Attention-Based Generative Neural Image Compression on Solar Dynamics Observatory

Ali Zafari[†], Atefeh Khoshkhahtinat[†], Piyush M. Mehta[‡], Nasser M. Nasrabadi[†], Barbara J. Thompson[§], Daniel da Silva[§], Michael S. F. Kirk[§]

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

[‡]Dept. 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,daniel.e.dasilva,michael.s.kirk}@nasa.gov

***Abstract*—NASA's Solar Dynamics Observatory (SDO) mission gathers 1.4 terabytes of data each day from its geosynchronous orbit in space. SDO data includes images of the Sun captured at different wavelengths, with the primary scientific goal of understanding the dynamic processes governing the Sun. Recently, end-to-end optimized artificial neural networks (ANN) have shown great potential in performing image compression. ANN-based compression schemes have outperformed conventional hand-engineered algorithms for lossy and lossless image compression. We have designed an ad-hoc ANN-based image compression scheme to reduce the amount of data needed to be stored and retrieved on space missions studying solar dynamics. In this work, we propose an attention module to make use of both local and non-local attention mechanisms in an adversarially trained neural image compression network. We have also demonstrated the superior perceptual quality of this neural image compressor. Our proposed algorithm for compressing images downloaded from the SDO spacecraft performs better in rate-distortion trade-off than the popular currently-in-use image compression codecs such as JPEG and JPEG2000. In addition we have shown that the proposed method outperforms state-of-the art lossy transform coding compression codec, i.e., BPG.**

***Index Terms*—Learned lossy image compression, solar dynamics observatory, generative adversarial network, attention**

I. Introduction

Image compression using artificial neural networks (ANN) has shown great potential to be applied on a wide variety of different areas since their first appearance [1]. In the past couple of years, they have outperformed most of the hand-engineered codecs such as JPEG [2] and JPEG2000 [3] in terms of rate-distortion (RD) performance [4]. One major advantage of ANN-based compression algorithms is that they can be developed on any ad-hoc dataset to do better compression than general codecs [4].

Although it is believed that the ultimate trade-off in image compression is between the rate and distortion, recent studies have shown that there is a third role governing the visual quality of compressed images known as perception [5]. Generative Adversarial Networks (GANs) are known for their high-quality reconstructed images by enforcing the ANN to capture the distribution of their input image. Hence, to improve the perceptual quality of reconstructed image at the receiver, GANs have been applied to image compression networks in the literature [6].

Another venue of works to improve the performance of Convolutional Neural Networks (CNNs) is attention mechanism. With its unprecedented influence in natural language processing [7], attention has found its way in computer vision and object detection/classification tasks [8], [9]. We have utilized both of these improvements in learned image compression networks to enhance the performance in terms of rate-distortion-perception trade-off [5]. As shown in Figure 1, although the attention mechanism can reach better performance compared with other compression standards, augmenting it with a GAN will lead to better perceptual quality.

Contributions of This Work. In this work we have investigated the application of recently successful learned image compression methods in the field of solar imaging. We have shown that these neural compression schemes could easily outperform traditional and currently-in-use image codecs. In addition, we have proposed a curated attention module to improve the RD tradeoff performance in state-of-the art neural compression architectures. We have also utilized adversarial training to encourage the decoder of our neural network to preserve the distribution of the solar images during the reconstruction process.

The remainder of the paper is organized as follows. Section II reviews the neural-based compression methods and the importance of compression on SDO mission. Section III describes our proposed method. The experiments and ablation studies are discussed in section IV with a conclusion at section V.

II. Related Work

A. Neural Image Compression

Transform coding based image compression algorithms share four main steps to compress an image [10]. First, encoding the images from their input space (e.g., RGB) to an uncorrelated space. Second, quantizing to discard less significant information from the data in its uncorrelated domain. At

This research is based upon work supported by the National Aeronautics and Space Administration (NASA), via award number 80NSSC21M032.

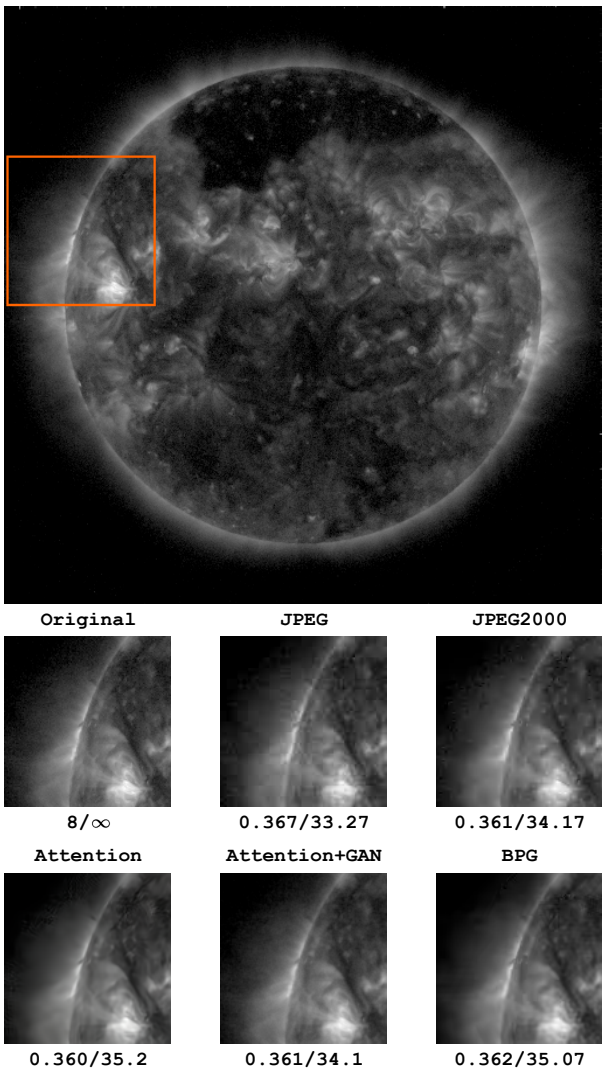


Fig. 1. Visual comparison of proposed compression schemes (Attention only and GAN+Attention) to other standard codecs. Reported performance in terms of bit-rate/distortion [bpp↓/PSNR↑]. GAN outputs are visually closer to the original input unless their lower performance in terms of PSNR. *Best viewed on screen.*

the third step, an entropy coding will be utilized to losslessly encode the quantized samples into a stream of ones and zeros. This bitstream will be the compressed image. Final step occurs at the receiving end (or at reconstructing step), which is responsible to decode the quantized values to the original space of the input image. The first and most widely used architecture to mimic this scenario in deep neural networks, is the convolutional autoencoder, which has shown its superiority in the literature [11]. Both the encoding and decoding part of the traditional transform coding, could be imitated by an autoencoder [12].

End-to-end optimizing of the neural networks are capable of handling various tasks [13], [14], [15], [16] if the learning objective chosen to be differentiable. In an end-to-end optimisation of an autoencoder, problems arise when we want to do quantization on its bottleneck. It is worth mentioning that quantization is the essential part of compression. Merely doing dimensionality reduction cannot necessarily result in discarding the redundant information, which is necessary to attain high compression ratios [17].

ANNs are optimized using gradient descent algorithms which update the parameters of the network by back-propagating the gradients of the loss function. Thus, all the operations performed in it must be differentiable. As a result, we need to approximate hard discrete quantization with a soft continuous operation. To do so, several approaches have been proposed in the literature. [1] used recurrent neural networks to directly binarize the latent code stochastically, while [17] used an approach similar to straight through estimator [18] by back-propagating the gradients of identity function and rounding to the nearest integer in the forward pass. By this continuous approximation, the network parameters can be successfully learned with backpropagating the gradient of the loss function. The most widely used approach is proposed by [19], inherited from [20], they showed that adding independent and identically distributed uniform noise in the range of scalar quantization can be interpreted equivalently as doing scalar quantization on the bottleneck. By doing so, we can optimize the differential entropy of the continuous approximation as a variational upper bound [21] to reduce the entropy of the bottleneck. Low entropy messages are compressed more efficiently into bitstreams [22], [23].

In classical image compression schemes, to get the best out of the quantization process, the first step was to apply an invertible linear transform and translate the image into decorrelated coefficients using Discrete Cosine Transform (DCT). By doing so, scalar quantization could reach a reasonable performance close to vector quantization [10]. The application of vector quantization in ANN-based compression has been investigated by [24], with the cost of complicated training procedure. On the other hand, it has been shown [11], [12] that a joint-optimized learned nonlinear transform, i.e., neural network, followed by scalar quantization can ideally approximate a parametric form of vector quantization.

Replacing the actual quantization with uniform noise approximation in the bottleneck of a vanilla autoencoder during training of the network [25], will transform it to a Variational Aueoencoder (VAE) [26]. The only difference is on the chosen prior. In autoencoder-based image compression, the Gaussian prior of the VAE is replaced with unit uniform distribution centered on integer numbers.

B. Solar Dynamics Observatory (SDO) Mission

*1) **Image Compression on SDO Data**:* Advances in sensor technology and an increasing desire for a deeper understanding of the space environment (Sun to Earth and beyond) have led to an explosion of data volume in recent years (unprecedented spatial and/or temporal resolution as well as multi-spectral data). As a result, it requires new innovative data compression algorithms. The SDO mission transmits 1.4 TB of data (most of it images of the sun at different wavelengths) each day to

the ground station [27], which shows the importance of compression on transmitting and more importantly on archiving this huge amount of data [28]. In [29] authors pointed out the essential need to do lossy image compression on petabyte size data gathered from Solar missions. They studied usage of JPEG2000 which is a transform coding compression scheme based on discrete wavelength transform on SDO data.

*2) **Imagery Instruments on SDO Spacecraft**:* SDO data are captured using three instruments onboard that gather data from the Sun 24 hours a day. The *Helioseismic and Magnetic Imager* (HMI) was created to investigate oscillations and the magnetic field at the solar surface, or photosphere [30]. The *Atmospheric Imaging Assembly* (AIA) on SDO takes full-sun images (1.3 solar diameters) of the solar corona at a spatial resolution of near 1 arcsec, with an image cadence of 12 seconds for multiple wavelengths [31]. To better understand variations on the timeframes that affect Earth's climate and near-Earth space, the *Extreme ultraviolet Variability Experiment* (EVE) analyzes the solar Extreme UltraViolet (EUV) irradiance with high spectral precision [32].

*3) **Machine/Deep Learning on SDO Data**:* Recently [33] has gathered a portion of SDO raw data and cleaned it as machine-learning ready dataset to be used in developing new learning-based methods on SDO mission data. From here now on, we call this dataset *SDOML*. Based on this dataset, [34] used a U-Net in a GAN network to translate AIA multi-spectral (94, 171, 193 Å) images to a specific wavelength (211 Å). As another machine learning work on SDOML dataset, [35] proposed deep neural networks as a means to auto-calibrate the instrument degradation on SDO imagery instruments. A conditional GAN is used in [36] to translate downloaded HMI images from SDO to AIA images. More details on the SDOML dataset will be presented in section IV-A.

III. Methods

A. Generative Image Compression

Autoencoder based learned image compression networks, like the one we have proposed in Figure 2, generally consist of two major parts. First the encoder/decoder network, and second the bottleneck entropy estimation network. The latter is discussed in depth in section III-A2. According to Figure 2, the network input ($\boldsymbol{x}$) and output ($\boldsymbol{x'}$) relations can be summarized as follows

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

in which, $\lfloor \cdot \rceil$ denotes quantizing the real valued input to the nearest integer number. $\hat{\boldsymbol{y}}$ is the quantized latent variable and $\hat{\boldsymbol{z}}$ is its quantized hyper-prior. Encoder and decoder nonlinear transforms are represented by g_a and g_s with their learned parameters, $\boldsymbol{\phi_g}$ and $\boldsymbol{\theta_g}$, respectively. The subscripts a and s refer to *analysis* and *synthesis* as they are common words in the area of transform coding based compression. h_a is the analysis transform to get the hyper-priors of the entropy estimation model, leaned by its parameters $\boldsymbol{\phi_h}$.

*1) **Learning Objective**:* Any learned image compression network tries to tackle with rate-distortion trade-off, governed by a Lagrangian coefficient λ which can be described as

$$R + \lambda D, \tag{2}$$

where R and D correspond to the estimated entropy of the latent code and reconstruction distortion, respectively. Estimated entropy of the quantized bottleneck represents the rate term which is desired to be minimized during the training of the neural network. The probability distribution of the latent code is variationally approximated by hyper-prior $\boldsymbol{z}$. Then the quantized $\hat{\boldsymbol{z}}$ is transmitted alongside the compressed image as a side-information. Therefore, the entropy of both should be optimized as defined below

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

where $\boldsymbol{\theta_h}$ and $\boldsymbol{\psi}$ are parameters of learned entropy model on latent code ($\hat{\boldsymbol{y}}$) and hyper-prior ($\hat{\boldsymbol{z}}$), respectively.

In Eq. 2, D accounts for the distortion between input and output image of the network which can be measured by any desired metric. The prevalently used criterion to measure distortion between input and output is the Mean Squared Error (MSE), which is heavily criticized because of reconstructing blurry images. Efforts have been made to propose metrics which can adhere perceptually to human visual system, e.g. Multi Scale Structural SiMilarity Index (MS-SSIM) [37]. Even these metrics have shown weaknesses when intensely scrutinized [38].

Recently, perceptual-aware metrics based on features generated by pre-trained neural networks have been proposed. Learned Perceptural Image Patch Similarity (LPIPS) introduced by [39] uses trained AlexNet/VGGNet features to compare patches of an image with a corresponding reference. In training our neural compressor we will fortify its reconstruction loss by exploiting this perceptual metric.

To make the reconstruction closer to the input image, we also consider adversarial training of our decoder network. Generative Adversarial Networks (GANs) [40] consisting of a generator and a discriminator sub-network, are able to follow the distribution of data at reconstruction instead of just trying to find the nearest pixel values in order to decrease the distortion. In our network the decoder plays the role of the generator. Then the discriminator forces the decoder output to preserve the distribution of the input image at the reconstructed image. The proposed objective to be optimized is a combination of distortion and perception as follows

$$\begin{aligned} D = \mathbb{E}_{\boldsymbol{x} \sim p_X}[&\lambda_{recon} MSE(\boldsymbol{x}, \boldsymbol{x'}) \\ &+ \lambda_{perc} LPIPS(\boldsymbol{x}, \boldsymbol{x'}) \\ &- \lambda_{adv} \log D(\boldsymbol{x'}, \boldsymbol{y})]. \end{aligned} \tag{4}$$

To make the adversarial training feasible we need the discriminator to judge whether its input sample came from the true distribution of data or is a fake generated one. The

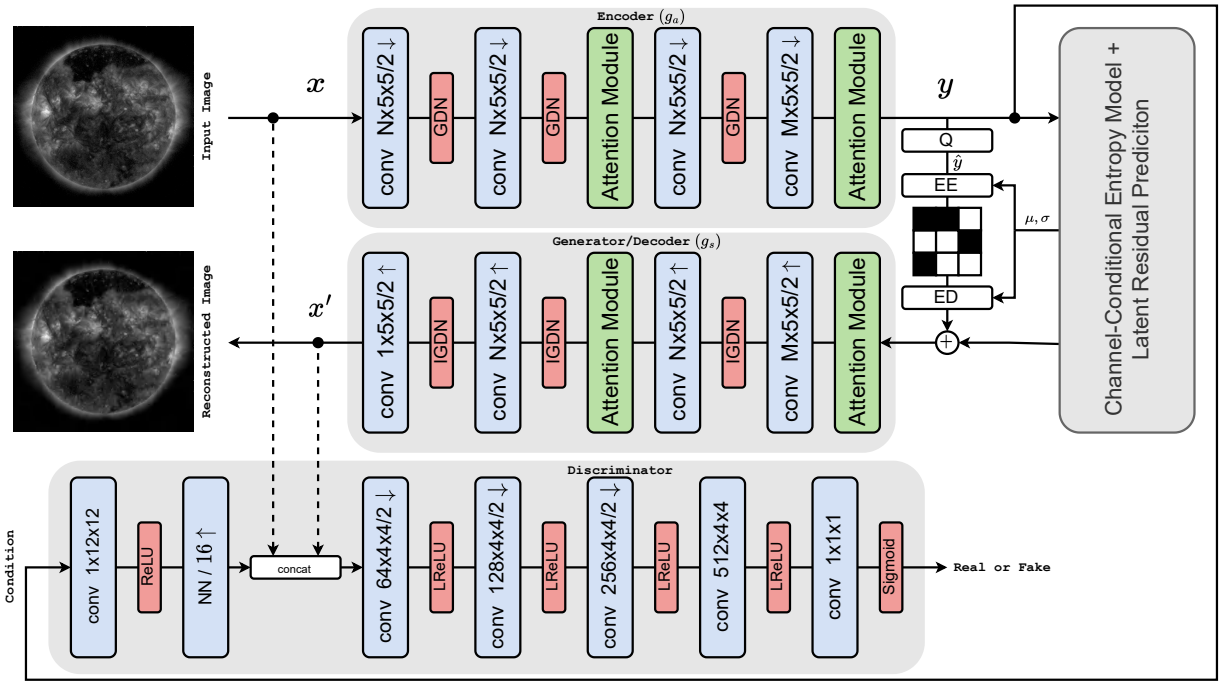


Fig. 2. Network architecture. Input image is down-scaled by a factor of 16 to get the latent code and up-sampled in reverse to get the reconstructed image. A conditional discriminator encourages the generator (decoder) for better perceptual quality. Number of channels in encoder and decoder are set by $N = 192$ and $M = 320$. Q performs the scalar quantization. EE and ED indicate entropy encoder and decoder, respectively. μ and σ are predicted parameters of the latent code probability distribution, defined by entropy estimation model. GDN and IGDN correspond to Generalized Divisive Normalization nonlinearity and its inverse, discussed in section IV-B.

discriminator will need to be optimized by a separate auxiliary loss, given as

$$\begin{aligned} L_{disc.} =& \mathbb{E}_{\boldsymbol{x}\sim p_X}[-\log(D(\boldsymbol{x},\boldsymbol{y})] \\ &+\mathbb{E}_{\boldsymbol{x}\sim p_X}[-\log(1-D(\boldsymbol{x'},\boldsymbol{y}))]. \end{aligned} \tag{5}$$

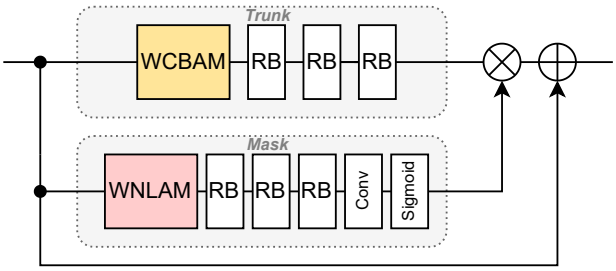
It has been shown analytically that the distortion is in direct trade-off with perception [41]. GANs are the solutions to find a better perception quality by losing an acceptable amount of distortion. In [5] a third term in this tradeoff was introduced as the rate in the lossy compression scheme. More detailed experiments have been adopted in [6] to prove this idea in practice. Therefore, it would be an expected behavior to have a lower Peak Signal to Noise Ratio (PSNR) value on a decoder trained adversarially in contrast to a decoder trained merely on distortion metrics.

*2) **Entropy Estimator Model**:* Performance of any learned image compression scheme heavily depends on how well it can estimate the true entropy of the bottleneck. So the objective will be to minimize the cross entropy between the two. To make the entropy estimation possible, several probability estimation methods have been proposed in the literature, including empirical histogram density estimation [24], [17], piecewise linear models [11], conditioning on a latent variable (hyper-prior) [25] and context modelling based on autoregressive models. [42].

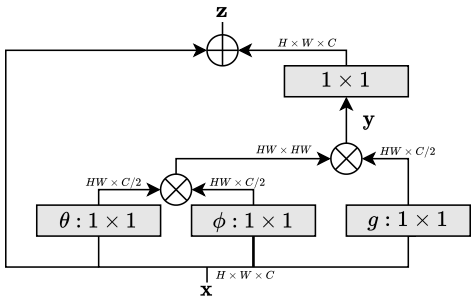
From a high-level overview, entropy estimation models can be divided into two main categories, Forward Adaptation (FA) and Backward Adaptation (BA) models. The former suffers from low capacity to capture all dependencies in the probability distribution of the latent code and the latter's disadvantage is that decoding process cannot be parallelized. Learned FA models [25], [12] will only use the information provided during the encoding of the image, while BA methods based on autoregressive models [42] need information from the decoded message as well. To take advantage of both of these models [43] we define the conditional probability of the latent code as

$$P_{\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}) = \prod_i P(\hat{\boldsymbol{y}}_{\boldsymbol{i}}|\hat{\boldsymbol{y}}_{\boldsymbol{j}<\boldsymbol{i}}, \hat{\boldsymbol{z}}; \boldsymbol{\theta_h}). \tag{6}$$

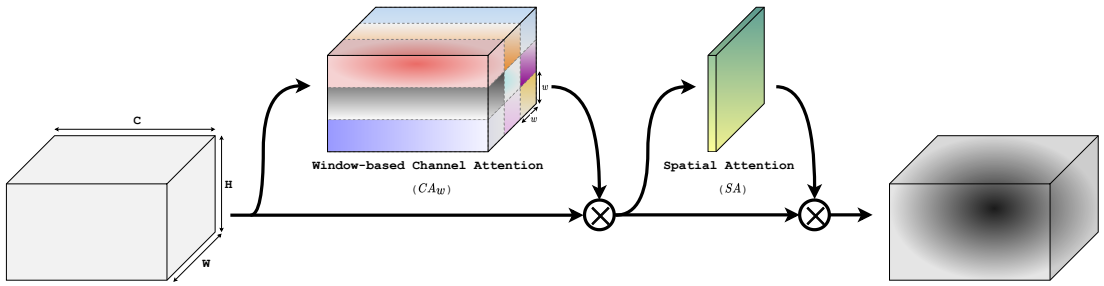
Conditioning on quantized hyper-prior, i.e., $\hat{\boldsymbol{z}}$, as a side-information is an example of FA and conditioning on all previously decoded elements of the latent space, i.e., $\hat{\boldsymbol{y}}_{\boldsymbol{j}<\boldsymbol{i}}$, is an example of BA. BA performance have been improved in [43] by letting the conditioning exist only between slices of channels in the bottleneck. In contrast to spatial autoregressive modeling in [42], [43] only considers the conditioning of the probabilities on the channels and it showed that by doing this the decoding process could be reasonably parallelized. We have used the same approach in [43] to estimate the entropy and minimize it during the training.



(a) Attention module with skip connection. RB denote residual block.



(b) Window-based non-local attention module (WNLAM).



(c) Window-based convolutional block attention module (WCBAM). A feature map of C channels with spatial dimensions of H×W. w is the window size to calculate channel attention on.

Fig. 3. Attention module architecture.

B. Attention Assisted Image Compression

When it comes to computer vision, deep convolutional neural networks are the de facto standard despite their poor performance on capturing long range dependencies. CNNs will have problems if it is required to simultaneously capture a few characteristics from non-neighboring pixels. It has been determined that the local nature of kernel sliding on just a few pixels of the input image is the primary cause of this degradation [44].

Efforts have been made to help CNNs capture more robust representation of the input image. One naive solution is to make the network deeper but other problems will arise on training such networks which has led to the introduction of deep residual networks (ResNet) [45]. Although increasing the parameters of network will generally lead to richer representation and better performance, it will make training of such networks harder. Attention mechanisms have been proposed to address this issue of CNNs without making the network deeper. In [46] authors have proposed a single module to be included in between sequential convolutional layers, consisting of two branches, *trunk*, to process local features and *mask* to decide which of the local features in the trunk are more important to be passed to the next convolutional layer, as in Figure 3(a).

In contrast to local attention, [47] first discussed how non-local attention can be viewed as a special case of non-local algorithm which was traditionally used as a method to denoise images [48]. The idea was to find similar pixels/patches in the image/feature map and replace it with a weighted sum over all the others, with higher weights for more similar ones. It can be inferred from [47] that Vision Transformers (ViT) [8] are all special cases of non-local attention mechanism.

Non-local attention block (Figure 3(b)) helps the *mask* branch to efficiently learn the most informative parts of features (in the *trunk*) for the task in hand [49]. The authors in [49] also added a skip connection to help the output feature maps be richer. This skip connection prevents vanishing gradients as well. Another recently proposed simple tweak to incorporate attention in CNNs has been introduced in [50]. It is an enhanced version of Squeeze-and-Excitation network [51], to apply attention both on spatial and channels feature maps separately. This way of applying attention is simpler and computationally more efficient.

Discussed attention mechanisms have been employed in deep learned neural compression networks as well. [52] applied residual attention, then [53] improved their work by adding a non-local attention to the mask of the residual attention. To improve further, [54] applied non-local attention limited to small windows of the feature maps. This window-based attention attained better results in the compression area. Here we propose to use two kinds of attention mechanisms in a window-based manner.

*1) **Window-based Non-Local Attention Module (WNLAM)**:* Non-local attention block as shown in Figure 3(b) is composed of a weighted sum (weights calculated as a softmax)

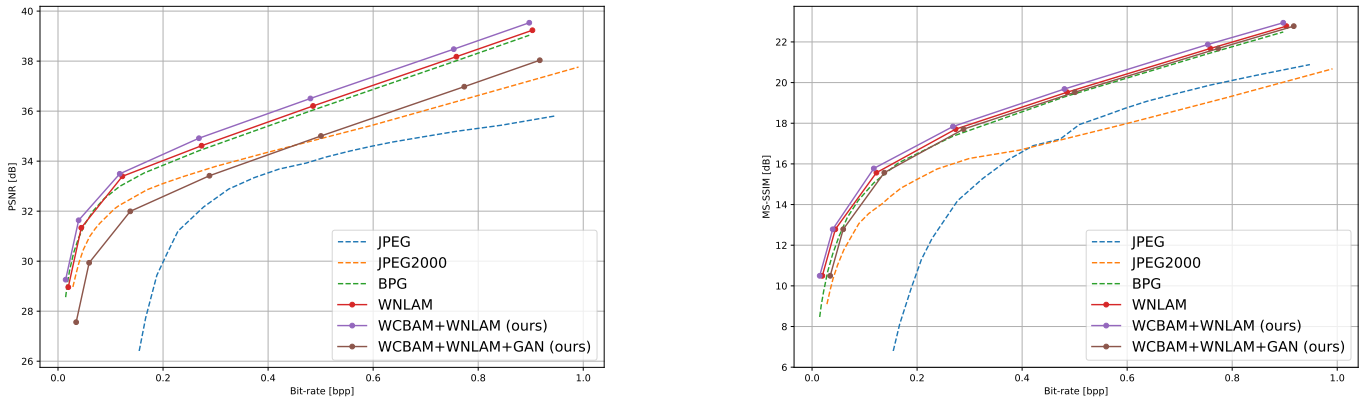


Fig. 4. Rate distortion curves aggregated over test set described on section IV-A. On the left PSNR is calculated from MSE by $10\log_{10}\frac{255^2}{MSE}$. On the right MS-SSIM is reported in logarithmic scale by $-10\log(1-m)$ to show the differences better, in which m is the MS-SSIM in the range of zero to one.

over linear transformed version of $\mathbf{x}$, i.e., $g(\mathbf{x})$.

$$\mathbf{y}_i = \frac{1}{\sum_{\forall j} e^{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}} \sum_{\forall k} e^{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_k)} g(\mathbf{x}_k), \tag{7}$$

where $g(.)$ is a linear transformation (W_g) implemented by a 1×1 convolution layer defined as $g(\mathbf{x_k}) = W_g \mathbf{x_k}$. The weights of the sum in Eq. 7 are calculated by the measure of similarity in embedding space of the input, i.e., $\theta(\mathbf{x}_i) = W_\theta \mathbf{x_i}$ and $\phi(\mathbf{x}_k) = W_\phi \mathbf{x_k}$.

As the final operation in non-local attention, $\mathbf{z}_i$ is calculated by a linear transformation (W_z) added to the original $\mathbf{x}_i$ as follows

$$\mathbf{z}_i = W_z \mathbf{y}_i + \mathbf{x}_i. \tag{8}$$

In image compression, restoring edges and high-frequency content is more important than representing the global features in the latent representation. Consequently, naive non-local attention mechanism performs worse than local attentions which are able to capture local redundancies and preserve details on the reconstructed image [54].

*2) **Window-based Convolutional Block Attention Module (WCBAM):*** A simple to implement kind of attention in CNNs is the convolutional block attention module (CBAM) which has shown great benefit in classification tasks [50]. It carries out two attention mechanisms. First, the channel attention (CA) guides the network to only consider channels with higher importance for the desired task. Second, the spatial attention (SA) dictates the network where to pay attention more. Here we propose to utilize this attention module in a window-based manner. Instead of globally considering the whole spatial dimensions of each channel, we focus only on a cropped window size of w, as shown in Figure 3(c).

Applying WCBAM mechanism on the input features X can be summarized as

$$\begin{aligned} X_{CA} &= CA_w \odot X, \\ X_{CA,SA} &= SA \odot X_{CA}, \end{aligned} \tag{9}$$

where CA_w reweighs the channels over each window. Then SA is multiplied on each refined channel to highlight the important spatial content.

Window based channel attention is calculated by passing the average and max pool through a shared fully connected network (F), as in Eq. 10

$$CA_w = sigmoid(F(Avg(X_w)) + F(Max(X_w))). \tag{10}$$

Afterwards, the spatial attention weights (SA) will be derived by concatenating average and max pool passed through a convolutional layer as

$$SA = sigmoid(Conv([Avg(X_{CA}), Max(X_{CA})])). \tag{11}$$

WCBAM helps the network to capture global dependencies which may be needed during transforming the image from pixel space to feature space, specially those which the window-based non-local attention are incapable of.

Transformers as Attention Modules: The superiority of models based on Transformers which are a special kind of non-local attention mechanism, has been recently proven [8], [9]. Although Transformers have shown great benefit in image classification or object detection tasks, their naive application in image compression networks has failed [54]. The whole purpose of transformers is to capture long-range dependencies in an image, while the ultimate goal in image compression is to capture dependencies in order to summarize the dependencies efficiently in the latent code.

IV. Experiments

A. Dataset

SDOML includes images of the sun at wavelengths 94, 131, 171, 193, 211, 304, 335, 1600, 1700 Å at a cadence of 6 minutes. We downsampled the images to a cadence of 1 hour to avoid the dependency between training samples. In addition, to prevent biases of the images with respect to solar variations at different stages of the solar cycle, we followed the same approach proposed by [34] to divide the dataset based on the month they are taken. Images of January to August are

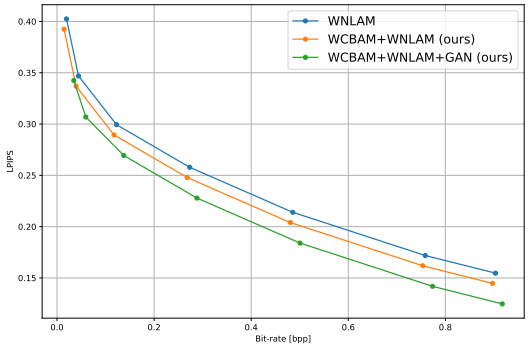


Fig. 5. Rate distortion curve. Distortion is measured by LPIPS metric (lower is better) described in section III-A1. As it can be seen, GAN performance in generating high quality image can be quantified by this metric.

chosen for training and September to December are reserved for testing. The results reported in this section are all based on this portion of dataset.

B. Implementation Details

As nonlinearity in our neural network, we have utilized computationally efficient [55] version of Generalized Divisive Normalization (GDN) [56]. As a result of GDN's local normalization, statistical dependencies are reduced in the feature maps. By exploiting GDN instead of more conventional nonlinearities like ReLU, the feature maps will be decorrelated. Ideally, scalar quantization of a set of decorrelated features will present compression performance close to parametric vector quantization [12]. During the evaluation phase, entropy coding of the latent integer values was realized by asymmetric numeral systems [57].

Seven models have been trained with $\lambda \in \{0.0015, 0.0035, 0.0070, 0.0125, 0.0250, 0.0410, 0.0550\}$ governing the rate-distortion trade off as in Eq. 2 for 100 epochs to train each model. We have used Adam [58] optimizer on batches of size 16 consisting of randomly cropped 256×256 patches out of the original images of 512×512. Initial value of the learning rate is set to 10^{-4} and annealed during the training to 1.2×10^{-6}.

All common reconstruction losses have been blamed for not being close to human perceptual vision [59]. L1 loss pays more attention to edges and high frequency areas of the image, L2 loss results in blurry reconstruction and SSIM/MS-SSIM losses will have an effect of not reconstructing minute details like text in images. Training our autoencoder based on MSE and LPIPS will result in outperforming even the state of the art hand-engineered codec, i.e., BPG [60], as shown in Figure 4.

The general lower performance of our GAN network is a common issue addressed in [41]. The PSNR or MS-SSIM are unable to capture the perceptual quality of the generated image in a GAN. The perceptual quality of GAN network reconstructions is discussed in section IV-C, measured by perceptual metrics.

C. Ablation Study

To investigate how much attention modules contribute to the performance of our neural compressor, we have trained three separate networks for each of the seven targeted bit-rates discussed in section IV-B. The first architecture has only the WNLAM module (Figure 4) whose performance in terms of PSNR and MS-SSIM has been improved by adding the WCBAM attention module.

As emphasized in Figure 1, the adversarially trained decoder results in better visual quality on the reconstructed image than the autoencoder only trained with attention mechanism. Conventional metrics like PSNR and MS-SSIM are unable to capture the higher perceptual quality of the GAN reconstructed images. We empirically found that LPIPS can show the merit of adversarially trained network. As it is shown in Figure 5, LPIPS values correspond to the human judgment of the quality of reconstructed images.

V. Conclusion

In this work, we have shown how an effective image compression scheme based on trainable neural networks could be utilized for ad-hoc applications like images from NASA's SDO mission. In addition, we explored the effectiveness of attention mechanisms in an adversarially trained neural network to improve performance of compression in terms of rate-distortion-perception trade-off.

References

[1] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, "Variable rate image compression with recurrent neural networks," in *4th International Conference on Learning Representations*, 2016.

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

[3] 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.

[4] Y. Yang, S. Mandt, and L. Theis, "An introduction to neural data compression," *CoRR*, vol. abs/2202.06533, 2022.

[5] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*. PMLR, 2019.

[6] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing*, 2017.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] H. A. Dehkordi, A. S. Nezhad, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "Multi-expert human action recognition with hierarchical super-class learning," *Knowledge-Based Systems*, vol. 250, p. 109091, 2022.

[14] M. S. E. Saadabadi, S. R. Malakshan, S. Soleymani, M. Mostofa, and N. M. Nasrabadi, "Information maximization for extreme pose face recognition," *arXiv preprint arXiv:2209.03456*, 2022.

[15] M. Akyash, H. Mohammadzade, and H. Behroozi, "Dtw-merge: A novel data augmentation technique for time series classification," *arXiv preprint arXiv:2103.01119*, 2021.

[16] H. Kashiani and S. B. Shokouhi, "Visual object tracking based on adaptive siamese and motion estimation network," *Image and Vision Computing*, vol. 83-84, pp. 17–28, 2019.

[17] 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.

[18] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.

[19] 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.

[20] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, 1998.

[21] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," in *International Conference on Learning Representations*, 2016.

[22] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[23] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Morph detection enhanced by structured group sparsity," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 311–320.

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

[25] 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.

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

[27] "A guide to the mission and purpose of nasa's solar dynamics observatory," 2010. [Online]. Available: https://sdo.gsfc.nasa.gov/assets/docs/SDO_Guide.pdf

[28] P. Chamberlin, W. D. Pesnell, and B. Thompson, *The Solar Dynamics Observatory*. Springer Science & Business Media, 2012.

[29] C. E. Fischer, D. Müller, and I. De Moortel, "JPEG2000 image compression on solar EUV images," *Solar Physics*, vol. 292, 2017.

[30] J. Schou, P. H. Scherrer, R. I. Bush, R. Wachter, S. Couvidat, M. C. Rabello-Soares, R. S. Bogart, J. Hoeksema, Y. Liu, T. Duvall *et al.*, "Design and ground calibration of the helioseismic and magnetic imager (HMI) instrument on the solar dynamics observatory (SDO)," *Solar Physics*, vol. 275, no. 1, 2012.

[31] J. R. Lemen, D. J. Akin, P. F. Boerner, C. Chou, J. F. Drake, D. W. Duncan, C. G. Edwards, F. M. Friedlaender, G. F. Heyman, N. E. Hurlburt *et al.*, "The atmospheric imaging assembly (AIA) on the solar dynamics observatory (SDO)," in *The solar dynamics observatory*, 2011.

[32] T. Woods, F. Eparvier, R. Hock, A. Jones, D. Woodraska, D. Judge, L. Didkovsky, J. Lean, J. Mariska, H. Warren *et al.*, "Extreme ultraviolet variability experiment (EVE) on the solar dynamics observatory (SDO): Overview of science objectives, instrument design, data products, and model developments," *The solar dynamics observatory*, 2010.

[33] 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.

[34] V. Salvatelli, S. Bose, B. Neuberg, 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," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[35] L. F. G. dos Santos, S. Bose, V. Salvatelli, B. Neuberg, 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.

[36] A. Dash, J. Ye, and G. Wang, "High resolution solar image generation using generative adversarial networks," *arXiv preprint arXiv:2106.03814*, 2021.

[37] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, 2009.

[38] J. Nilsson and T. Akenine-Möller, "Understanding SSIM," *arXiv preprint arXiv:2006.13846*, 2020.

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, 2014.

[41] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6228–6237.

[42] D. Minnen, J. Ballé, and G. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing*, 2018.

[43] 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.

[44] 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.

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

[46] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *CVPR*, 2017, pp. 3156–3164.

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

[48] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005.

[49] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *International Conference on Learning Representations*, 2019.

[50] 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.

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

[52] L. Zhou, Z. Sun, X. Wu, and J. Wu, "End-to-end optimized image compression with attention mechanism," in *CVPR Workshops*, June 2019.

[53] 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*, vol. 30, 2021.

[54] R. Zou, C. Song, and Z. Zhang, "The devil is in the details: Window-based attention for image compression," in *CVPR*, 2022.

[55] N. Johnston, E. Eban, A. Gordon, and J. Ballé, "Computationally efficient neural image compression," Google Research, Tech. Rep., 2019.

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

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

[59] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Computational Imaging*, 2017.

[60] F. Bellard, "Bpg image format," https://bellard.org/bpg/.