摘要

视觉媒体数据已经成为当前信息社会中最主要的信息载体之一，随着高分辨率采集设备的普及以及流媒体应用规模的持续增长，图像与视频数据量不断增加，存储容量与传输带宽面临的压力也在逐渐提升。在压缩算法层面，编码框架已经从传统标准编码器逐渐发展到基于变分自编码器（Variational Autoencoder，VAE）的端到端神经编码器，并进一步扩展到基于扩散模型的生成式编码框架，率失真性能和感知重建能力均得到了一定改善，压缩系统对于视觉内容的表示与重建能力也随之得到了提升。在这一过程中，如何对编码过程进行更加稳定和精细的质量控制，并在不同场景下稳定达到目标质量，开始成为视觉压缩领域新的研究重点。

目前，码率控制算法已经经过了较长时间的发展，在实际系统中也已经较为成熟，但围绕质量目标的控制研究仍然相对有限。随着视觉压缩应用场景不断增加，质量可控性的不足开始逐渐影响实际系统的使用效果。大规模云转码系统通常需要保证输出结果稳定满足设定的感知质量标准，自适应码率流媒体需要依赖稳定的质量结果构建码率阶梯，而面向海量用户生成内容的在线分发场景，也要求不同来源、不同内容复杂度的视频在压缩后能够达到平台规定的最低质量要求。在远程医疗、遥感观测等场景中，压缩结果还需要尽可能保留视觉细节，以支持后续的影像分析与判读。因此，编码系统不仅需要提高压缩效率，还需要能够围绕给定质量目标进行稳定调节。与此同时，质量评价指标也逐渐从传统的像素级客观质量，扩展到以视频多方法评估融合（Video Multi-method Assessment Fusion，VMAF）为代表的感知质量指标，以及以核初始距离（Kernel Inception Distance，KID）为代表的分布级语义质量指标。随着评价指标不断接近人类视觉感知特性，不同质量层次之间的控制差异也开始变得更加明显。

本文围绕视觉压缩中的质量控制问题展开研究，相关工作从编码前的参数决策逐步延伸到编码过程中的质量调控，并进一步扩展到生成式编码框架下的语义质量保障，主要工作和贡献如下。

（1）**基于量化失真建模与预分析的压缩信号质量控制方法。**以峰值信噪比（Peak Signal-to-Noise Ratio，PSNR）作为质量度量，本文研究了编码前根据目标质量反向确定编码参数的质量精确达标控制问题。理论分析发现，传统编码器与变码率 VAE 编码器中的量化参数与 PSNR 之间都存在较稳定的线性关系，二者在量化机制上也具有一定统一性。在此基础上，本文提出了一种基于预分析的客观质量控制框架，通过轻量级网络在编码器外部完成样本自适应的参数估计，进而建立了从目标 PSNR 到编码参数之间的高效映射。整个框架采用编码器无关的预分析设计，因此能够适配不同类型

的异构编码架构。实验结果表明，该方法在多个数据集上的平均质量控制误差低于 2%，并且在从端到端 VAE 编码器向传统编码器 x265 的迁移过程中仍然保持了较好的有效性。

（2）**基于预分析建模和渐进式拟合的压缩感知质量控制方法。**在客观信号质量控制的基础上，本文进一步将控制目标扩展到与人眼感知更加一致的感知融合指标 VMAF。由于 VMAF 本身具有明显的非线性特征，同时还涉及多种质量特征的融合建模，传统方式难以直接建立稳定映射关系。针对这一问题，本文提出了三阶段渐进式控制算法，通过逐步预测 VMAF 子指标和最终质量结果，隐式建立样本自适应的量化—质量关系并推导编码参数。同时结合两遍反馈修正机制，对量化参数整数离散化带来的控制误差进行进一步补偿。实验结果表明，该算法在多个数据集上的最低平均控制误差达到 0.330%，能够实现较为精确的感知质量控制。

（3）**基于生成式模型和隐空间解耦的压缩语义质量控制方法。**当质量控制目标进一步扩展到语义层面后，压缩系统不仅需要维持像素保真和感知一致性，还需要尽可能保证场景结构、目标类别以及关键内容在重建结果中的可辨识性。强码率约束会进一步放大语义质量需求与传统编码机制之间的矛盾，现有压缩算法仅依赖编码参数调节已经难以继续维持稳定的语义质量。为此，本文提出了基于扩散变换器（Diffusion Transformer，DiT）先验的统一生成式压缩框架，并利用潜在表示分组策略对空间语义与时域运动进行解耦。在此基础上，本文利用深度压缩模块实现帧内锚点隐变量的极致压缩，同时利用统一潜在去噪模块在零额外码率条件下完成帧内语义增强与帧间运动合成。实验结果表明，在以 KID 作为核心语义质量指标，并以学习型感知相似度（Learned Perceptual Image Patch Similarity，LPIPS）和深度图像结构纹理相似度（Deep Image Structure and Texture Similarity，DISTS）作为辅助参考指标的条件下，该方法在极低码率区间仍然能够将语义质量稳定控制在可接受范围内，其性能明显优于现有方法。

本文的三项工作均围绕视觉压缩中的质量控制问题展开，整体上形成了从解析关系建模、渐进式质量拟合到生成式先验约束的递进式控制思路，并分别对应客观信号质量、感知质量以及语义质量条件下复杂程度不断提升的控制任务。相关方法从不同层面对压缩过程中的质量稳定性与目标达成能力进行了研究，也进一步扩展了视觉压缩在质量约束场景下的控制方式。

关键词：图像视频压缩；质量控制；渐进拟合；生成式模型；隐空间解耦

Research on Image and Video Compression Quality Control Algorithms

Shaokang Wang (Computer Applied Technology)
Directed by Xiaodong Xie

ABSTRACT

Visual media data has become one of the most important carriers of information in modern society. With the widespread adoption of high-resolution acquisition devices and the continuous growth of streaming media applications, the volume of image and video data continues to increase, placing growing pressure on storage capacity and transmission bandwidth. At the compression algorithm level, compression frameworks have evolved from traditional standardized codecs to end-to-end neural codecs based on Variational Autoencoders (VAEs), and further toward generative coding frameworks based on diffusion models. Rate-distortion performance and perceptual reconstruction capabilities have improved to some extent, enhancing the ability of compression systems to represent and reconstruct visual content. In this context, how to achieve more stable and fine-grained quality control during the encoding process and consistently attain target quality across different application scenarios has gradually become an important research focus in the field of visual compression.

Currently, bitrate control algorithms have undergone considerable development and are relatively mature in practical systems. However, research on quality-oriented control remains relatively limited. As visual compression applications continue to expand, the lack of quality controllability is gradually affecting the usability of practical systems. Large-scale cloud transcoding systems typically need to ensure that the output consistently satisfies predefined perceptual quality requirements. Adaptive bitrate streaming relies on stable quality performance to construct bitrate ladders. In addition, online distribution platforms for massive user-generated content require videos from different sources and with varying content complexity to meet minimum quality requirements after compression. In scenarios such as telemedicine and remote sensing, compressed results also need to preserve visual details as much as possible to support subsequent image analysis and interpretation. Therefore, compression systems need not only to improve compression efficiency but also to provide stable adjustment

toward given quality objectives. Meanwhile, quality evaluation metrics have gradually expanded from traditional pixel-level objective metrics to perceptual quality metrics represented by Video Multi-method Assessment Fusion (VMAF) and distribution-based semantic quality metrics represented by Kernel Inception Distance (KID). As quality evaluation metrics continue to move closer to human visual perception, the differences in control requirements across different quality levels are becoming increasingly pronounced.

This thesis focuses on the quality control problem in visual compression, extending from parameter decision-making before encoding to quality regulation during the encoding process, and further to semantic quality assurance under the generative coding framework. The main work and contributions are as follows.

(1) **Signal quality control based on quantization distortion modeling and pre-analysis**. Using Peak Signal-to-Noise Ratio (PSNR) as a quality metric, this thesis investigates the problem of precise quality control by inversely determining coding parameters based on the target quality before encoding. Theoretical analysis reveals a relatively stable linear relationship between quantization parameters and PSNR in both traditional codecs and variable VAE-based codecs, and both exhibit a degree of uniformity in their quantization mechanisms. Based on this, this thesis proposes an objective quality control framework based on pre-analysis. This framework uses a lightweight network to perform sample-adaptive parameter estimation outside the encoder, thereby establishing an efficient mapping from the target PSNR to the coding parameters. The entire framework employs an encoder-independent pre-analysis design, thus adapting to different types of heterogeneous coding architectures. Experimental results show that the proposed method achieves an average quality control error of less than 2% on multiple datasets and maintains good effectiveness during the migration from an end-to-end VAE codec to a traditional x265 codec.

(2) **Perceptual quality control based on pre-analysis modeling and progressive fitting**. Building upon objective signal quality control, this thesis further extends the control objective to the perceptual fusion metric VMAF, which is more consistent with human visual perception. Due to the significant nonlinear characteristics of VMAF and its involvement in the fusion modeling of multiple quality features, traditional methods struggle to directly establish a stable mapping relationship. To address this issue, this thesis proposes a three-stage progressive control algorithm. By progressively predicting VMAF sub-indices and the final quality result, it implicitly establishes a sample-adaptive quantization-quality relationship and derives the encoding parameters. In addition, a two-pass feedback correction mechanism fur-

ther compensates for the control error caused by the integer discretization of the quantization parameters. Experimental results show that the algorithm achieves a minimum average control error of 0.330% on multiple datasets, demonstrating relatively accurate perceptual quality control.

(3) **Semantic quality control based on generative models and latent space disentanglement**. When quality control objectives are further extended to the semantic level, compression systems not only need to maintain pixel fidelity and perceptual consistency, but also need to ensure the recognizability of scene structure, target categories, and key content in the reconstruction results as much as possible. Strong bitrate constraints further amplify the contradiction between semantic quality requirements and traditional coding mechanisms, and existing compression algorithms can no longer maintain stable semantic quality by relying solely on coding parameter adjustments. To address this, this thesis proposes a unified generative compression framework based on the Diffusion Transformer (DiT) prior, and utilizes a latent representation grouping strategy to decouple spatial semantics from temporal motion. Based on this, this thesis uses a deep compression module to achieve extreme compression of intra-frame anchor latent variables, and a unified latent denoising module to complete intra-frame semantic enhancement and inter-frame motion synthesis under zero extra bitrate conditions. Experimental results show that, with Kernel Inception Distance (KID) as the core semantic quality indicator and Learned Perceptual Image Patch Similarity (LPIPS) together with Deep Image Structure and Texture Similarity (DISTS) as auxiliary reference indicators, this method can still keep the semantic quality stable within an acceptable range in the extremely low bitrate range, and its performance is significantly better than existing methods.

All three studies in this thesis all focus on the quality control problem in visual compression, forming a progressive control approach from analytical relation modeling and incremental quality fitting to generative prior constraints. These approaches correspond to increasingly complex control tasks under objective signal quality, perceptual quality, and semantic quality conditions. From different perspectives, the proposed methods investigate the stability of quality regulation and the capability of achieving target quality during the compression process, while further extending the control paradigms of visual compression under quality-constrained scenarios.

KEYWORDS: Image and Video Compression; Quality Control; Progressive Fitting; Generative Models; Latent Decoupling