

This work was written as part of one of the author's official duties as an Employee of the United States Government and is therefore a work of the United States Government. In accordance with 17 U.S.C. 105, no copyright protection is available for such works under U.S. Law.

Public Domain Mark 1.0

<https://creativecommons.org/publicdomain/mark/1.0/>

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

Multi-spectral Entropy Constrained Neural Compression of Solar Imagery

Ali Zafari[†], Atefeh Khoshkhahtinat[†], Piyush M. Mehta[‡], Nasser M. Nasrabadi[‡]

Barbara J. Thompson[§], Michael S. F. Kirk[§], Daniel da Silva[§]

[†]Department of Computer Science & Electrical Engineering, West Virginia University, WV USA

[‡]Department of Mechanical & Aerospace Engineering, West Virginia University, WV USA

[§]NASA Goddard Space Flight Center, MD USA

{az00004, ak00043}@mix.wvu.edu, {piyush.mehta, nasser.nasrabadi}@mail.wvu.edu

{barbara.j.thompson, michael.s.kirk, daniel.e.dasilva}@nasa.gov

Abstract—Missions studying the dynamic behaviour of the Sun are defined to capture multi-spectral images of the sun and transmit them to the ground station in a daily basis. To make transmission efficient and feasible, image compression systems need to be exploited. Recently successful end-to-end optimized neural network-based image compression systems have shown great potential to be used in an ad-hoc manner. In this work we have proposed a transformer-based multi-spectral neural image compressor to efficiently capture redundancies both intra/inter-wavelength. To unleash the locality of window-based self attention mechanism, we propose an inter-window aggregated token multi head self attention. Additionally to make the neural compressor autoencoder shift invariant, a randomly shifted window attention mechanism is used which makes the transformer blocks insensitive to translations in their input domain. We demonstrate that the proposed approach not only outperforms the conventional compression algorithms but also it is able to better decorrelates images along the multiple wavelengths compared to single spectral compression.

Index Terms—multi-spectral neural image compression, inter-window token aggregation, shift invariant self-attention

I. INTRODUCTION

Data compression is inevitable when there is a need for transmitting huge amount of data in a limited bandwidth. Neural network based compression algorithms have shown great potential on replacing traditional codecs [1]. Instead of relying on hand-engineered linear transforms, *e.g.*, DCT in JPEG [2], neural compressors can be trained on an arbitrary set of images depending on the task in hand. This major advantage encourages their usage in any other field to reach better trade-off for the ad-hoc application of data compression algorithms.

Data intensiveness of space missions studying the Sun for the goal of space weather analysis and prediction, sets stringent constraints on the bandwidth usage for data communication [3]. To meet the bandwidth requirements in such missions, lossy compression should be investigated which opens a lot of room for efficient data transmission. As an example, Solar Dynamics Observatory (SDO) captures images of the Sun at the resolution of 4096×4096 in nine different wavelengths at a cadence of 12 seconds which results in transmitting more than 1.4 terabytes of data each day to the the ground station [4]. Transmitting this huge amount of data justifies

the investigation for a multi-spectral compression algorithm to remove redundancies over the different frequencies and effectively save transmission cost [5], [6].

Convolution-based neural compression systems [7], [8] although showing great potential in replacing traditional codecs, are being replaced by transformer-based networks [9], [10], [11], [12] after the successful application of vision transformers in the computer vision community [13], [14]. Despite their great ability to capture long-range dependencies, to have a feasible implementation, their global self-attention mechanism must be bridled which comes with a deterioration in performance. This performance degradation need to be addressed properly to make use of these powerful architectures in the low-level computer vision algorithms [15], [16], [17], [18].

To get the most out of a multi-spectral neural image compressor we used transformers with window-based local self attention which are modified to achieve better rate-distortion performance. First by aggregating the keys and queries inter-windows we unleash limited capacity of local window attention mechanism to let the transformer module decide on which token outside its current window to attend. By doing so, we have kept the computation complexity not increased quadratically and simultaneously, enhanced the global dependency capturing of the model. One other important direction is to make the transformer network be able to be shift invariant with respect to its input. Shifted window self-attention [14], [19], [20] is not able to preserve this vital feature for the task of image compression. To make the transformer network insensitive to shifted input, we sample the shift size randomly during training which enforces the network not to distinguish between different shift sizes and operate in a translation invariant mode.

As a sample visualization of our multi-spectral data and the achieved rate-distortion performance of the proposed method, we refer to Figure 1. For further discussions and analysis, this paper is organized as follows: Section II provides a review of neural compression autoencoders and their potential application for a solar mission, with a specific focus on SDO mission and its multi-spectral data. Section III presents our proposed method, highlighting the architecture, aggregated window self

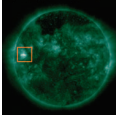
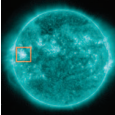
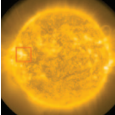
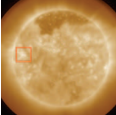
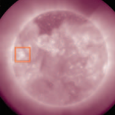
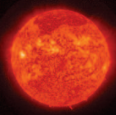
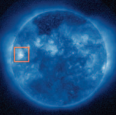

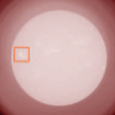
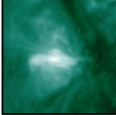
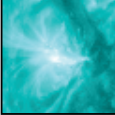
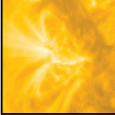
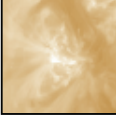
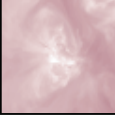
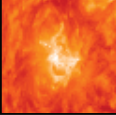
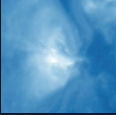
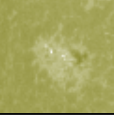

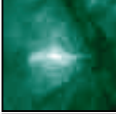
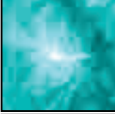
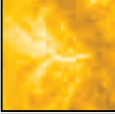
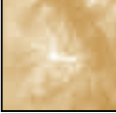

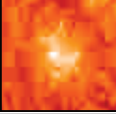



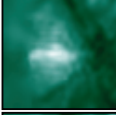
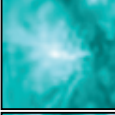

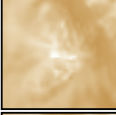

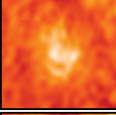

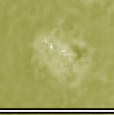

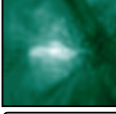
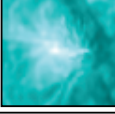
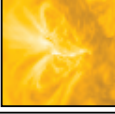
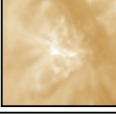
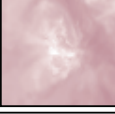
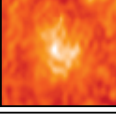
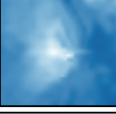
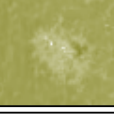

										Bitrate bpp ↓	PSNR dB ↑
Original										8	∞
JPEG										0.41	32.6
Single-spectral										0.41	34.4
Multi-spectral										0.4	35.6
	94	131	171	193	211	304	335	1600	1700	Angstrom	

Fig. 1. Visual comparison of compressing multi-spectral images using our proposed multi-spectral compression against traditional image compression JPEG and also a neural compression which compress images single-spectrally. Proposed multi-spectral compression can achieve the best rate-distortion performance [bpp↓/PSNR↑]. Section III discusses more about the advantages of multi-spectral compression over other methods. *Best viewed on screen.*

attention mechanism, and shift invariance transformer blocks. Experimental results and ablation studies are described in Section IV, showcasing the performance of our method on the SDO dataset. Section V concludes the paper by summarizing our findings.

II. RELATED WORK

In the first part of this Section we will review the works done in the field of neural image compression and the usage of transformers, then in the second part we focus on the solar dynamics observatory mission.

A. Rate-Distortion Optimized Neural Compression

Rate-Distortion Variational Autoencoders (RD-VAE) have dominated the neural-based learned image compression algorithms [21], [7], [22], [23], [1]. The idea is to replace the Gaussian prior and posterior with uniform distribution to have the hard scalar quantization be simulated by adding unit uniform noise to the bottleneck [24], [25]. Another modification to the vanilla VAEs will be to add a learnable prior entropy model [23] and use it to measure the cross entropy between the true distribution of the latent code and the shared prior between encoder and decoder. We follow the same paradigm but with transformers as nonlinear analysis and synthesis transforms [11], [10], [9], [26].

Vision Transformers [13] have shown a great potential in replacing the CNN-based transforms in the learned transform coding architectures. Authors in [9] applied transformer blocks only in the main analysis and synthesis transforms. [11] proposed to use the idea of local window self-attention to improve

the performance of the CNN architecture. Augmenting both main and hyper transforms has shown better rate-distortion trade-off [10] by using local self attention augmented by shifting windows as first proposed under the name of Swin transformer block [14]. Sequential [27] and Parallel-wise [12] mixing of transformer blocks with CNNs were studied as well, where the latter presented the state of the art in terms of rate-distortion performance of general neural image compression.

The Transformer model possesses several advantageous properties for constructing robust data-driven models. Firstly, it can effectively capture dependencies that span long distances within the data [28]. Secondly, it exhibits minimal inductive bias, enabling greater flexibility in accommodating vast amounts of data [13]. Lastly, its high parallelism greatly benefits the training and inference processes of large-scale models [29], [30], [28]. Consequently, the Transformer has not only revolutionized natural language processing but has also displayed promising advancements in the field of computer vision.

Within the computer vision community, there has been a significant proliferation of vision transformers recently [14], [16], [31]. Notably, self-attention, the fundamental building block of these models, has been a popular subject of research [32], [33], [34], [35], [36]. Unlike convolution, which is inherently localized in its operations, attention's pivotal characteristic is its global receptive field [37], [32], [38], [39], [17]. This empowers vision transformers to capture long-range dependencies [40]. However, this advantageous feature comes at the expense of increased computational complexity and

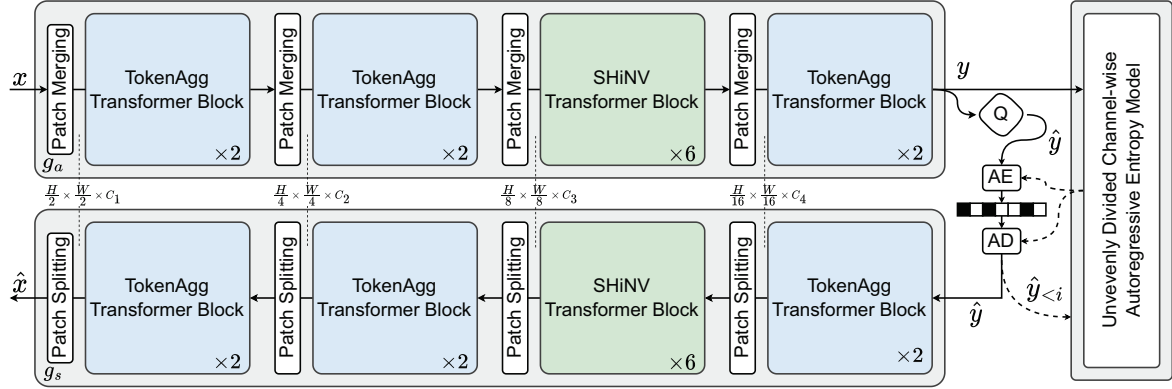


Fig. 2. Transformer-based network architecture. Multi-spectral image \mathbf{x} is fed into consecutive transformer blocks to generate the scalar quantized bottleneck ($\hat{\mathbf{y}}$). Token dimensions in each stage are defined as $(C_1, C_2, C_3, C_4) = (160, 256, 352, 448)$. Afterwards the quantized latent is further compressed/decompressed losslessly by arithmetic encoder/decoder (AE/AD), respectively. The black and white checkerboard box represents the bitstream of compressed latent features. Decoded latent is then used to reconstruct the lossy compressed image ($\hat{\mathbf{x}}$). To estimate the entropy of the latent code, a channel-wise autoregressive model is utilized which is divided unevenly over the channel dimension in which each division of latent code $\hat{\mathbf{y}}_i$ has a probability dependent on its previous divisions $\hat{\mathbf{y}}_{<i}$. Analysis and synthesis transformer-based nonlinear transforms are denoted by g_a and g_s , respectively. Token Aggregated (TokenAgg) and Shift iNvariance (SHiNV) transformer blocks are described in Section III-B.

memory requirements, as attention calculates pairwise token affinity across all spatial locations, resulting in substantial memory footprints. This work proposes two enhancements upon local shifted window vision transformers, as thoroughly explained in Section III-B.

B. Space Missions Studying the Solar Atmosphere

The advancement of sensor technology and the growing need for a deeper understanding of the space environment, ranging from the Sun to Earth and beyond, has resulted in a massive increase in data volume. This includes data with unprecedented spatial and/or temporal resolution, as well as multi-spectral information. Consequently, innovative data compression algorithms are required to handle this vast amount of data efficiently. Solar Dynamics Observatory (SDO) is a recently designated space mission to study the atmosphere of the Sun by observing it continuously. SDO carries Atmospheric Imaging Assembly on-board which captures 4K resolution images at 10 different wavelengths at a cadence of 10 seconds [4]. Downloaded images include seven Extreme UltraViolet (EUV) bands of 94, 131, 171, 193, 211, 304, 335 Å in addition to two visible wavelengths 1600 and 1700 Å [41]. This instrument merely transmits about 1.4 Tera bytes of data each day orbiting the earth [42], [4]. This huge amount of imagery data calls for compression mechanisms to be introduced specially designed for these types of data-intensive missions [41], [43], [44]. The crucial need for lossy image compression on petabyte-scale data obtained from solar missions has been emphasized by applying JPEG-2000 [45] and learning-based [5], [6], [46] methods to compress SDO images although non has considered utilizing the redundancies over the spectrum dimension, which is the topic of this work.

Learning-based analysis of solar data has gained attention recently by the work of Galvez et al. [47], which provided a set of SDO data ready for machine/deep learning analysis.

A portion of raw data from the Solar Dynamics Observatory (SDO) was collected and processed to create a machine-learning ready dataset known as SDOML. This dataset was specifically curated to facilitate the development and evaluation of learning-based methods using SDO mission data.

Building upon the SDOML dataset, Salvatelli et al. [48] employed a U-Net architecture within a Generative Adversarial Network (GAN) framework. Their approach aimed to translate multi-spectral images from the Atmospheric Imaging Assembly (AIA) instrument of SDO, captured at wavelengths 94, 171, and 193 Å, to a specific target wavelength of 211 Å. This work focused on spectral image translation using deep learning techniques. Another machine learning study conducted on the SDOML dataset was proposed by Santos et al. [49]. They utilized deep neural networks to address the auto-calibration of instrument degradation in SDO imagery. By leveraging the SDOML dataset, their approach aimed to automatically compensate for degradation effects in the instrument imagery, improving the accuracy and reliability of the captured data. In a different application, Dash et al. [50] employed a conditional GAN to perform image translation from the Helioseismic and Magnetic Imager (HMI) images downloaded from SDO to AIA images. This translation allowed them to generate AIA-like images from the HMI instrument, thus expanding the capabilities of HMI images through the use of deep learning techniques.

III. METHODS

A. Transform Coding Neural Image Compression

Autoencoder-based image compression networks [23], [51], such as the illustrated architecture in Figure 2, typically comprise two main components. The first component is the encoder/decoder network, while the second component is the bottleneck entropy modeling network. The details of the latter

network are extensively explained in Section III-A2. Referring to Figure 2, we can summarize the relationship between the network input (\mathbf{x}) and output (\mathbf{x}') as follows:

$$\begin{aligned}\mathbf{x}' &= g_s(\hat{\mathbf{y}}; \theta_g), \\ \hat{\mathbf{y}} &= \lfloor g_a(\mathbf{x}; \phi_g) \rfloor, \\ \hat{\mathbf{z}} &= \lfloor h_a(\mathbf{y}; \phi_h) \rfloor,\end{aligned}\quad (1)$$

where the output image \mathbf{x}' is generated by applying the synthesis transform g_s to the quantized latent variable $\hat{\mathbf{y}}$, controlled by the learned parameters θ_g . The quantized latent variable $\hat{\mathbf{y}}$ is obtained by quantizing the output of the analysis transform g_a applied to the input image \mathbf{x} using the learned parameters ϕ_g . The quantized hyper-prior $\hat{\mathbf{z}}$ is obtained by quantizing the output of the analysis transform h_a applied to the latent variable \mathbf{y} , using the learned parameters ϕ_h . The subscripts a and s indicate that g_a and g_s represent analysis and synthesis transforms, respectively, commonly used in the terminology of transform coding-based compression algorithms.

1) **Entropy-Constrained Distortion Minimization:** Any learned image compression network aims to balance rate and distortion, which is represented by the Lagrangian parameter λ in the equation:

$$R + \lambda D, \quad (2)$$

where R represents the estimated entropy of the latent code, and D corresponds to the reconstruction distortion. During network training, the goal is to minimize the rate term, which is the estimated entropy of the quantized bottleneck. The probability distribution of the latent code is approximated by the hyper-prior \mathbf{z} . The quantized $\hat{\mathbf{z}}$ is transmitted along with the compressed image as side-information. Thus, both the entropy of the latent code and the hyper-prior need to be optimized, as defined below:

$$R = \mathbb{E}_{\mathbf{x} \sim p_X} [-\log_2 P_{\hat{\mathbf{y}}|\hat{\mathbf{z}}}(\hat{\mathbf{y}}|\hat{\mathbf{z}}; \theta_h) - \log_2 P_{\hat{\mathbf{z}}}(\hat{\mathbf{z}}; \psi)], \quad (3)$$

In Equation 2, D represents the distortion between the input and output images of the network, which can be measured using various metrics. The Mean Squared Error (MSE) is commonly used but has been criticized for producing blurry reconstructions. Alternative metrics such as Multi Scale Structural Similarity Index (MS-SSIM) have been proposed to align with the human visual system but have their limitations when closely examined.

2) **Bottleneck Entropy Modeling:** The performance of a learned image compression scheme relies heavily on its ability to accurately estimate the true entropy of the bottleneck. The objective is to minimize the cross-entropy between the estimated and true entropies. Various probability estimation methods have been proposed in the literature, including empirical histogram density estimation [21], piecewise linear models [22], conditioning on a latent variable (hyper-prior) [23], and context modeling based on autoregressive models [51].

At a high level, entropy estimation models can be categorized into two main types: Forward Adaptation (FA) models and Backward Adaptation (BA) models. FA models have limited capacity to capture all dependencies in the probability

distribution of the latent code, while BA models suffer from the inability to parallelize the decoding process. Learned FA models utilize only the information provided during the encoding of the image, whereas BA methods based on autoregressive models require information from the decoded message as well. To leverage the advantages of both types of models, a conditional probability formulation is defined:

$$P_{\hat{\mathbf{y}}|\hat{\mathbf{z}}}(\hat{\mathbf{y}}|\hat{\mathbf{z}}) = \prod_i P(\hat{\mathbf{y}}_i | \hat{\mathbf{y}}_{j < i}, \hat{\mathbf{z}}; \theta_h). \quad (4)$$

Conditioning on the quantized hyper-prior $\hat{\mathbf{z}}$ as side-information represents a form of FA, while conditioning on all previously decoded elements of the latent space $\hat{\mathbf{y}}_{j < i}$ represents a form of BA.

The performance of BA models has been improved in [8] by introducing conditioning only between slices of channels in the bottleneck. Unlike spatial autoregressive modeling in [51], [8] considers conditioning probabilities only on the channels. This approach enables reasonable parallelization of the decoding process. We have employed the same approach as [8], but by dividing channels in an uneven set of groups [52], to estimate and minimize the entropy during training as shown in Figure 2.

B. Transformer-based Nonlinear Transforms

Nonlinear Transform coding [25] will be the choice when it comes to complex data distributions such as natural images or, in our case, multi-spectral images. Transformers [28], [13], which are based on self-attention mechanism, are replacing traditional convolution-based backbones in the deep neural networks used in computer vision tasks [53], [14], [16]. Although used as powerful representation learner, two issues of the self-attention need to be addressed to perform well when it comes to the task of image compression. First its notorious quadratic computation complexity which is addressed in the literature by sparsifying global attention, limiting it to a local window [14], widening the window by dilation [18] and make the structure hierarchical [31]. Second issue which is more concerned for the task of image compression is the translation invariance. In the task of compression, in contrast to object detection or classification, it is necessary for the encoder and decoder nonlinear transforms to extract representative features of the whole image no matter of its spatial location. Even in contrast to object detection, minute local details could be of more importance compared to low frequency predictable regions. Here we briefly set the terminology of self-attention mechanism and in sections III-B2 and III-B3 describe our proposed solutions to ameliorate its performance.

1) **(Preliminaries) Self-Attention Mechanism:** Let's assume we have an input \mathbf{X} of size $\mathbb{R}^{N \times C}$. To apply the self-attention mechanism on N tokens of dimension C , three linear transformations of it are calculated which are called query ($\mathbf{Q} \in \mathbb{R}^{N \times C}$), key and value ($\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N_{KV} \times C}$). Then the self-attended input will be as follows:

$$\text{SA}(\mathbf{X}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}}\right)\mathbf{V}. \quad (5)$$

More commonly, the self-attention is divided into multiple heads to bring different aspects of the input into attention. Having each head $\mathbf{H}_i \in \mathbb{R}^{N \times \frac{C}{n}}$, its corresponding query, key and value are calculated as

$$\begin{aligned} \mathbf{Q}_i &= \mathbf{Q} \mathbf{W}_i^Q, \\ \mathbf{K}_i &= \mathbf{K} \mathbf{W}_i^K, \\ \mathbf{V}_i &= \mathbf{V} \mathbf{W}_i^V, \end{aligned} \quad (6)$$

and then each head is self-attended as

$$\mathbf{H}_i = \text{Softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{C/n}}\right) \mathbf{V}_i, \quad (7)$$

where linear weights for each head are $\mathbf{W}_{1..n}^Q, \mathbf{W}_{1..n}^K, \mathbf{W}_{1..n}^V \in \mathbb{R}^{C \times \frac{C}{n}}$. To have the final output of multi-head self-attention, all the heads need to be concatenated and passed through a final linear ($\mathbf{W}^H \in \mathbb{R}^{C \times C}$) transformation as follows:

$$\text{MHSA}(\mathbf{X}) = [\mathbf{H}_1; \dots; \mathbf{H}_n] \mathbf{W}^H. \quad (8)$$

To keep the equations uncluttered, we resume our discussion by only having a single head, while in practice our work can be easily extended to multi-heads.

2) Inter-Window Token Aggregated: Window-based self-attention [14] with a fixed window size of \mathfrak{W} first partitions tokens into groups of size \mathfrak{W}^2 resulting in $\mathbf{X}^{\mathfrak{W}} \in \mathbb{R}^{\frac{HW}{\mathfrak{W}^2} \times \mathfrak{W}^2 \times C}$. Therefore there is a total of $\frac{HW}{\mathfrak{W}^2}$ windows, where in each of them there is \mathfrak{W}^2 tokens to be fed into the self attention module, as shown in Figure 3. On each window-partition, self-attention is then applied by the mechanism described in Section III-B1. Correspondingly we will have $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{\frac{HW}{\mathfrak{W}^2} \times \mathfrak{W}^2 \times C}$, where each of them are the outputs of linear transformations as follows:

$$\begin{aligned} \mathbf{Q} &= \mathbf{X}^{\mathfrak{W}} \mathbf{W}^Q, \\ \mathbf{K} &= \mathbf{X}^{\mathfrak{W}} \mathbf{W}^K, \\ \mathbf{V} &= \mathbf{X}^{\mathfrak{W}} \mathbf{W}^V, \end{aligned} \quad (9)$$

where weights have dimensions $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{C \times C}$.

Although the window-partitioning helps to reduce the computational complexity, it has deleterious effects on the long-range dependency modeling capability of the vanilla self-attention. Here, we propose to exploit the inter-window dependencies by representing each window with a single candidate, which is the average of key and value tokens in that window. The candidates are then used to search for similarities beyond the intra-window. There will be a $\frac{HW}{\mathfrak{W}^2}$ number of candidates, defined as follows:

$$\mathbf{Q}^{\mathfrak{W}}, \mathbf{K}^{\mathfrak{W}} \in \mathbb{R}^{\frac{HW}{\mathfrak{W}^2} \times C}. \quad (10)$$

After having the candidates, a resemblance matrix is utilized to measure the similarity inter-windows by comparing dot-product similarity between candidates,

$$\mathfrak{R} := \mathbf{Q}^{\mathfrak{W}} (\mathbf{K}^{\mathfrak{W}})^T \in \mathbb{R}^{\frac{HW}{\mathfrak{W}^2} \times \frac{HW}{\mathfrak{W}^2}}, \quad (11)$$

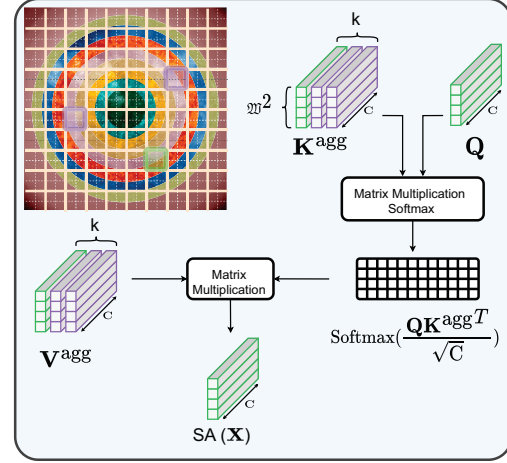


Fig. 3. Inter-window token aggregated attention module. Top- k closet windows (assumed $k = 3$, $\mathfrak{W} = 2$ for visualization) (violet) to be attended for the targeted window (green). (multi-spectral input is shown as hyper-disks for visualization purposes only.)

where its *top-k* indices will be selected for the inter-window aggregated self-attention. The numbers tokens are kept limited by setting value of k to have a reasonable amount of computations and have the computational complexity not being quadratically increasing with respect to the size of the input. The inter-window aggregated keys and values will be:

$$\begin{aligned} \mathbf{K}^{\text{agg}} &= \text{agg}[\text{top-}k(\mathbf{K}, \mathfrak{R})] \in \mathbb{R}^{\frac{HW}{\mathfrak{W}^2} \times k\mathfrak{W}^2 \times C}, \\ \mathbf{V}^{\text{agg}} &= \text{agg}[\text{top-}k(\mathbf{V}, \mathfrak{R})] \in \mathbb{R}^{\frac{HW}{\mathfrak{W}^2} \times k\mathfrak{W}^2 \times C}. \end{aligned} \quad (12)$$

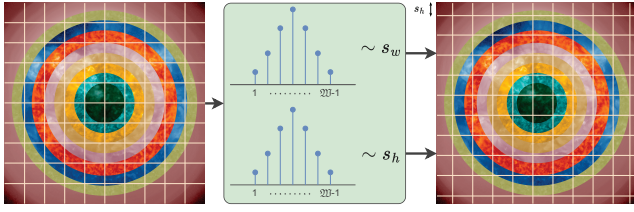
Finally the self-attention for each window is calculated by its intact query and its aggregated top- k most similar keys and values as follows:

$$\text{SA}(\mathbf{X}) = \text{Softmax}\left(\frac{\mathbf{Q} \mathbf{K}^{\text{agg}T}}{\sqrt{C}}\right) \mathbf{V}^{\text{agg}}. \quad (13)$$

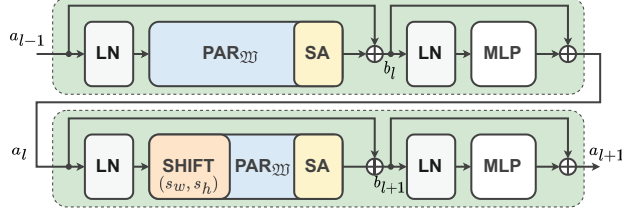
3) Randomly Sampled Shift Invariant Self-Attention:

When it comes to the comparison between different architectures, self-attention lacks one very important feature compared to CNNs. CNNs possess unique features such as inherent locality and weight sharing, which lead to the advantageous quality of shift invariance. These two characteristics are crucial in image compression, as pixels within nearby regions tend to display significant correlations, and achieving translation invariance is a desired trait for potential image compression techniques.

Conversely, transformer-based image compression lack compatibility with shift invariance and locality. The strategies employed by transformers to reconstruct high-resolution images can be categorized into two main patterns: the utilization of small patches with global attention [54], and the application of large patches with local attention [20]. However, neither of these patterns fulfills the demands for shift invariance or locality.



(a) Randomly sampled height and width shift size from categorical distribution with support of integer values less than window size, i.e., $1, \dots, 255 - 1$. (multi-spectral input is shown as hyper-disks for visualization purposes only.)



(b) Shift iNvariance transformer block including two consecutive transformers. First block is the vanilla window-based self-attention while the second one implements the randomly shifted window self-attention.

Fig. 4. Shift invariaant transformer blocks.

The objective of the shift invariance self-attention is to enhance the transformer's ability to recognize patterns regardless of their location during translation, while also effectively utilizing local connections. The traditional transformer structure consists of consecutive layers of self-attention (SA) and multi-layer perceptron (MLP). To enhance efficiency, the transformer typically employs local window-based attention, and a shifted window strategy is used to facilitate connections between windows. More specifically, the feature map is divided into non-overlapping windows, and self-attention is calculated within each local window, as shown in Figure 4, and described by the following equations:

$$\begin{aligned} b_l &= \text{SA}(\text{PAR}_{255}(\text{LN}(a_{l-1}))) + a_{l-1}, \\ a_l &= \text{MLP}(\text{LN}(b_l)) + b_l, \\ b_{l+1} &= \text{SA}(\text{PAR}_{255}(\text{SHIFT}(\text{LN}(a_l); s_w, s_h))) + a_l, \\ a_{l+1} &= \text{MLP}(\text{LN}(b_{l+1})) + b_{l+1}, \end{aligned} \quad (14)$$

where the height and width shift sizes are denoted by s_h and s_w , respectively. In local window self-attention, the shifting sizes are fixed to half of the window size, i.e., $s_h = s_w = \frac{255}{2}$. This lets the consecutive transformer blocks to capture similarities over the windows [14]. The requirement to increase the number of transformer blocks to make the shifting window capture all the dependencies outside of a local window makes the vanilla shifted window attention less feasible when we expect a shift invariance functionality. To let the window-based self-attention mechanism be performed shift-invariance we use non-deterministic shift sizes during training as shown in Figure 4(a), where the height and width shift sizes are sampled randomly from a pre-defined categorical distribution which is peaked on the half value of window size, i.e., $\frac{255}{2}$.

Any other distribution is possible to be chosen, and further investigation of this could be found in Section IV-C.

IV. EXPERIMENTS

A. Dataset

The dataset used in our experiments is based on the SDO images described in the work of Galvez et al. [47]. The dataset consists of images of the Sun captured at various wavelengths, including 94, 131, 171, 193, 211, 304, 335, 1600, and 1700 Å, with a temporal cadence of 6 minutes.

To reduce temporal dependencies between training samples, we downsampled the images to a cadence of 1 hour. This downsampling helps to decrease the correlation between consecutive images, allowing for more diverse training samples. The dataset provides a comprehensive collection of solar images, capturing different aspects of the Sun's atmosphere at various wavelengths. This diversity enables us to develop and evaluate our image compression methods using a wide range of spectral information. To address potential biases related to solar variations at different stages of the solar cycle, we adopted a specific data division strategy inspired by the approach proposed in Salvatelli et al. [55]. We divided the dataset based on the months in which the images were captured.

Specifically, we selected images from January to August of the years 2015 to 2018 for the training set. This range ensures coverage of different months throughout the years, helping to mitigate any biases introduced by seasonal variations in solar activity. On the other hand, we reserved images from September to December of the same years for the testing set. This separation enables us to evaluate the performance of our models on unseen data from different time periods.

In total, the training set consists of 15,768 images where each image contains 9 wavelengths at its channel dimension, while the test set comprises 3,315 images. The results reported in this section are based on the evaluation of our models on the test set, providing insights into their performance on previously unseen data.

B. Implementation Specifications

We trained a total of seven models, each with a different hyper-parameter λ governing the rate-distortion trade-off as defined in Equation (2). The chosen values for λ were empirically determined as $\{0.0015, 0.0035, 0.0070, 0.0125, 0.0250, 0.0410, 0.0550\}$, and each model was trained for 200 epochs.

For training, we used the Adam optimizer [56] with a batch size of 8. The training data consisted of randomly cropped patches of size 256×256 from the original 512×512 images. The initial learning rate was set to 10^{-4} and annealed during training to 1×10^{-5} to facilitate convergence.

During the evaluation phase, we performed entropy coding of the latent integer values using range asymmetric numeral systems coding [57]. It's important to note that this entropy coding is lossless and does not impact the measured performance or functionality of the algorithm during training. It

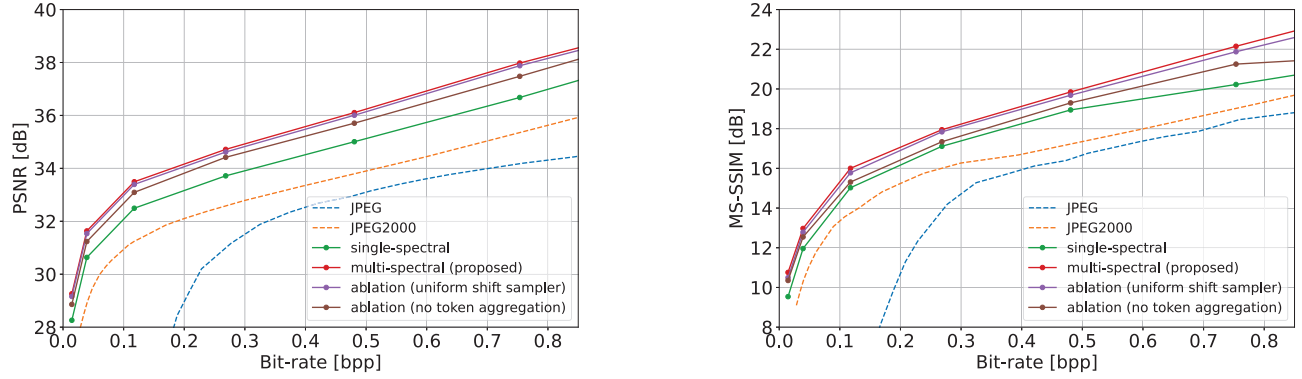


Fig. 5. The rate distortion curves aggregated over the test set are detailed in Section IV-A. The left side presents PSNR (Peak Signal-to-Noise Ratio), computed from MSE (Mean Squared Error) using the formula $10 \log_{10} \frac{255^2}{MSE}$. On the right side, MS-SSIM (Multi-Scale Structural Similarity Index) is displayed on a logarithmic scale using $-10 \log(1 - m)$ to enhance the visibility of differences. Here, m represents the MS-SSIM value within the range of zero to one.

is only during the evaluation phase that entropy coding is applied, as it allows for comparison with standard codecs such as JPEG [2] and JPEG-2000 [58].

C. Ablation Study

To verify the impact of both aggregating tokens inter-windows and non-deterministic shift sizes, we have conducted two isolated ablation studies. In the first set of models, the token-aggregation transformer blocks are replaced with vanilla Swin transformer blocks [14] to see how aggregation of windows could play more effective role than the local window-limited attentions even when the shifts are in place. The second ablative study is concerned with how the distribution of the shift sizes could affect the rate-distortion performance of the multi-spectral compressor. We have used a uniform distribution with support set of $\{1, \dots, \mathfrak{W} - 1\}$ instead of the pre-defined categorical distribution with the same support set, in the original network. As is shown in Figure 5, the choice of seeing outside of the local window (by aggregating tokens) has much tangible influence than how we choose to sample the shift sizes of the windows.

V. CONCLUSION

In this work a multi-spectral transformer-based neural image compression algorithm were proposed to effectively improve the rate-distortion performance and save the communication cost in data-intensive missions studying solar dynamics. Downloaded data from these missions possess high redundancy over the frequency spectrum and we proposed mechanisms to remove redundancies effectively. Inter-window token aggregation and shift invariance window partitioning were added to the transformer blocks to better de-correlate the high dimensional multi-spectral images of the Sun.

ACKNOWLEDGMENT

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.

REFERENCES

- [1] Y. Yang, S. Mandt, and L. Theis, "An introduction to neural data compression," *Foundations and Trends® in Computer Graphics and Vision*, vol. 15, no. 2, pp. 113–200, 2023. [Online]. Available: <http://dx.doi.org/10.1561/0600000107>
- [2] G. K. Wallace, "The JPEG still picture compression standard," *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.
- [3] P. Chamberlin, W. D. Pesnell, and B. Thompson, *The Solar Dynamics Observatory*. Springer Science & Business Media, 2012.
- [4] J. R. Lemen, A. M. Title, D. J. Akin, P. F. Boerner, C. Chou, J. F. Drake, D. W. Duncan, C. G. Edwards, F. M. Friedlaender, G. F. Heyman *et al.*, "The atmospheric imaging assembly (AIA) on the solar dynamics observatory (SDO)," *Solar Physics*, vol. 275, pp. 17–40, 2012.
- [5] 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.
- [6] A. Zafari, A. Khoshkhahtinat, N. Nasrabadi, and P. Mehta, "Neural image compression on solar dynamics observatory," *The Third Triennial Earth-Sun Summit (TESS)*, vol. 54, no. 7, 2022.
- [7] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *2016 Picture Coding Symposium, PCS 2016*. IEEE, 2016.
- [8] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *IEEE International Conference on Image Processing, ICIP 2020, Abu Dhabi, United Arab Emirates, October 25-28, 2020*. IEEE, 2020, pp. 3339–3343.
- [9] Y. Bai, X. Yang, X. Liu, J. Jiang, Y. Wang, X. Ji, and W. Gao, "Towards end-to-end image compression and analysis with transformers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 104–112.
- [10] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *International Conference on Learning Representations*, 2022.
- [11] R. Zou, C. Song, and Z. Zhang, "The devil is in the details: Window-based attention for image compression," in *CVPR*, 2022.
- [12] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-cnn architectures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 388–14 397.
- [13] 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 *9th International Conference on Learning Representations*. OpenReview.net, 2021.

- [14] 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 *ICCV*. IEEE, 2021.
- [15] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4794–4803.
- [16] 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.
- [17] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unified transformer for efficient spatiotemporal representation learning," *arXiv preprint arXiv:2201.04676*, 2022.
- [18] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *The Eleventh International Conference on Learning Representations*, 2022.
- [19] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global context vision transformers," in *International Conference on Machine Learning*. PMLR, 2023, pp. 12 633–12 646.
- [20] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 683–17 693.
- [21] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *5th International Conference on Learning Representations*. OpenReview.net, 2017.
- [22] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *5th International Conference on Learning Representations*. OpenReview.net, 2017.
- [23] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *6th International Conference on Learning Representations*, 2018.
- [24] J. Ballé, "Efficient nonlinear transforms for lossy image compression," in *2018 Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 248–252.
- [25] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, "Nonlinear transform coding," *IEEE J. Sel. Top. Signal Process.*, 2021.
- [26] 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.
- [27] M. Lu, P. Guo, H. Shi, C. Cao, and Z. Ma, "Transformer-based image compression," in *2022 Data Compression Conference (DCC)*, 2022, pp. 469–469.
- [28] 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.
- [29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [30] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," *OpenAI*, 2018.
- [31] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.
- [32] K. M. Choromanski, V. Likhoshesterov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Q. Davis, A. Mohiuddin, L. Kaiser, D. B. Belanger, L. J. Colwell, and A. Weller, "Rethinking attention with performers," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=Ua6zuk0WRH>
- [33] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, "Biformer: Vision transformer with bi-level routing attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 323–10 333.
- [34] H. Taghavi, A. El Shafei, and A. Nasiri, "Liquid cooling system for a high power, medium frequency, and medium voltage isolated power converter,"
- [35] J. Xiao, X. Fu, F. Wu, and Z.-J. Zha, "Stochastic window transformer for image restoration," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9315–9329, 2022.
- [36] P. Schober, S. N. Estiri, S. Aygun, A. H. Jalilvand, M. H. Najafi, and N. TaheriNejad, "Stochastic computing design and implementation of a sound source localization system," *IEEE Journal on Emerging and*
- Selected Topics in Circuits and Systems*, vol. 13, no. 1, pp. 295–311, 2023.
- [37] P. Ramachandran, N. Parmarand, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Advances in Neural Information Processing*, 2019.
- [38] M. M. Rahman, M. A. Farahani, and T. Wuest, "Multivariate time-series classification of critical events from industrial drying hopper operations: A deep learning approach," *Journal of Manufacturing and Materials Processing*, vol. 7, no. 5, p. 164, 2023.
- [39] R. Nematirad, A. Ahmadisharaf, and A. Lashgari, "Forecasting the performance of us stock market indices during covid-19: Rf vs lstm," *arXiv preprint arXiv:2306.03620*, 2023.
- [40] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=HJlnc1rKPB>
- [41] D. Savage, "A guide to the mission and purpose of nasa's solar dynamics observatory," *NASA Goddard Space Flight Center*, 2010. [Online]. Available: https://sdo.gsfc.nasa.gov/assets/docs/SDO_Guide.pdf
- [42] P. Chamberlin, W. D. Pesnell, and B. Thompson, *The Solar Dynamics Observatory*. Springer Science & Business Media, 2012.
- [43] A. Sarlak and Y. Darmani, "An approach to improve the quality of service in dtn and non-dtn based vanet," *Journal of Information Systems and Telecommunication (JIST)*, vol. 4, no. 32, p. 240, 2021.
- [44] P. Bhuvella and A. Nasiri, "Design methodology for a medium voltage single stage llc resonant solar pv inverter."
- [45] C. E. Fischer, D. Müller, and I. De Moortel, "JPEG2000 image compression on solar EUV images," *Solar Physics*, vol. 292, 2017.
- [46] A. Zafari, A. Khoshkhahtinat, P. M. Mehta, N. Nasrabadi, B. J. Thompson, D. da Silva, and M. Kirk, "Attention-based generative neural image compression on solar dynamics observatory," in *103rd AMS Annual Meeting*. AMS, 2023.
- [47] R. Galvez, D. F. Fouhey, M. Jin, A. Szenicer, A. Muñoz-Jaramillo, M. C. M. Cheung, P. J. Wright, M. G. Bobra, Y. Liu, J. Mason, and R. Thomas, "A machine-learning data set prepared from the NASA solar dynamics observatory mission," *The Astrophysical Journal Supplement Series*, may 2019.
- [48] V. Salvatelli, L. F. G. dos Santos, S. Bose, B. Neuberger, M. C. M. Cheung, M. Janvier, M. Jin, Y. Gal, and A. G. Baydin, "Exploring the limits of synthetic creation of solar euV images via image-to-image translation," *The Astrophysical Journal*, vol. 937, no. 2, p. 100, oct 2022.
- [49] L. F. G. dos Santos, S. Bose, V. Salvatelli, B. Neuberger, M. C. M. Cheung, M. Janvier, M. Jin, Y. Gal, P. Boerner, and A. G. Baydin, "Multi-channel auto-calibration for the atmospheric imaging assembly using machine learning," *CoRR*, vol. abs/2012.14023, 2020.
- [50] A. Dash, J. Ye, and G. Wang, "High resolution solar image generation using generative adversarial networks," *arXiv preprint arXiv:2106.03814*, 2021.
- [51] D. Minnen, J. Ballé, and G. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing*, 2018.
- [52] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5718–5727.
- [53] S. Mohamadi, G. Doretto, and D. Adjeroh, "Deep active ensemble sampling for image classification," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 4531–4547.
- [54] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 299–12 310.
- [55] V. Salvatelli, S. Bose, B. Neuberger, L. F. dos Santos, M. Cheung, M. Janvier, A. G. Baydin, Y. Gal, and M. Jin, "Using U-Nets to create high-fidelity virtual observations of the solar corona," *NeurIPS*, 2019.
- [56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, 2015.
- [57] J. Duda, "Asymmetric numeral systems: entropy coding combining speed of huffman coding with compression rate of arithmetic coding," *arXiv preprint arXiv:1311.2540*, 2013.
- [58] 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.