MLIC^{++}: Linear Complexity Multi-Reference Entropy Modeling for Learned Image Compression

Wei Jiang [1] **Ronggang Wang** [1 2]

Abstract

Recently, multi-reference entropy model has been proposed, which captures channel-wise, local spatial, and global spatial correlations. Previous works adopt attention for global correlation capturing, however, the quadratic cpmplexity limits the potential of high-resolution image coding. In this paper, we propose the linear complexity global correlations capturing, via the decomposition of softmax operation. Based on it, we propose the MLIC^{++}, a learned image compression with linear complexity for multi-reference entropy modeling. Our MLIC^{++} is more efficient and it reduces BD-rate by 12.44% on the Kodak dataset compared to VTM-17.0 when measured in PSNR.

1. Introduction

In recent years, we have seen much progress in entropy model design (Minnen et al., 2018; Minnen & Singh, 2020; He et al., 2022; Jiang et al., 2022). Most entropy models capture correlations in one dimension, however, there are channel-wise, local spatial, and global spatial correlations, leading to sub-optimal performance. To overcome this limitation, Jiang *et al.* introduce the multi-reference entropy model (MEM) (Jiang et al., 2022) and propose the learned image compression models MLIC and MLIC^{+}. In MLIC and MLIC^{+}, the latent representation $\hat{\boldsymbol{y}}$ is divided into slices $\{\hat{\boldsymbol{y}}^0, \hat{\boldsymbol{y}}^1, \cdots\}$ (Minnen & Singh, 2020) for channel-wise correlation capturing. For the i-th slice, they propose the checkerboard attention for local spatial correlations capturing with two-pass decoding, where the slice is divided into anchor part $\hat{\boldsymbol{y}}^i_{ac}$ and non-anchor part $\hat{\boldsymbol{y}}^i_{na}$. In addition, they propose to use the attention map of the previous slice to predict the global correlations in the current slice. The process is $\mathrm{softmax}(\hat{\boldsymbol{y}}^{i-1}_{na}(\hat{\boldsymbol{y}}^{i-1}_{ac})^\top)\hat{\boldsymbol{y}}^i_{ac}$. The main drawback of MLIC and MLIC^{+} is the quadratic complexity of global correlations capturing, caused by the softmax operation in attention, which specifies the order of matrix calculation. In this paper, we investigate the global correlations capturing with linear complexity. We decompose the softmax operation (Shen et al., 2021) into two softmax operations $\mathrm{softmax}_2(\hat{\boldsymbol{y}}^{i-1}_{na})\mathrm{softmax}_1(\hat{\boldsymbol{y}}^{i-1}_{ac})^\top\hat{\boldsymbol{y}}^i_{ac}$. Since $\mathrm{softmax}_2(\hat{\boldsymbol{y}}^{i-1}_{na})\mathrm{softmax}_1(\hat{\boldsymbol{y}}^{i-1}_{ac})^\top \geq 0$, it can be treated as the global similarity. Such decomposition enables global correlations capturing with linear complexity. Based on such decomposition, we propose linear complexity intra-slice and inter-slice global spatial context modules, and propose learned image compression model MLIC^{++}, denoting linear complexity multi-reference entropy modeling for learned image compression.

2. Related Works

Learned image compression (Theis et al., 2017; Ballé et al., 2020) aims to optimize the trade-off between bit-rate and distortion. Given a specific Lagrange multiplier λ, analysis transform g_a, and synthesis transform g_s, the optimization target is

$$\mathcal{L} = \mathcal{R}(\hat{\boldsymbol{y}}) + \lambda \times \mathcal{D}(\boldsymbol{x}, \hat{\boldsymbol{x}}), \tag{1}$$

where the $\boldsymbol{x}$ is the input image, $\boldsymbol{y}$ is the latent representation, $\hat{\boldsymbol{y}} = \lceil g_a(\boldsymbol{x}) \rfloor$,$\lceil \cdot \rfloor$ is quantization, $\hat{\boldsymbol{x}} = g_s(\hat{\boldsymbol{y}})$.

Ballé *et al.* (Ballé et al., 2017) propose to use convolutional layers to build g_a and g_s for non-linear transform. They adopt adding uniform noise $\mathcal{U}(-0.5, 0.5)$ to approximate quantization during training. Later, Ballé *et al.* (Ballé et al., 2018) introduce hyperprior $\hat{\boldsymbol{z}}$ to estimate entropy parameters. To further capture correlations within $\hat{\boldsymbol{y}}$, Minnen *et al.* (Minnen et al., 2018) adopt a serial pixel-cnn-like (Van den Oord et al., 2016) context model. To accelerate decoding, Minnen *et al.* (Minnen & Singh, 2020) propose to capture channel-wise contexts by group the symbol channels into several chunks, while He *et al.* adopt the checkerboard pattern (He et al., 2021) for two-pass decoding.

Recently, multi-reference entropy modeling has been explored. Some works (Ma et al., 2021; He et al., 2022) aim to capture local spatial and channel-wise contexts, however, they ignore the global spatial correlations. Jiang *et*

[1]Shenzhen Graduate School, Peking University [2]Pengcheng Laboratory. Correspondence to: Wei Jiang <wei.jiang1999@outlook.com>, Ronggang Wang <rgwang@pkusz.edu.cn>.

Published as a workshop paper at ICML 2023 neural compression workshop.

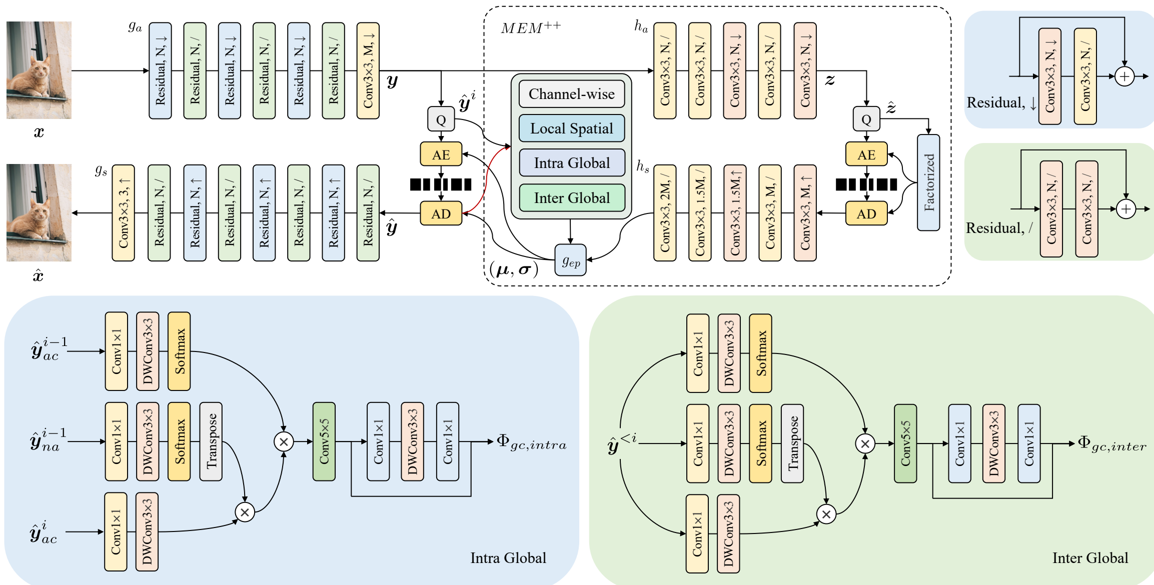

Figure 1. The overall architecture of MLIC and MLIC$^+$. ↓ means down-sampling. ↑ means up-sampling. / means stride equals 1. Red line is the dataflow during decoding. $\boldsymbol{x}$ is the input image and $\hat{\boldsymbol{x}}$ is the reconstructed image. $\boldsymbol{y}$ is the latent representation and $\hat{\boldsymbol{y}}$ is the quantized latent representation. $\hat{\boldsymbol{y}}^i$ is the i-th slice of $\hat{\boldsymbol{y}}$. $\hat{\boldsymbol{y}}^i_{ac}$ is the anchor part of $\hat{\boldsymbol{y}}^i$. $\hat{\boldsymbol{y}}^i_{na}$ is the non-anchor part of $\hat{\boldsymbol{y}}^i$. We use "Channel-wise" to denote channel-wise context module, "Local Spatial" to denote local spatial context module, "Intra Global" to denote intra-slice global context module, "Inter Global" to denote inter-slice global context module. We set M to 320 and set N to 192 in our MLIC^{++}.

al. propose MLIC and MLIC$^+$ (Jiang et al., 2022) to capture local spatial, global spatial, and channel-wise contexts, however, the quadratic computational complexity of global spatial contexts capturing makes it hard to be employed for high-resolution image coding.

3. Method

3.1. Overall Architecture

The overall framework of our MLIC^{++} is similar with MLIC$^+$ (Jiang et al., 2022). We briefly introduce our framework first. In our framework, we adopt the transform modules of MLIC$^+$, a simplified version of the transform modules of Cheng'20 (Cheng et al., 2020). We adopt the same transform modules to prove the effectiveness of our proposed linear complexity multi-reference entropy model MEM^{++}. In our MEM^{++}, we adopt the checkerboard attention (Jiang et al., 2022) for local spatial context capturing and use the same settings of MLIC$^+$ (Jiang et al., 2022) for channel-wise context capturing . we use Φ_h to denote the hyperprior. We first divide the latent representation $\hat{\boldsymbol{y}}$ into slices $\{\hat{\boldsymbol{y}}^0, \hat{\boldsymbol{y}}^1, \cdots\}$ (Minnen & Singh, 2020). We take the i-th slice as an example. We divide the $\hat{\boldsymbol{y}}^i$ into anchor part $\hat{\boldsymbol{y}}^i_{ac}$ and non-anchor part $\hat{\boldsymbol{y}}^i_{na}$. We extract channel-wise contexts Φ^i_{ch} from slices $\hat{\boldsymbol{y}}^{<i}$. $\hat{\boldsymbol{y}}^i_{ac}$ is local-spatial-context-free. We extract local spatial context Φ^i_{lc} of $\hat{\boldsymbol{y}}^i_{na}$ from $\hat{\boldsymbol{y}}^i_{ac}$. Slice different slices share the similar global similarity (Jiang et al., 2022; Guo et al., 2021), We extract the intra-slice global context $\Phi^i_{gc,intra}$ of $\hat{\boldsymbol{y}}^i_{na}$ from $\hat{\boldsymbol{y}}^i_{ac}$ via the global similarity of $\hat{\boldsymbol{y}}^{i-1}$. Besides, we extract the inter-slice global context $\Phi^i_{gc,inter}$ from slices $\hat{\boldsymbol{y}}^{<i}$ via the global similarity of slices $\hat{\boldsymbol{y}}^{<i}$. Therefore, the rate of $\hat{\boldsymbol{y}}^i_{ac}$ and the rate of $\hat{\boldsymbol{y}}^i_{na}$ are:

$$\begin{aligned}\mathcal{R}^i_{ac} &= \mathbb{E}\left[-\log_2 p_{\hat{\boldsymbol{y}}^i_{ac}}\left(\hat{\boldsymbol{y}}^i_{ac}|\Phi_h, \Phi^i_{ch}, \Phi^i_{gc,inter}\right)\right] \\ \mathcal{R}^i_{na} &= \mathbb{E}\left[-\log_2 p_{\hat{\boldsymbol{y}}^i_{na}}\left(\hat{\boldsymbol{y}}^i_{na}|\Phi_h, \Phi^i_{ch}, \Phi^i_{gc,intra}, \Phi^i_{gc,inter}\right)\right]\end{aligned} \tag{2}$$

3.2. Explicit Global Context in MLIC$^+$

We take the intra-slice global context module in MLIC$^+$ (Jiang et al., 2022) as an example. To extract the intra-slice global context of $\hat{\boldsymbol{y}}^i_{na} \in \mathbb{R}^{L\times C}$, global similarity between $\hat{\boldsymbol{y}}^{i-1}_{na} \in \mathbb{R}^{L\times C}$ and $\hat{\boldsymbol{y}}^{i-1}_{ac} \in \mathbb{R}^{L\times C}$ is employed to predict the global correlations between $\hat{\boldsymbol{y}}^i_{na} \in \mathbb{R}^{L\times C}$ and $\hat{\boldsymbol{y}}^i_{ac} \in \mathbb{R}^{L\times C}$, where $L = H \times W$. C is the channel number of $\hat{\boldsymbol{y}}^i$, H is the height of $\hat{\boldsymbol{y}}^i$, and W is the width of $\hat{\boldsymbol{y}}^i$.

$$\text{Attention} = \underbrace{\text{softmax}\left(\frac{\hat{\boldsymbol{y}}^{i-1}_{na}\left(\hat{\boldsymbol{y}}^{i-1}_{ac}\right)^\top}{\sqrt{d_k}}\right)}_{\text{non-negative}}\hat{\boldsymbol{y}}^i_{ac} \tag{3}$$

In MLIC$^+$(Jiang et al., 2022), the attention map is employed as a similarity metric, because of the non-negative property of the attention map. The process is shown in Equation 3. However, attention map leads to **quadratic** complexity. The

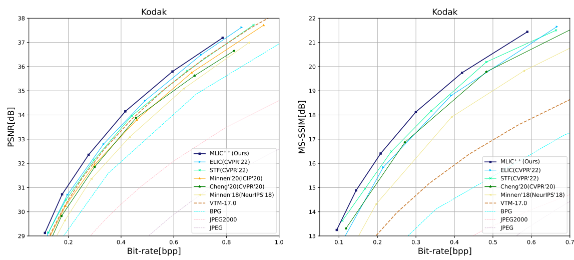

Figure 2. PSNR-Bit-rate curve (opt.MSE) and MS-SSIM-Bit-rate curve (opt.MS-SSIM) on Kodak dataset.

Methods	Kodak PSNR	Kodak MS-SSIM
VTM-17.0 (VTM, 2022)	0.00	0.00
SwinT-Charm (ICLR'22) (Zhu et al., 2022)	-1.73	–
STF (CVPR'22) (Zou et al., 2022)	-2.48	-47.72
ELIC (CVPR'22) (He et al., 2022)	-5.95	-44.60
Contextformer (ECCV'22) (Koyuncu et al., 2022)	-5.77	-46.12
MLIC (Arxiv'22) (Jiang et al., 2022)	-8.05	-49.13
MLIC$^+$ (Arxiv'22) (Jiang et al., 2022)	-11.39	-52.75
MLIC^{++} (Ours)	$\mathbf{-12.44}$	$\mathbf{-53.22}$

Table 1. BD-Rate $(\%)$ comparison for PSNR (dB) and MS-SSIM (dB). The anchor is VTM-17.0 Intra.

complexity of Equation 3 is $O(CL^2)$. The quadratic complexity makes it quite difficult to apply MLIC$^+$ to high-resolution images.

3.3. Linear Attention for Implicit Global Context

In Equation 3, the quadratic complexity is caused by the softmax operation, which leads to computing $\hat{\boldsymbol{y}}_{na}^{i-1}\hat{\boldsymbol{y}}_{ac}^{i-1}$ first. To solve the quadratic complexity, we employ efficient attention (Shen et al., 2021). In efficient attention (Shen et al., 2021), we employ the softmax operation on $\hat{\boldsymbol{y}}_{na}^{i-1}$ in row and the softmax operation on $\hat{y}_{ac}^{i-1}$ in column.

$$\text{Attention} = \underbrace{\text{softmax}_2\left(\hat{\boldsymbol{y}}_{na}^{i-1}\right)\text{softmax}_1\left(\hat{\boldsymbol{y}}_{ac}^{i-1}\right)^{\top}}_{\text{non-negative}}\hat{\boldsymbol{y}}_{ac}^{i} \tag{4}$$

The process is illustrated in Equation 4. Since we use softmax operation on $\hat{\boldsymbol{y}}_{na}^{i-1}$ and $\hat{\boldsymbol{y}}_{ac}^{i-1}$ separately, we can compute $\text{softmax}_1\left(\hat{\boldsymbol{y}}_{ac}^{i-1}\right)^{\top}\hat{\boldsymbol{y}}_{ac}^{i}$ first. The complexity of it is $O(C^2L)$, which is linear with the resolution.

We explain why such linear attention works for global correlation capturing. In Equation 4, we can get $\text{softmax}_2\left(\hat{\boldsymbol{y}}_{na}^{i-1}\right)\text{softmax}_1\left(\hat{\boldsymbol{y}}_{ac}^{i-1}\right) \geq 0$. The non-negative property makes it can be treated as a similarity metric. The metric is implicit because we compute $\text{softmax}_1\left(\hat{\boldsymbol{y}}_{ac}^{i-1}\right)\hat{\boldsymbol{y}}_{ac}^{i}$ first in practice.

3.4. Improvements on Global Context Modules

We illustrate the intra-slice global context module and inter-slice global context module in Figure 1. In MLIC$^+$, there is no position encoding in intra-slice and inter-slice global spatial context modules. In our modules, we employ a 1×1 convolutional layer for embedding and a 3×3 depth-wise convolutional layer for learnable position embedding. Following MLIC$^+$ (Jiang et al., 2022), we also adopt a 5×5 convolutional layer to aggregate the global correlations among adjacent symbols. We employ a residual bottleneck (Jiang et al., 2023) to fuse the global context further instead of a MLP in MLIC$^+$ (Jiang et al., 2022). The difference between our intra-slice global context module and our inter-slice global context module is the input. The input of intra-slice global context module is $\hat{\boldsymbol{y}}_{na}^{i-1}$, $\hat{\boldsymbol{y}}_{ac}^{i-1}$, and $\hat{\boldsymbol{y}}_{ac}^{i}$. Different from MLIC$^+$ (Jiang et al., 2022), we capture inter-slice global contexts from slices $\hat{\boldsymbol{y}}^{<i}$, instead of using $\hat{\boldsymbol{y}}_{ac}^{i}$ as approximation and capture inter-slice global contexts from $\hat{\boldsymbol{y}}^{i-1}$. This is because the different slices share the similar global correlations and capturing inter-slice global contexts from more slices leads to better performance. Note that the skip connection is deprecated because the global context capturing is based on calculating the similarity first and then calculating the weighted sum based on the similarity. Since the depth of our attention block is 1, deprecating the skip connection will not leads to gradient vanishing. Deprecating the skip connection also reduces the complexity. Based on the linear complexity attention and improvements on the architecture, our MLIC^{++} even performs better than MLIC$^+$ at some bit-rates with lower complexity.

4. Experiments

4.1. Setup

We build our MLIC^{++} on Pytorch (Paszke et al., 2019) and CompressAI (Bégaint et al., 2020). We train our MLIC^{++} on 10^5 images with a resolution larger than 512×512 from ImageNet (Deng et al., 2009), COCO (Lin et al., 2014), DIV2K (Agustsson & Timofte, 2017), and Flickr2K (Lim et al., 2017). We set λ to $\{18, 35, 67, 130, 250, 483\}\times10^{-4}$ for MSE and set λ to $\{2.4, 4.58, 8.73, 16.64, 31.73, 60.5\}$ for MS-SSIM (Wang et al., 2003). We adopt the training strategy of MLIC and MLIC$^+$ (Jiang et al., 2022) to train our MLIC^{++}.

4.2. Performance

We evalute the performance of our MLIC^{++} on Kodak (Kodak, 1993). We use VTM-17.0 (VTM, 2022) as anchor. We compare our MLIC^{++} with recent models (Jiang et al.,

Input Resolution	MLIC (Jiang et al., 2022)			MLIC^{+} (Jiang et al., 2022)			MLIC^{++}(Ours)		
	Enc.Tot (s)	Dec.Tot (s)	Peak Mem (GB)	Enc.Tot (s)	Dec.Tot (s)	Peak Mem (GB)	Enc.Tot (s)	Dec.Tot (s)	Peak Mem (GB)
512×512	0.0984	0.1376	0.8929	0.1676	0.2355	1.9759	0.1687	0.2363	1.9603
768×768	0.1564	0.2167	1.1808	0.2795	0.3640	2.5027	0.2286	0.3083	2.2089
1024×1024	0.2429	0.3101	1.6873	0.4748	0.5854	3.6110	0.3204	0.4177	2.5483
1536×1536	0.4858	0.5929	3.7926	1.0460	1.1515	9.1465	0.6266	0.7604	3.5151
2048×2048	0.8896	1.0281	9.9478	2.0211	2.1461	24.3720	1.0998	1.2339	4.8625
2560×2560	OM	OM	>32 (OM)	OM	OM	>32 (OM)	1.6511	1.8568	6.5996

Table 2. Complexity comparison among MLIC (Jiang et al., 2022), MLIC^{+} (Jiang et al., 2022) and MLIC^{++}. "Enc.Tot" and "Dec.Tot" mean the total encoding and decoding time, including entropy coding time. "Peak Mem" means the peak memory during the encoding and decoding. "OM" means out of memory. We evaluate on a single Tesla V100-32G GPU and a Xeon(R) 8260 CPU.

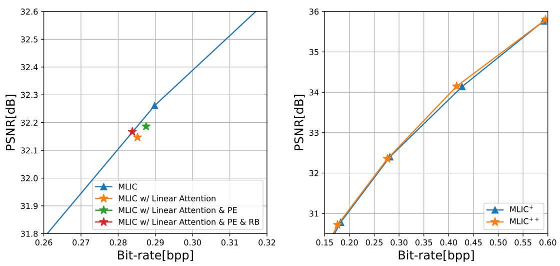

Figure 3. Ablation Studies on Kodak dataset. "PE" means positional encoding. "RB" means residual bottleneck.

2022; Koyuncu et al., 2022; He et al., 2022; Zou et al., 2022; Zhu et al., 2022; Minnen & Singh, 2020; Cheng et al., 2020). The results are illustrated in Table 1 and Figure 2. Our MLIC^{++} outperforms other models and achieves state-of-the-art performance. However, it has to be admitted that, the difference in rate-distortion performance between MLIC^{++} and MLIC^{+} is very small.

4.3. Complexity

We evaluate the complexity on encoding time, decoding time, and peak memory consumption. We compare our MLIC^{++} with MLIC and MLIC^{+} (Jiang et al., 2022). We report the results in Table 2. On 2048×2048 images, the encoding time and decoding time of our MLIC^{++} is one half of that of MLIC^{+} (Jiang et al., 2022), and MLIC^{++} consumes one-fifth of the peak memory consumed by MLIC^{+} (Jiang et al., 2022). On 2560×2560 images, MLIC and MLIC^{+} (Jiang et al., 2022) cannot encode and decode on a Tesla V100-32G, while our MLIC^{++} can encode and decode successfully and only consume 6.6 GB Memory. Our MLIC^{++} is more efficient.

4.4. Ablation Studies

We conduct ablation studies on Kodak (Kodak, 1993). We evaluate the contribution of linear complexity intra-slice global context model on MLIC (Jiang et al., 2022). We evaluate the contribution of linear complexity inter-slice global context model on MLIC^{+} (Jiang et al., 2022). The ablation results are illustrated in Figure 3. Our linear complexity intra-slice global spatial context module leads to almost no performance degradation. Compared to the inter-slice global spatial context module in MLIC^{+}, our MLIC^{++} performs better at some bit-rates. The almost no performance drop can be attributed to learnable positional encoding, and residual bottlenecks. Our MLIC^{++} performs better than MLIC^{+} at some bit-rates because MLIC^{++} capture inter-slice global contexts from more slices.

5. Limitations

One drawback of our MLIC^{++} is its transform modules. We find the transform modules lead to slight performance degradation when the bit-rate on Kodak (Kodak, 1993) ≥ 1. We think employing better transform modules (Jiang et al., 2023; He et al., 2022; Zou et al., 2022; Zhu et al., 2022) will address this problem. The other drawback is the parameters. In MLIC^{++}, we use separate modules for each slice, which leads to more parameters, however, it is harmless because the parameters have no high relevance with FLOPs, encoding time, and decoding time.

6. Conclusion

In this paper, we propose linear complexity intra-slice and inter-slice global context modules, which further improve the performance and reduce the complexity. We build the multi-reference entropy model MEM^{++} with the support of linear complexity intra-slice and inter-slice global context modules. Based on MEM^{++}, we obtain state-of-the-art model MLIC^{++}. To make our MLIC^{++} more practical, in the future, we will investigate the asymmetrical design (Yang & Mandt, 2023) between analysis and synthesis transform, and lighter multi-reference entropy model.

7. Acknowledgement

This work is financially supported by National Natural Science Foundation of China U21B2012 and 62072013, Shenzhen Science and Technology Program-Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents project(Grant No. RCJC20200714114435057) , Shenzhen Science and Technology Program-Shenzhen Hong Kong joint funding project (Grant No. SGDX20211123144400001), this work is also financially supported for Outstanding Talents Training Fund in Shenzhen.

References

Agustsson, E. and Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1122–1131, 2017.

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression. In *Int. Conf. on Learning Representations*, 2017.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. In *Int. Conf. on Learning Representations*, 2018.

Ballé, J., Chou, P. A., Minnen, D., Singh, S., Johnston, N., Agustsson, E., Hwang, S. J., and Toderici, G. Nonlinear transform coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):339–353, 2020.

Bégaint, J., Racapé, F., Feltman, S., and Pushparaja, A. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Guo, Z., Zhang, Z., Feng, R., and Chen, Z. Causal contextual prediction for learned image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

He, D., Zheng, Y., Sun, B., Wang, Y., and Qin, H. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14771–14780, 2021.

He, D., Yang, Z., Peng, W., Ma, R., Qin, H., and Wang, Y. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. *arXiv preprint arXiv:2203.10886*, 2022.

Jiang, W., Yang, J., Zhai, Y., and Wang, R. Multi-reference entropy model for learned image compression. *arXiv preprint arXiv:2211.07273*, 2022.

Jiang, W., Ning, P., and Wang, R. Slic: Self-conditioned adaptive transform with large-scale receptive fields for learned image compression. *arXiv preprint arXiv:2304.09571*, 2023.

Kodak, E. Kodak lossless true color image suite (photocd pcd0992), 1993. URL http://r0k.us/graphics/kodak/.

Koyuncu, A. B., Gao, H., Boev, A., Gaikov, G., Alshina, E., and Steinbach, E. Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pp. 447–463. Springer, 2022.

Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Ma, C., Wang, Z., Liao, R., and Ye, Y. A cross channel context model for latents in deep image compression. *arXiv preprint arXiv:2103.02884*, 2021.

Minnen, D. and Singh, S. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 3339–3343. IEEE, 2020.

Minnen, D., Ballé, J., and Toderici, G. D. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, pp. 10771–10780, 2018.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Shen, Z., Zhang, M., Zhao, H., Yi, S., and Li, H. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3531–3539, 2021.

Theis, L., Shi, W., Cunningham, A., and Huszár, F. Lossy image compression with compressive autoencoders. In *Int. Conf. on Learning Representations*, 2017.

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

VTM. Versatile video coding reference software version 17.0 (vtm-17.0), 2022. URL https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tags/VTM-17.0.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Yang, Y. and Mandt, S. Asymmetrically-powered neural image compression with shallow decoders. *arXiv preprint arXiv:2304.06244*, 2023.

Zhu, Y., Yang, Y., and Cohen, T. Transformer-based transform coding. In *International Conference on Learning Representations*, 2022.

Zou, R., Song, C., and Zhang, Z. The devil is in the details: Window-based attention for image compression. In *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.