arXiv:2302.09560v1 [eess.IV] 19 Feb 2023

Deep Selector-JPEG：满足人类视觉标准的用于图像分类计算机视觉自适应 JPEG 图像压缩

Hossam Amer、*Sepideh Shaterian* 和 *IEEE Fellow En-hui Yang*

滑铁卢大学　　电气与计算机工程系，滑铁卢，安大略省，加拿大 N2L 3G1

摘要

由于存储和带宽资源有限，使用深度神经网络（DNN）的计算机视觉（CV）应用通常会采用针对人类视觉（HV）进行优化的 JPEG 对输入图像编码。本文提出了 Deep Selector-JPEG，这是一种面向图像分类并满足人类视觉标准的自适应 JPEG 压缩方法。对于每幅图像，Deep Selector-JPEG 会自适应地选择质量因子（QF）对图像进行压缩，从而在一组图像上针对各种深度神经网络分类器，在压缩比（CR）与深度神经网络分类器准确率（率-准确率性能）之间实现良好的权衡；同时，这类压缩图像的多尺度结构相似性高于由人类视觉预先确定的阈值的概率较高。Deep Selector-JPEG 通过轻量级或重量级选择器架构进行设计。实验结果表明，与相同压缩比下的 JPEG 相比，在满足人类视觉约束的同时，Deep Selector-JPEG 在 ImageNet 验证集上对所有测试的深度神经网络分类器均实现了更好的率-准确率性能；在相同压缩率下，分类准确率提升了 0.2% 至 1%。在更高压缩率下，Deep Selector-JPEG 还能够大致保持原始分类准确率。

索引词— 图像压缩、深度学习、JPEG、人类视觉、计算机视觉、图像分类

1. 引言

几乎每天，计算机视觉（CV）应用都需要存储或交换大量图像。由于能够在无需任何领域知识的情况下，从输入图像的原始像素中提取所需特征，深度学习（DL）已成为这些计算机视觉应用的关键技术 [1]。例如，在图像分类任务中，为了提取这些特征，深度神经网络（DNN）使用反向传播学习算法来学习非线性激活函数的参数。这些函数逐步转换输入图像的原始像素，从而生成输出预测标签 [1]。凭借这一能力，深度学习在图像分类领域展现出成功，其在大规模、高质量的 ImageNet 数据集上的准确率稳步提升，从 63.3% 增至 90.2% [1, 2]。

输入深度神经网络的大规模图像数据集的原始像素通常来自图像采集、编码、存储/传输和解码流程。这意味着这些原始像素确实经过有损压缩，以满足存储和带宽要求。

自 20 世纪 80 年代末以来，JPEG 编解码器一直广泛用于图像，通过称为质量因子（QF）的参数控制压缩率与人类感知质量之间的权衡。然而，JPEG 很少关注计算机视觉[3,4]。我们进行了一项实验，结果表明，如果使用 QF = 10 压缩 ImageNet 验证集中的所有图像，压缩比可达 11 倍，但深度神经网络分类器的分类准确率会下降 ≈8-10%。即使根据人类视觉（HV）判断，QF=10 下的图像感知质量可以接受，分类准确率下降 ≈8-10% 对计算机视觉而言仍可能过于显著，难以接受。同样，表 1 展示了一个来自 ImageNet 验证集的图像实例，该图像以不同质量因子输入 Inception V3（IV3）模型。该表表明，与原始图像相比，较低 QF 的 JPEG 压缩版本可能使真实标签（GT）的排名更高，同时保持合理的人类视觉质量 [5, 6]。因此，理想的做法是自适应地选择大规模图像数据集中每张输入图像的 QF，从而改善 JPEG 压缩比与深度神经网络分类准确率（排名准确率（RA）性能）之间的权衡，同时保持一定的人类感知质量。那么，问题当然是：该如何实现?

表 1：来自 ImageNet 并输入 Inception V3 的图像 ID#651。

	原始	质量因子=60	质量因子=40	质量因子=10
PSNR	Inf	35.2	33.5	29.9
多尺度结构相似性	1.0	0.98	0.97	0.95
真实标签排名	2	2	2	1

文献中有两类方法用于回答上述问题 [7–10]。第一类方法成功地采用与人类视觉和计算机视觉约束联合进行端到端训练的神经网络，或仅采用人类视觉约束进行训练 [8, 11–17]。后一类中的一些论文

本研究部分经费由加拿大自然科学与工程研究委员会通过资助项目 RGPIN203035-16 提供，并由加拿大研究讲席计划提供。

同时面向人类视觉和计算机视觉，联合重新训练用于计算机视觉任务的深度神经网络以及用于压缩的神经网络。用于计算机视觉任务的深度神经网络可能需要大量训练资源，也可能并非总是已知，这使得联合训练颇具挑战。此外，这些方法并未在多个用于图像分类任务的深度神经网络上进行评估。第二类方法重新设计了传统编解码器的量化表，以适应人类视觉和不同的计算机视觉任务 [9, 10, 18, 19]。然而，这些论文并非都能在满足人类视觉标准的同时，为图像分类自适应地选择每张输入图像的 JPEG 质量因子，并在不同深度神经网络下对整个 ImageNet 进行测试。

本文提出了 Deep Selector-JPEG，这是一种自适应 JPEG 压缩方法，利用深度神经网络分类器在满足人类视觉标准的同时，面向图像分类中的计算机视觉任务。对于训练集中的每幅图像，Deep Selector-JPEG 会标记可用于压缩图像的可行 QF 集合。可行 QF 集合根据以下两个标准确定：(1) 使用任意可行 QF 进行 JPEG 压缩后图像的多尺度结构相似性以高概率大于或等于多尺度结构相似性阈值，其中，该概率的计算方式是假设图像是从某个图像集（例如 ImageNet 验证集）中随机且均匀抽取的；(2) 使用给定 QF 压缩后的图像，其真实标签的排名不低于原始图像在 ImageNet 中真实标签的排名。对于每个给定的 QF，我们分别训练一个独立的二元深度神经网络分类器，以预测该给定 QF 是否可行。Deep Selector-JPEG 通过轻量级或重量级选择器架构进行设计。在推理时，Deep Selector-JPEG 会选择最小的可行 QF 来压缩原始图像。

实验结果表明，与在 10 个不同的图像分类测试深度神经网络上具有相同压缩率的 JPEG 相比，Deep Selector-JPEG 架构在整个 ImageNet 验证集上对所有测试的深度神经网络分类器均实现了更优的 RA性能；在满足人类视觉约束的同时，在相同压缩率下，分类准确率增益介于 $\approx$0.2% 和 $\approx$1% 之间。Deep Selector-JPEG 能够在更高压缩率下保持原始分类准确率

2. DEEP SELECTOR-JPEG：面向图像分类并符合人类视觉标准的自适应 JPEG 压缩

2.1. Deep Selector-JPEG：图像分类的构建与考量，并符合人类视觉标准

Deep Selector-JPEG 旨在实现图像分类，同时通过从一组可行的量化因子中为每张输入图像自适应地选择一个量化因子，以满足人类视觉标准。令$L = \{10, 20, 30, 40, 50, 60, 70, 80, 90\}$[1]。若某个量化因子满足以下两个约束，则它在$L$中是可行的：

1. 使用该量化因子压缩后的图像，其多尺度结构相似性以较高概率高于人类预先设定的目标多尺度结构相似性 [20]。2. 使用该量化因子压缩后的图像的真实标签排名，与原始图像在 ImageNet 集合中的排名相同或更高。

基于上述约束，我们的训练集为 $T = \{(x_1, q_1) \cdots, (x_N, q_N)\}$，其中 X 是 ImageNet 训练集中的所有原始图像构成的集合；对于每个原始图像 x_i，q_i 是一个二进制向量 $q_i = (q_{i,1}, \cdots, q_{i,n})$，其中$q_{i,j} = 1$ 表示 QF_j 对于原始图像 x_i是可行的，其中 X 表示所有原始图像构成的集合。

为满足第一个可行性约束，我们线下为来自原始 ImageNet 图像的每个 $QF \in L$ 创建了一个压缩图像簇。对于给定的目标多尺度结构相似性，我们计算每个 QF 簇中多尺度结构相似性大于或等于目标多尺度结构相似性的图像所占百分比，即命中率。随后，我们移除 L 中命中率低于 90% 的 QF。对于第二个约束，真实标签$q_{i,j}$ 理想情况下应由人类确定。也就是说，给定原始图像 x_i 及其 JPEG 压缩图像（压缩率为QF_j），人类应判断该压缩图像是否仍会使人相信，真实标签的排名相对于其原始图像仍至少得以保持（即满足可行性）。由于图像集规模庞大，这项任务十分艰巨。为缓解这一困难，我们转而使用预训练深度神经网络分类器 S替代人工标注者。

如图1所示，Deep Selector-JPEG 通过“深度”监督学习，利用两种深度神经网络架构来确定每个假设 y_j：$X \rightarrow \{0, 1\}$。对于这两种形式，我们都将给定的、包含 n 个质量因子的问题分解为 n 个相互独立的二元分类问题。因此，对于每张原始图像 $x \in X$，每个假设 y_j 都会被训练为预测其对应的 $QF_j, j = 1, 2, \cdots, n$ 在 L 内是否可行。就架构而言，第一种形式会冻结预训练的图像分类深度神经网络 S'，但保留最后两个模块，并将冻结部分用作所有二元分类器共用的特征提取器。基于这些共用特征，每个由 S' 的最后两个模块组成的 h_j 都会在训练过程中被训练为独立地作出自己的二元分类决策，y_j。为了进一步提升排名准确率性能，在第二种形式中，会针对每个 y_j 独立训练整个 S'。

对于每个二元分类器，使用以下二元交叉熵损失函数来获得优化后的权重W_j^*：

$$W_j^* = \arg\min_{W_j} -\frac{1}{N}\sum_{i=1}^{N} pr_j * q_{i,j}\log(p_j(x_i)) + (1 - q_{i,j})\log(1 - p_j(x_i)) \quad (1)$$

其中 pr_j 是精确率常数这一超参数，用于调整分类器召回率与精确率之间的权衡。pr_j 越低，精确率越高，召回率越低。与.....中一样

[1]这组质量因子值仅作为示例使用。然而，本文的思想可以应用于任意一组质量因子值。此外，在本示例中，质量因子 $= 10$被视为人类可接受的最低压缩质量。

图1， p_i（x_i）是 Sigmoid 函数（σ（。））的输出，有两种形式：

$$\text{Form One: } p_j(x_i) = \sigma(h_j(FE(x_i); W_j^T)) \quad (2)$$

$$\text{Form Two: } p_j(x_i) = \sigma(h_j(x_i; W_j^T)) \quad (3)$$

根据公式 (2) 或 (3)，对于每个图像 x_i， QF_j的预测 ON/OFF（记作 $y_j(x_i)$）为：

$$y_j(x_i) = \begin{cases} 1 & p_j(x_i) \geq DT_j \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

其中， DT_j 是另一个超参数，称为 QF_j 的决策阈值。较高的 DT_j 意味着较高的精确率和较低的召回率。推理时，Deep Selector-JPEG 最终会选择用于 JPEG 压缩的最低可行质量因子。

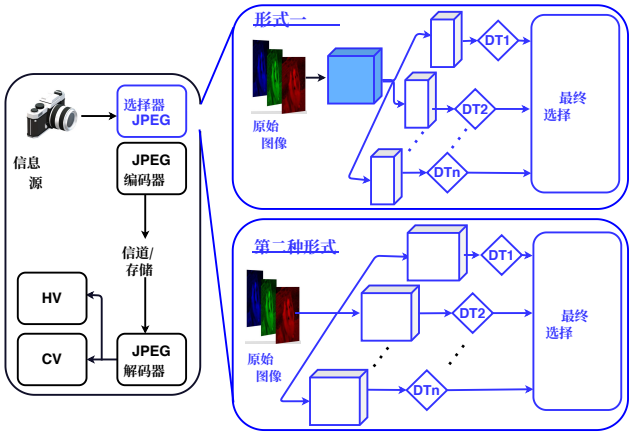


图 1：Deep Selector-JPEG 的构成与架构。

2.2. 训练描述

我们通过随机梯度下降在 TensorFlow 中训练了所提出的架构。我们使用两块 NVIDIA GeForce RTX 2080 Ti GPU 进行多 GPU 训练，批大小为 100，共训练两个 epoch，每个 epoch 包含 20000 个步骤。用于标注的预训练深度神经网络分类器 S 设置为 IV3 或 MobileNetV2，分别代表重量级和轻量级架构。精确率常数和决策阈值的取值范围分别为 0.2 至 0.7 和 0.5 至 0.9。对于 Deep Selector-JPEG 的第一种形式深度神经网络架构，我们生成了两个选择器，即 *IV3-TwoLayers* 和 *MobileNet-TwoLayers*，其底层深度神经网络 S' 分别为 IV3 和 MobileNet V2。在这些选择器中，我们冻结了除最后两个模块之外的所有层。对于 Deep Selector-JPEG 的第二种形式深度神经网络架构，将完整的 MobileNetV2 架构训练为 S'，用于构成二元分类*MobileNet-Full*。IV3 选择器使用来自 IV3 的标签进行训练，而 MobileNetV2 选择器则使用来自 MobileNetV2 的标签进行训练。

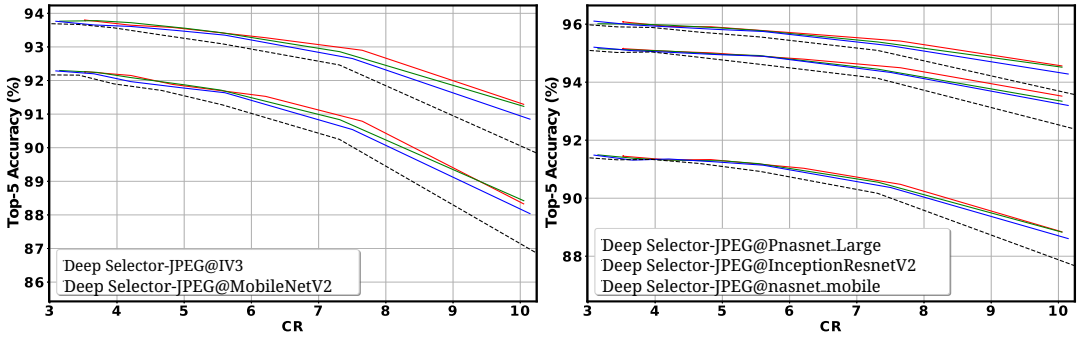
3. 实验结果

本节在整个 ImageNet 验证集上评估相同压缩率下 Deep Selector-JPEG 相对于 JPEG 的 RA性能。JPEG 的 RA曲线通过绘制 9 个压缩数据簇上的压缩率和测试深度神经网络的分类准确率得到；每个压缩数据簇都是将 ImageNet 验证集中的所有图像分别以 {10,20,..90} 的质量因子进行压缩而生成的。压缩率数值均以 Imagenet 验证集中的高质量原始图像为基准。将整个 ImageNet 验证集作为实验输入刺激后，我们在第 3.1 节中考察了人类视觉，并通过开展 4 组子实验，确保输出的压缩数据集满足以下四种约束之一：多尺度结构相似性 $\geq$ 0.8、0.85、0.9 或 0.95，且命中率为 90%。这四种多尺度结构相似性约束分别确保每幅输出压缩图像的峰值信噪比大于 26dB、28dB、30dB 和 32dB。对于这些多尺度结构相似性约束， L 中质量因子的起始值分别设置为 10、20、40 和 60。第 3.1 节展示了 Deep Selector-JPEG 在不同深度神经网络上的 RA性能一致性，并将我们的结果与 DeepN-JPEG [9]进行了比较。在第 3.2 节中，我们讨论 Selector-JPEG 的复杂度。

3.1. HumanVision 准则下的压缩率-准确率结果、一致性及与其他方法的比较

通过将 IV3-TwoLayers、MobileNet-Full 或 MobileNet-TwoLayers 选择器在不同压缩率下选择的质量因子应用于 ImageNet 数据集上 10 个分类性能各异的热门深度神经网络分类器，评估了 Deep Selector-JPEG 的排名准确率性能。图 2 展示了在多尺度结构相似性为 $\geq$ 0.8（峰值信噪比为 $\geq$ 26 dB）且命中率为 90% 时，相对于 JPEG 的准确率提升。在压缩率为 =3.5x 时，我们观察到相较于 JPEG，准确率略有提升，约为 0.2%。这一增益逐渐增加，在压缩率达到 =10.05x 时约为 1%。值得注意的是，在压缩率为 =10.05x 时，将 IV3-TwoLayers 应用于表现出色的深度神经网络分类器 Pnasnet Large，可使 Top-5 准确率提升 0.9%（基线的 Top-5 准确率为 =93%，压缩率为 =10.05x）。此外，该图表明，在多尺度结构相似性为 $\geq$ 0.8 时，面对不同的测试深度神经网络，排名准确率性能均能稳定提升。由于 MobileNet-Full 的容量较大，其准确率提升略高于 MobileNet-TwoLayers。我们还注意到，将通过深度神经网络 S 的标注进行训练的选择器应用于同一深度神经网络分类器时，Deep Selector-JPEG 的性能增益略高于应用于其他深度神经网络时的增益。在这些实验中，我们还观察到，在较高压缩率下，仍可保持原始分类准确率。

Deep Selector-JPEG 在 MS-SSIM $\geq$ 0.85（峰值信噪比 $\geq$ 28dB）和压缩率 =6.9x 为 0.3% 时，Top-5 准确率也持续提升。此外，在压缩率 =4.4x 时，



（a）从上到下：IV3、MobileNetV2　（b）从上到下：Pnasnet Large、InceptionResNetV2、Nasnet

（c）从上到下：IV4、ResNetV2-101、MobileNetV1　（d）从上到下：ResNetV2-50、IV1

图 2：将 Deep Selector-JPEG 应用于所有测试深度神经网络时的压缩率与 Top-5 准确率；MS-SSIM $\geq$ 0.8（峰值信噪比 $\geq$ 26dB）。黑色点线表示 JPEG，红色、绿色和蓝色曲线分别表示 IV3-TwoLayers、MobileNet-Full 和 MobileNet-TwoLayers。每个子图中的图例列出了从上到下排列的测试深度神经网络顺序。例如，图 2(d) 中顶部的一组率失真曲线对应 ResNet-50 V2，底部的一组率失真曲线对应 IV1。

准确率最高可提升 0.1%。这是因为在这种情况下，我们的质量因子范围选项有限；MS-SSIM $\geq$ 0.95（峰值信噪比 $\geq$ 32dB）时也存在同样的问题。未来，我们希望能从更大的范围内自适应地选择质量因子，为 Deep Selector-JPEG 提供更多选择，从而可能进一步提升准确率。

相较于 DeepN-JPEG，Deep Selector-JPEG 在压缩率为=3.5x 时，能够使 Top-5 准确率接近 DNN 分类器在原始图像上的准确率，这与 DeepN-JPEG 的结果相似（参见图 2 中的准确率与所测试的 10 个 DNN 的原始准确率）。这一相似结果仅通过自适应地选择单个参数 QF 实现。我们预计，如果像 DeepN-JPEG 一样，为每幅图像自适应地选择完整的量化表，还能获得更大的增益，这将留待未来研究。值得注意的是，Deep Selector-JPEG 通过为每幅输入图像自适应地选择 QF，使压缩率和 MS-SSIM 约束得到满足，从而改善了 JPEG 在不同 DNN 上的整体率失真权衡。

3.2. 时间与复杂度分析

Deep Selector-JPEG 的复杂度通过乘加运算（MACs）和总实际运行时间来评估，前者是常用的深度神经网络复杂度指标。总实际运行时间包括选择器推理时间和 JPEG 编码时间。在进行 9 种 QF 选择时，IV3-TwoLayers 的乘加次数为 8819.2M，MobileNet-Full 为 4869.9M。MobileNet-TwoLayers 的乘加次数为 541.5M。将整个验证集压缩至压缩率 =10.05x 和多尺度结构相似性 $\geq$ 0.8（峰值信噪比 $\geq$ 26dB）时，MobileNet-TwoLayers、MobileNet-Full 和 IV3-TwoLayers 仅使 JPEG 的耗时分别增加 2%、4.3% 和 3.74%。其他压缩率和人类视觉约束下也呈现相似的耗时结果。MobileNet-TwoLayers 选择器的复杂度相对较低，且其排名准确率表现与其他提出的选择器相当，因此能够部署在边缘设备上。

4. 结论

本文提出了 Deep Selector-JPEG，这是一种在满足人类视觉标准的同时服务于图像分类的自适应 JPEG 压缩方法。Deep Selector-JPEG 基于轻量或重量级深度神经网络架构设计。对于每张原始图像，Deep Selector-JPEG 会自适应地选择质量因子来压缩图像；与相同压缩率下的 JPEG 相比，在压缩比介于 0.2% 和 ≈1% 之间时，该方法能够提升分类准确率。此外，Deep Selector-JPEG 还能在更高压缩率下维持原始分类准确率。Deep Selector-JPEG 的自适应选择理念可迁移至其他能够生成不同质量级别图像的编解码器，例如 HEVC [21], VVC [22], 和基于深度学习的编解码器。因此，本文的另一项拓展是将这一选择理念应用于其他编解码器和计算机视觉任务。

5. 参考文献

[1] Y. LeCun、Y. Bengio 和 G. Hinton，《深度学习》，《自然》，第 521 卷，第 436 页，2015 年。

[2] H. Pham、Q. Xie、Z. Dai 和 Q. V. Le，《元伪标签》，*arXiv* 预印本 *arXiv:2003.10580*，2020 年。

[3] G. K. Wallace，《JPEG 静态图像压缩标准》，*IEEE* 消费电子学汇刊，第 38 卷，第 1 期，第 xviii–xxxiv 页，1992 年 2 月。

[4] E.-h. Yang 和 L. Wang，《运行长度编码、霍夫曼编码和量化表的联合优化，并完全兼容基线 JPEG 解码器》，*IEEE* 图像处理汇刊，2008 年。

[5] H. Amer 和 E.-h. Yang，《通过时间传播长度估计，为低延迟 HEVC 自适应选择量化参数》，《信号处理：图像通信》，第 115826 页，2020 年。

[6] E.-H. Yang、H. Amer 和 Y. Jiang，《压缩有助于深度学习中的图像分类》，《熵》，第 23 卷，第 7 期，第 881 页，2021 年。

[7] Q. Shen、J. Cai、L. Liu、H. Liu、T. Chen、L. Ye 和 Z. Ma，“Codedvision：通过端到端学习实现图像理解与压缩的联合处理”，载于*2018* 多媒体信息处理进展。

[8] Y. Hu、S. Yang、W. Yang、L.-Y. Duan 和 J. Liu，“面向人类与机器视觉的编码：一种可扩展图像编码方法”，载于*2020 IEEE* 多媒体与博览会国际会议 *(ICME)*。IEEE，2020 年，第 1–6 页。

[9] Z. Liu、T. Liu、W. Wen、L. Jiang、J. Xu、Y. Wang 和 G. Quan，“DeepN-JPEG：一种适用于基于 JPEG 的图像压缩的深度神经网络框架”，载于第 *55* 届年度设计自动化会议论文集。ACM，2018 年，第 18 页。

[10]X. Xie 和 K.-H. Kim，《面向物联网边缘计算机视觉、具有有界深度神经网络感知损失的源压缩》，发表于第 *25* 届移动计算与网络年度国际会议，2019 年，第 1–16 页。

[11] J. Ball´e、D. Minnen、S. Singh、S. J. Hwang 和 N. Johnston，《采用尺度超先验的变分图像压缩》，*arXiv* 预印本 *arXiv:1802.01436*，2018 年。

[12] D. Minnen、J. Ball´e 和 G. Toderici，《用于学习式图像压缩的联合自回归先验与分层先验》，*arXiv* 预印本 *arXiv:1809.02736*，2018 年。

[13] L.Theis, W. Shi, A. Cunningham, and F. Husz´ar, “使用压缩自编码器进行有损图像压缩”，*arXiv* 预印本 *arXiv:1703.00395*，2017。

[14] G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, and M. Covell，“使用循环神经网络进行全分辨率图像压缩”，收录于*IEEE*计算机视觉与模式识别会议论文集，2017年，第5306–5314页。

[15] D. Minnen, G. Toderici, M. Covell, T. Chinen, N. John-ston, J. Shor, S. J. Hwang, D. Vincent, and S. Singh，“使用分块深度网络进行空间自适应图像压缩”，收录于*2017*年*IEEE*国际图像处理会议（*ICIP*）。IEEE，2017年，第2796–2800页。

[16]N. Johnston、D. Vincent、D. Minnen、M. Covell、S. Singh、T. Chinen、S. J. Hwang、J. Shor 和 G. Toderici，《通过预热和空间自适应比特率改进循环网络的有损图像压缩》，发表于*IEEE* 计算机视觉与模式识别会议论文集，2018 年，第 4385–4393 页。

[17] L. D. Chamain, F. Racap´e, J. B´egaint, A. Pushparaja 和 S. Feltman，《面向机器的端到端优化图像压缩：一项研究》，载于*2021*年数据压缩会议（*DCC*）。IEEE，2021年，第163–172页。

[18] B. Brummer 和 C.De Vleeschouwer，《针对任务和内容调整 JPEG XS 的增益与优先级》，载于*IEEE/CVF*计算机视觉与模式识别会议研讨会论文集，2020年，第164–165页。

[19]N. Ghamsarian、H. Amirpourazarian、C. Timmerer、M. Taschwer 和 K. Sch¨offmann，《使用卷积神经网络对白内障手术视频进行基于相关性的压缩》，载于第*28*届*ACM*国际多媒体会议论文集，2020年，第3577–3585页。

[20] V. Kostina 和 S. Verd´u，《有限码长区间内的定长有损压缩》，*IEEE* 信息论汇刊，第 58 卷，第 6 期，第 3309–3338 页，2012 年。

[21] V. Sze、M. Budagavi 和 G. J. Sullivan，《高效视频编码（*HEVC*）》。Springer International Publishing，2014 年。

[22] A. Wieckowski、J. Ma、H. Schwarz、D. Marpe 和 T. Wiegand，《面向即将推出的多功能视频编码（ VVC ）标准的快速划分决策策略》，载于*2019 IEEE* 图像处理国际会议（*ICIP*）。IEEE，2019 年，第 4130–4134 页。