High-Fidelity Variable-Rate Image Compression via Invertible Activation Transformation

Shilv Cai, Zhijun Zhang, Liqun Chen*, Luxin Yan, Sheng Zhong, and Xu Zou. 2022. High-Fidelity Variable-Rate Image Compression via Invertible Activation Transformation. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503161.3547880

ABSTRACT

Learning-based methods have effectively promoted the community of image compression. Meanwhile, variational autoencoder (VAE) based variable-rate approaches have recently gained much attention to avoid the usage of a set of different networks for various compression rates. Despite the remarkable performance that has been achieved, these approaches would be readily corrupted once multiple compression/decompression operations are executed, resulting in the fact that image quality would be tremendously dropped and strong artifacts would appear (see Figure 1). Thus, we try to tackle the issue of high-fidelity fine variable-rate image compression and propose the Invertible Activation Transformation (IAT) module. We implement the IAT in a mathematical invertible manner on a single rate Invertible Neural Network (INN) based model and the quality level (QLevel) would be fed into the IAT to generate scaling and bias tensors. IAT and QLevel together give the image compression model the ability of fine variable-rate control while better maintaining the image fidelity. Extensive experiments demonstrate that the single rate image compression model equipped with our IAT module has the ability to achieve variable-rate control without any compromise. And our IAT-embedded model obtains comparable rate-distortion performance with recent learning-based image compression methods. Furthermore, our method outperforms the state-of-the-art variable-rate image compression method by a large margin, especially after multiple re-encodings.

CCS CONCEPTS

• Computing methodologies → Artificial intelligence.

KEYWORDS

Image Compression; Variable-Rate; Fidelity Maintenance

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

1 INTRODUCTION

Lossy image compression is one crucial technology due to the increasing volume of visual data in such a multimedia explosion era. This task aims at lowering data redundancy while maintaining visual fidelity and supporting efficient data storage and transmissions. To this end, many classical image compression standards (e.g., JPEG [49], JPEG2000 [41], Webp [17], BPG [7], and Versatile Video Coding (VVC) [26]) have been proposed and widely used in practical applications. Recently, learning-based image compression methods have started to show superiority in terms of common metrics, e.g., PSNR and MS-SSIM. These methods [4, 5, 45] make use of the powerful nonlinear transformation capability of DNNs, and perform end-to-end learning by a large number of high-quality images with a rate-distortion cost. However, despite the exciting progress, the learning-based image compression still remains challenging once variable-rate compression adaptation is needed. Most of them require training multiple single-rate models for different rates, resulting in a high cost of model storage and training.

To remedy the issue, a large number of VAE-based variable-rate image compression methods [9, 11, 13, 43, 44, 55] have been proposed. The researchers first try to achieve discrete rate adaptation using one single model. Choi et al. [11] introduced conditional convolution and achieved variable rate through two-stage training. Yang et al. [55] proposed the modulated autoencoder and achieved discrete adjustable compression rates by different Lagrange multipliers. Chen et al. [9] inserted a set of scaling factors directly before the quantizer to achieve the discrete adjustable compression rates. However, the performance of these methods would be dropped when conducting finer variable-rate control. Thus, the topic of fine rate adaptation has attracted more attention recently. Sun et al. [44] obtained continuously adjustable compression rate by linear interpolation. Cui et al. [13] achieved continuous compression rate control by exponential interpolation. Song et al. [43] conditioned on quality map and achieved the variable rate, which requires semantic segmentation labels for training. Though these methods have the ability of fine variable-rate compression control,
they need additional gain modules or semantic labels to maintain the performance.

Besides, it would be particularly interesting for a variable-rate compression model if the fidelity of images could be maintained while being transmitted multiple times between numerous entities under various compression rates, especially in the current multimedia society (e.g., one person may download a compressed image from Instagram and then send it to his friend via WhatsApp under another re-encoding). However, state-of-the-art VAE-based variable-rate approaches (e.g., Song et al. [43]) would be readily corrupted once multiple compression/decompression operations are executed, resulting in the fact that image quality would be tremendously dropped. Strong artifacts and color shifts would appear, as shown in Figure 1. The main reason is that the autoencoder transforms the image to a low-dimensional latent space and irreversibly discards information before quantization, imposing an implicit limitation on the reconstruction quality. To alleviate information loss, Invertible Neural Networks [19, 54] have gained much attention to effectively preserve fidelity. It is worth noting that VAE-based variable-rate methods cannot be directly fused into the INN-based framework since implementations of their variable-rate control do not satisfy the bijective mapping property. Nevertheless, there is no research on INN-based variable-rate methods to the best of our knowledge. Inspired by this, we construct a variable-rate image compression model which can maintain the fidelity, especially after multiple re-encodings, by exploring the invertibility.

To sum up, we propose an Invertible Activation Transformation (IAT) module based on the INN framework. This module exhibits a mathematical invertible property to avoid discarding any information in the latent space to maintain high fidelity. Notice that it is the initial work to extend the mathematical invertibility to the variable-rate image compression. Moreover, the proposed image compression method attempts to achieve finer control of multiple variable rates, by presenting a compatible tensor-based Lagrange multiplier to train the whole model. The contributions of our proposed method are 3-folded:

- We propose an effective yet neat framework, equipped with the INN-based Invertible Activation Transformation (IAT) module, to achieve the high fidelity of reconstructed images, especially after multiple variable-rate image compression/decompression operations, in a mathematical invertible manner. This issue is rarely investigated so far.
- The proposed model tuned rate-distortion loss and achieved fine variable-rate control through the quality level.
- Extensive experiments demonstrate the superiority of our proposed methods in rate-distortion performance, fidelity maintenance, and fine rate adaptation over three datasets, including Kodak [12], CLIC [47], and DIV2K [1].

## 2 RELATED WORK

In recent years, the application of neural networks in image compression has attracted widespread attention. The variational autoencoder (VAE) [4, 5, 8, 10, 18, 21, 29, 34, 35, 37, 38, 45, 57, 19, 20, 33, 54] and Generative Adversarial Networks (GAN) [2, 24, 36, 42, 52] based methods have achieved surprising results.

### 2.1 Learned Single Rate Image Compression

The VAE-based framework is used as a nonlinear transformation coding model, which is the main approach in the learned image compression method. The works [4, 5, 45] were the first to use CNN for end-to-end image compression and inspired many learning-based image compression methods. The work [5] introduced a hyperprior entropy model to capture the zero-mean Gaussian distribution of the latent representations. The works [29, 37] used the Gaussian
2.2 Learned Variable Rate Image Compression

Initially, LSTM networks [25, 46, 48] control different compression rates by the different number of iterations. The more iterations, the clearer the reconstructed image would be. However, the LSTM-based approach cannot outperform JPEG2000 [41] in rate-distortion performance and would not obtain continuous compression rates. In addition, the iterative procedure is very time-consuming and thus not suitable for practical applications. Then Choi et al. [11] introduced conditional convolution in the autoencoder framework to achieve variable-rate adaptation with a single model through two-stage training.

However, while variable rate is achieved, the rate-distortion performance degrades and there is a dilemma in choosing the appropriate Lagrange multiplier and quantization step size for forward inference. Yang et al. [55] proposed a modulated autoencoder that achieves discrete adjustable compression rate with a single model by different Lagrange multipliers. Thesis et al. [45] first trained the model with high bits per pixel(bpp) and then fixed the network model parameters to train the scaling parameters for different compression rates. However, the network model suffered from incongruity with the scaling parameters, especially in low bpp cases. Chen et al. [9] inserted a set of scaling factors directly before the quantizer to achieve the discrete variable compression rate.

Recently, research has been conducted on continuous compression rate adjustable [13, 43, 44]. The work [13] introduced a series of vector pairs for coarse compression rate control, and then achieve continuous compression rate control by exponential interpolation. Sun et al. [44] extended the work [11], which obtained a continuously adjustable compression rate by linear interpolation. Song et al. [43] conditioned the quality map by spatial feature transform (SFT) [51] to control different compression rates.

VAE-based variable-rate approaches have been extensively researched. However, those methods suffer from severe information distortion after multiple operations of compression/decompression for the same image. The distortion becomes more explicit as the number of operations increases.

2.3 Invertible Neural Networks

Invertible neural networks (INNs) are generative models that transform complex distributions into simple ones, allowing for accurate and efficient probability density estimation. INNs have a bijective mapping of input and output, which is ideal for image compression. NICE [14] introduced a flexible architecture that can learn highly nonlinear bijective transformations to represent data with simple distributions. Based on NICE [14], RealNVP [15] further extended the idea of hierarchical and combinatorial transformations, which used affine coupling and a multi-scale framework. Kingma et al. [28] proposed a generative flow model based on a 1 x 1 invertible convolutional network with a significant improvement in log-likelihood on a standard benchmark dataset, having the advantages of exact controllability of log-likelihood, the tractability of exact inference.
of latent representations, and parallelizability of training and synthesis. Aridzzone et al. [3] demonstrated that the validity of INNs is suitable not only for synthetic data but also for two practical applications in medicine and astrophysics. SRFlow [32] has designed a conditional normalizing flow architecture to solve the ill-posed problem in the super-resolution task. Xiao et al. [53] proposed an invertible rescaling network (IRN), which constructs a bijective transform to effectively implement the reconstruction of low-resolution images into high-resolution images.

INN greatly alleviates the information loss problem for better image compression, as in [19, 20, 33, 54]. But no one has specifically studied variable-rate image compression with a single model based on the INN framework.

3 METHODOLOGY

3.1 Framework

Our image compression approach is depicted in Figure 2. The proposed method implements fine variable-rate modulation in an invertible neural network framework, which involves the invertible activation transformation (IAT) module to control different compression rates through different quality levels. We present the detailed procedure of the model in the following: Firstly, the source image $x \in \mathbb{R}^{3 \times H \times W}$ is enhanced by the dense block module [23] to generate a nonlinear representation of $u \in \mathbb{R}^{3 \times H \times W}$, where $H$ and $W$ denote the height and width of the input image respectively. Then the forward pass of the Invertible Neural Network section, which is equipped with the proposed IAT module, transforms $u$ to a latent representation, conditioned on the quality level $L \in \mathbb{R}^{H \times W}$ to control the compression rate. This latent representation would be further fed into the Attention Channel Squeeze module to reduce the number of channels and obtain the potential representation $y$. This procedure could be formulated by a parametric analysis transform function, i.e.,

$$y = g_{a}(x, L), \quad (1)$$

the discrete latent features $\hat{y}$ are obtained by quantification of $y$, i.e., $\hat{y} = Q(y)$. We use the quantizer $Q(\cdot)$ in Ballé et al. [5] to model the quantized latent representation $\hat{y}$ approximately by adding the uniform noise $U(-0.5, 0.5)$ to the latent representation $y$ during training and rounding the latent representation $y$ during testing. The context entropy model generates parameters $\mu$ and $\sigma$ of the Gaussian entropy model that approximates the distribution of quantified latent representation $\hat{y}$ to support the entropy encoding. We use range asymmetric numeral system [16] to losslessly compress latent representation $\hat{y}$ and $\hat{z}$ into bitstreams.

The inverse calculation takes the quantified latent representation $\hat{y}$ and the quality level $L$ as the input, and reconstructs the decompressed images by a parametric synthesis transform, which is formulated as follows:

$$\hat{s} = g_{s}(\hat{y}, L). \quad (2)$$

3.2 Invertible Activation Transformation

We proposed the invertible activation transformation (IAT) module to enhance the invertible neural network, which efficiently generates the desired compressed representation conditional on the quality level $L$. The proposed IAT module can achieve variable-rate adaption on a single model while maintaining the image fidelity, especially after multiple compression/decompression operations, in a mathematical invertible manner.

The forward transform of the IAT module is illustrated by pink arrows on the top of Figure 3. The inputs are the quality level $L$ and the feature $s$. The element-wise activation parameters $y \in \mathbb{R}^{c \times h \times w}$ and $\beta \in \mathbb{R}^{c \times h \times w}$ are then calculated by the IAT module from the quality level $L$ via convolutional operations. These activation parameters would be applied to the feature $s$ via the Equation 3 to generate the feature $e$.

$$e = (s \odot \beta) \oplus y, \quad (3)$$

where $\odot$ denotes the Hadamard product and $\oplus$ denotes the addition by element. $c, h,$ and $w$ are the channel, height, and width of the feature, respectively.

The inverse transform of the IAT module is illustrated by green arrows at the bottom of Figure 3. The input quality level $L$ and features $\hat{e}$ are applied to obtain the feature $\hat{s}$. This inverse transform is formulated by Equation 4,

$$\hat{s} = (\hat{e} \odot y) \ominus \beta, \quad (4)$$

where $\ominus$ denotes the subtraction in element-order, $\odot$ denotes the division by elemental order. Once the quality level $L$ is the same in both forward and inverse procedures, the invertibility of the operation between the features $s$ and $e$ can be guaranteed.

In the previous work [9], a set of scaling factors was inserted directly before the quantizer to achieve the discrete adjustable compression rate. In our algorithm, the activation parameters are element-wise, which means that IAT module is computed as a spatial feature transform rather than a simple channel weighting. Moreover, the IAT module is attached after each invertible block which is initially proposed in RealNVP [15] and adopted by baseline model [54], not just inserted before the quantizer. These adjustments not only make fine variable-rate adaptation available but also turn out to the better performance, the experiment "Impact of the QLevel Representation" in section 4.4 shows its effectiveness, and the results are shown in Figure 7.
3.3 Fine Variable-Rate Control
Unlike interpolation-based methods [13, 44] for obtaining finer compression rates, our method achieves the fine compression rate adaptation directly by modulating the quality level \( L \), which is more convenient when controlling the compression rate by only one parameter instead of two. Compared to Song et al. [43], our method does not require additional semantic labels, either.

The goal of lossy image compression is to minimize the length of the bits stream and the distortion between the source image \( x \) and the reconstructed image \( \hat{x} \). The optimization function is always expressed in the rate-distortion loss: \( R + \lambda D \), where \( \lambda \) is the Lagrange multiplier which determines the trade-off between the rate \( R \) and the distortion \( D \). Theoretically, as long as the set of Lagrangian multiplier \( \lambda \) is large enough, it is possible to achieve fine compression rate control, but in practice, the computational cost is too high. For interpolation-based methods, the Lagrangian multiplier \( \lambda \) is a scalar. Thus, at each iteration during training, only one element in a finite set of \( \lambda \) would be randomly selected for optimization. In order to further promote the R-D performance of our model, we use a tensor instead of the scalar \( \lambda \). Our optimization function implements fine variable-rate control by minimizing the rate-distortion loss \( R + \Lambda \odot \bar{D} \), where dimensions of \( \Lambda \in \mathbb{R}^{C \times H \times W} \) and the distortion \( \bar{D} \in \mathbb{R}^{C \times H \times W} \) are the same as the dimension of the original input image. \( \odot \) denotes the Hadamard product. In this formulation, \( \Lambda \) is a tensor and no longer a finite set of constant scalars. Thus, \( \bar{D} \) measures pixel-wise distortion and is defined as \( \bar{D} = \sum_{i=1}^{T} \frac{\lambda_i (x_i - \hat{x}_i)^2}{T} \). \( T \) indicates the number of image pixels, \( \lambda_i \) is the Lagrangian multiplier, \( x_i \) and \( \hat{x}_i \) denote one pixel of the original and reconstructed image, respectively.

\( \Lambda \) is simply calculated from the quality level \( L \) via a monotonically increasing function: \( \Lambda = V(L) \), where \( V : \mathbb{R}^N \rightarrow \mathbb{R}^T \), \( V(L) = \theta \cdot e^{\tau L} \), \( \theta = 0.0012 \), \( \tau = 4.382 \), the process of dimensioning from \( \mathbb{R}^N \rightarrow \mathbb{R}^T \) is done by direct replication between channels. \( L = \{l_i\}_{i=1}^{N}, l_i \in [0, 1] \), \( N = H \times W, T = C \times H \times W \). \( C, H, \) and \( W \) denote the channel, height, and width of the source image \( x \), respectively. Under such a paradigm, we implement this pixel-wise distortion constraint by randomly generating values of each element of the tensor \( \Lambda \) via the quality level \( L \) during training. This is equivalent to increasing the number of \( \lambda \) values selected at each iteration. So, the fine variable-rate control can be obtained by feeding exact quality levels during the testing.

As in other learning-based method [5], the log-likelihood of the coded features \( \hat{y} \) is estimated by a probabilistic model to replace the true compression rate \( R \). Finally, the training loss would be:

\[
\text{Loss} = -\log p_g(\hat{y} | \hat{z}) - \log p_p(\hat{z} | \Lambda) + \frac{\sum_{i=1}^{T} \lambda_i (x_i - \hat{x}_i)^2}{T},
\]

(5)

where \( \hat{y} \) and \( \hat{z} \) are quantized latent representations and side information respectively. \( p_g(\hat{y} | \hat{z}) = \mathcal{N}(\mu, \sigma^2) \), \( \mu \) and \( \sigma \) denote the estimates of the mean and standard deviation of the quantified latent representation \( \hat{y} \). \( p_p(\hat{z} | \Lambda) = \mathcal{N}(\mu_1, \sigma_1^2) \), \( \mu_1 \) and \( \sigma_1 \) denote the estimates of the mean and standard deviation of the quantified side information \( \hat{z} \). The side information usually represents the hyper-prior originally proposed in [5] and refers to the extra stream \( \hat{z} \) generated by the "Context Entropy Model" in Figure 2. It is worth noting that this loss function would be degraded to the standard rate-distortion optimization function if all elements of the tensor quality level \( L \) are the same.

In addition, our method can be trained on arbitrary unlabeled data instead of requiring semantic segmentation labels corresponding to the original data, which is different from Song et al. [43], for training the model.

4 EXPERIMENTS
4.1 Implementation Details
Details For Training. In our implementation, the network of Xie et al. [54] is adopted as our basic architecture. The training datasets contain Flicker 2W [31] and COCO [30]. Flicker is used to train the network which has the context model, COCO is used to train the network without the context model. Our network is trained on 256x256 randomly cropped patches and discards images less than 256px in height or width during data pre-processing. In training, the quality level \( L \) needs to be sent to the INN section as a condition during the forward and inverse transform. The quality level \( L \) takes a uniform value tensor between (0,1) during the training and is randomly sampled between (0,1) during the training. Our implementation relies on Pytorch [39] and an open-source CompressAI PyTorch library [6]. All experiments were conducted on RTX 3090 GPU and trained for about 2.5M iterations with batch size 8. Adam optimizer [27] is used to optimize the parameters, there were multistage learning rates \{1e−4, 5e−5, 1e−5, 5e−6, 6.1e−6, 5e−7\} that changed with boundaries \{1000000, 1300000, 1600000, 1900000, 2200000, 2500000\}.

Details For Testing. We evaluate the rate-distortion performance on three commonly used datasets. The Kodak [12] contains 24 lossless images with a size of 768 x 512. The CLIC Professional Validation dataset [47] comprises 41 high-quality images with much higher resolution. The DIV2K validation dataset [1] contains 100 images with high resolutions of 2K. We draw curves based on the rate-distortion performance to compare the coding efficiency of different methods. We also calculate the area under the rate-distortion curve to observe the performance difference more effectively.

4.2 Rate-Distortion Performance
To verify the validity of the proposed approach, we conduct rate-distortion (RD) performance experiments on three datasets, i.e., Kodak [12], CLIC [47], and DIV2K [1]. We compare our approach with five recent state-of-the-art learning-based image compression methods [10, 22, 43, 50, 54] and two classical codec methods, BPG [7] and VCC [26]. The results of learning-based methods are collected from their official GitHub pages or their papers. The VCC approach is implemented by the official Test Model VTM 12.1 with the intra-profile configuration from the official GitHub page to test images. Both VVC and BPG software were configured with the YUV444 format to maximize compression performance.

All comparable results are demonstrated in Figure 4. It is seen that our approach achieves the best results with commonly used metrics PSNR and MS-SSIM on three datasets. Compared with the baseline method [54], our approach achieves comparable R-D performance on the Kodak dataset (Figure 4 (a)(d)) and outperforms the baseline on both the CLIC dataset (Figure 4 (b)(e)) and the DIV2K dataset (Figure 4 (c)(f)). This means that our approach achieves
the variable-rate adaptation based on the single rate method [54] without sacrificing any performance, verifying the effectiveness of the IAT module. It is worth noting that the CLIC dataset and DIV2K dataset are high-resolution images, implying that our method is more effective on high-resolution images. Our approach empowers the network model with variable-rate in addition to improving the algorithmic performance of the original model. To further compare the performance between the baseline [54] and our method, we calculate their corresponding area under curve (AUC) values, as shown in Table 1. The results show that our approach outperforms the single rate model method by Xie et al. [54] in terms of the aggregated AUC metric.

In addition, our proposed method could achieve variable-rate image compression with a fine granularity. To verify the effectiveness of fine variable-rate control, we illustrate multiple performances of fine variable-rate control within the low and high bpp range in Table 2. In practice, classical image codecs provide hundreds of variable-rate RD points to meet the basic requirement of the application. Compared with that, our method obtains at least 1000 effective variable-rate RD points with a very fine PSNR and MS-SSIM. We achieved the fine-rate control compared with the classical image codecs BPG [7] and VTM 12.1 [26], the comparative results refer to the supplementary material.

Table 1: Area under curve (AUC) of our method and Xie et al. [54](Baseline) on different datasets about PSNR and MS-SSIM. The bpp range is determined by the intersection of two methods. Our approach makes a single-rate baseline compression model achieve the variable-rate ability and even outperforms the baseline in R-D performance.

| Dataset  | Xie et al. [54] | Ours |
|----------|----------------|------|
|          | AUCPSNR | AUCCMS−SSIM | AUCPSNR | AUCCMS−SSIM |
| Kodak    | 32.7866 | 16.5030 | 32.7883 | 16.5036 |
| CLIC     | 23.5896 | 11.7463 | 23.7082 | 11.8571 |
| DIV2K    | 28.0998 | 14.7868 | 28.2138 | 14.8901 |

4.3 Fidelity for Re-encoding

In order to verify the high fidelity, our method is compared with the latest VAE-based variable-rate method by Song et al. [43]. This method does not use context model and has available source code. To make a fair comparison, we remove the context model and add the non-local attention module [8] to the hyperprior layer. Figure 5 illustrates the results of multiple compression/decompression operations on the same image with different compression rates. With operations increasing, our proposed method shows higher fidelity.
Figure 5: Qualitative results after different numbers of compression/decompression operations under various rates. The two images (kodim1.png and alexander-shustov-73.png) are from the Kodak dataset and the CLIC dataset, respectively. Severe artifacts and color shifts would appear in the state-of-the-art VAE-based approach [43] once multiple operations are executed, in contrast to better fidelity maintenance of our approach. Please refer to the supplementary material for more cases. N indicates the number of compression/decompression operations. Best viewed in color.

Table 2: Variable-rate control experiments over the Kodak dataset. Our approach can finely control the compression rate within the whole bpp range (no matter low or high).

| LOW    |          | HIGH     |          |
|--------|----------|----------|----------|
| BPP    | PSNR(dB) | MS-SSIM(dB) | BPP    | PSNR(dB) | MS-SSIM(dB) |
| 0.28181| 31.6951  | 14.5015  | 1.02433 | 38.3226  | 21.2580  |
| 0.28265| 31.7071  | 14.5153  | 1.02587 | 38.3312  | 21.2664  |
| 0.28342| 31.7177  | 14.5263  | 1.02733 | 38.3388  | 21.2717  |
| 0.28416| 31.7291  | 14.5377  | 1.02910 | 38.3468  | 21.2819  |
| 0.28500| 31.7435  | 14.5517  | 1.03071 | 38.3548  | 21.2903  |
| 0.28576| 31.7538  | 14.5639  | 1.03250 | 38.3625  | 21.2995  |
| 0.28659| 31.7657  | 14.5765  | 1.03406 | 38.3703  | 21.3087  |
| 0.28734| 31.7761  | 14.5874  | 1.03564 | 38.3767  | 21.3190  |
| 0.28808| 31.7884  | 14.5952  | 1.03733 | 38.3872  | 21.3291  |
| 0.28880| 31.8004  | 14.6092  | 1.03885 | 38.3943  | 21.3355  |

Figure 6 (a)(c) show the rate-distortion performance after multiple operations of compression/decompression with different compression rates. Both approaches change from high to low bpp ranges, our method in the set of bpp \{1.0267, 1.0116, 0.9949, 0.9784, 0.9619, 0.9456, 0.9292, 0.9127, 0.8965, 0.8809, 0.8658, 0.8507, 0.8357, 0.8206, 0.8056, 0.7907 \}; Song et al. [43] in the set of bpp \{1.0392, 1.0249, 1.0091, 0.9932, 0.9768, 0.9606, 0.9449, 0.9287, 0.9128, 0.8968, 0.8813, 0.8658, 0.8505, 0.8351, 0.8201, 0.8052\}. It is clearly seen that our method outperforms Song et al. [43], after multiple compression/decompression operations. Figure 6 (b)(d) show the rate-distortion performance by multiple operations with the fixed compression rate. Both approaches achieve a bit rate of 0.791 bpp for all steps. Also, our method achieves better results significantly, compared with Song et al. [43] and baseline [54]. The results indicate that our proposed IAT module is powerful to maintain image fidelity, which is important for practical applications. It is noteworthy that compression methods capable of high-fidelity in re-encoding are of great importance in the video production pipeline, as image/video content may be edited/composited by different people or at different times, requiring re-encodings in the process.

4.4 Discussion

Impact of the QLevel Representation. To further analyze the effectiveness of the tensor-based QLevel representation of our IAT module, we conducted an ablation study by modifying the quality level representation. We compared our approach with the baseline method [54] and the simplified version of our method, which modifies the quality level from tensor to scalar, similar to [9]. Comparative results are shown in Figure 7. The results indicate that the proposed tensor-based quality level obtains better performance,
compared with the scalar factor ones, which only provides channel-wise weighted computations on latent representation.

**Impact of Gain Components.** The context model [29, 35, 37] and the non-local attention module [8] are commonly used in the learned-based image compression methods to further reduce statistical redundancy within the latent features and improve the probabilistic estimation ability of the network. We conduct an ablation study to evaluate the impact of the context model and non-local attention module on our method in the Kodak dataset, as shown in Figure 8. We start from a baseline without the context model and non-local attention module, *i.e.*, W/O CM (context model) and W/O NLAM (non-local attention module), and plot the rate-distortion performance in green color. Then, we add the non-local attention module (blue color) and context model (red color) to evaluate the performance. We can observe that using the context model achieves the best results, while it requires high computational costs (codec process takes about 233 seconds on an Intel(R) Core(TM) i9-10900K CPU). In addition, our method outperforms Song et al. [43] without the context model and non-local attention module, demonstrating the effectiveness of the proposed method.

![Figure 6: Successive re-encodings on the Kodak dataset. (a) and (c): Compression rates of each compression/decompression operation are different. (b) and (d): The compression rate is fixed. Our approach outperforms baseline [54] and Song et al. [43] (a SOTA variable-rate approach) by a large margin to show the superiority of fidelity maintenance especially when multiple operations are executed.](image)

![Figure 7: Impact of the QLevel Representation. The scale factor method (green line) is similar to Chen et al. [9]. Our proposed tensor-based QLevel representation achieves better performance than simply using a scalar to control the compression rate.](image)

![Figure 8: Impact of Gain Components. W/O represents ‘without’, W represents ‘with’, CM represents ‘context model’, and NLAM represents ‘non-local attention module’.](image)

**5 CONCLUSION**

In this paper, we propose a high-fidelity variable-rate image compression method by introducing the Invertible Activation Transformation (IAT) module. The IAT module, implemented in a mathematical invertible manner, as a feature activation transform layer of the invertible neural network, has the ability of fine variable-rate control by feeding the quality level (QLevel) to generate the scaling and bias tensors while better maintaining the image fidelity. Extensive experiments demonstrate that the single rate model equipped with our IAT module is able to achieve fine variable-rate control without any performance compromise. Thanks to the mathematical invertibility of our approach, fewer artifacts or color shifts would have appeared and the fidelity of reconstructed images is better maintained, especially when multiple re-encodings are executed under various compression rates.

**ACKNOWLEDGMENTS**

This work was supported by the National Natural Science Foundation of China (NSFC) grant No. 62176100, the Special Project of Science and Technology Development of Central guiding Local-Central Guidance on Local Science and Technology Development Fund of Hubei Province grant 2021BEE056 and the National Key Laboratory Foundation of China grant No. 6142113200307.
A SUPPLEMENTARY MATERIAL

A.1 Example Images

This section holds all the results of our tests on the Kodak dataset and examples of the CLIC Professional Validation dataset (Due to file size limitation of supplementary material.), with each set of plots indicating, from left to right, 1, 16, and 31 compression/decompression operations. Please refer to Figure 9 ∼ Figure 20.

A.2 Compared with classical codecs

Figure 21 realizes fine rate control compared with the classical image codecs. Figure 21 also shows that our proposed method provides better reconstruction quality on PSNR, MS-SSIM, with fewer artifacts in visual perception.

Figure 9: (a) is kodim02 and (b) is kodim03 in Kodak dataset

Figure 10: (a) is kodim05 and (b) is kodim06 in Kodak dataset

Figure 11: (a) is kodim07 and (b) is kodim08 in Kodak dataset

Figure 12: (a) is kodim11 and (b) is kodim12 in Kodak dataset

Figure 13: (a) is kodim13 and (b) is kodim14 in Kodak dataset

Figure 14: (a) is kodim15 and (b) is kodim16 in Kodak dataset

Figure 15: (a) is kodim20 and (b) is kodim21 in Kodak dataset

Figure 16: (a) is kodim22 and (b) is kodim24 in Kodak dataset

Figure 17: (a) is felix-russell-saw-140699 and (b) is amy-zhang-15940 in the CLIC dataset

Figure 18: (a) is roberto-nickson-48063 and (b) is benjaminsloth-lindgreen-705 in the CLIC dataset
Figure 19: (a) is kodim04, (b) is kodim09, (c) is kodim10, and (d) is kodim17 in Kodak dataset.

Figure 20: (a) is kodim18 and (b) is kodim19 in Kodak dataset.

Figure 21: Visualization of sample images (ales-krivec-15949.png from CLIC dataset) reconstructed by BPG, VTM12.1, and Ours. The quantizer parameters (QPs) are used in BPG and VTM 12.1 to realize variable rate control. By adjusting the quality level (QL), our proposed method matches the rates of BPG and VTM 12.1 while outperforming them.