北京大学

博士研究生学位论文

题目：生成式图像编码方法研究

姓　　名：常建慧

学　　号：2101111559

院　　系：计算机学院

专　　业：计算机应用技术

研究方向：人工智能与媒体计算

导　　师：高　文　教授

马思伟　教授

☑学术学位　　□专业学位

二〇二四 年 六 月

摘要

在当今互联网快速发展的背景下，得益于移动互联网和高性能计算设备的广泛普及、人工智能技术的深度应用，推动了内容创作和编辑工具的持续革新，图像和视频内容的创作与分享已经无处不在。视觉内容的创作方式从专业制作内容（Professionally Generated Content, PGC）、用户生成内容（User Generated Content, UGC）发展到人工智能生成内容（Artificial Intelligence Generated Content, AIGC），促使服务于视觉感受的图像和视频数量呈现出爆炸性增长，对图像和视频压缩技术提出了新的挑战。随着压缩比的提高，数据信号的损失不断增大，图像视频重建质量随之下降。在极低码率条件下，重建结果难以保持关键的结构信息，视觉质量严重受损。而生成模型在数据驱动下能够深入理解和模拟复杂数据分布，展现出了以稀疏表示为条件生成复杂视觉内容的强大能力，为图像视频编码提供了在极低码率下实现高质量重建的新途径。

在当前与生成式编码相关的探索中，一种是在传统编码范式的基础上引入生成对抗损失，以在码率、失真和感知质量之间实现更优的权衡；另一种则是利用生成模型的采样生成能力实现图像重建，但现有工作的效果尚不理想，亟需探索更高效准确的生成式建模与编码方法。本研究深入研究了生成式图像编码方法，围绕编码端的高效压缩域表示建模及优化、解码端条件生成机制设计及失真优化这两大关键问题展开研究，旨在极低码率下实现高视觉质量的重建，最大化挖掘生成式编码的压缩潜力。本文的主要贡献点如下：

（1）针对极低码率编码的挑战，提出了基于结构-纹理分层的生成式编码框架。该框架打破了传统压缩范式依赖单一压缩域表示的局限，将视觉内容分解并建模为空域相关的稀疏结构表示以及空域无关低维紧凑的纹理表示，不仅有效降低了码率，还实现了对关键结构信息的显式保留，优化了重建图像的结构失真。在解码端，设计了一种结构与纹理层次融合的生成模型，实现了基于结构表示与纹理表示合成目标重建图像。广泛的实验证明，该方法能够在极低码率（< 0.1 bpp）范围以 LPIPS（Learned Perceptual Image Patch Similarity）为评价指标超越最新编码标准 VVC 的编码性能，主观重构质量超过了 VVC 感知优化方法、经典端到端神经网络编码方法、以及其他引入生成对抗损失的端到端编码方法。此外，得益于结构纹理分层生成式编码，提出方法能够实现更好的视觉重建质量、灵活的内容编辑以及对各种下游视觉任务的支持。

（2）针对复杂内容场景难重建的挑战，提出了语义先验引导的分层生成式编码方法，旨在将图像编码为紧凑、语义感知的视觉分解表示。提出基于语义先验信息构建语义级细粒度纹理表示，并提出基于注意力机制的非均匀纹理表示建模与对比一致表

示空间优化方法，获得了面向主观感知的高效特征表示，提升了复杂场景的重建质量；针对现有生成式编码方法缺乏可学习的码率约束的问题，基于对纹理表示统计特性的分析提出了跨通道熵模型建模纹理表示码率约束，实现了特征高效编码与率失真联合优化，显著提高了编码效率。实验结果显示，该方法在不同语义场景中相比传统视频编码标准 VVC 实现了 31.19% 至 59.92% 的性能提升，并且支持人脸视频编码、图像编辑及分析等多种应用。此外，通过结合生成式编码与残差信号补偿，该模型在至多 0.4 bpp 的码率范围内主观压缩性能优于 VVC。最后，在手机端部署所提出的生成式编码模型，验证了该编码方案在移动设备上的实际可行性。

（3）针对解码端不同生成模型难以兼容的挑战，本文提出了一种融合模型先验知识的生成式编码框架。该框架中的编码器不依赖于特定的生成模型，能够提取更为通用的压缩域表示。此外，通过自适应特征映射网络实现压缩域表示与不同生成模型之间的兼容和适配。最后，通过将压缩信息融入预训练扩散模型的生成过程中，利用模型先验知识提高压缩性能。以上设计使得所提方法成为具有更强的灵活性、兼容性和可迭代性的新型的生成式编码方法。进一步，针对生成式编码中存在的失真问题，提出了交叉注意力驱动的条件融合方法、生成域隐变量预测增强技术和基于统计分布的图像增强算法，显著提升了图像重构质量。此外，针对编码过程中的信息冗余问题和编解码流程无法实现端到端联合优化的问题，提出了联合隐变量在线优化和编码器离线优化的双阶段优化策略。最后，针对极低码率下重构图像存在的结构失真问题，通过深入分析压缩域隐变量，提出了一种适用于低码率编码的结构一致性优化算法，以在统一的压缩框架内显式保留和优化结构特征，有效减少了极低码率编码时重构图像的结构失真。实验结果表明，相较于最新一代视频编码标准 VVC，所提方法实现了 87.75% 的性能提升；与最新的端到端压缩方法相比，达到了 78.35% 的性能提升，展示了生成式编码的显著优势。此外，实验结果表明本章方法能够兼容其他预训练生成模型，并且在真实拍摄的场景以及更高分辨率的图像上展现出良好的适用性。

综上所述，本文致力于研究面向极低码率的生成式编码方法，从高效视觉表示建模与优化出发，提出了基于结构-纹理分层的生成式编码框架以及语义先验引导的分层生成式编码方法，实现了千倍压缩与高主观质量重建，并在移动端验证了该方案的可行性。进一步地，将研究焦点转向如何有效利用生成扩散模型的先验知识，构建更灵活、更易于迭代优化的生成式编码新方法，并探索生成式编码的性能潜力。所提编码方案在极低码率条件下，比最新一代视频编码标准 VVC 压缩性能提升了 87.75%，展现了生成式编码技术巨大应用潜力，推动了编码技术在更高效率和更佳主观视觉体验方向的发展，为未来智能图像视频压缩技术的发展研究贡献了新的思路和方法。

关键词：图像编码，生成模型，极低码率，视觉质量

Research on Generative Image Compression Methods

Jianhui Chang (Computer Application Technology)
Directed by: Prof. Wen Gao and Prof. Siwei Ma

ABSTRACT

In the context of the rapid development of the internet today, the widespread adoption of mobile internet and high-performance computing devices, along with the deep application of artificial intelligence technology, has driven continuous innovation in content creation and editing tools. Image and video content creation and sharing have become ubiquitous. The methods of creating visual content have evolved from Professionally Generated Content (PGC) and User Generated Content (UGC) to Artificial Intelligence Generated Content (AIGC), leading to an explosive growth in the number of images and videos catering to visual experiences, posing new challenges to image and video compression technologies. As the compression ratio increases, the loss of data signals grows, and the quality of image and video reconstruction decreases accordingly. Under ultra-low bit-rate conditions, it becomes challenging to maintain key structural information, resulting in severe degradation of visual quality. However, generative models, driven by data, can deeply understand and simulate complex data distributions, demonstrating a strong capability to generate complex visual content conditioned on sparse representations, providing a new pathway for achieving high-quality reconstruction of images and videos at ultra-low bit rates.

In the current exploration of generative coding, one approach introduces adversarial loss based on traditional coding paradigms to achieve better trade-offs between bit rate, distortion, and perceptual quality. Another approach leverages the sampling and generation capabilities of generative models for image reconstruction. However, the existing work has not yielded satisfactory results, necessitating the exploration of more efficient and accurate generative modeling and coding methods. This research delves into generative image coding methods, focusing on efficient compressed domain representation modeling and optimization on the encoding end, and conditional generation mechanism design and distortion optimization on the decoding end, aiming to achieve high visual quality reconstruction at ultra-low bit rates and maximize the compression potential of generative coding. The main contributions of this thesis are as follows:

(1) To address the challenges of ultra-low bit rate coding, a generative coding framework based on structural-textural layering is proposed. This framework breaks the limitation of traditional compression paradigms that rely on a single compressed domain representation by decomposing and modeling visual content into sparse structural representations that are spatially correlated and compact textural representations that are spatially uncorrelated. This not only effectively reduces the bit rate but also explicitly preserves key structural information, optimizing the structural distortion of the reconstructed image. On the decoding end, a generative model that fuses structural and textural hierarchies is designed to synthesize the target reconstructed image based on structural and textural representations. Extensive experiments demonstrate that this method can surpass the latest coding standard VVC in terms of LPIPS (Learned Perceptual Image Patch Similarity) at ultra-low bit rates (<0.1 bpp), and its subjective reconstruction quality exceeds VVC perceptual optimization methods, classic end-to-end neural network coding methods, and other end-to-end coding methods that introduce adversarial loss. Additionally, thanks to structural-textural layered generative coding, the proposed method achieves better visual reconstruction quality, flexible content editing, and supports various downstream visual tasks.

(2) To tackle the challenge of reconstructing complex content scenes, a semantic prior-guided layered generative coding method is proposed, aiming to encode images into compact, semantically-aware visual decomposition representations. The method constructs semantic-level fine-grained textural representations based on semantic prior information and proposes an attention mechanism-based non-uniform textural representation modeling and contrastive consistency representation space optimization method, achieving efficient feature representations oriented towards subjective perception and enhancing the reconstruction quality of complex scenes. Addressing the lack of learnable bit rate constraints in existing generative coding methods, a cross-channel entropy model is proposed based on the statistical characteristics of textural representations to model the bit rate constraint, achieving efficient feature encoding and rate-distortion joint optimization, significantly improving coding efficiency. Experimental results show that this method achieves a performance improvement of 31.19% to 59.92% compared to the traditional video coding standard VVC across different semantic scenes, and supports various applications such as facial video coding, image editing, and analysis. Moreover, by combining generative coding with residual signal compensation, this model outperforms VVC in subjective compression performance within a bit rate range of up to 0.4 bpp. Finally, the proposed generative coding model was deployed on mobile devices, verifying the

practical feasibility of this coding scheme on mobile devices.

(3) Addressing the challenge of incompatibility among different generative models on the decoding end, this thesis proposes a generative coding framework that integrates model prior knowledge. The encoder in this framework is not dependent on specific generative models and can extract more general compressed domain representations. Additionally, an adaptive feature mapping network is designed to achieve compatibility and adaptation between compressed domain representations and different generative models. Finally, by integrating compressed information into the generation process of pre-trained diffusion models, the model's prior knowledge is utilized to enhance compression performance. These designs make the proposed method a new generative coding method with stronger flexibility, compatibility, and iterative optimization capabilities. Furthermore, addressing the distortion problem in generative coding, a cross-attention-driven conditional fusion method, generative domain latent variable prediction enhancement technique, and statistical distribution-based image enhancement algorithm are proposed, significantly improving image reconstruction quality. Moreover, addressing the issue of information redundancy in the coding process and the lack of end-to-end joint optimization in the encoding and decoding process, a dual-stage optimization strategy of joint latent variable online optimization and encoder offline optimization is proposed. Finally, to address structural distortion in reconstructed images under ultra-low bit rates, a structural consistency optimization algorithm suitable for low bit rate coding is proposed by deeply analyzing the compressed domain latent variables, explicitly preserving and optimizing structural features within a unified compression framework, effectively reducing structural distortion in reconstructed images under ultra-low bit rate coding. Experimental results show that compared to the latest video coding standard VVC, the proposed method achieves an 87.75% performance improvement; compared to the latest end-to-end compression methods, it achieves a 78.35% performance improvement, demonstrating significant advantages of generative coding. Furthermore, experimental results show that the method proposed in this chapter can be compatible with other pre-trained generative models and exhibits good applicability to real-shot scenes and higher-resolution images.

In summary, this thesis focuses on generative coding methods for ultra-low bit rates, proposing a structural-textural layered generative coding framework and a semantic prior-guided layered generative coding method, achieving thousand-fold compression and high subjective quality reconstruction, and verifying the feasibility of this scheme on mobile devices. Furthermore, the research focuses on effectively utilizing prior knowledge of generative dif-

fusion models to construct a more flexible, easily iteratively optimized new generative coding method, exploring the performance potential of generative coding. The proposed coding scheme achieves an 87.75% compression performance improvement compared to the latest video coding standard VVC under ultra-low bit rate conditions, demonstrating the significant application potential of generative coding technology, driving the development of coding technology towards higher efficiency and better subjective visual experience, and contributing new ideas and methods to the development and research of future intelligent image and video compression technologies.

KEY WORDS: Image compression, generative models, extremely low bitrate, visual quality