ALIC: Adaptive Fusion Entropy Model for Learned Image Compression

Lingxue Li[1,2], Meiqin Liu[1,2*], Yifan Zhang[1,2], Qi Tang[1,2], Chao Yao[3], Yao Zhao[1,2]
[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China
[2] Visual Intelligence + X International Cooperation Joint Laboratory of MOE, Beijing Jiaotong University
[3]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
Emails: {lingxueli, mqliu, ifzhang, qitang, yzhao}@bjtu.edu.cn, yaochao@ustb.edu.cn

***Abstract*—Recently, learned image compression algorithms have achieved significant performance. The entropy model is crucial for improving the rate-distortion performance by estimating the probability distribution of latent representation. In this paper, we propose an adaptive fusion entropy model for learned image compression (ALIC). To explore the correlation between channel and global spatial features, an adaptive fusion entropy model (AFEM) is designed. AFEM first slices the latent representation along the channels and leverages the adaptive channel fusion context module (ACFC) to capture correlations between the decoded and current slices. Subsequently, AFEM uses the adaptive spatial fusion context module (ASFC) to further divide the current slice into encoding pixels and reference pixels, thus improving the accuracy of probability estimation. The attention map and modulation parameter are introduced in ACFC and ASFC to interact with channel and spatial features. In addition, the variable-rate residual transformer (VResFormer) is proposed to control dynamic bit-rate by selectively modulating the high-frequency component according to coefficient weight and bias. Experimental results indicate that our ALIC outperforms other learned image compression algorithms. Our ALIC saves 5.89% bit-rate compared with VVC (4:4:4) on Kodak dataset.**

***Index Terms*—Learned image compression, Entropy model, Variable-rate**

I. Introduction

Early compression standards, such as JPEG [1], JPEG2000 [2] and BPG [3] use linear transformations to reduce the redundant information. Recently, the learned image compression algorithms have achieved remarkable progress [4]–[6] even surpassing VVC (4:4:4) [7]. Most learned image compression models are based on variational autoencoders (VAEs) [8] and follow the paradigm consisting of transformation, quantization, entropy coding, and inverse transformation.

The current entropy models aim to reduce the redundancy of information by precisely estimating the probability distribution. For example, a conditional Gaussian mixture model [9] was devised to optimize the autoregressive model based on the hyperprior model [10]. Subsequently, a Gaussian mixture model (GMM) [11] was proposed to estimate the distribution of complex data by mixing multiple Gaussian distributions. A Gaussian-Laplacian-Logistic mixture model (GLLMM) [6], which combined multiple distributions, could estimate the probability distribution more accurately and further reduce redundancy. However, the autoregressive model significantly improves the performance but is plagued by high complexity because of the strict sequential decoding process. To address this issue, Minnen et al. [12] proposed a channel-wise autoregressive entropy model to accelerate serial processing and He et al. [13] introduced a parallel decoding checkerboard context model from a spatial perspective. Recent algorithms utilized attention mechanisms to enhance correlation modeling and reduce redundancy in latent representations. CBAM [14] and DAT [15] refined the characteristics by applying attention operations along the channel and spatial dimensions. IBVC [16] proposed a conditional spatio-temporal contextual decoder that reduced redundant information through the channel attention mechanism. ELIC [17] proposed a space-channel ConTeXt model, which divided the channel context into uneven slices to achieve an effective reduction in bit-rate. However, these algorithms ignored the relationship between channel dimension and global spatial dimension and limited the performance of image compression.

Hence, we propose an Adaptive Fusion Entropy Model (AFEM) combining the channel context and global spatial context to further eliminate redundancy. Specifically, AFEM mainly consists of two key components: the Adaptive Channel Fusion Context Module (ACFC) and the Adaptive Spatial Fusion Context Module (ASFC). To separately process inter-slice information, ACFC integrates depthwise convolution (DW-Conv) with channel attention mechanisms. By referring to the previous slices, the probability distribution of the current slice can be accurately estimated. To handle intra-slice information, ASFC leverages the Swin-Transformer to model long-term dependencies. A modulation mechanism is introduced in ACFC and ASFC to enable the interaction between DW-Conv and attention operations, fully exploiting the complementary context. Moreover, a variable-rate residual transformer (VResFormer) is proposed in the encoder-decoder network. Specifically, it contains a transformer branch and a variable-rate branch. The latter adjusts the high-frequency components of the former according to the controllable parameter λ, adaptively adjusting the content for compression and achieving variable-rate. We evaluate our ALIC on Kodak and CLIC datasets in terms of MSE and MS-SSIM. The experimental results show that our ALIC outperforms the traditional codecs and other learned compression algorithms. Specifically, by calculating the BD-Rate of MSE result on Kodak dataset, our ALIC saves 5.89% bit-rate compared with VVC (4:4:4).

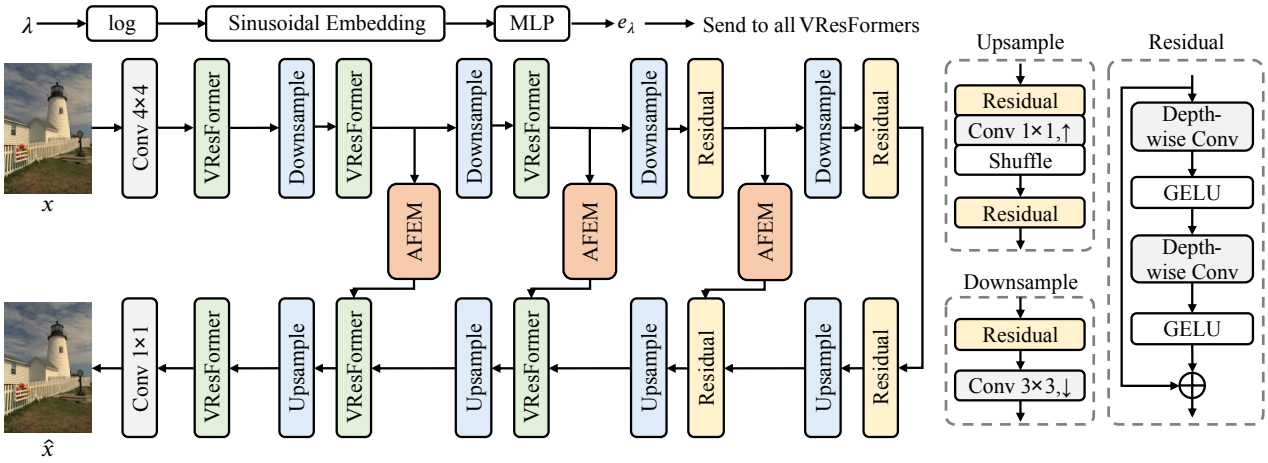

Fig. 1: The overview framework of our proposed ALIC. AFEM represents the proposed adaptive fusion entropy model and VResFormer represents the proposed variable-rate residual transformer. The controllable parameter λ controls the adjustment of high-frequency details to enable variable-rate compression. x is the input image and $\hat{x}$ is the reconstructed image.

II. Method

A. Overall Architecture

The overview of the proposed learned image compression framework is depicted in Fig. 1. The input image x first passes through the downsampling layer and the proposed variable-rate residual transformer (VResFormer) to generate the latent representation y. Then, the proposed adaptive fusion entropy model (AFEM) accurately estimates the probability distribution. Finally, the decoded latent representation $\hat{y}$ is used to gradually reconstruct the image $\hat{x}$.

B. Adaptive Fusion Entropy Model

To better utilize channel and spatial information in probability estimation, we propose an Adaptive Fusion Entropy Model (AFEM) shown in Fig. 2. The latent representation $\hat{y}$ is divided into several slices along channel dimensions [12]. There is a high correlation between adjacent slices, which allows the model to reconstruct the current slice more precisely by using the previous decoded information. For the i-th slice $\hat{y}_i$, the checkerboard mask is further applied to divide it into anchor part $\hat{y}_i^a$ and non-anchor part $\hat{y}_i^{na}$. Non-anchor $\hat{y}_i^{na}$ is decoded with reference to the contents of anchor $\hat{y}_i^a$. To explore correlations between and within slices, we design the Adaptive Channel Fusion Context Module (ACFC) and the Adaptive Spatial Fusion Context Module (ASFC), which will be elaborated on in detail below.

1) Adaptive Channel Fusion Context Module: To improve the accuracy of probability estimation, the Adaptive Channel Fusion Context Module (ACFC) is devised to utilize the correlation between slices. As shown in Fig. 3 (left), ACFC adopts a two-branch structure. In DW-Conv branch, the input feature f passes through DW-Conv to efficiently extract context information $f_d \in \mathbb{R}^{H\times W\times C}$, which is defined as:

$$f_d = DConv(f) \tag{1}$$

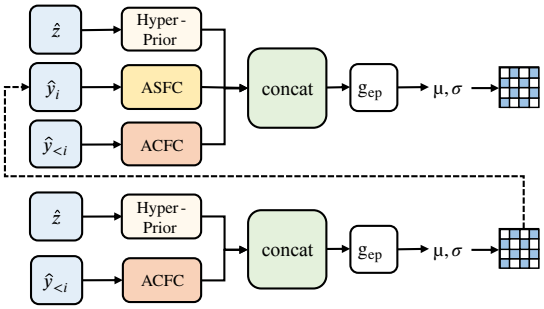

Fig. 2: The structure of the adaptive fusion entropy model.

where $DConv(\cdot)$ represents DW-Conv. The attention branch first performs pooling operations to calculate channel attention map $m_c \in \mathbb{R}^{1\times 1\times C}$. The above operation is formulated as:

$$m_c = \sigma(MLP(P_{avg}(f)) + MLP(P_{max}(f))) \tag{2}$$

where P_{avg} and P_{max} are respectively the average pooling and the max pooling, $MLP(\cdot)$ is the multi-layer network and $\sigma(\cdot)$ denotes the sigmoid function. Then, the input feature f is multiplied by m_c to adjust the weight of each channel and generate the output feature $f_c \in \mathbb{R}^{H\times W\times C}$, which is formulated as:

$$f_c = f \times m_c \tag{3}$$

To realize the adaptive feature fusion, we propose a modulation method for the interaction between the attention branch and the DW-Conv branch. The modulation parameter f_d' is obtained via the convolution operation, which is defined as:

$$f_d^{'} = \sigma(Conv(f_d)) \tag{4}$$

where $Conv(\cdot)$ represents the convolution operation. Subsequently, the modulation parameter f_d' and the channel attention map m_c are used as the adaptive weights of feature fusion to obtain the output feature f_{ch} of ACFC, which is defined as:

$$f_{ch} = Conv((f_d \times m_c) \oplus (f_c \times f_d^{'})) \tag{5}$$

where $\oplus$ represents element-wise addition.

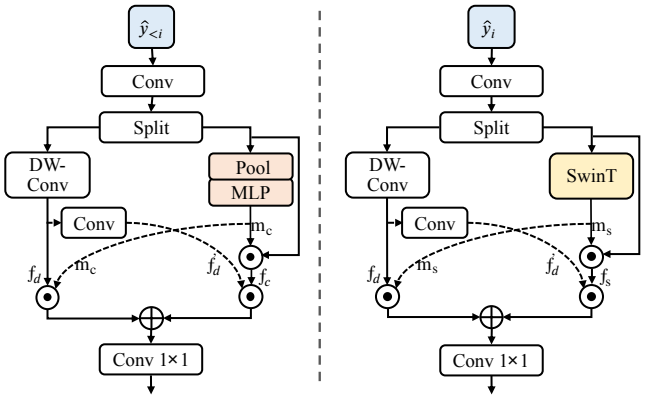

Fig. 3: The structure of adaptive channel fusion context module (left) and adaptive spatial fusion context module (right).

2) Adaptive Spatial Fusion Context Module: To make better use of the global spatial information in each slice, an Adaptive Spatial Fusion Context Module (ASFC) is designed as shown in Fig. 3 (right). In the attention branch, the feature f is processed by Swin-Transformer and sigmoid function σ to obtain the spatial attention map m_s, which is formulated as:

$$m_s = \sigma(Swin(f)) \tag{6}$$

where $Swin(\cdot)$ denotes Swin-Transformer. The attention map m_s re-weights feature f in the spatial dimension to obtain the output feature f_s, which is formulated as:

$$f_s = f \odot m_s \tag{7}$$

where $\odot$ represents element-wise multiplication. Similar to ACFC, the local feature $f_d \in \mathbb{R}^{H\times W\times C}$ can be obtained via the DW-Conv branch. After that, the proposed modulation method is also utilized to achieve adaptive feature fusion between the attention branch and the DW-Conv branch. Finally, the modulation parameter f_d' and the spatial attention map m_s are employed as the adaptive weights of feature fusion to obtain the output feature f_{sp} of ASFC, which is defined as:

$$f_{sp} = Conv((f_d \times m_s) \oplus (f_s \times f_d^{'})) \tag{8}$$

C. Variable-Rate Residual Transformer

In each stage of encoder-decoder networks, the Variable-Rate Residual Transformer (VResFormer) is designed to assemble global and local information [18] and achieve the variable-rate [19], as illustrated in Fig. 4. The transformer branch performs SW-MSA to focus on the long-term dependencies between features. The variable-rate branch adaptively ignores high-frequency components according to the demands of bit-rates. It utilizes AdaLN [4] to generate coefficients w, b for modulating high-frequency features.

$$w, b = split(Linear(GELU(e_\lambda))) \tag{9}$$

where $GELU(\cdot)$ and $Linear(\cdot)$ denote the GELU operation and linear operation respectively. e_λ is respectively split into weight w and bias b. Next, the input feature f_{bn} of the AdaLN block is modulated by the obtained coefficients. The output feature f_{ad} of the AdaLN block is calculated as:

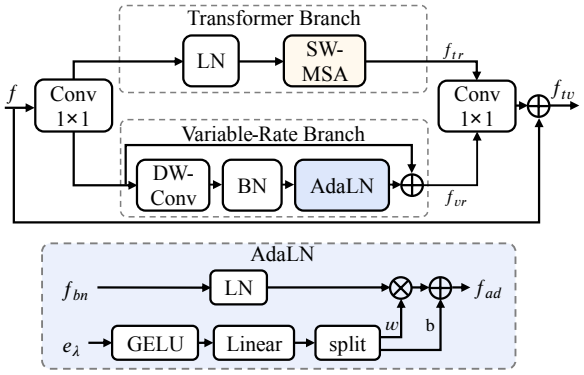

Fig. 4: The structure of variable-rate residual transformer.

$$f_{ad} = w \times LN(f_{bn}) + b \tag{10}$$

where $LN(\cdot)$ represents the layer normalization.

D. Loss Function

Our learned image compression method aims to minimize the rate-distortion loss. The trade-off between rate and distortion is determined by the Lagrange multiplier λ. The loss function of our framework is defined as follows:

$$L = R(\hat{y}) + \lambda D(x, \hat{x}) \tag{11}$$

where $R(\cdot)$ is utilized to calculate bit-rate for representation compression, $D(\cdot)$ measures the distortion between the original image x and the reconstructed image $\hat{x}$.

III. Experiments

A. Experimental Settings

We use COCO 2017 [20] as the training dataset, which contains 118,287 images with dimensions of 640×420. We evaluate our model on Kodak [21] and CLIC [22] datasets. The Kodak dataset comprises 24 images with the resolution of 512×768 or 768×512. The CLIC dataset consists of 41 images with the resolution of 2048×1370.

During the training process, the images are cropped to patches with the size of 256×256. The batch size is set to 16 to balance the memory and the training efficiency. We use the Adam optimizer with a learning rate of 4×10^{-4} to ensure stable convergence. All experiments are conducted on a RTX 3090 GPU for 200 epochs. We train our models using MSE and MS-SSIM loss.

B. Comparisonal Results

We compare ALIC with several representative image compression algorithms including traditional codecs [3], [7] and learned-based algorithms [4], [6], [11], [23]–[28]. The results of PSNR and SSIM (values are converted to decibels by $-10log_{10}(1-\text{MS-SSIM})$ for clarity) are shown in Fig. 5. To intuitively evaluate the advancement of our ALIC, we further compare the BD-Rate (The lower is the better) and BD-PSNR

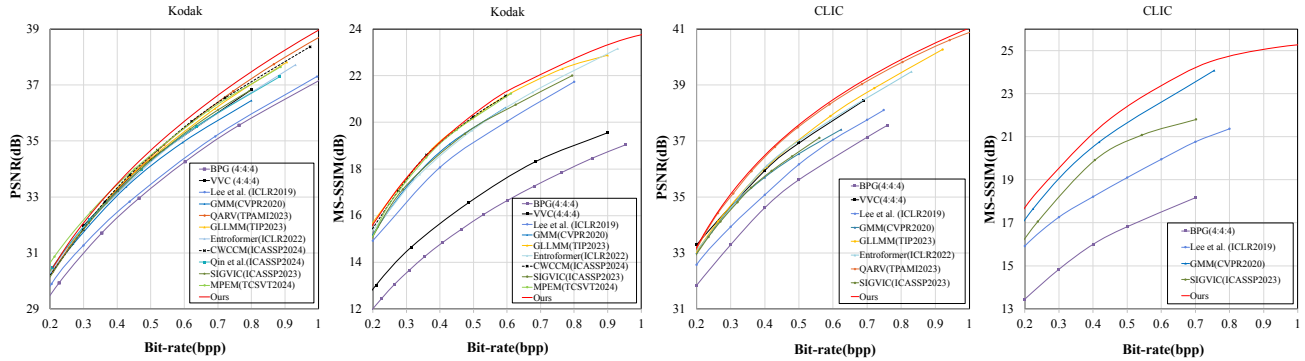

Fig. 5: The RD curves, using MSE and MS-SSIM as the distortion term of loss function, on Kodak and CLIC datasets.

TABLE I: BD-Rate and BD-PSNR on Kodak and CLIC.

Methods	Kodak		CLIC	
	BD-PSNR	BD-Rate	BD-PSNR	BD-Rate
GLLMM [6]	-0.070dB	2.55%	0.092dB	-1.56%
GMM [11]	-0.350dB	8.13%	-0.269dB	6.64%
Lee et al. [23]	-0.864dB	21.59%	-0.796dB	20.22%
CWCCM [24]	0.134dB	-3.53%	-	-
SIGVIC [25]	-0.091dB	2.23%	-0.209dB	5.29%
MPEM [26]	0.131dB	-1.88%	-	-
QARV [4]	0.085dB	-1.74%	0.374dB	-7.91%
Entroformer [27]	-0.021dB	-1.92%	0.134dB	-2.71%
Qin et al. [28]	-0.093dB	1.64%	-	-
ALIC (Ours)	**0.334dB**	**-5.89%**	**0.481dB**	**-9.80%**

(The higher is the better). The results are shown in TABLE I, with VVC (4:4:4) as the anchor.

On Kodak dataset, our ALIC outperforms GLLMM by 0.404 dB in PSNR and reduces 8.44% bit-rate. ALIC outperforms GMM by 0.684 dB in PSNR and reduces 14.02% bit-rate. ALIC achieves 0.334 dB in PSNR and reduces 5.89% bit-rate. On CLIC dataset, ALIC yields 0.481 dB in PSNR and saves 9.80% bit-rate. Obviously, our ALIC can achieve the best performance in PSNR.

C. Subjective Performance

The subjective performance of our ALIC is shown in Fig. 6. Our ALIC recovers more intricate details in enlarged regions such as the texture of the petals and the window blinds. It indicates that our ALIC excels in visual quality.

D. Ablation Study

To compare the performance of different components, we conduct the following ablation studies sequentially. The experimental results are shown in TABLE II.

The results show that adding ACFC and ASFC reduces the BD-Rate by 1.565% and 1.253% respectively. Scheme C is reduced by 2.688% in BD-Rate, showing that our modules can effectively utilize channel and spatial information to accurately estimate the probability distribution. Scheme D saves 3.904% in BD-Rate, which proves that VResFormer can realize flexible bit-rate control. Compared with the network without any proposed components, BD-Rate shows progressive improvements.

Fig. 6: Visual quality of reconstructed images.

TABLE II: Results of ablation study on Kodak dataset.

Schemes	ACFC	ASFC	VResFormer	BD-Rate
A	✓			-1.565%
B		✓		-1.253%
C	✓	✓		-2.688%
D	✓	✓	✓	-3.904%

IV. Conclusion

In this paper, we propose an Adaptive Fusion Entropy Model for learned image compression (ALIC). The Adaptive Fusion Entropy Model (AFEM) is designed with the incorporation of the Adaptive Channel Fusion Context Module (ACFC) and the Adaptive Spatial Fusion Context Module (ASFC) to capture the correlations between slices and within slices separately. In addition, the Variable-Rate Residual Transformer (VResFormer) is designed to adapt and ignore the high-frequency component to meet the need for flexible rate control. Experimental results show that our ALIC outperforms other compared algorithms, achieving significant storage savings and improvements in flexibility.

Acknowledgments

This work is supported by the Talent Fund of Beijing Jiaotong University (2024XKRC023) and the National Natural Science Foundation of China (62372036, 62120106009, 62332017).

REFERENCES

[1] G. K. Wallace, "The JPEG still Picture Compression Standard," *CACM*, vol. 34, no. 4, pp. 30–44, 1991.

[2] D. S. Taubman, M. W. Marcellin, and M. Rabbani, "JPEG2000: Image compression fundamentals, standards and practice," *JEI*, vol. 11, no. 2, pp. 286–287, 2002.

[3] F. Bellard, "BPG image format (2014)," *URL http://bellard. org/bpg/.[Online, Accessed 2016-08-05]*, vol. 1, no. 2, 2016.

[4] Z. Duan, M. Lu, J. Ma, Y. Huang, Z. Ma, and F. Zhu, "QARV: Quantization-aware resnet vae for lossy image compression," *TPAMI*, 2023.

[5] W. Jiang, J. Yang, Y. Zhai, P. Ning, F. Gao, and R. Wang, "MLIC: Multi-reference entropy model for learned image compression," in *ACM MM*, 2023, pp. 7618–7627.

[6] H. Fu, F. Liang, J. Lin, B. Li, M. Akbari, J. Liang, G. Zhang, D. Liu, C. Tu, and J. Han, "Learned image compression with discretized gaussian-laplacian-logistic mixture model and concatenated residual modules," *arXiv:2107.06463*, 2021.

[7] 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," *TCSVT*, vol. 31, no. 10, pp. 3736–3764, 2021.

[8] D. P. Kingma, "Auto-encoding variational bayes," *arXiv:1312.6114*, 2013.

[9] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *ANIP*, vol. 31, 2018.

[10] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv:1802.01436*, 2018.

[11] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *CVPR*, 2020.

[12] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *ICIP*, 2020.

[13] D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard context model for efficient learned image compression," in *CVPR*, 2021.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *ECCV*, 2018.

[15] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," in *ICCV*, 2023.

[16] C. Xu, M. Liu, C. Yao, W. Lin, and Y. Zhao, "IBVC: Interpolation-driven b-frame video compression," *PR*, vol. 153, p. 110465, 2024.

[17] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *CVPR*, 2022.

[18] M. Liu, C. Xu, C. Yao, C. Lin, and Y. Zhao, "JNMR: Joint non-linear motion regression for video frame interpolation," *TIP*, 2023.

[19] Y. Zhang, M. Liu, C. Xu, Q. Tang, C. Yao, and Y. Zhao, "TLVC: Temporal bit-rate allocation for learned video compression," in *ICME*, 2024.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.

[21] "Kodak PhotoCD dataset." 2013. [Online]. Available: http://r0k.us/graphics/kodak/.

[22] "CLIC dataset." 2019. [Online]. Available: http://www.compression.cc/.

[23] J. Lee, S. Cho, and S.-K. Beack, "Context-adaptive entropy model for end-to-end optimized image compression," *arXiv:1809.10452*, 2018.

[24] H. Fu, F. Liang, J. Liang, Z. Fang, G. Zhang, and J. Han, "Efficient learned image compression with selective kernel residual module and channel-wise causal context model," in *ICASSP*, 2024.

[25] J. Liang, M. Liu, C. Yao, C. Lin, and Y. Zhao, "SIGVIC: Spatial importance guided variable-rate image compression," in *ICASSP*, 2023.

[26] C. Li, S. Yin, C. Jia, F. Meng, Y. Tian, and Y. Liang, "Multirate progressive entropy model for learned image compression," *TCSVT*, 2024.

[27] Y. Qian, M. Lin, X. Sun, Z. Tan, and R. Jin, "Entroformer: A transformer-based entropy model for learned image compression," *arXiv:2202.05492*, 2022.

[28] P. Qin, Y. Bao, F. Meng, W. Tan, C. Li, G. Wang, and Y. Liang, "Leveraging redundancy in feature for efficient learned image compression," in *ICASSP*, 2024.