基于端到端学习的图像编码研究及进展

陈积敏[1]，林泽昊[2*]

[1]南京森林警察学院，江苏 南京 210023；

[2]东华大学信息科学与技术学院，上海 201620

摘要 在图像大数据应用背景下，伴随着硬件技术的高速发展，基于深度学习的图像视频编码技术逐渐成熟。基于端到端学习的压缩框架因能更高效地对原始图像数据进行紧致表达，在学术界和工业界都得到了广泛的关注。系统地总结了基于端到端学习的图像压缩框架中的核心模块如变换、量化、熵编码和损失函数的研究现状，对其研究进展和关键技术进行了概括性的介绍，并对前沿研究成果进行了性能比较。

关键词 图像处理；图像压缩；端到端学习；深度学习

中图分类号 TN919.81 **文献标志码** A **doi**：10.3788/LOP57.220002

End-to-End Learning-Based Image Compression: A Review

Chen Jimin[1], Lin Zehao[2*]

[1] *Nanjing Forest Police College, Nanjing, Jiangsu* 210023, *China*;

[2] *College of Information Science and Technology, Donghua University, Shanghai* 201620, *China*

Abstract In the big data era, we have witnessed the explosive growth of deep learning based image and video compression technologies. Such end-to-end learning-based compression frameworks have demonstrated promising efficiency for compact representation of original image data, and attracted a vast attention from both academia and industry. A systematic review of transformation, quantization, entropy coding, and loss function used in end-to-end learning-based image compression framework is introduced in this work. The research progress and key technologies are briefly introduced, as well as the comparative studies of coding performance for existing methods with leading efficiency.

Key words image processing; image compression; end-to-end learning; deep learning

OCIS codes 100.2000; 100.2960

1 引 言

现今，图像视频应用愈来愈广泛，如以视频为主要信息载体的远程教育、远程医疗、互联网直播和视频会议等极大地方便了人们的工作和生活。但这些应用的蓬勃发展对网络带宽带来了诸多挑战：一方面是用户基数庞大，仅中国在线直播用户已破五亿人次并仍不断上升；另一方面是人们对高质量（如高清、超高清）图像视频的持续渴求。因此，人们迫切需要更高效的图像视频压缩技术来面对和解决这些严峻的挑战。

近年来，随着硬件技术的不断发展，基于深度学习的图像视频编码研究受到广泛关注，技术发展逐渐成熟[1-3]。目前深度学习中流行的循环神经网络（RNN）、卷积神经网络（CNN）、生成对抗网络（GAN）等架构在图像视频编码方面得到广泛的应用，如 Ballé 等[4]于 2016 年提出的基于端到端学习的 CNN 框架的图像编码取得了与 JPEG2000 相媲美的性能，Rippel 和 Bourdev[5]于 2017 年在图像编码框架中引入 GAN 优化，实现了低码率时超越 HEVC Intra 的高主观质量重建。

前期应用神经网络技术的图像视频编码方法主

收稿日期：2019-12-18；**修回日期**：2020-02-17；**录用日期**：2020-04-17

基金项目：国家林业局软科学项目（2017-R06）、江苏省哲学社会科学优秀科技创新团队（生态环境保护执法）

* **E-mail**：lzhtocoffee@163.com

要优化压缩框架中的编码工具[6]，其主要思路都是将传统混合编码框架[7]和混合视频编码框架(HVC)[8]中的模块与神经网络相结合来提升性能，但基于模块改进的性能提升有限且实现复杂。因此有必要采用端到端学习来实现整体性能的提升以满足日益增长的压缩需求。

端到端学习省去了传统方法中需要手动设计、联合优化多个模块的复杂步骤，而是将输入通过一个多层叠加的神经网络，学习输入与输出的映射关系，并以此得到对应的输出。端到端学习的重要特点之一是让“数据说话”。由神经网络组成的编解码器需要大量图像数据训练模型，而当前大数据时代的海量图像视频资源很有效地解决了这个问题，让基于学习的图像视频编码成为可能；同时，基于学习的编码还具有以下的优点，如能够进行建模和分解图像中的语义信息并将其以更高效的形式存储和生成结构化码流[9]。

基于端到端学习的图像编码研究是从Ballé[4,10-11]、Toderici[12-13]、Theis[14]等的研究开始的。最初的研究在端到端学习的压缩框架中引入了基于CNN或者RNN的自编码器，实现了可观压缩率下图像编码的整体优化重建。为了更好地表达图像相关性，研究者们一方面拓展研究了多种不同结构的自编码器，包括变分自编码器(VAE)、多尺度自编码器(MSAE)等，另一方面引入超先验信息进行融合预测以提升端到端学习能力。Ballé等[11]证明了超先验模型中率失真优化可等效为最小化数据分布的KL散度(Kullback-Leibler divergence)，Zhou等[15]提出了基于注意力机制的超先验自编码器。

除上述框架的创新外，神经网络技术也被拓展应用于图像编码中的变换、量化、熵编码等核心模块以提升性能。变换从传统的线性离散余弦变换(DCT)和小波变换逐步发展到非线性变换，如广义分歧归一化(GDN)变换。在端到端学习框架中，变换等效为利用逐层卷积来提取特征激活(fmaps)；量化由标量量化逐步发展到矢量量化，实现从round函数到现在流行的软量化[15-17]和格型矢量量化[18-19]，这样可在满足数据压缩的同时保证反向传播梯度可导；熵编码从最初JPEG使用的Huffman编码[20]，发展到基于超先验和递归近邻概率混合预测的算术编码[21]，实现编码性能的大幅提升；此外，损失函数从广泛使用的L1损失和L2损失，发展到改善收敛不稳定和局部最优解问题的交叉熵损失[16,22]以及改善图像主观质量的感知损失[15,23-24]和对抗损失[23-25]，再到现在的复合损失函数[15,18,23,24]。

现有对基于神经网络的图像视频编码研究的综述[26]，其清楚地梳理了基于神经网络的图像视频编码研究及进展，但是对于端到端学习的研究描述相对较少，因此本文系统地回顾了基于端到端学习的图像编码的相关研究成果和最新进展，以对图像编码的研究综述进行进一步的扩展，对现有的前沿研究进行性能比较和分析，并对该领域的技术研究进行多方面的展望。

2 基于端到端学习的图像编码

基于端到端学习的图像编码研究的发展历程如图1所示，前期代表人物为Toderici[12-13]和Ballé等[4,10-11]，他们在2016年就分别将迭代RNN和非线性变换融入端到端学习压缩框架中，极大地推动了该领域的发展。至今，已衍生出了该方向的很多研究，如2017年Rippel和Bourdev[5]提出了嵌套金字塔解构和生成对抗模块的深度自编码器，并引入对抗损失进行端到端训练；2018年Mentzer等[22]在自编码器结构中引入上下文模型，利用卷积实现前后概率预测以更好地优化率失真性能，Ballé等[11]之后提出了基于高斯比例混合(GSM)的熵模型，Minnen等[27]在此基础上将GSM推广至以超先验为条件的均值和方差预测高斯混合模型(GMM)；2019年Chen等[28]在VAE结构的基础上提出一种新的非局部注意力优化和基于上下文模型的图像压缩算法(NLAIC)等，实现了在峰值信噪比(PSNR)和结构相似度(SSIM)两种失真度量下的领先编码性能。可见大多数基于端到端学习的图像编码方法都依赖于自编码器的训练，以充分利用空间相关性和数据统计分布，可在码率和失真之间取得良好的平衡，并且可以针对任意失真指标进行快速优化，拥有可媲美甚至超过现有国际图像编码标准(如JPEG、JPEG2000、HEVC Intra等)的压缩性能。

传统的图像编码器可分为变换、量化和熵编码三个独立模块，而端到端学习则是将三个模块联合进行整体优化。本节将分别论述基于端到端学习的图像编码框架中这三个模块和损失函数的研究现状及进展，并对各模块中使用的方法进行了分析与对比。

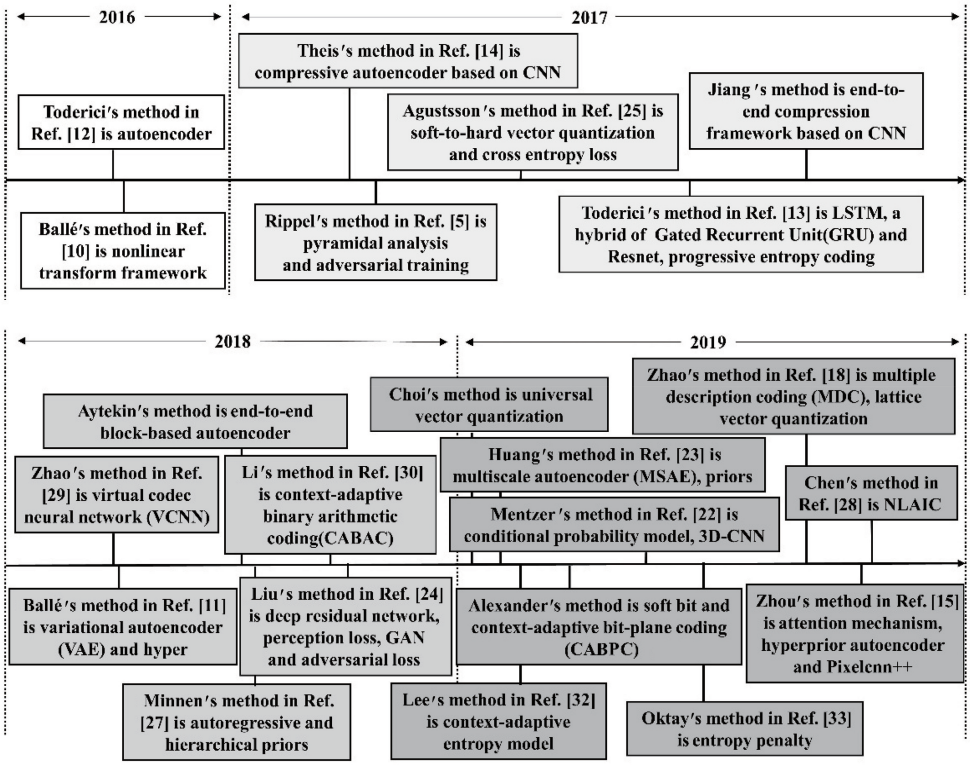

图 1 基于端到端学习的图像编码技术发展历程

Fig. 1 Technical roadmap of end-to-end learning-based image compression

2.1 变 换

图像变换编码将空域图像像素转换为变换域系数，实现能量聚集的紧致表达，以达到压缩的目的。大多数压缩方法都使用正交线性变换来降低数据的相关性。在传统的变换方法中，最早针对信号冗余解耦优化的线性变换可以追溯至 KL 变换和主成分分析法（PCA）。之后国际图像编码标准 JPEG 和 JPEG2000 分别使用的离散余弦变换和小波变换也均为线性变换。

但是正交线性变换中线性滤波器响应的联合统计量呈现了很强的高阶依赖性，为解决此问题可联合局部非线性进行增益控制。近几年，端到端学习将非线性变换融入图像压缩框架中。其中，Ballé 等[4,10]提出了基于非线性变换编码的端到端学习框架（图 2），将图像强度向量 $\boldsymbol{x}$ 先通过分析变换 $\boldsymbol{y}=g_{\mathrm{a}}(\boldsymbol{x};\boldsymbol{\phi})$（其中 $\boldsymbol{\phi}$ 为学习参数向量）映射到编码域，再通过量化处理得到离散值向量 $\boldsymbol{q}$，之后进行熵编码，相对应地，由离散值向量 $\boldsymbol{q}$ 估计连续值向量 $\hat{\boldsymbol{y}}$，应用生成变换 $\hat{\boldsymbol{x}}=g_{\mathrm{s}}(\hat{\boldsymbol{y}};\boldsymbol{\theta})$（其中 $\boldsymbol{\theta}$ 为学习参数向量），并进行像素重建；编码决策通过率失真优化性能，常见的失真度量包括均方误差（MSE）和 SSIM，也可引入感知失真等进行性能优化，最后端到端学习系统通过优化学习参数向量 $\boldsymbol{\phi}$ 和 $\boldsymbol{\theta}$ 来最小化码率 R 和失真 D 的加权和 $R+\lambda D$，其中，λ 控制码率和失真的平衡。分析变换分为三个阶段：卷积、下采样和 GDN 变换，作为其逆变换的生成变换也分为三个阶段：仿射卷积、上采样和 GDN 逆（IGDN）变换，且两类变换中的上下采样操作均可通过卷积来实现，从而提高了计算效率。感知变换中归一化拉普拉斯金字塔模型（NLP）与 GDN 的组合考虑了图像局部亮度和对比度的误差，相较于采用 MSE 优化 DCT 的传统方法而言，在相似重建质量的情况下降低了码率。

现今，自编码器被越来越广泛地用于图像压缩中。这些研究利用单个自编码器或循环自编码器在瓶颈层生成 fmaps，用于后续的量化和熵编码[24]。典型的自编码器结构包含三个部分：编码器、表示压缩数据的瓶颈和解码器，将这三个部分级联并进行端到端训练。由于传统 JPEG 等算法中线性变换对空间相关性和压缩数据分布的利用不够充分，使用深度卷积神经网络可实现非线性变换，对图像分布进行更好的冗余解耦，实现更紧致的特征表达并实现更好的压缩[14,18,29-30]。

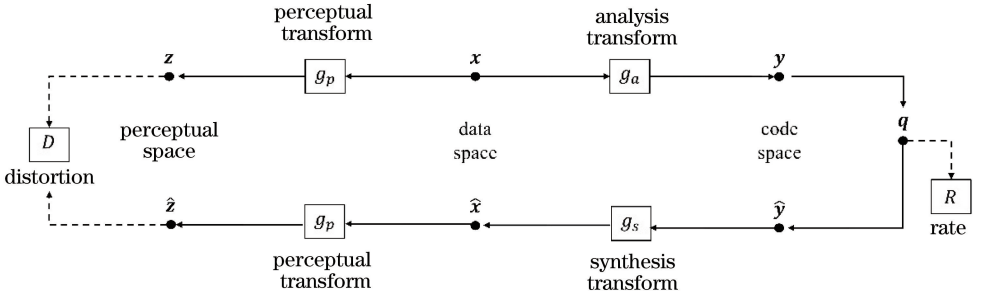

图 2 基于非线性变换的端到端学习图像编码框架[10]

Fig. 2 Nonlinear transform based end-to-end learning image compression[10]

传统自编码器结构复杂、时间复杂度高，受 Shi 等[31]的工作启发，Theis 等[14]提出基于卷积神经网络的压缩式自编码器(CAE)，对图像先进行卷积以提取特征再进行上采样，并在解码器中采用了子像素结构，该方法适用于高分辨率图像压缩并可大幅度提升计算效率。对于图像编解码而言，可以通过级联多个卷积神经网络进行定义分析和生成变换，并允许以端到端学习的方式联合优化非线性的编码器和解码器。因此绝大多数研究都采用了不同的卷积神经网络进行非线性变换的设计，如 Zhao 等[29]利用由 CNN 组成的特征描述神经网络(FDNN)在低维空间对真实图像(ground-truth image)进行有效的描述以大幅减少图像所包含的数据量，Li 等[30]用多个卷积层定义了非线性分析变换。

综上，基于端到端学习的图像编码框架中的变换方法从先前的正交线性变换发展到非线性变换。现有的图像变换编码的主要作用在于提取特征以进行更紧致的表达，且使用深度卷积神经网络的自编码器是如今变换编码的主流方法。

2.2 量　　化

在传统的图像压缩框架中，量化参数与图像质量和码率(压缩率)息息相关。而在端到端学习的图像压缩框架中，量化将变换后的特征激活值由浮点数转换为规则定点数，作为后续熵编码的输入。最常用的规则量化方法是取整函数——round 函数。

由于目标失真函数主要使用梯度下降法优化端到端编码中的率失真，反向传播中要求量化函数全局可导[6]，所以基于端到端学习的图像压缩研究一直围绕着解决量化的不可导问题(量化不连续，其导数在任何地方都为零或无穷大)而展开。Ballé 等[4,10]使用加性均匀噪声源代替了标量量化器实现全局可导。

$$\hat{\boldsymbol{y}}_i = \boldsymbol{q}_i = \mathrm{round}(\boldsymbol{y}_i), \tag{1}$$

$$P_{\boldsymbol{q}_i}(n) = (p_{\boldsymbol{y}_i} * \mathrm{rect}), n \in \mathbf{Z}, \tag{2}$$

$$\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i + \Delta\boldsymbol{y}_i, \tag{3}$$

$$p_{\tilde{\boldsymbol{y}}_i} = p_{\boldsymbol{y}_i} * \mathrm{rect} = \int_{n-\frac{1}{2}}^{n+\frac{1}{2}} p_{\boldsymbol{y}_i}(t)\mathrm{d}t, n \in \mathbf{Z}, \tag{4}$$

$$p_{\boldsymbol{y}_i}(n) = P_{\boldsymbol{q}_i}(n), n \in \mathbf{Z}, \tag{5}$$

式中：i 为索引，遍历向量中的所有元素；$\boldsymbol{y}_i$ 为图片强度向量 $\boldsymbol{x}_i$ 经过分析变换后的结果；$\hat{\boldsymbol{y}}_i$ 和 $\boldsymbol{q}_i$ 为 $\boldsymbol{y}_i$ 经过量化后得到的向量；n 指第 n 个量化区间；$P_{\boldsymbol{q}_i}$ 为 $\boldsymbol{q}_i$ 的概率质量函数；$p_{\boldsymbol{y}_i}$ 为 $\boldsymbol{y}_i$ 的密度；$*$ 代表连续卷积；rect 为 $\left[-\frac{1}{2},\frac{1}{2}\right]$ 的均匀分布；$\Delta\boldsymbol{y}_i$ 为均匀噪声源，即 rect；$\tilde{\boldsymbol{y}}_i$ 为加上均匀噪声后的 $\boldsymbol{y}_i$。

证明过程如下：通常假设量化区间为 1，对于传统的标量量化，利用 round 函数得到(1)式，此时量化得到的 $\boldsymbol{q}_i$ 的概率质量函数可表示为(2)式；而如果给 $\boldsymbol{y}$ 加上独立均匀噪声，该过程可描述为(3)式，加上噪声后的 $\tilde{\boldsymbol{y}}$ 的密度函数可表示为(4)式。观察(2)式和(4)式可知两者相等，即得到(5)式，因此证明了均匀噪声可以模拟量化的过程，并使全局具有了可微性。Liu 等[24]延续此方法，利用离散 $P_{\boldsymbol{q}_i}$ 的分段线性逼近来保证它的连续且可导。

为了便于图像传输、实现对压缩率的分层精细控制，Toderici 等[12]通过使用随机二值量化学习到相较于有许多冗余位的标准浮点层更有效的表示。此研究将二值化分为两步：第一步是在连续区间 $[-1,1]$ 内产生具有一定比特数的输出，第二步是将实值表达作为输入，得到 $\{-1,1\}$ 的离散输出：

$$b(x) = x + \varepsilon,\ b(x) \in \{-1,1\}, \tag{6}$$

$$\varepsilon = \begin{cases} 1-x, & \text{with probability of } \dfrac{1+x}{2} \\ -1-x, & \text{with probability of } \dfrac{1-x}{2} \end{cases}, \tag{7}$$

式中：$b(x)$ 为二值化函数；x 为实值；ε 等同于量化噪声。

Rippel 等[5]和 Li 等[30]延续了 Toderici 等提出

的二值量化的量化方法，并在反向传播中引入代理函数以对二值运算进行近似，使其具有可导性，从而解决了零梯度问题。

图 3 为 Theis 等[14]的研究，展示了在 JPEG 压缩框架中采用以上不同量化方法得到的图像质量。图 3(a)为压缩前裁剪的图像，图 3(b)为使用 round 函数进行量化后得到的图像，但图像出现了由于 DCT 系数取整所引起的块效应，图 3(c)为采用了随机量化得到的图像，该方法类似于 Toderici 等提出的二值化量化，而图 3(d)则为采用了 Ballé 等提出的用加性均匀噪声源代替标量量化器的方法得到的图像，可以看出图像中出现了高频噪声。

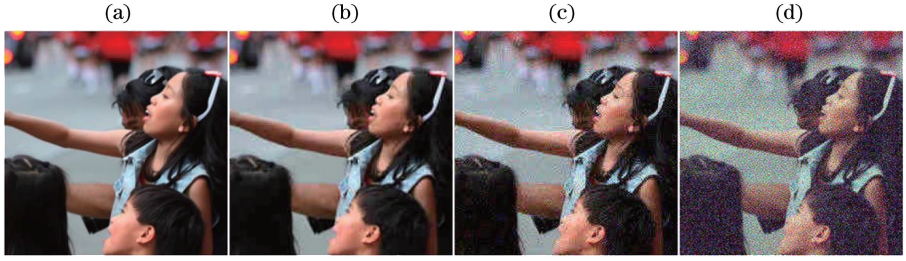

图 3 在 JPEG 压缩框架中采用不同量化方法得到的图像[14]。(a)原始图像；(b)取整；(c)随机取整；(d)加性噪声

Fig. 3 Images obtained by different quantization methods using JPEG compression[14]. (a) Original image; (b) rounding; (c) stochastic rounding; (d) additive noise

Theis 等[14]提出的量化方法具有与上述方法相同的前向传递，且具有不会改变解码器梯度的优点。他们将反向传播的 round 函数的导数替换为一个光滑的近似，将其导数定义为

$$\frac{\mathrm{d}}{\mathrm{d}x}[y]:=\frac{\mathrm{d}}{\mathrm{d}x}r(y), \tag{8}$$

式中：y 为 x 经过变换后的输出；$[y]$ 为 y 经过 round 函数量化后的值；r 为替换的光滑近似。

采用取整函数的量化将浮点数转换为整数，会显著地降低重建的图像质量，因此 Agustsson 等[16]在图像压缩的背景下探讨了矢量量化，提出软到硬(soft-to-hard)的量化方法，让网络结合权重学习量化级，将其应用于更广泛的问题中，并证明了矢量量化比标量量化更具优势。传统的矢量量化需要占据大量的存储空间且需要进行复杂的近邻搜索，对编码器复杂度要求过高，因此 Zhao 等[18]提出多描述格型矢量量化，应用对称结构避免了复杂的邻搜索。

由此可知，为解决量化的不可导问题，最常见的方法是随机近似和用光滑导数近似的 round 方法。如今矢量量化相较于标量量化成为更具竞争力的量化方法，提出的软矢量和格型矢量的量化方法可在保证重建质量的同时又使模型具有可微性。

2.3 熵编码

熵编码通过减少统计冗余进一步提升图像压缩率。早期的深度熵编码算法能够使压缩性能得到一定的提升，但在基于端到端学习的超先验模型出现后，熵编码能够提供的压缩性能愈来愈高。

常用的熵编码大多是变长编码(VLC)，其中包括 Huffman 编码和算术编码[21]。就现有的国际图像编码标准而言，除使用 Huffman 编码的 JPEG 以外，其余的国际图像标准都选择使用算术编码，目前算术编码因其可以在一个定义良好的上下文中表现出更高的压缩率且能将量化后的 fmaps 转化为码流的优点，已成为更具竞争力的熵编码选择。

基于端到端学习的图像压缩的重要组成部分之一是用于隐式表达的可训练熵模型，因为隐式表达的实际分布是未知的，熵模型提供了通过近似分布来估计编码隐式表达所需比特的方法，从而显著提高了基于神经网络的图像压缩性能[32]。熵模型是由 Ballé 等[4]和 Theis 等[14]首次提出的，他们的工作对之后的研究做出了极大的贡献。前者认为隐式表达的熵模型为非参数模型，而后者采用 GSM 模型，其共同点在于尽可能地学习统计分布。为减少先验和边缘信息的不匹配，Ballé 等[11]在先前研究的基础上，通过在隐式表达的局部尺度参数上引入超先验来捕获空间依赖性，得到更好的模型匹配，从而增强熵模型、提升压缩性能。他们根据输入自适应估计表达尺度，并使用 GSM 的方法将压缩的超先验作为辅助信息添加到生成的码流中，从而允许解码器使用条件熵模型。同年，受概率生成模型的启发，Minnen 等[27]提出了包含超先验和递归近邻概率融合的自编码器，如图 4 所示，以两种方式扩展了基于 GSM 的熵模型：一是将 GSM 模型推广至 GMM，在不增加模型复杂度的情况下呈现出更好的率失真性能；二是将递归模型与超先验模型相结

合。这两种结构可以互补，从而更好地利用隐式表达的概率结构。近期，Oktay 等[33]又提出了一种基于端到端学习的神经网络权值的压缩方法，该方法先在隐式空间进行重参数化，在训练过程中采用概率模型对参数表示施加熵惩罚[33]，完成训练后使用算术编码压缩隐式表达。

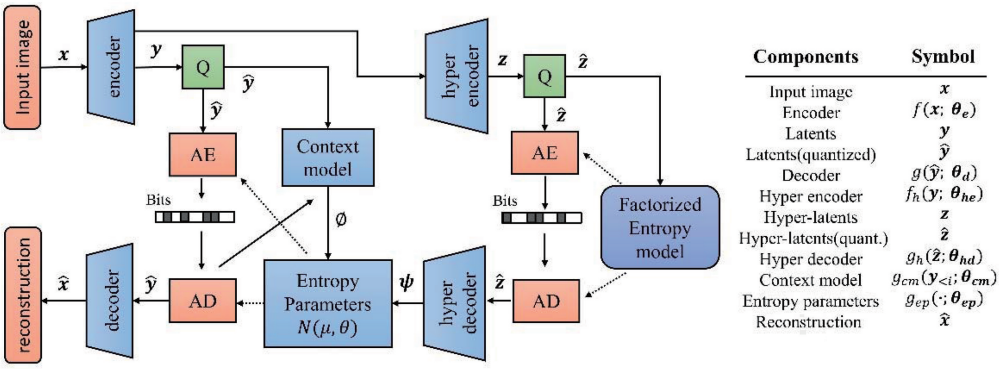

图 4 超先验和递归近邻概率融合的自编码器[27]

Fig. 4 Autoencoder based on hyperprior and recursive nearest neighbor probability fusion[27]

在 Ballé 等的研究基础上，Lee 等[32]发现压缩性能在本质上是取决于熵模型的容量的，因此为扩大熵模型的容量，提出了一种上下文自适应熵模型的框架，根据是否需要额外的比特分配，使用两种类型的上下文：比特消耗上下文和无比特消耗上下文。通过使用以上熵模型来更准确地估计每个隐式表达的分布，从而更有效地减少相邻隐式表达之间的空间依赖性，达到提高性能的目的，该研究实现的图像压缩性能在 PSNR 和多尺度结构相似性（MS-SSIM）方面优于图像压缩的国际标准 BPG。

综上所述，传统图像压缩框架的熵编码基本都是 Huffman 编码和算术编码，后者相较于前者而言能够准确地呈现概率分布，更能逼近香农提出的理论熵值。随着超先验的加入，结合了神经网络和算术编码的混合熵模型因为有着能够近似分布、估计编码隐式表达所需的比特以及克服隐式表达分布未知的问题这几个优势，因此被广泛应用于基于端到端学习的图像编码框架中。

2.4 损失函数

损失函数作为深度学习中不可或缺的一部分，是用来指导网络训练的重要工具。其原理是通过对比预测值和真实值，根据所得的损失并考虑模型求导数的难易程度、梯度下降的效率、采用的算法等因素，选择出合适的损失函数，以减小损失，并对模型进行相应的优化。

最典型的损失函数有 L2 损失函数（MSE）和 L1 损失函数（平均绝对误差）。现有研究中的损失函数都以此为基础进行拓展和优化[13,29,34]，为解决传统损失函数最后一层权重的梯度和激活函数的导数直接相关导致收敛速度不稳定及易造成局部最优解的问题，Agustsson 等[16]和 Mentzer 等[22]引入了基于 softmax 计算的交叉熵损失函数。L2 损失函数易产生边缘模糊并造成细节信息的丢失，因此为得到更好的主观图像质量，感知损失和对抗损失得到了更多的应用[15,23-25]，如 Huang 等[23]通过引入感知损失函数并在端到端学习的压缩框架中采用对抗损失训练来实现极限压缩，Zhou 等[15]提出了由 L2 损失和感知损失函数组成的损失函数来训练自编码器。

为获得更好的率失真性能，损失函数的研究向着多个损失函数联合的复合损失函数发展，如 Huang 等[23-24]在 MSE 和 MS-SSIM 这两个常用的损失函数的基础上提出感知损失和对抗损失，以改善重建图像的主观质量。感知损失可以优化特征域失真以大幅增加感知信息，对抗损失可实现更快的收敛和更稳定的性能，在较低码率时可解决模糊和轮廓问题，最终的损失函数由 L1 损失、L2 损失、感知损失和对抗损失以一定的权值相加而得。

$$L_{\text{percept}}=\frac{1}{N}\sum_{n=1}^{N}\left\|\Psi(Y_n)-\Psi(X_n)\right\|^2, \tag{9}$$

$$L_{\text{generator}}=-D(Y_n), \tag{10}$$

$$L_{\text{discriminator}}=D(Y_n)-D(X_n)+\beta L_{\text{penalty}}, \tag{11}$$

$$L_{\text{final}}=L_2+\lambda_1 L_{\text{R}}+\lambda_2 L_{\text{percept}}+\lambda_3 L_{\text{generator}}, \tag{12}$$

式中：Y_n 为解码后的图像；X_n 为输入图像；n 为一个批次中的第 n 张图像；$L_{\text{discriminator}}$ 为判别神经网络的损失函数；L_{final} 为最终的损失函数；λ_1 为码率损失的参数；L_{R} 为码率损失；N 为 VGG 某层特征图的通道数；$\Psi(Y_n)$ 和 $\Psi(X_n)$ 分别为输入图像和重建

图像；D 为判别神经网络；L_{percept} 为感知损失；L_{penalty} 用来提升 WGAN(Wasserstein GAN)性能的惩罚；$L_{\text{generator}}$ 为对抗损失；β 为惩罚的参数，$\beta=10$；λ_2 为感知损失的参数，$\lambda_2=0.003$；λ_3 为对抗损失的参数，$\lambda_3=0.0001$。

L2 损失函数因其计算简便，是回归损失函数中最常用的误差函数，而之后提出的交叉熵损失函数解决了传统损失函数收敛速度慢、已陷入局部最优解的问题，MS-SSIM 损失函数则运用了结构信息的思想。研究发现复合损失函数相比单一损失函数可得到更好的率失真性能。感知损失可以增加感知信息，对抗损失则能使模型收敛得更快，得到更稳定的性能，两者与常用的损失函数相结合都可呈现出更高的图像质量。

3 性能比较

与传统的图像编码技术相比，深度学习在训练的阶段因其需要大量的数据量而呈现巨大的计算量和较高的时间复杂度。而现有的应用(如云计算)可以进行并行化，采用多 GPU 可实现数据和模型的并行，加速深度学习的训练。训练好的模型复杂度相对较低，硬件技术的发展和高效深度学习架构的设计[35]有望实现实时图像编解码。同时，如今众多的移动设备，如华为 Mate30、苹果 iPhone 都已经全面支持深度学习模块，极大地促进了基于学习的图像编解码的实际应用。除运算量这一评价指标外，图像质量是如今图像编码研究中的主要评价指标，本节先介绍了常用的图像质量评价指标，并从客观和主观两个角度对现有的基于端到端学习的几个领先研究与传统国际编码图像标准进行了图像质量性能的比较。

3.1 评价指标

图像压缩旨在生成低码率的高质量图像，因此衡量图像压缩性能好坏的标准有两个：码率和重建质量。

现有的图像质量评价指标可分为主观和客观两大类。主观质量评价是通过人眼对图像质量的直观反映来打分(rating)，而客观指标是在人眼视觉感知的研究基础上，通过构建相应的数学模型来量化图像质量。下面介绍三种广泛应用于图像压缩领域的客观质量评价指标，并对三种客观质量评价指标进行分析与对比。

1) 峰值信噪比

PSNR 是最早被广泛使用的图像质量客观评价指标。PSNR 是通过计算像素值的绝对误差所得到的，代表了像素级失真，其计算公式为

$$P_{\text{SNR}}=10\lg\frac{(2^m-1)^2}{E_{\text{MSE}}},\tag{13}$$

式中：m 为每像素的比特数；E_{MSE} 为原图像和处理后图像之间的均方误差，E_{MSE} 越小则 P_{SNR} 越大，说明重建质量越好。

2) 结构相似性

SSIM[36]是一种衡量两张图片相似度的指标，对于原图像和处理后的图像中的 $N\times N$ 窗口，SSIM 的计算公式为

$$\text{SSIM}(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)},\tag{14}$$

式中：μ_x、μ_y 分别是 x 和 y 的均值；σ_x^2、σ_y^2 是 x 和 y 的方差；σ_{xy} 是 x 和 y 的协方差；c_1、c_2 是稳定数值的系数。

3) 多尺度结构相似性

MS-SSIM[37]是对 SSIM 改进后的一项新指标，改进点在于通过多阶下采样实现对多尺度的计算，MS-SSIM 的计算公式为

$$l(x,y)=\frac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1},\tag{15}$$

$$c(x,y)=\frac{2\sigma_{xy}+c_2}{\sigma_x^2+\sigma_y^2+c_2},\tag{16}$$

$$s(x,y)=\frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3},\tag{17}$$

$$\text{MS-SSIM}(x,y)=l_M(x,y)\alpha_M\cdot\prod_{j=1}^{M}c_j(x,y)\beta_j s_j(x,y)\gamma_j,\tag{18}$$

式中：MS-SSIM(x,y)为多尺度结构相似性；$l(x,y)$为亮度对比；$c(x,y)$为对比度对比；$c_j(x,y)$为第 j 个尺度上的对比度对比；$s(x,y)$为结构相似度对比；$s_j(x,y)$为第 j 个尺度上的结构相似度对比；μ_x、μ_y 分别为 x 和 y 的均值；σ_x^2、σ_y^2 分别为 x 和 y 的方差；σ_{xy} 为 x 和 y 的协方差；c_1、c_2、c_3 为稳定数值的系数。原始图像的尺寸为 1，最大的尺度为 M。α_M、β_j、γ_j 用来调整各部分的占比，其中，亮度对比 $l_M(x,y)$只在尺度 M 上计算。

对比以上三种图片质量客观评价指标，PSNR 表征像素值的绝对误差，而人眼对像素的误差感知受多方面影响，因此会出现 PSNR 的结果和人眼的主观感受不匹配的情况，但由于 PSNR 计算简便且评价直接，如今仍被广泛地使用。而 SSIM 和 MS-SSIM 表征结构信息感知变化，如亮度、对比度和结

构相似度，其计算包括：用均值估计亮度，用标准差估计对比度，用协方差估计结构相似度。相较于PSNR，SSIM 和 MS-SSIM 更贴近于主观质量评价结果。而 MS-SSIM 相较于 SSIM 引入了多尺度的计算，被证明可以更好地度量失真[38]。

3.2 前沿研究的性能比较

将 Rippel（2017）[5]、Minnen（2018）[27]、Ballé（2018）[11]、Mentzer（2018）[22]和 Chen 等[28]提出的 NLAIC 这几个前沿研究与图像压缩的国际标准 BPG（YCbCr 4∶4∶4）、JPEG2000 和 JPEG（4∶2∶0）分别进行客观质量评价和主观质量评价。

1）客观质量评价

通过对比在公开的 Kodak 数据集上各先进模型呈现的平均率失真性能来进行性能比较。图 5 显示了以 MS-SSIM 和 PSNR 作为客观质量评价指标时的性能，其中 BPG444 为 BPG（YCbCr 4∶4∶4），JPEG420 为 JPEG（4∶2∶0）。图 5（a）以 MS-SSIM 指标表示了像素级失真，而图 5（b）以 PSNR 指标表示了结构相似性，这里采用 $-10\lg(1-d)$ 的形式表示在 dB 单位下的原始 MS-SSIM（d），d 表示原始 MS-SSIM（d）。从图中可看出，各先进研究中的模型性能都已超过了包括 BPG（YCbCr 4∶4∶4）在内的国际图像编码标准，并且各研究模型在低码率下的性能提升幅度较小，但在高码率下性能得到了很大的提升，相比较而言，近期 Chen 等提出的 NLAIC 模型表现出最好的性能。

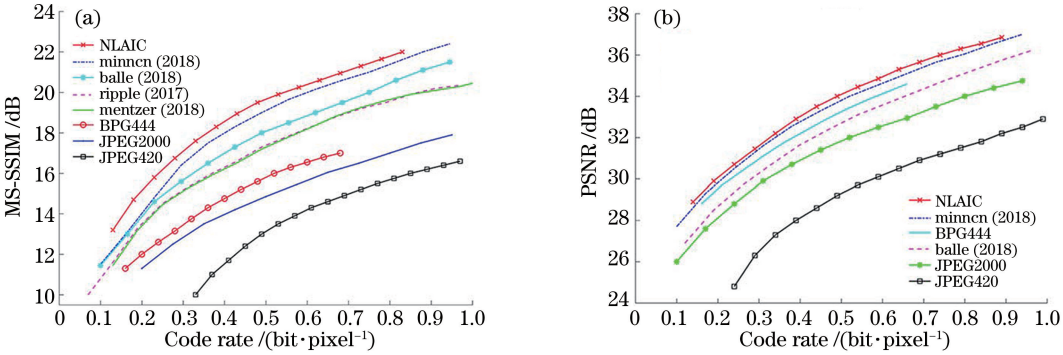

图 5 客观评价。（a）用 MS-SSIM（dB）测量像素级失真；（b）PSNR 用于评估结构相似度

Fig. 5 Objective evaluation. (a) Pixel-level distortion measured by MS-SSIM (dB); (b) PSNR used for structural similarity evaluation

2）主观质量评价

根据客观质量评价结果，Chen 等提出的模型相比以上提到的前沿模型有更好的结果，因此在 BDS500 数据集上进一步评估了该模型。图 6 显示了在相近码率下不同图像编解码器的重建结果，可看出在低码率的情况下图 6（a）中的 JPEG2000 在很多位置尤其是人脸部分出现块效应和模糊效应，而之后的国际编码标准解决了低码率时重建质量较低的问题，可看到图 6（b）中的 BPG 大大减弱了块效应，但仍存在一定的模糊效应，而NLAIC在相对

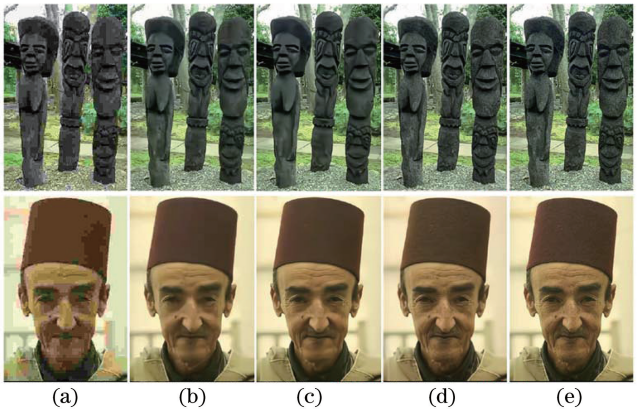

图 6 主观评价。（a）JPEG420；（b）BPG444；（c）NLAIC MSE opt；（d）NLAIC MS-SSIM opt；（e）原始图像

Fig. 6 Subjective evaluation. (a) JPEG420; (b) BPG444; (c) NLAIC MSE opt; (d) NLAIC MS-SSIM opt; (e) original image

较小的码率下呈现出更好的质量，尤其是图 6(d)的 NLAIC MS-SSIM opt 与图 6(e)的原始图像相比在无块效应和模糊效应的情况下保留了原始图像更多的纹理信息。

4 结 论

总结了基于端到端学习的图像编码框架中变换、量化、熵编码及损失函数的研究及进展。在变换方面，变换方法从先前的正交线性变换发展到非线性变换，在端到端学习中可等效为利用逐层卷积来提取特征；在量化方面，主流方法有随机近似和光滑导数近似的 round 方法、软到硬量化和格型量化等矢量量化方法；熵编码方法从最初的 JPEG 使用的 Huffman 编码发展到基于超先验和递归近邻概率混合预测的算术编码；在损失函数方面，最流行的是 L2 损失和 MS-SSIM 损失，同时交叉熵损失、感知损失、对抗损失和复合损失函数的出现使得优化性能不断提高。将现有的领先研究进行比较，在客观和主观质量评价中，以图的方式直观地呈现了领先研究中提到的模型与现有国际图像编码标准的性能差异。

基于端到端学习的图像编码的研究只有短短几年的历史，尽管已经取得卓越性能，但仍有较多环节需进一步改进和完善：1)在网络结构方面，应进一步探索新的神经网络模型在图像编码上的应用效果，如 InceptionNet、可变性卷积等；2)在码率分配方面，可进一步根据图像内容差异性、特征图的分布等，研究自适应的码率分配，以保证将码率更多地分配在对视觉质量影响大的信息上；3)在性能方面，目前的研究更注重压缩比和重建质量的提升，忽略了对计算复杂度的研究和优化，但在实际应用中复杂度是不可忽略的性能指标，因此，不仅需要研究深度学习方法复杂度的客观衡量标准，还应以此为指导，简化网络，实现更高效的计算；4)在硬件加速方面，应充分考虑硬件实现的需求，对网络进行改造，从而为将来的图像压缩芯片提供基础；5)在应用方面，随着智能化需求的增加，图像更多地被应用于计算机视觉，因此，可以研究图像压缩和计算机视觉之间的共通特性，探索更有利于计算机理解和识别的图像压缩方法。

参 考 文 献

[1] Kong F Q, Zhou Y B, Shen Q, et al. End-to-end multispectral image compression using convolutional neural network[J]. Chinese Journal of Lasers, 2019, 46(10): 1009001.
孔繁锵, 周永波, 沈秋, 等. 基于卷积神经网络的端到端多光谱图像压缩方法[J]. 中国激光, 2019, 46(10): 1009001.

[2] Wang H J, Jin T, Men K. Application of FA-LMBP hybird neural network algorithm in image compression[J]. Laser & Optoelectronics Progress, 2019, 56(19): 191005.
王海军, 金涛, 门克内木乐. FA-LMBP 混合神经网络算法在图像压缩中的应用[J]. 激光与光电子学进展, 2019, 56(19): 191005.

[3] Cheng D Q, Cai Y C, Chen L L, et al. Multi-scale convolutional neural network reconstruction algorithm based on edge correction[J]. Laser & Optoelectronics Progress, 2018, 55(9): 091003.
程德强, 蔡迎春, 陈亮亮, 等. 边缘修正的多尺度卷积神经网络重建算法[J]. 激光与光电子学进展, 2018, 55(9): 091003.

[4] Ballé J, Laparra V, Simoncelli E P. End-to-end optimized image compression[EB/OL]. (2016-11-05)[2019-12-18]. https://arxiv.org/abs/1611.01704.

[5] Rippel O Bourdev L. Real-time adaptive image compression[EB/OL]. (2017-05-16)[2019-12-18]. https://arxiv.org/abs/1705.05823.

[6] Jia C M, Zhao Z H, Wang S S, et al. Neural network based image and video coding technologies[J]. Telecommunications Science, 2019, 35(5): 32-42.
贾川民, 赵政辉, 王苫社, 等. 基于神经网络的图像视频编码[J]. 电信科学, 2019, 35(5): 32-42.

[7] Habibi A. Hybrid coding of pictorial data[J]. IEEE Transactions on Communications, 1974, 22(5): 614-624.

[8] Forchheimer R. Differential transform coding: a new hybrid coding scheme [C] // Picture Coding Symposium(PCS-81), [S. l.]:[s. n.], 1981: 15-16.

[9] He T Y. End-to-end image and video compression[D]. Hefei: University of Science and Technology of China, 2019.
何天宇. 端到端的图像视频压缩研究[D]. 合肥: 中国科学技术大学, 2019.

[10] Ballé J, Laparra V, Simoncelli E P. End-to-end optimization of nonlinear transform codes for perceptual quality [C] // 2016 Picture Coding Symposium (PCS), December 4-7, 2016,

Nuremberg, Germany. New York: IEEE Press, 2016: 1-5.

[11] Ballé J, Minnen D, Singh S, et al. Variational image compression with a scale hyperprior [EB/OL]. (2018-05-01) [2019-12-18]. https://arxiv.org/abs/1802.01436.

[12] Toderici G, O'Malley S M, Hwang S J, et al. Variable rate image compression with recurrent neural networks [EB/OL]. (2016-03-01) [2019-12-18]. https://arxiv.org/abs/1511.06085.

[13] Toderici G, Vincent D, Johnston N, et al. Full resolution image compression with recurrent neural networks [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5435-5443.

[14] Theis L, Shi W Z, Cunningham A, et al. Lossy image compression with compressive autoencoders [EB/OL]. (2017-03-01) [2019-12-18]. https://arxiv.org/abs/1703.00395.

[15] Zhou L, Sun Z H, Wu X J, et al. End-to-end optimized image compression with attention mechanism [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, [S. l.]: [s.n.], 2019.

[16] Agustsson E, Mentzer F, Tschannen M, et al. Soft-to-hard vector quantization for end-to-end learning compressible representations [EB/OL]. [2019-12-18]. https://www.cnblogs.com/lucifer1997/p/11203729.html.

[17] Alexandre D, Chang C P, Peng W H, et al. Learned image compression with soft bit-based rate-distortion optimization [C]//2019 IEEE International Conference on Image Processing (ICIP), September 22-25, 2019, Taipei, Taiwan, China. New York: IEEE Press, 2019: 1715-1719.

[18] Zhao L J, Bai H H, Wang A H, et al. Multiple description convolutional neural networks for image compression [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(8): 2494-2508.

[19] Bai H H, Zhu C, Zhao Y. Optimized multiple description lattice vector quantization for wavelet image coding [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2007, 17(7): 912-917.

[20] Huffman D. A method for the construction of minimum-redundancy codes [J]. Proceedings of the IRE, 1952, 40(9): 1098-1101.

[21] Rissanen J, Langdon G. Universal modeling and coding [J]. IEEE Transactions on Information Theory, 1981, 27(1): 12-23.

[22] Mentzer F, Agustsson E, Tschannen M, et al. Conditional probability models for deep image compression [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4394-4402.

[23] Huang C, Liu H J, Chen T, et al. Extreme image coding via multiscale autoencoders with generative adversarial optimization [C]// 2019 IEEE Visual Communications and Image Processing (VCIP), December 1-4, 2019, Sydney, Australia. New York: IEEE Press, 2019: 1-4.

[24] Liu H J, Chen T, Shen Q, et al. Deep image compression via end-to-end learning [EB/OL]. (2018-06-05) [2019-12-18]. https://arxiv.org/abs/1806.01496.

[25] Agustsson E, Tschannen M, Mentzer F, et al. Generative adversarial networks for extreme learned image compression [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27 - November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 221-231.

[26] Ma S W, Zhang X F, Jia C M, et al. Image and video compression with neural networks: a review [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(6): 1683-1698.

[27] Minnen D, Ballé J, Toderici G. Joint autoregressive and hierarchical priors for learned image compression [EB/OL]. (2018-09-08) [2019-12-18]. https://arxiv.org/abs/1809.02736.

[28] Chen T, Liu H J, Ma Z, et al. Neural image compression via non-local attention optimization and improved context modeling [EB/OL]. (2019-10-11) [2019-12-18]. https://arxiv.org/abs/1910.06244.

[29] Zhao L J, Bai H H, Wang A H, et al. Learning a virtual codec based on deep convolutional neural network to compress image [J]. Journal of Visual Communication and Image Representation, 2019, 63: 102589.

[30] Li M, Zuo W M, Gu S H, et al. Learning convolutional networks for content-weighted image compression [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York:

IEEE Press, 2018: 3214-3223.

[31] Shi W Z, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 1874-1883.

[32] Lee J, Cho S, Beack S K. Context-adaptive entropy model for end-to-end optimized image compression [EB/OL]. (2019-05-06) [2019-12-18]. https://arxiv.org/abs/1809.10452.

[33] Oktay D, Ballé J, Singh S, et al. Scalable model compression by entropy penalized reparameterization [EB/OL]. (2020-02-16) [2020-04-17]. https://arxiv.org/abs/1906.06624.

[34] Johnston N, Vincent D, Minnen D, et al. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4385-4393.

[35] Chen Y H, Yang T J, Emer J, et al. Eyeriss v2: a flexible accelerator for emerging deep neural networks on mobile devices[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2019, 9(2): 292-308.

[36] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[37] Wang Z, Simoncelli E P, Bovik A C. Multiscale structural similarity for image quality assessment[C]// The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, November 9-12, 2003, Pacific Grove, CA, USA. New York: IEEE Press, 2003: 1398-1402.

[38] Dosselmann R, Yang X D. A comprehensive assessment of the structural similarity index [J]. Signal, Image and Video Processing, 2011, 5(1): 81-91.