

Multi-Scale Invertible Neural Network for Wide-Range Variable-Rate Learned Image Compression

Hanyue Tu, Siqi Wu, Li Li, *Member, IEEE*, Wengang Zhou, *Senior Member, IEEE*, and Houqiang Li, *Fellow, IEEE*

Abstract—Autoencoder-based structures have dominated recent learned image compression methods. However, the inherent information loss associated with autoencoders limits their rate-distortion performance at high bit rates and restricts their flexibility of rate adaptation. In this paper, we present a variable-rate image compression model based on invertible transform to overcome these limitations. Specifically, we design a lightweight multi-scale invertible neural network, which bijectively maps the input image into multi-scale latent representations. To improve the compression efficiency, a multi-scale spatial-channel context model with extended gain units is devised to estimate the entropy of the latent representation from high to low levels. Experimental results demonstrate that the proposed method achieves state-of-the-art performance compared to existing variable-rate methods, and remains competitive with recent multi-model approaches. Notably, our method is the first learned image compression solution that outperforms VVC across a very wide range of bit rates using a single model, especially at high bit rates. The source code is available at <https://github.com/hytu99/MSINN-VRLIC>.

Index Terms—Learned image compression, invertible neural network, wide-range bit rates

I. INTRODUCTION

IMAGE compression is a fundamental task in the field of image processing, aiming at reducing the storage and transmission cost while preserving the visual quality of an image. In the past decades, traditional image/video compression standards have been developed, such as JPEG [1], JPEG2000 [2], HEVC [3], VVC [4]. The lossy image compression process can be generally divided into three steps, including transform, quantization, and entropy coding. Quantization parameters (QP) are used to control the trade-off between compression efficiency and visual quality. A wide range of quantization parameters QP values are supported by the traditional image

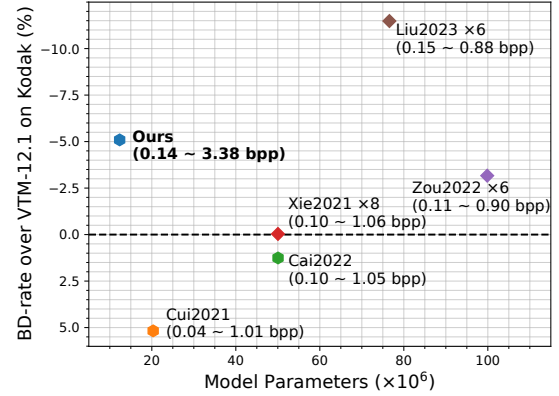


Fig. 1. Comparison of compression performance and model size. BD-rate over VTM-12.1 on Kodak is reported. A lower BD-rate indicates better performance. Our method, Cui2021 [5] and Cai2022 [6] are variable-rate models, while Xie2021 [7], Zou2022 [8] and Liu2023 [9] are fixed-rate models that require multiple models for different rates (“ $\times n$ ” means that n models are required). Our method surpasses VTM-12.1 across a wide range of rates with only a single lightweight variable-rate model.

codecs, which allow the users to determine the desired compression level by adjusting the QP values based on their own needs.

Recently, deep learning-based image compression [5], [6], [9]–[19] has shown promising progress, some of which have outperformed VVC. The state-of-the-art learned image compression methods [8], [9], [20] are usually based on the variational autoencoder [21] architecture. Although learned image compression has shown outstanding performance, it still has several limitations. Firstly, most of the state-of-the-art learned image compression methods [9], [20] require training multiple fixed-rate models for different target bit rates, which increases the training and storage costs when rate adaptation is needed. Besides, the supported bit rate range of learned image codecs is quite limited when compared with traditional image codecs. To tackle the first issue, variable-rate image compression methods [5], [6], [22] are proposed to utilize a single model for multiple bit rates, where the rate adaptation is achieved by conditioning on the Lagrange multiplier [22] or a quality map [6], [23], or introducing gain units in latent space [5]. Nonetheless, it is worth noting that these variable-rate image compression methods still only support a relatively narrow range of low bit rates.

This work was supported in part by the Natural Science Foundation of China under Grants 62171429 and 62021001, and in part by the Fundamental Research Funds for the Central Universities under Grant WK3490000007. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

Hanyue Tu, Li Li, Wengang Zhou and Houqiang Li are with the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application, University of Science and Technology of China, Hefei, 230027, China (e-mail: tuhanyue@mail.ustc.edu.cn; lilil@ustc.edu.cn; zhwg@ustc.edu.cn; lihq@ustc.edu.cn).

Siqi Wu is with the Department of Electrical Engineering, University of Missouri, Columbia, MO 65211 (email: siqiwu@mail.missouri.edu).

Corresponding authors: Li Li and Houqiang Li.

The main reason behind the constraints of autoencoder-based image compression methods is the inevitable information loss caused by the autoencoder itself. The encoder transforms the input image to a more compact latent representation, which has fewer elements than the input image. Therefore, the upper bound of the reconstruction quality is limited inherently. An alternative solution is to adopt the invertible neural network [6], [24] as the transform. With the property of invertibility, the latent representation contains all the information of the input image, and information loss only arises during the quantization process in the latent space. Therefore, the reconstruction quality is not bounded, enhancing the performance at high bitrates. Invertible transform is widely used in traditional image codecs, such as Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT). Previous studies [7], [24] have explored the development of the end-to-end compression scheme based on invertible transform. However, it is also claimed that invertible neural networks have limited nonlinear transformation capacity. To address this issue, Ma *et al.* [24] relied on a time-consuming RNN-based autoregressive context model, while Xie *et al.* [7] introduced extra modules for feature enhancement and dimensionality reduction, which disrupts the bijective property between the input image and the latent representation. Nevertheless, current invertible transform-based learned image compression methods still exhibit inferior performance when compared to autoencoder-based approaches. The potential of adopting invertible transforms in learned image compression remains largely unexplored.

In this paper, we present a lightweight variable-rate image compression model based on the invertible neural network, which could cover a wide range of bit rates, ranging from 0.1 to 3.4 bpp on Kodak [25] with a single model. As shown in Fig. 2, the invertible transform maps the input image to multi-scale latent representations, which are quantized and mapped back to the pixel domain through the reverse of the forward transform. The multi-scale structure reduces the parameters of the network without sacrificing performance and effectively preserves high-frequency information in low-level latent representation. To establish an efficient and practical compression scheme, we propose a multi-scale spatial-channel context model for accurate entropy modeling. Specifically, multi-layer masked convolutional layers are designed to enlarge the receptive field of the spatial context model while maintaining causality. In addition, to achieve flexible rate adaption, we extend the gain units proposed in [5] to better align with our multi-scale structure.

We evaluate the proposed compression scheme on Kodak image set [25], CLIC professional validation dataset [26], and old Tecnick dataset [27]. The experimental results demonstrate that our method outperforms VVC across a wide range of bit rates, using a single model, and also shows competitive performance with the state-of-the-art multi-model learned image compression methods. Another advantage of invertible image compression methods is that it is robust for multiple re-encoding, as demonstrated in [6], [28]. Our method also outperforms previous compression methods in terms of the fidelity for re-encoding.

The main contributions of our paper are summarized as follows:

- Instead of utilizing the commonly-used autoencoder-based architecture, we adopt invertible transform to develop a lightweight learned image codec that supports a wide range of bitrates with high compression efficiency.
- We design a multi-scale invertible neural network (INN) by stacking a series of simple yet powerful invertible units, which include an invertible convolutional layer and an affine coupling layer to enhance the model's representational capacity. The multi-scale structure reduces the parameters of our model without sacrificing performance.
- To ensure efficient compression of the multi-scale latent representation, a multi-scale spatial-channel context model with multi-layer masked convolutional layers is designed for accurate entropy modeling.
- The proposed method surpasses VTM-12.1 across a wide range of bit rates using a single model with only 12M parameters, especially at very high bit rates. With the property of invertibility, our method also shows the best performance on maintaining rate-distortion performance after multiple re-encodings.

II. RELATED WORK

A. Learned Image Compression

In recent years, there has been remarkable progress in leveraging deep learning for image compression. Most advanced end-to-end image compression methods are built upon the autoencoder architecture [10]–[12], [29], [30]. Ballé *et al.* [10] proposed the first autoencoder-based image compression framework and utilized additive uniform noise as a continuous approximation of quantization to enable end-to-end training. A lot of follow-up methods improve the rate-distortion performance by using more accurate entropy coding modeling and more advanced non-linear transformation. The hyperprior [11] is introduced to model the entropy model as a conditioned Gaussian distribution. Minnen *et al.* [12] utilized an autoregressive context model to capture the spatial correlation of latent representations. Li *et al.* [31] used 3D masks to extract spatial-context causal context autoregressively. Channel-wise context model [32] and checkerboard context model [20], [33], [34] are introduced to improve the computational efficiency of context prediction. Ballé *et al.* [30] presented an overview of nonlinear transform coding with both theoretical analysis and empirical study. Wang *et al.* [35] utilized Laplacian pyramid structure and compressed multi-scale residual images for better texture restoration. Akbari *et al.* [36] proposed to employ octave-based residual blocks in image compression framework. Recently, attention mechanism [8], [13], [37], [38] and transformer-based blocks [34], [39]–[41] are incorporated into the network architectures to further enhance image compression performance. Liu *et al.* [9] proposed a parallel transformer-CNN mixture block and incorporated a transformer-based attention module into the channel-wise entropy model.

Most learned image compression methods need to train and store multiple models for different rate operating points.

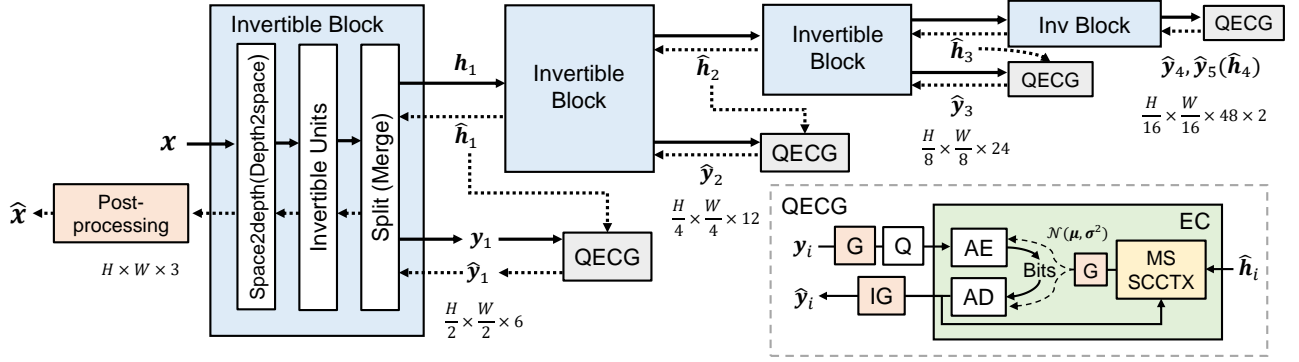


Fig. 2. Framework of our proposed invertible transform-based image compression model. Four-level invertible blocks map the input image to multi-scale latent representations $\{y_1, y_2, y_3, y_4, y_5\}$, which are quantized and passed through the reverse process of the network to reconstruct the image. The post-processing module is used to compensate for the quantization loss. “QECG” includes the process of quantization and entropy coding with gain units. “G” and “IG” denote gain units and inverse gain units, respectively. “MS-SCCTX” is the proposed multi-scale spatial-context model. “Q” represents quantization. “AE” and “AD” stand for arithmetic encoding and decoding, respectively.

Several studies have investigated the potential of using one single model to achieve *variable-rate* compression. Theis *et al.* [42] introduced and finetuned scale parameters for different rates while keeping the parameters of the pre-trained autoencoder fixed. Choi *et al.* [22] proposed a conditional autoencoder, which is conditioned on the Lagrange multiplier. Besides, mixed quantization training is utilized in [22] to allow finer rate adaptation by tuning the quantization bin size. Song *et al.* [23] proposed a quality map-guided compression model to support spatially-adaptive variable rates. Sun *et al.* [43] introduced an interpolation channel attention module to achieve fine rate control. Cui *et al.* [5] designed gain units to rescale the latent representations and used the exponent interpolation to enable continuous rate adaptation. Cai *et al.* [6] proposed an invertible activation transformation module, which takes the quality level as input to generate element-wise scaling and bias tensors.

B. Invertible Neural Network

An invertible neural network is a special kind of neural network, which is composed of several bijective mappings. Therefore, the input can be losslessly recovered given the output through a reverse process of the neural network. Invertible neural networks have been widely studied in the context of normalizing flows [44], [45], which employs a sequence of invertible mapping to transform the complex distributions of input data into simpler ones. The commonly used invertible architectures are additive coupling layers [45] and affine coupling layers [46]. Instead of directly applying the transformation to the input \mathbf{x} , the input $\mathbf{x} \in \mathbb{R}^D$ is first split into $(\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^d \times \mathbb{R}^{D-d}$. Then the additive coupling layer can be formulated as $(\mathbf{y}_1, \mathbf{y}_2) = (\mathbf{x}_1, \mathbf{x}_2 + f(\mathbf{x}_1))$, and the affine coupling layer can be formulated as $(\mathbf{y}_1, \mathbf{y}_2) = (\mathbf{x}_1, \mathbf{x}_2 \odot \exp(s(\mathbf{x}_1)) + f(\mathbf{x}_1))$. The coupling layers are composed in an alternating pattern. Kingma *et al.* [47] proposed to replace the fixed permutation in the flow model with a learned invertible 1×1 convolution.

Invertible neural networks are demonstrated effective in other research domains as well. Gomez *et al.* [48] proposed reversible residual network (RevNet), where the activations of

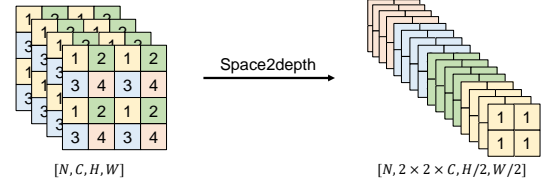


Fig. 3. Invertible down-scaling using the space-to-depth module.

each layer can be exactly reconstructed from the activations of the next layer. Therefore, the activations for most layers need not be stored in memory. Jacobsen *et al.* [49] built a fully invertible architecture i-RevNet upon RevNet and showed that it can achieve the same classification accuracy compared to similar non-invertible architectures. Xiao *et al.* [50] designed an invertible rescaling network to transform high-resolution images into low-resolution images and a case-agnostic latent variable.

Invertible transform is also widely used in traditional hybrid image codecs, such as Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), and the lifting scheme [51]. Recently, some works also developed learned image compression based on invertible architecture. Helming *et al.* [28] explored the potential of normalizing flows for lossy image compression. The model covers a wide range of quality levels but the performance degrades at low bit rates. Ma *et al.* [24], [52] designed a trainable wave-like transform to support both lossless and lossy image compression. Xie *et al.* [7] presented an enhanced invertible encoding network to mitigate the information loss problem, which used an attentive channel squeeze layer and a feature enhancement module for better compression performance. Li *et al.* [53] designed a reversible autoencoder with a guarantee of stability for image reconstruction. Cai *et al.* [6] introduced invertible activation transformation and extended [7] to a variable-rate model.

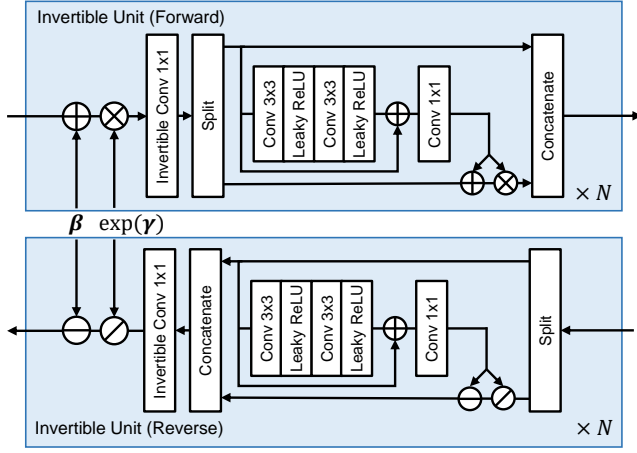


Fig. 4. Architecture of the invertible unit. We stack N invertible units within each invertible block.

III. PROPOSED METHOD

A. Framework

The framework of our proposed image compression model based on invertible transform is illustrated in Fig. 2. We employ a multi-scale architecture. The input image $x \in \mathbb{R}^{H \times W \times 3}$ is first invertibly down-scaled by the space-to-depth [33] module. We divide the input image into non-overlapping 2×2 patches. Each 2×2 patch is flattened to a one-dimensional vector through a series of reshaping and permuting operations, as shown in Fig. 3. The down-scaled image $x_{ds} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 12}$ are then passed through a sequence of invertible units. This sequence enables invertible computations and captures the dependencies within the data. The output is split along the channel dimension, generating the first latent representation y_1 and another hidden representation h_1 , which is formulated as:

$$(y_1, h_1) = IB_1(x), \quad (1)$$

where IB represents the invertible block. The hidden representation h_1 is then passed as input to the subsequent invertible block at the next level. This recursive process is defined as follows:

$$\begin{aligned} (y_{i+1}, h_{i+1}) &= IB_{i+1}(h_i), 1 \leq i \leq 2, \\ (y_4, y_5) &= IB_4(h_3). \end{aligned} \quad (2)$$

Four-level invertible blocks are used in our model and we obtain the multi-scale latent representations $\mathbf{Y} = \{y_1, y_2, y_3, y_4, y_5\}$.

The latent representations \mathbf{Y} are quantized and encoded into bitstream. During the image reconstruction phase, the decoded latent representations $\hat{\mathbf{Y}}$ go through the reverse process of the invertible blocks. At the end of the decoding process, a post-processing module is utilized to compensate for the information loss resulting from the quantization. The decoding process is formulated as:

$$\hat{x} = P(IB_1^{-1}(IB_2^{-1}(IB_3^{-1}(IB_4^{-1}(\hat{y}_5, \hat{y}_4, \hat{y}_3), \hat{y}_2), \hat{y}_1))), \quad (3)$$

where IB^{-1} represents the reverse process of the invertible block, and P denotes the post-processing module.

B. Architecture of Invertible Units

At each level, the invertible block is composed of N invertible units. The architecture of a single invertible unit is shown in Fig. 4, which is designed based on the flow step in the generative model [47]. In the forward process, given the input tensor t , a data-independent channel-wise normalization is applied first. Learnable vectors β and γ are initialized as $-E(t)$ and $-\log(\sqrt{\text{Var}(t)})$. The mean and variance are calculated over the spatial dimension per channel.

Different from previous invertible transform-based image compression methods [7], [24], which split the tensor into two chunks and directly input the two chunks into multiple sequential coupling layers, we utilize an invertible 1×1 convolutional layer before each coupling layer. The invertible 1×1 convolutional layer fuses the local features in the channel dimension, facilitating more frequent exchange of information and enhancing the representation power of the transformation. After the invertible 1×1 convolutional layer, an affine coupling layer is applied. Denoting the input of the coupling layer as $u \in \mathbb{R}^{h \times w \times C}$, we first evenly split it into two parts, u_1 and u_2 , each with a size of $h \times w \times \frac{C}{2}$. Then the first tensor u_1 is fed into a residual block and a 1×1 convolutional layer is used to generate the bias and scale parameters that are applied to u_2 . At last, u_1 and the updated u'_2 are concatenated again and passed through the next invertible unit. This process can be formulated as follows:

$$\begin{aligned} t' &= (t + \beta) \otimes \exp(\gamma), \\ u &= t' \cdot W, \\ u_1, u_2 &= \text{Split}(u), \\ \text{bias, scale} &= \text{Split}(\text{Conv}_{1 \times 1}(\text{ResBlock}(u_1))), \\ u'_2 &= (u_2 + \text{bias}) \odot \exp(2\sigma(\text{scale}) - 1), \\ u' &= \text{Concatenate}(u_1, u'_2) \end{aligned} \quad (4)$$

where \otimes denotes channel-wise multiplication, \odot denotes Hadamard product, and $\sigma(\cdot)$ represents the sigmoid function.

Correspondingly, we can formulate the reverse process of the invertible unit as:

$$\begin{aligned} (u_1, u'_2) &= \text{Split}(u'), \\ \text{bias, scale} &= \text{Split}(\text{Conv}_{1 \times 1}(\text{ResBlock}(u_1))), \\ u_2 &= u'_2 \odot \exp(1 - 2\sigma(\text{scale})) - \text{bias}, \\ u &= \text{Concatenate}(u_1, u_2) \\ t' &= u \cdot W^{-1}, \\ t &= t' \otimes \exp(-\gamma) - \beta, \end{aligned} \quad (5)$$

C. Multi-scale Spatial-channel Context Model

Previous invertible transform-based image compression methods [6], [7], [24] adopted the time-consuming serial auto-regressive context models, which are actually impractical for establishing an effective compression scheme. In this paper, we design a multi-scale spatial-channel context model to efficiently encode the multi-scale latent representations $\mathbf{Y} = \{y_1, y_2, y_3, y_4, y_5\}$. Our method encodes the latent representations \mathbf{Y} from low resolution to high resolution. The entropy model for quantized latent \hat{y}_i is illustrated in

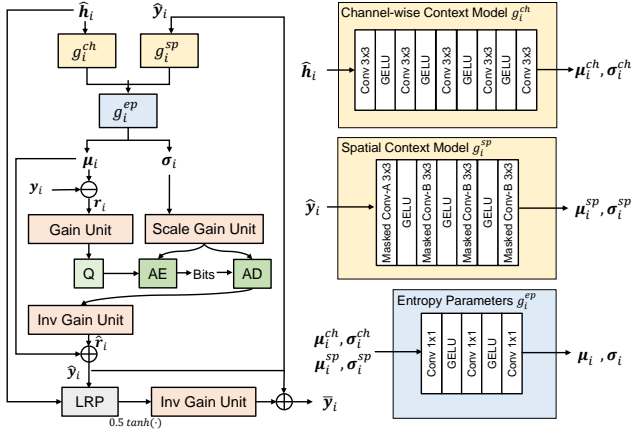


Fig. 5. Proposed spatial-channel context model with extended gain units for entropy coding. The architecture of the latent residual prediction (LRP) module is the same as the channel-wise context model.

Fig. 5. Each element in \hat{y}_i is modeled as a single Gaussian distribution. The channel-wise Gaussian parameters mean μ_i^{ch} and scale σ_i^{ch} are conditioned on the previously decoded lower-resolution latent codes $\{\hat{y}_{>i}\}$. For slice \hat{y}_5 , μ_5^{ch} is set to zero and σ_5^{ch} is defined as learnable parameters. To align the resolution for channel-wise context prediction, we utilize the reverse process of the invertible transform to reconstruct the hidden representation \hat{h}_i , which has the same shape as \hat{y}_i . The channel-wise context model g_i^{ch} takes \hat{h}_i as input and produces the channel-conditional μ_i^{ch} and σ_i^{ch} .

To reduce the spatial redundancy, we further incorporate the checkerboard context model proposed in [33] into the entropy model of each latent code. The latent code \hat{y}_i is split into two parts, $\hat{y}_i^{\text{anchor}}$ and $\hat{y}_i^{\text{non-anchor}}$, in a checkerboard pattern, and the entropy model for $\hat{y}_i^{\text{non-anchor}}$ is conditioned on $\hat{y}_i^{\text{anchor}}$. Previous methods [20], [33] utilized a single-layer 5×5 checkerboard-masked convolution to model the spatial correlation, as shown in Fig. 6a. This is sufficient in previous autoencoder-based models with lower-resolution latent representations, where each element corresponds to 16 pixels in the input image. However, in our multi-scale architecture, the low-level latent representation is larger, requiring a context model with a larger receptive field.

We design a multi-layer spatial context model by stacking four 3×3 masked convolutional layers. Specifically, the first layer uses mask A while the consecutive layers use mask B, as shown in Fig. 6b. We visualize the weighted receptive field of our multi-layer masked convolutional layers in Fig. 6c. To visualize the receptive field, we set all the convolution weights to 1 and biases to 0, and then calculate the gradient magnitude of the input with respect to the output center pixel. Our multi-layer spatial context model preserves the causality while benefiting from a larger receptive field and increased nonlinearity. Finally, the entropy parameter model g_i^{ep} aggregates the estimated parameters from g_i^{ch} and g_i^{sp} . In

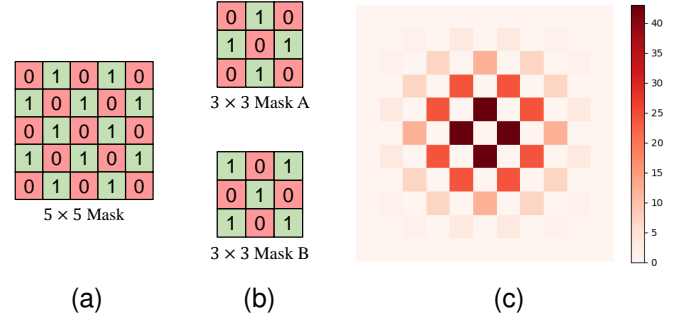


Fig. 6. (a) 5×5 mask used in [20], [33]. (b) Two types of 3×3 masks used in our spatial context model. (c) Visualization of the weighted receptive field of our spatial context model, which consists of one masked convolution-A and three masked convolution-B.

summary, our context model can be formulated as:

$$\begin{aligned} \mu_5, \sigma_5 &= g_5^{ep}(\mathbf{0}, \sigma_5^{ch}, g_5^{sp}(\hat{y}_i^{\text{anchor}})), \\ \mu_i, \sigma_i &= g_i^{ep}(g_i^{ch}(\hat{h}_i), g_i^{sp}(\hat{y}_i^{\text{anchor}})), 1 \leq i \leq 4, \\ \hat{h}_i &= IB_{i+1}^{-1}(\hat{h}_{i+1}, \hat{y}_{i+1}) = IB_{>i}^{-1}(\hat{y}_{>i}), \end{aligned} \quad (6)$$

where $IB_{>i}^{-1}$ represents the reverse process of the transform starting from the $i+1$ -th invertible blocks onwards. The spatial context term g_i^{sp} is set to zero for all the elements in $\hat{y}_i^{\text{anchor}}$.

D. Rate Adaptation with Extended Gain Units

Gain units are introduced in [5] to achieve continuous rate adaptation by rescaling the magnitude of the latent representation using channel-wise multiplication before quantization and entropy parameter estimation. However, the original gain units are unsuitable for our multi-scale architecture because our channel-wise context model takes as input the reconstructed hidden representation \hat{h}_i , instead of the scaled latent representations, as explained in Sec. III-C. The inconsistency in magnitude between the input and output can adversely affect the training process of the channel-wise context model. In our approach, the entropy parameters are all calculated based on the unscaled hidden and latent representations. We place the gain units and the inverse gain units immediately before and after the quantization process, which is formulated as:

$$\hat{y}_{i,q} = IG_i(Q(G_i(\mathbf{y}_i - \mu_i, q)), q) + \mu_i \quad (7)$$

where $Q(\cdot)$ represents the quantization process, $G_i(\cdot)$ and $IG_i(\cdot)$ represent the gain process and inverse gain process, respectively, and q denotes the quality value, which can take any arbitrary value within a predefined range using the exponential interpolation [5].

In practical implementation, it is the gained residual $Q(G_i(\mathbf{y}_i - \mu_i, q))$ to be encoded, instead of $\hat{y}_{i,q}$. The quantized gained residual is modeled as a single Gaussian distribution. In order to estimate the rates in a variable-rate model, we also apply the gain process to the estimated scale parameter σ_i , as below:

$$\begin{aligned} P(\hat{y}_{i,q} | \hat{y}_{>i,q}) &\sim P(Q(G_i(\mathbf{y}_i - \mu_i, q)) | \hat{y}_{>i,q}) \\ &\sim \mathcal{N}(\mathbf{0}, G_i'(\sigma_i, q)) \end{aligned} \quad (8)$$

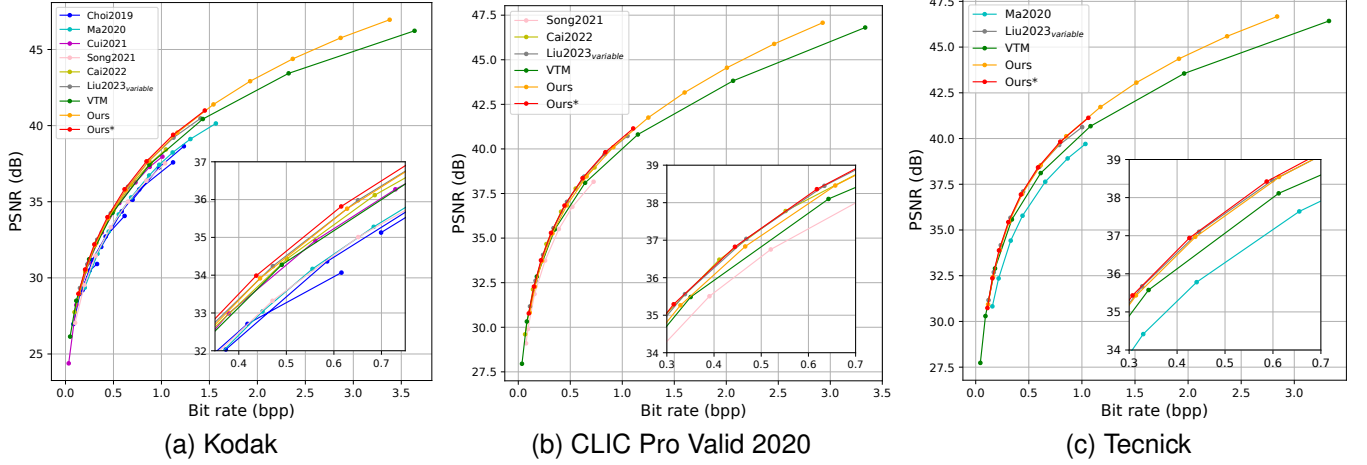


Fig. 7. Rate-distortion performance comparison between our method and variable-rate image compression methods in a wide range of bit rates. “Ours” represents our model with $\lambda_{max} = 1.8$, while “Ours*” represents our model with $\lambda_{max} = 0.18$.

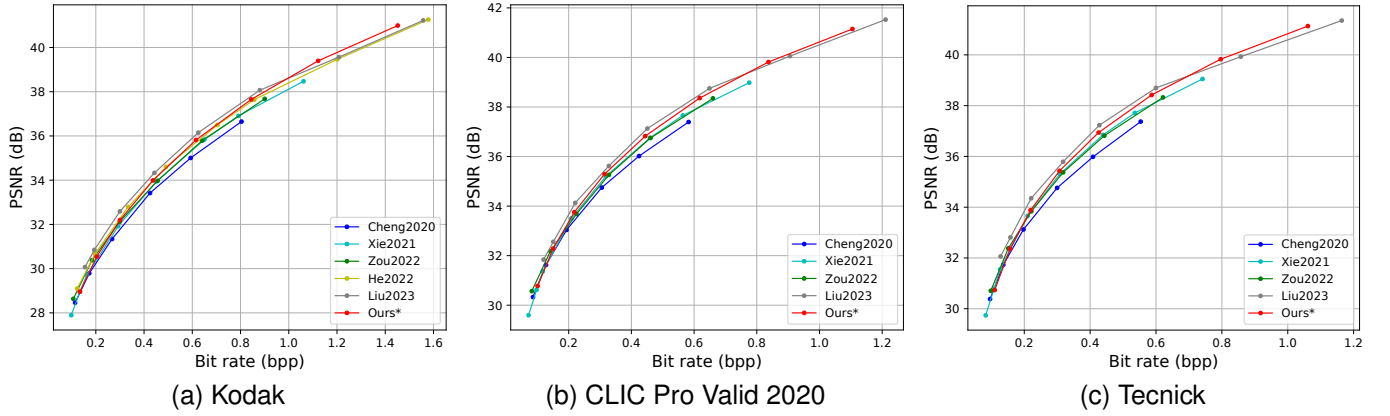


Fig. 8. Rate-distortion performance comparison between our method and fixed-rate image compression methods. “Ours*” represents our model with $\lambda_{max} = 0.18$.

Latent residual prediction (LRP) [32] aims to reduce the quantization error by predicting the residual in latent space. The predicted residual is scaled by the inverse gain units because the quantization error of \hat{y}_i falls in the range of $(IG_i(-0.5, q), IG_i(0.5, q))$.

During the training process of our model, we randomly sample the quality values q from the predefined set $\mathcal{S}_q = \{0, 1, 2, \dots, q_{max}\}$. Each quality value q corresponds to a Lagrange multiplier λ_q in the rate-distortion optimization problem. The loss function of the proposed variable-rate model is defined as:

$$\begin{aligned} \mathcal{L} &= \sum_{q \in \mathcal{S}_q} \sum_{i=1}^5 R(\hat{y}_{i,q}) + \lambda_q D(\mathbf{x}, \hat{\mathbf{x}}_q) \\ &= \sum_{q \in \mathcal{S}_q} \sum_{i=1}^4 \mathbb{E}[-\log_2(P(\hat{y}_{i,q}|\hat{y}_{>i,q}))] \\ &\quad + \mathbb{E}[-\log_2(P(\hat{y}_{5,q}))] + \lambda_q D(\mathbf{x}, \hat{\mathbf{x}}_q), \end{aligned} \quad (9)$$

where $D(\cdot)$ represents the distortion term such as mean squared error (MSE). During the inference time, we can select

a specific quality value $q \in [1, q_{max}]$ to achieve the desired compression bit rate.

IV. EXPERIMENTS

A. Experimental Setup

1) *Training Details*: Our variable-rate model is trained on the Flickr2W [54] dataset. The training set consists of 20511 images and the remaining 200 images are used for quick validation. During training, We randomly crop 256×256 patches and use a batch size of 8. We employ the Adam optimizer [55] with an initial learning rate of 1×10^{-4} . The model is trained for a total of 750 epochs (approximately 2M iterations), with the learning rate decreasing to 1×10^{-5} at the 675th epoch and 1×10^{-6} at the 720th epoch.

For the network architecture, we utilize four invertible blocks, each consisting of four invertible units. For the first three invertible blocks, the channel number of the intermediate layer is set to 128, while for the last invertible block, it is set to 192. The channel number of the intermediate layer in the channel-wise context model, spatial context model, and latent residual prediction model are all set to 128. For the

post-processing network, we utilize a lightweight U-Net [56] containing 1.6M parameters.

For rate adaptation, the max quality value q_{max} is set to 11, and λ_q in Eq. (9) belongs to $\{0.0018, 0.0035, 0.0067, 0.0130, 0.0250, 0.0483, 0.0932, 0.1800, 0.320, 0.569, 1.012, 1.8\}$ for MSE distortion. This model covers a wide range of bit rates similar to the traditional codec VTM. We also train another model with q_{max} equal to 7 ($\lambda_{max} = 0.18$) to obtain a comparable bitrate range with recent fixed-rate methods for a fair comparison. Another approach for a fair comparison is training other methods with bigger q_{max} . However, training multiple fix-rate models for all other methods is rather time-consuming and we cannot guarantee the optimal training schedule for their methods. We first train the model with a narrower bit rate range and fine-tune it to obtain a wider-range model. The initial learning rate for fine-tuning is set to $1e-5$. We find such a fine-tuning process can benefit the RD performance at lower bit rates for the model with a wider bit rate range, compared to training from scratch.

2) *Evaluation*: We evaluate our model on three datasets, including Kodak image set [25] containing 24 images with size of 768×512 , CLIC professional validation dataset [26] containing 41 images with 2K resolution, and old Tecnick dataset [27] containing 100 images with size of 1200×1200 .

B. Comparison with the SOTA Methods

1) *Rate-distortion Performance*: We compare our method with state-of-the-art variable-rate image compression methods [5], [6], [22]–[24], fixed-rate methods [7]–[9], [13], [20], and the traditional codec VVC (VTM-12.1) [4]. Specifically, the methods [6], [7], [24] are also based on invertible neural networks. For the latest fixed-rate method Liu2023 [9], we modify it to obtain a variable-rate version using the same gain units as in our model and train two additional fixed-rate models at higher rates. The training configuration is the same as ours. The RD performance comparison is shown in Fig. 7 and Fig. 8.

When compared to variable-rate methods, our model with $\lambda_{max} = 1.8$ outperforms the traditional codec VVC across a wide range of rates. In particular, as shown in Fig. 7, the PSNR gain is more pronounced at higher bit rates. Previous methods only focus on a limited rate range. When only considering low rates, our model with $\lambda_{max} = 0.18$ (denoted as Ours*) further enhances the compression performance, surpassing all the existing variable-rate methods.

In comparison to fixed-rate methods, our model still shows competitive performance, surpassing most of the existing learned image compression methods, as shown in Fig. 8. Our model performs slightly worse than the latest method [9], but outperforms it at higher bit rates. This can be attributed to the fact that our model is built on the invertible neural network and the latent representations to be compressed contain all the information of the input image. It is worth noting that we still use our single variable-rate model here.

We also provide the MS-SSIM rate-distortion performance of our model optimized for both MSE and MS-SSIM distortion metrics. The model optimized for MS-SSIM is fine-tuned from the MSE-optimized model with a learning rate of $1e-5$. The

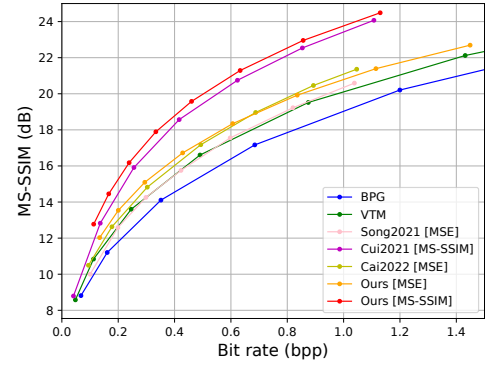


Fig. 9. MS-SSIM rate-distortion performance comparison between our method and variable-rate image compression methods on Kodak dataset. MS-SSIM is converted to decibels $-10 \log_{10}(1 - \text{MS-SSIM})$ for better visual clarity.

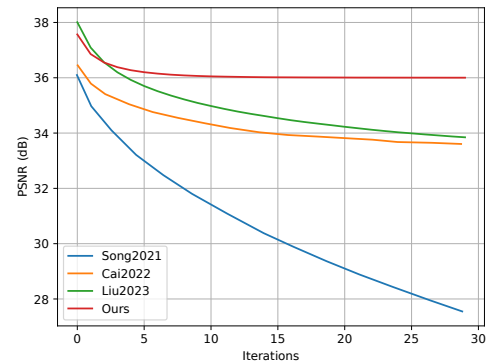


Fig. 10. Compression performance after multiple re-encoding operations on Kodak. All models achieve comparable bit rates.

results on Kodak are shown in Fig. 9, where MS-SSIM (dB) is calculated by $-10 \log_{10}(1 - \text{MS-SSIM})$. We compare our method with recent variable-rate image compression methods [5], [6], [23], as well as two traditional image codecs VVC (VTM-12.1) and BPG. Some previous methods only report results for MSE-optimized models. Our models optimized for MSE and MS-SSIM both outperform VVC by a large margin, and our model optimized for MS-SSIM achieves the best RD performance among the compared methods.

2) *Fidelity of Re-encoding*: Invertible neural network (INN)-based image compression is better at preserving the image quality after multiple re-encoding operations [6]. To verify the fidelity of re-encoding of our method, We compare it with the state-of-the-art INN-based method [6] and VAE-based methods [9], [23], as shown in Fig. 10. Both Song2021 [23] and Cai2022 [6] achieve a bit rate of around 0.791 bpp, as reported in [6]. The initial bit rate of our model is 0.832 bpp, which gradually decreases and stabilizes at 0.786 bpp. The rate of the model from [9] starts at 0.869 bpp and gradually decreases to 0.772 bpp. Our method achieves the best performance in maintaining the image quality after multiple re-encoding iterations, which is important in practical applications. We also provide a qualitative comparison between our method and Liu2023 [9] in Fig. 11. Our method preserves the texture details even after 30 re-encoding operations. The slight degradation mainly comes from the

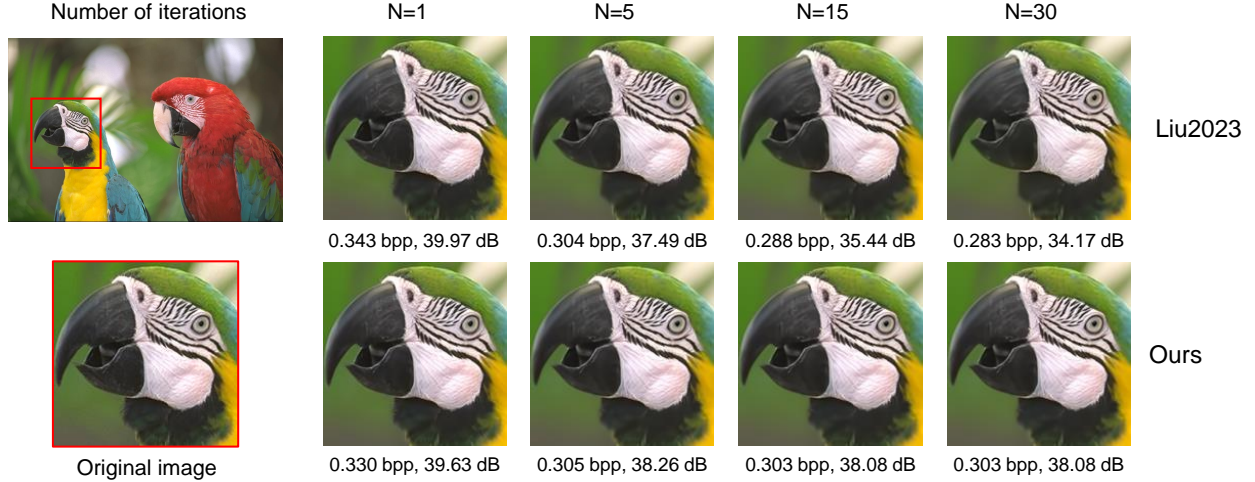


Fig. 11. Qualitative results after different re-encoding operations with our method and autoencoder-based method Liu2023 [9].

TABLE I

COMPLEXITY AND COMPRESSION PERFORMANCE OF LEARNED IMAGE COMPRESSION METHODS. BD-RATE AGAINST VVC (VTM-12.1) ON KODAK IS REPORTED. A LOWER BD-RATE INDICATES BETTER PERFORMANCE. “ $\times N$ ” MEANS THAT N MODELS ARE REQUIRED FOR DIFFERENT RATES.

Method	Enc. (s)	Dec. (s)	Params. (M) $\times N$	BD-rate (%)
Xie2021 [7]	2.944	7.127	50.03 \times 8	-0.035
Zou2022 [8]	0.146	0.151	99.86 \times 6	-3.167
Liu2023 [9]	0.207	0.199	76.57 \times 6	-11.481
Ours	0.189	0.191	12.34 \times 1	-5.091

additional dequantization modules in the decoding process, including LRP and post-processing modules, which have a minor impact on re-encoding performance but are crucial for RD performance.

3) *Complexity*: We evaluate the complexity and BD-rate against VVC of our method and recent learned image compression methods in Tab. I. All the models are tested on Kodak with a single RTX 3090 GPU. Our method stands out for its significantly fewer parameters and achieves a comparable inference latency against the state-of-the-art methods, *i.e.*, decoding an image with 768×512 pixels in 200 ms. In addition, our method excels in terms of training speed. For training, our model processes a batch of 8 images in 35.4 ms, which is approximately one-third the processing time of [9].

The BD rate of our method is higher than [9]. However, our model covers a much wider continuous rate range from 0.14 to 3.38 bpp using only a single variable-rate model, while Liu2023 [9] requires six separate models and can only handle a narrow rate range from 0.15 to 0.88 bpp. Note that the total size of the model parameters of our method is only $12.34 / (76.57 \times 6) = 2.6\%$ of that of [9]. We attribute this significant advantage in model size to the inherent priors of the invertible neural network.

C. Ablation Study

1) *Settings*: For ablation study experiments, we train each model for 600K iterations. The learning rate starts at 1×10^{-4}

TABLE II

ABLATION STUDIES ON KODAK DATASET. THE ANCHOR IS OUR MODEL WITH $\lambda_{max} = 0.18$

Method	Params. (M)	BD-rate (%)
Ours*	12.339	0.00
w/o normalization	12.338	+2.23
w/o invertible 1×1 conv	12.290	+3.38
w/ additive coupling	12.281	+2.30
w/ 5×5 masked conv	10.654	+1.57
w/ 7×7 masked conv	10.912	+1.23
w/ 9×9 masked conv	11.255	+1.84
w/ 11×11 masked conv	11.684	+2.09
w/o spatial context	10.386	+9.38
w/o post-processing	10.680	+5.38
w/o LRP	9.703	+6.96
w/o post-processing & LRP	8.044	+10.37

and drops to 1×10^{-5} at 500K iterations. All models are evaluated on Kodak dataset. The results are shown in Tab. II.

2) *Analysis of the invertible units*: Each invertible unit comprises a learnable channel-wise normalization, an invertible 1×1 convolutional layer and an affine coupling layer, as shown in Fig. 4. We first remove the normalization layer, leading to performance degradation of 2.23% BD-rate. The learnable normalization layer stabilizes the training process and improves performance with negligible parameters.

The invertible 1×1 convolutional layer in each invertible unit fuses the features channel-wisely. Previous invertible neural network-based image compression methods [7], [24] exchanged the order of the two split tensors, *i.e.*, $(u_1, u_2) \rightarrow (u_2, u_1)$, before coupling layers, which can be seen as reversing the tensor u along channel dimension. We replace the invertible 1×1 convolutional layer with the reversing operation. The 1×1 convolutional layer improves the RD performance by 3.38% BD-rate, which demonstrates the importance of sufficient information exchange. Note that all the invertible 1×1 convolutional layers necessitate 0.049M parameters, which account for only 0.4% of the total parameters.

We use the affine coupling layer to enhance the nonlinearity capacity of the invertible transform. By replacing the affine coupling transform with an additive coupling transform, *i.e.*,

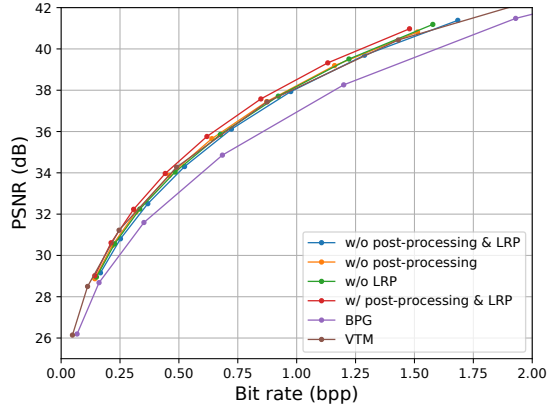


Fig. 12. Effects of the post-processing network and the latent residual prediction model.

removing the scale term in Eq. (4) and Eq. (5), the performance degrades by 2.30% BD-rate. Note that we use a single residual block to generate both the bias and scale parameters, instead of two separate blocks as in [7]. Therefore, the affine coupling layer is more efficient with similar parameters as the additive coupling layer.

3) *Analysis of Spatial Context Model*: To investigate the effect of the proposed multi-layer spatial context model, we develop two variants of our model: one without the spatial context model and another using a single-layer $k \times k$ masked convolutional layer. Removing the spatial context model leads to severe performance degradation, which shows that the spatial correlation is important for accurate entropy modeling of the multi-scale latent representation. A single-layer masked convolutional layer can also enhance the compression performance, but the performance is still not as good as our multi-layer spatial context model. We set the kernel size of the single-layer masked convolutional layer to 5, 7, 9, 11 and observe that the 7×7 masked convolutional layer achieves the best performance, which indicates that simply increasing the kernel size of the single-layer convolutional layer can not always improve the RD performance. The stacked 3×3 masked convolutional layers in our proposed multi-layer spatial context model can not only enlarge the receptive field, but also increase nonlinearity while maintaining low complexity. We can adjust the intermediate channel dimensions to enhance the capacity of the context model, which is not achievable with a single convolution layer.

4) *Analysis of the dequantization modules*: We evaluate the effects of the dequantization modules, namely the latent residual prediction (LRP) model and the post-processing network by removing them from our model and retraining it. These two modules reduce the quantization error in latent space and pixel space, respectively. It is worth noting that dequantization modules play a crucial role in invertible neural network-based image compression methods. This is because the encoder and the decoder share the same parameters, while in other autoencoder-based image compression methods, the dequantization process is actually integrated within the decoder's parameters. The results are shown in Tab. II and Fig. 12. The RD performance drops significantly without

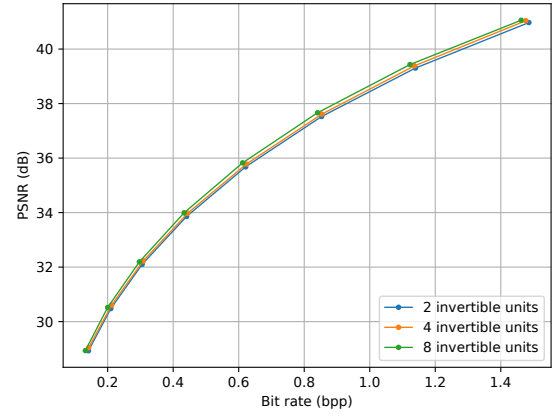


Fig. 13. Ablation study on the number of invertible units.

the dequantization modules. Nevertheless, it still achieves comparable performance to VTM and outperforms the BPG by a considerable margin. Both the LRP module and post-processing network contribute to enhancing the compression performance.

5) *Effect of the number of invertible units*: The number of invertible units N is set to 4 in each invertible block by default. We also conduct experiments with 2 and 8 invertible units per block to explore the effects of shallower or deeper invertible transforms. The results are shown in Sec. IV-C5. It can be observed that using 2 invertible units slightly decreases the RD performance. When using 8 invertible units, the RD performance is further enhanced, which suggests that our approach can benefit from more complex transform. Nevertheless, since we aim to develop a lightweight model, we maintain the setting of 4 invertible units in this paper.

V. CONCLUSION

In this paper, we propose a variable-rate image compression method based on a lightweight invertible neural network. Specifically, we design a multi-scale spatial-channel context model to enlarge the receptive field and extend gain units for flexible rate adaptation. Through extensive experiments, we demonstrate that the proposed method outperforms VVC and most existing learned image compression methods, covering a wider rate range with fewer parameters. These results demonstrate the significant potential of employing invertible transform in learned image compression.

REFERENCES

- [1] G. K. Wallace, "The jpeg still picture compression standard," *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.
- [2] D. S. Taubman, M. W. Marcellin, and M. Rabbani, "Jpeg2000: Image compression fundamentals, standards and practice," *Journal of Electronic Imaging*, vol. 11, no. 2, pp. 286–287, 2002.
- [3] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.
- [4] J. V. E. T. (JVET), "Vvc official test model vtm," 2021. [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM

- [5] Z. Cui, J. Wang, S. Gao, T. Guo, Y. Feng, and B. Bai, "Asymmetric gained deep image compression with continuous rate adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10532–10541.
- [6] S. Cai, Z. Zhang, L. Chen, L. Yan, S. Zhong, and X. Zou, "High-fidelity variable-rate image compression via invertible activation transformation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2021–2031.
- [7] Y. Xie, K. L. Cheng, and Q. Chen, "Enhanced invertible encoding for learned image compression," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 162–170.
- [8] R. Zou, C. Song, and Z. Zhang, "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*, 2022, pp. 17492–17501.
- [9] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-cnn architectures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14388–14397.
- [10] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations*, 2017.
- [11] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.
- [12] D. Minnen, J. Ballé, and G. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *NeurIPS*, 2018.
- [13] 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.
- [14] Y. Mei, L. Li, Z. Li, and F. Li, "Learning-based scalable image compression with latent-feature reuse and prediction," *IEEE Transactions on Multimedia*, vol. 24, pp. 4143–4157, 2022.
- [15] S. Yang, Y. Hu, W. Yang, L.-Y. Duan, and J. Liu, "Towards coding for human and machine vision: Scalable face image coding," *IEEE Transactions on Multimedia*, vol. 23, pp. 2957–2971, 2021.
- [16] P. Zhang, M. Wang, B. Chen, R. Lin, X. Wang, S. Wang, and S. Kwong, "Learning-based compression for noisy images in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.
- [17] C. Gao, D. Liu, L. Li, and F. Wu, "Towards task-generic image compression: A study of semantics-oriented metrics," *IEEE Transactions on Multimedia*, vol. 25, pp. 721–735, 2023.
- [18] Y. Zhang, H. Lin, J. Sun, L. Zhu, and S. Kwong, "Learning to predict object-wise just recognizable distortion for image and video compression," *IEEE Transactions on Multimedia*, vol. 26, pp. 5925–5938, 2024.
- [19] W. Jiang, P. Ning, J. Yang, Y. Zhai, F. Gao, and R. Wang, "Llic: Large receptive field transform coding with adaptive weights for learned image compression," *IEEE Transactions on Multimedia*, pp. 1–15, 2024.
- [20] 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.
- [21] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [22] Y. Choi, M. El-Khamy, and J. Lee, "Variable rate deep image compression with a conditional autoencoder," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3146–3154.
- [23] M. Song, J. Choi, and B. Han, "Variable-rate deep image compression through spatially-adaptive feature transform," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2380–2389.
- [24] H. Ma, D. Liu, N. Yan, H. Li, and F. Wu, "End-to-end optimized versatile image compression with wavelet-like transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1247–1263, 2022.
- [25] E. Kodak, "Kodak lossless true color image suite," 1999. [Online]. Available: <http://r0k.us/graphics/kodak/>
- [26] CLIC, "Workshop and challenge on learned image compression," 2021. [Online]. Available: <https://clic.compression.cc/2021/tasks/index.html>
- [27] N. Asuni and A. Giachetti, "Testimages: a large-scale archive for testing visual devices and basic image processing algorithms," in *STAG*, 2014, pp. 63–70. [Online]. Available: <https://sourceforge.net/projects/testimages/files/OLD/>
- [28] L. Helming, A. Djelouah, M. Gross, and C. Schroers, "Lossy image compression with normalizing flows," *arXiv preprint arXiv:2008.10486*, 2020.
- [29] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Energy compaction-based image compression using convolutional autoencoder," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 860–873, 2020.
- [30] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, "Nonlinear transform coding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 339–353, 2020.
- [31] C. Li, J. Luo, W. Dai, C. Li, J. Zou, and H. Xiong, "Spatial-channel context-based entropy modeling for end-to-end optimized image compression," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2020, pp. 222–225.
- [32] 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.
- [33] D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard context model for efficient learned image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14771–14780.
- [34] M. Lu, F. Chen, S. Pu, and Z. Ma, "High-efficiency lossy image coding through adaptive neighborhood information aggregation," *arXiv preprint arXiv:2204.11448*, 2022.
- [35] J. Wang, Y. Duan, X. Tao, M. Xu, and J. Lu, "Semantic perceptual image compression with a laplacian pyramid of convolutional networks," *IEEE Transactions on Image Processing*, vol. 30, pp. 4225–4237, 2021.
- [36] M. Akbari, J. Liang, J. Han, and C. Tu, "Learned multi-resolution variable-rate image compression with octave-based residual blocks," *IEEE Transactions on Multimedia*, vol. 23, pp. 3013–3021, 2021.
- [37] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, and Y. Wang, "End-to-end learnt image compression via non-local attention optimization and improved context modeling," *IEEE Transactions on Image Processing*, vol. 30, pp. 3179–3191, 2021.
- [38] J.-H. Kim, B. Heo, and J.-S. Lee, "Joint global and local hierarchical priors for learned image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5992–6001.
- [39] Y. Qian, M. Lin, X. Sun, Z. Tan, and R. Jin, "Entroformer: A transformer-based entropy model for learned image compression," in *International Conference on Learning Representations*, 2022.
- [40] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *International Conference on Learning Representations*, 2022.
- [41] A. B. Koyuncu, H. Gao, A. Boev, G. Gaikov, E. Alshina, and E. Steinbach, "Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression," in *European Conference on Computer Vision*. Springer, 2022, pp. 447–463.
- [42] L. Theis, W. Shi, A. Cunningham, and F. Huszar, "Lossy image compression with compressive autoencoders," in *International Conference on Learning Representations*, 2017.
- [43] Z. Sun, Z. Tan, X. Sun, F. Zhang, Y. Qian, D. Li, and H. Li, "Interpolation variable rate image compression," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 5574–5582.
- [44] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.
- [45] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.
- [46] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.
- [47] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *arXiv preprint arXiv:1807.03039*, 2018.
- [48] A. N. Gomez, M. Ren, R. Urtasun, and R. B. Grosse, "The reversible residual network: Backpropagation without storing activations," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 2211–2221.
- [49] J.-H. Jacobsen, A. Smeulders, and E. Oyallon, "i-revnet: Deep invertible networks," *arXiv preprint arXiv:1802.07088*, 2018.
- [50] M. Xiao, S. Zheng, C. Liu, Y. Wang, D. He, G. Ke, J. Bian, Z. Lin, and T.-Y. Liu, "Invertible image rescaling," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 126–144.
- [51] W. Sweldens, "The lifting scheme: A construction of second generation wavelets," *SIAM journal on mathematical analysis*, vol. 29, no. 2, pp. 511–546, 1998.

- [52] H. Ma, D. Liu, R. Xiong, and F. Wu, “iwave: Cnn-based wavelet-like transform for image compression,” *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1667–1679, 2020.
- [53] S. Li, W. Dai, Z. Zheng, C. Li, J. Zou, and H. Xiong, “Reversible autoencoder: A cnn-based nonlinear lifting scheme for image reconstruction,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 3117–3131, 2021.
- [54] J. Liu, G. Lu, Z. Hu, and D. Xu, “A unified end-to-end framework for efficient deep image compression,” *arXiv preprint arXiv:2002.03370*, 2020.
- [55] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [56] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. Springer, 2015, pp. 234–241.