Learning-Based Conditional Image Compression

Tianma Shen[1], Wen-Hsiao Peng[2], Huang-Chia Shih[3], and Ying Liu[1,†]
[1]Department of Computer Science and Engineering, Santa Clara University, USA
[2]Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan
[3]Department of Electrical Engineering, Yuan Ze University, Taiwan
Emails: tshen2@scu.edu, wpeng@cs.nctu.edu.tw, hcshih@saturn.yzu.edu.tw, yliu15@scu.edu

***Abstract*—In recent years, deep learning-based image compression has achieved significant success. Most schemes adopt an end-to-end trained compression network with a specifically designed entropy model. Inspired by recent advances in conditional video coding, in this work, we propose a novel transformer-based conditional coding paradigm for learned image compression. Our approach first compresses a low-resolution version of the target image and up-scales the decoded image using an off-the-shelf super-resolution model. The super-resolved image then serves as the condition to compress and decompress the target high-resolution image. Experiments demonstrate the superior rate-distortion performance of our approach compared to existing methods.**

***Index Terms*—conditional coding, deep learning, entropy model, hyperprior, image compression, super resolution, vision transformer**

I. Introduction

Learning-based image compression [1]–[8] has demonstrated higher coding efficiency than traditional compression algorithms [9]–[11]. Existing learned image compression schemes usually adopt an end-to-end trained compression network with a specifically designed entropy model, utilizing either hyperpriors [1], [2], or both hyperpriors and context models [3]–[8]. Most recently, conditional coding has emerged as a new paradigm for learning-based video coding [12]–[18]. While traditional video codecs and prior learned video compression models leverage inter-frame correlations by compressing the residue between the target frame $\mathbf{I}_t$ and the motion-compensated reference frame $\mathbf{I}_c$, conditional video coding directly compresses $\mathbf{I}_t$ under the condition of $\mathbf{I}_c$. According to (1),

$$H\left(\mathbf{I}_t - \mathbf{I}_c\right) \geq H\left(\mathbf{I}_t - \mathbf{I}_c \mid \mathbf{I}_c\right) = H\left(\mathbf{I}_t \mid \mathbf{I}_c\right) \tag{1}$$

the residual entropy $H\left(\mathbf{I}_t - \mathbf{I}_c\right)$ is no less than the conditional entropy $H\left(\mathbf{I}_t \mid \mathbf{I}_c\right)$. Therefore, conditional coding potentially can save bit rates compared to residue coding. In this work, we will extend conditional coding to the field of image compression to improve the rate-distortion performance of a learning-based image codec. Our main contributions are summarized as follows:

- We propose a learning-based conditional image compression model which adopts super-resolved images as the conditional information.
- We propose a new multi-scale cross-attention transformer structure in the conditional coding architecture, such that the target images can be compressed with multi-scale conditional information extracted from super-resolved images.
- Experimental studies on popular image compression datasets demonstrate that our proposed method offers state-of-the-art performance visually and quantitatively.

II. Related Work

End-to-end learned image compression has gained significant attention to achieve improved compression performance [1]–[8]. In particular, hyperpriors were introduced [1], [2] to estimate the distribution of latent representation. Context models [3], [5], [6], [8] were developed to exploit correlations between current coding data and previously decoded data for more effective entropy coding. For example, channel context model was used in [3], [5] and spatial context model was used in [6]. To utilize both spatial and channel correlations, the spatial-channel auto-regressive context model (SC-AR CM) [8] splits feature maps into spatial-channel chunks, which are entropy encoded and decoded sequentially in a channel-first order, followed by a 2D zigzag spatial order.

Transformer structures were also investigated in learned image compression and demonstrated exceptional rate-distortion performance [4]–[8], surpassing convolutional neural network (CNN)-based image compression, because the self-attention mechanism of transformers effectively captures global dependencies within images. For instance, Entroformer [6] utilizes transformers in the hyper encoder, hyper decoder, and context model. STF [5] adopts the Swin transformer [19] in the main encoder and decoder. Based on STF, SC-AR CM [8] proposed transformer structures in the main encoder/decoder, in the spatial-channel context model, and in the latent residual prediction module, which further improved the rate-distortion performance.

Recently, conditional coding has emerged as a promising learning-based video coding framework [12]–[18]. Instead of coding the residue between the target frame and the motion-compensated reference frame, it directly encodes the target frame using the reference frame as a condition. For example,

†Corresponding author.
This work is supported in part by the National Science Foundation under Grant ECCS-2138635 and the NVIDIA Academic Hardware Grant.

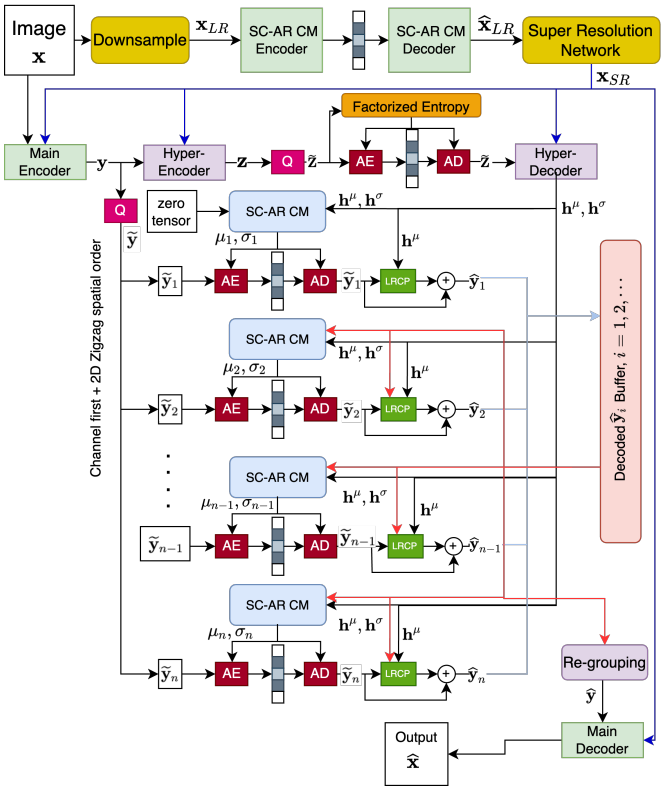

Fig. 1. The overall architecture of the proposed model. AE and AD are the arithmetic encoder and arithmetic decoder. Q represents quantization. Super Resolution Network is the large SwinIR [21] model pre-trained on the DIV2K dataset [22].

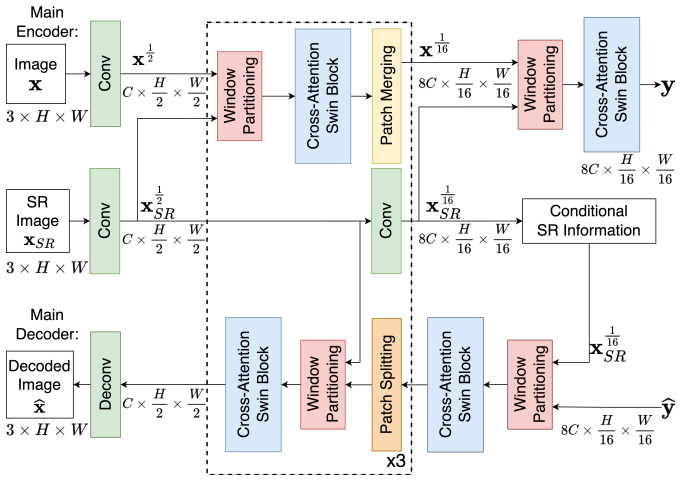

Fig. 2. The architecture of the main encoder and main decoder. The dimension on each arrow represents the output tensor dimension.

DCVC [12] uses feature-domain context as the condition, and DCVC-TCM [14] learns multi-scale temporal contexts as the condition. Besides, conditional P-frame coding [16] and conditional B-frame coding [17], [18] were developed using augmented normalizing flows [20]. Such conditional coding paradigms effectively improve the video coding efficiency.

III. The Proposed Method

A. The Overall Architecture

Fig. 1 shows the overall architecture of our proposed conditional learned image codec. Instead of directly compressing the target image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, our conditional coding framework has two steps. The first step adopts the method in [8] to compress a low-resolution image $\mathbf{x}_{LR} \in \mathbb{R}^{3\times \frac{H}{4}\times \frac{W}{4}}$, which is $4\times$ down-sampled from $\mathbf{x}$, followed by a super-resolution network which upscales the decoded low-resolution image $\hat{\mathbf{x}}_{LR}$ to generate the super-resolved image $\mathbf{x}_{SR} \in \mathbb{R}^{3\times H\times W}$.

The second step adopts a conditional coder to compress $\mathbf{x}$. It consists of a pair of main encoder and decoder, and a pair of hyper encoder and decoder, both of which take $\mathbf{x}_{SR}$ as the conditional information. The main encoder compresses $\mathbf{x}$ into a latent representation $\mathbf{y}$, which is then quantized as $\widetilde{\mathbf{y}}$. Then, $\widetilde{\mathbf{y}}$ is split into spatial-channel chunks $\widetilde{\mathbf{y}}_1$, $\widetilde{\mathbf{y}}_2$, $\cdots$, which are coded in a sequential manner using the spatial-channel auto-regressive context model (SC-AR CM) [8]. On the decoder side, the hyper decoder integrates the same conditional information $\mathbf{x}_{SR}$ to generate $\mathbf{h}^{\mu}$ and $\mathbf{h}^{\sigma}$, which serve as the input of the SC-AR CM to estimate the distribution parameters μ_i, σ_i, $i = 1, \cdots, n$ to assist the entropy coding of the spatial-channel chuncks $\widetilde{\mathbf{y}}_i$, $i = 1, \cdots, n$. Utilizing the latent residual cross-attention prediction (LRCP) module [8], our scheme effectively reduces the quantization error to get decoded latent representations $\hat{\mathbf{y}}_i$, $i = 1, \cdots, n$, which are re-grouped into $\hat{\mathbf{y}}$. Finally, the main decoder decompresses $\hat{\mathbf{y}}$ with conditional information $\mathbf{x}_{SR}$ to generate the decoded image $\hat{\mathbf{x}}$.

B. The Main Encoder and Decoder

Fig. 2 shows the main encoder and decoder with the proposed cross-attention transformer-based conditional coding architecture. Convolution blocks (Conv) are used to extract conditional information at different scales from $\mathbf{x}_{SR}$, and these multi-scale conditional features are fed into different stages of the main encoder and decoder. The input image $\mathbf{x}$ and the super-resolved image $\mathbf{x}_{SR}$ are each processed by a convolutional layer to spatially downsample the image and perform channel-wise linear projection. The resultant feature maps of dimension $C \times \frac{H}{2} \times \frac{W}{2}$ are then partitioned into non-overlapping windows and serve as the input of the subsequent cross-attention-based shifted-window transformer block (Cross-Attention Swin Block).

Fig. 3 shows the details of the Cross-Attention Swin Block. The feature map $\mathbf{x}^c$ extracted from the original image $\mathbf{x}$ provides the source of key and value, while the conditional feature map $\mathbf{x}^c_{SR}$ extracted from the super-resolved image $\mathbf{x}_{SR}$ provides the source of query. The query guides the compression of $\mathbf{x}$ by specifying what information in $\mathbf{x}$ should be retained. By incorporating both the window multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA) within the Cross-Attention Swin Block, the model can effectively reduce computational complexity while establishing correlations among non-overlapping windows.

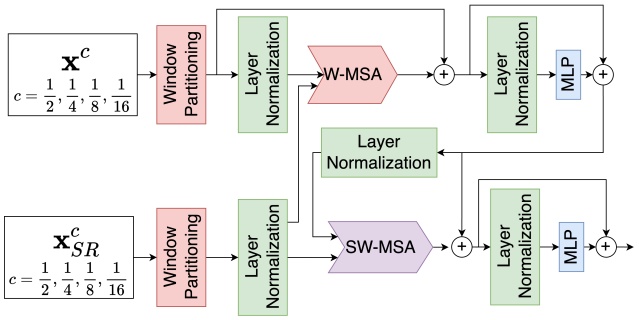

Fig. 3. The architecture of the Cross-Attention Swin Block.

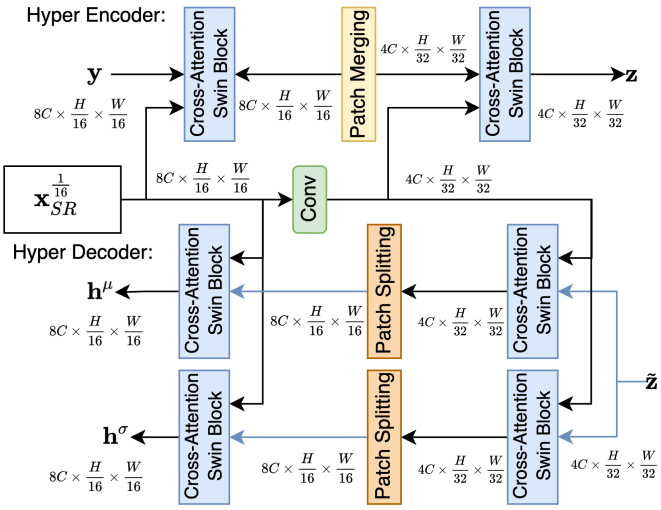

Fig. 4. The architecture of the hyper encoder and hyper decoder.

C. The Hyper Encoder and Decoder

Fig. 4 shows the architecture of the hyper encoder and decoder, which also adopt the Cross-Attention Swin Block to further process the latent representation $\mathbf{y}$ with multi-scale conditional information extracted from $\mathbf{x}_{SR}$. The hyper decoder has two branches, one generating $\mathbf{h}^\mu$, another generating $\mathbf{h}^\sigma$. As shown in Fig. 1, $\mathbf{h}^\mu$ and $\mathbf{h}^\sigma$ are used in the SC-AR CM to estimate distribution parameters μ_i, σ_i, $i = 1, \cdots, n$ for spatial-channel chuncks $\widetilde{\mathbf{y}}_i$, $i = 1, \cdots, n$. Besides, $\mathbf{h}^\mu$ also serves as an input of the LRCP modules.

D. Training Loss

We trained two models: Ours(PSNR) and Ours(MS-SSIM) for the proposed scheme. Ours(PSNR) uses the mean squared error (MSE) as the distortion loss as shown in (2), such that the trained model protects pixel-level fidelity. In contrast, Ours(MS-SSIM) adopts the multi-scale structural similarity (MS-SSIM) in the distortion loss as shown in (3), such that the trained model protects more structural similarity and perceptual quality. Besides, $\mathcal{R}_{\widetilde{\mathbf{y}}}$ and $\mathcal{R}_{\widetilde{\mathbf{z}}}$ represent the bit rates measured by the entropy of $\widetilde{\mathbf{y}}$ and $\widetilde{\mathbf{z}}$, respectively.

$$\mathcal{L}_{\text{PSNR}} = \mathcal{R}_{\widetilde{\mathbf{y}}} + \mathcal{R}_{\widetilde{\mathbf{z}}} + \lambda \times \text{MSE}(\mathbf{x}, \widehat{\mathbf{x}}), \tag{2}$$

$$\mathcal{L}_{\text{MS-SSIM}} = \mathcal{R}_{\widetilde{\mathbf{y}}} + \mathcal{R}_{\widetilde{\mathbf{z}}} + \lambda \times \big(1 - \text{MS-SSIM}(\mathbf{x}, \widehat{\mathbf{x}})\big). \tag{3}$$

IV. Experimental Studies

A. Datasets

The proposed model was trained using the OpenImages dataset [23], which is known for its diverse distribution of images. From the original training set, we randomly selected 300,000 images to form our training set. To evaluate the performance of the proposed model, as well as existing methods, we employed three benchmark test datasets: Kodak [24], Tecnick [25], and CLIC [26]. These datasets encompass images of different resolutions.

B. Comparison with Other Methods

In our experimental studies, we compare the performance of the proposed model with several state-of-the-art learning-based image compression methods: SC-AR CM [8], STF [5], Entroformer [6], and Coarse2Fine [2]. They represent the latest advances in image compression techniques. We also include traditional image compression methods, namely BPG [10] and VVC Intra (VTM 19.2) [11], for comparison. They serve as baselines that have been widely used in practical applications.

We quantitatively evaluate the rate-distortion (RD) performance of the proposed model and existing methods. The bit rates are measured in bits per pixel (BPP), and the bit rates of our proposed model include the bits to encode both $\mathbf{x}$ and $\mathbf{x}_{LR}$. The distortion between the decoded and ground-truth images is measured using the peak signal-to-noise ratio (PSNR) and the MS-SSIM.

Fig. 5 presents the PSNR and MS-SSIM curves for different bit rates on the Kodak, Tecnick, and CLIC datasets, respectively. The results demonstrate that Ours(PSNR) consistently achieves the highest PSNR values across all datasets, indicating its superior performance in terms of pixel fidelity. Besides, Ours(MS-SSIM) achieves the highest MS-SSIM values for all datasets and all bit rates, showing its superiority in preserving structural information, especially at low bit rates. Moreover, Ours(PSNR) also offers higher MS-SSIM values than existing learning-based methods SC-AR CM, STF, Entroformer, and Coarse2Fine, and traditional image codecs VVC Intra and BPG.

Fig. 6 shows the visual results. We present a sample decoded image from each test dataset. We show an enlarged area of each sample image, allowing for a close-up view of the decoding quality. By comparing the proposed model with SC-AR CM, STF, Coarse2Fine, VTM, and BPG, it is evident that the proposed model excels in recovering fine details and textures in the decoded images. Notably, the proposed model demonstrates superior performance in preserving the edges of the window in the Kodak image, capturing the line of the tiles in the Tecnick image, and accurately representing the flower of the man's tattoo in the CLIC image. Below each enlarged decoded image patch are the corresponding BPP,

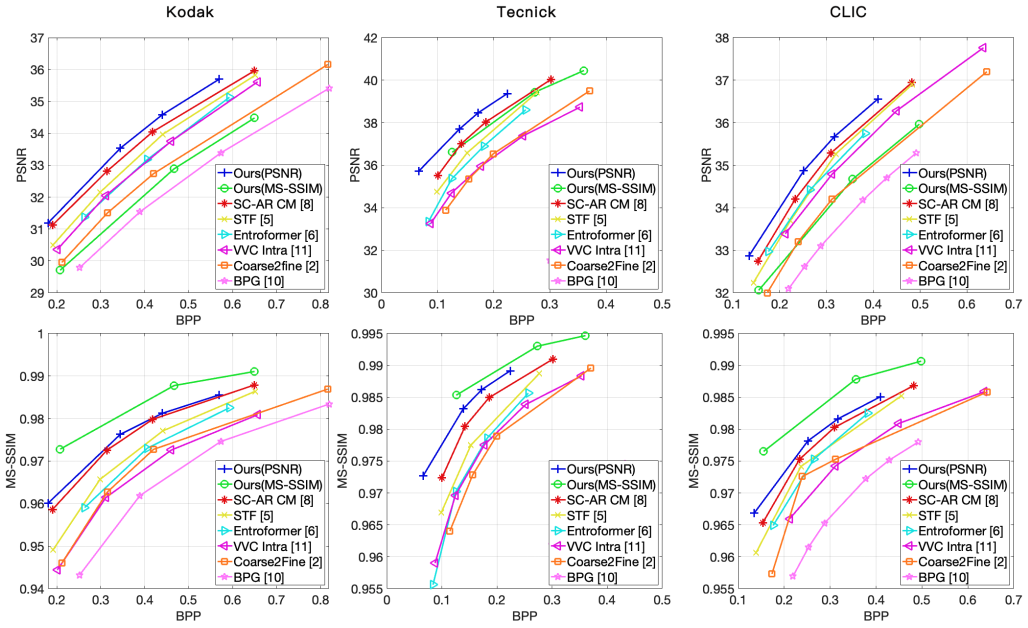

Fig. 5. The rate-distortion curves of our proposed models and existing methods on the Kodak, Tecnick, and CLIC datasets.

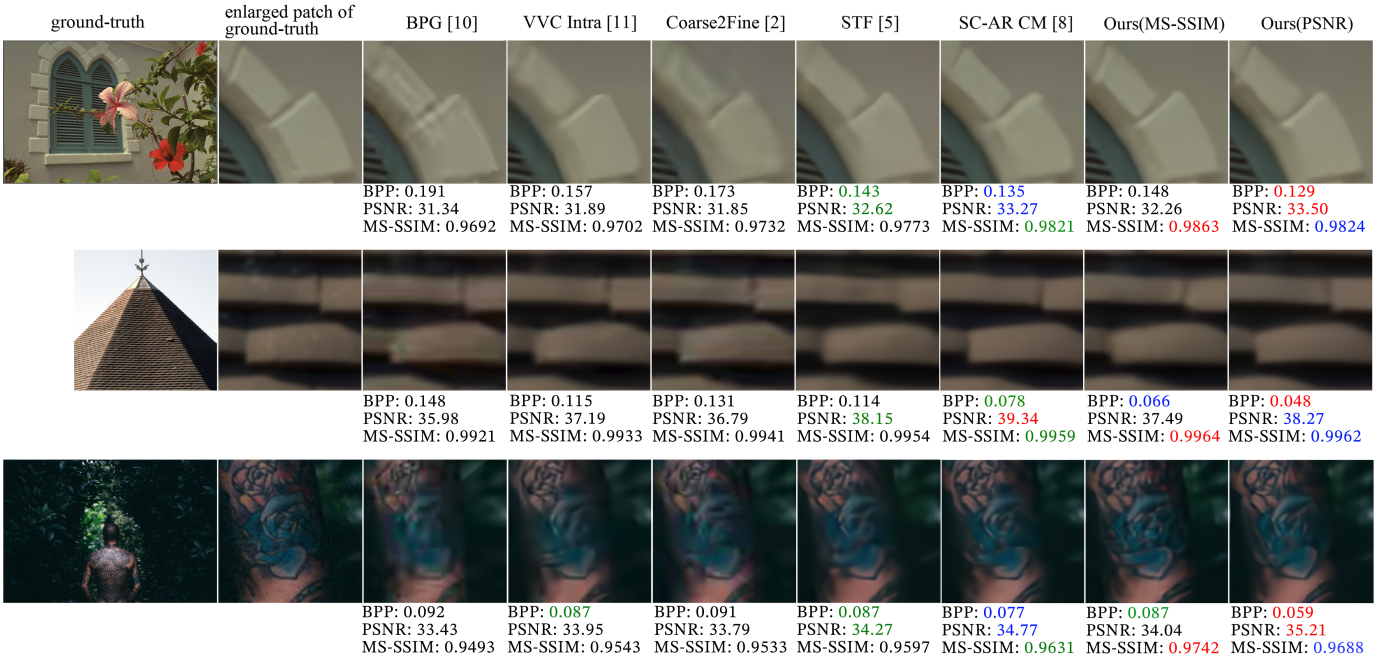

Fig. 6. Sample images decoded by our proposed models and existing methods on the Kodak (1st row), Tecnick (2nd row), and CLIC (3rd row) datasets. The best, 2nd best, and 3rd best BPP/PSNR/MS-SSIM values are shown in red, blue, and green, respectively.

PSNR (dB), and MS-SSIM values. We observe that for all three sample images, Ours(PSNR) achieves the lowest BPP, the highest or 2nd highest PSNR, and the 2nd highest MS-SSIM. Ours(MS-SSIM) achieves the highest MS-SSIM values with competitively low bit rates.

V. Conclusions

The proposed learning-based conditional image compression model introduces a novel transformer-based approach for learned image compression. By incorporating super-resolution images as the conditional information, the model achieves reduced bit rates while maintaining high decoding quality. In terms of future studies, we will conduct complexity analysis and investigate conditional coding frameworks for machine task-oriented visual coding.

References

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

Representations, Toulon, France, April 2017, pp. 1199–1108.

[2] Y. Hu, W. Yang, and J. Liu, "Coarse-to-fine hyper-prior modeling for learned image compression," in *Proc. The Thirty-Fourth AAAI Conference on Artificial Intelligence*, Mar. 2020, pp. 11 013–11 020.

[3] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *Proc. IEEE International Conference on Image Processing*, Abu Dhabi, United Arab Emirates, Nov. 2020, pp. 3339–3343.

[4] M. Lu, P. Guo, H. Shi, C. Cao, and Z. Ma, "Transformer-based image compression," in *Proc. Data Compression Conference*, Snowbird, UT, USA, Jul. 2022, p. 469.

[5] R. Zou, C. Song, and Z. Zhang, "The devil is in the details: Window-based attention for image compression," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, Oct. 2022, pp. 17 471–17 480.

[6] Y. Qian, X. Sun, M. Lin, Z. Tan, and R. Jin, "Entroformer: A transformer-based entropy model for learned image compression," in *Proc. The Tenth International Conference on Learning Representations*, Virtual Event, Aug. 2022, pp. 1169–1181.

[7] A. B. Koyuncu, H. Gao, A. Boev, G. Gaikov, E. Alshina, and E. Steinbach, "Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression," in *Proc. Computer Vision–ECCV*, Tel Aviv, Israel, Nov. 2022, pp. 447–463.

[8] T. Shen and Y. Liu, "Learned image compression with transformers," in *Proc. Big Data V: Learning, Analytics, and Applications*, vol. 12522. Florida, US: SPIE, 2023, pp. 10–20.

[9] G. K. Wallace, "The jpeg still picture compression standard," *IEEE Trans. Consumer Electronics*, vol. 38, no. 01, pp. xviii – xxxiv, Feb. 1992.

[10] F. Bellard, "BPG image format," http://bellard.org/bpg/ (Accessed: 2022-1-18).

[11] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, Aug. 2021.

[12] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," in *Proc. Advances in Neural Information Processing Systems*, virtual, December 6-14, 2021, pp. 18 114–18 125.

[13] J. Li, B. Li, and L. Yan, "Hybrid spatial-temporal entropy modelling for neural video compression," in *Proc. The 30th ACM International Conference on Multimedia*, Lisboa, Portugal, October 10 - 14,2022, pp. 1503–1511.

[14] X. Sheng, J. Li, B. Li, L. Li, D. Liu, and Y. Lu, "Temporal context mining for learned video compression," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–12, 2021.

[15] J. Li, B. Li, and Y. Lu, "Neural video compression with diverse contexts," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, June 18-22, 2023, pp. 22 616–22 626.

[16] Y.-H. Ho, C.-P. Chang, P.-Y. Chen, A. Gnutti, and W.-H. Peng, "Canf-vc: Conditional augmented normalizing flows for video compression," in *European Conference on Computer Vision*. Springer, 2022, pp. 207–223.

[17] M.-J. Chen, Y.-H. Chen, and W.-H. Peng, "B-canf: Adaptive b-frame coding with conditional augmented normalizing flows," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 22–27, 2023.

[18] D. Alexandre, H.-M. Hang, and W.-H. Peng, "Hierarchical b-frame video coding using two-layer canf without motion coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, June 2023, pp. 10 249–10 258.

[19] 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 *Proc. 2021 IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, May 2021, pp. 9992–10 002.

[20] Y.-H. Ho, C.-C. Chan, W.-H. Peng, H.-M. Hang, and M. Domański, "Anfic: Image compression using augmented normalizing flows," *IEEE Open Journal of Circuits and Systems*, vol. 2, pp. 613–626, 2021.

[21] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proc. IEEE/CVF International Conference on Computer Vision Workshops*, Montreal, BC, Canada, October 11, 2021, pp. 1833–1844.

[22] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proc. the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 126–135.

[23] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, and A. Kolesnikov, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, Jul. 2020.

[24] E. Kodak, "Kodak lossless true color image suite (PhotoCD PCD0992)," http://r0k.us/graphics/kodak/. (Accessed: 2022-1-18).

[25] N. Asuni and A. Giachetti, "Testimages: A large data archive for display and algorithm testing," *Journal of Graphics Tools*, vol. 17, no. 4, pp. 113–125, Oct. 2013.

[26] "Workshop and challenge on learned image compression," https://www.compression.cc/ (Accessed: 2022-1-18).