摘要

随着元宇宙、数字孪生等技术的迅速发展，社交、娱乐、教育等领域都经历了革命性的变革。传统的物理空间不再是人们交流、学习和工作的唯一场所，而是与虚拟空间相互融合，共同构建出一个混合现实世界。在这种虚实融合的环境中，人类动作的真实化、自然化呈现至关重要。单目三维人体姿态估计技术提供了一种低成本、高效的方法来实现这一目标。该技术旨在从单张图像或单个视频中定位人体关节点在三维空间中的位置，以获取人体动作表征。这一任务有助于增强对人体运动信息的理解，进而辅助创建出高度逼真的数字化形象，为各种应用场景提供更真实、更生动的体验。

然而，现有三维人体姿态估计算法依赖场景、动作类别、样本数目受限的数据集进行训练，它们难以应对实际场景中的复杂情况，例如自遮挡、复杂的前背景关系以及罕见的动作等。本文将围绕单目三维人体姿态估计从学术研究转化为实践应用中遇到的鲁棒性、泛化性和自适应性问题展开研究。本文的主要工作可以总结为以下四个方面：

（1）针对现有三维姿态估计模型对裁切不鲁棒的问题，提出基于相对信息编码的三维人体姿态估计方法。相对信息编码可分为位置信息编码与时域信息编码。前者利用二维姿态的相对坐标来编码位置信息，以增强输入和输出分布之间的一致性。具有不同绝对二维位置的相同姿势可以映射到同一个共有表示。后者通过建立一段时间内当前姿势和其他姿势之间的联系来编码时域信息，这有助于模型关注当前姿势前后的运动变化。此外，还提出一种多阶段优化方法对整体框架进行训练。上述对位置和时域信息的编码方式有利于提升鲁棒性，抵抗全局运动对预测结果的干扰，同时能提升局部小范围运动的预测性能。在多个数据集上的实验表明，所提出方法相较同期最优方法在三维人体姿态估计任务上的整体性能提高 6.3%。

（2）针对现有方法难以泛化到真实场景数据上的问题，提出基于自监督预训练的三维人体姿态估计方法。本文利用丰富的二维数据来辅助三维姿态估计的学习，将训练过程分为预训练和微调两个阶段。在预训练阶段，对二维姿态进行随机遮挡，并让网络通过自监督的方式对原始动作序列进行恢复，以增强模型学习动作时空关联的能力。在微调阶段，对前一阶段训练完成的编码器进行微调，并结合一个多到一帧聚合模块进行三维姿态预测的学习。此外，还提出一种时域下采样策略，有效地降低数据冗余。所提出方法通过第一阶段的预学习有效降低二阶段的优化难度，从而提升三维人体姿态估计的准确性与泛化性。在多个数据集上的实验结果显示，所提出方法相较同期最优方法在三维人体姿态估计任务上的整体性能提高 32.0%。

（3）针对现有方法难以自适应地聚合多个三维姿态假设的问题，提出基于扩散模型的姿态假设生成与关节级姿态假设聚合方法。基于扩散模型的假设生成方法从高斯分布中进行采样并去噪，以获取多个准确度、多样性更强的三维姿态假设。关节级姿态假设聚合方法以三维姿态假设的重投影和二维输入关键点之间的差异作为引导信息，通过最小化重投影误差或者使用一个假设选择网络来获取最终的输出结果。所提出方法使用更小的粒度进行假设聚合，提升了姿态估计的自适应性。在多个数据集上的实验结果显示，所提出方法相较同期最优方法在单姿态输出与多姿态输出三维人体姿态估计任务上的整体性能分别提高 10.6% 与 10.5%。

（4）搭建了一套数字人动作迁移系统与一套中国古画“静转动”系统来对方法的有效性进行验证。在数字人动作迁移系统中，使用三维人体姿态估计方法提取人体姿态，将其迁移到虚拟人物模型上，并在两种全息平台上进行展示。在古画“静转动”系统中，使用三维人体姿态估计方法提取自定义视频中的动作数据，并将其用于驱动中国传统古画中的人物进行运动。

综上所述，本文以解决姿态估计中面临的鲁棒性、泛化性、自适应性问题为目的，分别提出三种三维人体姿态估计方法，有效提升了方法的准确性与效率，为相关领域的学术研究提供了新的思路。并将该方法用于数字人动作生成与沉浸式中国古典文物数字化创作中，推动了三维人体姿态估计任务的实际应用。

关键词：计算机视觉，三维人体姿态估计，鲁棒性，泛化性，自适应性