摘要

近年来，人工智能领域快速发展，基于深度学习的视频编解码算法逐渐成为多媒体领域的研究热点。随着视频通话、网络会议、在线直播等应用的普及，针对以人物为核心的视频进行高效压缩以满足低延迟、低带宽需求已成为亟待解决的问题。然而，目前主流的深度视频压缩方法遵循传统混合编码框架，模块设计复杂，编解码效率较低，且在低码率下解码视频的主观视觉质量较差。另外，人物视频包含人脸视频与人体视频，二者的分布特性与自然场景中的其他种类视频有较大差别，针对人物视频的压缩方法仍缺乏深入的研究。随着深度生成网络的性能不断提升，现有的生成式模型可以基于人物的高级特征表示合成高质量的人物图像和视频。本文以深度生成模型为基础，探究人物视频的高效压缩方法，分别从人物运动关键信息、人体分层特征两个角度进行研究，并利用深度生成模型构建可实际应用的人物视频压缩系统。本文主要工作包含以下三个方面：

- **提出了基于运动关键信息的生成式人物视频压缩方法**。针对人物生成模型中运动特征存在冗余的问题，提出了一种基于运动关键信息的生成式人物视频压缩框架。通过运动估计网络得到源视频中人物关键区域的运动变换矩阵，提取具有不同重要程度运动特征表示。分别对运动表示应用不同的量化步长和熵模型，保留人物主要运动信息并过滤次要运动信息，降低运动信息冗余。将解压缩后的运动表示与输入关键帧通过深度生成网络进行融合，按顺序重建得到解码视频序列。在不影响解码视频主观质量的前提下，本方法节省了传输运动特征时所需码率的 65%。
- **提出了基于分层特征的生成式人体视频压缩方法**。针对低码率下人体视频高效压缩的难题，提出了一种基于人体分层特征融合的生成式编解码框架。通过将源视频分解为结构和纹理特征表示，实现对视频中人体的建模。提取纹理信息时，只采用关键帧的特征信息作为全局的纹理特征表示，降低编码开销。通过两阶段的训练方式，使生成模块得以利用从生成帧到当前重建帧的光流信息，得到更加精细的重建结果。为保证解码视频的纹理一致性，提出了使用语义层面的对比学习方法；为保证解码视频的时序一致性，提出了融合生成帧光流信息的标准化模块。实验结果表明，该方法在同码率下与主流视频编码方法 VVC 相比，重建主观质量提升了 50%。
- **构建生成式人物视频压缩系统**。基于所提出的人物视频压缩方法，构建了深度生成式人物视频压缩系统，使用包含视频通话、在线直播等多种应用场景下的

自采集数据集进行训练与测试。通过测试不同条件下人物视频的压缩效果，并与主流编解码框架进行对比，进一步论证了深度生成模型对于人物视频编码的有效性与可行性。

综上所述，本文从深度生成模型出发，从运动关键信息生成和分层特征生成两方面进行研究，构建端到端生成式人物视频压缩框架。提出的方法在对高级特征信息（如运动特征、结构特征、语义特征等）进行高效编码的同时，进一步提升了解码视频的重建质量，为生成式视频编码的应用提供了新思路。

关键词：视频生成，深度视频编码，生成-对抗网络，对比学习

Deep Generative Model Based Human Video Coding

Ruofan Wang (Computer Application Technology)
Directed by: Prof. Ronggang Wang and Prof. Siwei Ma

ABSTRACT

In recent years, the field of artificial intelligence has experienced rapid development, resulting in the emergence of deep-learning-based video coding algorithms as a research hot-spot in the area of multimedia. As video calls, online conferences, and live streaming have become increasingly popular, the demand for efficient video compression techniques that accommodate low-latency and low-bandwidth requirements has become urgent. However, mainstream deep video compression methods follow a traditional hybrid coding framework that is characterized by complex module designs, low encoding and decoding efficiency, and poor subjective visual quality of decoded videos at low bitrates. Moreover, Human videos encompass both facial and body videos, which have distribution characteristics that differ significantly from those of other types of videos in natural scenes. Consequently, there remains a lack of in-depth research on compression methods for human-oriented videos. With the development of deep generative networks, generative models can now synthesize realistic images and videos of people based on high-level feature representations. This thesis explores efficient compression for human-oriented videos from two perspectives, namely, the key information of human motion and the layered features of the human body, building a practical human video compression system using deep generative models. The contributions of this thesis can be summarized in three aspects:

- **A generative video compression method for human videos based on key motion information is proposed.** A generative video compression framework for human videos is proposed to tackle the issue of redundant motion feature representation in human generation models. The encoder obtains the motion transformation matrix of the key regions of people in the source video using a motion estimation network. The matrix is then subjected to principal component analysis to extract motion feature representations of different importance levels. By applying different quantization step sizes and entropy models to the motion representation, minor motion information can be filtered out while retaining the core motion information of the person, thereby reducing the redundancy of motion information. The decompressed motion representation is fused with the input key frame by a deep generative network, and the decoded video sequence is reconstructed sequentially. The proposed method saves 65% of the required bit rate for transmitting motion features without compromising the subjective quality of the decoded video.

- **A generative video compression method for human body videos based on hierarchical feature is proposed.** A generative video coding framework is proposed for efficient compression of human body videos at low bitrates. The framework leverages hierarchical feature fusion to achieve texture consistency in reconstructed videos. Specifically, the source video is decomposed into structural and texture feature representations, with only key frame feature information used as the global texture feature representation to reduce encoding costs. Through a two-stage training process, the generative module utilizes optical flow information from generated frames to achieve refined reconstruction results. A contrastive learning method at the semantic level is proposed to ensure the texture consistency of the decoded video, and a normalization module that fuses optical flow information of the generated frames is proposed to ensure the temporal consistency of the decoded video. Experimental results show that the proposed method outperforms the mainstream video coding method VVC in subjective quality of reconstructed videos by 50%.
- **A generative human video compression system is constructed.** Based on the proposed video compression method for human videos, a deep generative video compression system was developed. The system was trained and tested on self-collected datasets containing various application scenarios, such as video calling and live streaming. The compression performance of human body videos was evaluated under different conditions and compared with mainstream encoding-decoding frameworks. The results further demonstrate the effectiveness and feasibility of the deep generative model for encoding human body videos.

In summary, this thesis presents an end-to-end deep generative video compression framework for human videos based on deep generative models, which is studied from the perspectives of generated motion key information and hierarchical features. The proposed method efficiently encodes the extracted high-level feature information, such as motion features, structural features, and semantic features, while simultaneously enhancing the reconstruction quality of decoded videos. This approach provides novel insights for the application of generative video coding frameworks.

KEY WORDS: Video Generation, Deep Learning Based Video Coding, Generative Adversarial Network, Contrastive Learning