The Devil Is in the Details: Window-based Attention for Image Compression

Renjie Zou [1]   Chunfeng Song [1]   Zhaoxiang Zhang [1,2,*]

[1] National Laboratory of Pattern Recognition (NLPR) of
Institute of Automation, Chinese Academy of Sciences (CASIA) &
University of Chinese Academy of Sciences (UCAS)

[2] Centre for Artificial Intelligence and Robotics, HKISI_CAS

{zourenjie2020, chunfeng.song, zhaoxiang.zhang}@ia.ac.cn

Abstract

Learned image compression methods have exhibited superior rate-distortion performance than classical image compression standards. Most existing learned image compression models are based on Convolutional Neural Networks (CNNs). Despite great contributions, a main drawback of CNN based model is that its structure is not designed for capturing local redundancy, especially the non-repetitive textures, which severely affects the reconstruction quality. Therefore, how to make full use of both global structure and local texture becomes the core problem for learning-based image compression. Inspired by recent progresses of Vision Transformer (ViT) and Swin Transformer, we found that combining the local-aware attention mechanism with the global-related feature learning could meet the expectation in image compression. In this paper, we first extensively study the effects of multiple kinds of attention mechanisms for local features learning, then introduce a more straightforward yet effective window-based local attention block. The proposed window-based attention is very flexible which could work as a plug-and-play component to enhance CNN and Transformer models. Moreover, we propose a novel Symmetrical TransFormer (STF) framework with absolute transformer blocks in the down-sampling encoder and up-sampling decoder. Extensive experimental evaluations have shown that the proposed method is effective and outperforms the state-of-the-art methods. The code is publicly available at https://github.com/Googolxx/STF.

1. Introduction

Image compression is a fundamental and long-standing research topic in image processing field. With the ever increasing visual applications, lossy image compression is a
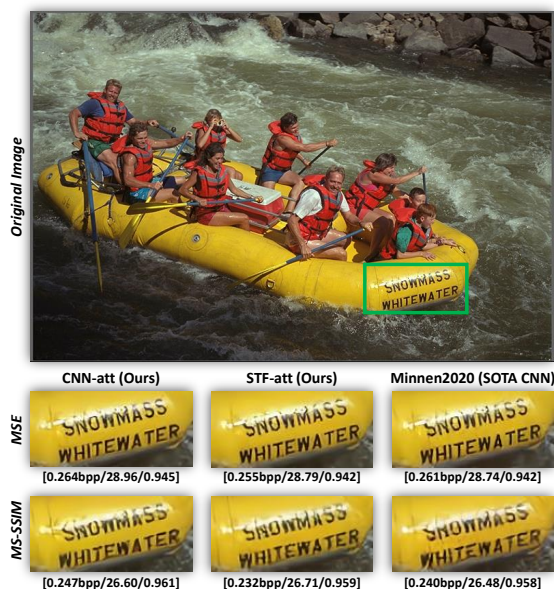

Figure 1. Comparison of image reconstruction by different methods, including CNN + window-attention (CNN-att), Symmetrical TransFormer + window-attention (STF-att), and previous SOTA CNN model (MBT). The upper shows results optimized for MSE, while the lower is optimized for MS-SSIM. It is obvious that the proposed window-based attention is effective for both supervision, especially with the STF model, showing that the proposed window-based attention model could internally allocate more bits on high contrast areas and achieve better visual quality. The metrics are [bpp↓/PNSR↑/MS-SSIM↑].

vital technique for storing images and videos efficiently in limited hardware resources. Classical lossy image compression standards including JPEG [44], JPEG2000 [40], BPG [8], and VVC [9] follow a similar coding scheme: transforming, quantization, and entropy coding. However, these image compression standards rely heavily on the handcrafted rules, which means they are not expected to be of the ultimate solution for image compression.

---

* Corresponding Author

In recent years, the learned image compression based on variational auto-encoder (VAE) [24] has achieved better rate-distortion [41] performance than conventional lossy image compression methods on metrics of Signal-to-Noise Ratio (PSNR) and Multi-Scale-Structural Similarity Index Measure (MS-SSIM) [46], showing great potential for practical compression use. Here, we would briefly introduce the general pipeline [5] of the VAE-based methods. For encoding, the VAE-based image compression methods use a linear and nonlinear parametric analysis transform to map the images to a latent code space. After quantization, entropy estimation modules predict the distributions of latents, then the lossless Context-based Adaptive Binary Arithmetic Coding (CABAC) [30] or Range Coder [31] compresses the latents into the bit stream. Meanwhile, hyper-prior [6], auto-regressive [34] priors and Gaussian Mixture Model (GMM) [15] allow the entropy estimation modules to more precisely predict distributions of latents, and achieve better rate-distortion (RD) performance. For decoding, lossless CABAC or Range Coder decompresses the bit stream, then the decompressed latents are mapped to reconstructed images by a linear and nonlinear parametric synthesis transform. Further, there are also some works [21,27,48] designing post-processing networks for better quality of reconstruction. Combining above sequential units, those models could be trained end-to-end. Although great progresses have been achieved, one core problem of above CNN-based model is that the original convolutional layer is designed for the high-level global feature distillation, rather than the low-level local detail restoration. As shown in the right side of Fig.1, even the SOTA CNN model is still affected by the weak local detail learning ability which would inevitably limit further performance improvement.

Inspired by the success of attention mechanism in natural language processing (NLP) and computer vision tasks such as the image classification and semantic segmentation, many researchers apply the non-local attention mechanism to guide the adaptive processing of latent features, which could help the compression algorithm allocate more bits to challenging areas (i.e., edges, textures) for better RD performance. However, such non-local attentions still not change the intrinsic global-aware character of the CNN structure. Recent studies [13,18,29,42] have demonstrated that transformer [43] can be successfully applied to vision tasks with competitive even better performance compared with convolutional neural networks (CNNs). Those attention-based networks, such as Vision Transformer [18] and Swin Transformer [29] take the advantages of attention mechanism to capture global dependency. However, we intuitively find that the global semantic information in image compression is not as effective as in other computer vision tasks. Instead, spatially neighboring elements have stronger correlation.

Following above discussions, this paper explores to ad-

dress the detail missing problem from two aspects, i.e., studying *the local-aware attention mechanism* and introducing *a novel transformer-based framework*. **First**, we comprehensively study how to combine the neural networks with the attention mechanism for designing local-aware lossy image compression architectures. Through conducting a set of comparative experiments based on a global attention mechanism and a local attention mechanism, we have verified our aforementioned guess, that the local attention is more suitable for the local texture reconstruction. Afterward, we present a flexible attention module combined with neural networks to capture correlations among spatially neighboring elements, namely window-attention. As shown in Fig.1, the proposed attention module could work as a plug-and-play component to enhance CNN and Transformer models. **Second**, although the transformer-based models have achieved great success in a variety of computer vision tasks, there are still great challenges applying transformer model in image compression, e.g., no up-sampling units, fixed attention model. To this end, we propose a novel Symmetrical TransFormer (STF) framework with absolute transformer blocks in down-sampling encoder and up-sampling decoder, which may be the first exploration of designing up-sampling transformer, especially for the image compression task. Extensive experimental results show that our methods are superior to the state-of-the-art (SOTA) image compression methods in essential metrics. The main contributions of this paper are summarized as follows:

- We extensively study the local-aware attention mechanism, and find that it is crucial to combine the global structure learned by neural networks and the local texture mined by the attention units.
- We present a flexible window-based attention module to capture correlations among spatially neighboring elements, which could work as a plug-and-play component to enhance CNN or Transformer models.
- We design a novel Symmetrical TransFormer (STF) framework with absolute transformer blocks in both down-sampling encoder and up-sampling decoder.
- Extensive experimental evaluations have shown that the proposed methods are effective and outperform the SOTA image compression methods.

2. Related Works

Learned Image Compression. Recently, learned image compression models [5,6,15–17,19,27,32,34,35] based on CNNs exhibit a fast development trend and achieve significant breakthroughs. For VAE-based architectures, [5] firstly proposes an end-to-end optimized model for image compression. [6] incorporates a hyper-prior to effectively capture spatial dependencies in the latent representation based on [5]. Inspired by the success of auto-regressive priors in probabilistic generative models, [34] further enhances the

entropy model in [6] by adding an auto-regressive component. In addition to this, [15] enhances the network architecture by using residual blocks and adding a simplified attention module with the Gaussian Mixture Model (GMM) replacing the mostly used Single Gaussian Model (SGM). Though the GMM-based entropy model performs better in RD performance, the adopted probability distribution function (PDF) and cumulative distribution function (CDF) have to be dynamically generated for every single element in encoding and decoding, thereby introducing much redundancy and making it time-consuming. In contrast, the SGM-based entropy model can build fixed PDF and CDF tables for entropy coding, which is less computationally expensive. To minimize serial processing steps in auto-regressive context models, [35] proposes a channel-wise auto-regressive entropy model. Some approaches [2,33] use the Generative Adversarial Networks (GANs) to directly learn the distribution of images and prevent compression artifacts. For GAN-based architectures, image compression is a rate-distortion-perception trade-off task.

Attention Mechanism. Attention mechanism mimics the internal process of biological observation, devoting more attention resources to the key regions to obtain more details and suppress other useless information. Non-local attention mechanism [45] has been proven beneficial in various vision tasks. For learned image compression, [28] applies the non-local attention to generate implicit importance masks to guide the adaptive processing of latent features, while [15] simplifies the attention mechanism by removing the non-local block.

Transformer-based Model. Inspired by the success of Transformer architectures [43] in the natural language processing, there are many works exploring the potentiality of Transformer in computer vision tasks. [18] conducts image classification with transformer architecture. [13] implements transformer-based models in detection. [14] proposes a universal pre-training approach for image processing tasks. [29] proposes a hierarchical Transformer computed with shifted windows and achieves SOTA performance in image classification, semantic segmentation, and object detection. However, as far as our knowledge, there is no transformer related work for image compression. In this paper, we explore how to design a transformer-based architecture for learned image compression to achieve comparable RD performance.

Normalization in Neural Networks. Generalized divisive normalization (GDN) [4] is a milestone in learned image compression. GDN is highly efficient in Gaussianizing the local joint statistics of natural images. We have conducted some comparative experiments by replacing GDN with batch normalization [22], layer normalization [3], and channel normalization [33]. We found that replacing GDN in CNN would result in a huge drop in RD performance.

However, we also found that GDN is unstable in a deep Transformer-based architecture. In addition, GDN is not compatible with the attention mechanism in Transformer blocks.

3. Method

3.1. Formulation

Since our method is built upon the hyper-prior architecture [6,34] and channel-wise auto-regressive entropy model [35], we would briefly introduce the basic pipeline for better understanding.

The encoder E maps a given image x to a latent y. After quantization Q, $\hat{y}$ is the discrete representation of the latent y. Then $\hat{y}$ is mapped back to the reconstructed image $\hat{x}$ using the decoder D. The main process is formulated as:

$$
\begin{aligned}
y &= E(x;\phi) \\
\hat{y} &= Q(y) \\
\hat{x} &= D(\hat{y};\theta)
\end{aligned}
\tag{1}
$$

where ϕ and θ are trainable parameters of the encoder E and decoder D. Quantization Q inevitably introduces clipping errors of the latent ($error = y - Q(y)$), which leads to distortion of the reconstructed image. Following previous work [35], in the training phase, we also modify the quantization error by rounding and adding the predicted quantization error.

We model each element $\hat{y}_i$ as a single Gaussian distribution with its standard deviation σ_i and mean μ_i by introducing a side information $\hat{z}_i$. The distribution $p_{\hat{y}_i|\hat{z}_i}$ of $\hat{y}_i$ is modeled by a SGM-based entropy model:

$$
p_{\hat{y}_i|\hat{z}_i}(\hat{y}_i|\hat{z}_i) = \mathcal{N}(\mu_i, \sigma_i^2)
\tag{2}
$$

The loss function of image compression model is:

$$
\begin{aligned}
\mathcal{L} =& R + \lambda \cdot D \\
=& \mathbb{E}_{x \sim p_x}[-\log_2 p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z}) - \log_2 p_{\hat{z}}(\hat{z})] \\
& + \lambda \cdot \mathbb{E}_{x \sim p_x}[d(x, \hat{x})]
\end{aligned}
\tag{3}
$$

where λ controls the trade-off between rate and distortion, R is the bit rate of latents $\hat{y}$ and $\hat{z}$, $d(x, \hat{x})$ is the distortion between the raw image x and reconstructed image $\hat{x}$.

3.2. Window-based Attention

Most of previous approaches [15, 28, 49] apply the attention mechanism to generate attention masks based on a global receptive field. Attention mechanism has also been successfully adopted in many computer vision tasks, such as the image classification, semantic segmentation and object detection. However, intuitively, global semantic information in image compression is not as practical as in those computer vision tasks.
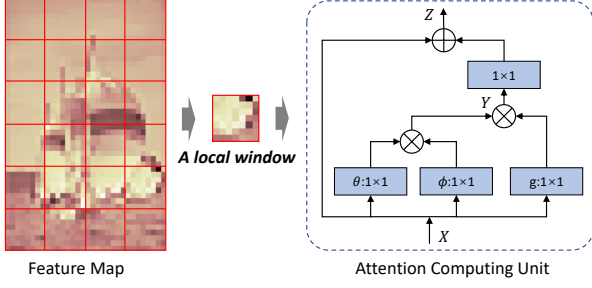
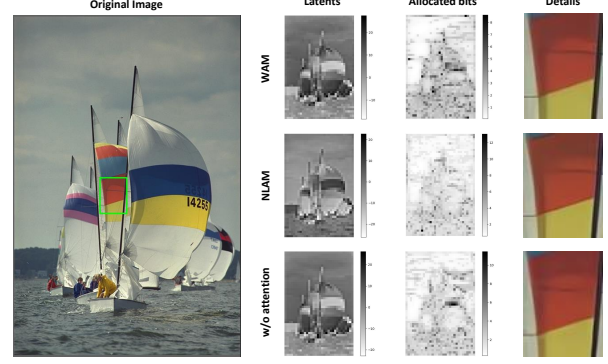Figure 2. An illustration of the window-based attention. Attention masks are computed in a local window.

Figure 3. Visualization of WAM, NLAM, and w/o attention module for the channel with maximal entropy. It shows that our WAM focuses on high contrast regions (sailboats) and allocates more bits on those areas while fewer bits on low contrast regions (sky and clouds). Instead, the bits are allocated evenly in the w/o attention module and NLAM.

Attention in Non-overlapped Windows. We notice that generating attention masks based on spatially neighboring elements can improve RD performance with less computing cost. For efficient modeling and focusing on spatially neighboring elements, we propose the window-based attention. As illustrated in Fig. 2, the feature map is divided into windows of $M \times M$ in a non-overlapping manner. We compute the attention map in each window separately, X_i^k and X_j^k are i-th and j-th elements in k-th window, as below:

$$Y_i^k = \frac{1}{C(X^k)} \sum_{\forall j} f(X_i^k, X_j^k) g(X_j^k) \qquad (4)$$

$$\textbf{with } f(X_i^k, X_j^k) = e^{\theta(X_i^k)^T \phi(X_j^k)},$$
$$C(X^k) = \sum_{\forall j} f(X_i^k, X_j^k), \qquad (5)$$
$$g(X_j^k) = W_g X_j^k$$

Here, $\theta(X_i^k) = W_\theta X$ and $\phi(X_i^k) = W_\phi X$, where W_θ and W_ϕ are cross-channel transforms. $f(\cdot)$ is an embedded Gaussian function. $C^k(X)$ is a normalizing factor. For given i and k, $\frac{1}{C^k(X)} f(X_i^k, X_j^k)$ is the $softmax$ computation along the dimension j in k-th window. Residual connection is necessary for such attention mechanism, the output is as follows:

$$Z_i^k = W_z Y_i^k + X_i^k \qquad (6)$$

where W_z is a weight matrix for computing the position-wise embedding on Y_i^k, given in Eq.4.

Window Attention Module. Liu et al. [28] propose the Non-Local Attention Module (NLAM), which consists of a cascade of a Non-local Block and regular convolutional layers. We replace the Non-local Block with Window Block for focusing on regions with high contrast. Fig. 4 (b) shows our Window Attention Module (WAM). We visualize the channel with the highest entropy for WAM, NLAM, and w/o attention module, as shown in Fig. 3. In the 3-rd column, it's obviously observed that WAM can allocate more

bits in complex regions (high contrast) and fewer bits in simple regions (low contrast). In addition, the WAM reconstructed image has sharper texture details. NLAM with global receptive field often leads to evenly allocation of bits on different regions, which is not consistent with the expectation in [28].

3.3. CNN-based Architecture

As shown in Fig. 4, our CNN-based architecture is established upon [35]. We enhance the encoder and decoder by inserting the proposed WAMs, respectively. The WAMs help allocate bits on different areas more rationally and internally with negligible computation overhead. It is simply yet could significantly improve RD performance.

3.4. Transformer-based Architecture

Inspired by the success of Transformer architectures [18, 29] in computer vision and our aforementioned experimental result that the local attention helps allocate bits rationally and improve RD performance, we further propose a novel Transformer architecture for learned image compression, as shown in Fig. 5.

Rethinking the Transformer. Since our goal is to verify whether combining self-attention layers and MLPs can achieve comparable performance with original CNN-based architectures, we design a novel Symmetrical Transformer (STF) framework without convolution layers in the encoder and decoder. The difficulties of designing the Transformer model for learned image compression are the following:

- Previous works are mostly based on the CNNs to eliminate the spatial redundancy and capture the spatial structure. Directly splitting image to patches may result in space redundancy within each patch.
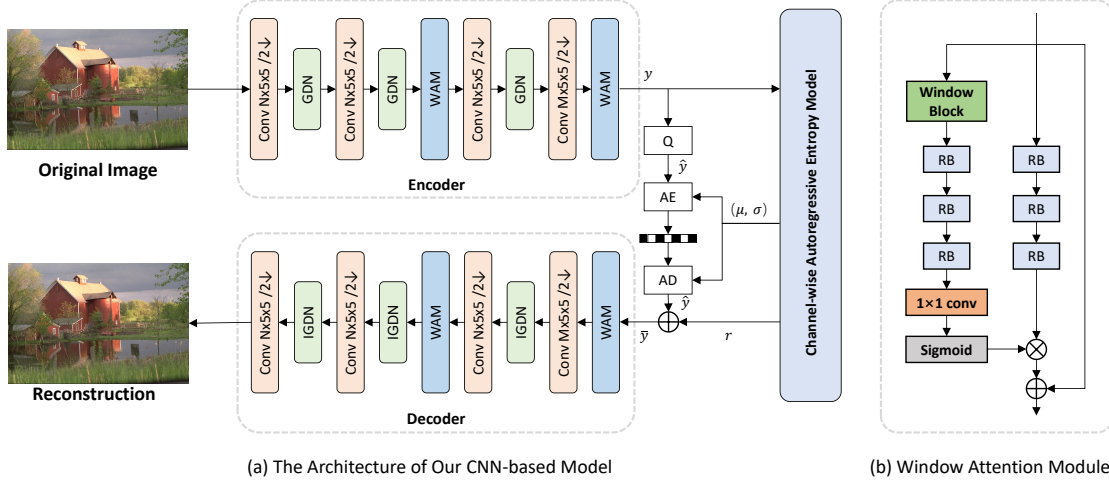
Figure 4. (a) The architecture of our proposed CNN-based model. We adopt the architecture of [35] for the sake of subsequent comparison. IGDN is the inverse GDN. (b) RB is the residual block which consists of 1×1 and 3×3 convolutional layers.

- GDN is the most commonly used normalization and nonlinear activation function in image compression. However, GDN is unstable in the deep Transformer-based architecture. Furthermore, we have found GDN and the attention mechanism in the transformer are incompatible.

- Based on our previous analysis and experimental results, calculating the attention map on a large field is not optimal.

To address above concerns, we choose a small patch size to avoid the space redundancy within each patch. We use the LN for normalization, which is most common used in Transformer. GELU is adopted as the nonlinear activation function in our Transformer architecture. Inspired by [29], we compute the attention map within local windows. The advantage of our Transformer architecture is that it could focus on spatially neighboring patches while gradually expanding the receptive field, with acceptable computational complexity.

Transformer-based Encoder. We split the raw image $x \in R^{3 \times H \times W}$ into patches with patch size N. A linear embedding layer is applied on the raw patches to generate a feature map $f_p \in R^{C \times \frac{H}{N} \times \frac{W}{N}}$ with C channels. The feature map $f_p \in R^{C \times \frac{H}{N} \times \frac{W}{N}}$ is reshaped into a sequence $f_s \in R^{P^2 \times C}$, where $P = \frac{HW}{N^2}$ is the number of patches. Then the sequence f_s will be input into Transformer blocks and patch merging layers. Following the structure in [29], the former calculate attention masks in a window for feed-forwarding. At the same time, the latter down-sample the resolution of features and doubles channels of features.

Transformer-based Decoder. We design a symmetrical decoder consisting of multiple Transformer blocks, patch splitting layers and a de-embedding layer. The patch splitting layers up-sample the resolution of features and halve

channels of features. The de-embedding layer maps the feature map to the reconstructed image $\hat{x}$.

Entropy Model. To more effectively and efficiently predict the probability distribution of the latent. We use an SGM-based channel-wise auto-regressive entropy model [35].

4. Experiments

4.1. Experimental Setup

Training. We implement our proposed CNN-based architecture and Transformer-based architecture in CompressAI platform [7]. For training, we randomly choose 300k images from the OpenImages dataset [26], and randomly crop them with the size of 256×256. All models are trained for 1.8M steps using the Adam optimizer [23] with a batch size of 16. The initial learning rate is set to 1×10^{-4} for $120k$ iterations, and drops to 3×10^{-5} for another $30k$ iterations, 1×10^{-5} for the last $30k$ iterations.

Our models are optimized using two quality metrics (MSE and MS-SSIM) as supervisions. Following the same settings in [7], when the model is optimized for MSE, the lambda values λ belongs to {0.0018, 0.0035, 0.0067, 0.0130, 0.025, 0.0483}. When the model is optimized for MS-SSIM, the lambda values λ belongs to {2.4, 4.58, 8.73, 16.64, 31.73, 60.50}.

For our CNN-based models, the channel number of latent and hyper latent are set to 320 and 192, respectively. For our Transformer-based models, patch size is 2, window size is 4, and channel number C is 48. Typically, our models have same hyper-parameters for different λ.

Evaluation. We evaluate our CNN-based models and STF models by calculating the average RD performance (PSNR and MS-SSIM) on the commonly used Kodak image set [25] and CLIC professional validation dataset [1].
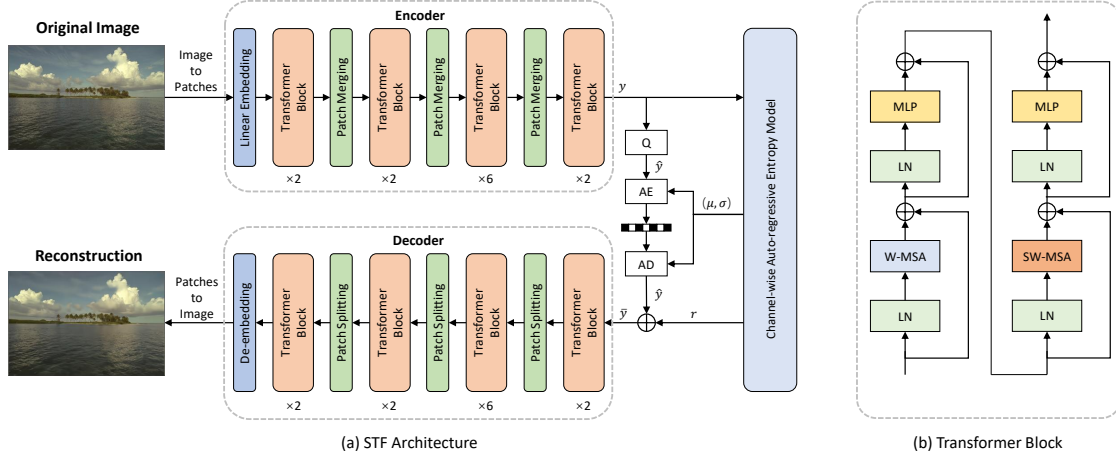
Figure 5. (a) The architecture of our proposed Transformer-based model. The patch merging layer and patch splitting layer consist of linear layers and LN layers. (b) W-MSA [29] is the window-based multi-head self-attention, and SW-MSA [29] is the shifted window-based self-attention, respectively.

We compare our methods with influential learned compression methods, including the context-free hyperprior model (Ballé2018) [6], auto-regressive hyper-prior model (Minnen2018) [34], and auto-regressive hyper-prior model with GMM and simplified attention (Cheng2020) [15]. Refer to appendix for RD curves covering a wide range of conventional and ANN-based compression methods.

4.2. Comparison with the SOTA Methods

RD Performance. Fig. 6 shows the comparison results on Kodak dataset. When trained for MSE and measured by PSNR, the performances of our CNN-based models and STF models are very close, and could outperform other learned compression methods. In contrast, when trained and measured by MS-SSIM, our CNN-based models and STF models only have slight improvements. As mentioned in [6], MS-SSIM has the effect of attenuating the error in regions with high contrast, and boosting the error in regions with low contrast. But the fact is not as expected, it frequently assigns more details to the regions with low contrast (e.g., grass and hair) and removes details from regions with high contrast (e.g., text and salient objects). Attention mechanism focuses more on high contrast regions, and consequently allocates more bits on them. The contradiction may result in an inconspicuous improvement for our attention-based models when optimized for MS-SSIM.

As illustrated in Fig. 7, the comparison results on the CLIC professional validation dataset indicate the same conclusion. It shows the robustness of our CNN-based models and STF models.

Visual Quality. Fig. 8 shows the example of reconstructed images (*kodim07.png*) by our methods and the compression standards JEPG, BPG, and VVC (VTM 9.1). Our reconstructed images retain more details with approxi-

Method	Enc (s)	Dec (s)	PSNR↑	bpp↓
Cheng2020 [15]	8.49	14.49	35.12	0.595
Minnen2018 [34]	16.19	21.16	35.09	0.640
Our CNN	**0.12**	**0.12**	**35.91**	**0.650**
Our STF	**0.15**	**0.15**	**35.82**	**0.651**

Table 1. Comparison of the averaged encoding and decoding time with Minnen2018 [34] and Cheng2020 [15] on Kodak dataset using GPUs (TITAN V). Note that the results of Cheng2020 are based on a light implementation (without attention module and Gaussian mixture likelihoods) in CompressAI [7].

mate bpp. When optimized for MS-SSIM, our CNN-based models and STF models yield significant improvements in visual qualities.

4.3. Codec Efficiency Analysis

Spatially Auto-regressive (AR) context model is effective and follow-up studies often use it to enhance the RD performance. However, the AR model sequentially encodes and decodes each spatial symbol, which significantly slows down the codec efficiency on GPUs and TPUs. GMM-based entropy model has the same defect. Although it can more precisely estimate PDFs and CDFs of latents, generating CDFs and PDFs dynamically would sacrifice the coding efficiency, while SGM-based entropy model has fixed PDFs and CDFs tables. Therefore, we use channel-conditional (CC) models [35] as the auto-regressive context model along the channel dimension for better parallel processing, and adopt the SGM-based entropy model for more efficient coding efficiency.

In Table 1, we evaluate the inference latency of our methods and those time-consuming models [15,34] on the Kodak
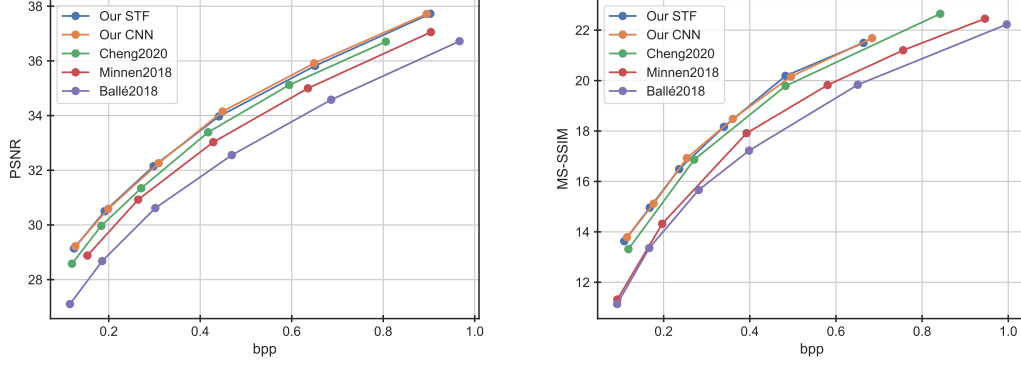
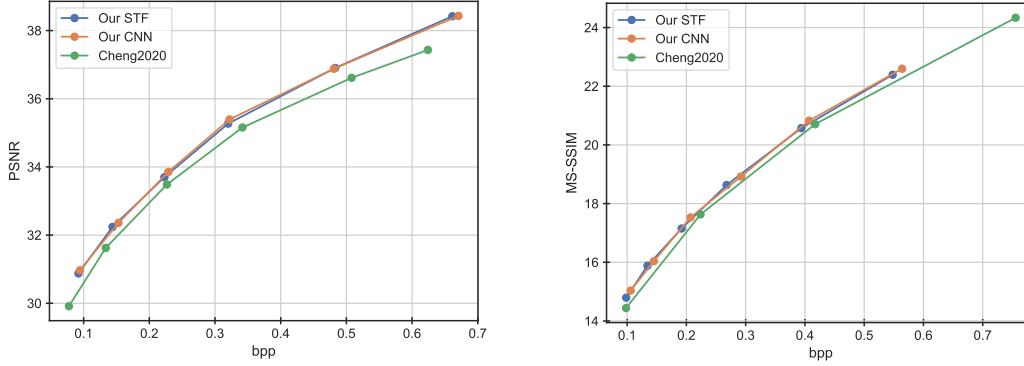Figure 6. RD Performance evaluation on the Kodak dataset, which contains 24 high quality images.

Figure 7. RD Performance evaluation on CLIC Professional Validation dataset, which contains 41 high resolution and high quality images.

Method	λ	PSNR ↑	bpp ↓
Baseline	*0.0035*	30.32	0.198
Baseline + NLAM	*0.0035*	30.42	0.202
Baseline + WAM	*0.0035*	**30.58**	**0.199**
Baseline	*0.0130*	33.87	0.454
Baseline + NLAM	*0.0130*	33.95	0.467
Baseline + WAM	*0.0130*	**34.10**	**0.448**

Table 2. **Contrast Experiment.** Baseline is Minnen2020 (SOTA CNN-based model) [35].

dataset.

4.4. Evaluating the Effects of Window Attention

To prove our conclusion that focusing on spatially neighboring elements can achieve better RD performance, we present contrast experiments by removing the WAMs or using NLAMs, as shown in Table.2. Ablation study in Fig. 9 shows that the proposed window-based attention is effective and could enhance current SOTA CNN-based model [35].

4.5. Discussion

More Efficient Window Attention Designing. Our CNN-based architecture still have room for improvement,

because the window-based attention is not powerful enough for capturing the structural information. Besides, our experimental results show that GDN and attention mechanism are incompatible. We guess that GDN is more of a nonlinear activation function than a normalization block. Our follow-up experiments show that directly using convolution layers to generate attention masks can also achieve comparable RD performance.

Compatible Normalization for Transformer. In our Transformer-based architecture, we use the layer normalization (LN) as default. The drawback is that LN rescales the responses of linear filters in the network so as to keep it in a reasonable operating range with an identical rescaling factor across all spatial locations, which may destroy the Gaussian distributions of elements. When computing the attention map, the LN is necessary to rescale the response range. It seems contradictory but implies that Transformer-based architecture has more potential in learned image compression. Moreover, we are looking forward to combining the Transformer and convolution blocks, our results so far show that they are complementary.

Human Perception. Learned image compression models directly optimize metrics such as PSNR or MS-SSIM and achieve high RD performance. However, some works [11,12,36–39] propose neither of PSNR and MS-SSIM cor-

Figure 8. Visualization of the reconstructed image (*kodim07.png*) from Kodak dataset. The metrics are [bpp↓/PNSR↑/MS-SSIM↑].
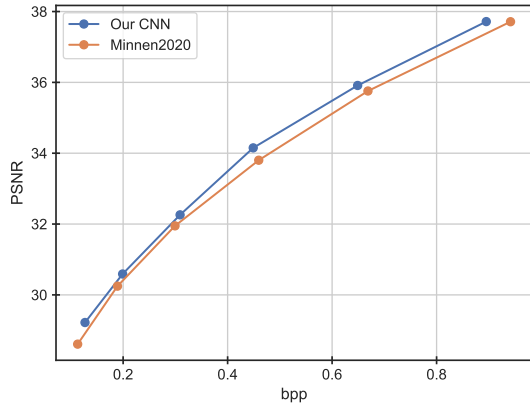

Figure 9. Ablation study of window-based attention. Our model is based on Minnen2020 with WAM. The RD points of Minnen2020 are obtained from [35], whose models were trained for 5M steps on the dataset consisting of 2M web images.

relate well with human perception. We found that models optimized for MSE would result in blurred image, and models optimized for MS-SSIM would remove details from regions with high contrast (e.g., text and salient objects), as shown in Fig.1. [39] proposed a new metric, which is learned on perceptual similarity data specific to image compression. Inspired by the mathematical definition of perceptual quality in [11], [12] studied the rate-distortion-perception trade-off. For a more impartial evaluation, perceptual quality measure are vital for practical applications(e.g.,LPIPS [47], FID [20], KID [10]).

5. Conclusion

In this paper, we have extensively studied the local-aware attention mechanism and have found that it is crucial to combine the global structure learned by neural networks and the local texture mined by the attention unit, we have presented a flexible window-based attention module to capture correlations among spatially neighboring elements, which could work as a plug-and-play component to enhance CNN or Transformer models. Furthermore, we have proposed a novel Symmetrical TransFormer (STF) framework with absolute transformer blocks in both down-sampling encoder and up-sampling decoder. Extensive experimental results show that the proposed methods are effective and exceed the state-of-the-art (SOTA) RD performance. In future, we will deeply explore other factors that affect the local detail reconstruction in image compression, such as the convolution kernel shaping and the normalization mode.

Acknowledgements

References

[1] Workshop and challenge on learned image compression, 2020. https://www.compression.cc/. 5

[2] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *CVPR*, 2019. 3

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. Density modeling of images using a generalized normalization transformation. In *ICLR*, 2016. 3

[5] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *ICLR*, 2017. 2

[6] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *ICLR*, 2018. 2, 3, 6

[7] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020. 5, 6

[8] Fabrice Bellard. Bpg image format, 2014. https://bellard.org/bpg/. 1

[9] Bross Benjamin, Chen Jianle, Liu Shan, and Wang Ye-Kui. Versatile video coding. In *JVET*, 2020. 1

[10] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018. 8

[11] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 7, 8

[12] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *ICML*, 2019. 7, 8

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 3

[14] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 3

[15] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *CVPR*, 2020. 2, 3, 6

[16] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In *ICCV*, 2019. 2

[17] Ze Cui, Jing Wang, Shangyin Gao, Tiansheng Guo, Yihui Feng, and Bo Bai. Asymmetric gained deep image compression with continuous rate adaptation. In *CVPR*, 2021. 2

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2, 3, 4

[19] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *CVPR*, 2021. 2

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 8

[21] Jianhua Hu, Ming Li, Changsheng Xia, and Yundong Zhang. Combine traditional compression method with convolutional neural networks. In *CVRP workshops*, 2018. 2

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 3

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[24] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[25] Eastman Kodak. Kodak lossless true color image suite (photocd pcd0992), 1993. http://r0k.us/graphics/kodak/. 5

[26] Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Dhyanesh Narayanan, Chen Sun, Gal Chechik, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from* https://github.com/openimages, 2016. 5

[27] Jooyoung Lee, Seunghyun Cho, and Munchurl Kim. An end-to-end joint learning scheme of image compression and quality enhancement with improved entropy minimization. *arXiv preprint arXiv:1912.12817*, 2019. 2

[28] Haojie Liu, Tong Chen, Peiyao Guo, Qiu Shen, Xun Cao, Yao Wang, and Zhan Ma. Non-local attention optimized deep image compression. *arXiv preprint arXiv:1904.09757*, 2019. 3, 4

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2, 3, 4, 5, 6

[30] Detlev Marpe, Heiko Schwarz, and Thomas Wiegand. Context-based adaptive binary arithmetic coding in the h. 264/avc video compression standard. *IEEE Transactions on circuits and systems for video technology*, 13(7):620–636, 2003. 2

[31] G Nigel N Martin. Range encoding: an algorithm for removing redundancy from a digitised message. In *Proc. Institution of Electronic and Radio Engineers International Conference on Video and Data Recording*, 1979. 2

[32] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *CVPR*, 2018. 2

[33] Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *arXiv preprint arXiv:2006.09965*, 2020. 3

[34] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *NeurIPS*, 2018. 2, 3, 6

[35] David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *ICIP*, 2020. 2, 3, 4, 5, 6, 7, 8

[36] Yash Patel, Srikar Appalaraju, and R Manmatha. Deep perceptual compression. *arXiv preprint arXiv:1907.08310*, 2019. 7

[37] Yash Patel, Srikar Appalaraju, and R Manmatha. Human perceptual evaluations for image compression. *arXiv preprint arXiv:1908.04187*, 2019. 7

[38] Yash Patel, Srikar Appalaraju, and R Manmatha. Hierarchical auto-regressive model for image compression incorporating object saliency and a deep perceptual loss. *arXiv preprint arXiv:2002.04988*, 2020. 7

[39] Yash Patel, Srikar Appalaraju, and R Manmatha. Saliency driven perceptual image compression. In *Proceedings of the IEEE/CVF WACV*, 2021. 7, 8

[40] Majid Rabbani. Jpeg2000: Image compression fundamentals, standards and practice. *Journal of Electronic Imaging*, 11(2):286, 2002. 1

[41] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001. 2

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3

[44] Gregory K Wallace. The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1), 1992. 1

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[46] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003. 2

[47] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8

[48] Lei Zhou, Chunlei Cai, Yue Gao, Sanbao Su, and Junmin Wu. Variational autoencoder for low bit-rate image compression. In *CVRP workshops*, 2018. 2

[49] Lei Zhou, Zhenhong Sun, Xiangji Wu, and Junmin Wu. End-to-end optimized image compression with attention mechanism. In *CVPR workshops*, 2019. 3