南京大学

本 科 毕 业 论 文

院　　系　电子科学与工程学院

专　　业　微电子科学与工程

题　　目　多尺度特征表达的图像压缩网络的设计与优化

年　　级　2017 级　学　　号　171830578

学生姓名　张凌宇

指导老师　沈秋　职　　称　副研究员

提交日期　2021 年 5 月 21 日

南京大学本科毕业论文（设计）
诚信承诺书

《多尺度特征融合的图像压缩网络设计与优化》

本人郑重承诺：所呈交的毕业论文（设计）（题目：）是在指导教师的指导下严格按照学校和院系有关规定由本人独立完成的。本毕业论文（设计）中引用他人观点及参考资源的内容均已标注引用，如出现侵犯他人知识产权的行为，由本人承担相应法律责任。本人承诺不存在抄袭、伪造、篡改、代写、买卖毕业论文（设计）等违纪行为。

作者签名：张浩宇

学号：171850578

日期：2021.6.8

南京大学本科毕业论文（设计）开题报告

填表人签名　张凌宇　填表时间：2021 年 3 月 2 日

学生姓名	张凌宇	学号	171830578
院系专业	微电子科学与工程	手机号	18851131567
指导教师姓名 (可按实际指导情况填写两位)	沈秋	职称	副研究员
导师所在单位	☑校内___电子学院　□校外_____　____		
毕设类型	□毕业论文　☑毕业设计（含毕业作品）		
论文题目	多尺度特征表达的图像压缩网络的设计与优化		

一、**研究背景及意义**（附参考文献，不少于 600 字）

在当今的信息时代，人们每天都会产生天文数字级别的信息量。这些信息数字化后以数据的形 式存储和传输。人们可以感知的数据形式包括图像、音频、视频等。为了能够在有限的磁盘空间存 储、在有限的带宽内传输这些数据，需要我们找到都空间角度更“高效的”数据表示形式，即在保持 感知体验的同时，用更少的比特数来存储它们。压缩就是寻找更高效表现形式的直接手段。

压缩分为可完全复原的无损压缩和不可逆的有损压缩。无损压缩虽然在编码和解码的过程中不损失任何信息，能够完全恢复原始数据，其可压缩的程度非常有限。而有损压缩通过牺牲一定程度(甚至很多时候对人眼来难以分辨)的视觉体验，可以换取很大程度的空间节省，有着更大的压缩潜力和更广泛的应用场景，是领域内主要关注的方向。

我们可以通过以下方法，从模型角度来描述有损压缩[1]:图像在压缩和解压过程中存在于两个 不同的域:数据域和压缩域。在数据域中，数据以未压缩的原始状态存储，比如图片的 RGB 三通道 表示。压缩的古过程中，数据域的数据通过一系列的变换，提取了原始图像的某些或者全部的重要 特征，用更精简的方式存储在了压缩域中，比如通过离散余弦变换获得的图像中不同频率信号的参 数。解压的过程则是仅利用压缩域的数据，通过一系列变换去尽可能还原出原始的数据。

一个压缩算法主要有两个目标，也对应着它的两类评价指标。一方面，我们希望图像的还原度 高。具体则是希望经过压缩和解压后重构出的图像在观感上和原始图像差别足够小，即失真率小。 对失真率的描述常用的指标包括峰值信噪比(PSNR)、结构相似性(SSIM)、和多尺度结构相似性等(MS-SSIM)。另一方面，我们希望压缩域的数据的比特数相比原图像

足够小，即压缩比高。失真率 与压缩存在一个权衡，二者难以同时达到极限，这是压缩问题的主要矛盾点，也是该领域的主要研 究方向。

学习获得的压缩算法与传统压缩算法相比，在适应各种特殊应用(如医学图像、双目图像等) 中有着更大的灵活性[27]。基于深度学习的压缩算法的另一优势是有潜力获得更具可读性的压缩域数 据。大数据时代，大量的图像数据都由机器只能处理。便于直接在压缩域进行机器视觉任务的压缩 算法有着广泛的应用场景。编码后，无需解码重构而直接在压缩域进行搜索、分类、识别、分割等 任务，可以极大节省计算资源。这不仅要求可观的压缩比、一定限度的失真率，同时要求各中视觉 任务在压缩域的数据上有着较高的精度。原理上这是可行的，因为用于推断的神经网络结构常常与 用于压缩的网络结构相似，编码器理论上具备提取出与任务相关特征的能力[27]。

综上所述，寻找一个失真率-压缩比总体令人满意，同时适应压缩域也具备良好可读性的算法有 很大的应用前景，也对图像存储、传输技术有着重要影响。

二、**国内外研究现状**（文献综述，附参考文献，不少于 800 字）

目前日常生活里都用的图像压缩算法之一是 JPEG。它的主要方法是首先把 RBG 空间的图像变换 到 YCbCr 空间，将对人更敏感的亮度和色差信息分开。然后将图像分割 8x8 像素块，进行离散余弦变 换，分解成直流信号和不同频率交流信号。再根据头文件中的量化表将这些信号的分量大小进行量化，对低频的信号保留度高，对高频信号进行一定程度的惩罚，最终取整到整数。按 z 字形顺序排列 量化的数据，将参数由低频列到高频，通过哈夫曼熵编码完成压缩。其他常见的传统图像压缩算法 包括基于小波变换的 JPEG2000、带有预测机制的 Webp 以及视频编码衍生出的 BPG 等。

压缩的核心问题是找到空间角度最高效特征提取方法，也就是找到一个数据域到压缩域的最优 变换。可以证明 KL 变换是最优的线性变换，而离散余弦变换是 KL 变换的易于计算近似。然而这些方 法都局限于线性变换，实验证明如果用拟合能力更强大的非线性变换可以达到更好的效果[2][14]。神 经网络是一种拟合能力强大的非线性变换方法。将基于神经网络的编码器和解码器引入图像压缩问 题为该领域开创了一片新大陆。

一个最早提出的神经网络用于压缩的模型是自编码器。将一个大尺寸输入图像，通过宽度节点数 逐渐减小的神经网络，最后只通过少数节点输出(称为瓶颈层)，作为编码器。将这个输出作为输 入连到一个节点数逐渐变大的神经网络，作为解码器。将得到的输出与原始输入用某个损失函数 (如均方误差)计算距离，然后用梯度下降等方法训练。训练好参数使得

损失函数降得很低后，可 以发现原来的大尺寸的输入可以用少数节点的瓶颈表示。实际应用时，将自编码器的前半部分和后 半部分拆开来，分别交给编码和解码方即可。该方法因存在许多限制条件而无法直接代替标准压缩 算法，例如对于每个目标压缩比，需要不同的瓶颈层节点数，要专门训练多个网络;神经网络框架 固定后，对输入尺寸有特定要求等。并且原理上自编码器是一种降维的方法，而降维并不一定对应 着最优的数据压缩。尽管如此，对自编码器的框架进行改进被证明可以得到性能非常良好的压缩算 法。目前基于深度学习的图像压缩算法主要针对编码器解码器网络结构、可微量化函数、熵预测模 型等部分进行改进优化;同时也有很多引入自编码器框架之外的机制也有很大的进展，如对图像内 容自适应的编码、超先验机制、基于 GAN 的生成模型等。

基于神经网络的图像压缩方法通常由以下几个部分组成:编码器网络、量化函数、熵预测模型、 解码器网络。对编码器网络结构进行改进的工作中，Toderici 等[3][4]首先提出结合循环神经网络的自 编码器，通过渐进式的残差构建，得到了压缩比可调的压缩算法，在低比特率段重建效果好于 JPEG。随后对比联合 LSTM 和 GRU 与残差网络的结合，探索了不同残差构建方法进一步提高了性 能，获得了全比特率段优于 JPEG 的表现。Johnston 等[8]在该结构基础上提出预启动机制提高了编码 器的表达效率，引入了空间自适应压缩比机制，使得压缩比可以随图像局部的复杂程度不同而进行 动态调整，并使用了 SSIM 加权的损失函数，获得了更好的压缩表现。Theis 等[10]提出基于卷积神经 网络的自编码器，并采用基于取整的量化方式，得到了可适应任何尺寸的，压缩能力与 JPEG2000 相 近的方法。Rippel 等[11]提出了基于金字塔分析的自编码器模块、一个自适应编码模块、预期编码长 度正则化，并增加了多尺度对抗训练来补充模型，性能超过了此前所有基于 CNN 的模型，并相对轻 量级，可在 GPU 上实时解码。Wen 等[34]介绍了一种基于多尺度卷积的自编码器压缩方法，采用 InceptionNet 结合金字塔尺寸调整的卷积网络架构作为编码器，在解码器引入残差网络提升其非线 性，并设独立熵模型来预测特征分布。该方法借助 InceptionNet 的多尺度特征提取能力实现压缩。

在量化函数的改进工作中，Balle´ 等[1]从概率模型角度描述压缩问题，用归一化的噪声替代取整 量化，解决了反向传播时量化器不可微的问题，并选择联合优化失真率和压缩比，实现了端到端的 模型优化。Agustsson等[12]通过向量量化取代标量量化，反向传播时对soft松弛的函数微分，实现 了量化赋值由软到硬的控制。

熵预测模型相关工作中，Mentzer等[6]提出3D-CNN，通过背景信息context模型对压缩域

特征分 布的熵进行建模，卷积自编码器利用context模型对特征的熵进行预测，context模型则不断更新学会 编码符号和压缩域特征的关系。Minnen[17]引入了随图像适应的多项分布字典，提供随不同图像自适 应调整的side信息，辅助图像块熵的预测。基于超先验机制的方法中，Balle´ 等[18]在自编码器的基础上增加了一个超先验模块，产生的信号作为熵模型的先验知识向编码器和解码器传输辅助信息，调 整熵模型的信号，降低了它与熵模型的不匹配程度。Minnen[19]在此基础上把单高斯模型推广至高斯 混合模型，同时生成一个以超先验为条件的均值和尺度参数，并结合自回归模型可以在不增加比特 数要求，只是用已被解码的特征的情况下更准确地对熵模型进行建模。

根据图像内容进行自适应的编码方法中，Li 等[9]首先提出图像不同区域应采用不同的压缩比，根 据内容加权的重要性图在指引与内容相关的比特分配。Baig[25]等指出假设每个分割出的待编码图块 之间是互相独立的是不合理的，提出一种发掘邻域相关性的模块，通过邻域信息学习预测图像内 容，节省下所需要存储信息所需的比特数。Minnen 等[13]提出了结合了神经网络和图像质量敏感的比 特率适应机制，通过的块状结构进行 context 预测，保持了分辨率的灵活性和局域信息共享，同时大 大简化了毕业率适应的实现。Lee 等[21]指出不仅特征的标准差可以从图像中邻域预测，其均值也可 以，并且同时预测特征的标准差和均值后的压缩方法比仅预测标准差更有效。

在基于 GAN 的重构图像生成方法中，Santurkar 等[16]把解码器描述为生成函数，提出使用 GAN 学习在缩略图上的生成模型，再利用该模型作为解码器进行缩略图压缩。Agustsson[15]等采用了部 分原信息保留，纹理通过语义信息直接利用 GAN 生成，达到了 50%以上的压缩比。也有考虑对失真 图像进行后处理的，如 Galteri[7]等用 GAN 训练了一个卷积残差网络，作为一个模拟人眼的评判器， 无需一个直接描述图像质量的具体损失函数，生成去除人工失真点的图像。Mentzer[23]认为各种观 感评价指标各有缺点，考虑把专门设计失真损失，而是基于 GAN 来最小化原图和重构图像分布的差 异，并在各种失真损失函数上进行优化后验证了压缩比-失真率-观感权衡的存在。

在观感评价指标算法的问题上，Chinen[20]等采集了大量未被训练的人眼评价数据，训练了基于 VGG-16 的 10 参数网络来拟合数据，起到模拟人眼观感的作用。Blau 等[24]则从数学上证明不仅失真 率和压缩比存在不可兼得的权衡，观感和失真率在推向极端的条件下也是相互矛盾的。

压缩域数据分析相关工作中，Duan 等[28][29]总结了基于紧凑描述子的，应用于推断和

搜索的视 频压缩标准。Torfason 等[27]结合了用于推断的 ResNet、DeepLab 和 Theis 等[10]提出的用于压缩的自 编码器框架，针对压缩域图像的分类和语义分割问题进行训练。结果证明压缩域推断精度与数据域 相当，并尽需要更少的操作数即可完成，对图像压缩和分类联合优化时，二者表现同时提高。

参考文献：

[1] J .Balle´ and E. P. Simoncelli. End to end optimized image compression. In ICLR, 2017.

[2] J. Balle´ , V. Laparra and E. Simoncelli. Density modeling of images using a generalized normalization transformation. In ICLR, 2016.

[3] G. Toderici, S. Malley, S. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell and R. Sukthankar. Variable rate image compression with recurrent neural networks. In ICLR, 2016.

[4] G. Toderici Damien, V. Johnston, S.Hwang, D. Minnen, J. Shor and M. Covell. Full resolution image compression with recurrent neural networks. In CVPR, 2017.

[5] K. Gregor, F. Besse, Danilo J. Rezende I. Danihelka and D. Wierstra. Towards conceptual compression. arXiv preprint arXiv:1604.08772v1, 2016.

[6] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte and L. V. Gool. Conditional probability models for deep image compression. In CVPR, 2018.

[7] L. Galteri, L.Seidenari, M. Bertini and A. D. Bimbo. Deep generative adversarial compression artifact removal. In ICCV, 2017.

[8] 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. In CVPR, 2018.

[9] M. Li, S. Gu, W. Zuo, D. Zhao and D. Zhang. Learning convolutional networks for content-weighted image compression. In CVPR, 2018.

[10] L. Theis, W. Shi, A. Cunningham and F. Huszar. Lossy image compression with compressive autoencoders. arXiv preprint arXiv:1703.00395v1, 2017.

[11] O. Rippel and L. Bourdev. Real-time adaptive image compression. arXiv preprint arXiv:1705.05823v1, 2017.

[12] E. Agustsson, F. Mentzer, M. Tschannen, L. Benini, L.Cavigelli, R. Timofte, and L. V. Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In CVPR,

2018.

[13] D. Minnen, G. Toderici, M. Covell, T. Chinen, N. Johnston, J. Shor, S.J. Hwang, D. Vincent and S. Singh. Spatially adaptive image compression using a tiled deep network. In ICIP, 2017.

[14] J. Balle´ . Efficient nonlinear transforms for lossy image compression. arXiv preprint arXiv:1802.00847v2, 2018.

[15] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte and L. V. Gool. Generative adversarial networks for extreme learned image compression. In ICCV, 2019.

[16] S. Santurkar, D. Budden, and N. Shavit. Generative compression. In PCS, 2018.

[17] D. Minnen, G. Toderici, S. Singh, S. J. Hwang and M. Covell. Image-dependent local entropy models for learned image compression. arXiv preprint arXiv:1805.12295v1, 2018.

[18] J. Balle´ , S. J. Hwang, D. Minnen and N. Johnston. Variational image compression with a scaled hyperprior. In ICLR, 2018.

[19] D. Minnen, J. Balle´ and G. Toderici. Joint autoregressive and hierarchical priors for learned image compression. In NIPS, 2018.

[20] T. Chinen, J. Balle´ , C. Gu, S. J. Hwang, S. Ioffe, N. Johnston, T. Leung, D. Minnen, S. O' Malley, C. Rosenberg and G. Toderici. Towards a semantic perceptual image metric. In ICIP, 2018.

[21] J. Lee, S.n Cho and S. Beack. Context-adaptive entropy model for end-to- end optimized image compression. In ICLR, 2019.

[22] S. Singh, S. Abu-El-Haija, N. Johnston, J. Balle´ , A. Shrivastava and G. Toderici. End-to-end learning of compressible features. arXiv preprint arXiv:2007.11797v1, 2020.

[23] F. Mentzer, G. Toderici, M. Tschannen and E. Agustsson. High-fidelity generative image compression. arXiv preprint arXiv:2006.09965v3, 2020.

[24] Y. Blau and T.r Michaeli. The perception-distortion tradeoff. arXiv preprint arXiv:1711.06077v4, 2020.

[25] M. H. Baig, V. Koltun and L. Torresan. Learning to inpaint for image compression. In CVPR, 2017.

[26] H. Liu, T. Chen, Q. Shen, Tao Yue, and Z. Ma. Deep image compression via end-to-end learning. In CVPR, 2018.

[27] R. Torfason, F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte and L. V. Gool. Towards

image understanding from deep compression without decoding. In ICLR, 2018.

[28] L. Duan, V. Chandrasekhar, S. Wang, Y. Lou, J. Lin, Y. Bai and T. Huang. Compact descriptors for video Analysis: the emerging MPEG standard. arXiv preprint arXiv:1704.08141v1, 2017.

[29] L. Duan, J. Lin, J. Chen, T. Huang and W. Gao, "Compact Descriptors for Visual Search," in IEEE MultiMedia, vol. 21, no. 3, pp. 30-40, July-Sept. 2014, doi: 10.1109/MMUL.2013.66.

[30] H. Li, Y. Guo, Z.Wang, S. Xia and W. Zhu. AdaCompress: Adaptive Compression for Online Computer Vision Services. arXiv preprint arXiv:1909.08148v1, 2019.

[31] J. Zhong and C. Pun, "An End-to-End Dense-InceptionNet for Image Copy-Move Forgery Detection," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 2134-2146, 2020, doi: 10.1109/ TIFS.2019.2957693.

[32] P. Bhattacharya and U. Z.¨ lzer. Convolutional Neural Network with Inception Blocks for Image Compression Artifact Reduction. In IJCNN, 2020.

[33] L. Duan, J. Liu, W. Yang, T. Huang and W. Gao. Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent Analytics. arXiv preprint arXiv:2001.03569v1, 2020.

[34] S. Wen, J. Zhou, A. Nakagawa, K. Kazui, and Z. Tan. Variational autoencoder based image compression with pyramidal features and context entropy model. In CVPR, 2019.

三、主要研究或解决的问题和拟采用的方法

深度学习方法的引入，凭借神经网络强大的非线性拟合能力，让图像压缩算法性能得到了很大 的提升。目前主流的压缩算法主要关注压缩比-失真率的损失优化，在 PSNR 和 MS-SSIM 等和人眼相 关的图像质量指标上的表现。然而如今大量的图像都是通过机器智能处理，未经解压的的压缩域数 据直接处理更是重要的研究方向。

得到失真率压缩比优良，同时压缩域特征具备可读性的压缩算法，需要在编码时不仅考虑到压缩比和可复原度，还需要考虑到压缩形式的特征对于机器视觉算法是否具有可读性。本工作拟通过编码时的多尺度特征提取来在保持压缩性能的同时，得到对于机器具有可读性的特征。

拟采用的方法：

1.设计基于 inceptionnet 多尺度特征学习框架的编码器。编码器基于 InceptionNet 架构，通过不 同尺寸卷积，得到串联的不同尺度的具有机器可读性特征图。与自编码器压缩网络结合，学习到能 够表达视觉任务所需特征的压缩模型。

2.引入注意力机制，考虑机器视觉任务精度，实现自适应码率分配。

3.利用基于 GAN 的生成网络对高层语义信息的保留，提高压缩域数据在视觉任务上的精度。

四、工作进度计划（每两周为一个单位）

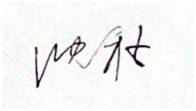
3.1 — 3.28 ： baseline 模型实现，改进算法设计

3.29 — 4.25：改进算法实现，进行实验、数据记录整理

4.26 — 5.10：完成报告撰写与答辩准备

指导教师意见（不少于 50 个字）

考虑计算机视觉需求的图像压缩算法是符合人工智能应用发展趋势的新方向，具有较大的探索空间和应用潜力。多尺度特征表达有望在该研究中取得良好效果。该选题在创新性、理论性和实用性方面都符合专业培养目标，题目难度和工作量适中。

签名： 

2021 年 3 月 4 日

院系意见：

审核通过

院系负责人签名：

年 月 日

南京大学本科毕业论文（设计）中期检查表

论文题目	中文：多尺度特征表达的图像压缩网络设计与优化
	外文： Design and Optimization of a Multi-scale Representation based Image Compression Network
学号：171830578	姓名：张凌宇
所在院系：电子科学与工程	专业：微电子科学与工程
指导教师：沈秋	职称：副研究员
计划完成时间：2021 年 5 月 15 日	

论文（设计）的进度计划：

4.15 - 4.19 完善 Baseline 工作的实验与调试

4.20 - 4.26 搭建多尺度特征编码和自回归模块

4.27 - 5.11 实验

5.1 - 准备答辩/撰写论文

已经完成的内容：

本设计目标为通过多尺度特征提取进行压缩与具有可读性的压缩算法设计。该算法实现上由以下部分组成：

一、多尺度卷积编码器：

基于 InceptionNet[2][3]的结构，输入图像通过不同尺寸的降采样卷积得到不同尺度的特征信息，将其串联作为自编码器的瓶颈层。由于不同尺度特征得到分离，这种瓶颈层有潜力在机器视觉任务上具有超越一般压缩算法压缩域数据的表现。

二、量化函数

降瓶颈层数据取整量化，以便熵编码为二进制比特流用于通信。训练过程中，为了解决取整函数不可微分性对反向传播造成的障碍，改为添加归一化噪声。

三、解码器：

与编码器对称，形成自编码器架构。

四、超先验熵模型：

基于 Balle2018[1]的结构，用一个超先验网络对编码器得到的特征分布的方差进行预测，作为指导性先验知识辅助熵编码。

五、自回归模块：

基于 Minnen2018[4]的结构，利用自回归模型对编码器得到的特征分布的均值进行预测，辅助熵编码。

目前已完成部分：

Baseline 压缩网络结构：

1.编码器：Balle16[1]和 Minnen18[4]中的一般卷积编码器

2.量化函数：训练时加入归一化噪声、测试时取整

3.解码器：Balle16[1]和 Minnen18[4]中的一般卷积解码器

4.超先验网络：Balle18[5]和中的超先验卷积自编码器

5.超先验熵模型：Balle18[5]中的 factorized 熵模型

实验内容：联合损失中取不同的 lamda 值调整失真、压缩比权衡，分别进行训练。通过每一个工作点，绘制失真-压缩比(RD)曲线。

实验结果：得到了接近 Balle2018 的 RD 曲线。

正在实验的部分：

1.Wen[3]中的金字塔型多尺度编码器

2.Minnen[4]中的自回归 Context 模型

待完成实验的部分：

1.Wen[3]中的残差解码器

2.Wen[3]中均值、方差独立预测的 Context 模型和熵参数模型

3.瓶颈层视觉任务实验

参考文献

[1] Johannes Balle´, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. arXiv preprint arXiv:1611.01704, 2016.

[2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich. Going Deeper with Convolutions. In CVPR, 2014.

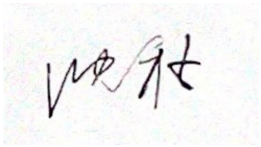
[3] Sihan Wen, Jing Zhou, Akira Nakagawa, Kimihiko Kazui, and Zhiming Tan. Variational autoencoder based image compression with pyramidal features and context entropy model. In CVPR, 2019

[4] David Minnen, Johannes Balle´, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In Advances in Neural Information Processing Systems, pages 10771–10780, 201

[5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, Nick Johnston. Variational image compression with a scale hyperprior. In ICLR, 2018.

指导教师意见：（对目前完成情况的指导性意见，不少于 100 字）

该生对课题展开了丰富的调研和学习，对理论框架和现有方法掌握熟练，初步完成了基础实验方法的复现和测试，并在此基础上提出可行的改进思路，后续工作计划安排合理，有望按期完成毕业设计，取得优异的成绩。建议在实验中加强对比实验和消融实验。



指导教师签字：

2021 年　4 月　23 日

南 京 大 学

本科生毕业论文（设计、作品）指导教师评阅意见

指导教师评语：（不少于 300 字）

考虑计算机视觉需求的图像压缩算法是符合人工智能应用发展趋势的新方向，具有较大的探索空间和应用潜力．多尺度特征表达包含更完善、更结构化的图像特征，被证明在计算机视觉任务上具有良好的性能．因此，探索多尺度特征在图像压缩算法上的应用效果，以及对计算机视觉任务的影响，具有创新意义和实用价值，同时也具有较高的挑战性．

该生通过广泛调研和算法复现与实验，熟练掌握了深度学习图像编码的理论与方法，并在此基础上实现了基于多尺度特征的图像压缩算法，取得了良好的应用效果．

该论文结构完整，内容充实，研究意义阐述清晰，国内外现状分析详实，创新方法切实可行，研究内容表述完整，实验设置合理，对照实验充足，结果分析中肯，未来工作的展望依据充分．该研究工作有望为人机协同中的图像压缩编码的研究工作奠定良好的基础．

指导教师签名：

2021 年 6 月 日

南 京 大 学

本科生毕业论文（设计、作品）评阅教师评阅意见

评阅教师评语：（不少于300字）

答辩人毕业设计《多尺度特征融合的图像压缩网络设计与优化》介绍了基于深度学习的一种图像压缩算法。该方法结合了多尺度特征，有较好的应用前景。论文选题具有实际意义，内容充分，说明作者较好地掌握了专业知识和进行实验的必要方法。作者阅读了较多文献，对领域发展熟悉，掌握了相关工作的原理。文中尚存的问题是研究点不够清晰明确，压缩网络和视觉应用的关系和篇幅有待更好的平衡，文章结构有改进空间。另外，需要注意实验的严谨性准确，以及图题字数和文献格式统一性。总之，该生按要求完成了毕业论文任务书所规定的内容，论文撰写基本符合规范，论述基本准确。答辩时可见答辩人对题目有较好的理解。论文中除对方法的描述外应更多体现自己的分析和思考。未能达到预期结果的实验和想法也应加以分析。希望未来工作引入更多创新性。

评阅教师签名：张丽红

年 月 日

南 京 大 学

本科生毕业论文（设计、作品）答辩记录、成绩评定

答辩记录：

评：你题目的重点是压缩网络设计？

答：是的

评：那论文应该放更多篇幅在压缩上。文章中对context model翻译得不准确。BSVR上Minnen的实验结果比BPG相差较多，可能存在问题

答：好的，我回去修改检查一下。

评：如果有不成功的实验也可以写。看得出来你对你的题目有一定理解，这些分析也应该写到论文里。另外格式上图题字号的时候应该写入正文，参考文献格式要更统一。

答：明白了，谢谢老师。

答辩记录人签名： 习政宇

答辩小组评语：（不少于100字）

答辩人毕业设计《基于变换域残差的图像压缩网络设计与优化》介绍了基于深度学习的一种图像压缩算法。论文选题具有实际意义，内容充分。答辩人对题目有较好的理解，希望论文中除对方法的描述外能更多体现自己的分析和思考，未来工作中引入更多创新性。

答辩小组成员： 王飞 周湘 张丽丽 朱昊

成绩 83.67

组长签名： 张丽丽

答辩时间： 2021 年 5 月 24 日

南京大学本科生毕业论文(设计、作品)中文摘要

题目：**多尺度特征表达的图像压缩网络的设计与优化**

院系：电子科学与工程学院
专业：微电子科学与工程
本科生姓名：张凌宇
指导老师（姓名、职称）：沈秋 副研究员

摘要：

基于深度学习的图像压缩编码方法近年来取得了一系列进展。主流的框架是通过神经网络学习一个从图像信号域到压缩域的非线性变换，该过程提取了重建图像所需的关键特征；再通过反变换将经过量化的特征解压重建出图像。然而当今的许多场景下，图像在被人接受前就会经过计算机的处理，如分类、目标检测、召回、语义分割等。而图像的解压过程常常是视觉任务中最占用计算资源的步骤之一，因此能在压缩域直接执行视觉任务的压缩算法也成为了重要的研究方向。图像压缩的编码和视觉任务的分析有着相似的特征提取过程。受此启发，为了提高特征提取的效率并得到适合视觉任务的特征，本设计提出将基于 Inception Block 的多尺度特征提取编码器应用于压缩域视觉任务。模型以自编码器为基本架构，编码器分四个支路的不同卷积层提取了不同尺度的特征信息，在小尺寸瓶颈层经过量化后传入解码器。量化的特征可在一个基于混合高斯模型的熵参数预测模型的指导下进行熵编码，得到二进制比特流。熵参数模型的信息来源于一个自回归背景信息预测模型和由一个超先验网络提供的边信息。特征的机器可读性通过以 ResNet18 为主干网络的分类器进行测试。图像算法的压缩性能和多尺度特征的分类精度分别在 Kodak 公开数据集和 Pascal VOC2012 上验证。结果显示，编码器在达到目前界内高水准压缩性能的同时，提取到的多尺度特征分类精度与重建 RGB 图像相当，且省去解码步骤后大大减小了计算复杂度。

关键词：深度学习；图像压缩；多尺度特征

南京大学本科生毕业论文(设计、作品)英文摘要

THESIS: Design and Optimization of a Multi-scale Representation based Image Compression Network

DEPARTMENT: School of Electronic Science and Engineering
SPECIALIZATION: Electronic Information Science and Technology
UNDERGRADUATE: Lingyu Zhang
MENTOR: Professor Qiu Shen

ABSTRACT:

Deep Learning based image compression methods have achieved series of improvements in recent years. The mainstream framework is to utilize a neural network to learn a non-linear transform from the signal domain of images to a latent domain, extracting essential features for image reconstruction. A reverse transform is then learned to decompress the quantized latent representations into the reconstructed image. In many scenarios, however, images are processed by computers before being received by humans, going through classification, object detection, retrieval, semantic segmentation and so on. Moreover, decompressing is often one of the most time consuming steps in computer vision tasks. Therefore, an image compression algorithm that has a latent domain where vision tasks are directly executable is an important field of research. Inspired by the fact that image compression and vision task analysis have similar feature extraction processes, this study proposes an image compression method, applying a multi-scale feature extraction encoder based on Inception blocks for vision task compatible latent generation. We designed a model based on an autoencoder framework. The encoder extracts features with different scales by different convolution layers though four branches. The bottleneck layer with smaller size is quantized and passed into the decoder. Quantized representations can be entropy coded into storable and transmittable bit streams under the instructions of an entropy parameter estimation model based on a Gaussian Mixture Model, whom receives data processed by an autoregressive context prediction model and side information provided by a hyperprior net. Understandability of the representations are evaluated through a classifier with ResNet18 as backbone. The performance of image compression and classification with compressed representation are validated on the Kodak open dataset and Pascal VOC2012 set, respectively. Results show that the compressor yields a state of art performance while extracting compressed latents that achieves classification accuracies comparable to that of reconstructed RGB images, greatly reducing computational time.

KEY WORDS: Deep Learning, Image Compression, Multi-scale Features

目　录

1　绪论 …… 1

1.1 研究背景 …… 1

1.2 研究问题 …… 2

1.3 研究现状 …… 3

1.4 本文工作 …… 4

2　相关工作 …… 5

2.1 引言 …… 5

2.2 传统图像压缩算法 …… 5

2.3 基于深度学习的图像压缩算法 …… 6

3　多尺度特征表达的图像压缩模型 …… 9

3.1 端到端优化的图像压缩框架 …… 9

3.2 多尺度变分自编码器 …… 11

3.2.1　变分自编码器 …… 11

3.2.2　卷积编解码器 …… 12

3.2.3　泛化归一化层 …… 13

3.2.4　Inception 模块 …… 14

3.2.5　多尺度编码器 …… 15

3.3 量化函数 …… 16

3.4 熵模型 …… 17

3.4.1　熵编码 …… 17

3.4.2　混合高斯模型 …… 17

3.4.3　超先验网络 …… 18

3.5 背景信息预测 …… 20

3.6 损失函数 …… 21

4　实验结果与分析 …… 22

4.1 压缩性能试验 …… 22

4.1.1　数据集与预处理 …… 22

4.1.2　模型参数 …… 22

4.1.3　评价标准 …… 23

4.1.4　训练细节 …… 24

4.1.5　结果分析 …… 24

4.2 特征可读性实验 26
4.2.1 图像分类问题 26
4.2.2 深度残差网络 27
4.2.3 实验设置 29
4.2.4 分类结果与分析 30
4.2.5 特征可视化分析 32
5 总结与展望 33
5.1 本文工作总结 33
5.2 未来工作展望 33
参考文献 **34**
致 谢 **39**

1 绪论

1.1 研究背景

当今信息时代，人们每天都会产生天文数字级别的信息量。这些信息数字化后以数据的形式存储和传输。数字多媒体技术中，人们可以感知的数据形式包括图像、音频、视频等。为了能够在有限的磁盘空间存储、在有限的带宽内快速传输这些数据，需要我们找到空间角度更“高效”的数据表示形式，即：在尽可能小地影响感知体验的同时，用更少的比特数来存储它们。压缩就是寻找更高效表示形式的直接手段。图像作为人们传递信息的主要载体之一，常常是对人最直观，而所需存储空间和带宽最大的一类数据。打开网页时，一般图像加载速度总慢于文字。因此对图像进行一定程度的压缩是信息交互、处理的必要步骤和关键技术，是信息时代高速发展的基础支撑。

过去数十年，图像压缩算法得到了很大发展。1992 年提出的 JPEG[2]（联合图像专家组）算法迄今为止仍是使用最广泛的图像压缩标准，可在视觉体验影响甚微的情况下达到十倍以上压缩比。JPEG2000、WebP、BPG 等算法更近一步提高了压缩性能。尽管如此，这些压缩算法的变换步骤都是人工设计的，在需要更高压缩比时会产生非自然图像块，影响观感。

近年来，在大数据和硬件算力提高的基数上，人们发现并验证了神经网络强大的表达能力，深度学习在计算机视觉、自然语言处理等方向得到了空前的发展和应用。在图像压缩问题中引入神经网络来学习非线性变换为该领域带来了新一轮发展。图像压缩的关键是提取图像特征，去除冗余信息。在图像处理领域的应用中，卷积神经网络的提出是里程碑式的，成为无数深度学习视觉应用的基础框架。其特点在于强大的抽象表达能力，能够提取出图像不同层面上的特征。这些特征是对图像进行分类、目标检测的关键，同样也是图像压缩重建过程的核心。

另一方面，很多应用场景下图像被人眼接受前会先被计算机处理，如分类，识别，召回，语义分割。在被计算机处理前，图像往往需要先被解压。然而压缩形式的图像特征理论上具备着恢复图像的必要信息，因此具备跳过解压

而直接执行视觉任务的潜力。如何提取出兼具可读性和可重建图像的特征是一个重要的研究方向。

自然图像中存在着从全局到局部细节上不同尺度的特征信息，综合这些特征才能对图像有完整的表达。多尺度的特征的提取在视觉应用和图像压缩领域都有着很大的研究价值，是研发高压缩性能、无需解压即可执行视觉任务的图像压缩算法的可靠方案。

1.2 研究问题

图像压缩的根本目的是用更少的比特数去存储最有价值的信息，从而节约图像存储所需的硬盘空间和传输所需的带宽。

从是否丢失信息的角度，压缩分为无损压缩和有损压缩。无损压缩虽然在编码和解码的过程中不损失任何信息，能够完全恢复原始数据，其可压缩的程度往往非常有限。而有损压缩通过去除对人眼不敏感的冗余信息，可以换取很大程度的空间节省，有着更大的压缩潜力和更广泛的应用场景，是领域内主要关注的方向。

我们通过以下方法来描述有损压缩的模型[1]：图像在压缩和解压过程中存在于两个不同的域：信号域和压缩域。在信号域中，数据以未压缩的原始状态存储，比如图片的 RGB 三通道表示。压缩的过程中，信号域的数据通过一系列的变换，提取了原始图像的某些或者全部的重要特征，用更精简的方式存储在了压缩域中，比如通过 DCT（离散余弦变换）获得的图像中不同频率信号的参数，通过量化舍弃冗余信息。解压的过程则是仅利用压缩域的数据，通过一系列反变换去尽可能还原出原始的信号域数据。一个有损压缩算法主要有两个目标，也对应着它的两项评价指标。一方面，我们希望图像的还原度高。具体则是希望经过压缩和解压后重构出的图像在观感上和原始图像差别足够小，即失真率小。对失真率的衡量常用的指标包括峰值信噪比 (PSNR)、结构相似性[3](SSIM)、和多尺度结构相似性[4](MS-SSIM) 等。另一方面，我们希望压缩域数据的比特数足够小，即码率低。码率通常以（比特/像素）为单位。失真率与码率存在一个权衡，优化该权衡是有损图像压缩的主要目标。

在此基础上，为了降低计算复杂度，不完全解码或无需解码即可在视觉任务上获得良好表现对压缩编码所提取特征的可读性提出了要求。压缩域数据对

计算机的可读性可直接通过视觉任务中（例如图像分类）的精度来衡量。

综上所述，寻找一个失真率-码率权衡优化，同时压缩域特征执行视觉任务有可观精度的算法有很大的应用前景，是本工作研究的主要问题。

1.3 研究现状

有损图像压缩中，找到最优（即能得到最小的失真率-码率联合损失）的量化方式是一个高维的难解问题，因此传统的图像编码算法通常将图像线性变换到一个特征空间，再进行简单的取整量化和解码。近数十年，基于变换编码的图像压缩方法取得了很多成果。该方法的关键是找到空间角度最高效特征提取方法，也就是找到一个信号域到压缩域的最优变换。可以证明 KL 变换（Karhunen-Loève Transform）因其完全去相关性，是能量最小化角度的最优线性变换。基于此，应用极为广泛的 JPEG 采用了 DCT 变换。DCT 变换是 KL 变换的一种近似，而其有快速算法使得 JPEG 在能够获得相对可观的压缩性能的同时又得以在实际应用中获得优势。

其他常见的传统图像压缩算法包括基于小波变换的 JPEG2000、带有预测机制的 Webp 以及视频编码衍生出的 BPG 等。然而这些方法都局限于线性变换，实验证明如果通过学习优化一个非线性变换，把数据空间“扭曲”到更适合简单取整量化的状态，可以达到更好的效果[1][19]。神经网络是一种拟合能力强大的非线性变换函数。将基于神经网络的编码器和解码器引入图像压缩问题为该领域开创了一片新大陆。一个最早提出的神经网络用于压缩的模型是自编码器[6]。将一个较大尺寸输入图像，输入每层节点数逐渐减小的神经网络，最后只通过少数节点输出（称为瓶颈层），作为编码器。将这个输出作为输入连到一个节点数逐渐增大的神经网络，作为解码器。将得到的输出与原始输入用某个损失函数（如均方误差）计算距离，然后用梯度下降等方法训练优化。实际应用时，将自编码器的前半部分和后半部分拆开来，分别交给编码和解码方。然而自编码器将大尺寸的输入用少数节点数存储，本质上只是对数据的降维，不等价于优化失真率-码率。自编码器也存在许多限制条件而无法直接代替标准压缩算法，例如对于每个目标码率，需要不同的瓶颈层节点数，要专门训练多个网络；神经网络框架固定后，对输入尺寸有特定要求等。

尽管如此，对自编码器的框架进行改进被证明可以得到性能非常良好的压

缩算法。目前基于深度学习的图像压缩算法主要针对编码器解码器网络结构、可微量化函数、熵预测模型等部分进行改进优化；同时，很多引入自编码器框架之外的机制也取得不错的进展，如对图像内容适应的编码、超先验熵模型、基于 GAN 生成模型的解码器等。

深度学习获得的压缩算法与传统压缩算法相比，在适应各种特殊应用（如医学图像、双目图像等）中有着更大的灵活性[30]。基于深度学习的压缩算法的另一优势是有潜力获得更具可读性的压缩域数据。针对传统图像视频压缩域视觉应用中，紧凑的特征描述子通过消除空间冗余、同时利用视频纹理编码信息，在支持匹配和检索高性能的同时带来较大的压缩比。而基于深度学习的压缩方法里，人们发现以机器视觉为目标和以图像压缩为目标的特征提取所采用的神经网络结构相似。压缩得到的特征理论上包含视觉任务所需要的关键信息，因此可以凭借神经网络的灵活性改动视觉任务的网络结构来适应以压缩特征为直接输入。

1.4 本文工作

本设计中，以自编码器为框架，融合超先验熵参数模型和背景信息预测模型的图像压缩算法为基础，结合了多尺度特征提取编码，获得了兼具高压缩性能和特征计算机可读性的图像压缩模型。

文章组织结构上，第一章为绪论部分，介绍了图像压缩的研究背景和关键热点问题、对目前图像压缩算法研究现状进行了说明，并介绍了本文的主要工作和贡献。

第二章为本设计相关工作的介绍。主要概括了传统图像压缩方法并详细介绍了近年来基于深度学习的图像压缩算法框架下各模块上的进展工作。最后阐述了具有压缩域可读性的算法重要性和相关工作。

第三章中展开描述了端到端优化的图像压缩模型理论基础和实现方法，介绍了基于多尺度特征表达的图像压缩算法。

第四章详细介绍了模型压缩性能和视觉任务支持性的实验过程并对结果进行了分析

第五章是对本文工作的总结和展望。

2 相关工作

2.1 引言

早期的图像压缩方法主要基于各种人工手动设计的线性变换、量化，再通过熵编码获得比特流。随着深度学习方法的引入，凭借神经网络强大的非线性拟合能力，图像压缩算法性能得到了很大的提升。目前主流的压缩算法主要关注失真率-码率的损失优化，在 PSNR 和 MS-SSIM 等和客观图像质量指标上的表现。压缩算法中的编解码网络结构、熵参数预测、比特率分配、量化函数的优化改进等都是研究的热点。

压缩域视觉任务相关工作中，以设计针对压缩特征为输入的视觉任务网络为主。图像压缩、视觉任务网络联合优化方法也取得了良好结果。

2.2 传统图像压缩算法

近三十年应用最广泛的图像压缩算法采用的是 JPEG 标准[2]。JPEG 基于 DCT 变换（离散余弦变换），是 KLT 的近似。KLT 给出了线性变换中可能获得的最优基底，因此理论上适合用于图像高低频信息分离。JPEG 算法主要方法如图4−12所示：首先把 RBG 空间的图像变换到 YCbCr 空间，将对人更敏感的亮度和色差信息分开。然后将图像分割 8×8 像素块，进行 DCT 变换，分解成直流信号和不同频率交流信号。人眼对不同频率的图像信号敏感度不同，通常对高频信息敏感度低，因此这些信息可以视为冗余，根据头文件中的量化表，通过量化惩罚或去除，最终取到整数。然后从左上角起，按 z 字形顺序排列量化的数据，将参数由低频到高频排列，通过哈夫曼熵编码完成压缩。

JPEG2000[7] 是联合图像专家组制定的新编码标准，基于离散小波变换（DWT）。与 DCT 相比，DWT 不再要求固定尺寸的图像分块，不同区域允许得到不同的空-频分辨率，局部性更好。JPEG2000 相比 JPEG 具有更好的压缩性能，支持有损和无损压缩，应用更加灵活。

基于预测的编码方法中，Webp[36] 是谷歌在 2010 年提出的图片压缩标准，

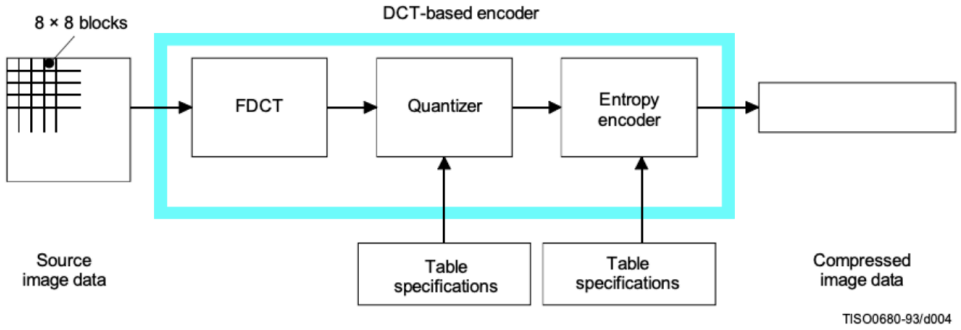


图 2-1: **JPEG 图像编码算法流程**[2]

衍生自视频编码标准 V8。与 JPEG 相比，同样具有图像分块、DCT 变换和熵编码。区别在于引入了预测机制，根据附近已编码图像块进行填充，预测了填充值与真实值的差。同时，熵编码采用算术编码替代哈夫曼编码，获得了更高的压缩率。此外，Fabrice Bellard 在 2014 年提出了由 HEVC 视频编码衍生的 BPG[8] 算法，性能相比 JPEG 也有着大幅度的提高。

2.3 基于深度学习的图像压缩算法

基于深度学习的图像压缩方法通常由以下几个部分组成：编解码器网络、量化函数、熵参数模型。

编解码器网络：对编解码器网络结构进行改进的工作中，Toderici[9] 等首先提出结合循环神经网络的自编码器，通过渐进式的残差构建，得到了压缩比可调的压缩算法，在低码率段重建效果好于 JPEG。随后[10] 对比联合 LSTM 和 GRU 与残差网络的结合，探索了不同残差构建方法后进一步提高了性能，获得了较宽码率段优于 JPEG 的表现。Johnston 等[13] 在该结构基础上提出预启动方法提高了编码器的表达效率，引入了空间自适应码率机制，使得码率可以随图像局部的复杂程度不同而进行动态调整，并使用了 SSIM 加权的损失函数，获得了更好的压缩表现。Theis[15] 等提出基于卷积神经网络的自编码器，采用基于取整的量化方式，得到了可适应任何尺寸的，压缩能力与 JPEG2000 相近的方法。Rippel[16] 等提出了一个基于金字塔分析的自编码器模块、一个自适应编码模块、预期编码长度正则化，并增加了多尺度对抗训练来补充模型，性能超过了此前所有基于 CNN 的模型，并相对轻量级，可在 GPU 上实时解码。Wen

等[33] 介绍了一种基于多尺度卷积的自编码器压缩方法，采用 Inception 模块结合金字塔尺寸调整的卷积网络架构作为编码器，在解码器引入残差网络提升其非线性，并设均值、方差独立预测的熵模型来预测特征分布。

量化器：在量化函数的改进工作中，Ballé 等[1] 从概率模型角度描述压缩问题，用归一化的噪声替代取整量化，解决了反向传播时量化器不可微的问题，实现了端到端的模型优化。Agustsson 等[17] 通过向量量化取代标量量化，反向传播时对 soft 松弛的函数微分，实现了量化赋值由软到硬的控制。

熵模型：熵参数预测模型相关工作中，Mentzer 等[11] 提出 3D-CNN，通过背景信息 (context) 模型对压缩域的特征分布进行建模，卷积自编码器利用 context 模型对特征的熵进行预测，context 模型则不断更新学习编码符号直接的相关性。Minnen[18] 引入了适应图像的多项分布字典，提供随不同图像自适应调整的边信息，辅助图像块码率的预测。基于超先验机制的方法中，Ballé 等[23] 在自编码器的基础上增加了一个超先验模块，预测的边信息作为熵模型的先验知识向熵编码、解码器提供指导，调整熵模型参数，降低了它与真实特征分布的不匹配程度。Minnen 等[24] 在此基础上把高斯混合尺度模型推广至混合高斯模型，同时生成一个以超先验为条件的均值和方差参数，并结合自回归模型可以在不增加比特数要求，只是用已被解码的特征的情况下更准确地对熵模型进行建模。

内容自适应：根据图像内容进行自适应的编码方法中，Li 等[14] 首先提出图像不同区域应赋予不同的码率，根据内容加权的重要性图指引比特分配。Baig 等[29] 指出假设每个分割出的待编码图块之间是互相独立的是不合理的，提出一种发掘邻域相关性的模块，通过邻域信息学习预测图像内容，节省下存储信息所需的比特数。Minnen 等[22] 提出了结合了神经网络和图像质量敏感的比特率适应机制，通过的块状结构进行 context 预测，保持了分辨率的灵活性和局域信息共享，同时大大简化了码率适应的实现。Lee 等[26] 指出不仅特征的标准差可以从图像中邻域预测，其均值也可以，并且同时预测特征的标准差和均值后的压缩方法比仅预测标准差更有效。

基于生成对抗网络：在基于 GAN 的重构图像生成方法中，Santurkar 等[21] 把解码器描述为生成函数，提出使用 GAN 学习在缩略图上的生成模型，再利用该模型作为解码器进行缩略图压缩。Agustsson 等[20] 采用了部分原信息保

留，纹理借助语义信息直接利用 GAN 生成，达到了 50% 以上的压缩比。考虑对失真图像进行后处理的工作中，Galteri 等[12] 用 GAN 训练了一个卷积残差网络，作为一个模拟人眼的评判器，无需一个描述图像质量的显式损失函数，生成去除人工失真点的图像。Mentzer[27] 认为各种观感评价指标各有缺点，考虑不专门设计失真损失，而是基于 GAN 来最小化原图和重构图像分布的差异，并在各种失真损失函数上进行优化后验证了失真率-码率-观感权衡的存在。

观感指标：在观感评价指标算法的问题上，Chinen 等[25] 采集了大量未被训练的人眼评价数据，训练了基于 VGG-16 的 10 参数网络来拟合数据，起到模拟人眼观感的作用。Blau 等[28] 则从数学上证明不仅失真率和码率存在不可兼得的权衡，观感和失真率在推向极端的条件下也是相互矛盾的。

压缩域数据分析相关工作中，Duan 等[31][32] 总结了基于紧凑描述子的，应用于推断和搜索的视频压缩标准。Torfason 等[30] 结合了用于推断的 ResNet、DeepLab 和 Theis 等[15] 提出的用于压缩的自编码器框架，针对压缩域图像的分类和语义分割问题进行训练。结果证明压缩域推断精度与数据域相当，并仅需要更少的操作数即可完成，对图像压缩和分类或语义分割联合优化时，二者表现同时提高。虽然该工作中视觉任务精度接近重建 RGB 图像，其压缩性能仅在低码率端与 JPEG2000 相当，与当前基于深度学习的压缩算法有着较大差距。Shen 等提出 CodedVision[34] 和 CodedRetrieval[35] 框架，设计了能够提取同时适合压缩和机器理解的特征的网络模型，分别在压缩和分类、压缩和召回任务中获得了可观的表现。

本设计采用了基于 inception 模块的多尺度特征学习编码器。编码器基于 Inception 模块架构，通过对图像不同尺度进行不同尺寸的卷积，得到串联的多尺度特征图。与自编码器压缩网络结合，学习到能够表达视觉任务所需特征的压缩模型。本工作通过编码时的多尺度特征提取来在保持压缩性能的同时，得到对于机器具有可读性的特征。

3　多尺度特征表达的图像压缩模型

3.1　端到端优化的图像压缩框架

本设计中的图像压缩模型以可端到端优化的卷积自编码器为主要架构。编码器部分由一个卷积神经网络构成，该网络输入 x 尺寸大于输出尺寸。通过多个卷积核对图像扫描得到不同特征 y。特征经过量化得到 $\hat{y}$ 再通过返卷积解码器网络得到重建图像 $\hat{x}$。为了解决取整量化函数不可微的问题，本文在训练时采用添加归一化噪声来替代取整量化。

存储小尺寸多通道特征数据的称为瓶颈层。在该层进行量化去除冗余，得到能用有限比特数存储的离散特征数据。为了能得到可存储和通讯的二进制比特流，需要对特征进行无损的熵编码。该过程中每一个特征对应于编码中的一个符号，算法需根据符号出现的先验概率才能得到最优编码。本文使用一个熵参数模型来预测特征的先验分布。当预测的熵模型和实际符号分布完全一致时，熵编码可以达到理论的香农信息熵极限。实际上，该架构可以被证明是一个变分自编码器。本文中用混合高斯模型对熵模型建模，即假设不同特征符号具有不同的均值和方差。

可以发现，当量化特征在被熵解码后，从预测编码的思想可知应可从中提取到一定的局域相关性信息。为了利用该信息，本文采用了[24] 所述自回归模型来借助已熵解码特征进行基于上下文背景信息 (context information) 的预测。

同时，搭建另一个自编码器网络，用于学习提取特征中的空间相关性信息 z。量化后的该信息 $\hat{z}$ 作为边信息加入到总的存储比特流中，用于修正对熵参数的预测。由于熵模型提供的是熵编码的先验知识，而上述自编码器网络提供的是指导熵模型的先验知识，因此将其称为超先验网络[23]。

超先验网络和自回归模型的输出参数 ψ 和 ϕ 通过一个熵模型预测网络计算得到对压缩特征均值 μ 和方差 σ 的预测。据此我们可以通过计算特征码长的期望得到香农信息熵。

模型的损失函数分为两项，即失真率和码率，分别衡量着图像压缩重建后的效果和压缩的程度，也正是图像压缩的两个目标。失真率由原图像与重建图

像之间的均方误差 (MSE) 计算求得。码率由熵模型所预测的香农信息熵计算求得。失真率和码率之间存在一个权衡 (trade off)，而不同的应用场景对失真率和码率有不同的要求，因此可以通过设置一个拉格朗日乘子作为超参数 λ 来控制压缩特性。由此可得联合损失函数 $L = R + \lambda D$，具体可写作：

$$L = E(-log_2 P(\hat{y})) + \lambda \|x - \hat{x}\|_2 \tag{3-1}$$

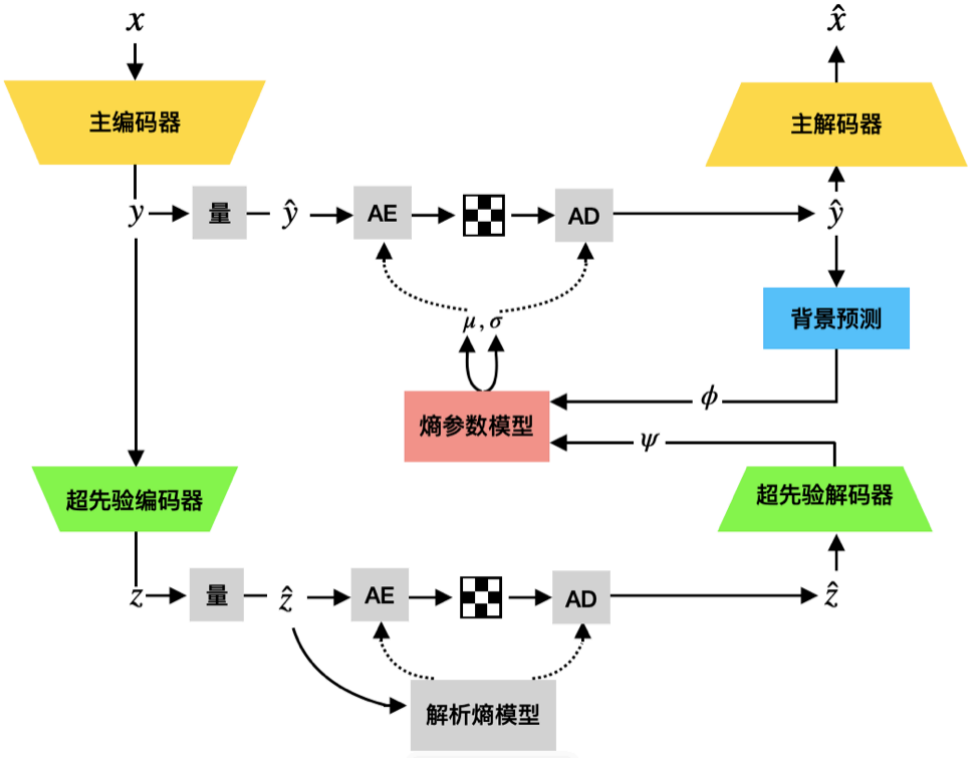


图 3-1: **端到端优化的图像压缩模型。**

设计完整框架如图 3-1 所示。图中 x 为输入图像、$\hat{x}$ 为重建图像；y 为编码器提取的特征，量化（或加入归一化噪声）后得到 $\hat{y}$；AE 和 AD 分别为熵编码和熵解码步骤；z 为超先验编码器提取的超先验空间信息，量化（或加入归一化噪声）后得到 $\hat{z}$；超先验网络所预测的参数为 ψ，自回归网络所预测的参数为 ϕ；熵参数模型所预测的特征分布参数 μ、σ 分别为 $\hat{x}$ 的均值和标准差。反向传播时，对联合损失进行优化，实现整个模型的端到端优化。

3.2 多尺度变分自编码器

3.2.1 变分自编码器

基于神经网络的变换压缩方法中，人们将信号域图像非线性变换到压缩域，再对其反变换得到重建图像，实现该过程的框架即为自编码器。自编码器是一种无监督学习，学习了观察量到隐变量的变换与其对应反变换。然而自编码器本质上是对数据进行了降维，而图像压缩的根本目标是尽力减小压缩域数据的信息熵和失真率，二者并不等价。我们需要更符合图像压缩本质目标的数学模型。

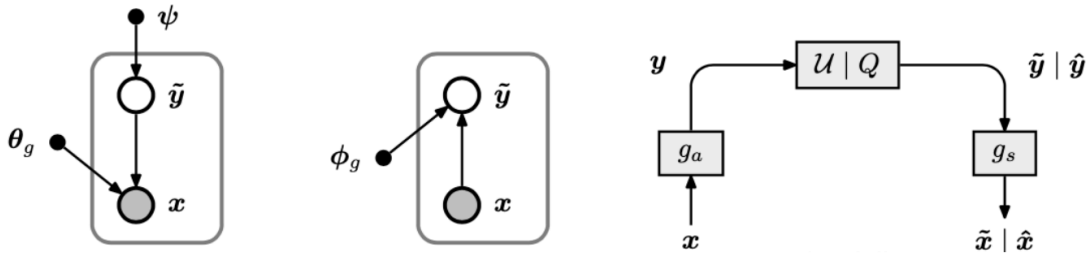


图 3-2: **图像压缩所采用的变分自编码器框架[1]。图左为生成模型；中间为推断模型；图右为变分自编码器**

贝叶斯推断问题描述的是通过已知的观察值 x 和生成模型 $p_{x|y}(x|y)$ 去推断隐变量 y 的后验分布 $p_{y|x}(y|x)$。通常该分布没有解析解。变分推断将变分法应用于此，构造一个参数化的概率模型（如高斯分布）去近似隐变量的后验分布。KL 散度是用于衡量两个分布的距离公式。Kingma 和 Welling[37] 提出可以假设一个参数化的分布 $q(y|x)$，设法最小化 $q(y|x)$ 和 $p_{y|x}(y|x)$ 的 KL 散度，这样把推断问题转化为了优化问题：

$$D_{KL}[q||p_{y|x}] = E_{y\sim q} log q(y|x) - E_{y\sim q} log p_{x|y}(x|y) - E_{y\sim q} log p_y(y) + const. \tag{3-2}$$

我们可以把变分自编码器的优化目标与图像压缩损失函数进行对比。若用添加归一化噪声代替量化步骤（原理在 3.3 具体介绍），则第一项为常数；第二项最大似然对应着原始图像 x 和通过 y 生成的图像 $\hat{x}$ 的均方误差；第三项隐变量概率对数的期望有着信息熵的形式，对应着码率。可见，对以像素分布

x 为观察值、以特征分布 y 为隐变量的变分自编码器进行优化，在 D 为 MSE 时，等于对上文所述图像压缩模型的 $R+\lambda D$ 联合损失进行优化。该框架等同于一个变分自编码器。

与普通自编码器相比，变分自编码器学习的不再只是观察值到固定的隐变量的变换，而是观察值到隐变量所服从的分布的变换。解码时，从给定隐变量分布中采样出一个隐变量向量，作为解码网络的输入。本文的图像压缩模型将以变分自编码器为主干，进行优化。

3.2.2 卷积编解码器

上章提到传统编码算法中，出于计算考虑常用线性变换。然而，没有理由认为在线性变换得到的压缩域进行简单标量量化能等价于对原信号域进行“最优”量化。相反，在我们不知道最优变换的形式时，用非线性函数去拟合它有望取得良好的效果。

近年来，人工神经网络（ANN）在包括图像处理在内的许多领域得到了关键应用，其核心作用是在大量数据支持下作为一个强大的非线性函数拟合器。人工神经网络在结构上模仿了生物神经元，由输入层、隐藏层、输出层组成。每一个神经元接收来自前一层的多个输入，对其做加乘运算后，作为下一层神经元的输入，加乘的权重和偏差是不断更新的。然而即便任意多层加乘操作累积的结果仍然是一个线性函数，因此在每个神经元输出前，数据通过一个简单的非线性激活函数，来增强网络的非线性。输入数据传入输入层，通过多层加乘和激活计算后得到输出数据的过程称为前向传播。根据输出数据与待拟合函数的距离设置损失函数。链式法则给出了计算损失函数相对网络中每一个待定参数的梯度。每次迭代，所有参数沿着梯度方向进行一定步长（学习率）的下降，该过程称为反向传播。随着前向反向传播不断迭代、网络参数不断更新，神经网络与目标函数越来越接近直至收敛。

由对于图像而言，最合适的神经网络变体为 Yann Lecun 提出的卷积神经网络[38]（CNN）。卷积神经网络每一层的神经元由多个卷积核组成，每一个卷积核分别对上层输入数据进行扫描卷积，提取出不同的特征，称为“特征图”，作为下一层的输入。扫描的计算机制使得卷积神经网络可以适应不同大小的输入图像，多核卷积的计算方法有着强大的抽象表达能力，使卷积神经网络具备学

习提取非常复杂特征的能力。

典型的卷积神经网络由卷积层、归一化层、池化层和非线性激活函数等组成。卷积层负责捕获图像空间信息，提取特征；批量归一化层（Batch Normalization Layer）将每一层特征图归一化到均值为 0、方差为 1 的分布上，这样可以使其更接近自然分布、加速训练；池化层引入了不变性，加强了对特征变化的容忍，并减少了参数，抑制过拟合；激活函数引入非线性，提高模型表达能力。

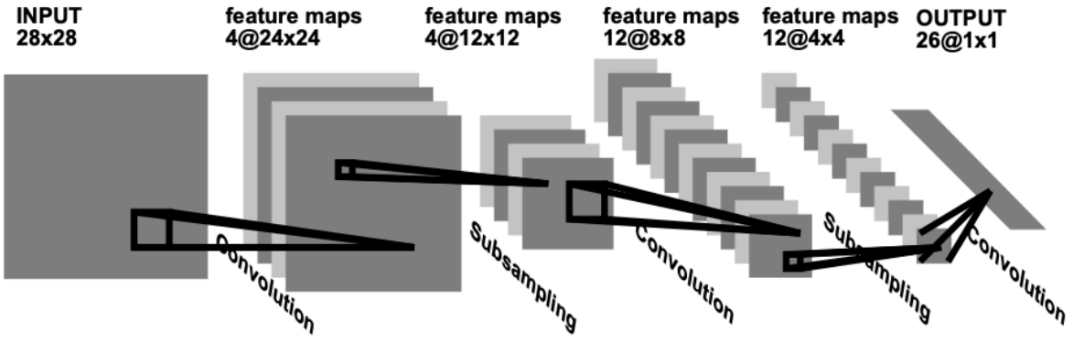


图 3-3: **卷积神经网络**[38]

本工作中的主编码器、解码器、超先验编码器、解码器、自回归网络、熵参数网络均采用了卷积神经网络。对比实验中参数设置参考了 Minnen18[24]，采用了四个卷积层组成的卷积神经网络作为主编码器，卷积核尺寸为 5×5，每层通道数（卷积核数）设置为 192、降采样率为 2。采用四个去卷积层组成的神经网络作为主解码器，每层卷积核参数同上，上采样率为 2。其中归一化非线性函数采用了 Ballé[1] 提出的泛化归一化层（GDN），具体见下节。

3.2.3 泛化归一化层

卷积神经网络中常用批归一化（Batch Normalization）来让特征图符合高斯特性，更易于训练。这样的归一化过程等于在特征图中引入的噪声，对于图像重建任务的不利的。本文在主编解码器中采用泛化归一化层[1]（generalized divisive normalization）。GDN 是一种参数化的、非线性归一化层。对其参数进行优化后，该层将输入数据转化为高斯分布。与 BN 相比，GDN 避免引入并有效消除噪声，更适合概率建模和图像重建问题。其计算方法如下：

$$y_i = \frac{x_i}{\beta_i + \sum (\gamma_{j,i}|x_i|^{\alpha})^{\epsilon}} \tag{3-3}$$

其中 x_i 为第 i 层特征图、y_i 为该层输出。$beta_i$ 和 $gamma_i$ 为需要学习的参数。文中取 $\alpha = 2, \epsilon = \frac{1}{2}$。GDN 是对自然图像的高效归一化，是 divisve 归一化[39]（Heeger，1992，一种局部增益控制，用于对神经元的非线性进行建模）的推广，在本文的主编解码器中起非线性归一化作用。

3.2.4 Inception 模块

提取更高维的特征、提高卷积神经网络在视觉应用中的性能最直接的方法之一是堆叠更深层的网络。然而更深层的网络所含有的巨大参数量带来了过拟合倾向和计算成本的增加，更重要的是会出现随深度增大的性能饱和。为此 Google[40] 在 ILSVRC14 中提出了基于 Inception 模块的 GoogLeNet，其主要思想是允许网络学习选择不同尺寸卷积核，以此从滤波器层面增大网络的稀疏性。ILSVRC14 中，其结构在明显提高精度的同时，相比 AlexNet 减少了 12 倍的参数量。

本文采用 Inception 模块用于编码器中的特征提取。模块中串联了尺寸为 7×7，5×5 和 3×3 的卷积核以及极大值池化层输出，得到了不同大小卷积出的特征图，最后通过 1x1 的卷积整合到 N 个通道，并通过 GDN 层归一化。

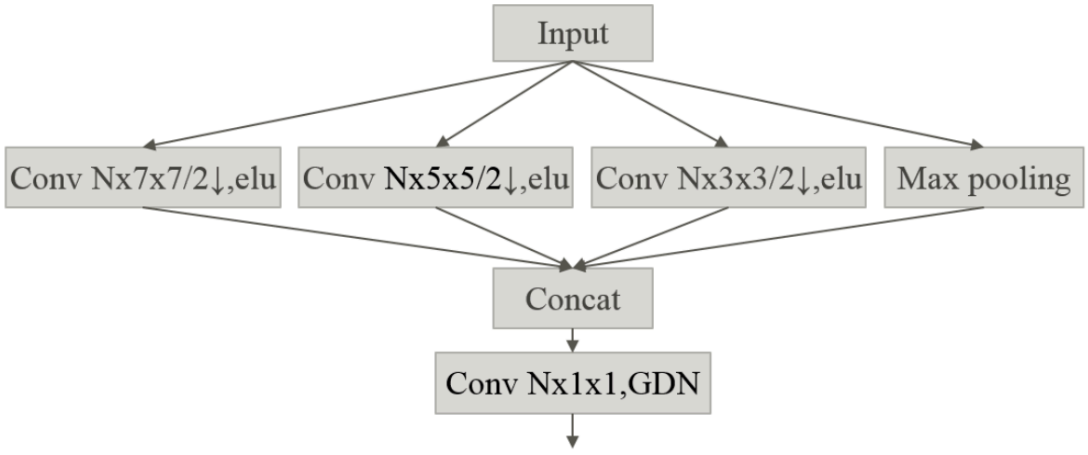


图 3-4: **本文采用的 Inception 模块[33]**

3.2.5 多尺度编码器

自然图像中存在的特征信息往往是具有不同尺度的，对不同尺度特征的分别提取在计算机视觉领域中有着重要的作用。例如图像分类中，图像中物体的大小是相对的，对图像进行不同尺度的特征扫描可以提高精度。相比普通卷积网络，多尺度特征的提取有助于提高机器视觉任务在压缩域中的性能。

本文实现了 Wen[33] 提出的金字塔 Resize 模型作为图像压缩的编码器。如图所示，首先用双线性插值把输入图像尺寸调整到 4 个不同的尺度（1:1, $\frac{1}{2}$:$\frac{1}{2}$, $\frac{1}{4}$:$\frac{1}{4}$, $\frac{1}{8}$:$\frac{1}{8}$），用卷积层在不同尺度上寻找结构特征，将四种尺度的特征图串联，即把各通道堆叠在一起。使用卷积网络时，网络深层一般获得全局粗略的信息，浅层获得局部细节信息。Inception 网络善于凭借不同卷积核发掘多尺度特征，因此在原尺寸图上使用四层 Inception 来提取全局高级信息。为了防止通道数爆炸，最后通过一个学习得到的权重层对 4xN 个不同尺度特征通道加权，最后经过通道数变换卷积层将通道数控制到 N 个。本文为了与 Minnen18 对比，将最后瓶颈层通道数设为相同的 N=192。

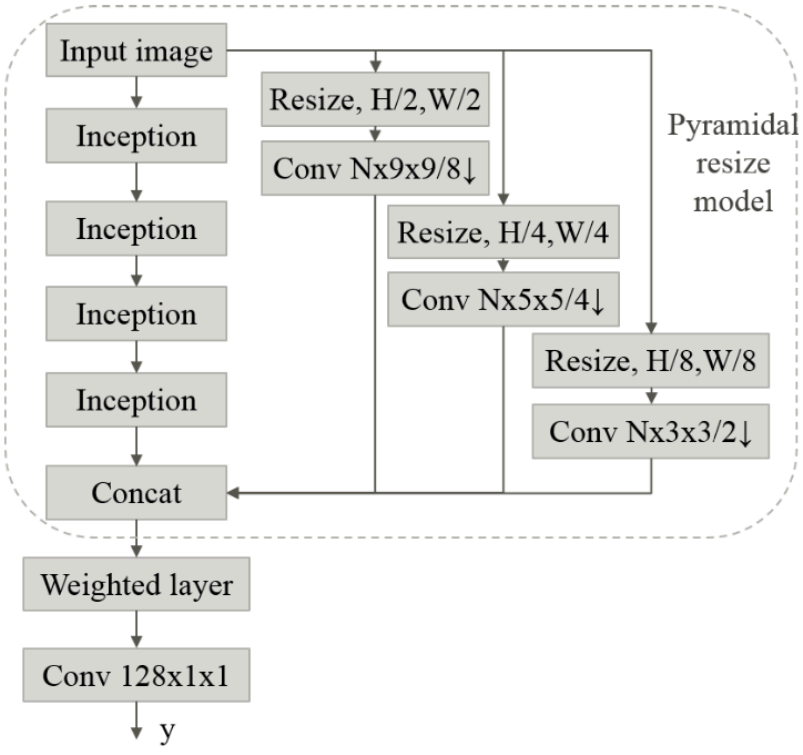


图 3-5: 本文采用多尺度编码器结构[33]

3.3 量化函数

具有量化步骤是有损压缩与无损压缩的最主要区别，也是码率节省的主要来源。在将图像变换到压缩域后，对特征进行取整量化。

$$\hat{y}_i = round(y_i) \tag{3-4}$$

其中 i 为瓶颈层所有空间、通道中的元素。该过程中，取整函数是不可微的，这将导致训练模型时，量化层梯度为 0 或者无穷。为了实现端到端优化，解决量化步骤反向传播时不可微问题，本文训练时采用[1] 中加入归一化噪声的方法替代量化，最终得到相应可微的代理损失函数。取归一化噪声宽度与取整区间长度相等。从概率模型角度看，量化后的特征分布是离散的，只有在可以取值的地方（整数值）有概率密度（迪拉克 δ 函数）。噪声对特征进行扰动给出了量化特征分布的连续松弛形式，松弛后的分布在原特征取值处有着和量化的特征分布相同的概率密度:

$$p_{\tilde{y}}(n) = P_{\hat{y}}(n), n \in Z^M \tag{3-5}$$

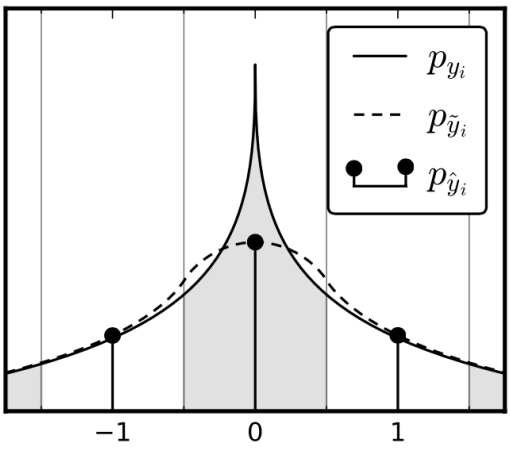


图 3-6: **添加归一化噪声是对量化特征分布的松弛[1]**

这样在训练时，特征服从一个连续分布。在反向传播时对其进行微分，该微分可近似为恒等变换，即可跳过量化器，实现网络的端到端优化。测试模型时，依旧正常取整量化即可。

3.4 熵模型

3.4.1 熵编码

特征经过量化后，需要通过熵编码转化为二进制文件。常见的熵编码包括香农编码、哈夫曼编码、算术编码等。熵编码的目标是用尽可能少的二进制位数去对消息进行可逆的编码。每一个离散的特征值视为一个待编码的符号，不同的符号需要编为不同的二进制码，因此分配的码长应该和符号出现概率成反比。根据香农信源编码原理，信源单符号输出的信息量期望由符号信息熵定义：

$$H(y) = \sum_{i}^{L} -p(y_i)log_2 p(y_i) \tag{3-6}$$

根据此可知，对于一个出现概率为 $p(y_i)$ 的符号 y_i，其最佳码长应该为 $-log_2 p(y_i)$。为了能够进行熵编码，我们需要知道待编码的特征符号的概率分布。然而这个分布由输入图像和主编码器的映射关系决定，是无法预知的。因此，我们建立一个对该分布的参数化假设，称为熵模型。

3.4.2 混合高斯模型

对特征符号分布的一种简单假设是单高斯模型。该模型认为隐变量（特征符号）服从一个高斯分布，所有特征符号有统一的均值和方差。这样的假设显然是过度简化的，不能对特征分布有很好的建模。

$$P(y_i) \sim N(x;\mu;\sigma) \tag{3-7}$$

Ballé 在 2018 年的工作中[23] 假设特征服从高斯混合尺度模型（Gaussian Scale Mixture），即所有特征有一致的均值和不同的方差。

$$P(y_i) \sim N(x;\mu;\sigma_i) \tag{3-8}$$

后 Minnen[24] 将其推广为混合高斯模型，假设特征有不同的均值和方差，需要分别预测。

$$P(y_i) \sim N(x; \mu_i; \sigma_i) \tag{3-9}$$

3.4.3 超先验网络

传统编码中使用边信息是常见的做法。边信息是从已有信息中提取的，用于进一步提高编码效率的指导性辅助信息。在图像编码中，边信息通常包含图像的某些空间相关性。HEVC 标准中，不同图像会进行不同的分割，分割结构作为边信息传递给解码器。相比之下，JPEG 只是固定的 8×8 像素块分割，而 HEVC 是变化的，解码器需要先解码边信息再根据边信息选择合适的熵模型解码特征块。

在基于深度学习的压缩方法中，熵模型的预测准确程度是影响压缩性能的关键。而不同的单个图像特征的分布与整个数据集图像特征分布常常相差很大，有必要针对不同的图像采用不同的熵模型（熵编码参数），该目标可以用额外的边信息来实现。我们希望额外添加的边信息总体长度短于熵模型提升带来的比特数下降。Ballé 在 2018 年提出[23] 可以用另一个自编码器预测熵模型参数，作为边信息加入比特流中。结果显示，边信息占用比特数很低，该机制的引入显著提高了压缩性能。

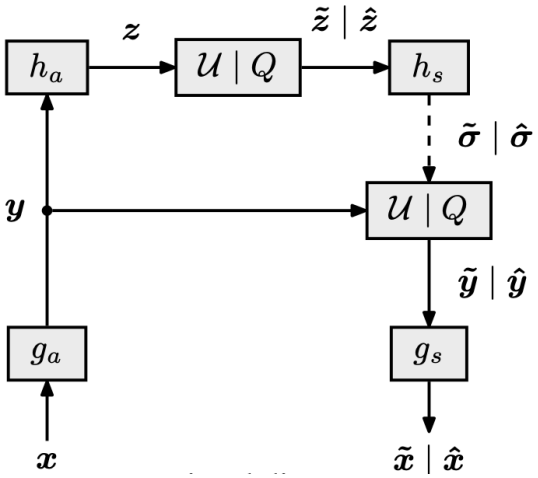


图 3-7: **带有超先验机制的图像压缩模型[23]**

该工作主要创新点在于，不再使用人工设计的边信息，而是学习一个熵模型参数的隐变量表示。通过端到端的优化保证总长度是最优的，平衡了边信息

长度和熵模型带来的性能提升。文中通过消融实验证明了超先验网络的确可以提取特征中的空间相关性。

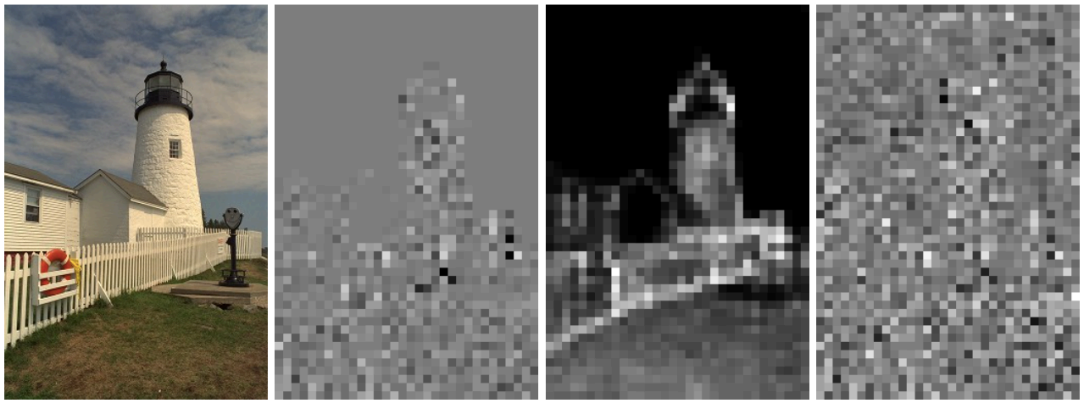


图 3-8: **Ballé18[23] 中对超先验机制作用的分析**

由于熵模型起自编码器中的先验作用，而学习到的边信息是熵模型的先验，因此称之为“超先验”。引入超先验边信息后，总码率项需要加入边信息所占的码率，联合损失函数变为：

$$R + \lambda D = E_{x\sim p_x}[-log_2 p_{\hat{y}}(\hat{y})] + E_{x\sim p_x}[-log_2 p_{\hat{z}}(\hat{z})] + \lambda E_{x\sim p_x}\|x - \hat{x}\|_2^2 \tag{3-10}$$

以 Kodak 数据集中一张灯塔图片为例。图 3-8 中左一为原图像；左二为通过分解熵模型网络提取的特征图，图中存在着明显的空间信息；右二为引入超先验网络后，预测的特征方差信息，可以看到超先验作为边信息携带了明显的空间信息；右一为引入超先验网络后的特征图，可见空间相关性大大减弱，这将带来极大的码率节省。

超先验网络所需预测信息较少，本文中分别采用三层卷积、去卷积组成超先验编码器与解码器，编码器第一层和解码器最后一层卷积核尺寸 3×3，其余为 5×5，总上下采样率均为 4。超先验编码器每层卷积核数均为 192，超先验解码器由于需要预测两倍通道数的熵参数信息，三个去卷积层通道数分别设为逐渐增大的 192、288、384。非线性激活函数采用 Leaky ReLU。

3.5 背景信息预测

图像中每一个点与其邻域之间是存在相关性的。例如平滑区域，领域间像素值会非常接近。因此领域存在着一定可以被利用的信息。特征图中这样的相关性也存在，用已解码的特征对未解码的特征进行预测可以带来一定程度的码率节省。

受此启发，Minnen2018 工作中[24] 提出，对已熵解码特征采取自回归式预测可以获得图像的一些上下文背景信息（context information），这部分的信息和超先验所预测的是互补的，二者结合可以更好地发掘隐变量的概率结构。这种自回归结构与生成模型 PixelCNN[41] 类似，每一个像素质（特征值）的概率以前面已生成的部分为条件，逐个生成：

$$p(y) = \prod_{i=1}^{n^2} p(y_i|y_1, y_2, ...y_{i-1}) \tag{3-11}$$

由于这个过程计算代价较为昂贵，本工作仅利用大小为 5×5 邻域内已解码特征。该过程通过带掩膜的卷积（MaskedCNN）实现，如下图所示。

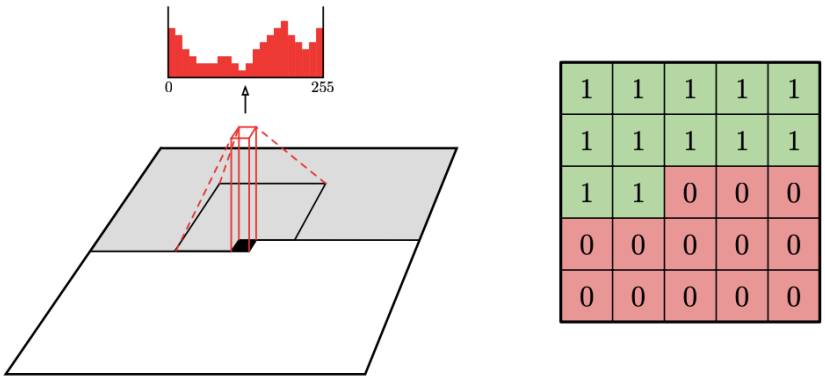


图 3-9: **PixelCNN[41] 所采用的掩膜卷积，每一个卷积核权重被乘以一个图右所示的掩膜，确保未被解码的信息不被利用**

Context 模型在接受者一端，当解码到第 i 个特征 y_i 时，仅能接触到已被解码的特征 $\hat{y}_{<i}$。混合高斯分布的均值和方差由超先验和 Context 模型一同预测。由于结合了 context 模型，超先验的编码器和解码器将会学习为不同的函数，因此不对超先验隐变量作假设，使用分解的熵模型。（由于边信息仅占文件很小一部分，叠加多层的超先验网络即便有效果，意义也不大）

自回归模型得到的对特征的预测需要超先验知识的修正。本文将自回归模型的输出与超先验网络的输出串联通过一个熵参数预测网络，切分得到对特征分布均值和方差的预测。

3.6 损失函数

联合损失函数由公式（3-10）给出。其由失真率项和码率项组成，码率包括主编码器特征码率和超先验特征码率。

失真率为重建图像与原始图像的均方误差。注意该距离也可以采用 MS-SSIM（多尺度结构相似性）来替代。由于 MS-SSIM 是可微的，可以直接优化。对 MS-SSIM 直接优化的模型在 MS-SSIM 上表现优于对 MSE 优化的模型。然而 Ballé18 指出对 MS-SSIM 优化并不能提高模型在 PSNR 上的表现，并且倾向于让模型分配更多码率在纹理上，而纹理不一定是图像中对人来说最重要的信息。本文采用使用更为广泛且有参照性的 MSE 计算失真。

码率计算根据特征熵模型参数进行预测。通过分布计算出估计的每个量化特征符号概率密度，通过香农信息熵计算其总信息量。设熵参数模型预测得到特征服从 $N(y|\mu_i,\sigma_i)$ 分布。则特征 $\hat{y}_i$ 的概率密度为（$\hat{z}_i$ 同理）：

$$p(\hat{y}_i)=\int_{\hat{y}_i-\frac{1}{2}}^{\hat{y}_i+\frac{1}{2}}N(y|\mu_i,\sigma_i)dy \tag{3-12}$$

总码率由特征平均信息熵除以像素数计算得到：

$$bpp_y=\frac{\sum_{i}^{L}-p(y_i)log_2p(\hat{y}_i)}{W\times H} \tag{3-13}$$

超先验特征同理。其中 W、H 为图像尺寸的宽和长。

4 实验结果与分析

4.1 压缩性能试验

4.1.1 数据集与预处理

本实验训练集采用自然图像组成的 ILSVRC2012 数据集，训练数据总共包含 1281167 个图像。训练前，对图像进行随机裁剪，得到 256×256 的固定大小图像块。再对图像进行随机水平翻转，增加多样性。

测试集采用了广泛应用于图像处理的柯达公开图像集，包含 24 张 512×768 像素的 RGB 图像。

4.1.2 模型参数

模型参数设置如下表所示。

<table>
<tr><td>主编码器
（对比）</td><td colspan="4">主编码器
（多尺度）</td><td>主解码器</td></tr>
<tr><td>Conv 5x5
cN s2</td><td>Inception cN s2</td><td>Resize H/2 W/
2</td><td></td><td rowspan="2"></td><td>Deconv 5x5
cN s2</td></tr>
<tr><td>GDN</td><td>Inception cN s2</td><td>Conv 9x9 cM
s8</td><td>Resize H/4 W/
4</td><td>IGDN</td></tr>
<tr><td>Conv 5x5
cN s2</td><td>Inception cN s2</td><td rowspan="2"></td><td>Conv 5x5 cM
s4</td><td>Resize H/8
W/8</td><td>Deconv 5x5
cN s2</td></tr>
<tr><td>GDN</td><td>Inception cN s2</td><td></td><td>Conv 3x3 cM
s2</td><td>IGDN</td></tr>
<tr><td>Conv 5x5
cN s2</td><td colspan="4">Concat</td><td>Deconv 5x5
cN s2</td></tr>
<tr><td>GDN</td><td colspan="4">Weighted Layer</td><td>IGDN</td></tr>
<tr><td>Conv 5x5
cN s2</td><td colspan="4">Conv 1x1 cP s2</td><td>Deconv 5x5
cN s2</td></tr>
</table>

超先验编码器	超先验解码器	背景预测模型	熵参数预测模型
Conv 3x3 cN s2	Deconv 5x5 cN s2	Masked 5x5 c2N s1	Conv 1x1 c$\frac{10}{3}$N s2
Leaky ReLU	Leaky ReLU		Leaky ReLU
Conv 5x5 cN s2	Deconv 5x5 c$\frac{3}{2}$N s2		Conv 1x1 c$\frac{8}{3}$N s2
Leaky ReLU	Leaky ReLU		Leaky ReLU
Conv 5x5 cN s2	Deconv 3x3 c2N s2		Conv 1x1 c2N s2

图 4-1: **其中卷积层标注格式为卷积核宽 × 卷积核高 c 通道数 s 步长。Deconv 代表反卷积。GDN、IGDN 分别代表泛化归一化层和反泛化归一化层。实验中 N、M 取 192**

4.1.3 评价标准

本文中采用两种常用的图像质量客观评价标准：峰值信噪比（Peak Signal to Noise Ratio，PSNR）和多尺度结构相似性（MS-SSIM）。PSNR 可以由均方误差计算，式中 MAX_I 为图像最大像素值，PSNR 值越高失真越小：

$$PSNR = 10log_{10}(\frac{MAX_I}{MSE}) \tag{4-1}$$

SSIM（结构相似性）假设人眼观看图像时，会提取其中的结构化信息，从亮度 $l(X,Y)$、对比度 $c(X,Y)$、结构 $s(X,Y)$ 角度对比图像。MS-SSIM 考虑了多尺度信息，更加贴合人眼的主观评价。

$$SSIM = l(X,Y)c(X,Y)s(X,Y) \tag{4-2}$$

其中：

$$l(X,Y) = \frac{2\mu_X\mu_Y + C_1}{\mu_X^2 + \mu_Y^2 + C_1} \quad c(X,Y) = \frac{2\sigma_X\sigma_Y + C_2}{\sigma_X^2 + \sigma_Y^2 + C_2} \quad s(X,Y) = \frac{\sigma_{XY} + C_3}{\sigma_X\sigma_Y + C_3} \tag{4-3}$$

式中 μ_X、μ_Y 为图像 X、Y 的均值；σ_X、σ_Y 为图像 X、Y 的标准差；σ_{XY} 为图像 X、Y 的协方差。MS-SSIM 可通过计算不同滑动窗口对应的 SSIM 取平均得到：

$$MSSSIM(X,Y) = \frac{1}{N}\sum_{k=1}^{N} SSIM(x_k, y_k) \tag{4-4}$$

4.1.4 训练细节

模型参数使用 Adam 优化器优化，初始学习率设置为 1×10^{-5}。使用学习率预热，在第 400 次迭代时，乘以 10 倍；在第 800 和第 3000 次迭代时，衰减至 0.5 倍。模型中的卷积层采用了 Xavier 权重初始化。

图像批量训练的大小为 14，由于训练集数据庞大，通常在第二个 epoch 即可收敛。为了获得不同失真率-码率权衡，本文训练了不同 $l\lambda$（0.002～0.1）取值下的模型。

4.1.5 结果分析

本实验对比了多尺度特征模型、Minnen18、JPEG、WebP、BPG 的 RD 曲线。每条曲线由对应压缩算法在取不同比特率时测得的 PSNR 和 MS-SSIM 绘制得到。

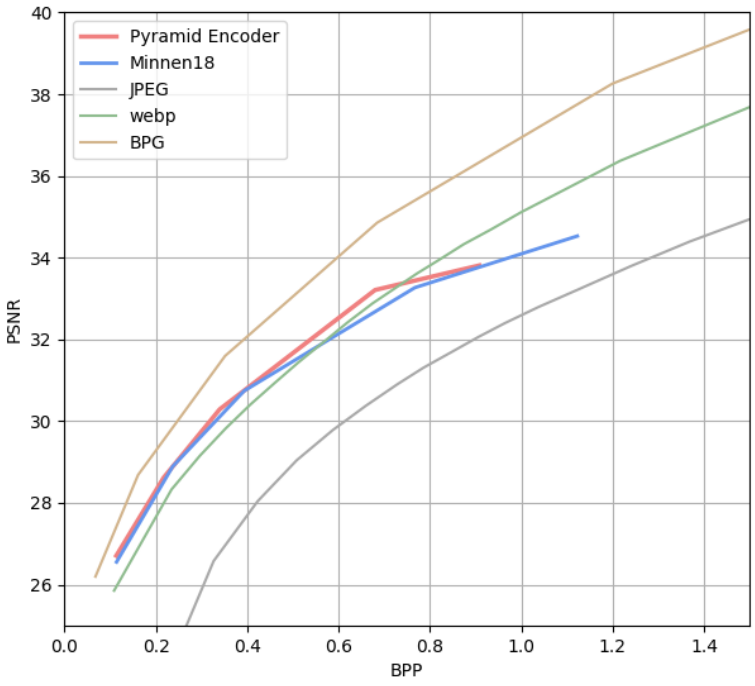


图 4-2: **不同编码算法的 BPP-PSNR 曲线对比，横轴为码率，纵轴为 PSNR**

可见基于深度学习的算法在 PSNR 上表现远好于 JPEG，与 WEBP 接近，不及 BPG。多尺度特征表达的压缩算法达到了与对比模型相近的压缩性能。

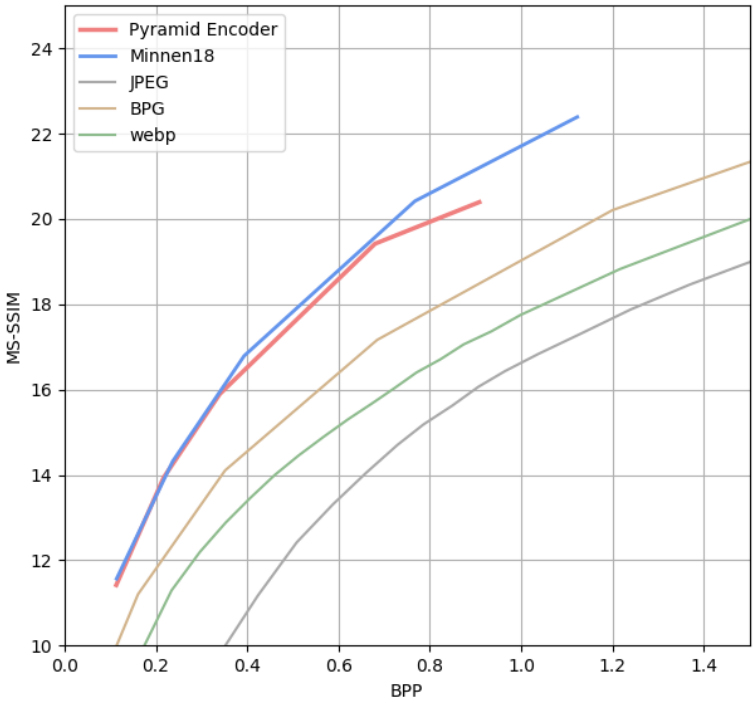


图 4-3: **不同编码算法的 BPP-MSSSIM 曲线对比，横轴为码率，纵轴为 MS-SSIM**

可见基于深度学习的算法在 MS-SSIM 上表现远好于 JPEG、WEBP 和 BPG。多尺度特征表达的压缩算法达到了与对比模型相近的压缩性能。

图 4-4: **采用 JPEG、WEPB 和多尺度自编码模型压缩的图像重建效果对比。bpp 为码率，ms-ssim 为多尺度结构相似性。注 *：JPEG 由于算法预设质量下届，能取得的最低码率为 0.173**

上图对比了 JPEG、WEBP 和多尺度自编码器模型的重建图像，可以看到在基本相同的码率下，基于多尺度编码器的图像压缩模型的重建图像人眼观感明显优于 WEBP 和 JPEG。

4.2 特征可读性实验

4.2.1 图像分类问题

监督学习中，图像分类是最常见的机器视觉任务之一。给定一组训练集和一组测试集，每个数据集中包含大量的图片，与每一张图片相匹配的有一个或多个标签，通常指示着图像中存在的事物属于哪个类别。算法在训练集进行训练，学习一个图像到标签的映射，然后以测试集图像为输入，检测其预测标签与正确标签的差别来衡量算法性能。图像分类算法的关键是提取有区分度的、鲁棒的数字化特征。每个图像可以用多个特征代表，作为高维特征空间的一点。图像分类则是对该空间进行划分，使得每张图可以通过其特征点于该空间所在位置得知其类别。传统的分类算法在特征提取上主要以设计有代表性的特征描述子为方向。空间划分上有常见的 K 最近邻点算法（KNN）、支持向量机（SVM）等。2012 年的 ImageNet 分类竞赛中，改进自 LeCun LeNet-5 的卷积神经网络模型 AlexNet[42] 以远高于第二名的性能获得了冠军，把基于深度学习的分类算法提高了一个大层次。2014 年，VGG[43] 凭借更深的层数所带来的

强大表达能力进一步提高了分类精度。此后 ResNet、GoogLeNet 等结构在减少参数的同时继续提高精度，成为主流的机器视觉任务框架。本文采用以 ResNet（深度残差网络）来检验压缩模型重建图像和压缩特征的分类效果。

4.2.2 深度残差网络

前文提到，堆叠更深层的卷积神经网络能提取更抽象的特征，是提高性能的直接办法，层数多到一定程度面临性能退化的问题。残差网络（Residual Network）的提出也是针对这一问题的创新。作者指出，网络层数加深导致的性能饱和是由于巨量的网络参数让优化变得困难。使用神经网络拟合一个复杂映射 H(x) 时，很多时候如果改为拟合其残差 F(x) = H(x)-x 会更容易。基于此，该工作设计了一种拟合目标映射残差的网络块，通过直接加入输入的恒等映射作为网络的捷径实现。这种方法未引入新参数且计算简单。实验证明 ResNet 因其残差结构，精度和收敛速度都高于对比网络。

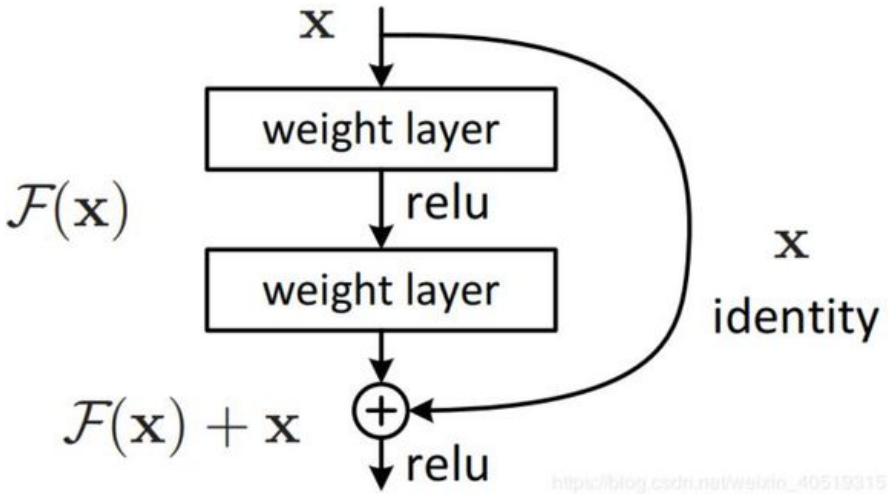


图 4-5: **残差网络块**[44]

本文采用 ResNet18 作为分类器主干。

对于重建 RGB 图像为输入的分类器，为了适应压缩网络特征图的通道数，第一层残差块输入通道数上由 64 改为 48，输入尺寸也改为 256×256。对于压缩特征，首先将其输入一个通道数变换卷积层，将通道数变换回原 4×N 个多尺度特征。分类时，该重建特征图与编码器中未经过权重层的 4×N 通道特征图计算 MSE，加入到分类损失中联合优化。本质上，该通道数变换层学习了一个 N 通道瓶颈层到 4×N 通道多尺度特征图的反变换。

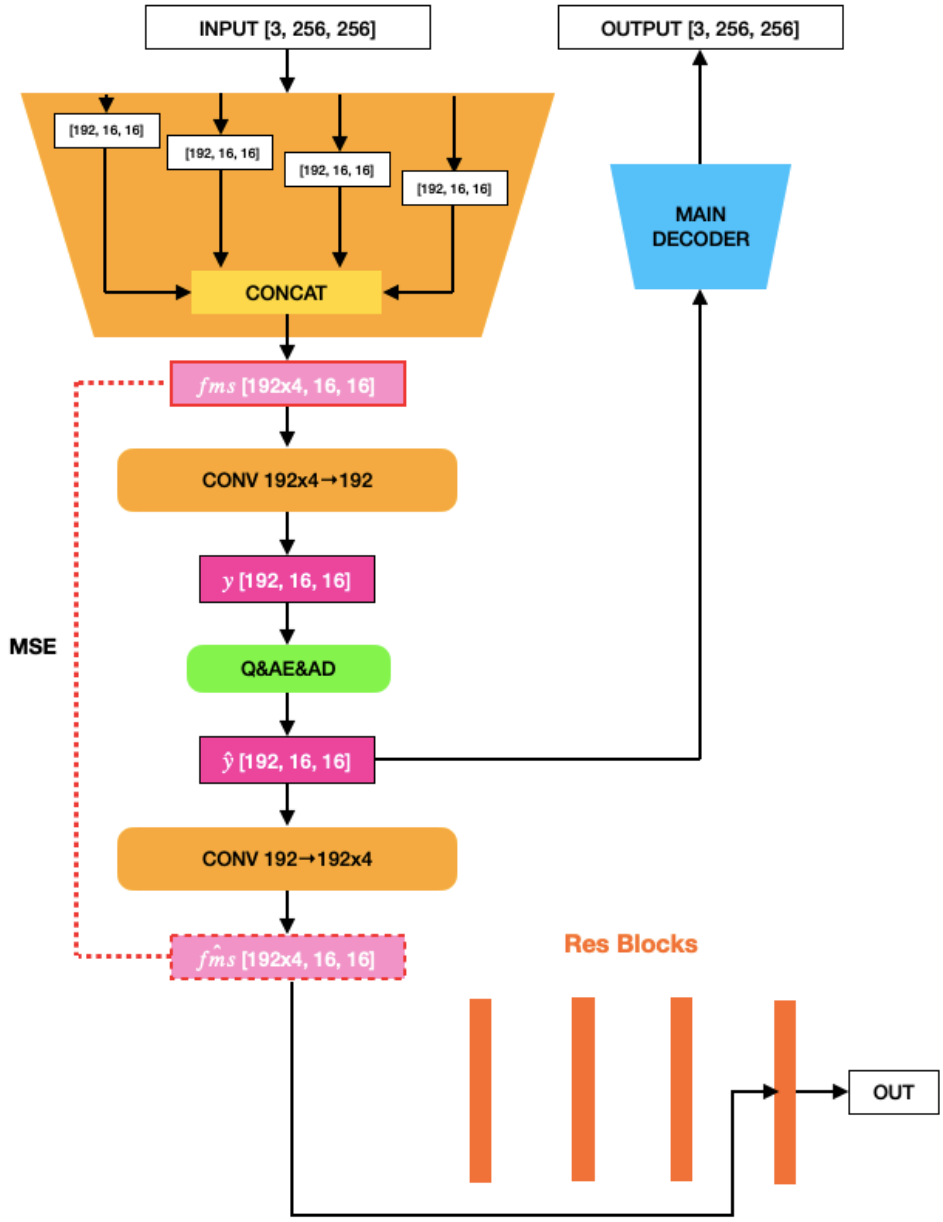


图 4-6: **压缩特征作为输入的分类网络结构。先训练压缩网络；然后固定住压缩网络，训练通道数变换层和分类器网络。**

对于以压缩特征为输入的分类器，本文对 ResNet18 结构进行类似 Torfason2018[30] 的改动，去除输入尺寸大于 16×16 的残差块，直接将特征数据输入最后一级残差块，成为 cResNet-k，其中 k 为输入层通道数。这样做的动机在于，分类器网络的越后面的卷积层处理的特征越抽象高维，而压缩得到的特征图本身就是图像信息的一种高维表达，压缩域特征图直接输入最后一个残差块不仅解决了尺寸问题，还大大降低了分类器的参数量，提高了计算效率。

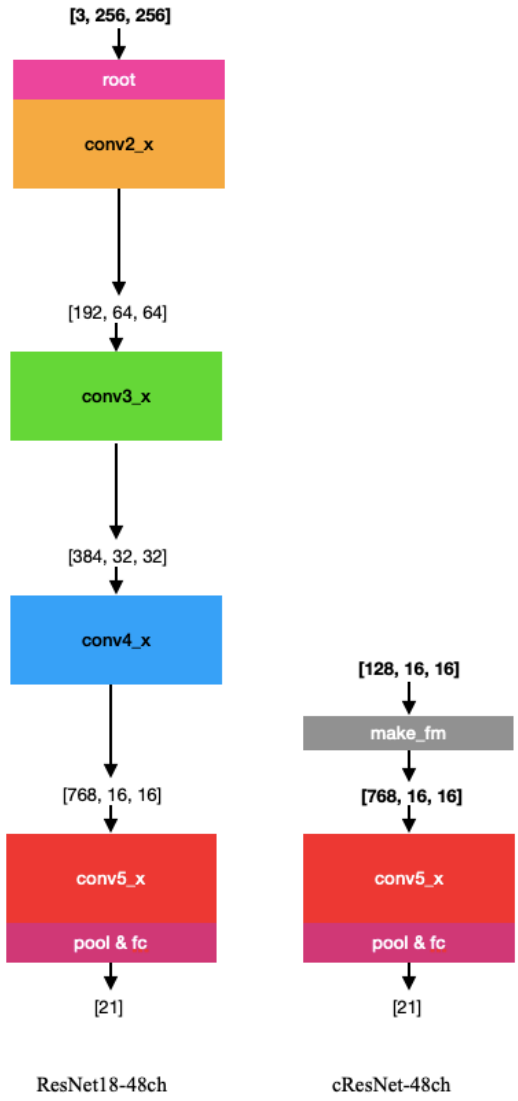


图 4-7: **ResNet18 与 cResNet 结构**

4.2.3 实验设置

本实验于 Pascal VOC2012 挑战数据集上，进行 20 个标签的多标签分类任务。选用 Adam 优化器，初始学习率设为 1×10^{-3}，第 80 和第 110 个 epoch 衰减至 0.1 倍。测试评价标准采用多标签图像分类常用的均值平均精度 mAP（mean Average Precision）。分类器网络结构如下表所示。

		RGB	压缩特征
卷积块	输出尺寸	ResNet-18-48ch	cResNet-48ch
conv2_x	64x64	3x3 maxpool 2↓ $\begin{bmatrix} 1\times1,48 \\ 3\times3,48 \\ 1\times1,192 \end{bmatrix}\times3$	无
con3_x	32x32	$\begin{bmatrix} 1\times1,96 \\ 3\times3,96 \\ 1\times1,384 \end{bmatrix}\times4$	无
conv4_x	16x16	$\begin{bmatrix} 1\times1,192 \\ 3\times3,192 \\ 1\times1,768 \end{bmatrix}\times6$	无
conv5_x	8x8	$\begin{bmatrix} 1\times1,384 \\ 3\times3,384 \\ 1\times1,1536 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1,384 \\ 3\times3,384 \\ 1\times1,1536 \end{bmatrix}\times3$
	1x1	average pool, 21d fc	

图 4-8: **分类器网络结构参数。对 RGB 图像采用 ResNet-18-48ch，；对压缩特征分类采用 cResNet-48ch，剥离了前三级的残差块。**

4.2.4 分类结果与分析

本设计中，多尺度模型和对比模型（Minnen18）分别训练了三个不同码率的模型，每个模型都分别进行重建图像的直接分类实验和特征分类实验。

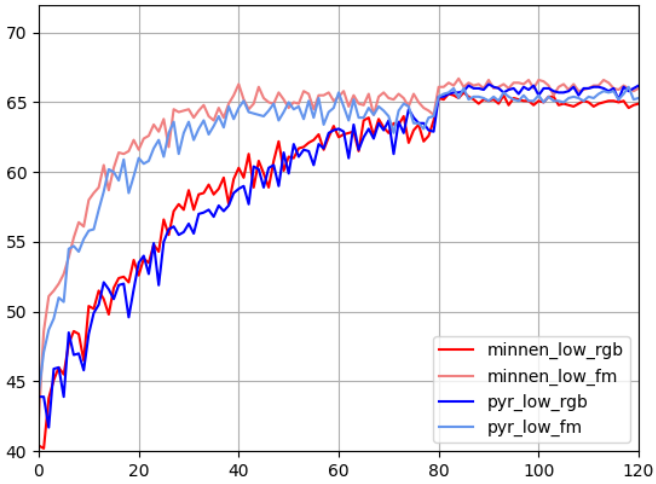


图 4-9: **低码率 (bpp=0.113) 模型的分类器收敛趋势**

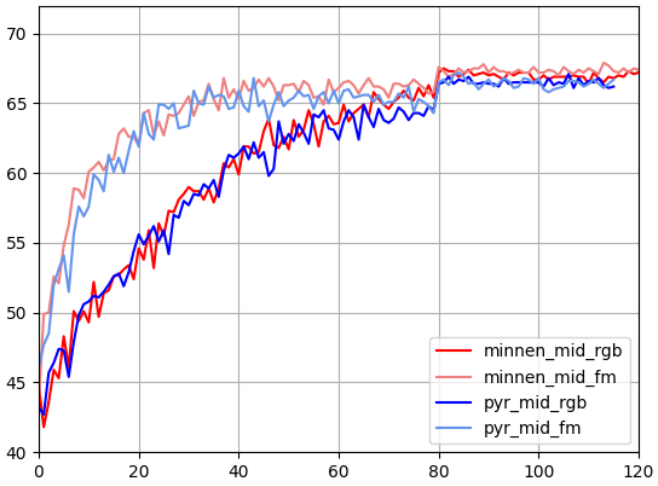


图 4-10: **中码率 (bpp=0.680) 模型的分类器收敛趋势**

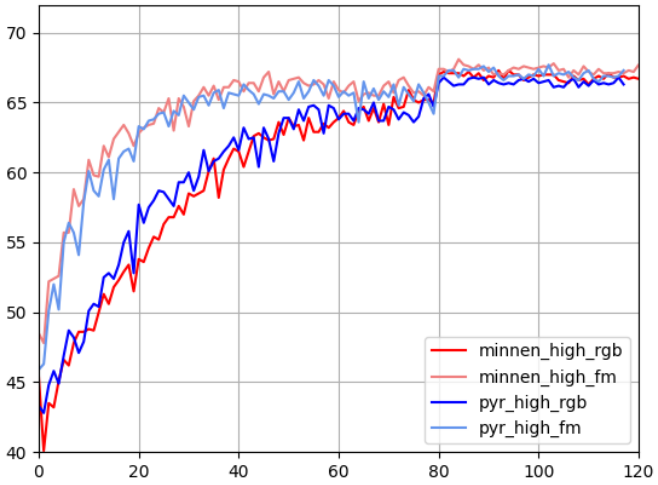
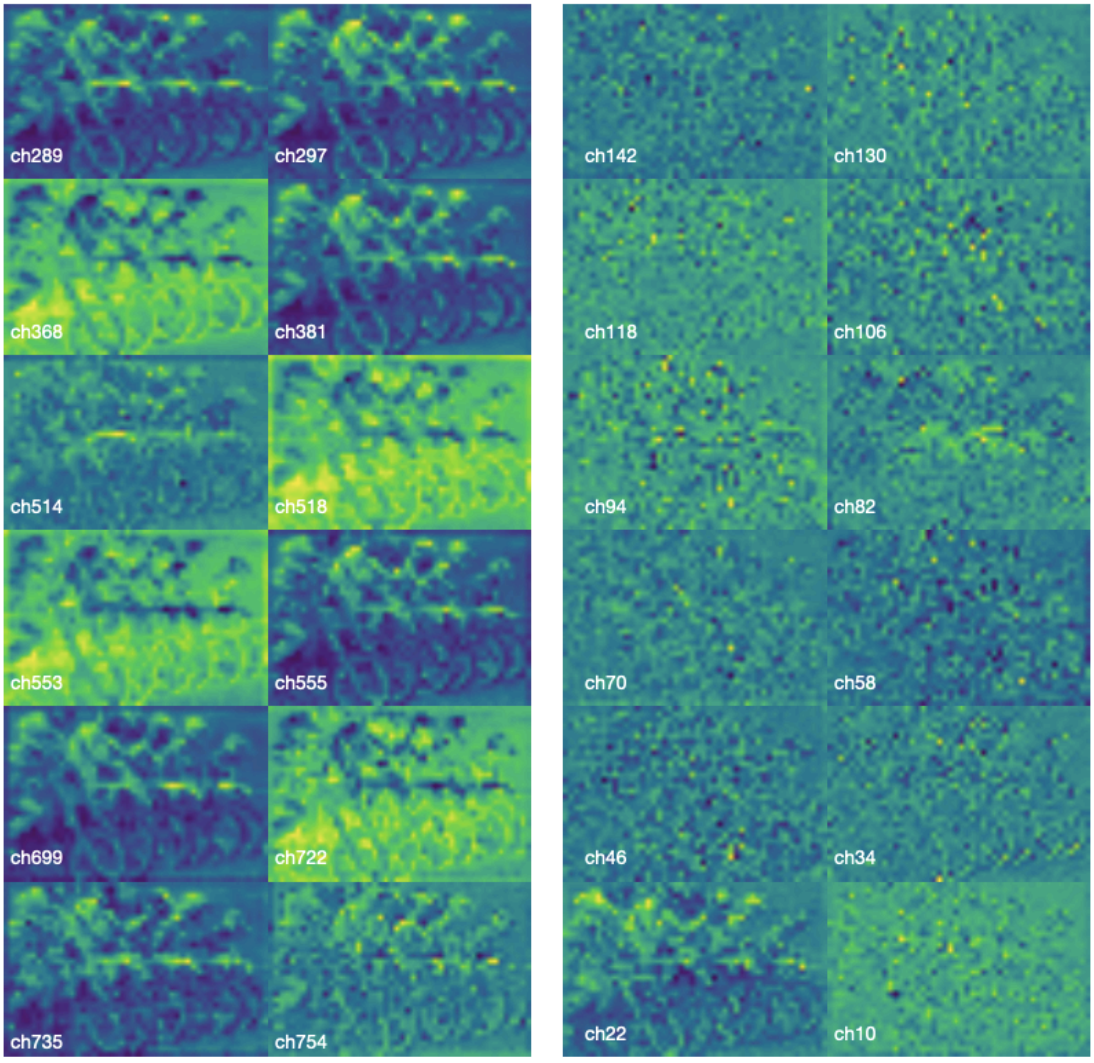


图 4-11: **高码率 (bpp=1.238) 模型的分类器收敛趋势**

可见各码率下，多尺度特征达到了和对比模型基本一致的分类精度。用压缩特征直接进行分类可以达到以重建图像为输入几乎一样的精度，且收敛速度更快。实验说明了该压缩模型的压缩域执行视觉任务的精度可以不低于重建图像，且无需完全解码。多尺度特征的分类性能与普通特征基本一致。

4.2.5 特征可视化分析

抽取了多尺度编码器和普通编码器的特征图进行可视化：



(a)多尺度编码器特征图　　(b)Minnen18特征图

图 4-12: **多尺度编码器和普通编码器的特征图。多尺度编码器特征图取自权重层前的 4xN 通道；Minnen18 取自瓶颈层。**

显然多尺度编码器与 Minnen18 相比，提取的特征有着更强的空间特征。超先验网络从特征中提取走了空间信息，导致了 Minnen18 瓶颈层空间相关性的缺失。这对于压缩是高效的，但限制了其压缩特征的机器可读性。而多尺度特征从瓶颈层恢复仅需一层卷积即可得到含有大量空间信息的特征图，对于视觉任务有着更大的潜力。

5 总结与展望

5.1 本文工作总结

本文调研了近年来基于深度学习的图像压缩算法进展，实现了一个端到端优化的图像压缩算法。该模型以自编码器为框架，通过卷积编码器提取图像特征，用归一化噪声替代量化获得可微的量化函数。量化的特征通过反卷积解码器解码出重建图像。一个混合高斯熵参数模型用于预测特征的分布，该预测分布可以计算码率的期望，从而实现优化。一个用于学习空间相关的边信息超先验自编码器，和一个基于自归回模型的上下文预测网络一同用于熵模型参数的预测。

本文还提出了结合多尺度表达的编码器的图像压缩模型，通过实验证明其压缩性能达到近前沿水平，并且提取出了更有视觉应用潜力的压缩域特征。

5.2 未来工作展望

未来工作中，考虑研究进一步优化编码器模型，通过改进各尺度特征在瓶颈层的权重等方法，提高其在高码率的压缩性能。应用方向上，研究如何在视觉应用中更好地利用多尺度特征的优势，使之能相较普通编码器提取的特征在分类、召回、语义分割等任务上获得更好的表现。综合编解码与视觉应用场景来看，如何改进编码器网络和视觉应用网络的适应性，实现多尺度编码与视觉应用更好的协同，也是未来的研究方向之一。

参考文献

[1] J. Ballé and E. P. Simoncelli. End to end optimized image compression. In ICLR, 2017.

[2] G. K. Wallace. The JPEG still picture compression standard. In IEEE Trans. CE,1992.

[3] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. In IEEE Trans. IP, vol. 13, pp. 600-612, 2004.

[4] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multi-scale structuralsimilarity for image quality assessment. In ACSSC'03, pp. 1398-1402, 2003.

[5] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks.[J].Science (New York, N.Y.),2006,313(5786):504-7.

[6] A. Skodras, C. Christopoulos and T. Ebrahimi. The JPEG 2000 still image compression standard. In IEEE Signal Processing Magazine, vol. 18, no. 5, pp. 36-58, Sept. 2001, doi: 10.1109/79.952804.

[7] F. Bellard. BPG Image format[EB/OL].https://bellard.org/bpg/,2014.

[8] G. Toderici, S. Malley, S. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell and R. Sukthankar. Variable rate image compression with recurrent neural networks. In ICLR, 2016.

[9] G. Toderici Damien, V. Johnston, S.Hwang, D. Minnen, J. Shor and M. Covell. Full resolution image compression with recurrent neural networks. In CVPR, 2017.

[10] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofter and L. V. Gool. Conditional probability models for deep image compression. In CVPR, 2018.

[11] L. Galteri, L. Seidenari, M. Bertini and A. D. Bimbo. Deep generative adversarial compression artifact removal. In ICCV, 2017.

[12] 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. In CVPR, 2018.

[13] M. Li, S.g Gu, W. Zuo, D. Zhao and D. Zhang. Learning convolutional networks for content-weighted image compression. In CVPR, 2018.

[14] L. Theis, W. Shi, A. Cunningham and F. Huszar. Lossy image compression with compressive autoencoders. arXiv preprint arXiv:1703.00395v1, 2017.

[15] O. Rippel and . Bourdev. Real-time adaptive image compression. arXiv preprint arXiv:1705.05823v1, 2017.

[16] E. Agustsson, F. Mentzer, M. Tschannen, L. Benini, L. Cavigelli, R. Timofte, and L. V. Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In CVPR, 2018.

[17] D. Minnen, G. Toderici, M. Covell, T. Chinen, N. Johnston, J. Shor, S.J. Hwang, D. Vincent and S. Singh. Spatially adaptive image compression using a tiled deep network. In ICIP, 2017.

[18] J. Ballé. Efficient nonlinear transforms for lossy image compression. arXiv preprint arXiv:1802.00847v2, 2018.

[19] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte and L. V. Gool. Generative adversarial networks for extreme learned image compression. In ICCV, 2019.

[20] S. Santurkar, D. Budden, and N. Shavit. Generative compression. In PCS, 2018.

[21] D. Minnen, G. Toderici, S. Singh, S. J. Hwang and M. Covell. Image-dependent local entropy models for learned image compression. arXiv preprint arXiv:1805.12295v1, 2018.

[22] J. Ballé, S. J. Hwang, D. Minnen and N. Johnston. Variational image compression with a scaled hyperprior. In ICLR, 2018.

[23] D. Minnen, J. Ballé and G. Toderici. Joint autoregressive and hierarchical priors for learned image compression. In NIPS, 2018.

[24] T. Chinen, J. Ballé, C. Gu, S. J. Hwang, S. Ioffe, N. Johnston, T. Leung, D. Minnen, S. O'Malley, C. Rosenberg and G. Toderici. Towards a semantic perceptual image metric. In ICIP, 2018.

[25] J. Lee, S. Cho and S. Beack. Context-adaptive entropy model for end-to-end optimized image compression. In ICLR, 2019.

[26] F. Mentzer, G. Toderici, M. Tschannen and E. Agustsson. High-fidelity generative image compression. arXiv preprint arXiv:2006.09965v3, 2020.

[27] Y. Blau and T. Michaeli. The perception-distortion tradeoff. arXiv preprint arXiv:1711.06077v4, 2020.

[28] M. H. Baig, V. Koltun and L. Torresan. Learning to inpaint for image compression. In CVPR, 2017.

[29] R. Torfason, F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte and L. V. Gool. Towards image understanding from deep compression without decoding. In ICLR, 2018.

[30] L. Duan, V. Chandrasekhar, S. Wang, Y. Lou, J. Lin, Y. Bai and T. Huang. Compact descriptors for video Analysis: the emerging MPEG standard. arXiv preprint arXiv:1704.08141v1, 2017.

[31] L. Duan, J. Lin, J. Chen, T. Huang and W. Gao. Compact Descriptors for Visual Search. In IEEE MultiMedia, vol. 21, no. 3, pp. 30-40, July-Sept. 2014, doi: 10.1109/MMUL.2013.66.

[32] S. Wen, J. Zhou, A. Nakagawa, K. Kazui, and Z. Tan. Variational autoencoder based image compression with pyramidal features and context entropy model. In CVPR, 2019.

[33] Q. Shen, J. Cai, L. Liu, H. Liu, T. Chen, L. Ye, and Z. Ma. CodedVision: Towards Joint Image Understanding and Compression via End-to-End Learning. In PCM, 2018.

[34] L. Liu, H. Liu, T. Chen, Q. Shen and Z. Ma. Codedretrieval: Joint Image Compression and Retrieval with Neural Networks. In VCIP, 2019, pp. 1-4, doi: 10.1109/VCIP47243.2019.8965918.

[35] Rabbat, Richard, "WebP, a new image format for the Web". Chromium Blog.

[36] Kingma, Diederik P. and M. Welling. Auto-Encoding Variational Bayes. In: arXiv e- prints. Presented at the 2nd Int. Conf. on Learning Representations. arXiv: 1312.6114.

[37] Y. LeCun, P. Haffner, L. Bottou, Y. Bengio. Object recognition with gradient-based learning. Contour and Grouping in Computer Vision. 1999.

[38] Carandini, Matteo and Heeger, David J. Normalization as a canonical neural computation. Nature Reviews Neuroscience, 13, January 2012. doi: 10.1038/nrn3136.

[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In CVPR, 2015.

[40] A. van den Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves. Conditional image generation with pixelcnn decoders. In NIPS, 2016.

[41] A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks[J]. In NIPS, 2012, 25(2).

[42] K. Simonyan, A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[43] K. He, X. Zhang, S. Ren, J. Sun. Deep Residual Learning for Image Recognition. In CVPR, 2016.

致　　谢

感谢我的导师沈秋老师和刘霖枫师兄在毕设过程中给我提供的悉心指导、思维启示与有意义的讨论。进行毕业设计的四个月是我本科期间学到东西最多的一个学期，很感恩有这样的机会磨练心智，汲取宝贵的科研经验。

感谢学院老师们从大一以来的传道授业，我的学科基础由此奠定，是未来科研职业道路的基石。四年南大的学术氛围的熏陶也将成为影响我一生的气质。

感谢父母和朋友一直以来的支持和鼓励，本科四年将会是我永远不会忘怀的青春记忆。在这里带走与留下的都将永恒地刻在时空里。