

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

Haofeng Wang<sup>1,2,5</sup>, Fangtao Zhou<sup>3</sup>, Qi Zhang<sup>2</sup>, Zeyuan Chen<sup>2,4</sup>, Enci Zhang<sup>1</sup>, Zhao Wang<sup>5,2</sup>,

Xiaofeng Huang<sup>3\*</sup>, Siwei Ma<sup>2\*</sup>

**Abstract**—RGB-IR(RGB-Infrared) 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 double. 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 RGB-IR image pair and single-modality compression methods on LLVIP and KAIST datasets. For instance, the proposed framework achieves a 23.1% bit rate saving on LLVIP dataset compared to the state-of-the-art RGB-IR image codec presented at CVPR 2022.

## I. INTRODUCTION

Recently, RGB-IR images pairs captured within the same scene have been jointly applied to various practical scenarios[1], [2], [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], [6], [7], [8], [9], [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 images that use spatial geometry, IR images capture thermal properties and are less sensitive 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 of two modalities cannot be performed simultaneously, as one modality has to be decoded at first to serve as an anchor for compressing the other, 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 to simultaneously compress RGB-IR image pairs, leveraging the correlation between modalities to save bit rate. 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. Within CCEM, 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 RGB-IR image pair and single-modality compression methods on LLVIP [16] and KAIST datasets [17].

\* Corresponding authors.

<sup>1</sup>School of Electronic and Computer Engineering, Peking University, Shenzhen, China

<sup>2</sup>National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, China

<sup>3</sup>School of Communication Engineering, Hangzhou Dianzi University, Zhejiang, China

<sup>4</sup>Pengcheng Laboratory, Shenzhen, China

<sup>5</sup>Advanced Institute of Information Technology, Peking University, Zhejiang, China

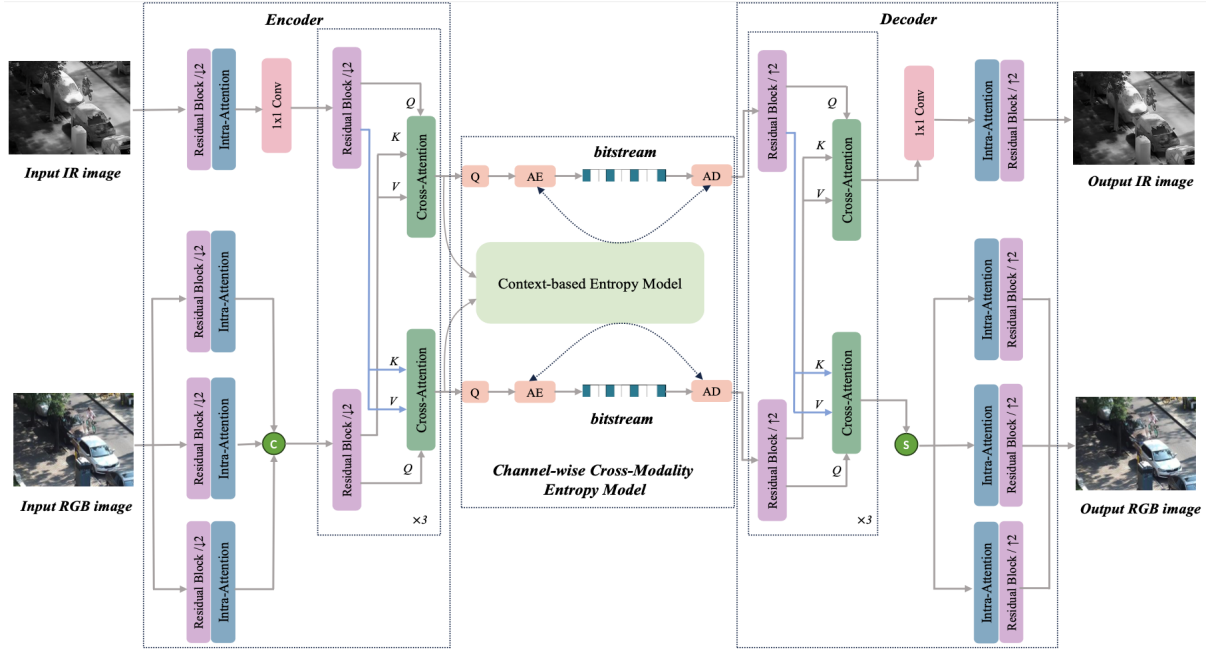


Fig. 1. The overall framework of the proposed method. The network consists of an encoder, a Channel-wise Cross-Modality Entropy Model and a decoder. AE, AD denote arithmetic encoding and decoding, respectively. Q denotes the quantizer, C and S denote concat and split operation, " $\uparrow 2$ " and "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

## 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 image is 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 [18] 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  $\theta$ ,  $\phi$  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.[19] introduced an entropy model based on spatial autoregressive prediction, surpassing the compression performance of H.265. To accelerate

decoding, another work[20] 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 predict entropy model parameters more accurately. For RGB-IR image pairs, leveraging cross-modality information, which has not been fully utilized, as prior context to enhance the accuracy of entropy model

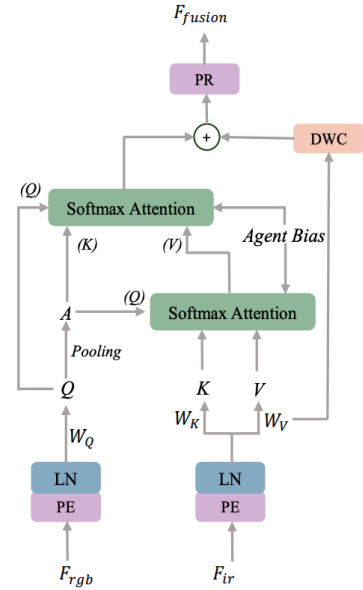


Fig. 2. The architecture of Low-frequency Context Fusion Block(LCFB). PE, PR, LN represent Patch Embedding, Patch Recovery, LayerNorm, respectively.

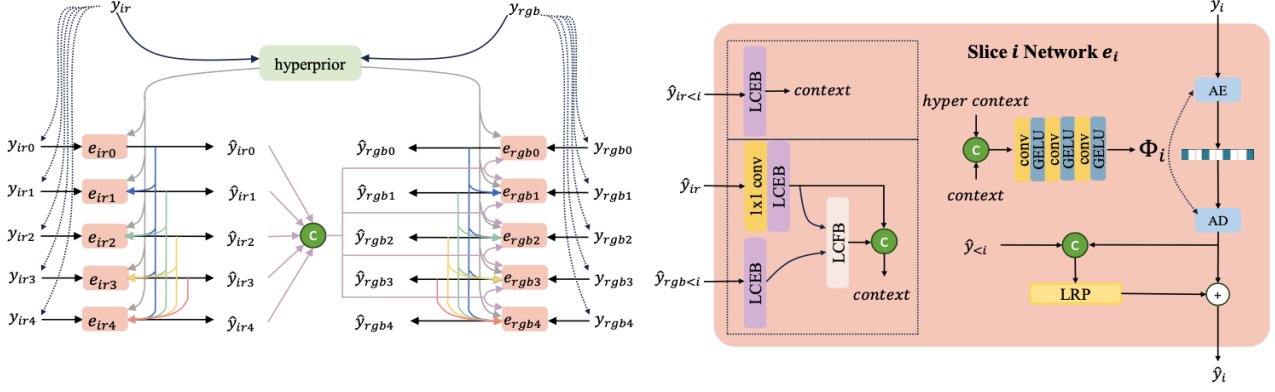


Fig. 3. The architecture of the proposed Channel-wise Cross-Modality Entropy Model. The latent representations are split into slices and sent to hyperprior model. The encoded slices are fed into Low-frequency Context Extraction Block (LCEB) and Low-frequency Context Fusion Block (LCFB) to extract global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

parameters 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[21]. 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. To verify this, We design the Low-frequency Context Extraction Block (LCEB) and Low-frequency Context Fusion Block (LCFB). The role of LCEB is to extract global low-frequency information within the modality, and since low-frequency information is typically distributed over large regions, it requires cross-regional global information exchange. Therefore, we adopt the Lite Transformer architecture[21], as it can model long-range dependencies globally, making it particularly suitable for capturing low-frequency information. As shown in Fig. 2, Instead of using a concatenation operation, we designed the LCFB based on agent-attention [22] to better aggregate global low-frequency information from two modalities. As agent-attention allows dynamic and selective weighting of information from both modalities, it enables more effective and context-aware fusion of low-frequency features, thereby enhancing the prediction of cross-modal entropy model parameters. The pipeline for processing the latent representations of the two modalities through the LCFB is as follows:

$$\begin{aligned}
 \{Q, K, V\} &= \{F_{rgb} \mathbf{W}^Q, F_{ir} \mathbf{W}^K, F_{ir} \mathbf{W}^V\}, \\
 A &= \text{Pooling}(Q), \\
 V' &= \text{softmax}\left(\frac{AK^\top}{\sqrt{d_k}} + B_1\right) V, \\
 F &= \text{softmax}\left(\frac{QA^\top}{\sqrt{d_a}} + B_2\right) V', \\
 F_{\text{fusion}} &= F + \text{DWC}(V).
 \end{aligned} \tag{2}$$

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[23]. We see the output  $F_{\text{fusion}}$  as the aggregated global low-frequency information from two modalities and use this context in entropy model to get more accurate entropy parameters.

Combining LCEB and LCFB, 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  $\{\mathbf{y}_m^0, \mathbf{y}_m^1, \dots, \mathbf{y}_m^N\}$ , 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  $i^{\text{th}}$  slice, we use the previous slices to extract context and predict entropy parameters. In particular, the slices from 1 to  $i-1$  are concatenated and processed through a 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  $j^{\text{th}}$  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 LCFB to derive cross-modality information. This additional cross-modality information is used to further improve the accuracy of the entropy model parameters 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:

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

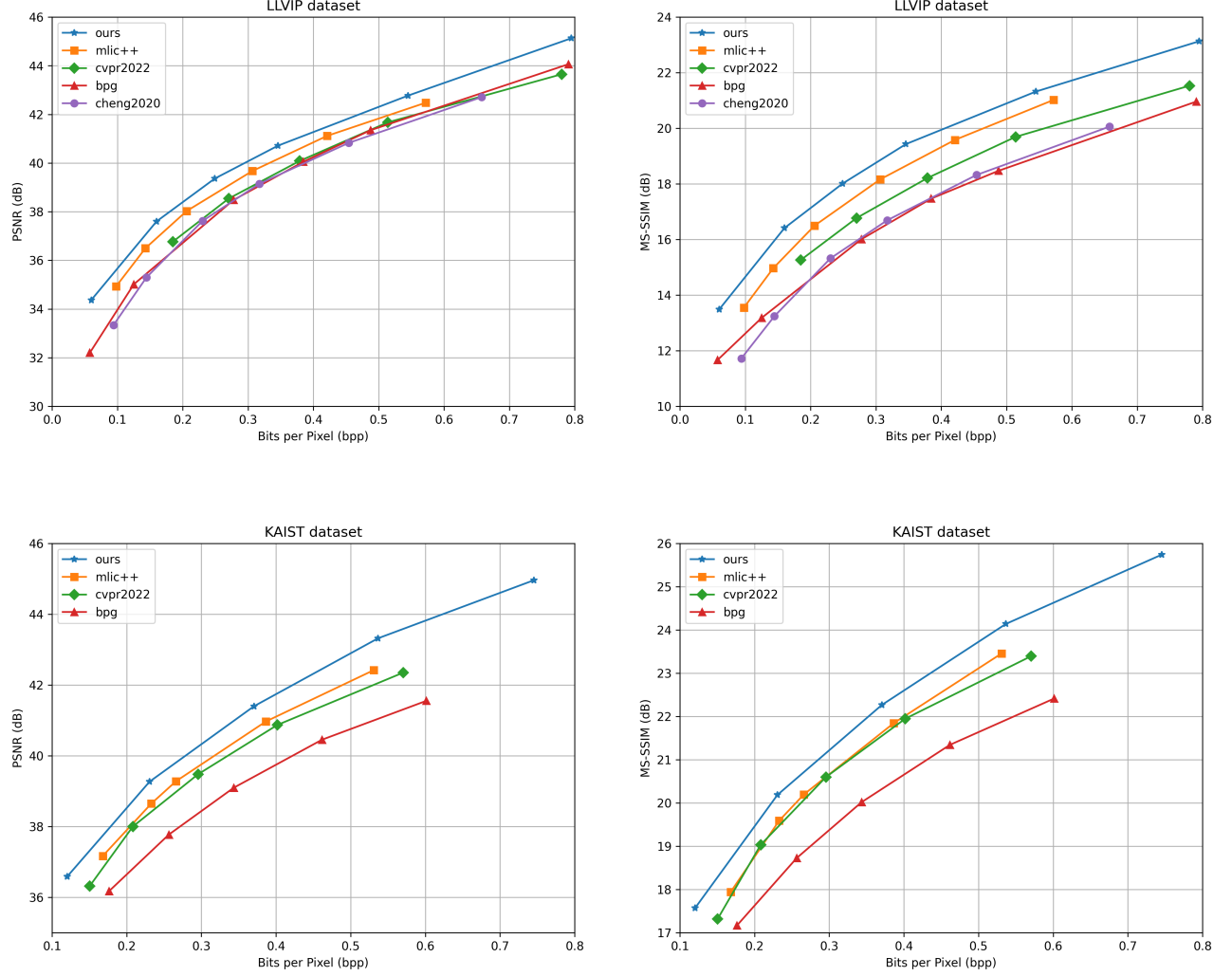


Fig. 4. Experimental results from different image compression approaches on the LLVIP and KAIST datasets.

### C. Loss Function

The loss function  $L$  of our framework is described as:

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

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 introduce state-of-the-art RGB-IR image compression method (Hereafter, referred to as CVPR2022)[13], for comparison with our model. Additionally, 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], traditional single-modality image compression method BPG[24]. To ensure a

fair comparison, we fine-tuned all end-to-end compression methods on the LLVIP and KAIST datasets. For single-modality codecs, which are primarily trained on RGB-IR images, we duplicated the single-channel IR images into three channels to preserve the original model structure during training with IR images. This approach follows the methods used in previous learning-based multimodality compression studies[12][13], where existing codecs were directly fine-tuned using IR images. Additionally, for the BPG codec, when encoding IR images, we set the pixel.format option to 0 (Grayscale). We use PSNR and MS-SSIM to assess the quality of compressed images and Bjontegaard delta rate (BD-Rate)[25] 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 all evaluation metrics are computed in the YUV420 domain.

**Training details** Considering that joint training of both modalities from the beginning would require the model to



Fig. 5. Visualization of reconstructions of 260299.png in LLVIP using different methods.

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 adopting this training method improves the model's performance by approximately 4%. Additionally, we set different hyperparameters  $\lambda$ , to control the bit rate, following the settings in CompressAI [26]. 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] and KAIST Pedestrian [17], two widely used RGB-IR datasets. For LLVIP, 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. On the KAIST Pedestrian dataset, we use set00 to set05 for training and the rest for testing. Both datasets we tested were collected under different **lighting conditions** (from day through night). Additionally, to our knowledge, the LLVIP dataset is currently the **highest resolution** (1280x1024) RGB-IR image pair dataset. The KAIST dataset contains a significant number of **high-dynamic scenes** because it includes many fast-moving objects in real-time traffic scenes.

During the two-stage training process, we followed the training schemes of [26], setting the values of  $\lambda$  to [0.0018, 0.0035, 0.0067, 0.0130, 0.0250, 0.0483]. We followed [13], treating the two modalities as equally important. Therefore, we did not introduce an additional weight in the loss function to balance the distortion between the two modalities. In the training process, images are randomly cropped to a size of  $256 \times 256$  and the batch-size is set to 4. The number of channels in the latent representation of each modality output by the Encoder is 320. Additionally, in the CCEM, the latent representation is divided into  $N = 5$  blocks. The structure of the hyperprior is consistent with that in [12].

The training goal of the initial stage training is to maintain the consistency of the overall model architecture and enhance the model's learning capability for the RGB modality during the early stages of training, we temporarily exclude the IR modality in the first stage. Thus, the loss function in the initial stage is similar to that of the second stage, except that the terms  $R_{ir}$  and  $D_{ir}$  from the second stage are removed, with all other hyperparameters remaining the same.

## B. Experiment Results

**Quantitative Results** We make a comparison of compression performance among various single-modality codecs and a state-of-the-art RGB-IR compression framework. Compared to other single-modality compression frameworks, our proposed framework shows a significant improvement in BD-rate performance. Specifically, our method outperforms CVPR2022[13] and MLIC++[10] by 23.1% and 14.6%, respectively. We plot the corresponding RD curves in Fig. 4 to more intuitively illustrate the performance gap between different codecs. The results clearly demonstrate that our proposed method significantly outperforms the other methods in terms of compression performance.

**Qualitative Results** As depicted in Fig. 5, our method exhibits superior subjective visual quality under the premise of using less bit rate. Specifically, after the local details are enlarged, our method can still retain the semantic information (such as the text on the roadside sign and the license plate number) of the original image.

**Modality-Specific Performance** As shown in Table I, the proposed method achieves significant performance improvements for **both RGB and IR modalities**. This is due to the efficient utilization of low-frequency information from both modalities in the proposed CCEM, which makes the parameter estimation in the entropy model more accurate.

**Running time and complexity** The model we proposed has 89.01M parameters, and compressing a RGB-IR image pair of size 1280x1024 requires 2.3GB of GPU memory. The FLOPs of the CCEM module reach 2.85 Mil/pixel. The model we proposed does not focus on real-time performance, but rather emphasizes Rate-Distortion (RD) performance. When tested on both datasets using a single NVIDIA 4090 machine, our model's average encoding time for an RGB-IR image pair is 881ms, and the average decoding time



TABLE I  
MODALITY-SPECIFIC BD-RATE (%) COMPARISONS ON LLVIP DATASET  
AND KAIST DATASET AGAINST BPG.

Methods	LLVIP		KAIST	
	RGB	IR	RGB	IR
Cheng2020	5.426	-1.716	10.335	-4.628
CVPR2022	-10.733	-3.262	-18.639	-21.289
MLIC++	-9.463	-18.034	-16.761	-25.622
Ours	<b>-19.89</b>	<b>-35.051</b>	<b>-27.8236</b>	<b>-39.001</b>

is 942ms. In comparison, prior sequential approaches, such as CVPR2022[13], reach an encoding time of 697ms and a decoding time of 601ms. Anchor-based methods, like MLIC++[10], reach an encoding time of 901ms and a decoding time of 978ms. While our method introduces slightly more running time costs, it achieves a significant improvement in RD performance.

**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 II shows that both proposed modules contribute to BD-rate performance, and our proposed CCEM significantly enhances compression efficiency.

TABLE II  
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

### C. Other Exploration and Analysis

In our CCEM, we extract contextual information from the decoded IR features through a series of transformations to assist in entropy parameters prediction for RGB features. In fact, we also conducted comparative experiments where RGB features were used to assist in entropy parameters prediction for IR features. Ultimately, on the LLVIP dataset, the former approach achieved approximately 8% bit rate savings compared to the latter.

To investigate the underlying reasons, we visualized the IR and RGB features in the entropy models of the two approaches. As illustrated in the second line of Fig. 6, the RGB features obtained with the assistance of decoded IR features exhibit more distinct structural information compared to the features without such assistance. Additionally, some regions experience texture enhancement, enabling more accurate entropy parameters prediction. This improvement is attributed to the low-frequency information extracted by our designed LCEB. Conversely, as illustrated in the first line of Fig. 6, when using the decoded IR features to assist the RGB feature, the resulting feature contain richer texture

information compared to the features without assistance. However, this approach also introduces noise in certain regions (as marked in the figure), which interferes with the accuracy of entropy parameters prediction.

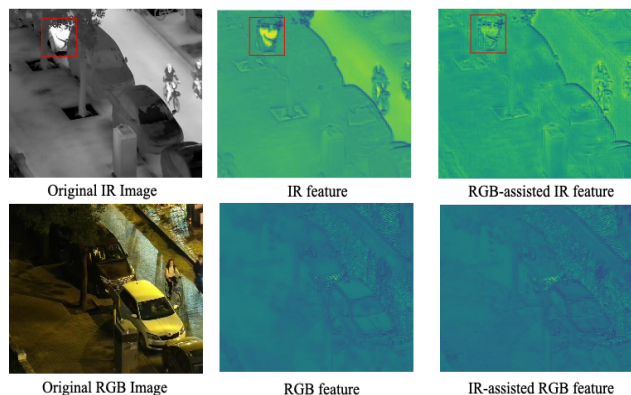


Fig. 6. Visualization of the RGB and IR feature in CCEM

## IV. CONCLUSIONS

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.-A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z.-G. Hou, "Cross-modality paired-images generation for rgb-infrared person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 144–12 151.
- [2] X. Zhang, X. Zhang, J. Wang, J. Ying, Z. Sheng, H. Yu, C. Li, and H.-L. Shen, "Tfdet: Target-aware fusion for rgb-t pedestrian detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [3] S. Lee, T. Kim, J. Shin, N. Kim, and Y. Choi, "Insanet: Intra-inter spectral attention network for effective feature fusion of multispectral pedestrian detection," *Sensors*, vol. 24, no. 4, p. 1168, 2024.
- [4] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3623–3632.
- [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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7939–7948.
- [7] Y. Hu, W. Yang, and J. Liu, "Coarse-to-fine hyper-prior modeling for learned image compression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 013–11 020.
- [8] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *International Conference on Learning Representations*, 2022.

- [9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5718–5727.
- [10] W. Jiang, J. Yang, Y. Zhai, P. Ning, F. Gao, and R. Wang, "Mlic: Multi-reference entropy model for learned image compression," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7618–7627.
- [11] F. Kong, G. Ren, Y. Hu, D. Li, and K. Hu, "Mixture autoregressive and spectral attention network for multispectral image compression based on variational autoencoder," *The Visual Computer*, vol. 40, no. 9, pp. 6295–6318, 2024.
- [12] H. Zheng and W. Gao, "End-to-end rgb-d image compression via exploiting channel-modality redundancy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7562–7570.
- [13] G. Lu, T. Zhong, J. Geng, Q. Hu, and D. Xu, "Learning based multi-modality image and video compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6083–6092.
- [14] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.
- [15] T. Zhao, M. Yuan, F. Jiang, N. Wang, 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, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvp: A visible-infrared paired dataset for low-light vision," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3496–3504.
- [17] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "Kaist multi-spectral day/night data set for autonomous and assisted driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.
- [18] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [19] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, 2018.
- [20] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3339–3343.
- [21] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5906–5916.
- [22] D. Han, T. Ye, Y. Han, Z. Xia, S. Pan, P. Wan, S. Song, and G. Huang, "Agent attention: On the integration of softmax and linear attention," in *European Conference on Computer Vision*. Springer, 2025, pp. 124–140.
- [23] D. Han, X. Pan, Y. Han, S. Song, and G. Huang, "Flatten transformer: Vision transformer using focused linear attention," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 5961–5971.
- [24] F. Bellard, "Bpg image format," Available: <http://bellard.org/bpg/>, 2018, accessed: Oct. 30, 2018.
- [25] G. Bjontegaard, "Calculation of average psnr differences between rd-curves," ITU-T, Tech. Rep. VCEG-M33, 2001.
- [26] 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.