摘要

随着虚拟现实、数字人通信和自由视角视频等沉浸式应用的快速发展，三维人体数据已成为支撑高临场感内容生成、交互呈现与远程通信的重要基础。与此同时，采集与重建技术的持续演进推动三维人体数据在空间分辨率、时间长度和视角维度上持续增长，使其传输与存储面临显著压力。三维人体数据的常见表示形式包括点云、网格和多视角视频等，其中点云、网格等表示形式因能够灵活刻画人体复杂的非刚性表面而被广泛采用，多视角视频则因能够直接支撑自由视角下的沉浸式观看体验，成为面向用户端呈现的重要载体。如何在有限带宽和存储资源条件下实现三维人体数据的高效编码与高质量重建，已成为沉浸式媒体处理中的关键研究问题。

传统压缩方法主要通过去除空间、时间和视角维度上的信号冗余来提升编码效率。然而，对于具有复杂拓扑结构、非刚性运动特性和稳定语义规律的三维人体数据，这类方法由于未能充分利用三维人体在空间结构、时域运动、外观语义以及跨视角观测等方面所蕴含的内在规律与稳定约束，在低码率或信息缺失条件下往往容易出现结构失真、时序漂移和属性退化等问题，难以同时兼顾编码效率、结构一致性与感知质量。针对上述挑战，本文围绕先验驱动的三维人体编码方法开展系统研究，分别从参数化结构表示、动作演化建模、纹理生成重建以及人体中心场景解耦表示四个层面出发，面向静态人体结构、动态人体结构、动态人体纹理以及人体中心多视角场景构建紧凑编码表示，并在解码端结合生成式重建机制实现高质量数据恢复。具体而言，本文的主要研究工作总结如下：

（1）提出了基于参数化模型的静态人体结构编码方法。针对高分辨率静态人体点云几何数据在低码率条件下易出现结构失真、局部细节难以稳定保留的问题，本文利用人体形态具有全局结构一致性和低维可描述性的特点，引入参数化人体模型对静态人体结构进行建模。通过将高维点云几何映射到低维参数空间，并对语义主体与局部几何残差进行分层编码，该方法在编码端实现了静态人体结构的紧凑表示，并在解码端结合生成式重建机制恢复人体几何细节。实验结果表明，该方法能够在显著降低编码开销的同时，有效保持人体结构的完整性与语义一致性。

（2）提出了基于动作建模的动态人体结构编码方法。针对动态人体点云几何序列中帧间冗余高、传统逐帧编码方式难以保持时序一致性的问题，本文从人体动作演化过程具有连续性、协调性和可预测性的特点出发，对人体形体信息与姿态变化进行解耦表示，将姿态变化编码为紧凑动作参数，并以此建立跨帧结构预测关系，从而大幅削减帧间冗余。该方法以动作建模为核心实现动态人体结构的紧凑编码，并在解码端通过生成式恢复抑制结构漂移和时序失真。实验结果表明，该方法能够有效降低动态

人体结构中的时域冗余，在提升重建连贯性的同时提高动态几何序列的编码效率。

（3）提出了基于扩散模型的动态人体纹理编码方法。针对动态人体点云属性数据在时间维度上冗余度高、传统逐帧属性编码方式易导致纹理模糊、细节退化和时域闪烁的问题，本文结合人体外观在区域语义分布和帧间演化上的稳定性，引入扩散模型构建面向动态人体纹理的生成式编码机制。该方法将动态纹理信息压缩为紧凑条件表示，并在解码端借助扩散模型以姿态与几何约束为引导完成高质量纹理重建，通过结构约束抑制帧间生成的不一致性。实验结果表明，该方法能够在较低码率下有效保持动态人体外观的纹理细节与时序稳定性，显著提升属性重建结果的感知质量。

（4）提出了基于解耦高斯表示的人体中心场景编码方法。在前述工作所建立的参数化人体表示与紧凑结构编码能力的基础上，本文进一步将研究对象由人体主体扩展到真实场景中的人体中心多视角视频。针对多视角场景中背景区域相对稳定而人体主体高度动态、传统统一建模方式难以同时兼顾压缩效率与自由视角重建质量的问题，本文构建了人体主体与场景背景解耦的高斯表示框架。其中，背景由稳定的场景表示建模，动态人体主体则通过紧凑的人体驱动表示进行描述，从而充分利用不同组成部分在时空统计特性上的差异，实现多帧、多视角数据的高效编码与解码端高质量重建。实验结果表明，该方法能够有效减少多视角视频中的冗余信息传输，在提升编码效率的同时，保障自由视角条件下人体主体的结构清晰度和运动连贯性。

综上，本文围绕三维人体数据在空间结构、动作演化、纹理表达和场景观测四个层面的先验规律，系统研究了静态人体结构、动态人体结构、动态人体纹理以及人体中心场景的编码问题，形成了一条由静态结构表示、动态动作建模、纹理生成重建到跨视角场景解耦表示的递进式技术路线。本文研究表明，充分挖掘三维人体先验并结合解码端生成式重建机制，有助于弥补传统信号级压缩方法在结构保持、时序一致性和感知质量方面的不足，为沉浸式场景下三维人体数据的高效传输、重建与应用提供了新的研究思路和技术支撑。

关键词：三维人体数据，先验驱动编码，动态建模，生成式重建，人体中心场景

Research on Prior-driven Compression Methods for 3D Humans

Ruoke Yan (Computer Application Technology)
Supervised by Prof. Siwei Ma

ABSTRACT

With the rapid development of immersive applications such as virtual reality, digital human communication, and free-viewpoint video, 3D human data has become a fundamental basis for high-presence content generation, interactive presentation, and remote communication. Meanwhile, continual advances in acquisition and reconstruction technologies have driven sustained growth in the spatial resolution, temporal duration, and viewpoint scale of 3D human data, imposing increasing pressure on transmission and storage. Common representations of 3D human data include point clouds, meshes, and multi-view videos. Point clouds and meshes are widely adopted for their flexibility in describing complex non-rigid human surfaces, while multi-view videos have become an important medium for user-facing presentation due to their ability to directly support immersive free-viewpoint viewing experiences. Therefore, efficient encoding and high-quality reconstruction of 3D human data under limited bandwidth and storage constraints have become key research problems in immersive media processing.

Traditional compression methods mainly improve coding efficiency by removing signal redundancy in the spatial, temporal, and viewpoint dimensions. However, 3D human data contains complex topological structures, non-rigid motion patterns, and stable semantic regularities. Since such methods do not fully exploit the intrinsic regularities and stable constraints of 3D humans in spatial structure, temporal motion, appearance semantics, and cross-view observations, they are prone to structural distortion, temporal drift, and attribute degradation under low-bitrate or incomplete-information conditions, making it difficult to simultaneously preserve coding efficiency, structural consistency, and perceptual quality. To address these challenges, this dissertation presents a systematic study of prior-driven encoding methods for 3D human data. Specifically, it investigates four aspects: parametric structural representation, motion evolution modeling, texture generation and reconstruction, and decoupled representation of human-centric scenes. Based on these perspectives, compact encoding representations are developed for static human structures, dynamic human structures, dynamic human textures,

and human-centric multi-view scenes. High-quality data recovery is further achieved through generative reconstruction mechanisms at the decoder side.

First, a parametric-model-based encoding method for static human structures is proposed. High-resolution static human point cloud geometry is prone to structural distortion at low bitrates, and its local details are difficult to preserve stably. To address this problem, this method exploits the global structural consistency and low-dimensional describability of human morphology, and introduces a parametric human model for static structure representation. By mapping high-dimensional point cloud geometry into a low-dimensional parameter space and hierarchically encoding the semantic body and local geometric residuals, the proposed method achieves a compact representation of static human structures at the encoder. At the decoder side, a generative reconstruction mechanism is employed to restore fine geometric details. Experimental results demonstrate that the proposed method significantly reduces the coding cost of human structures while effectively preserving structural integrity and semantic consistency.

Second, a motion-modeling-based encoding method for dynamic human structures is proposed. Dynamic human point cloud geometry sequences contain substantial inter-frame redundancy, and conventional frame-by-frame coding methods struggle to maintain temporal consistency. To address this problem, this method starts from the continuity, coordination, and predictability of human motion evolution. It decouples body shape information from pose variation and encodes pose changes into compact motion parameters. These parameters are then used to establish cross-frame structural prediction relationships, thereby substantially reducing inter-frame redundancy. Centered on motion modeling, the proposed method enables compact encoding of dynamic human structures, and suppresses structural drift and temporal distortion through generative reconstruction at the decoder side. Experimental results show that the proposed method effectively reduces temporal redundancy in dynamic human structures and improves the coding efficiency of dynamic geometry sequences while enhancing reconstruction coherence.

Third, a diffusion-model-based encoding method for dynamic human textures is proposed. Dynamic human point cloud attributes exhibit high temporal redundancy, and conventional frame-by-frame attribute coding often leads to texture blur, detail degradation, and temporal flickering. To address these problems, this method leverages the stability of human appearance in regional semantic distribution and inter-frame evolution, and introduces a diffusion model to build a generative encoding framework for dynamic human textures. The proposed method compresses dynamic texture information into compact conditional representations. At the decoder side, high-quality textures are reconstructed using pose and geometry constraints

as guidance, and structural constraints are applied to suppress inter-frame generation inconsistencies. Experimental results demonstrate that the proposed method effectively preserves texture details and temporal stability of dynamic human appearance at lower bitrates, significantly improving the perceptual quality of attribute reconstruction.

Finally, a decoupled-Gaussian-representation-based encoding method for human-centric scenes is proposed. Building upon the parametric human representation and compact structural encoding capabilities established in the preceding studies, this work further extends the research scope from human representations to human-centric multi-view videos in real-world scenes. In multi-view scenes, the background is relatively stable whereas the human is highly dynamic, making it difficult for conventional unified modeling methods to simultaneously achieve high compression efficiency and high-quality free-viewpoint reconstruction. To address this problem, this method constructs a decoupled Gaussian representation framework for the human and the scene background. The background is modeled with a stable scene representation, while the dynamic human is described by a compact human-driven representation. By exploiting the differences in spatiotemporal statistical properties between these components, the proposed method enables efficient encoding of multi-frame and multi-view data and achieves high-quality reconstruction at the decoder side. Experimental results show that the proposed method effectively reduces redundant information transmission in multi-view videos, improving coding efficiency while preserving the structural clarity and motion coherence of the human under free-viewpoint rendering.

In summary, this dissertation systematically investigates the encoding of static human structures, dynamic human structures, dynamic human textures, and human-centric scenes by exploiting prior regularities of 3D human data in spatial structure, motion evolution, texture representation, and scene observation. A progressive technical route is established, spanning static structural representation, dynamic motion modeling, texture generative reconstruction, and decoupled cross-view scene representation. The results of this dissertation demonstrate that fully exploiting 3D human priors and incorporating generative reconstruction mechanisms at the decoder side can help compensate for the limitations of traditional signal-level compression methods in structural preservation, temporal consistency, and perceptual quality. These findings provide new research insights and technical support for the efficient transmission, reconstruction, and application of 3D human data in immersive scenarios.

KEY WORDS: 3D human data, Prior-driven encoding, Motion modeling, Generative reconstruction, Human-centric scenes