End-to-End RGB-IR Joint Image Compression with Channel-wise Cross-modality Entropy Model

Haofeng Wang*†¶, Fangtao Zhou‡, Qi Zhang†, Zeyuan Chen†§, Enci Zhang*, Zhao Wang¶, Xiaofeng Huang‡, Siwei Ma†
*School of Electronic and Computer Engineering, Peking University, Shenzhen, China
†National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Peking, China
‡School of Communication Engineering, Hangzhou Dianzi University, Zhejiang, China
§Pengcheng Laboratory, Shenzhen, China
¶Advanced Institute of Information Technology, Peking University, Zhejiang, China

***Abstract*—RGB-IR image pairs are frequently applied simultaneously in various applications like intelligent surveillance. However, as the number of modalities increases, the required data storage and transmission costs also doubles. Therefore, efficient RGB-IR data compression is essential. This work proposes a joint compression framework for RGB-IR image pair. Specifically, to fully utilize cross-modality prior information for accurate context probability modeling within and between modalities, we propose a Channel-wise Cross-modality Entropy Model (CCEM). Among CCEM, a Low-frequency Context Extraction Block (LCEB) and a Low-frequency Context Fusion Block (LCFB) are designed for extracting and aggregating the global low-frequency information from both modalities, which assist the model in predicting entropy parameters more accurately. Experimental results demonstrate that our approach outperforms existing single-modality image compression methods on LLVIP dataset. Compared to MLIC++, the best-performing image codec on the Kodak dataset, our proposed framework achieves a bit rate saving of 14.6% for RGB-IR pair.**

***Index Terms*—Image compression, multi-modality image compression, cross-modality entropy model, RGB-IR joint image compression.**

I. INTRODUCTION

Recently, RGB-IR images pairs captured within the same scene have been jointly applied to various practical scenarios [1]–[3]. This is largely due to the fact that the advantages of RGB and IR modalities are complementary. RGB images, known for their high resolution and ability to capture fine details such as textures, are limited by the reliance on ambient lighting. [4] However, this limitation can be mitigated by incorporating IR images because of the low sensitivity to illumination changes. Nevertheless, the use of RGB-IR image pairs significantly increases the amount of data that needs to be transmitted and stored. Consequently, developing an efficient joint compression method for RGB-IR image pairs has become a crucial and challenging task.

Over the past decades, deep learning-based image compression methods [5]–[10] have been extensively developed, pushing the boundaries of rate-distortion performance. It is intuitive to compress RGB and IR modalities independently using these neural codecs. However, the redundancy between RGB and IR modalities is not fully exploited during the compression, thereby limiting the overall rate-distortion performance.

In recent years, several multi-modality data compression methods [11], [12] have been proposed. However, most of these methods are specifically designed for visible images paired with depth or hyperspectral images, which are not suitable for compressing RGB-IR image pairs due to the different distributions between modalities. For example, Unlike depth image which uses depth for spatial geometry, ir image integrates infrared data to capture thermal properties and reduce sensitivity to lighting. For RGB-IR image pairs, a learning-based multimodal image compression framework [13] leverages one modality as an anchor to assist in the encoding and decoding process of the other modality. While this approach enhances the compression performance of one modality, it does not leverage the cross-modality correlation in the context-based entropy model, thereby limiting the rate-distortion performance of both modalities, which is often necessary in practical applications where RGB-IR image pairs are used together [14], [15]. Besides, the compression cannot be performed simultaneously, as one modality has to be decoded at first to serve as an anchor for compressing another, which lowers the computation efficiency. Therefore, designing a framework capable of jointly compressing RGB-IR image pairs by fully exploiting cross-modality correlations as prior information to enhance performance remains a challenge.

In this paper, our main contribution is to propose a dual-branch learning-based RGB-IR joint image compression framework. We design a Channel-wise Cross-modality Entropy Model (CCEM) to fully utilize cross-modality prior information for accurate context probability modeling within and between modalities. Specifically, we propose Low-frequency Context Extraction Block(LCEB) and Low-frequency Context Fusion Block(LCFB) to extract and aggregate low-frequency prior information to further reveal the dependency between the modalities. Besides, unlike previous learning-based method for RGB-IR image pair compression, our approach does not require decoding one modality's image to be an anchor for compressing another. According to the experimental results, our proposed framework attains state-of-the-art performance compared to existing single-modality image compression methods

Xiaofeng Huang and Siwei Ma are corresponding authors for this work.

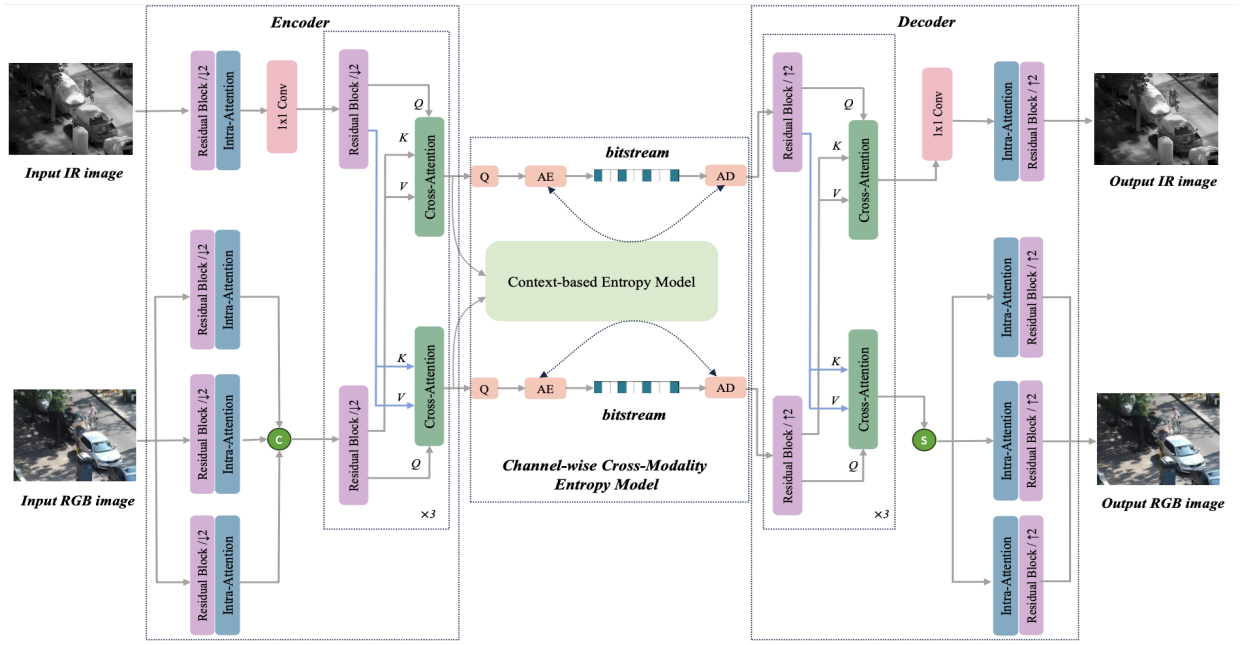

Fig. 1. The overall framework of the proposed method. The network consists of an encoder, a Channel-wise Cross-Modality Entropy Model and a decoder. AE, AD denote arithmetic encoding and decoding, respectively. Q denotes the quantizer, C and S denote concat and split operation, "↑ 2" and "↓ 2" denote upsampling and downsampling by a factor of two, respectively.

on LLVIP dataset [16].

II. PROPOSED METHOD

A. Overall Architecture

The overall architecture of our RGB-IR joint compression framework is illustrated in Fig. 1. We use a transformer-based encoder-decoder architecture. Before compression, the RGB and IR image are converted to the YUV420 format, and the Y, U, V, and IR channels are used as inputs of the model. First, the input channels are individually fed into the Encoder for feature extraction. We use a residual network [6] combined with a self-attention-based module [17] to obtain feature maps y^y, y^u, y^v, and y^{ir} for each input channel. The feature maps from the Y, U, and V channels are then concatenated to form a unified YUV feature y^{yuv}. We use cross-attention to embed cross-modality information within the latent representations y^{yuv} and y^{ir}. Subsequently, y^{yuv} and y^{ir} are quantized to $\hat{y}^{yuv}$ and $\hat{y}^{ir}$, and fed into the proposed Channel-wise Context-based Cross-modality Entropy Model for accurate symbol probability prediction. Finally, $\hat{y}^{yuv}$ and $\hat{y}^{ir}$ are input into the decoder for upsampling and image reconstruction. We denote the encoder, quantizer, decoder as $g_a(\cdot)$, $Q(\cdot)$, and $s_a(\cdot)$, respectively. The overall process can be formulated as:

$$y^i = g_a(x^i;\theta), \hat{y^i} = Q(y^i), \hat{x^i} = g_s(\hat{y^i};\phi) \quad (1)$$

where x^i and $\hat{x^i}$ represents one of the input and output channels and θ, ϕ are learnable parameters.

B. Channel-Wise Cross-modality Entropy Model

The entropy model, plays a key role in boosting compression performance by estimating the distribution of the latent representation. Minnen et al. [17] introduced an entropy model based on spatial autoregressive prediction, surpassing the compression performance of H.265. To accelerate decoding, another work [18] have proposed to split the latent representation into multiple slices and leveraging inter-channel correlations to autoregressively predict the entropy model parameters for each slice. Based on this, MLIC++ [10] incorporates multiple perspectives of context information as multi-references to

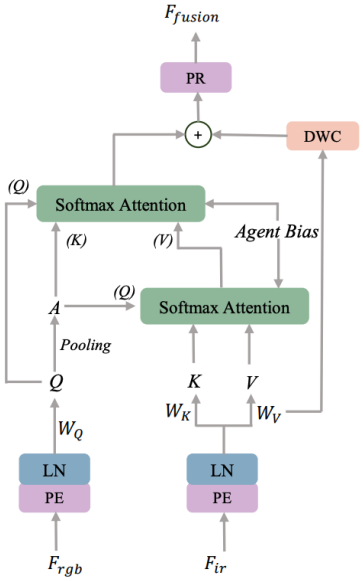

Fig. 2. The architecture of Low-frequency Context Fusion Block(LCFB). PE, PR, LN represent Patch Embedding, Patch Recovery, LayerNorm, respectively.

predict entropy model parameter more accurately. For RGB-IR image pairs, utilizing cross-modality information as a prior context to enhance the accuracy of entropy model parameter prediction is a natural and worthwhile problem to explore.

The global low-frequency information of RGB images and IR images from the same scene is highly similar [20]. Therefore, it is reasonable to infer that, in the compression of RGB-IR image pairs, extracting and aggregating the global low-frequency information from both modalities as a conditional prior will enable the context-based entropy model to predict the parameters of the entropy model more accurately, thereby effectively reducing the bit rate. We designed the Low-frequency Context Extraction Block (LCEB) and Low-frequency Context Fusion Block (LCFB). The LCEB adopts the same structure as the Lite Transformer [20] and loads its pre-trained model at the start of training. As shown in Fig. 2, Instead of using a concatenation operation, we designed the LCFB based on agent-attention [21] to better aggregate global low-frequency information from two modalities. The pipeline for processing the latent representations of the two modalities through the LCFB is as follows:

$$
\begin{aligned}
&\{Q, K, V\} = \left\{F_{rgb}\mathbf{W}^{Q}, F_{ir}\mathbf{W}^{K}, F_{ir}\mathbf{W}^{V}\right\}, \\
&A = \mathrm{Pooling}(Q), \\
&V' = \mathrm{softmax}\left(\frac{AK^{T}}{\sqrt{d_k}} + B_1\right)V, \\
&F = \mathrm{softmax}\left(\frac{QA^{T}}{\sqrt{d_a}} + B_2\right)V', \\
&F_{fusion} = F + \mathrm{DWC}(V).
\end{aligned}
\tag{2}
$$

where F_{rgb} and F_{ir} represent feature of input slices. $\mathbf{W}^{Q}$, $\mathbf{W}^{K}$, $\mathbf{W}^{V}$ are linear projection matrices to map the input features into query Q, key K, and value V spaces, respectively. d, B represent the dimension and relative position bias , respectively. DWC is a depth-wise convolution module [22]. We see the output F_{fusion} as the aggregated global low-frequency information from two modalities and use this context in entropy model to get more accurate entropy parameters.

In our proposed method, we design a Channel-wise Cross-modality Entropy Model (CCEM) for more accurate probability estimation. The architecture of CCEM is shown on the left side of Fig. 3. The latent representation generated from the encoder is fed into a hyperprior model to obtain spatial prior information. Additionally, the latent representation is divided into slices $\{\hat{\boldsymbol{y_m}}^0, \hat{\boldsymbol{y_m}}^1, \cdots \hat{\boldsymbol{y_m}}^N\}$, where m represents one of input modalities. For the IR latent representation, the first slice uses only the hyperprior as context to predict entropy model parameters. For the i^{th} slice, we use the previous slices to exact context and predict entropy parameters. In particular, the slices from 1 to $i-1$ are concatenated and processed through a Low-frequency Context Extraction Block (LCEB) to extract global low-frequency information. The global low-frequency context and hyperprior context are then used to predict entropy parameters. For the RGB latent representation, in addition to the above, the j^{th} slice is processed by concatenating the preceding $j-1$ slices with the global low-frequency information from the previously obtained IR latent representation and input into a Low-frequency Context Fusion Block (LCFB) to derive cross-modality information. This additional cross-modality information is used to further improve the accuracy of the entropy model parameter prediction. Specifically, we denote $\hat{y}_{ir}$ and $\hat{y}_r$ as the latent representation of two modalities. $\hat{z}$ represents the side information extracted from hyperprior. The probability distribution of the latent variables $p_{\hat{y}_{ir}}$ and $p_{\hat{y}_r}$ can be formulated as:

$$
\begin{aligned}
p_{\hat{y}_{ir}|\hat{z}_{ir}}(\hat{y}_{ir}|\hat{z}_{ir}) &= \prod_{i=1}^{N} p_{\hat{y}^i_{ir}|\hat{y}^{<i}_{ir},\hat{z}_{ir}}(\hat{y}^i_{ir}|\hat{y}^{<i}_{ir},\hat{z}_{ir}), \\
p_{\hat{y}_r|\hat{y}_{ir},\hat{z}_r}(\hat{y}_r|\hat{y}_{ir},\hat{z}_r) &= \prod_{i=1}^{N} p_{\hat{y}^i_r|\hat{y}^{<i}_r,\hat{y}_{ir},\hat{z}_r}(\hat{y}^i_r|\hat{y}^{<i}_r,\hat{y}_{ir},\hat{z}_r).
\end{aligned}
\tag{3}
$$

C. Loss Function

The loss function L of our framework is described as:

$$
L = R_{ir} + R_{rgb} + \lambda(D_{ir} + D_{rgb}). \tag{4}
$$

where R_{ir} and R_{rgb} are the bit rate cost of two modalities, they can be calculated by the probability distribution of latent representations. D_{rgb} and D_{ir} are calculated as the pixel-wise mean square error (MSE) between compressed and original image.

III. Experiments

A. Experiment Details

Baseline and Metric We compare our model with the best-performing single-modality codec on the Kodak dataset, MLIC++ [10], and the classic end-to-end codec, Cheng2020 [6]. Additionally, we introduce traditional single-modality image compression method BPG [23], for comparison with our model. We use PSNR to assess the quality of compressed images and Bjontegaard delta rate (BD-Rate) [24] to evaluate the rate-distortion performance. Considering our ultimate goal is the joint compression of RGB-IR image pairs, we compare the average PSNR of both modalities and the corresponding BD-rate with the baseline models. Note that our evaluation metrics are computed in the YUV420 domain.

Training strategy and details Considering that joint training of both modalities from the beginning would require the model to simultaneously process multiple channels from two modalities, it's difficult to learn the features of each modality and their cross-modality correlations. During model training, we propose a two-stage training method. In the first stage, we focus on training for compressing the RGB data. Specifically, after converting the RGB modality to YUV, we input the Y channel data into the proposed model for training. This approach ensures that the model can effectively extract features from the RGB modality in the early stages. After completing the first stage, we proceed to jointly optimize both the RGB and IR modalities. Experimental results show that

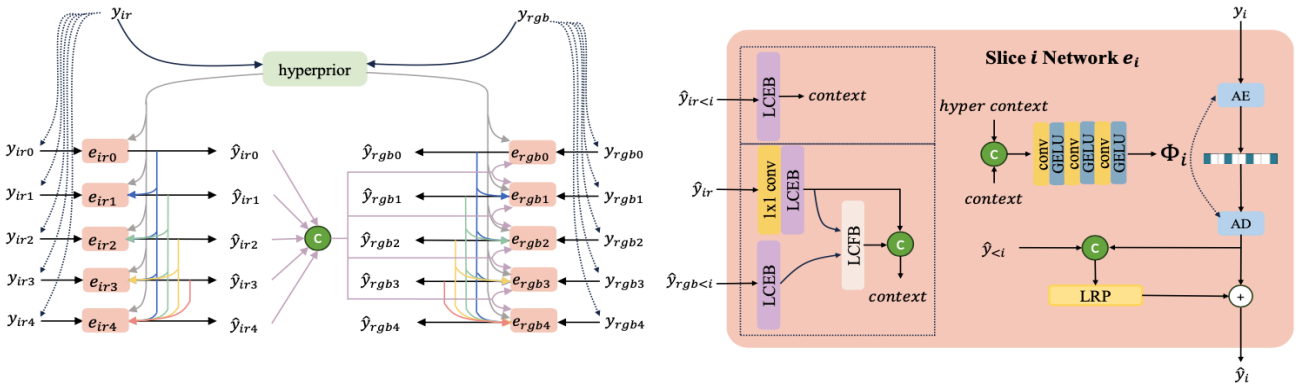

Fig. 3. The architecture of the proposed Channel-wise Cross-Modality Entropy Model. The latent representations are split into slices and sent to hyperprior model. The encoded slices are fed into Low-frequency Context Extraction Block (LCEB) and Low-frequency Context Fusion Block (LCFB) to extract global low-frequency prior, then in slice entropy model e_i, hyperprior context and global low-frequency context are used to predict entropy parameter. LRP represents latent residual prediction module. C denotes concatenate operation.

adopting this training method improves the model's performance by approximately 4% BD-rate. Additionally, we set different hyperparameters λ, to control the bit rate, following the settings in CompressAI [25]. During training, we use the Adam optimizer, and the learning rate gradually decreases from 1e-4 to 1e-5 throughout each stage. We conduct training and testing on LLVIP [16], a widely used RGB-IR dataset. Training is performed on the dataset's 12,000+ training images for 150 epochs in each stage, and testing is carried out on its 3,400+ pairs of test images.

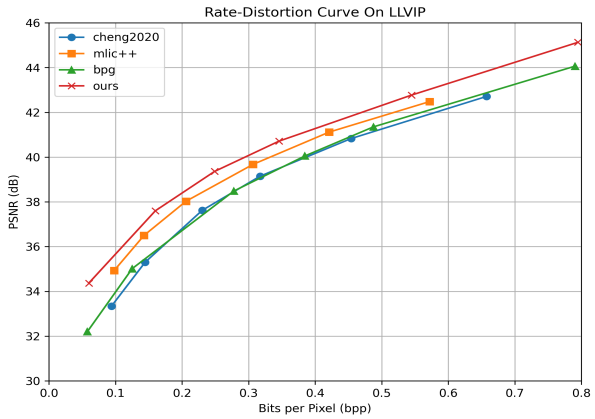

Fig. 4. Experimental results from different image compression approaches on the LLVIP dataset.

B. Experiment Results

Quantitative Results We make a comparison of compression performance among various learning-based codec and BPG on the LLVIP dataset. Note that, to ensure a fair comparison, we re-deployed and retrained the other end-to-end compression frameworks on the LLVIP dataset. Compared to other single-modality compression frameworks, our proposed framework shows a significant improvement in BD-rate performance. Specifically, our method outperforms MLIC++ and BPG by 14.6% and 26.8%, respectively. We plot the corresponding RD curves in Fig. 4 to more intuitively illustrate the performance gap between different codecs. The results clearly demonstrates that our proposed method significantly outperforms the other methods in terms of compression performance.

TABLE I
ABLATION STUDY OF EACH COMPONENT IN CHANNEL-WISE CROSS-MODALITY ENTROPY MODEL

Model	BD-Rate(%)
baseline	-
Channel-wise Cross-modality Entropy Model	-19.34
baseline + LCEB	-6.92
baseline + LCFB	-9.17

Ablation Study: To demonstrate the effectiveness of the proposed LCEB and LCFB modules, we conducted experiments by removing each module individually and comparing the results. Table I shows that both proposed modules contribute to BD-rate performance, and our proposed CCEM significantly enhances compression efficiency.

IV. Conclusion

In this paper, we propose a joint compression framework for RGB-IR image pair. Specifically, to remove cross-modality redundancy and save bit-rate, we introduce the Channel-wise Cross-modality Entropy Model (CCEM). Within CCEM, we design the Low-frequency Context Extraction Block (LCEB) and the Low-frequency Context Fusion Block (LCFB) based on the similarity of low-frequency information between RGB and IR images. These blocks effectively capture both intra-modality and cross-modality priors, thus assisting the entropy model in predicting symbol probability estimates more accurately. Comparative experiments and ablation studies confirm the effectiveness of the proposed method.

References

[1] G. Wang, S. Zhang, J. Lu, and X. Hu, "Cross-modality paired-images generation for RGB-infrared person re-identification," Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, pp. 11837–11844, 2020.

[2] X. Zhang, J. Liu, W. Li, H. Wang, and Z. Chen, "Tfdet: Target-aware fusion for rgb-t pedestrian detection," IEEE Trans. Neural Netw. Learn. Syst., 2024, doi: 10.1109/TNNLS.2024.

[3] S. Lee, Y. Kim, J. Park, and K. Han, "INSANet: INtra-INter spectral attention network for effective feature fusion of multispectral pedestrian detection," Sensors, vol. 24, no. 4, p. 1168, 2024.

[4] Wang, Guan'an, et al. "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[5] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," arXiv preprint, arXiv:1611.01704, 2016.

[6] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 7939–7948.

[7] Y. Hu, W. Yang, and J. Liu, "Coarse-to-fine hyper-prior modeling for learned image compression," in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, pp. 11042–11049, 2020.

[8] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in Proc. Int. Conf. Learn. Represent., 2022.

[9] D. He, Z. Liu, H. Yu, Z. Ma, and J. Wang, "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 5718–5727.

[10] W. Jiang, Z. Li, F. Ma, and Y. Liu, "Mlic: Multi-reference entropy model for learned image compression," in Proc. 31st ACM Int. Conf. Multimedia, 2023, pp. 2247–2256.

[11] F. Kong, Y. Li, Z. Liu, and S. Ma, "Mixture autoregressive and spectral attention network for multispectral image compression based on variational autoencoder," Vis. Comput., vol. 1, pp. 1–24, 2023.

[12] H. Zheng and W. Gao, "End-to-end rgb-d image compression via exploiting channel-modality redundancy," in Proc. AAAI Conf. Artif. Intell., vol. 38, no. 7, 2024.

[13] G. Lu, L. Zhu, Z. Ma, and C. Yang, "Learning based multi-modality image and video compression," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 15745–15755.

[14] J. Liu, X. Wang, W. Zhang, and D. Li, "Multispectral deep neural networks for pedestrian detection," arXiv preprint, arXiv:1611.02644, 2016.

[15] T. Zhao, M. Yuan, and X. Wei, "Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion," arXiv preprint, arXiv:2401.10731, 2024.

[16] X. Jia, D. Zhang, Q. Gao, and J. Guo, "LLVIP: A visible-infrared paired dataset for low-light vision," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 15459–15468.

[17] 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. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 10012–10022.

[18] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in Adv. Neural Inf. Process. Syst., vol. 31, 2018, pp. 10771–10780.

[19] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in Proc. IEEE Int. Conf. Image Process. (ICIP), 2020, pp. 3339–3343.

[20] Z. Zhao, H. Qin, Z. Zhang, S. Liu, L. Zhang, and T. Mei, "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 20722–20732.

[21] D. Han, F. Wang, and J. Wu, "Agent attention: On the integration of softmax and linear attention," arXiv preprint, arXiv:2312.08874, 2023.

[22] D. Han, F. Wang, and J. Wu, "Flatten transformer: Vision transformer using focused linear attention," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2023, pp. 15618–15627.

[23] F. Bellard, "BPG image format," [Online]. Available: http://bellard.org/bpg/. Accessed: Oct. 30, 2018.

[24] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," VCEG-M33, ITU-T, 2001.

[25] J. Bégaint, F. Racapé, S. Feltman, and A. Pushparaja, "CompressAI: A PyTorch library and evaluation platform for end-to-end compression research," arXiv preprint, arXiv:2011.03029, 2020.