End-to-End RGB-D Image Compression via Exploiting Channel-Modality Redundancy

Huiming Zheng[1, 2], **Wei Gao**[1, 2*]

[1] School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen, China
[2] Peng Cheng Laboratory, China
hmzheng@stu.pku.edu.cn, gaowei262@pku.edu.cn

Abstract

As a kind of 3D data, RGB-D images have been extensively used in object tracking, 3D reconstruction, remote sensing mapping, and other tasks. In the realm of computer vision, the significance of RGB-D images is progressively growing. However, the existing learning-based image compression methods usually process RGB images and depth images separately, which cannot entirely exploit the redundant information between the modalities, limiting the further improvement of the Rate-Distortion performance. With the goal of overcoming the defect, in this paper, we propose a learning-based dual-branch RGB-D image compression framework. Compared with traditional RGB domain compression scheme, a YUV domain compression scheme is presented for spatial redundancy removal. In addition, Intra-Modality Attention (IMA) and Cross-Modality Attention (CMA) are introduced for modal redundancy removal. For the sake of benefiting from cross-modal prior information, Context Prediction Module (CPM) and Context Fusion Module (CFM) are raised in the conditional entropy model which makes the context probability prediction more accurate. The experimental results demonstrate our method outperforms existing image compression methods in two RGB-D image datasets. Compared with BPG, our proposed framework can achieve up to 15% bit rate saving for RGB images.

Introduction

RGB-D images are an important 3D data format. It has been widely used in 3D scene reconstruction (Zollhöfer et al. 2018), salient object detection (Liao et al. 2020; Gao et al. 2021), robotics and autonomous navigation, medical imaging and health monitoring, environmental monitoring, and other fields. Unlike RGB images, depth images contain information about the distance to the surface of the scene object from the viewpoint, which provides depth information among 3D scenes. Therefore, the RGB-D joint analysis methods are popular in computer vision tasks. However, these methods (Bozic et al. 2020; Ji et al. 2020; Liu, Zhang, and Han 2020; Shi et al. 2022; Tian et al. 2020; Xiang et al. 2021) use additional modality, which will bring supernumerary storage and transmission costs. Therefore, designing an efficient RGB-D image compression method is an important and challenging work.

Deep learning-based image compression has been developing for several years. Numerous works (Minnen and Singh 2020; Gao et al. 2020; Wu et al. 2021; Lee et al. 2022; Wu and Gao 2022; Zhu et al. 2022; Tao et al. 2023; Jiang et al. 2023) have been put forth to improve rate-distortion performance and optimize coding framework. In addition, some open-source algorithm libraries (Bégaint et al. 2020; Gao et al. 2023) also efficiently promote the prosperity of the field. However, existing methods focus on single image compression, ignoring the direct interactivity of RGB and depth modalities. Modality redundancy is not adequately considered, limiting rate-distortion performance improvement. Besides, knowledge-guided compression is one of the most relevant topics. The coding framework can use additional information from the data source itself or analyze additional information from its own modules to better eliminate redundancy. Stereo image compression framework (Deng et al. 2021) adopts homography transformation to remove view redundancy. Light field image compression framework (Zhao et al. 2018) leverages the inherent similarity of light field images to remove the redundancy of different perspectives. 360° image compression framework (Li et al. 2022) utilizes a latitude adaptive coding scheme to allocate variant numbers of bits for different regions. Although these methods explore modality redundancy removal to some extent, they can not achieve a higher compression ratio in the RGB-D image compression owing to the significant difference in the distribution between RGB images and depth images. Therefore, it is necessary to develop a compression framework dedicated to RGB-D images.

In this paper, we proposed an efficient learning-based RGB-D image compression network by exploiting the redundant information between the modalities and channels. Most learning-based methods usually sample and compress images in RGB domain, while our method chooses to sample images in YUV domain in order to remove spatial redundancy in transform domain for depth images. In addition, we design Intra-Modality Attention (IMA) in feature extraction module and Cross-Modality Attention (CMA) in the main encoder module to eliminate channel redundancy and modality redundancy separately. We adopt Context Prediction Module (CPM) and Context Fusion Module (CFM)

*Corresponding author

in the conditional entropy model to sufficiently excavate the coherence between the two modalities and utilize cross-modality prior information, which provides more accurate probability prediction information for entropy coder. Experimental results prove that our proposed network attains better rate-distortion performance on several widely used RGB-D datasets compared with the single image compression methods. The contributions of our proposed method can be summarized as follows:

- We propose a learning-based RGB-D image compression framework by exploiting the redundant information between the channels and modalities. The framework is conducted in the YUV domain rather than RGB domain, which is conducive to the elimination of spatial redundancy for depth images.
- Intra-Modality Attention and Cross-Modality Attention are designed to remove cross-channel redundancy and cross-modality redundancy for higher compression ratio. To be specific, multi-head self attention and multi-head cross attention are integrated into the module for more efficient cross-channel and cross-modality information interaction.
- Conditional Context-based Entropy Model is adapted to reveal the dependency between the modalities. In addition, Context Prediction Module and Context Fusion Module are elaborately designed for efficient probability prediction.
- According to the experimental results, our proposed framework achieves SOTA performance when compared with existing image compression methods in two RGB-D image datasets.

Related Work

Deep Learning-based Image Compression

Over the recent years, with the development of deep learning, numerous works (Ballé, Laparra, and Simoncelli 2016; Ballé et al. 2018; Minnen, Ballé, and Toderici 2018; Toderici et al. 2015) focus on learning-based image compression have been presented. Ballé et al. (Ballé, Laparra, and Simoncelli 2016) proposed the most widely used lossy image compression framework. The model is constructed based on an autoencoder. The final rate-distortion (RD) performance surpassed JPEG and JPEG-2000. On this basis, Ballé et al. (Ballé et al. 2018) then introduced hyperprior model, which captured spatial redundancy between feature maps. Minnen et al. (Minnen, Ballé, and Toderici 2018) utilized single PixelCNN (Van Den Oord, Kalchbrenner, and Kavukcuoglu 2016) layer to model autoregressive priors and combined autoregressive priors with hyperprior to achieve better symbol probability prediction. The above frameworks are based on Convolutional Neural Network (CNN). In addition, Toderici et al. (Toderici et al. 2015) proposed a variable rate compression framework based on Recurrent Neural Network (RNN). Although it is feasible to extend the above methods to RGB-D image compression frameworks, most of these methods are devoted to single modality redundancy removal. The further improvement of compression efficiency is limited. It is urgent to design a novel compression framework for RGB-D images.

Stereo Image Compression

With the continuous improvement of 3D vision technology, stereo image compression has become one of the hot topics in the image compression field. A great deal of works (Liu, Wang, and Urtasun 2019; Huang et al. 2021; Wödlinger et al. 2022; Deng et al. 2021) have sprung up in recent years. Liu et al. (Liu, Wang, and Urtasun 2019) first proposed a deep learning-based stereo image compression network (DSIC). In this work, a parameter skip function is proposed to exploit the dependencies between two perspectives. Wödlinger et al. (Wödlinger et al. 2022) proposed a scheme to compress images by exploiting the similarity of the stereo images. The right image utilized the latent shifting information from the encoded left image for extreme bit rate savings. Deng et al. (Deng et al. 2021) introduced a homography estimation based stereo image compression network, called HESIC. To map the left image to the right image, the homography matrix is adapted to achieve homologous transformation. The residual information from different views is encoded. The above approaches take full advantage of the similarities between the left and right views. However, for RGB-D images, similar features between modalities are more difficult to extract. In addition, some transformation methods, such as homologous transformation and affine transformation, are probably not suitable for RGB-D images. Therefore, targeted works are still urgently needed in the learning-based RGB-D compression field.

Image Compression with Attention Mechanism

Attention mechanisms have been introduced to image compression for a long time. In self attention mechanism, each input pixel interacts with the others to calculate its dependencies with the others, and then uses the dependencies to allocate different weights to each position. Cheng et al. (Cheng et al. 2020) first used simplified non-local attention module and integrated it into the network architecture to improve performance. Chen et al. (Chen et al. 2021) embedded more efficient non-local attention module into the whole framework for reducing time and space complexity, and used attention mechanism to generate implicit masks for adaptive bit allocation. Zou et al. (Zou, Song, and Zhang 2022) introduced a simpler and more efficient window-based local attention block, which can fully take advantage of global structures and local textures in the transformer-based structure. The above methods typically use attention mechanisms in a single modality. The redundancy information between modalities is not fully explored. Therefore, it is necessary to design new attention modules to eliminate the redundancy between RGB and depth modalities.

Methodology

Overview

The overall architecture of our RGB-D image compression framework is presented in Fig. 1. The network is based on transformer architecture (Liu et al. 2021). The input RGB

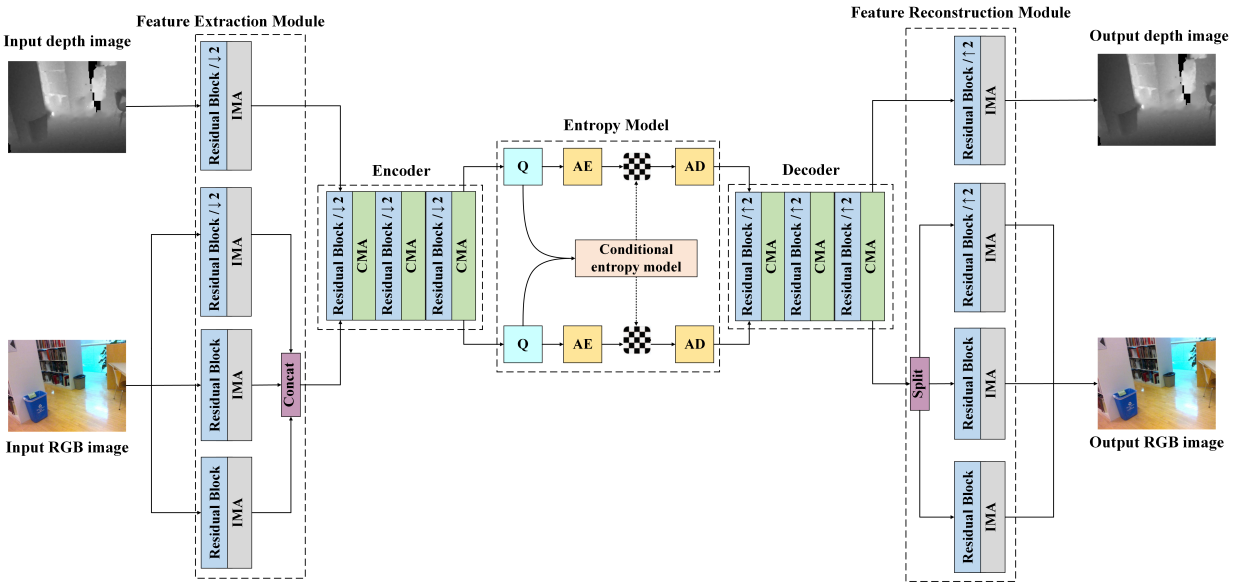

Figure 1: The overall network architecture of the proposed method. The input depth image and input RGB images are splitted into four channels. The framework consists of feature extraction module, encoder, entropy model, decoder and feature reconstruction module. Here, AE donates arithmetic encoding, AD donates arithmetic decoding, Q donates the quantizer, "↑ 2" represents upsampling by a factor of two, "↓ 2" represents downsampling by a factor of two.

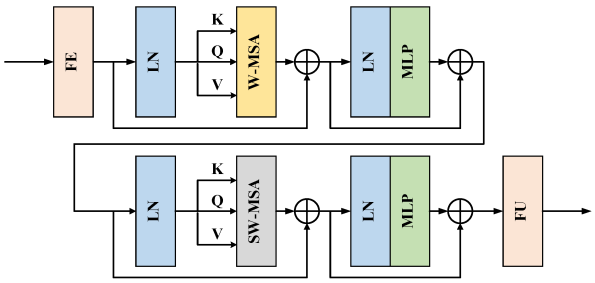

Figure 2: The architecture of Intra-Modality Attention (IMA).

and depth images are transformed into 4 channels in the YCbCr subsampled color space. The weight and height of U and V channels are half of Y's weight and height in RGB images. The depth images only retain Y channel information. We donate y, u, v, d as the input channels. First, the input channels are fed into feature extraction module to eliminate channel redundancy. The feature maps y^{ex}, u^{ex}, v^{ex}, d^{ex} are obtained from y, u, v, d respectively after feature extraction. Then we concat y^{ex}, u^{ex}, v^{ex} for the next stage input yuv^{ex}. In the encoder stage (analysis transform), a dual-branch network is presented for the input yuv^{ex} and d^{ex}. The proposed Cross-Modality Attention allows the latent representations to learn cross-modality information from each other. After the encoder stage, the latent representations yuv^a and d^a are sent to quantizer. The quantized latent representations $\widehat{yuv^a}$ and $\widehat{d^a}$ are then sent into the conditional entropy model for accurate symbol probability prediction. In the decoder side (synthesis transform), $\widehat{yuv^a}$ and $\widehat{d^a}$ are fed into the dual-branch decoder framework for feature restoration and upsampling. Feature maps yuv^s and d^s are obtained after the decoding process. At last, in the feature reconstruction module, the feature map yuv^s are splited into Y,U,V channels y^{re}, u^{re}, v^{re}. Detail restoration and reconstruction are conducted in the feature reconstruction module. We donate the feature extraction module, encoder, quantizer, decoder, feature reconstruction module as $E(\cdot)$, $g_a(\cdot)$, $Q(\cdot)$, $g_s(\cdot)$, $R(\cdot)$, respectively. The main encoding-decoding process except hyperprior can be formulated as:

$$\begin{aligned} i^{ex} &= E(i), \\ i^a &= g_a(i^{ex}), \\ \widehat{i^a} &= Q(i^a), \\ i^s &= g_s(\widehat{i^a}), \\ i^{re} &= R(i^s), \end{aligned} \tag{1}$$

where i represents one of the input y, u, v, d.

Intra-Modality Attention

In our proposed framework, we use Intra-Modality Attention in the feature extraction module and feature reconstruction module to reduce the channel redundancy. The framework of IMA is shown in Fig. 2. The main framework is based on two successive Swin Transformer Blocks (Liu et al. 2021).

Given an input feature map with the dimensions $H \times W \times C$, the window-based attention first reshapes the feature map to the size of $\frac{HW}{M^2} \times M^2 \times C$, while M represents the window size. $\frac{HW}{M^2}$ windows are obtained from the operation. Then, self-attention is adopted to each window. Three learn-

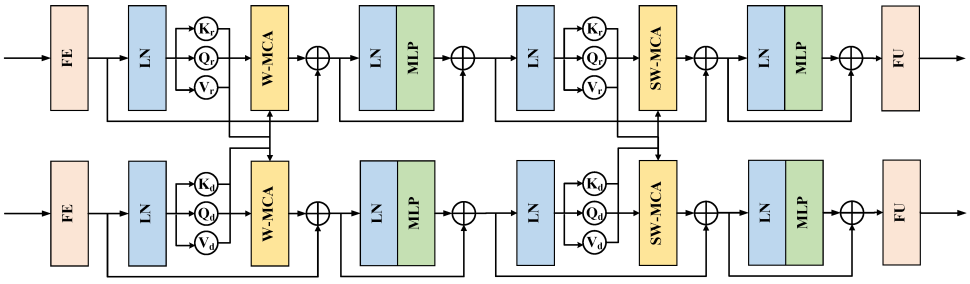

Figure 3: The architecture of Cross-Modality Attention (CMA).

able weight matrices $\mathbf{W}^Q$, $\mathbf{W}^K$, $\mathbf{W}^V$ shared the same parameters are multiplied to the local feature map F, in order to get query Q, key K, and value V, respectively. The process can be described as:

$$\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} = \left\{F\mathbf{W}^Q, F\mathbf{W}^K, F\mathbf{W}^V\right\}. \quad (2)$$

Then, the attention function calculates the dot-product of the query with each of the keys. The result includes a relative position bias for better computational complexity. A softmax operator is adopted to normalize the result for attention scores. The above process can be defined as:

$$\text{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{B}\right)\mathbf{V}, \quad (3)$$

where d is the dimension, B is the relative position bias. The main process of Intra-Modality Attention can be domenstrated as:

$$\begin{aligned}
\hat{\mathbf{z}}^l &= \text{W-MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)(Q, K, V)\right) + \mathbf{z}^{l-1}(Q), \\
\mathbf{z}^l &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^l\right)\right) + \hat{\mathbf{z}}^l, \\
\hat{\mathbf{z}}^{l+1} &= \text{SW-MSA}\left(\text{LN}\left(\mathbf{z}^l\right)(Q, K, V)\right) + \mathbf{z}^l(Q), \\
\mathbf{z}^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1},
\end{aligned} \quad (4)$$

where $\hat{\mathbf{z}}^l$ and $\mathbf{z}^l$ are the output features of (S)W-MSA and MLP blocks respectively. $\mathbf{z}^{l-1}$ is the input feature map. $\text{LN}(\cdot)$ is the LayerNorm function. $\text{W-MSA}(\cdot)$ represents the window based multi-head self attention, and $\text{SW-MSA}(\cdot)$ represents the shifted-window based multi-head self attention.

Cross-Modality Attention

Following the Intra-Modality Attention, we also design Cross-Modality Attention. The network architecture is shown in Fig. 3. Different from IMA that removes channel redundancy, CMA devotes to eliminating modality redundancy. In addition, CMA can further integrate the queries between different modalities. The framework of IMA and CMA are similar, the main difference is that CMA adopts multi-head cross attention rather than multi-head self attention to achieve information interaction between modalities. Given the input RGB features map $\mathbf{z}_r^{l-1}$ and depth feature map $\mathbf{z}_d^{l-1}$ in local windows, the complete process of the Cross-Modality Attention adapted to $\mathbf{z}_r^{l-1}$ can be defined as:

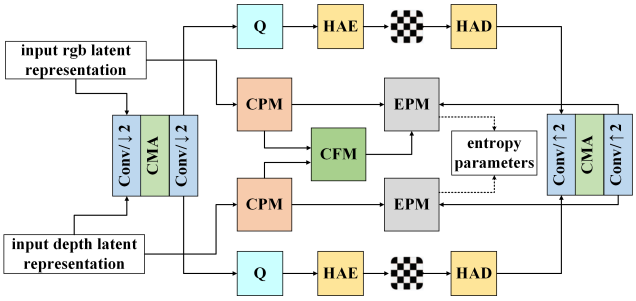

Figure 4: The architecture of Conditional Context-based Entropy Model.

$$\begin{aligned}
\hat{\mathbf{z}}_r^l &= \text{W-MCA}\left(\text{LN}\left(\mathbf{z}_r^{l-1}\right)(Q_r, K_d, V_d)\right) \\
&\quad + \mathbf{z}_r^{l-1}(Q_r), \\
\mathbf{z}_r^l &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}_r^l\right)\right) + \hat{\mathbf{z}}_r^l, \\
\hat{\mathbf{z}}_r^{l+1} &= \text{SW-MCA}\left(\text{LN}\left(\mathbf{z}_r^l\right)(Q_r, K_d, V_d)\right) \\
&\quad + \mathbf{z}_r^l(Q_r), \\
\mathbf{z}_r^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}_r^{l+1}\right)\right) + \hat{\mathbf{z}}_r^{l+1},
\end{aligned} \quad (5)$$

while the complete process of the Cross-Modality Attention adapted to $\mathbf{z}_d^{l-1}$ can be described as:

$$\begin{aligned}
\hat{\mathbf{z}}_d^l &= \text{W-MCA}\left(\text{LN}\left(\mathbf{z}_d^{l-1}\right)(Q_d, K_r, V_r)\right) \\
&\quad + \mathbf{z}_d^{l-1}(Q_d), \\
\mathbf{z}_d^l &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}_d^l\right)\right) + \hat{\mathbf{z}}_d^l, \\
\hat{\mathbf{z}}_d^{l+1} &= \text{SW-MCA}\left(\text{LN}\left(\mathbf{z}_d^l\right)(Q_d, K_r, V_r)\right) \\
&\quad + \mathbf{z}_d^l(Q_d), \\
\mathbf{z}_d^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}_d^{l+1}\right)\right) + \hat{\mathbf{z}}_d^{l+1},
\end{aligned} \quad (6)$$

where $\hat{\mathbf{z}}_r^l$ and $\hat{\mathbf{z}}_d^l$ represent the output RGB feature map and depth feature map of (S)W-MCA. $\mathbf{z}_r^l$ and $\mathbf{z}_d^l$ are the output features of MLP blocks. $\text{W-MCA}(\cdot)$ represents the window based multi-head cross attention, and $\text{SW-MCA}(\cdot)$ represents the shifted-window based multi-head cross attention.

Conditional Context-based Entropy Model

Traditional single image compression methods (Ballé et al. 2018; Cheng et al. 2020; Chen et al. 2021) usually utilize

hyperprior information as conditional prior. The probability density of one spatial location can be estimated by the known probability density of other locations. But for RGB-D images with cross-modality information, it is not enough for hyperprior to provide accessional information. In our proposed method, we adopt Conditional Context-based Entropy Model for more accurate probability estimation. The architecture of Conditional Context-Based Entropy Model is shown in Fig. 4. After the encoder stage, the latent representation is sent to hyper encoder and hyper decoder for spatial distribution information. Besides, it is also fed to Context Prediction Module (CPM) for context prior information. The output feature maps of CPM are then sent to Context Fusion Module (CFM) for cross-modality information aggregation. For depth latent representation, we estimate the entropy parameters using context and spatial priors. For more complex RGB latent representation, in addition to the former, we use supernumerary cross-modality information to improve probability prediction accuracy. To be specific, we donate $\tilde{y}_d$ as the likehoods of the depth latent representations and $\tilde{y}_r$ as the likehoods of the RGB latent representations. $\tilde{y}_d^i$ and $\tilde{y}_r^i$ represent the i-th element in $\tilde{y}_d$ and $\tilde{y}_r$. The estimated probability mass functions (PMFs) $q_{\tilde{\mathbf{y}}_d|\tilde{\mathbf{z}}_d}$ and $q_{\tilde{\mathbf{y}}_r|\tilde{\mathbf{y}}_d,\tilde{\mathbf{z}}_r}$ are shown in Eq. (7).

$$
\begin{aligned}
q_{\tilde{\mathbf{y}}_d|\tilde{\mathbf{z}}_d}\left(\tilde{\mathbf{y}}_d \mid \tilde{\mathbf{z}}_d\right) &= \sum_i q_{\tilde{y}_d^i|\tilde{y}_d^{<i}\tilde{\mathbf{z}}_d}\left(\tilde{y}_d^i \mid \tilde{y}_d^{<i}, \tilde{\mathbf{z}}_d\right), \\
q_{\tilde{\mathbf{y}}_r|\tilde{\mathbf{y}}_d,\tilde{\mathbf{z}}_r}\left(\tilde{\mathbf{y}}_r \mid \tilde{\mathbf{y}}_d, \tilde{\mathbf{z}}_r\right) &= \sum_i q_{\tilde{y}_r^i|\tilde{y}_r^{<i},\tilde{\mathbf{y}}_d,\tilde{\mathbf{z}}_r}\left(\tilde{y}_r^i \mid \tilde{y}_r^{<i}, \tilde{\mathbf{y}}_d, \tilde{\mathbf{z}}_r\right).
\end{aligned} \tag{7}
$$

Context Prediction Module and Context Fusion Module

In order to further model PMFs, context prediction module is adapted to accurately estimate context prior information. Mask Scaled Cosine Attention (MSCA) is adopted in the context prediction module. In addition, we proposed the context fusion module instead of concat operation to better aggregate cross-modality information. Mask Scaled Cross Cosine Attention (MSCCA) is integrated into the context fusion module in order to achieve information interaction between modalities. In order to ensure the serial encoding-decoding order, we use look ahead mask mechanism (Alcorn and Nguyen 2021) in the transformer architecture. Rather than scaled dot self attention, we adopt scaled cosine attention which makes the training of the model more stable. In addition, log-space continuous relative position bias is used instead of linear-space relative position bias for better reconstruction quality aiming at high-resolution images.

Loss Function

During the training phase, the loss function L is described as follows:

$$
L = R_r + R_d + \lambda(D_r + D_d), \tag{8}
$$

where D_r and D_d are the weighted mean-squared error (MSE) of YUV channels and depth channel. They can be formulated as:

$$
\begin{aligned}
D_d &= \mathrm{MSE_d}, \\
D_r &= (4\mathrm{MSE_y} + \mathrm{MSE_u} + \mathrm{MSE_v}))/6.
\end{aligned} \tag{9}
$$

R_r and R_d denote the bit rate cost, which can be calculated by the likelihood of latent representations. According to the configuration of YUV420 color domain, the weighting ratio between Y, U, and V is 4:1:1.

Experiments

Datasets

SUN-RGBD The SUN-RGBD dataset (Song, Lichtenberg, and Xiao 2015) is a widely used computer vision research dataset for indoor scene understanding and depth perception tasks. The dataset provides data such as RGB images, depth images, and semantic segmentation labels in indoor environments, and is suitable for lots of different computer vision tasks. The dataset contains 10,000 RGB-D images. For training, 8,000 image pairs were randomly selected, while 1,000 image pairs were chosen for validation, and an additional 1,000 image pairs were reserved for testing.

NYU-Depth V2 NYU-Depth V2 dataset (Chodosh, Wang, and Lucey 2019) comprises video sequences capturing diverse indoor scenes recorded by the RGB and depth cameras of Microsoft Kinect. It includes 1,449 annotated RGB images and depth images. The images are from 464 scenes in three cities. We divide the entire dataset into three parts, 1,159 image pairs for training, 145 image pairs for validating, and 145 image pairs for testing.

Experimental Details

Training Strategy We train the whole network jointly. The proposed network is based on the CUDA-enabled PyTorch implementation. We set different values for the hyperparameter λ to control the bit rate. The λ configuration is referred to CompressAI (Bégaint et al. 2020). Adam optimizer (Kingma and Ba 2014) is adopted in the training process. We initialize the learning rate to $1e-4$. It gradually decreases with the update of the model during training and eventually falls to $1e-5$. The batch size is set to 4. We train about 1000 epochs for each model. It costs about ten days for the training stage according to Tesla V100. The input training data is trimmed to the size of 256×256 convenient for model inference. The training data is mainly based on SUN-RGBD dataset. When the model is tested on the NYU-Depth V2 dataset, we finetune the pretrain model using training dataset in NYU-Depth V2 dataset for about 100 epochs.

Evaluation Metric We adopt PSNR as the evaluation metrics. PSNR is an objective metric to evaluate image quality, which reflects the signal fidelity of an image. Additionally, we compare the Bjontegaard delta rate (BD-Rate) (Bjontegaard 2001) in order to obtain the quantitative rate-distortion performance. Noted that the PSNR and BD-Rate metrics are evaluated in YUV420 domain.

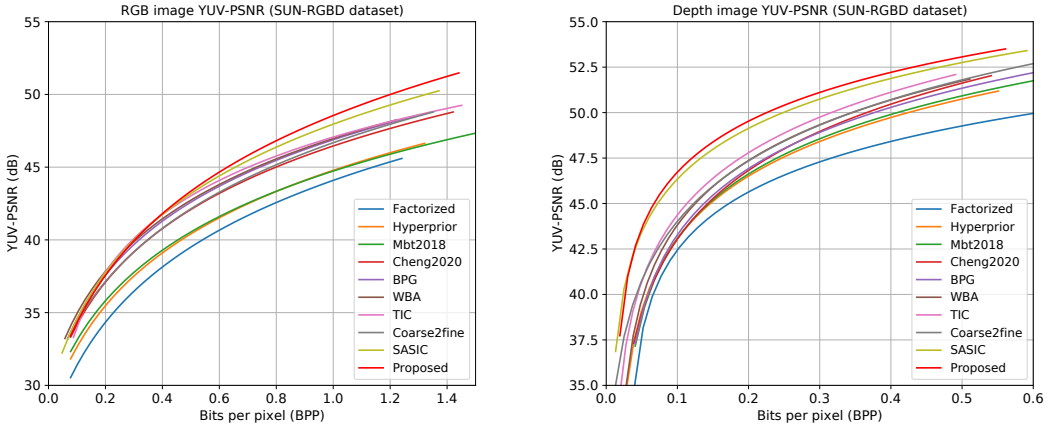

Figure 5: Rate-distortion curves for RGB images (left) and depth images (right) tested in SUN-RGBD dataset.

Methods	SUN-RGBD		NYU-Depth V2	
	RGB	Depth	RGB	Depth
Factorized	76.871	29.656	67.362	21.290
Hyperprior	46.893	15.706	38.214	9.816
Mbt2018	41.027	9.762	28.963	6.372
Cheng2020	8.753	-1.727	9.242	0.004
WBA	-4.623	-7.962	-6.352	-9.251
TIC	-1.555	-16.809	-3.375	-12.378
Coarse2fine	5.306	-18.567	3.957	-19.619
SASIC	-9.310	-25.918	-8.118	-30.462
Proposed	**-15.717**	**-36.244**	**-12.376**	**-38.971**

Table 1: BD-Rate (%) comparisons on SUN-RGBD dataset and NYU-Depth V2 dataset against BPG. The bold numbers represent the optimal results in this classification.

Baseline We compare our method against serval well-performing single image methods (WBA (Zou, Song, and Zhang 2022), TIC (Lu et al. 2021), Coarse2Fine (Hu, Yang, and Liu 2020)), a stereo image compression method (SASIC (Wödlinger et al. 2022)) and some classic learning-based methods (Factorized (Ballé, Laparra, and Simoncelli 2016), Hyperprior (Ballé et al. 2018), Mbt2018 (Minnen, Ballé, and Toderici 2018), Cheng2020attention (Cheng et al. 2020)). In addition, traditional single-modality image compression method BPG (Bellard 2018) is also compared with our proposed framework.

Experiment Results

Quantitative Results Table 1 presents the coding performance of the methods against BPG in two datasets. The BD-Rate value is negative, indicating that the coding performance of this algorithm is better than that of the benchmark algorithm. Otherwise, it is worse than the benchmark algorithm. To ensure a fair comparison, we employ the same training dataset and training methods as this model to train other learning-based methods. It is evident that our proposed method attains the the best RD performance. In comparison with the single image compression methods, the RD performance of our proposed method is significantly improved. To be specific, our approach offers over 10% gains against BPG on BD-Rate metric for RGB images in both datasets. In addition, we plot the RD curve to further visualize the performance gap between the various methods. Fig. 5 shows the YUV-RSNR result for RGB images and depth images in SUN-RGBD dataset. It indicates that our proposed framework surpassed other frameworks, showcasing the best RD performance. Besides, from Fig. 5, it is obvious that the compression effect of the model on depth images is obviously better than that on RGB images.

Model	BD-Rate(%)
Baseline	-
Baseline + Conditional Context-based Entropy Model	-7.56
Baseline + Proposed CPM	-4.18
Baseline + Proposed CFM	-2.35

Table 2: Ablation study of each component in conditional entropy model. Our entropy model is based on Mbt2018.

Qualitative Results In order to show the compression effect of each model more intuitively, we visualize the compressed image of each model in Fig. 6. It is important to note that, for the sake of fairness, we try to keep all models compress at the same bit rate. As depicted in Fig. 6, our method exhibits superior subjective visual quality under the premise of using less bit rate. After the local details are enlarged, our method can still retain the semantic information (such as the letters in the figure) of the original image.

Running Time and Complexity The number of our proposed model parameters is 69.03 M. For a RGB-D image pair with the input resolution of 256×256, the FLOPs reach 6.93 Mil/pixel. When we test our proposed model in both two datasets on the Tesla V100, the average encoding time

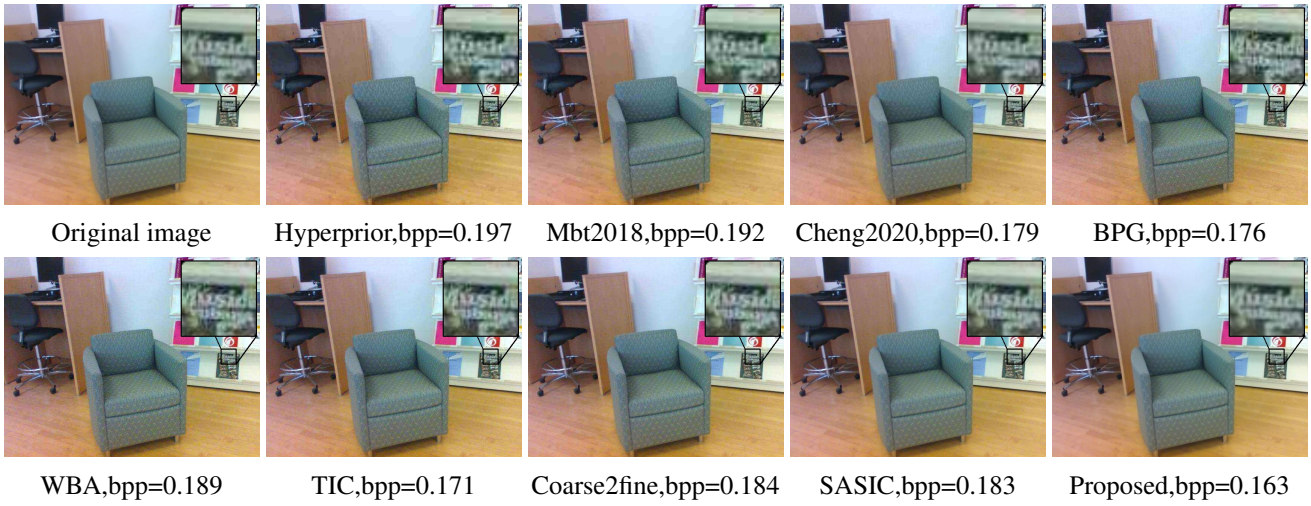

Figure 6: Visual quality comparison for the RGB image compression results.

Model	BD-Rate(%)
Baseline	-
Baseline + IMA	-3.98
Baseline + CMA	-5.16
Baseline + IMA + CMA	-6.71

Table 3: Ablation study of IMA and CMA. It should be noted that after removing the CMA, the upper and lower branches no longer have information interaction in the encoder and decoder. Therefore, we can change the dual-branch structure of this model to a single-branch structure.

and decoding time are 11.696 s and 8.582 s, respectively. Compared with other learning-based model, our method introduces additional computational cost, but obtains significant rate-distortion performance gain.

Ablation Study and Analysis

Case 1: Effectiveness of conditional entropy model. As illustrated in Table 2, We verify the validity of each module in the entropy model through substitution. It is conducted on the SUN-RGBD dataset. We can find that each module contributes to enhancing the overall coding performance. In addition, it is noticed that conditional context-based entropy model contributes most to the RD performance.

Case 2: Effectiveness of YUV domian compression. In order to verify that for RGB-D images, compression in YUV domain is more efficient, compared with the proposed framework, we design a framework that the original inputs are RGB images and depth images, instead of four channels. To ensure the fairness of the comparison experiment, we retain both IMA and CMA. The ablation experiments show that YUV domain compression methods have obvious performance gain compared with RGB domain compression algorithm when tested in the YUV domain.

Case 3: Effectiveness of IMA and CMA. We assess the efficacy of IMA and CMA, and the results are presented in Table 3. It is shown that each module improves the whole RD performance. It is noteworthy that the results are better when CMA is used alone than when IMA is used alone. The results imply the importance of different modality information interactions and cross-modality redundancy removal in RGB-D image compression.

Conclusion

In this paper, we propose a novel learning-based RGB-D image compression framework, which significantly improves the compression efficiency of RGB-D images. First, we convert input image pairs from the RGB domain to the YUV420 domain to eliminate spatial redundancy. Intra-Modality Attention (IMA) is designed in the feature extraction and feature reconstruction stage to reduce cross-channel redundancy. Then, Cross-Modality Attention (CMA) is adapted in the encoder and decoder to remove cross-modality redundancy. To leverage the prior information between modalities effectively, Conditional Context-based Entropy Model is adopted for better symbol probability estimation. In the entropy model, we change the Context Prediction Module (CPM) with Mask Scaled Cosine Attention (MSCA). Context Fusion Module (CFM) is also proposed to aggregate cross-modality information. The comparative experiment results and the ablation study confirms the effectiveness of the proposed method.

Acknowledgments

This work was supported by Natural Science Foundation of China (62271013, 62031013), Shenzhen Fundamental Research Program (GXWD20201231165807007-20200806163656003), Shenzhen Science and Technology Program (JCYJ20230807120808017), CAAI-MindSpore Open Fund, developed on OpenI Community (CAAIXSJLJJ-2023-MindSpore07).

References

Alcorn, M. A.; and Nguyen, A. 2021. baller2vec++: A look-ahead multi-entity transformer for modeling coordinated agents. *arXiv preprint arXiv:2104.11980*.

Ballé, J.; Laparra, V.; and Simoncelli, E. P. 2016. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*.

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

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

Bellard, F. 2018. Bpg image format. http://bellard.org/bpg/. Accessed: 2023-08-01.

Bjontegaard, G. 2001. Calculation of average PSNR differences between RD-curves. *VCEG-M33, Austin, TX, USA*.

Bozic, A.; Zollhofer, M.; Theobalt, C.; and Nießner, M. 2020. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7002–7012.

Chen, T.; Liu, H.; Ma, Z.; Shen, Q.; Cao, X.; and Wang, Y. 2021. End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Transactions on Image Processing*, 30: 3179–3191.

Cheng, Z.; Sun, H.; Takeuchi, M.; and Katto, J. 2020. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7939–7948.

Chodosh, N.; Wang, C.; and Lucey, S. 2019. Deep convolutional compressed sensing for lidar depth completion. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part I 14*, 499–513. Springer.

Deng, X.; Yang, W.; Yang, R.; Xu, M.; Liu, E.; Feng, Q.; and Timofte, R. 2021. Deep homography for efficient stereo image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1492–1501.

Gao, W.; Liao, G.; Ma, S.; Li, G.; Liang, Y.; and Lin, W. 2021. Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4): 2091–2106.

Gao, W.; Sun, S.; Zheng, H.; Wu, Y.; Ye, H.; and Zhang, Y. 2023. OpenDMC: An Open-Source Library and Performance Evaluation for Deep-learning-based Multi-frame Compression. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9685–9688.

Gao, W.; Tao, L.; Zhou, L.; Yang, D.; Zhang, X.; and Guo, Z. 2020. Low-rate image compression with super-resolution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 154–155.

Hu, Y.; Yang, W.; and Liu, J. 2020. Coarse-to-fine hyperprior modeling for learned image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11013–11020.

Huang, Z.; Sun, Z.; Duan, F.; Cichocki, A.; Ruan, P.; and Li, C. 2021. L3c-stereo: Lossless compression for stereo images. *arXiv preprint arXiv:2108.09422*.

Ji, W.; Li, J.; Zhang, M.; Piao, Y.; and Lu, H. 2020. Accurate RGB-D salient object detection via collaborative learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, 52–69. Springer.

Jiang, W.; Yang, J.; Zhai, Y.; Ning, P.; Gao, F.; and Wang, R. 2023. MLIC: Multi-Reference Entropy Model for Learned Image Compression. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7618–7627.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, J.-H.; Jeon, S.; Choi, K. P.; Park, Y.; and Kim, C.-S. 2022. DPICT: Deep progressive image compression using trit-planes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16113–16122.

Li, M.; Li, J.; Gu, S.; Wu, F.; and Zhang, D. 2022. End-to-End Optimized 360° Image Compression. *IEEE Transactions on Image Processing*, 31: 6267–6281.

Liao, G.; Gao, W.; Jiang, Q.; Wang, R.; and Li, G. 2020. Mmnet: Multi-stage and multi-scale fusion network for rgb-d salient object detection. In *Proceedings of the 28th ACM international conference on multimedia*, 2436–2444.

Liu, J.; Wang, S.; and Urtasun, R. 2019. Dsic: Deep stereo image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3136–3145.

Liu, N.; Zhang, N.; and Han, J. 2020. Learning selective self-mutual attention for RGB-D saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13756–13765.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lu, M.; Guo, P.; Shi, H.; Cao, C.; and Ma, Z. 2021. Transformer-based image compression. *arXiv preprint arXiv:2111.06707*.

Minnen, D.; Ballé, J.; and Toderici, G. D. 2018. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31.

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

Shi, Y.; Xu, X.; Xi, J.; Hu, X.; Hu, D.; and Xu, K. 2022. Learning to detect 3D symmetry from single-view RGB-D images with weak supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4882–4896.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 567–576.

Tao, L.; Gao, W.; Li, G.; and Zhang, C. 2023. AdaNIC: Towards Practical Neural Image Compression via Dynamic Transform Routing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16879–16888.

Tian, M.; Pan, L.; Ang, M. H.; and Lee, G. H. 2020. Robust 6d object pose estimation by learning rgb-d features. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 6218–6224. IEEE.

Toderici, G.; O'Malley, S. M.; Hwang, S. J.; Vincent, D.; Minnen, D.; Baluja, S.; Covell, M.; and Sukthankar, R. 2015. Variable rate image compression with recurrent neural networks. *arXiv preprint arXiv:1511.06085*.

Van Den Oord, A.; Kalchbrenner, N.; and Kavukcuoglu, K. 2016. Pixel recurrent neural networks. In *International conference on machine learning*, 1747–1756. PMLR.

Wödlinger, M.; Kotera, J.; Xu, J.; and Sablatnig, R. 2022. Sasic: Stereo image compression with latent shifts and stereo attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 661–670.

Wu, Y.; and Gao, W. 2022. End-to-end lossless compression of high precision depth maps guided by pseudo-residual. In *2022 Data Compression Conference (DCC)*, 489–489. IEEE.

Wu, Y.; Qi, Z.; Zheng, H.; Tao, L.; and Gao, W. 2021. Deep image compression with latent optimization and piecewise quantization approximation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1926–1930.

Xiang, Y.; Xie, C.; Mousavian, A.; and Fox, D. 2021. Learning rgb-d feature embeddings for unseen object instance segmentation. In *Conference on Robot Learning*, 461–470. PMLR.

Zhao, Z.; Wang, S.; Jia, C.; Zhang, X.; Ma, S.; and Yang, J. 2018. Light field image compression based on deep learning. In *2018 IEEE International conference on multimedia and expo (ICME)*, 1–6. IEEE.

Zhu, X.; Song, J.; Gao, L.; Zheng, F.; and Shen, H. T. 2022. Unified multivariate gaussian mixture for efficient neural image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17612–17621.

Zollhöfer, M.; Stotko, P.; Görlitz, A.; Theobalt, C.; Nießner, M.; Klein, R.; and Kolb, A. 2018. State of the art on 3D reconstruction with RGB-D cameras. In *Computer graphics forum*, volume 37, 625–652. Wiley Online Library.

Zou, R.; Song, C.; and Zhang, Z. 2022. The devil is in the details: Window-based attention for image compression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17492–17501.