

Learning a Single Tucker Decomposition Network for Lossy Image Compression with Multiple Bits-Per-Pixel Rates

Jianrui Cai, Zisheng Cao, and Lei Zhang, *Fellow, IEEE*

Abstract—Lossy image compression (LIC), which aims to utilize inexact approximations to represent an image more compactly, is a classical problem in image processing. Recently, deep convolutional neural networks (CNNs) have achieved interesting results in LIC by learning an encoder-quantizer-decoder network from a large amount of data. However, existing CNN-based LIC methods generally train a network for a specific bits-per-pixel (bpp). Such a “one-network-per-bpp” problem limits the generality and flexibility of CNNs to practical LIC applications. In this paper, we propose to learn a single CNN which can perform LIC at multiple bpp rates. A simple yet effective Tucker Decomposition Network (TDNet) is developed, where there is a novel tucker decomposition layer (TDL) to decompose a latent image representation into a set of projection matrices and a core tensor. By changing the rank of core tensor and its quantization, we can easily adjust the bpp rate of latent image representation within a single CNN. Furthermore, an iterative non-uniform quantization scheme is presented to optimize the quantizer, and a coarse-to-fine training strategy is introduced to reconstruct the decompressed images. Extensive experiments demonstrate the state-of-the-art compression performance of TDNet in terms of both PSNR and MS-SSIM indices.

Index Terms—Lossy Image Compression, Convolutional Neural Networks, Tucker Decomposition

I. INTRODUCTION

AS an indispensable step in many image processing applications, lossy image compression (LIC) is a classical yet still active topic. The goal of LIC is to reduce the image storage space without sacrificing much the image quality, and thus provide an economic solution to image storage and transmission systems. Recently, with the development of portable imaging devices and social media (*i.e.*, Facebook), billions of images are transmitted and stored daily on social networks [1]. The explosive growth of the amount of shared images on Internet raises higher requirements on LIC for more effective visual communication systems.

A typical LIC system contains mainly three modules: transformation (*e.g.*, an encoder and a corresponding decoder), quantization (*e.g.*, a quantizer), and encoding. To compress an image into bitstreams, conventional LIC methods firstly apply predefined transformations to transform an image into a sparse

domain, then perform lossy quantization on the transformed coefficients, followed by entropy coding [2]. Notwithstanding their demonstrated success, conventional LIC methods (*i.e.*, JPEG [3] and JPEG 2000 [4]) do not work very well for low bitrate image compression and may introduce visible artifacts such as blurring and ringing.

Deep convolutional neural networks (CNNs) have recently led to a series of breakthroughs in many vision problems [5], [6], [7], [8], [9]. The flexible non-linear modelling capability and powerful end-to-end training paradigm of CNN also make it a promising new approach to LIC. In the last several years, a flurry of CNN-based LIC methods have been proposed, including the study of network structures [10], [11], [12], [13], [14] as well as loss functions [15], [16], [17], [18]. Firstly, the end-to-end training manner enables CNN-based LIC systems to adaptively learn an effective encoder-decoder pair from a large amount of image data and in a larger context to represent more complex image structures, reducing the artifacts in the decompressed image. Secondly, by adopting specific loss functions (*i.e.*, perceptual metrics) in the training, the CNN compressors are able to strengthen certain desired aspects (*i.e.*, perceptual quality) of the decomposed image.

Despite the advantages of employing CNN for compression, there are still some challenges which limit the performance of CNN-based compressors. First, existing CNN-based LIC methods generally change the number of latent feature maps and/or quantized values to adjust the bpp rate. As a result, the network is trained dedicatedly for a specific bpp rate once at a time. Such a “one-network-per-bpp” problem limits the flexibility and applicability of CNNs to practical image compression systems. Second, because of the non-differentiable property of discrete operation, quantizer is hard to be updated during the end-to-end CNN network training. Therefore, the optimal decision boundaries of quantization levels are almost unreachable. Third, existing CNN based LIC methods usually adopt fixed quantization bins to discretize the latent image representation and treat each element of the latent image representation equally. Such a quantization scheme, however, ignores the prior knowledge that the local content is spatially variant in an image, and restricts the capability of CNNs in compressing complex image structures.

To address the aforementioned issues, in this work, we propose a new paradigm for deep LIC. More specifically, we proposed a deep Tucker Decomposition Network (TDNet) which takes the sparsity/low-rankness of latent image representations into consideration. The key component of TDNet is

This work is supported by Hong Kong RGC GRF grant (PolyU 152124/15E).

J. Cai and L. Zhang are with Department of Computing, The Hong Kong Polytechnic University, Kowloon Hong Kong (e-mail: {csjcai, cs1zhang}@comp.polyu.edu.hk).

Z. Cao is with Imaging Group of DJI Innovations Co., Ltd, Shenzhen, China (e-mail: zisheng.cao@dji.com).

a novel tucker decomposition layer (TDL), which decomposes the latent image representation into a set of projection matrices and a compact core tensor. By changing the rank of core tensor and its quantization levels, we can easily adjust the bpp rate of latent image representation, and thus a single CNN can be trained to compress and reconstruct images under multiple bpp rates. Besides, we propose an iterative non-uniform quantization strategy to obtain the optimal quantization boundaries based on the distribution of encoding coefficients. A coarse-to-fine training strategy is introduced to train a stable TDNet and reconstruct the decompressed images. Extensive experiments demonstrate that, our proposed TDNet trained with the mean-squared error (MSE) loss or the multi-scale structural similarity index (MS-SSIM) [19] loss can yield competitive results with state-of-the-art CNN-based LIC schemes but it uses only a single network to achieve this goal.

The contributions of this work are summarized as follows:

- (1) We propose an end-to-end trainable deep tucker decomposition network, namely TDNet, which, for the first time to the best of our knowledge, enables a single network to perform LIC at multiple bpp rates. Besides, we proposed the “all-in-one” training strategy to ensure the convergence of TDNet training.
- (2) We present an iterative non-uniform quantization scheme to obtain the quantization boundaries of the tensor decomposition coefficients, and adopt a variable-bits quantization scheme to discretize the latent image representation. By assigning different number of quantized bits to different chunks, the proposed TDNet can preserve the major data components and treats each element of the latent image representations differently.
- (3) We introduce the reconstruction sub-network to further enhance the deconvolution output by reproducing the details and textures lost in the encoding and TDL quantization process. The proposed methods demonstrate state-of-the-art PSNR/SSIM indices and visual quality.

The remainder of this paper is organized as follows. Section II provides a brief survey of related work. Section III introduces our proposed TDNet model. Section IV presents in detail the tucker decomposition layer. Section V presents the all-in-one training strategy. In Section VI, extensive experiments are conducted to evaluate TDNet. Finally, several concluding remarks are given in Section VII.

II. RELATED WORK

A. Traditional Lossy Image Compression

The most prevalent LIC method is JPEG (Joint Photographic Experts Group)¹, which first applies discrete cosine transform (DCT) to non-overlapping 8×8 image blocks, and then quantizes the transformed DCT coefficients in frequency domain using a predefined quantization table, followed by entropy coding such as Huffman coding and arithmetic coding [20]. As a significantly improved version of JPEG, JPEG2000 adopts the more powerful discrete wavelet transform (DWT), instead of DCT, to perform time-frequency analysis on images.

More specifically, JPEG2000 adopts the Cohen-Daubechies-Feauveau (CDF) 9/7 wavelet to decompose an image into multiple bands, and performs scalar-quantization on the DWT coefficients, followed by the Embedded Block Coding with Optimal Truncation (EBCOT) [21]. Another powerful LIC scheme is the so-called Better Portable Graphics (BPG) method², which is built upon the intra-frame encoding scheme of the High Efficiency Video Coding (HEVC) video compression standard³. It has been proved that BPG can produce smaller files for a given quality than JPEG and JPEG 2000.

Although these traditional LIC approaches have demonstrated their great success, they predefined transformations (*e.g.*, DCT and wavelet transform) to transform the image into some sparse domain for quantization. These conventional predefined transformations are limited in adaptively and effectively decomposing complex image structures, resulting in visual artifacts around image edges and textures, especially when the bpp rates are low. The deep neural network based LIC methods are then proposed to address these problems.

B. Deep Lossy Image Compression

Recently, deep neural networks have been investigated and achieved promising results in LIC. As a pioneering work, Toderici *et al.* adopted the recurrent neural network (RNN) to encode and decode images of size 32×32 [10], and they further extended the network to compress full-resolution images [11]. Built upon the architecture proposed in [10], [11], Johnston *et al.* [17] modified the recurrent architecture by introducing hidden-state priming to improve spatial diffusion, and replaced the MSE loss by MS-SSIM loss [19] to increase the visual quality of reconstructed images.

Different from the above methods which employ RNN, methods in [12], [13], [14], [15], [18], [16] rely on CNN based auto-encoder architectures. Ballé *et al.* [12] used generalized divisive normalization for joint nonlinearity to implement local gain control. Li *et al.* [14] learned a content-weighted importance map, according to which more bits are allocated to the region with rich content to preserve image edge and texture details. Rippel *et al.* [15] aggregated image information across different scales by exploiting the pyramidal decomposition strategy, and introduced the generative adversarial networks (GANs) [22] to sharpen the edge of reconstructed images. To alleviate the effect of vanishing gradient caused by non-differentiable quantization operation, Theis *et al.* [13] introduced a smooth approximation of the derivative of the rounding function. A soft-to-hard scheme is adopted in [18] to find assignments to the quantizer.

For all the aforementioned deep LIC methods, the bpp rate of latent image representation can only be adjusted by changing the number of latent feature maps and/or quantized values since the output of encoder should have the same size as the input of decoder. Thus, one network can only be trained to deal with a specific bpp rate, making these deep CNN-based LIC methods less flexible. In this work, we introduce a novel tucker decomposition layer into CNN, and present a TDNet

¹<https://jpeg.org/>

²<https://bellard.org/bpg/>

³<https://www.itu.int/rec/T-REC-H.265>

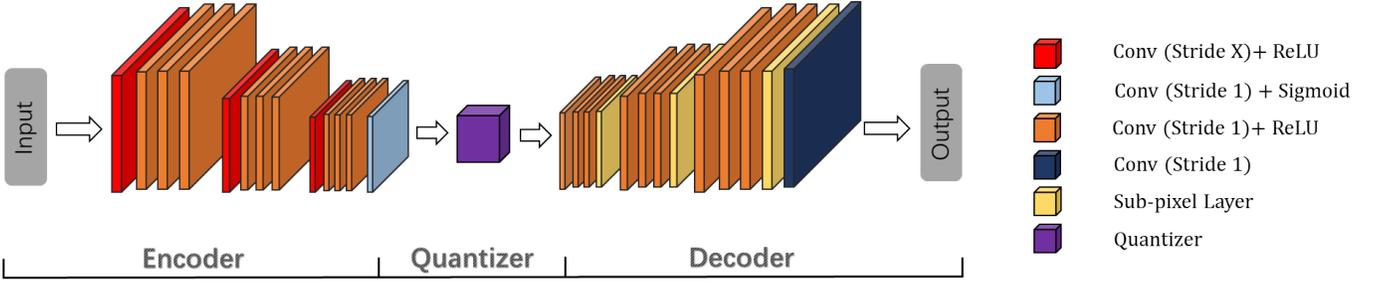


Fig. 1: Illustration of conventional lossy image compression network architecture.

scheme which enables a single network to tackle with multiple bpp rates for LIC.

III. DEEP LOSSY IMAGE COMPRESSION MODEL

In this section, we first summarize the pipeline of conventional CNN-based LIC methods, and then present the pipeline of our proposed TDNet. Finally, we present in detail the network architecture.

A. Overview of Conventional LIC Network Pipeline

Existing deep LIC networks can be generally formulated as a joint rate-distortion optimization process to learn an encoder, a quantizer, and a decoder. The architecture of those networks is shown in Figure 1. Given a set of training images $\{\mathbf{x}_i\}_{i=1}^N$, where N is the total number of training images, deep LIC methods aim to learn a nonlinear analysis transformation encoder $E(\cdot)$, a quantizer $Q(\cdot)$, and a nonlinear synthesis transformation decoder $F(\cdot)$. The encoder $E(\cdot)$ first converts an input image \mathbf{x}_i into a latent feature representation $\mathbf{z}_i = E(\mathbf{x}_i)$. Then, the quantizer $Q(\cdot)$ quantizes the features into discrete values $\hat{\mathbf{z}}_i = Q(\mathbf{z}_i)$, which can be losslessly encoded into a bitstream for transmission or storage. Once the bitstream is received by the decoder $F(\cdot)$, an approximation of the original image is obtained as $\hat{\mathbf{x}}_i = F(\hat{\mathbf{z}}_i)$. Overall, the deep image compression pipeline can be formulated as:

$$\hat{\mathbf{x}}_i = F(Q(E(\mathbf{x}_i, \Omega)), \Phi), \quad (1)$$

where Ω and Φ are the parameters of encoder $E(\cdot)$ and decoder $F(\cdot)$, respectively.

Given a certain compression ratio, the network is expected to learn the parameters Ω and Φ to minimize the distortion of the reconstructed image. Note that the compression ratio can be defined as $\alpha = \frac{C(\mathbf{x}_i)}{C(\hat{\mathbf{x}}_i)}$, where $C(\cdot)$ is the function to calculate the average number of bits to store a pixel of an image. Since $C(\mathbf{x}_i)$ is usually a constant for the original image \mathbf{x}_i without compression, we can adjust $C(\mathbf{z}_i)$ to change the compression ratio α . For most of the existing CNN based LIC methods, one can only change the number of feature maps and quantization levels to adjust $C(\mathbf{z}_i)$ of latent image representation \mathbf{z}_i . As a result, usually a specific network has to be trained for a certain compression ratio or bpp rate. For a new bpp rate, a new network has to be retrained by adjusting the number of latent representation feature maps and quantization levels.

B. Proposed LIC Network Pipeline

Our proposed TDNet is designed to achieve the objective of multiple bpp rates with a single network. The pipeline of TDNet is shown in Figure 2. Instead of directly quantizing the latent image representation into a bitstream as in conventional deep LIC methods, we introduce a novel tucker decomposition layer (TDL) to process the latent image representation. Denote by $T(\cdot)$ the decomposition operation of TDL, and by $T^{-1}(\cdot)$ the inverse operation. Given the latent image representation \mathbf{z}_i , we use $\{\mathbf{Y}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}\} = T(\mathbf{z}_i)$ to decompose the features into 3 orthogonal matrices $\{\mathbf{U}^{(n)}\}_{n=1}^3$ and a core tensor \mathbf{Y} , and then apply the quantizer $Q(\cdot)$ to quantize the decomposed components to generate bitstream. Once the bitstream is received, with $T^{-1}(\cdot)$ and $Q^{-1}(\cdot)$, we can de-quantize and reproduce the features by back-projecting the core tensor and 3 orthogonal matrices into the approximation $\hat{\mathbf{z}}_i = T^{-1}(Q^{-1}(\mathbf{Y}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}))$.

By changing the rank of core tensor in TDL, we can easily adjust the bpp rates and hence the compression ratio α while keeping the size of latent image representation unchanged. Once $\hat{\mathbf{z}}_i$ is received, we can obtain an approximation of the original input image by $\hat{\mathbf{x}}_i = R(D(\hat{\mathbf{z}}_i))$, where $R(\cdot)$ is the reconstruction network to reproduce the decompressed image, and $D(\cdot)$ is the deconvolutional process to up-sample the latent image representations to the size of original images. Together, the decoder can be presented as $R(D(\cdot))$. The pipeline of the proposed TDNet can be formulated as:

$$\hat{\mathbf{x}}_i = R(D(T^{-1}(Q^{-1}(Q(T(E(\mathbf{x}_i, \Omega))))), \Theta), \Pi), \quad (2)$$

where Ω is the parameter of encoder $E(\cdot)$, and Θ and Π are the parameters of decoder $D(\cdot)$ and $R(\cdot)$, respectively. Being optimized in an end-to-end manner, the network is expected to learn the parameters $\{\Omega, \Theta, \Pi\}$ to minimize the distortion of the reconstructed image.

C. Architecture of TDNet

Loss Function: The loss function of a LIC network defines how close or how similar the decompressed image is to the original image. Many existing deep CNN based LIC methods [16], [18] use the perceptual loss such as the MS-SSIM loss [19], [23] to strength the perceptual quality of the compressed image. The MS-SSIM loss can also be adopted into our TDNet to learn the LIC network. Refer to Figure 2, we require both the deconvolution output and the final output of the network

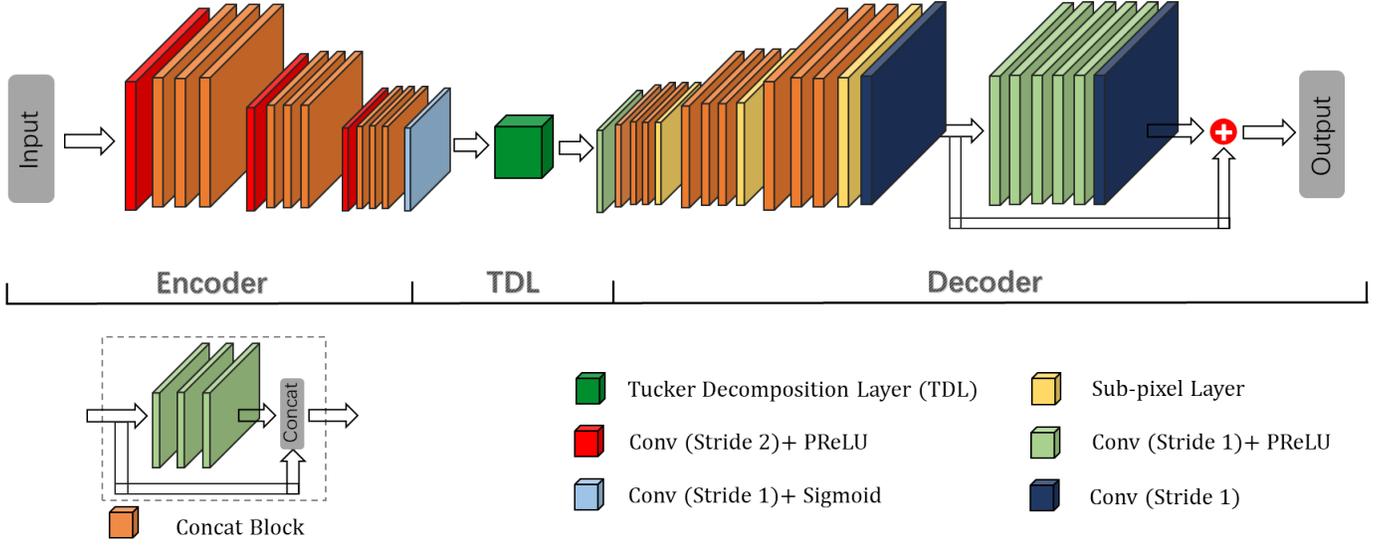


Fig. 2: Illustration of our proposed TDNet architecture.

to be similar to the original image, resulting in the following loss function:

$$L_{MS-SSIM} = 1 - MS-SSIM(D(T^{-1}(Q^{-1}(Q(T(E(\mathbf{x}_i, \Omega))))), \Theta)) - \lambda MS-SSIM(R(D(T^{-1}(Q^{-1}(Q(T(E(\mathbf{x}_i, \Omega))))), \Theta), \Pi)), \quad (3)$$

where \mathbf{x}_i refers to the i -th image in $\{\mathbf{x}_i\}_{i=1}^N$, and λ is a parameter to balance the loss between the intermediate deconvolution output and the final reconstruction.

Considering that most of the classical LIC methods such as JPEG and JPEG2000 take MSE as the objective to optimize, it is also important to validate whether a deep LIC network can achieve good MSE or equivalently PSNR measures. By minimizing the MSE of both the deconvolution output and the final output of the network, the MSE oriented loss function of the proposed network can be formulated as:

$$L_{MSE} = \frac{1}{N} \sum_i \|\mathbf{x}_i - D(T^{-1}(Q^{-1}(Q(T(E(\mathbf{x}_i, \Omega))))), \Theta)\|_2 + \frac{\lambda}{N} \sum_i \|\mathbf{x}_i - R(D(T^{-1}(Q^{-1}(Q(T(E(\mathbf{x}_i, \Omega))))), \Theta), \Pi)\|_2. \quad (4)$$

Encoder: Our encoder network consists of 3 types of layers, which are shown with 3 different colors in Figure 2. Instead of using Rectified Linear Unit (ReLU), we adopt Parametric Rectified Linear Units (PReLU) [24] as the activation function since it could improve the model fitting capability with little extra computational cost. Several convolution layers with a stride of 2 are utilized to downsample the feature maps, and the sigmoid function is used to project the data into the range of $[0, 1]$. Besides, concat operations are adopted to concatenate the feature maps of two layers to ensure maximum information flow. With the concat operations, we can tackle with the gradient vanishing problem in CNN training and ensure the TDNet to converge into the closest local minimum. Equipped with the concat connection, the network could also

TABLE I: Encoder network architecture.

Layer	Activation size
Input	$320 \times 320 \times 3$
Conv + PReLU ($3 \times 3 \times 64$, stride 2, pad 1)	$160 \times 160 \times 64$
Concat blocks $\times 3$ ($3 \times 3 \times 64$, stride 1, pad 1)	$160 \times 160 \times 256$
Conv + PReLU ($3 \times 3 \times 128$, stride 2, pad 1)	$80 \times 80 \times 128$
Concat blocks $\times 3$ ($3 \times 3 \times 64$, stride 1, pad 1)	$80 \times 80 \times 320$
Conv + PReLU ($3 \times 3 \times 256$, stride 2, pad 1)	$40 \times 40 \times 256$
Concat blocks $\times 3$ ($3 \times 3 \times 128$, stride 1, pad 1)	$40 \times 40 \times 640$
Conv + Sigmoid ($3 \times 3 \times 32$, stride 1, pad 1)	$40 \times 40 \times 32$

make full use of the hierarchical features from the previous layers. By stacking several convolutional layers, PReLU and Concat layers, the encoder network can transform the images into a compact domain with reduced redundancy. The detailed settings of the encoder network are summarized in Table I.

Tucker Decomposition Layer (TDL): Our TDL consists of two operations: $T(\cdot)$, which decomposes and quantizes the features into 3 orthogonal matrices $\{\mathbf{U}^{(n)}\}_{n=1}^3$ and a core tensor \mathbf{Y} , and $T^{-1}(\cdot)$, which de-quantizes and projects the core tensor and 3 orthogonal matrices back into the features. By setting the ranks $\{R_1, R_2, R_3\}$ of the three matrices and setting the quantization level M of the core tensor, the compression ratio of the network can be calculated as:

$$\alpha = \frac{C(\mathbf{x}_i)}{C(\mathbf{Y}) + C(\mathbf{U}^{(1)}) + C(\mathbf{U}^{(2)}) + C(\mathbf{U}^{(3)})}. \quad (5)$$

TABLE II: Decoder network architecture.

Layer	Activation size
Input	$40 \times 40 \times 32$
Conv + PReLU ($3 \times 3 \times 256$, stride 1, pad 1)	$40 \times 40 \times 256$
Concat blocks $\times 3$ ($3 \times 3 \times 128$, stride 1, pad 1)	$40 \times 40 \times 640$
Sub-pixel (upsampling factor: $\times 2$)	$80 \times 80 \times 160$
Concat blocks $\times 3$ ($3 \times 3 \times 64$, stride 1, pad 1)	$80 \times 80 \times 352$
Sub-pixel (upsampling factor: $\times 2$)	$160 \times 160 \times 88$
Concat blocks $\times 3$ ($3 \times 3 \times 64$, stride 1, pad 1)	$160 \times 160 \times 280$
Sub-pixel (upsampling factor: $\times 2$)	$320 \times 320 \times 70$
Conv ($3 \times 3 \times 3$, stride 1, pad 1)	$320 \times 320 \times 3$
Conv + PReLU $\times 5$ ($3 \times 64 \times 3$, stride 1, pad 1)	$320 \times 320 \times 64$
Conv ($3 \times 3 \times 3$, stride 1, pad 1)	$320 \times 320 \times 3$
Residual Sum	$320 \times 320 \times 3$

To change the compression rate, we can adjust the ranks of the decomposition matrices and the quantization levels $\{R_1, R_2, R_3, M\}$ of the core tensor, instead of retraining the network, and consequently achieve the goal of multiple bpp rates with a single network. More detail of the proposed TDL can be found in Section IV.

To ensure the end-to-end training of the network, the gradient of each component should be calculated for back-propagation. Though the tucker decomposition operation is non-differentiable and we cannot differentiate it with respect to its argument, based on the straight through estimator on gradient in [13], fortunately, we could set the derivative of tucker decomposition layer as:

$$\begin{cases} \frac{d}{dz}T(z) = 1; \\ \frac{\partial}{\partial \mathbf{Y}}T^{-1}(\mathbf{Y}, \mathbf{U}^{(n)}) = 1; \\ \frac{\partial}{\partial \mathbf{U}^{(n)}}T^{-1}(\mathbf{Y}, \mathbf{U}^{(n)}) = 1. \end{cases} \quad (6)$$

By setting the derivative to 1, the network can back propagate the loss from a decoder to an encoder. Thus, the whole network can be trained in an end-to-end manner.

Decoder: Our decoder network consists of a deconvolutional sub-network and a reconstruction sub-network, as shown in Figure 2. The deconvolutional sub-network basically mirrors the architecture of the encoder, and the stride of all convolutional layers is set to 1 since there is no need to downsample the feature maps. To ensure that the output image will have the same size as the input one, the sub-pixel layer [25] is

adopted to reshape and upsample feature maps. Usually, the deconvolution sub-network can deliver a rough approximation of the original image. The reconstruction sub-network aims to further enhance the deconvolution output by reproducing the missing details and textures in the encoding and TDL quantization process. Inspired by [8], [26], we propose to use the residual learning framework for the design of reconstruction sub-network. It consists of five convolutional layers, PReLUs and an element-wise addition operation. With the reconstruction sub-network, the final output image quality can be much refined. The detailed settings of the decoder network are summarized in Table II.

IV. TUCKER DECOMPOSITION LAYER

In this section, we first introduce some necessary notations and preliminaries of tensor decomposition, and then present in detail the proposed TDL.

A. Notations and Preliminaries

Denote by $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ an N -order tensor, and denote by $a_{i_1 \dots i_n \dots i_N}$ its elements, where $1 \leq i_n \leq I_n$. Let $\mathbf{B} \in \mathbb{R}^{J_n \times I_n}$ denote a matrix. The *mode- n product* of a tensor \mathbf{A} and a matrix \mathbf{B} can be defined as [27]:

$$\mathbf{C} = \mathbf{A} \times_n \mathbf{B}, \quad (7)$$

where the symbol \times_n denotes the tensor-times-matrix operation, and the *mode- n product* output $\mathbf{C} \in \mathbb{R}^{I_1 \times \dots \times I_{n-1} \times J_n \times I_{n+1} \times \dots \times I_N}$ is a tensor of order N . The elementwise representation of Eq. (7) can be written as:

$$\begin{aligned} \mathbf{C}(i_1, \dots, i_{n-1}, j_n, i_{n+1}, \dots, i_N) = \\ \sum_{k=1}^{I_n} \mathbf{A}(i_1, \dots, i_{n-1}, k, i_{n+1}, \dots, i_N) \mathbf{B}(j_n, k). \end{aligned} \quad (8)$$

The *mode- n product* can also be calculated by matrix multiplication:

$$\mathbf{C}_{(n)} = \mathbf{B} \mathbf{A}_{(n)}, \quad (9)$$

where $\mathbf{A}_{(n)} = \text{unfold}_n(\mathbf{A}) \in \mathbb{R}^{I_n \times (I_1 \dots I_{n-1} I_{n+1} \dots I_N)}$ and $\mathbf{C}_{(n)} = \text{unfold}_n(\mathbf{C}) \in \mathbb{R}^{J_n \times (I_1 \dots I_{n-1} I_{n+1} \dots I_N)}$, $1 \leq n \leq N$, are called *mode- n matrices*. Note that the operator $\text{unfold}_n(\cdot)$ is the process of reordering the elements of an n -way data array into a matrix. Conversely, the unfolding matrices along the n^{th} mode can be transformed back to the tensor by the $\text{unfold}_n(\cdot)$ operation.

For convenience, we define $\mathbf{A} \bar{\times}_n \{\mathbf{B}^{(j)}\}_{j=1}^N$ as [27]:

$$\begin{aligned} \mathbf{A} \bar{\times}_n \{\mathbf{B}^{(j)}\}_{j=1}^N \\ = \mathbf{A} \times_1 \mathbf{B}^{(1)} \times_2 \dots \times_{n-1} \mathbf{B}^{(n-1)} \times_{n+1} \mathbf{B}^{(n+1)} \dots \times_N \mathbf{B}^{(N)} \\ = \mathbf{A}_{(n)} (\mathbf{B}^{(N)} \otimes \dots \otimes \mathbf{B}^{(n+1)} \otimes \mathbf{B}^{(n-1)} \otimes \dots \otimes \mathbf{B}^{(1)}), \end{aligned} \quad (10)$$

where \otimes denotes the *Kronecker product*. The SVD of $\mathbf{A}_{(n)}$ is defined as:

$$\mathbf{A}_{(n)} = \mathbf{\Psi}^{(n)} \mathbf{\Sigma}^{(n)} \mathbf{V}^{(n)T}, \quad (11)$$

and the leading R_n -dimensional left singular subspace of $\mathbf{A}_{(n)}$ is defined as $\mathbf{\Psi}_{r_n}^{(n)} = \mathbf{\Psi}^{(n)}(:, 1 : R_n)$.

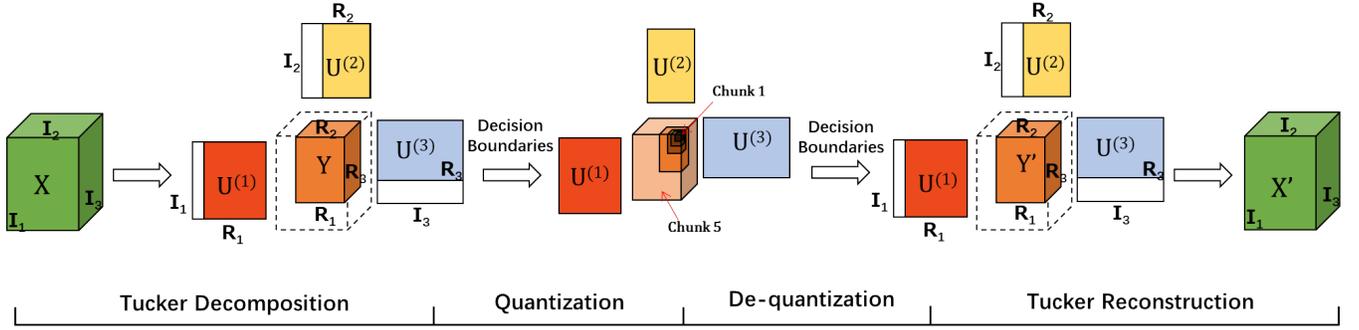


Fig. 3: Flowchart of the proposed Tucker decomposition layer.

B. Tucker Decomposition Layer

Deep CNNs have shown great potentials in LIC; however, it is still inconvenient to train a single network for LIC at multiple bpp rates with good performance. It is well-known that low-rank approximation is an effective way to reduce the redundancy of data. As a powerful low rank approximation approach, tensor decomposition, *e.g.*, Tucker decomposition [28] and CP decomposition [29], has been successfully used in various tasks, such as multispectral image restoration [30], [31], 3D image reconstruction [32], and higher-order web link analysis [33]. Inspired by the success of tensor decomposition methods, we introduce a novel TDL into the network architecture to achieve the goal that a single LIC network can perform image compression with multiple bpp rates. The flowchart of the proposed TDL is illustrated in Figure 3. It consists of 4 major components: 1) Tucker decomposition; 2) quantization; 3) de-quantization; and 4) Tucker reconstruction. The details of each component are described as follows.

Tucker decomposition: Tucker decomposition aims to decompose an N -order tensor $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ as an affiliation of N orthogonal bases $\{\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times R_n}\}_{n=1}^N$ and the associated core tensor $\mathbf{Y} \in \mathbb{R}^{R_1 \times R_2 \times \dots \times R_N}$, where $R_n \leq I_n$. It can be formulated as:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} \times_1 \mathbf{U}^{(1)T} \times_2 \mathbf{U}^{(2)T} \times_3 \dots \times_N \mathbf{U}^{(N)T} \\ \Leftrightarrow \mathbf{X} &\approx \hat{\mathbf{X}} = \mathbf{Y} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \dots \times_N \mathbf{U}^{(N)}. \end{aligned} \quad (12)$$

To find the optimal orthogonal matrices $\{\mathbf{U}^{(n)}\}_{n=1}^N$ and the core tensor \mathbf{Y} , we could minimize the error between the original data tensor \mathbf{X} and its approximation $\hat{\mathbf{X}}$, leading to the following optimization problem:

$$\underset{\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \dots, \mathbf{U}^{(N)}, \mathbf{Y}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{Y} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \dots \times_N \mathbf{U}^{(N)}\|_F^2. \quad (13)$$

Since $\|\mathbf{X}\|_F^2$ is a constant, according to [34], [27], [35], [36], Eq. (13) can be recast as an optimization problem to maximize $\|\mathbf{Y}\|_F^2$. We have:

$$\underset{\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \dots, \mathbf{U}^{(N)}}{\operatorname{argmax}} \|\mathbf{X} \times_1 \mathbf{U}^{(1)T} \times_2 \mathbf{U}^{(2)T} \times_3 \dots \times_N \mathbf{U}^{(N)T}\|_F^2. \quad (14)$$

To solve Eq. (14), we first employ the higher order singular value decomposition (HOSVD) [34] to initialize a set of basis factor matrices $\{\mathbf{U}_0^{(n)}\}_{n=1}^N$, then utilize the higher

order orthogonal iteration (HOOI) [37] to iteratively update the orthogonal matrices $\{\{\mathbf{U}_k^{(n)}\}_{n=1}^N\}_{s=1}^S$ until convergence, where s is the index of loop. With the obtained set of optimal orthogonal matrices $\{\mathbf{U}^{(n)}\}_{n=1}^N$, we can easily obtain the corresponding core tensor \mathbf{Y} by Eq. (12).

Specifically, for our TDNet the order of the feature tensor is $N = 3$. Given a 3-order tensor $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ and the desired rank of output $\{R_1, R_2, R_3\}$, we first compute the leading R_n -dimensional left singular subspace of $\mathbf{X}_{(n)}$ to initialize the basis factor matrices $\mathbf{U}_0^{(n)} \in \mathbb{R}^{I_n \times R_n}$, where $n = \{1, 2, 3\}$. Then, we rewrite Eq. (14) as follows to solve the n -th component matrix $\mathbf{U}^{(n)}$:

$$\begin{aligned} \underset{\mathbf{U}^{(n)}}{\operatorname{argmax}} & \|\mathbf{U}^{(n)T} \mathbf{X} \bar{\times}_{-n} \{\mathbf{U}^{(n)}\}_{n=1}^3\|_F^2 \\ \text{s.t.} & \mathbf{U}^{(n)T} \mathbf{U}^{(n)} = \mathbf{I}. \end{aligned} \quad (15)$$

The optimal solution $\mathbf{U}^{(n)}$ of Eq. (15) can be set as the leading R_n -dimensional left singular vectors of the matrix $\mathbf{X} \bar{\times}_{-n} \{\mathbf{U}^{(n)}\}_{n=1}^3$. By iteratively updating $\{\{\mathbf{U}_s^{(n)}\}_{n=1}^3\}_{s=1}^S$, we can obtain a set of final orthogonal matrices $\{\mathbf{U}^{(n)}\}_{n=1}^3$. With these final basis factors, the corresponding core tensor $\mathbf{Y} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ can be easily solved by Eq. (12).

Quantization and de-quantization: Since the core tensor \mathbf{Y} has both positive and negative values, we take one bit to represent the sign of the original value. Let $|\mathbf{Y}|$ denotes the absolute value of the core tensor. To obtain the optimal quantizer, we first train the TDNet (without quantization) in an end-to-end manner. After the network converges, we will have the core tensors for all the training images. With a set of core tensors \mathbf{Y} , we can easily compute $p(|\mathbf{Y}|)$, the probability density function (PDF) of the positive core tensor $|\mathbf{Y}|$. With the $p(|\mathbf{Y}|)$, the optimal quantizer can be solved as follows by minimizing the quantization error:

$$Q^*(|\mathbf{Y}|) = \underset{Q}{\operatorname{argmin}} \int p(|\mathbf{Y}|) (Q(|\mathbf{Y}|) - |\mathbf{Y}|)^2 d|\mathbf{Y}|. \quad (16)$$

Given a number M of decision intervals, the optimal quantizer is expected to find the set of decision boundaries $\{b_q\}_{q=0}^M$ and quantized values $\{\hat{Y}_q\}_{q=1}^M$. Solving the partial derivative of Eq.(16), we could have:

$$\hat{Y}_q = \frac{\int_{b_{q-1}}^{b_q} |\mathbf{Y}| p(|\mathbf{Y}|) d|\mathbf{Y}|}{\int_{b_{q-1}}^{b_q} p(|\mathbf{Y}|) d|\mathbf{Y}|}; \quad b_q = \frac{1}{2}(\hat{Y}_q + \hat{Y}_{q+1}). \quad (17)$$

Algorithm 1: Tucker Decomposition Layer

Input: A 3-order tensor $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$,
the desired output rank- (R_1, R_2, R_3) ,
a number of decision intervals M .

Output: An approximation of original data $\hat{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$,
a 3-order quantized core tensor $\hat{\mathbf{Y}} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$,
the orthogonal matrices $\{\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times R_n}\}_{n=1}^3$.

- 1 : $\mathbf{U}_0^{(n)} \leftarrow R_n$ leading left singular vectors of $\mathbf{X}^{(n)}$, for $n = 1, 2, 3$;
- 2 : **For** $s = 0, 1, 2, \dots$ (until converged), **do**:
- 3 : **For** $n = 1, 2, 3$, **do**:
- 4 : $\mathbf{U}_{s+1}^{(n)} \leftarrow \mathbf{U}_s^{(n)}$ by Eq. (15);
- 5 : **End for**;
- 6 : **End for**;
- 7 : Let $\{\mathbf{U}\} = \{\mathbf{U}_S\}$, where S is the index of the final result of step 2;
- 8 : Compute decision boundaries $\{b_q\}_0^M$ by Eq.(17);
- 9 : Divide $|\mathbf{Y}|$ into M non-overlapping chunks;
- 10 : $\hat{\mathbf{Y}} \leftarrow Q(|\mathbf{Y}|)$ by Eq. (18);
- 11 : $\hat{\mathbf{Y}} \leftarrow Q^{-1}(\hat{\mathbf{Y}})$ by Eq. (19);
- 12 : $\hat{\mathbf{X}} \leftarrow \hat{\mathbf{Y}}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}$ by Eq. (20);
- 13 : return $\hat{\mathbf{X}}, \{\mathbf{U}^{(n)}\}_{n=1}^3, \hat{\mathbf{Y}}$.

The optimal solutions of Eq.(17) can be easily solved by the Lloyd's algorithm [38], outputting the optimal quantizer $Q(\cdot)$ with decision boundaries $\{b_q\}_0^M$ and quantized values $\{\hat{y}_q\}_1^M$.

Considering that the core tensor has clear physical meanings (the larger singular values are generally associated with the major projection orientations), we adopt a variable-bits quantization scheme to allocate different quantized bits to the core tensor, which is useful in preserving the major edges and textures. More specifically, we first scan the positive core tensor $|\mathbf{Y}|$ in raster order, then utilize the decision boundaries $\{b_q\}_0^M$ to divide the core tensor into M non-overlapping chunks. Let C_m ($m \in [1, M]$) denote the index of the chunk. For the elements $|\mathbf{Y}_i|$ in chunk C_m , instead of using $\{\hat{y}_q\}_m$ as the quantized values, we define a new quantizer for it:

$$\bar{\mathbf{Y}}_i = Q(\mathbf{Y}_i) = \left\lfloor \frac{2^m}{MAX_{C_m} - MIN_{C_m}} \right\rfloor (|\mathbf{Y}_i| - MIN_{C_m}), \quad (18)$$

where MAX_{C_m} and MIN_{C_m} are the maximum and minimum values in chunk C_m , respectively, and m is the number of quantized bits in chunk C_m . In this way, chunk C_m would take $m + 1$ bits for the quantization. For example, if elements are in chunk C_1 , we quantize them into $M + 1$ bits, while if they are in chunk C_M , we quantize them into 2 bits.

Conversely, the de-quantization process can be readily formulated as:

$$\hat{\mathbf{Y}}_i = Q^{-1}(\bar{\mathbf{Y}}_i) = \left\lceil \frac{\bar{\mathbf{Y}}_i (MAX_{C_m} - MIN_{C_m})}{2^m} \right\rceil + MIN_{C_m}. \quad (19)$$

Considering that the precision of the orthogonal matrices $\{\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}\}$ is vital for the Tucker reconstruction and these 3 orthogonal matrices are relatively small compared with the image representation matrix, we simply use 6-bits to uniformly quantize them into discrete intervals, then adopt entropy coding to encode them.

Tucker reconstruction: With the de-quantized core tensor $\hat{\mathbf{Y}}$ and the orthogonal matrices $\{\mathbf{U}^{(n)}\}_{n=1}^3$, we can easily obtain an approximation of the original feature data $\hat{\mathbf{X}}$ by:

$$\hat{\mathbf{X}} \approx \hat{\mathbf{Y}} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}. \quad (20)$$

The overall TDL is summarized in Algorithm 1. With the proposed TDL, we can easily adjust the compression

Algorithm 2: All-in-One Training

Input: A set of training data $\{\mathbf{x}_i\}_{i=1}^N$,
 G groups of desired ranks $\{R_1, R_2, R_3\}_{g=1}^G$,
 G groups of decision intervals $\{M\}_{g=1}^G$.

Output: The optimal parameters $\{\Omega, \Theta, \Pi\}$.

- 1 : $\{\Omega, \Theta, \Pi\}_0 \leftarrow$ train encoder-decoder with Eq. (4) / (3);
- 2 : **For** $k = 1, 2, 3, \dots$ (until converged), **do**:
- 3 : $\{\{\mathbf{z}_i\}_{i=1}^N\}_k = E(\{\mathbf{x}_i\}_{i=1}^N, \Omega_{k-1})$;
- 4 : $\{\{\mathbf{Y}_i\}_{i=1}^N\}_k \leftarrow$ decompose $\{\{\mathbf{z}_i\}_{i=1}^N\}_k$ by Algorithm 1;
- 5 : **For** $g = 1, 2, 3, \dots, G$, **do**:
- 6 : $\{\{b_q\}_0^M\}_g \leftarrow$ compute the g -th group decision boundaries by Eq. (17);
- 7 : **End for**;
- 8 : Fixed the TDL;
- 9 : **For** $epoch = 1, 2, 3, \dots$ (until converged), **do**:
- 10 : $\hat{g} = epoch \bmod G$;
- 11 : $\{\{\Omega, \Theta, \Pi\}_{epoch+1}\}_k \leftarrow$ use $\{\{R_1, R_2, R_3, \{b_q\}_0^M\}_{(\hat{g}+1)}\}_k$ and Eq. (4) / (3) to update encoder-TDL-decoder;
- 12 : **End for**;
- 13 : Let $\{\Omega, \Theta, \Pi\}_k = \{\{\Omega, \Theta, \Pi\}_E\}_k$,
 where E is the index of the final result of step 9;
- 14 : **End for**;
- 15 : Let $\{\Omega, \Theta, \Pi\} = \{\Omega, \Theta, \Pi\}_K$,
 where K is the index of the final result of step 2;
- 16 : $\{\{\hat{\mathbf{z}}_i\}_{i=1}^N\}_{g=1}^G = T^{-1}(Q^{-1}(Q(T(E(\{\mathbf{x}_i\}_{i=1}^N, \Omega)))))$
 by Algorithm 1;
- 17 : $\{\Theta, \Pi\} \leftarrow$ update the decoder network with
 Eq. (4) / (3) and $\{\{\hat{\mathbf{z}}_i\}_{i=1}^N\}_{g=1}^G$;
- 18 : return $\{\Omega, \Theta, \Pi\}$.

ratio while keeping the size of latent image representation unchanged. Finally, we are able to train a single network to achieve LIC with multiple bpp rates.

The computational complexity of HOSVD [34] for an N -order tensor $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ is $\mathcal{O}(NI^{N+1})$, where $I_1 = I_2 = \dots = I_N = I$. As for the HOOI [37], it employs the HOSVD to initialize the tensors and converges after several iterations of alternating least squares (ALS). In this paper, the order number N is set to 3, I_i is set to $\{40, 40, 32\}$.

V. ALL-IN-ONE TRAINING

Instead of training a specific network for a certain compression ratio as in previous deep LIC methods [11], [12], [13], [14], [15], [16], [18], the proposed TDNet allows an all-in-one training strategy to enable a single network to compress an image at multiple compression ratio. More specifically, we first train an encoder-decoder network without the TDL to learn some initial parameters $\{\Omega, \Theta, \Pi\}_0$. We can then calculate a set of latent image representations of the input images $\{\mathbf{x}_i\}_{i=1}^N$ by $\{\mathbf{z}_i\}_{i=1}^N = E(\{\mathbf{x}_i\}_{i=1}^N, \Omega_0)$. Given G groups of desired output ranks and quantization levels $\{R_1, R_2, R_3, M\}_{g=1}^G$ and the obtained latent image representations $\{\mathbf{z}_i\}_{i=1}^N$, we can use the proposed TDL to calculate G groups of decision boundaries $\{\{b_q\}_0^M\}_{g=1}^G$. The TDL can then be initialized after obtaining the decision boundaries $\{b_q\}_0^M$.

With the initialized TDL, we can use Eq.(4) or Eq. (3) to jointly fine-tune the encoder-TDL-decoder network by minimizing the loss function. The latent image representations at multiple bpp rates will be taken into consideration during the training process. In each training epoch, we first decide which group of ranks and decision boundaries will be used by calculating $\hat{g} = \text{mod}(\text{epoch}, G)$, then take this group of desired output ranks and decision boundaries

$\{R_1, R_2, R_3, \{b_q\}_0^M\}_{\hat{g}+1}$ to update the TDL and fine-tune the parameters $\{\Omega, \Theta, \Pi\}$ of the whole network. To obtain G groups of optimal decision boundaries $\{\{b_q\}_0^M\}_{g=1}^G$ and network parameters $\{\Omega, \Theta, \Pi\}$, an iterative training scheme can be used, *i.e.*, fix the encoder-decoder network to update the TDL decision boundaries $\{\{b_q\}_0^M\}_{g=1}^G$ by solving Eq.(17), and fix the TDL to update the network parameters $\{\Omega, \Theta, \Pi\}$. Such an alternative optimization process continues till the loss function in Eq.(4) or Eq. (3) converges.

After the TDNet converges, we can use the optimal decision boundaries $\{\{b_q\}_0^M\}_{g=1}^G$ and network parameters $\{\Omega, \Theta, \Pi\}$ to compress and reconstruct images with different bpp rates. The all-in-one training scheme is summarized as Algorithm 2.

It should be noted that although our proposed ‘all-in-one’ joint training strategy can be used in the existing CNN-based LIC methods for ‘single-network-multiple-bpp’, the one with our proposed TDL can achieve better performance. Besides, since the quantization parameter M acts on all pixels, a small adjustment of M can change the bpp rates in a wide range for many existing methods, while our TDL can adjust the rank of core tensor to change the bpp rates with a finer step.

VI. EXPERIMENTAL RESULTS

In this section, we first present the experimental settings, including training and testing datasets, as well as parameter settings. We then discuss the performance of TDNet using a single network and multiple networks. Finally, we compare TDNet with state-of-the-art LIC methods.

A. Experimental Settings

Datasets: It is generally agreed that a larger scale training dataset which covers various image contents and structures will bring benefit to train a stable deep LIC network. Therefore, we mix the MS COCO test2017 dataset [39], the DIV2K dataset [40] and the Waterloo Exploration dataset [41] together as the training dataset. The MS COCO test2017 dataset contains 40,652 images which cover a great diversity of objects and scenes. The DIV2K dataset has 900 high-resolution images with complex structures and texture patterns. The Waterloo Exploration dataset contains 4,744 elaborately selected high quality natural images. We first crop these images into 320×320 patches⁴, then randomly flip them. With around 100,000 image patches, we could train a robust model for LIC with multiple bpp rates.

For the testing, we use two different test datasets for comprehensive evaluation: the Kodak PhotoCD dataset⁵, which contains 24 natural images, and the McMaster dataset [42], which contains 18 high quality images. Note that all those images are widely used for the evaluation of image processing methods and they are not included in the training dataset.

Parameter Settings: In the training phase, we set $G = 4$ groups of desired ranks and decision intervals

$\{R_1, R_2, R_3, M\}_{g=1}^4$ to train the TDNet: $\{\{38, 37, 28, 5\}, \{36, 35, 26, 4\}, \{34, 31, 23, 3\}, \{34, 30, 22, 3\}\}$. The mini-batch size is set to 7. We initialize the network weights by the method in [24] and adopt the Adam solver [43] to optimize the network parameters $\{\Omega, \Theta, \Pi\}$. The learning rate starts from $1e-4$ and is then fixed to $1e-5$ when the training error stops decreasing. The training is terminated when the training error does not decrease in 20 sequential epochs. For the other hyper-parameters of Adam, we utilize the default setting.

The parameter $\{R_1, R_2, R_3, M\}$ in this paper ranges from $\{40, 40, 32, 5\}$ to $\{32, 30, 22, 2\}$. We first utilized 100,000 image patches of size $320 \times 320 \times 3$ to train the network, then we obtained the latent image representations of size $40 \times 40 \times 32$ for each patch. With these 100,000 latent image representations, we used the HOOI algorithm to decompose them into a core tensor and its associated orthogonal matrices, from which the latent image representations can be reconstructed. We empirically found that when the rank of core tensor (R_1, R_2, R_3) is less than $(34, 30, 22)$, the PSNR between the original latent image representation and the reconstructed one will drop significantly. Based on this observation, the rank (R_1, R_2, R_3) is set to be greater than $(34, 30, 22)$. While for the quantization parameter M , we set it to $\{2, 3, 4, 5\}$. As mentioned above, the quantization parameter M is acting on all pixels, a small adjustment of M (*e.g.*, from 3 to 2) can change the bpp rates in a wide range. Thus, we adjust the rank of the core tensor (R_1, R_2, R_3) to change the bpp rates in a small range. Otherwise, we change the quantization parameter M . Though we use 4 groups of ranks and decision intervals $\{R_1, R_2, R_3, M\}$ to train the TDNet, to validate the generality of the trained network, in the testing phase we use 6 groups of ranks and decision intervals to test the performance of TDNet: $\{\{38, 37, 28, 5\}, \{36, 35, 26, 4\}, \{35, 32, 23, 4\}, \{34, 31, 23, 3\}, \{34, 30, 22, 3\}\}$ and $\{34, 30, 22, 2\}$. As we will see in the following sections, our TDNet achieves highly competitive performance.

We experimentally found that the alternative optimization process of our TDNet compressor (refer to Algorithm 2 please) will converge in less than $k = 4$ iterations. The parameter λ in our loss function Eq. (4) or Eq. (3) is set to 0.4 by experience. The network is trained in CAFFE [44] with an Nvidia Titan Xp GPU. In our PC with Intel(R) Core(TM) i9-7900X CPU @ 3.3GHz, 96G RAM, the training process costs about 3 days.

Entropy Coding Scheme: As for the entropy coding engine, we follow the previous CNN-based LIC methods to take the CABAC [45] as a separated module to losslessly encode the latent image representations. Specifically, we first use an exponential Golomb code to quantize the values \hat{Y}_i into binary decisions. We then integrate a non-optimized CABAC implementation provided in an open source software tool⁶, where most of the parameters are set as default. In this open source software tool, the context is formed by the previous n bits, where n is a tuneable parameter and we set $n = 4$, implying that we have 16 contexts in total. The updating rule of the probability within each context is standard: when the specific context is met, the corresponding

⁴We experimentally found that using the same network architecture, the larger the training patches are, the better the results would be. To trade off between GPU memory and compression results, we set the size of training patches as 320×320 .

⁵<http://r0k.us/graphics/kodak/>

⁶<https://github.com/jigar23/CABAC>

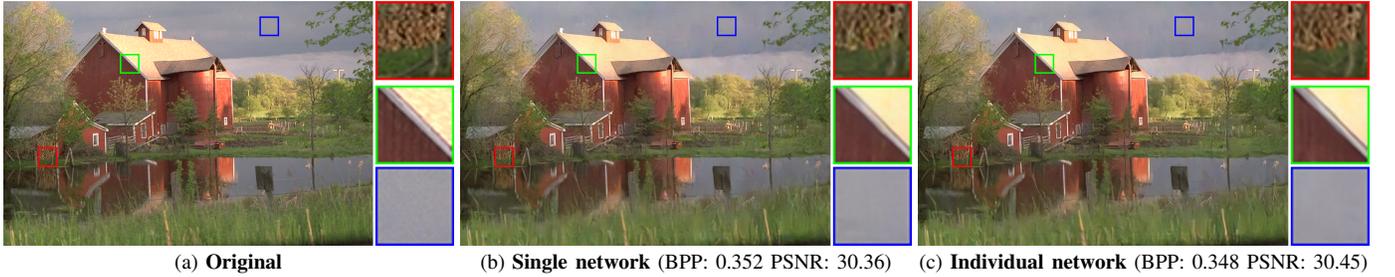


Fig. 4: Visual comparison on image “house” by the proposed single network and individual network.

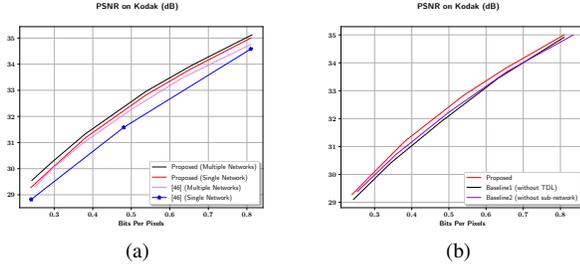


Fig. 5: Rate-distortion curves on the Kodak dataset.

probability mass function is updated according to current symbol. Furthermore, the data unit for CABAC coding is the binarized syntax elements of the whole image. In other words, we do not explicitly reset the context probabilities. Certainly, a tailored and optimized CABAC implementation (*e.g.*, through re-defining the context template, increasing the number of contexts, adaptively resetting the context probability, etc) could bring additional performance gain, at the cost of lower computational efficiency. However, since the goal of this paper is to design a CNN-based LIC method for image compression with multiple bpp rates, the in-depth discussion regarding this improvement strategy is beyond the scope of our current work.

B. Ablation Study

Single Network vs. Multiple Networks: The proposed TDNet allows an “all-in-one” training strategy to learn a single network to perform LIC at multiple bpp rates. To validate the effectiveness of our “single-network-multiple-bpp” scheme, in this section we also train six individual TDNet compressors for the six groups of $\{R_1, R_2, R_3, M\}_{g=1}^6$ (*i.e.*, six bpp rates), respectively, and compare the performance of single TDNet and multiple TDNet at different bpp rates.

Figure 4 compares the visual quality of compressed images *house* by those two training strategies at around 0.35bpp. We also show the zoom-in images of a smooth background area, a texture area and a large edge area. It can be seen that both the two training strategies produce good preservation of image edges and details, and they have very small visual difference.

Figure 5(a) shows the PSNR based rate-distortion curves by the single network and multiple networks on the Kodak dataset. Note that the six points on the curve of multiple networks is obtained by a TDNet trained at a specific bpp. As one can see, the rate-distortion curve of the trained single TDNet

is very close to the curve obtained by multiple networks. On average, its PSNR is only 0.181dB lower than that of multiple networks.

Effectiveness of TDL: To further evaluate the effectiveness of the proposed TDL, we compare the proposed TDNet with [46]. By changing the quantization levels M , we can utilize the proposed “single-network-multiple-bpp” training scheme to make [46] adaptive for multiple bpp rates. The rate-distortion curves on the Kodak dataset of our proposed TDNet and [46] are shown in Figure 5(a). One can see that the proposed TDL enables training a single network for multiple bpp rates. Besides, we also trained a baseline variant of our method (denoted by baseline 1), which uses the network described in Figure 2 but without the proposed TDL. The comparison results are shown in Figure 5(b).

Reconstruction Network: The reconstruction sub-network aims to further enhance the deconvolution output by reproducing the missing details and textures in the encoding and TDL quantization process. To evaluate the effectiveness of the reconstruction sub-network, we trained a baseline variant of our method (denoted by baseline 2), which uses the network shown in Figure 2 but without the proposed reconstruction sub-network. The comparison results are shown in Figure 5(b). One can see that the network with the reconstruction sub-network achieves better performance.

C. Results

We compare our proposed TDNet with both traditional LIC Codecs and CNN-based LIC methods.

Traditional LIC Codecs: The compared traditional LIC codecs include JPEG (implemented by libjpeg⁷), JPEG 2000 (implemented by Matlab) and the state-of-the-art compression format better portable graphics (BPG)⁸. Following [15], we use BPG with the setting of 4 : 4 : 4 chroma format.

Deep LIC Methods: The compared CNN-based LIC methods include Ballé *et al.* [12]⁹, Theis *et al.* [13]¹⁰, Li *et al.* [14]¹¹, Johnston *et al.* [17], Rippel & Bourdev [15], Mentzer *et al.* [16] and Minnen *et al.* [47]. Note that since the source codes of the above deep compressors [12], [13], [14], [17], [15], [16] are not available, we either digitize their rate-distortion

⁷<http://libjpeg.sourceforge.net/>

⁸<https://bellard.org/bpg/>

⁹<http://www.cns.nyu.edu/lcv/iclr2017/>

¹⁰http://theis.io/compressive_autoencoder/

¹¹<http://www2.comp.polyu.edu.hk/~15903062r/index.html>

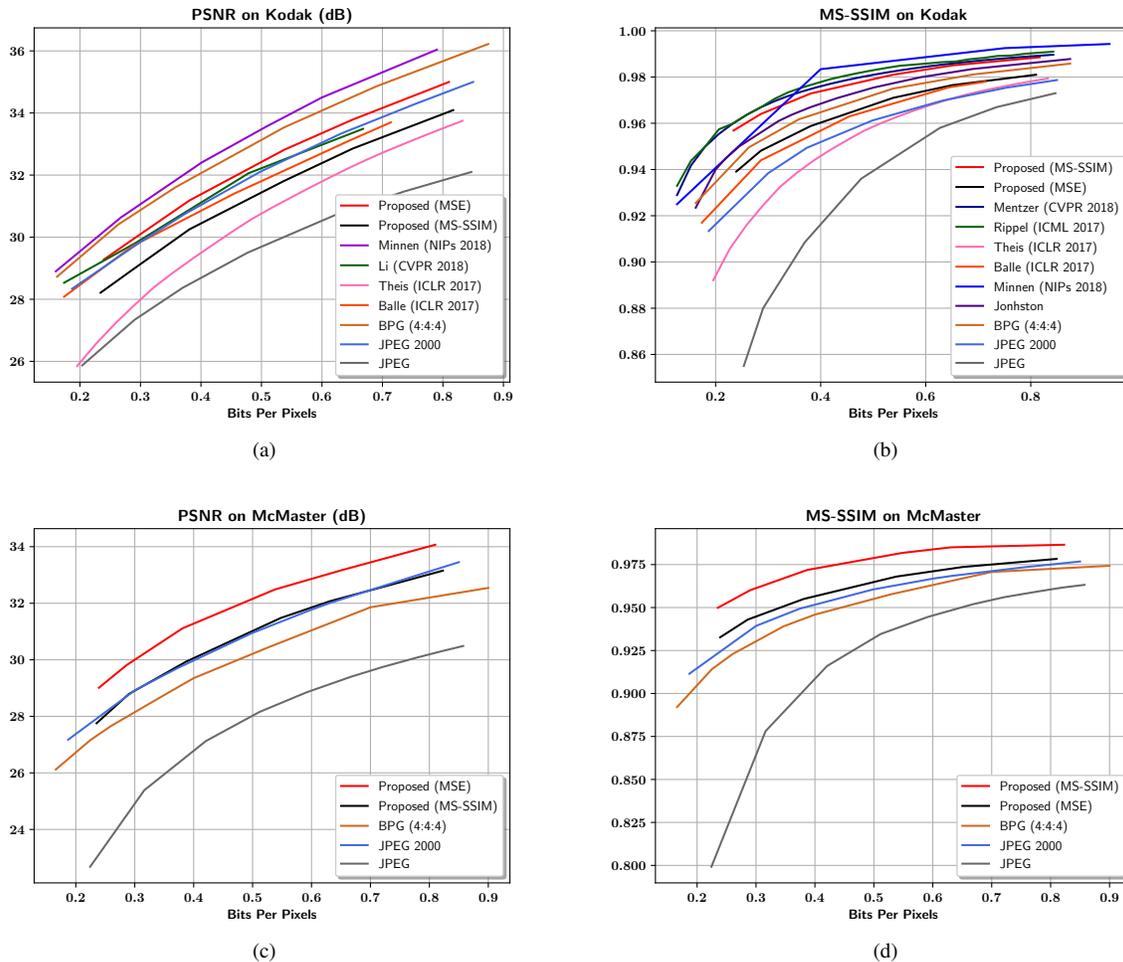


Fig. 6: Comparison of the rate-distortion curves on Kodak and McMaster datasets.

curves on the Kodak dataset from the original papers or copy the results from their websites.

Quantitative Evaluation: Most of the existing CNN-based LIC models [15], [16], [17] are optimized with the MS-SSIM loss [19], while traditional LIC methods (*i.e.*, JPEG, JPEG2000 and BPG) and some of the deep LIC methods [12], [13], [14], [47] are optimized in terms of PSNR. Therefore, we conduct the experiments to quantitatively evaluate the competing methods in terms of both PSNR and MS-SSIM indices for a more comprehensive comparison.

The PSNR and MS-SSIM based rate-distortion curves on the Kodak and McMaster datasets are summarized in Figure 6. As in previous works [10]-[18], the curves are interpolated based on a set of points [bpp, PSNR] and [bpp, MS-SSIM] for one method. Note that for some methods, only the points of [bpp, PSNR] or [bpp, MS-SSIM] are available on the Kodak dataset, and all existing deep LIC methods do not report their results on the McMaster dataset. Therefore, not all methods have all the four curves in Figure 6.

Figures 6(a) and 6(c) show the PSNR based rate-distortion curves on the Kodak and McMaster datasets, respectively. One can see that on the Kodak dataset, our TDNet (trained with MSE loss) achieves better result than JPEG2000 and the

recently deep LIC methods, including Ballé *et al.* [12], Theis *et al.* [13] and Li *et al.* [14], and significantly outperforms the prevalent compressor JPEG. Although the proposed TDNet does not show advantage over BPG in term of PSNR on the Kodak dataset, it achieves much better PSNR index than BPG on the McMaster dataset (see Figure 6(c)). Meanwhile, it is not a surprise that TDNet trained with MSE has much higher PSNR indices than TDNet trained with MS-SSIM.

Figures 6(b) and 6(d) show the MS-SSIM based rate-distortion curves on the Kodak and McMaster datasets, respectively. One can see that our TDNet largely outperforms the traditional codecs BPG, JPEG2000 and JPEG. It also significantly outperforms the methods of Johnston *et al.* [17], Ballé *et al.* [12] and Theis *et al.* [13], and achieves comparable performance to the state-of-the-art deep LIC methods Rippel & Bourdev [15] and Mentzer *et al.* [16].

As shown in Figures 6(a) and 6(b), the recent work in [47] reports better results than ours on the Kodak dataset; however, it employs a much deeper and wider network and is trained on a very large dataset, which consumes much higher computational cost than ours. Our tucker decomposition strategy can also be used in other CNN-based LIC methods. Since the focus of this paper is on the design of a single

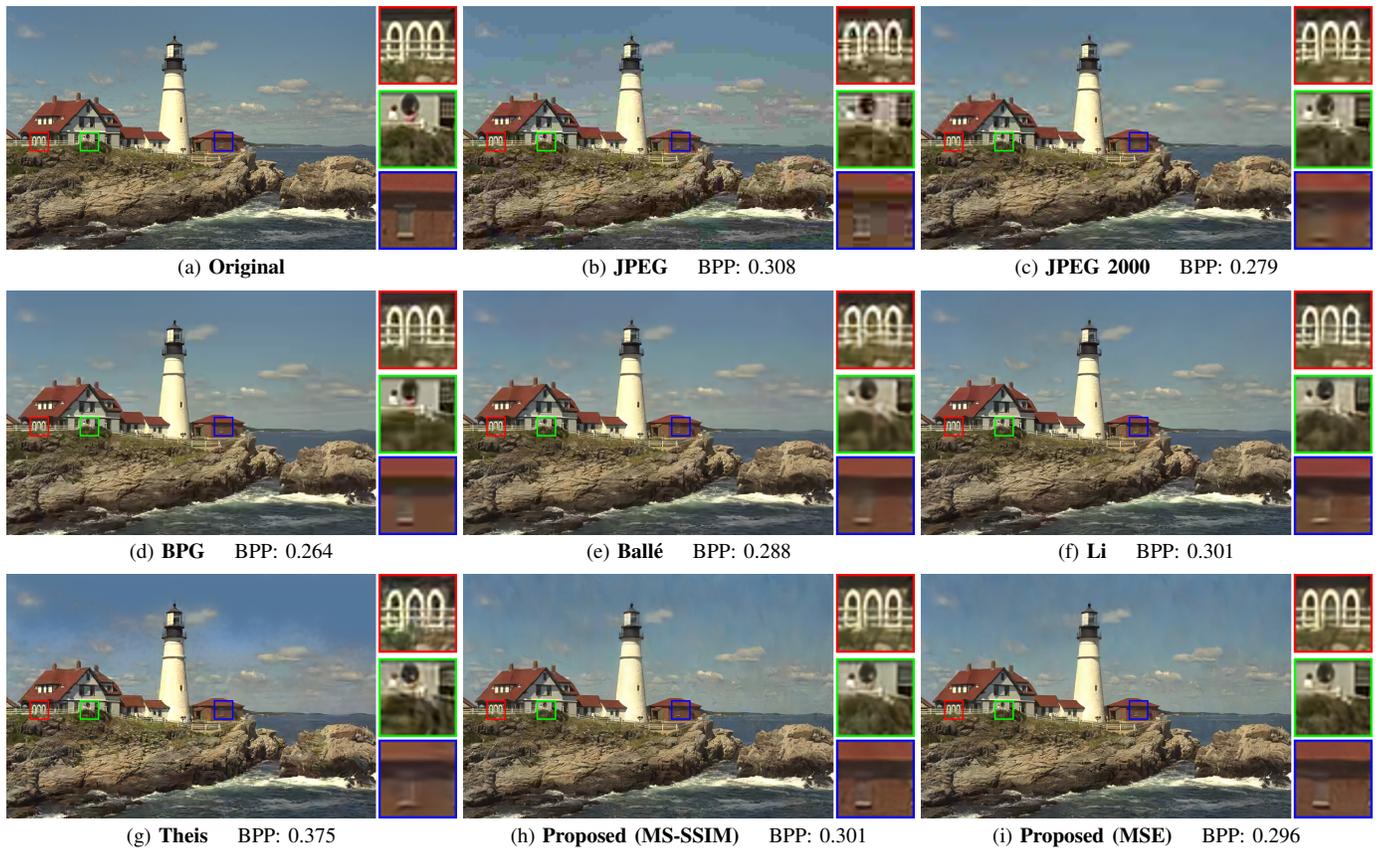


Fig. 7: Visual comparison on image “lighthouse” by different methods at a compression rate around 0.3bps (Image from Kodak dataset)

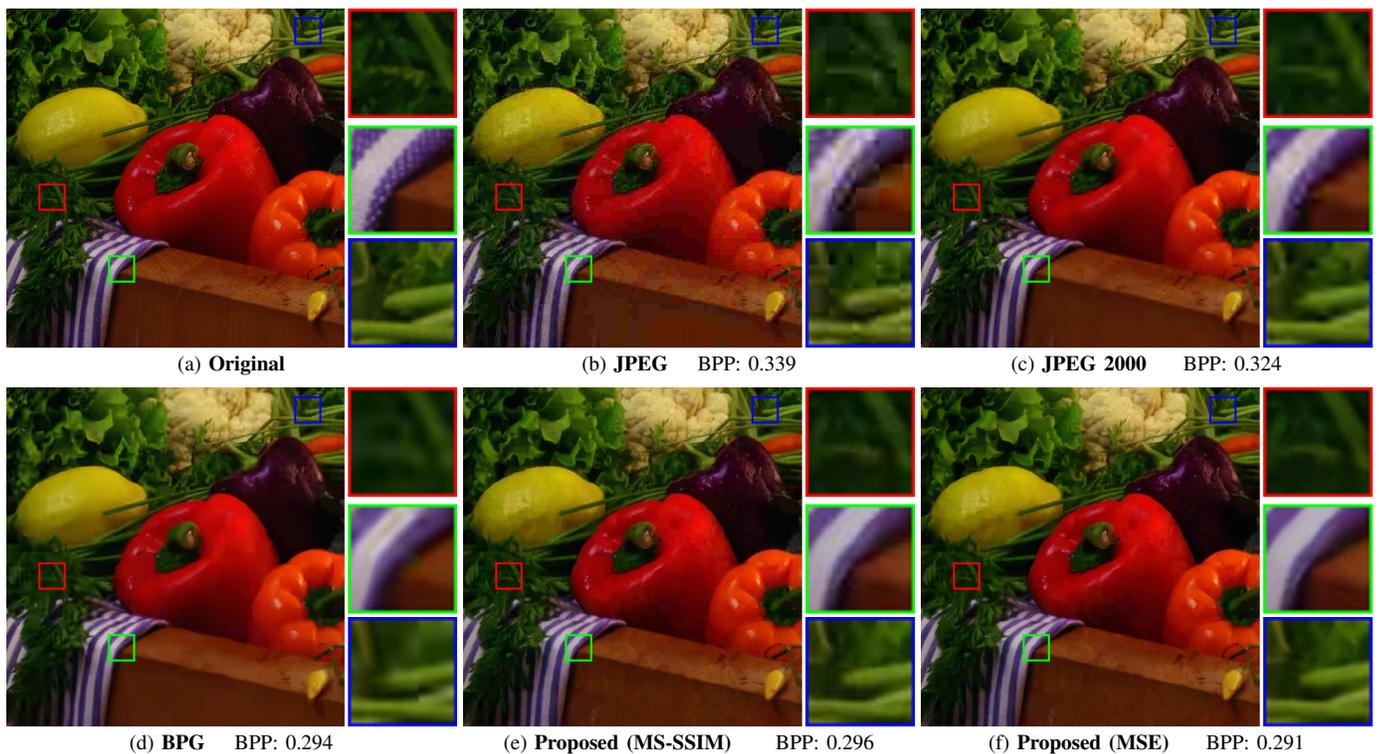


Fig. 8: Visual comparison on image “peppers” by different methods at a compression rate around 0.3bps (Image from McMaster dataset)

CNN which can perform LIC with multiple bpp rates, we prefer to use a lightweight network to verify the effectiveness of our proposed TDL. Note that our method achieves better performance than many previous CNN-based LIC methods, including those recent ones [12], [13], [14], [15], [46].

As shown in Figure 6, one can see that both our proposed TDNet and traditional LIC codecs (including JPEG, JPEG 2000 and BPG) achieve better quantitative results (in terms of both PSNR and MS-SSIM indices) on the Kodak dataset than the McMaster dataset. This is mainly because the images in the Kodak dataset are smoother, less saturated and have more redundant information than the images in the McMaster dataset [42], [48], [49], [50]. Specifically, as shown in [42], the Kodak images are smooth in chromatic gradient, have very high spectral correlation and low saturation (please refer to Table III).

TABLE III: Statistics of Kodak and McMaster datasets [42].

		Kodak	McMaster
Mean Spectral Correlation	G and R	0.8712	0.7445
	G and B	0.9050	0.7114
Mean Saturation		15.6	45.81
Mean Chromatic Gradient		1.78	4.54

Although the two datasets have different statistics, the proposed TDNet still achieves similar quantitative performance on them. In contrast, the quantitative results of those traditional LIC codecs (*i.e.*, JPEG2000 and BPG) on the McMaster dataset are much lower than the Kodak dataset. The reason may be that these traditional methods are less effective for sharp and saturated images. The proposed TDNet is trained on a large scale of images (including COCO testing dataset [39], the DIV2K dataset [40] and the Waterloo Exploration dataset [41]), which cover a variety of image structures and contents. Thus, it is more robust to compress images with different statistics.

Again, we would like to stress that all the competing CNN-based LIC methods here train a specific network for a certain bpp, while our proposed TDNet trains a single network to deal with multiple bpp rates.

Visual Quality Evaluation: We further compare the visual quality of images compressed by JPEG, JPEG 2000, BPG, Ballé *et al.* [12], Li *et al.* [14], Theis *et al.* [13] and our proposed TDNet (trained with MSE and MS-SSIM). Note that since the source codes of all existing deep LIC compressors are not available, we can only download the results of Ballé *et al.* [12], Li *et al.* [14] and Theis *et al.* [13] from their websites. The compressed images of other deep LIC methods are not available and thus cannot be compared.

Figure 7 shows the compressed images *lighthouse* by the comparison methods at a compression rate around 0.3bpp (note that Theis *et al.*, only provides the image *lighthouse* at 0.375bpp). One can see that noticeable blocky and ringing artifacts are inevitable in the reconstructed images by traditional JPEG and JPEG 2000 compression formats. While BPG, Theis *et al.* and Ballé *et al.* can produce much better visual quality, they still blur much the edges and over-smooth the textures

(see the zoom-in areas). Li *et al.*'s method can preserve better the sharp edges and detailed textures, but still generate some noticeable artifacts. Compared with these methods, the image compressed by our TDNet method is visually more pleasing with sharper edges and much less artifacts.

Figure 8 presents the visual comparison results on image *peppers* from the McMaster dataset at a compression rate around 0.3bpp. Note that since the results on this dataset are not available for all existing deep LIC methods, we only compare TDNet with JPEG, JPEG2000 and BPG. Again, one can see noticeable artifacts in the zoom-in areas for the traditional LIC methods. In contrast, the result produced by our proposed TDNet exhibits visually much more pleasing results.

VII. CONCLUSION AND FUTURE WORK

In this paper, we presented a simple yet effective Tucker Decomposition Network (TDNet) with a novel tucker decomposition layer (TDL), which can decompose a latent image representation into a set of matrices and one small core tensor for lossy image compression (LIC). By changing the rank of core tensor and its quantization levels, we could easily adjust the bits-per-pixel (bpp) rate of latent image representation, and consequently achieved the goal of using a single CNN model to cover a range of bpp rates. An iterative non-uniform quantization scheme was presented to optimize the quantizer, and an all-in-one training strategy was employed to train the TDNet. Compared with traditional LIC schemes and previous deep LIC compressors which use different networks to compress images at different bpp rates, our TDNet exhibits very competitive results on benchmark datasets by using a single network.

ACKNOWLEDGMENT

We gratefully acknowledge the support from NVIDIA Corporation for providing us the Titan X GPU used in this research. We would like to thank Dr. Jiantao Zhou for helpful advice on entropy coding. We also like to thanks the anonymous reviewers for constructive comments.

REFERENCES

- [1] X. Liu, G. Cheung, C. W. Lin, D. Zhao, and W. Gao, "Prior-based quantization bin matching for cloud storage of jpeg images," *IEEE Transactions on Image Processing*, 2018.
- [2] X. Liu, G. Cheung, X. Wu, and D. Zhao, "Random walk graph laplacian-based smoothness prior for soft decoding of jpeg images," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 509–524, 2017.
- [3] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3445–3462, 1993.
- [4] M. Rabbani and R. Joshi, "An overview of the jpeg 2000 still image compression standard," *Signal processing: Image communication*, vol. 17, no. 1, pp. 3–48, 2002.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

- [8] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.
- [9] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.
- [10] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthakar, "Variable rate image compression with recurrent neural networks," *arXiv preprint arXiv:1511.06085*, 2015.
- [11] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell, "Full resolution image compression with recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 5435–5443.
- [12] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *International Conference on Learning Representations*, 2017.
- [13] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," *International Conference on Learning Representations*, 2017.
- [14] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.
- [15] O. Rippel and L. Bourdev, "Real-time adaptive image compression," in *International Conference on Machine Learning*, 2017, pp. 2922–2930.
- [16] E. Agustsson, F. Mentzer, M. Tschannen, R. Timofte, and L. Van Gool, "Conditional probability models for deep image compression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.
- [17] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S. J. Hwang, J. Shor, and G. Toderici, "Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks," *arXiv preprint arXiv:1703.10114*, 2017.
- [18] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavignoli, R. Timofte, L. Benini, and L. V. Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Advances in Neural Information Processing Systems*, 2017, pp. 1141–1151.
- [19] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers*, 2004. *Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. Ieee, 2003, pp. 1398–1402.
- [20] G. K. Wallace, "The jpeg still picture compression standard," *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.
- [21] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The jpeg 2000 still image compression standard," *IEEE Signal processing magazine*, vol. 18, no. 5, pp. 36–58, 2001.
- [22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [23] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2017.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
- [25] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.
- [26] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [27] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.
- [28] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.
- [29] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Studies in Applied Mathematics*, vol. 6, no. 1-4, pp. 164–189, 1927.
- [30] Q. Xie, Q. Zhao, D. Meng, Z. Xu, S. Gu, W. Zuo, and L. Zhang, "Multispectral images denoising by intrinsic tensor sparsity regularization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1692–1700.
- [31] Q. Xie, Q. Zhao, D. Meng, and Z. Xu, "Kronecker-basis-representation based tensor sparsity and its applications to tensor recovery," *IEEE transactions on pattern analysis and machine intelligence*, 2017.
- [32] A. C. Sauve, A. Hero, W. L. Rogers, S. Wilderman, and N. Clinthorne, "3d image reconstruction for a compton spect camera model," *IEEE Transactions on Nuclear Science*, vol. 46, no. 6, pp. 2075–2084, 1999.
- [33] T. G. Kolda, B. W. Bader, and J. P. Kenny, "Higher-order web link analysis using multilinear algebra," in *Data Mining, Fifth IEEE International Conference on*. IEEE, 2005, pp. 8–pp.
- [34] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.
- [35] C. A. Andersson and R. Bro, "Improving the speed of multi-way algorithms: Part i. tucker3," *Chemometrics and intelligent laboratory systems*, vol. 42, no. 1-2, pp. 93–103, 1998.
- [36] T. G. Kolda, "Multilinear operators for higher-order decompositions." Sandia National Laboratories, Tech. Rep., 2006.
- [37] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-(r_1, r_2, \dots, r_n) approximation of higher-order tensors," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1324–1342, 2000.
- [38] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1979.
- [39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [40] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, vol. 3, 2017, p. 2.
- [41] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, 2017.
- [42] L. Zhang, X. Wu, A. Buades, and X. Li, "Color demosaicking by local directional interpolation and nonlocal adaptive thresholding," *Journal of Electronic imaging*, vol. 20, no. 2, p. 023016, 2011.
- [43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.
- [45] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the h. 264/avc video compression standard," *IEEE Transactions on circuits and systems for video technology*, vol. 13, no. 7, pp. 620–636, 2003.
- [46] J. Cai and L. Zhang, "Deep image compression with iterative non-uniform quantization," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 451–455.
- [47] D. Minnen, J. Ballé, and G. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in neural information processing systems*, 2019.
- [48] A. Buades, B. Coll, J. M. Morel, and C. Sbert, "Self-similarity driven demosaicking," *Image Processing on Line*, vol. 1, 2011.
- [49] Z. Fan, W. Xiaolin, Y. Xiaokang, Z. Wenjun, and Z. Lei, "Robust color demosaicking with adaptation to varying spectral correlations," *IEEE Transactions on Image Processing*, vol. 18, no. 12, pp. 2706–2717, 2009.
- [50] X. Li, B. Gunturk, and L. Zhang, "Image demosaicing: A systematic survey," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 6822, pp. 68 221J–68 221J–15, 2009.