Deep Image Compression with Iterative Non-uniform Quantization

Jianrui Cai and Lei Zhang

Dept. of Computing

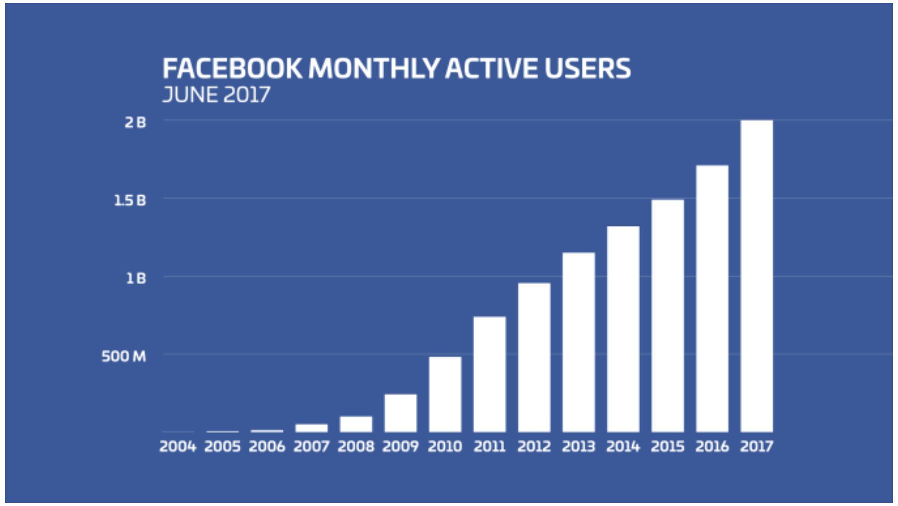
The Hong Kong Polytechnic University

Overview

- Introduction
- Proposed Method
- Experimental Results
- Conclusions

Introduction: motivation

Facebook



Instagram

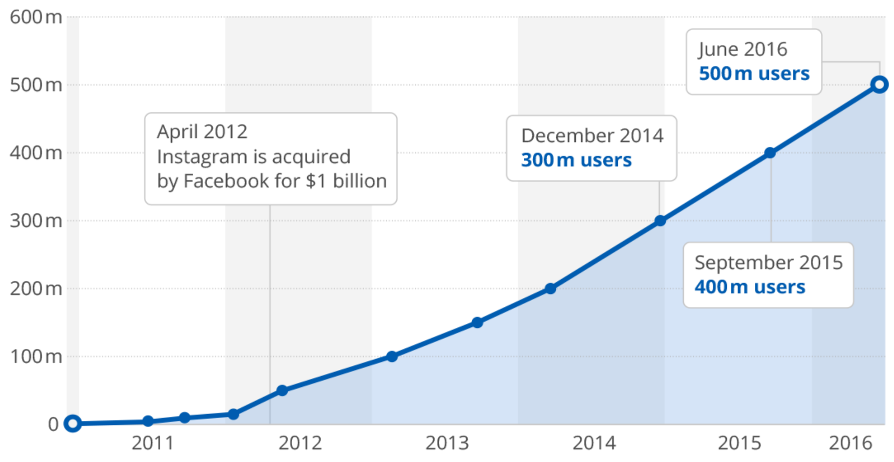


Image compression is a must to reduce the storage space and provide an economic solution to a wide range of image storage and transmission systems.

High → Low

Bit Per Pixel

Image Quality

Introduction : Pipeline

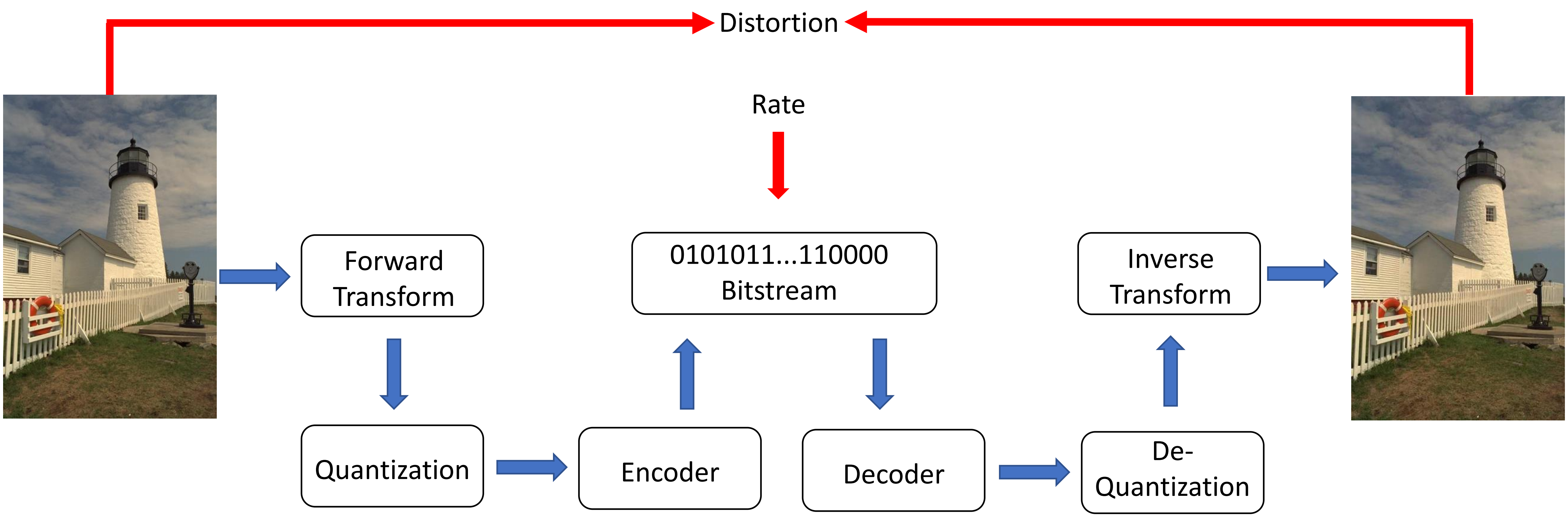
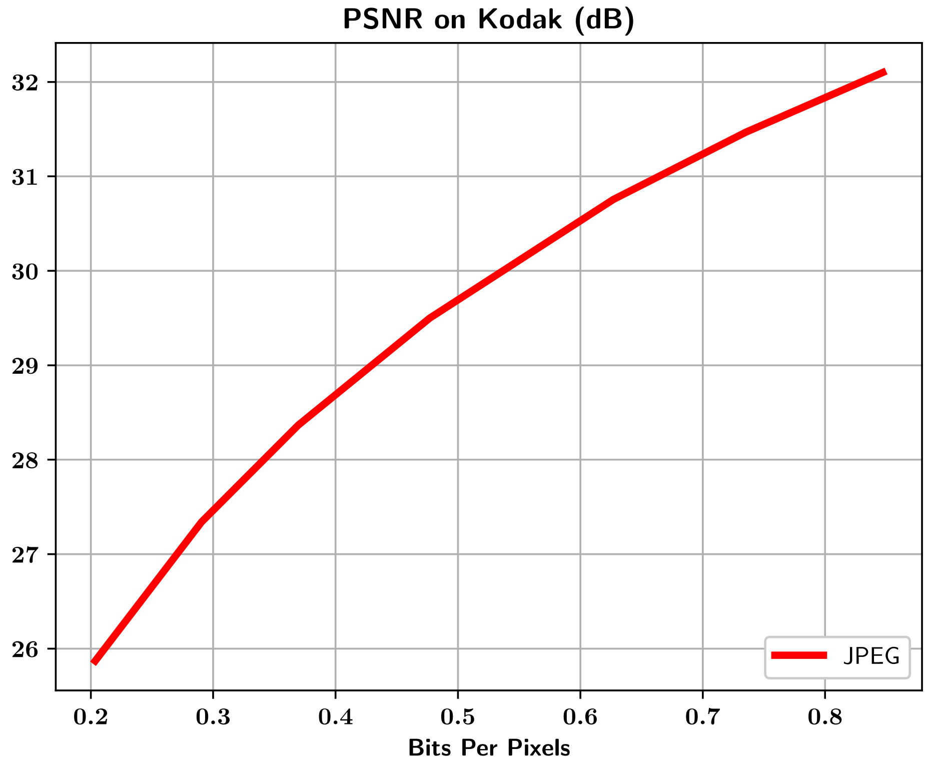


Image Compression Pipeline

Introduction : formulation

$$\text{Loss} = \arg\min \, [R(z) + \lambda \, D(X, Y)]$$

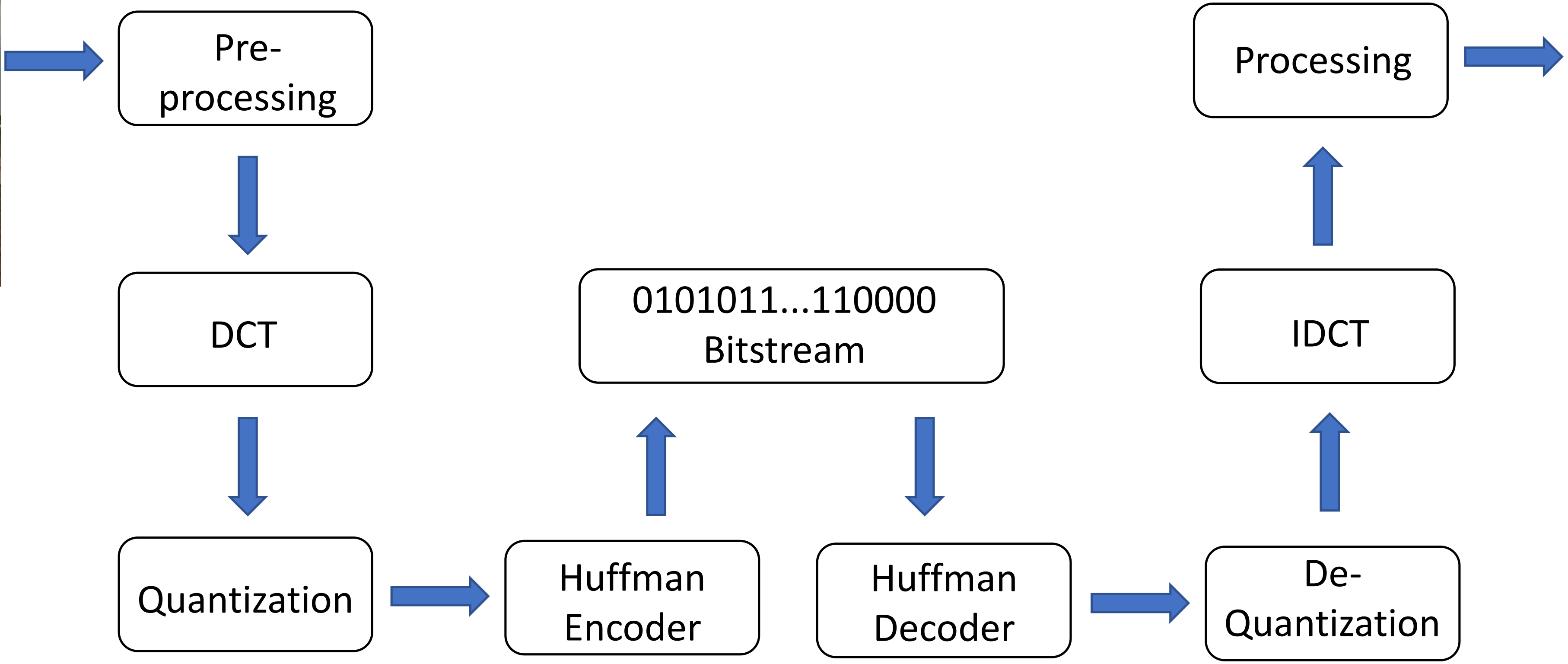
- X: the original image;
- Z: the latent image;
- Y: the reconstructed image;
- R(): the entropy of the latent image;
- D(): the distortion between two images.



Rate-Distortion curve

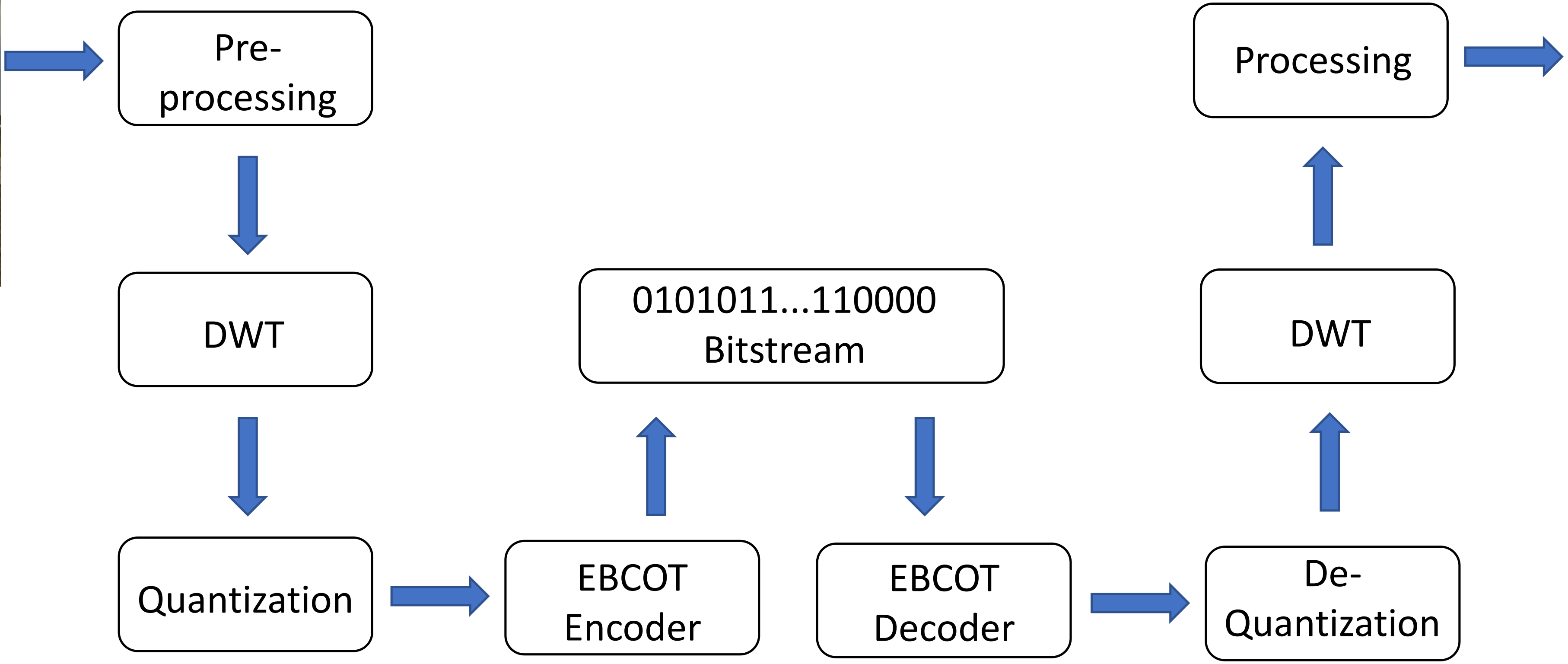
Introduction: JPEG

Pros & Cons: Simple but often with artifact (e.g., blocking)



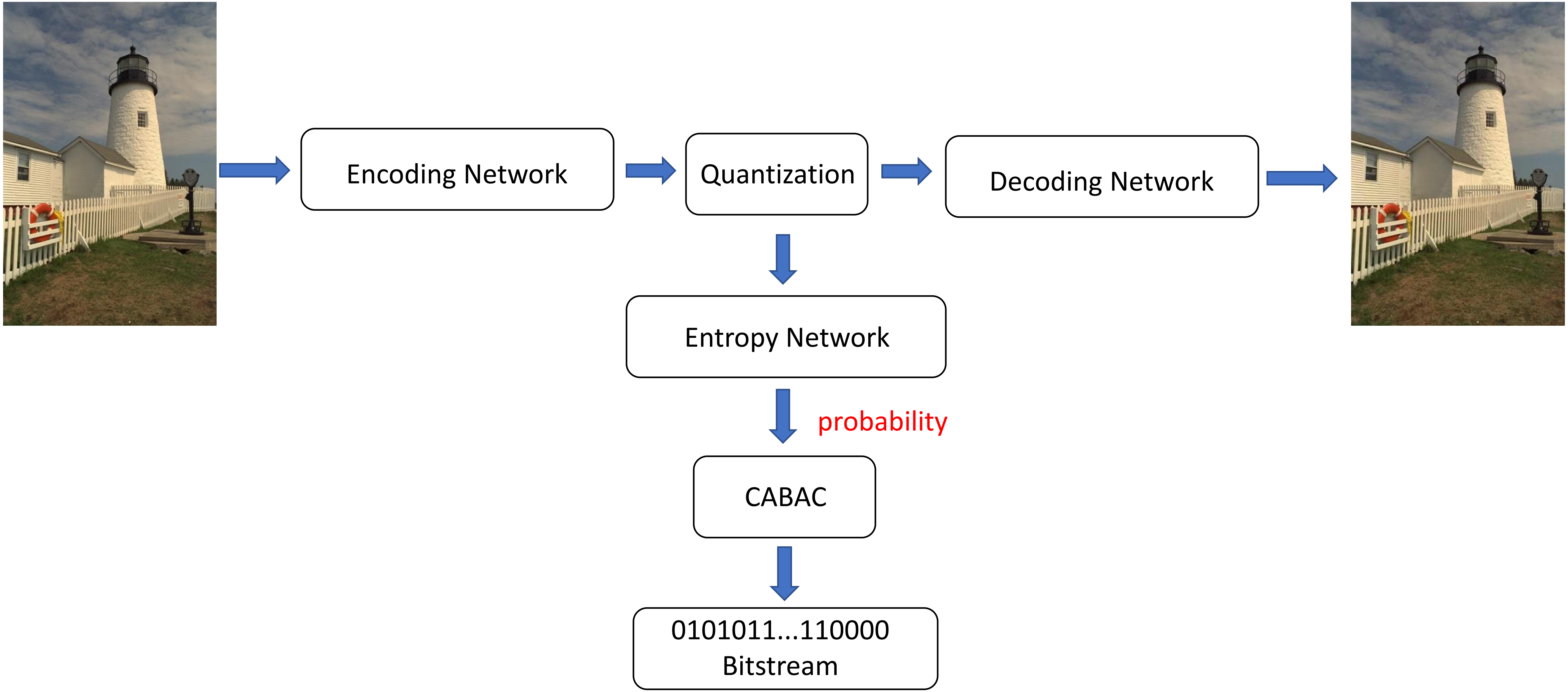
Pipeline of JPEG

Pros & Cons: Much improved but still with visual artifact (e.g., rings)



Pipeline of JPEG 2000

Introduction: deep neural networks



❖ Deep Neural Network

- **Pros:** end-to-end training; adaptively learn an effective encoder-decoder from a large amount of image data and in a larger context to represent more complex image structures; reduce the visual artifacts in the decompressed image.

- **Cons:** time-consuming, uniform quantization

Formulation

$$\hat{Y} = \mathbf{D}(\mathbf{Q}(\mathbf{E}(X, \Omega)), \Theta)$$

Reconstructed image

Decoder

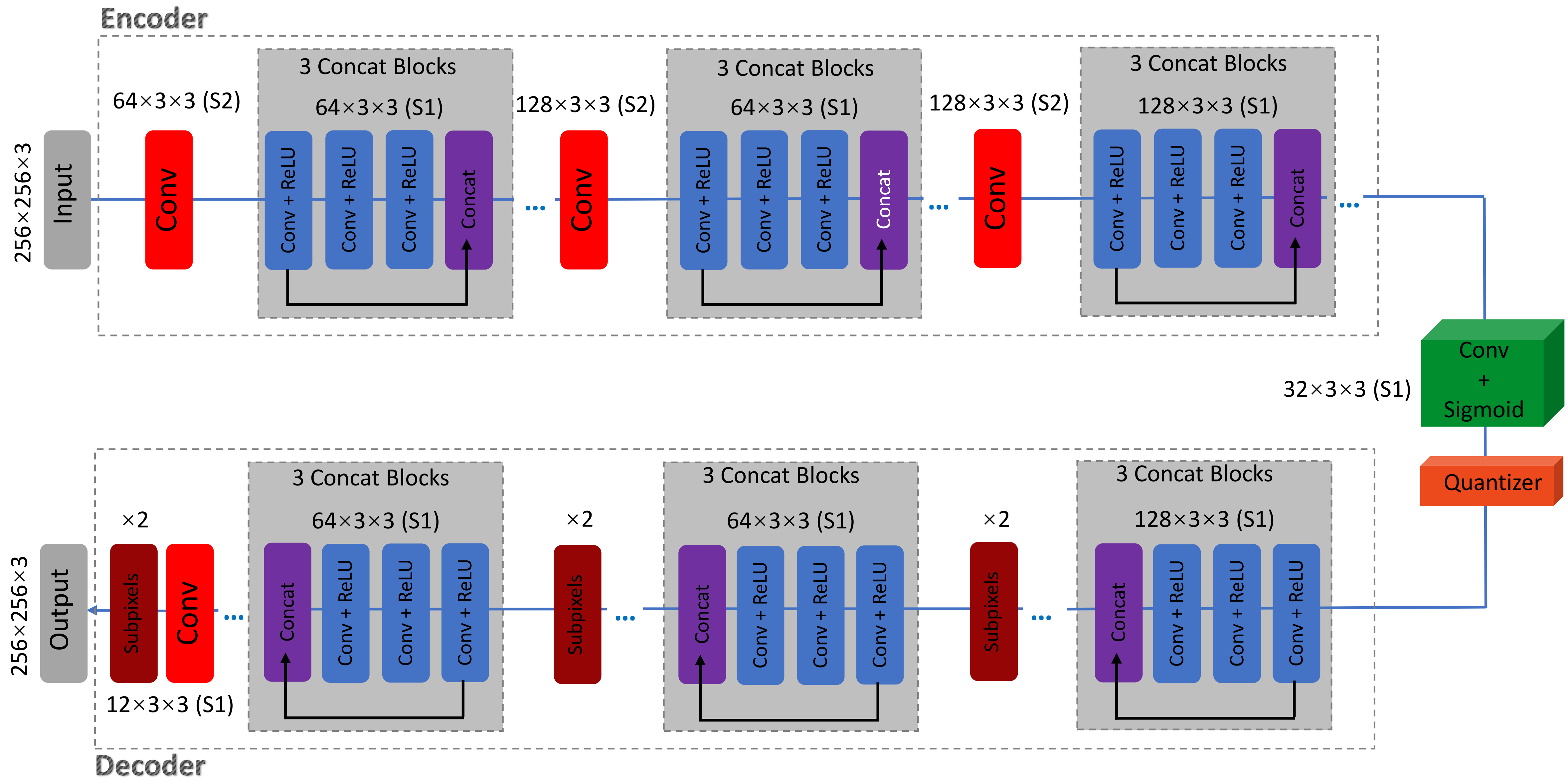
Quantizer

Encoder

Original image

Network parameters

Proposed Method: network architecture



Loss function

$$l(\Omega, \Theta) = \frac{1}{n} \sum_{i}^{n} \| x_i - \mathbf{D}(\mathbf{Q}(\mathbf{E}(x_i, \Omega)), \Theta) \|_1$$

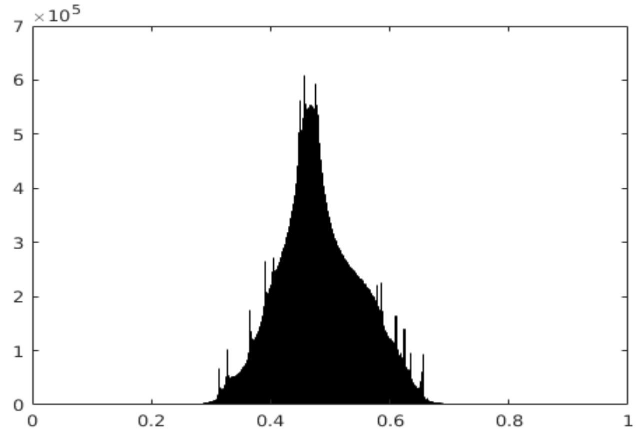
Proposed Method: alternative network and quantizer training

With the encoder, we have:

$$z = \mathbf{E}(x, \Omega)$$

z: the latent image feature

The probability density function (PDF) of z, denoted by P(z):



Proposed Method: alternative network and quantizer training

The optimal quantizer:

$$\mathbf{Q}^{\star}(z) = \underset{\mathbf{Q}}{\operatorname{argmin}} \int p(z)(\mathbf{Q}(z) - z)^2 \, dz.$$

quantized values z_q:

$$\hat{z}_q = \frac{\int_{b_{q-1}}^{b_q} z p(z) \, dz}{\int_{b_{q-1}}^{b_q} p(z) \, dz} \;;$$

decision boundaries b_q:

$$b_q = \frac{1}{2}(\hat{z}_q + \hat{z}_{q+1}).$$

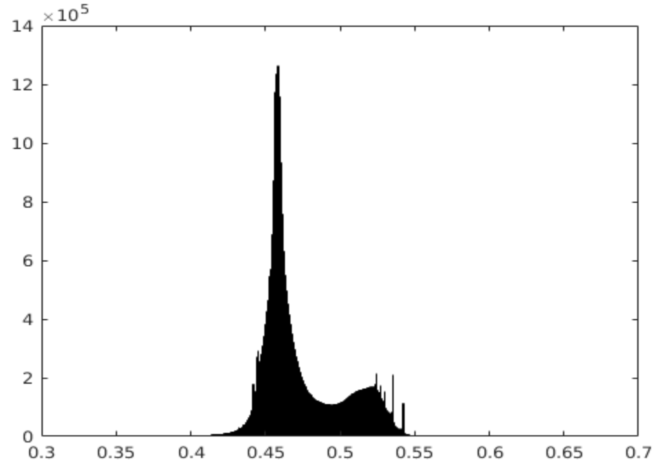
Proposed Method: alternative network and quantizer training

Fine-tune the network with the updated quantizer:

$$l(\Omega, \Theta) = \frac{1}{n} \sum_{i}^{n} \| x_i - \mathbf{D}(\mathbf{Q}(\mathbf{E}(x_i, \Omega)), \Theta) \|_1$$

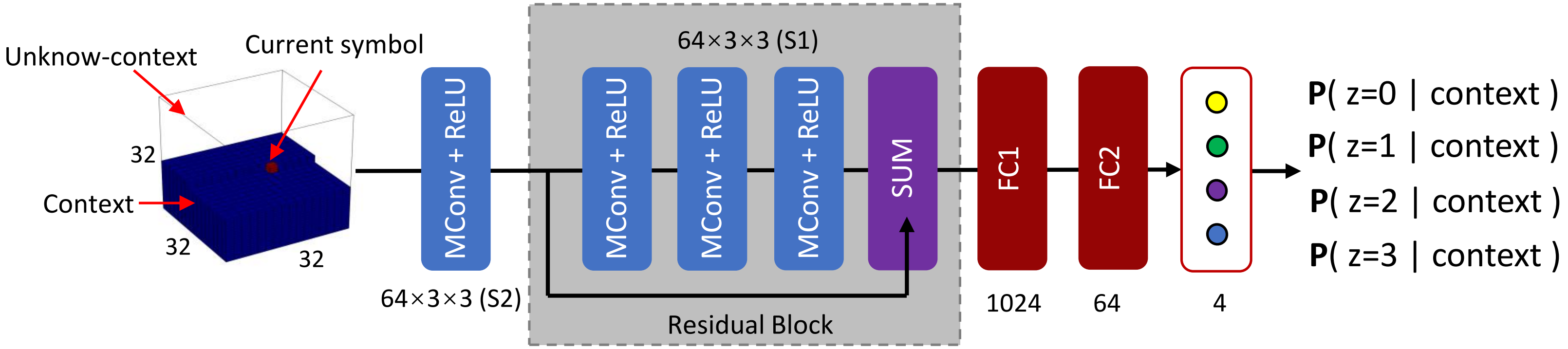
Updated latent image feature z: $z = \mathbf{E}(x, \Omega)$

Updated PDF of z:



... Alternative optimization

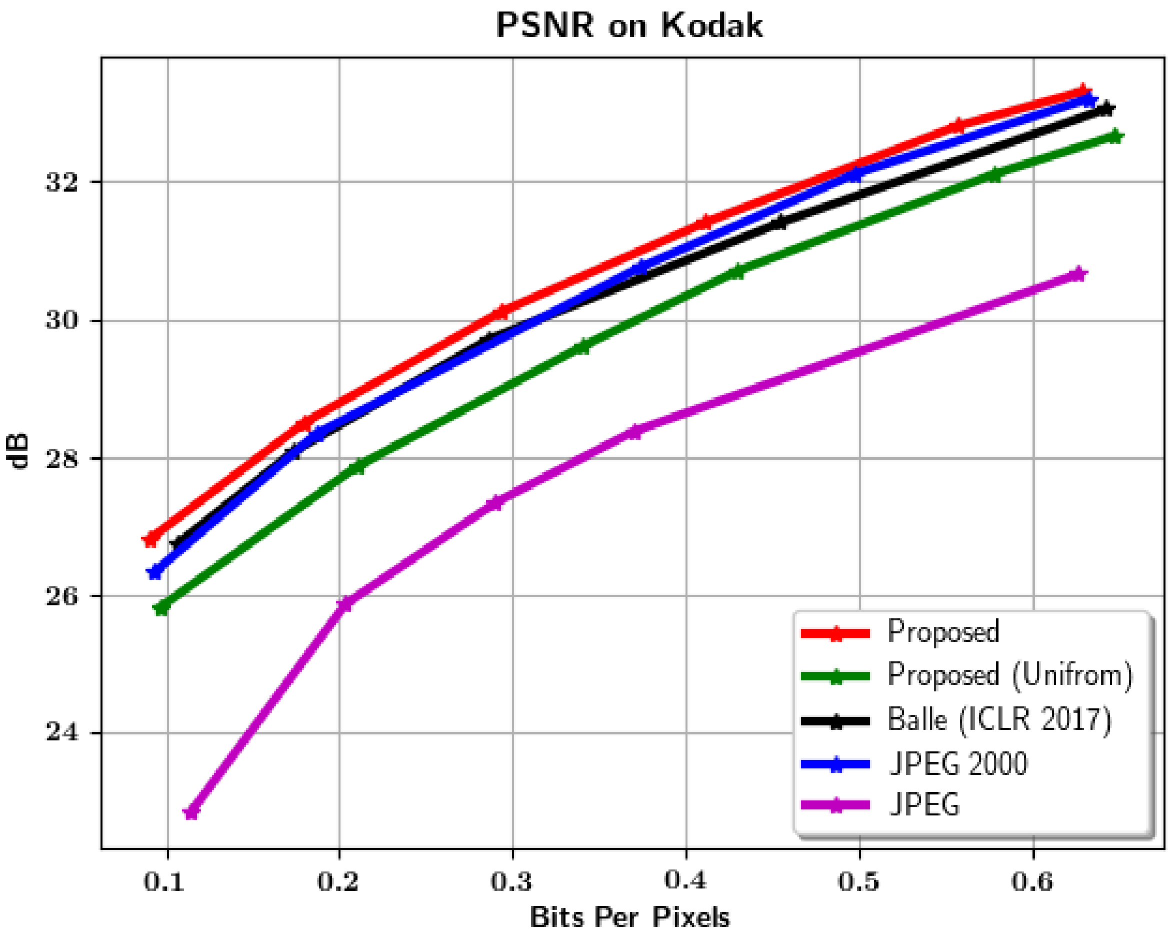
Proposed Method: entropy network



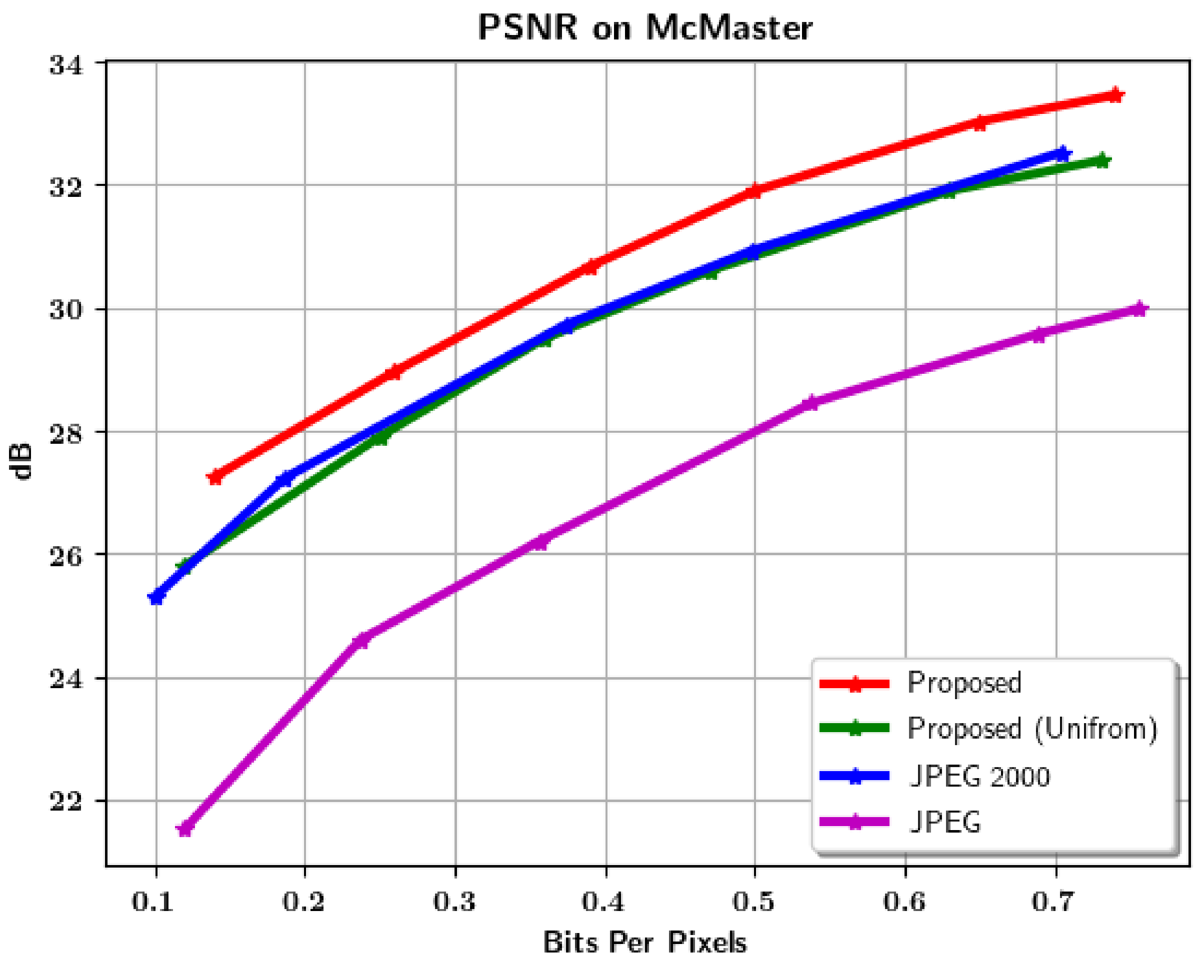
$$\mathbf{H}(\hat{z};\Pi) = -\frac{1}{C}\Big[\sum_{i=1}^{C}\sum_{j=0}^{3} 1\{\hat{z}^{(i)} = j\}\log(\mathbf{P}(\hat{z}^{(i)} = j|\hat{z}^{(i)};\Pi))\Big],$$

$$\mathbf{P}(\hat{z}^{(i)} = j|\hat{z}^{(i)};\Pi) = \frac{e^{\Pi_j^T \hat{z}^{(i)}}}{\sum_{k=0}^{3} e^{\Pi_k^T \hat{z}^{(i)}}}.$$

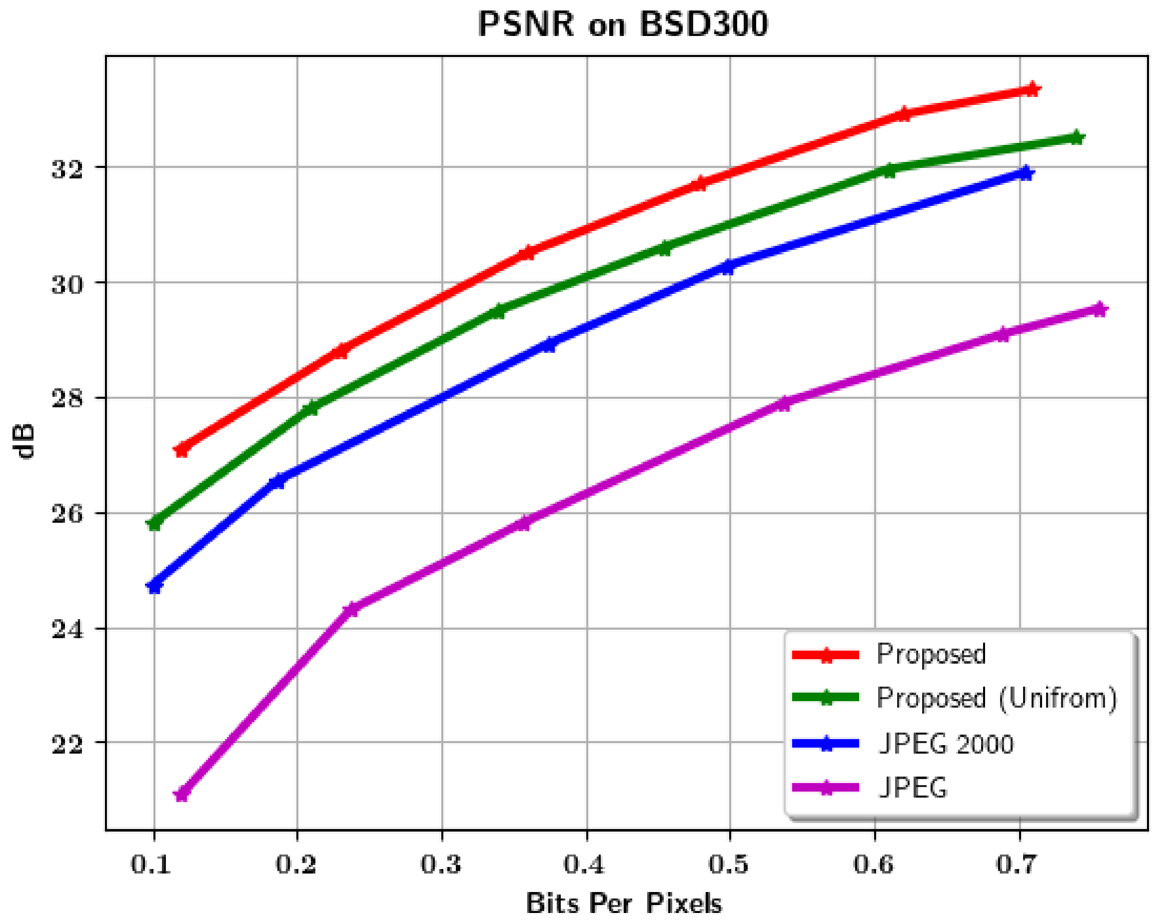
Experimental results: Kodak dataset



Experimental results: McMaster dataset



Experimental results: BSD300 dataset

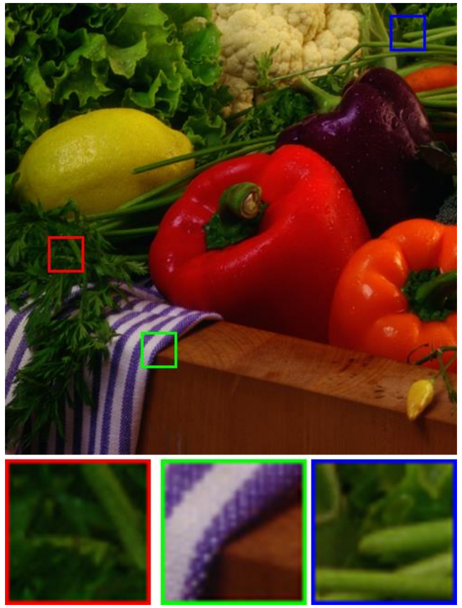


Experimental results: visual examples

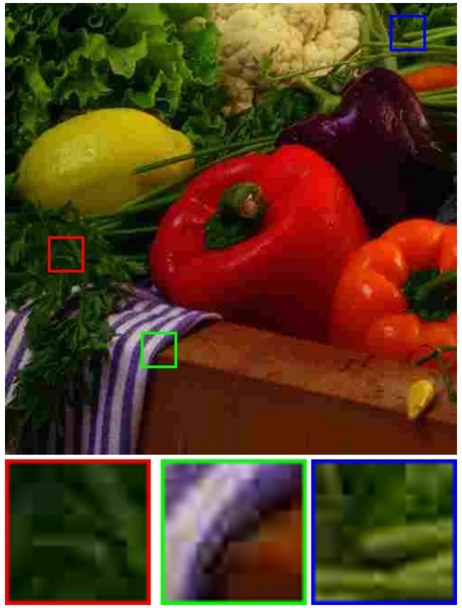
(a) **Original** (b) **JPEG** (BPP: 0.221 PSNR: 25.34db) (c) **JPEG2000** (BPP: 0.186 PSNR: 26.82db)

(d) **Ballé** (BPP: 0.175 PSNR: 27.04db) (e) **Our-Uniform** (BPP: 0.171 PSNR: 26.74db) (f) **Proposed** (BPP: 0.181 PSNR: 27.19db)

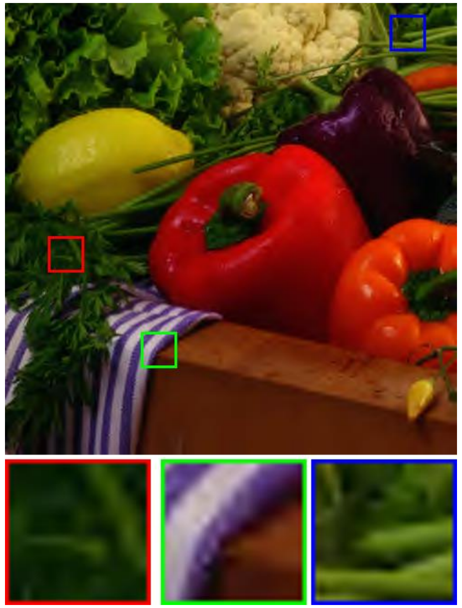
Experimental results: visual examples



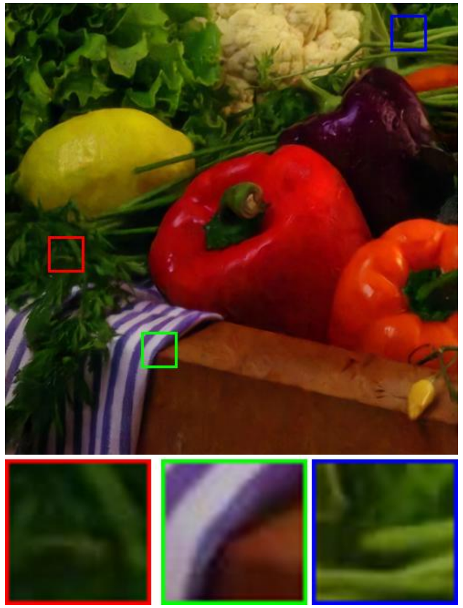
(a) original



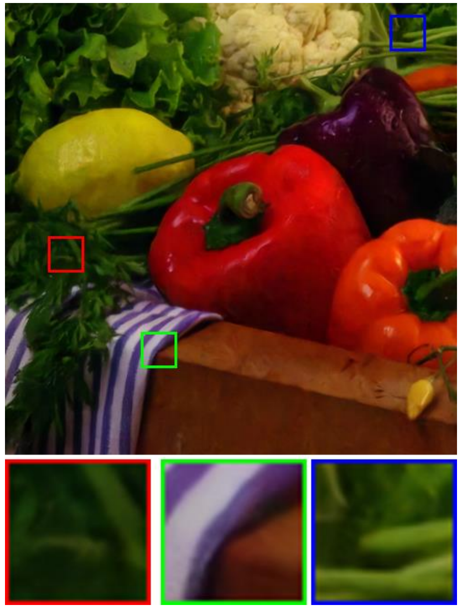
(b) JPEG



(c) JPEG 2000



(d) Uniform



(e) Proposed

Conclusions

- We presented an iterative non-uniform quantization scheme for deep image compression network.
- The quantizer and the encoder-decoder network are updated alternatively.
- Compared with previous deep compressors, our method exhibits better PSNR based rate-distortion curves, as well as better visual quality.

Thanks for your time!

Q & A

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Department of Computing
電子計算學系