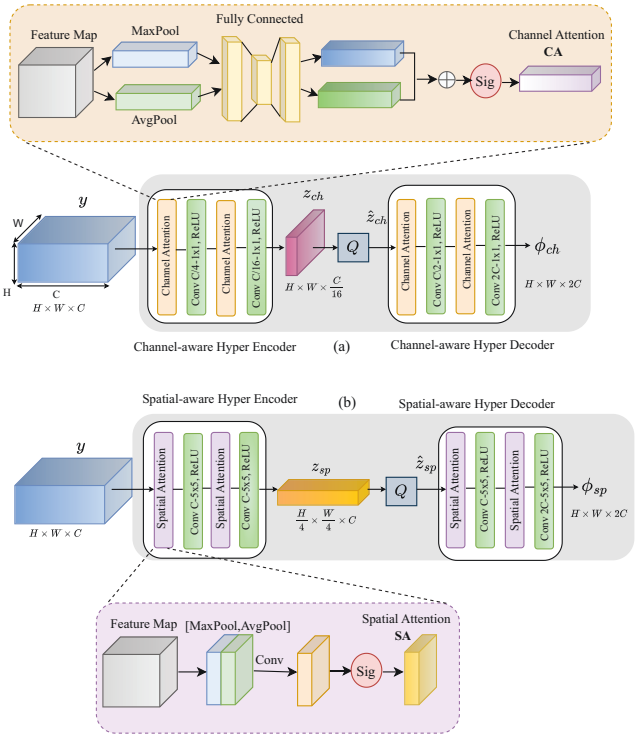

Fig. 6. (a) The architecture of Channel-aware Hyperprior Network. (b) The architecture of Spatial-aware Hyperprior Network. Sig represents the sigmoid function.

model is designed to extract spatial interactions among the latent representation elements. The hyperprior $\boldsymbol{z}_{\boldsymbol{sp}} \in R^{\frac{H}{4} \times \frac{W}{4} \times C}$, generated by utilizing this model, doesn't change the channel number of the latent representation; rather, it decreases the spatial resolution. The network adopted to output the spatial-aware hyperprior takes the form of an autoencoder architecture and consists of 5×5 convolution layers and spatial-attention [46] blocks. As depicted in Fig. 6(b), the latent representation $\boldsymbol{y}$ is fed into an encoder block to obtain the spatial-aware hyperprior $\boldsymbol{z}_{\boldsymbol{sp}}$, and the decoder component produces the output $\boldsymbol{\phi}_{\boldsymbol{sp}} \in R^{H \times W \times 2C}$.

C. *Quantization*

To facilitate end-to-end training feasibility, the quantization process requires a replacement with a soft differentiable function. In this study, we utilize uniform noise, which is added to the latent representations, as an approximation for the hard quantization operation [18]. Consequently, the conditional probability of each latent, $\hat{y}_i$ is modeled as a univariate Gaussian with its location and scale convolved with a unit uniform distribution:

$$P_{\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}, \boldsymbol{\theta}) = \prod_{i=1} (\mathcal{N}(\mu_i, {\sigma_i}^2) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2}))(\hat{y}_i), \quad (7)$$

where location μ_i and scale σ_i are determined by the entropy model, $\boldsymbol{\theta}$ denotes the parameters of the entropy model, $\hat{\boldsymbol{z}} =$

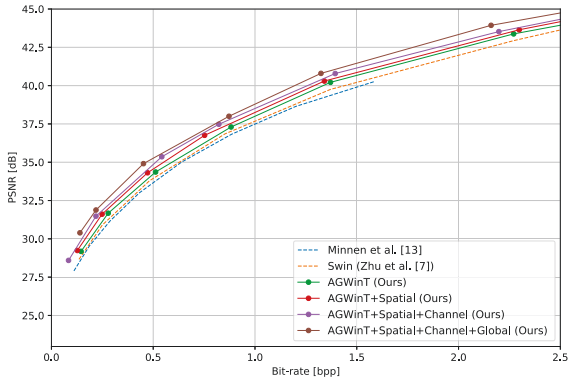

Fig. 7. Rate-distortion performance comparison of different methods on the Kodak dataset, comprising 24 images. Distortion is measured by PSNR.

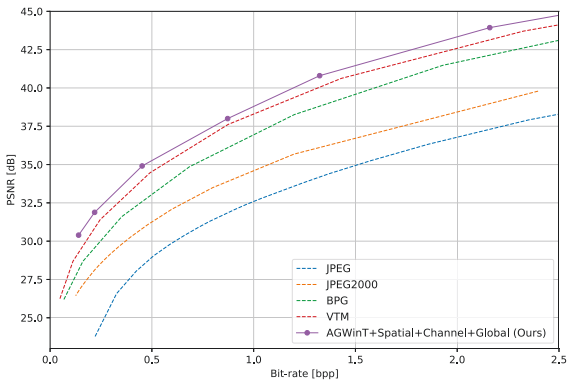

Fig. 8. Comparison of our full model with traditional image codecs in terms of rate-distorsion performance. The RD values are averaged over 24 images from the Kodak test set.

$\{\hat{\boldsymbol{z}}_{\boldsymbol{sp}}, \hat{\boldsymbol{z}}_{\boldsymbol{ch}}\}$ is the quantized hyperpriors.

D. *Loss Function*

Any learned image compression network attempts to jointly minimize the trade-off of rate and distortion, controlled by a Lagrangian multiplier which can be formulated as:

$$R + \lambda D, \tag{8}$$

where R denotes the estimated rate of the quantized latent representation and D is the distortion between the input image and reconstructed image. Since the probability distribution of the latent representation is conditionally estimated by hyperpriors, the rate term R consists of the estimated entropy of the latent representation and two hyperpriors. Therefore, the rate term can be described as below:

$$\boldsymbol{E}_{x\sim p_X}[-\log_2 P_{\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}) - \log_2 P_{\hat{\boldsymbol{z}}_{\boldsymbol{sp}}}(\hat{\boldsymbol{z}}_{sp})) - \log_2 P_{\hat{\boldsymbol{z}}_{\boldsymbol{ch}}}(\hat{\boldsymbol{z}}_{\boldsymbol{ch}}))], \tag{9}$$

where $\hat{\boldsymbol{z}} = \{\hat{\boldsymbol{z}}_{\boldsymbol{sp}}, \hat{\boldsymbol{z}}_{\boldsymbol{ch}}\}$ includes both hyperpriors and the non-parametric fully factorized entropy model [47] is employed to approximate the probability distribution of hyperpriors.

IV. EXPERIMENTS

A. *Implementation Details*

For our experiments, we utilize a combined set of $5,285$ high-resolution images from datasets DIV2K, Flickr2K, and CLIC2021 [48] as our training dataset. The evaluation set consists of images from the Kodak dataset [49]. During the training phase, we use a batch size of 16, which includes randomly cropped 256×256 patches. To cover a wide range of bitrates, we set the hyper-parameter λ to take values from the set $\{0.00125, 0.005, 0.01, 0.02, 0.04, 0.08\}$. All the models are trained for 100 epochs using the Adam optimizer [50] for a total of 2.4 million steps. The initial learning rate is set to 10^{-4} and gradually decreased to 1.2×10^{-6} during the training process.

B. *Ablation Study*

In this section, we conduct several ablation studies to investigate the effectiveness of our proposed Transformer-based autoencoder and the entropy model. In the first study, to assess the performance of AGWinT, we choose the Minnen *et al.* model [13], whose entropy model is comprised of hyperprior and local context, and replace the main convolutional autoencoder with the Swin Transformer-based autoencoder and AGWinT Transformer-based autoencoder. The results demonstrate that the AGWinT-based network outperforms both the convolutional and Swin-based architectures in terms of rate-distortion performance. This improvement can be attributed to the fact that AGWinT is capable of effectively capturing long-distance and local correlations, leading to enhanced compression performance.

In the second study, to validate the impact of decomposing hyperprior and adding spatial and channel attention, we build three different hyperprior models: existing hyperprior, the whole compression model is referred to as the "AGWinT" framework, "spatial-aware hyperprior", and "spatial-aware hyperprior + channel-aware hyperprior". Each of the mentioned hyperprior frameworks is combined with local context to form the entropy model. These entropy models are then integrated into compression models based on the AGWinT autoencoder. As depicted in Fig. 7, the "spatial-aware hyperprior + channel-aware hyperprior" model exhibits better performance compared to the other models. This enhancement is related to the improved modeling capabilities of the combined spatial and channel attention in exploiting the complex dependencies within the latent representations.

In the third study, we examine the effect of incorporating global context in the entropy model. Fig. 7 demonstrates that global context contributes to achieving a better result. This evaluation emphasizes that combining global context with local context leads to precise modeling correlations of the latent representation.

Fig. 9. Visual comparison of our proposed framework with other conventional image codecs, using the bit-rate/distortion [bpp↓/PSNR↑] as the metric. The results demonstrate that our method achieves lower distortion in terms of PSNR compared to the other codecs, highlighting its superior capability in preserving image quality.

C. *Comparison with Traditional Codecs*

We compare the rate-distortion (RD) performance of our full model, denoted as "AGWinT+Spatial+Channel+Global", with traditional video compression standards, such as JPEG [1], JPEG2000 [2], HEVC-Intra (BPG) [51], and VVC-Intra (VTM) [52]. The distortion is measured by the Peak Signal-to-Noise Ratio (PSNR). Fig. 8 illustrates that our proposed model outperforms the JPEG, JPEG2000, and BPG codecs in terms of RD performance. Additionally, it showed comparable performance to VTM.

D. *Visual Quality*

In Fig. 9, we present an example of a reconstructed image (kodim19.png) obtained using our proposed framework, as well as the compression standards JEPG [53], JEPG2000, BPG, and VTM. The visual comparison reveals that our reconstructed image retains significantly more details while achieving a similar bpp value. This observation highlights the superior performance of our approach in preserving image quality.

V. Conclusion

This paper presents a novel image compression algorithm using a Transformer-based autoencoder and a new entropy model. The main autoencoder consists of a local Transformer and aggregated-window Transformer blocks which are responsible to capture both long-term and short-term relationships for achieving a more compressible and decorrelated representation of the input image. Our proposed entropy model is comprised of two different hyperpriors: channel-aware hyperprior and spatial-aware hyperprior. These two hyperpriors aim to model the remaining spatial and cross-channel redundancies in the latent representation. The global context is integrated with local context and hyperpriors to exploit global information which assists in accurately estimating the distribution of the latent representation. Our experimental results demonstrate that our proposed network boosts compression efficiency.

References

[1] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, 1991.

[2] D. S. Taubman and M. W. Marcellin, *JPEG2000 - image compression fundamentals, standards and practice*, ser. The Kluwer international series in engineering and computer science. Kluwer, 2002.

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

[4] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, and Y. Wang, "End-to-end learnt image compression via non-local attention optimization and improved context modeling," *IEEE Transactions on Image Processing*, 2021.

[5] H. Fu, F. Liang, J. Lin, B. Li, M. Akbari, J. Liang, G. Zhang, D. Liu, C. Tu, and J. Han, "Learned image compression with discretized gaussian-laplacian-logistic mixture model and concatenated residual modules," *arXiv preprint arXiv:2107.06463*, 2021.

[6] R. Zou, C. Song, and Z. Zhang, "The devil is in the details: Window-based attention for image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[7] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *International Conference on Learning Representations*, 2022.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE International Conference on Computer Vision*, 2021.

[9] A. Zafari, A. Khoshkhahtinat, P. Mehta, M. S. E. Saadabadi, M. Akyash, and N. M. Nasrabadi, "Frequency disentangled features in neural image compression," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 2815–2819.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[11] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

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

[13] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, 2018.

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

[15] Y. Qian, Z. Tan, X. Sun, M. Lin, D. Li, Z. Sun, L. Hao, and R. Jin, "Learning accurate entropy model with global reference for image compression," in *International Conference on Learning Representations*, 2021.

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

[17] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, "Nonlinear transform coding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 339–353, 2020.

[18] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[19] J. Zhou, S. Wen, A. Nakagawa, K. Kazui, and Z. Tan, "Multi-scale and context-adaptive entropy model for image compression," *arXiv preprint arXiv:1910.07844*, 2019.

[20] Z. Cui, J. Wang, B. Bai, T. Guo, and Y. Feng, "G-vae: A continuously variable rate deep image compression framework," *arXiv preprint arXiv:2003.02012*, vol. 2, no. 3, 2020.

[21] Z. Cui, J. Wang, S. Gao, T. Guo, Y. Feng, and B. Bai, "Asymmetric gained deep image compression with continuous rate adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 532–10 541.

[22] 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.

[23] H. Liu, T. Chen, Q. Shen, and Z. Ma, "Practical stacked non-local attention modules for image compression." in *CVPR Workshops*, 2019, p. 0.

[24] 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.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[28] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8741–8750.

[29] A. Zafari, A. Khoshkhahtinat, P. M. Mehta, N. M. Nasrabadi, B. J. Thompson, D. Da Silva, and M. S. Kirk, "Attention-based generative neural image compression on solar dynamics observatory," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2022, pp. 198–205.

[30] M. A. Farahani, M. McCormick, R. Gianinny, F. Hudacheck, R. Harik, Z. Liu, and T. Wuest, "Time-series pattern recognition in smart manufacturing systems: A literature review and ontology," *arXiv preprint arXiv:2301.12495*, 2023.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[33] S. Mohamadi, G. Doretto, and D. A. Adjeroh, "Fussl: Fuzzy uncertain self supervised learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2799–2808.

[34] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[35] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 181–196.

[36] K. Safavigerdini, K. Nouduri, R. Surya, A. Reinhard, Z. Quinlan, F. Bunyak, M. R. Maschmann, and K. Palaniappan, "Predicting mechanical properties of carbon nanotube (CNT) images using multi-layer synthetic finite element model simulations," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 3264–3268.

[37] M. Tanhaeean, N. Nazari, S. H. Iranmanesh, and M. Abdollahzade, "Analyzing factors contributing to covid-19 mortality in the united states using artificial intelligence techniques," *Risk Analysis*, vol. 43, no. 1, pp. 19–43, 2023.

[38] N. A. Talemi, H. Kashiani, S. R. Malakshan, M. S. E. Saadabadi, N. Najafzadeh, M. Akyash, and N. M. Nasrabadi, "Aaface: Attribute-aware attentional network for face recognition," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 1940–1944.

[39] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 124–12 134.

[40] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 459–479.

[41] P. Ansari Bonab, J. Holland, and A. Sargolzaei, "An observer-based control for a networked control of permanent magnet linear motors under a false-data-injection attack," in *2023 6th IEEE Conference on Dependable and Secure Computing*. IEEE, 2023, in press.

[42] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal attention for long-range interactions in vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30 008–30 022, 2021.

[43] A. Hatamizadeh, H. Yin, J. Kautz, and P. Molchanov, "Global context vision transformers," *arXiv preprint arXiv:2206.09959*, 2022.

[44] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*, 2021.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[46] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

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

[48] "CLIC · challenge on learned image compression,," Available at http://compression.cc/tasks/, 2022.

[49] "Kodak image dataset,," Available at https://r0k.us/graphics/kodak/, 2022.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, 2015.

[51] F. Bellard, "BPG image format," *URL https://bellard. org/bpg*, vol. 1, no. 2, p. 1, 2015.

[52] "Versatile Video Coding Reference Software," Available at https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM, 2022.

[53] G. K. Wallace, "The jpeg still picture compression standard," *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.