天津大学
Tianjin University

元宇宙数字人的前沿进展与未来

—— 2023年度数字人领域进展报告

李坤
天津大学智能与计算学部

背景介绍

- 数字人技术的目标
  - 对人的**外观**数字化实现**逼真性** ➡ 真实感数字人重建与生成技术
  - 对人的**行为**感知实现**精准性** ➡ 运动捕捉与运动生成技术
  - 与周围**环境**和人实现**交互性** ➡ 推理与交互生成技术

汇报提纲

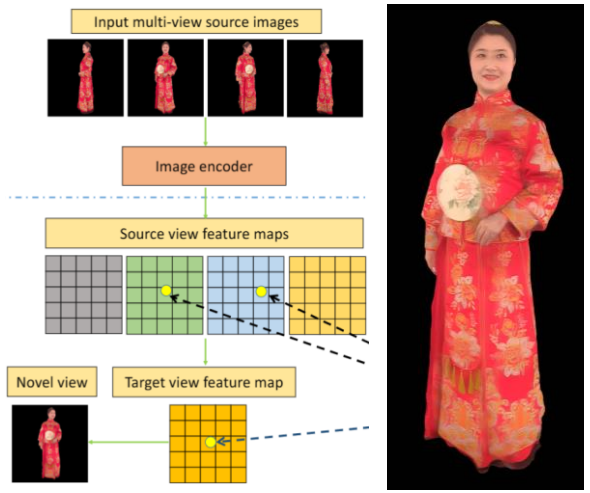
一 真实感数字人重建与生成

二 运动捕捉与运动生成

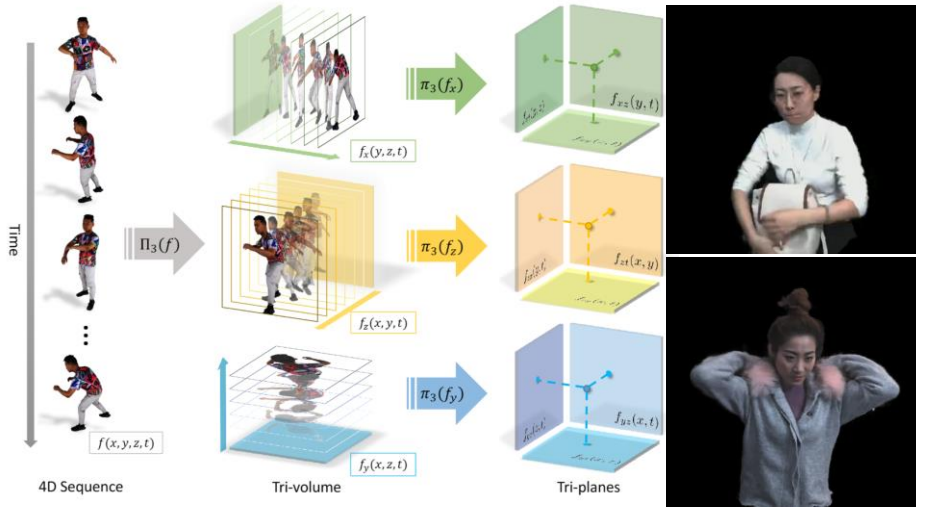
三 推理与交互生成

真实感数字人重建与生成：稀疏多视角

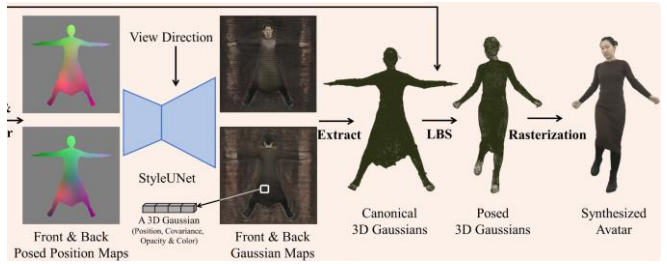
- 研究动机：视角间重叠少，导致传统方法不适用
- 解决思路：寻求同时易于回归或拟合和渲染的新表示
  - 隐式纹理 + 神经渲染：如HDhuman[1]，学习通用的编解码器，避免单人单训，实现静态和动态重建
  - NeRF + 三平面表示：如Tensor4D[2]，以三平面表示一般的4D-NeRF场，能以时序建模一般物体
  - 3D高斯溅射 + 变形场：如Animatable Gaussians[3]，学习参数化模板，解耦几何纹理与动作



HDhuman [TVCG 2023]



Tensor4D [CVPR 2023]

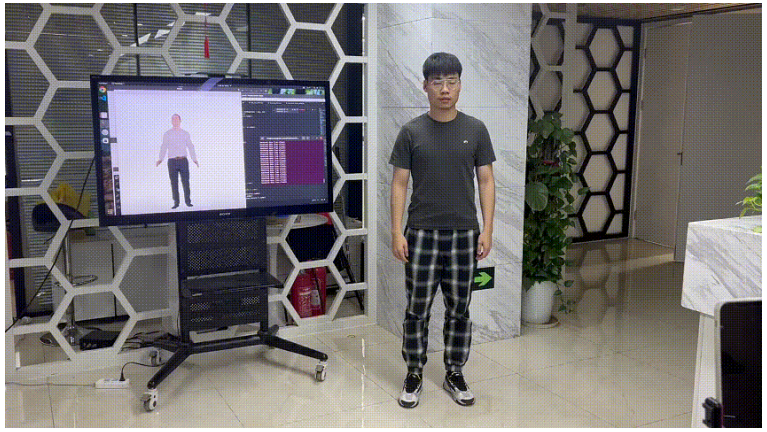


Animatable Gaussians [CVPR 2024]

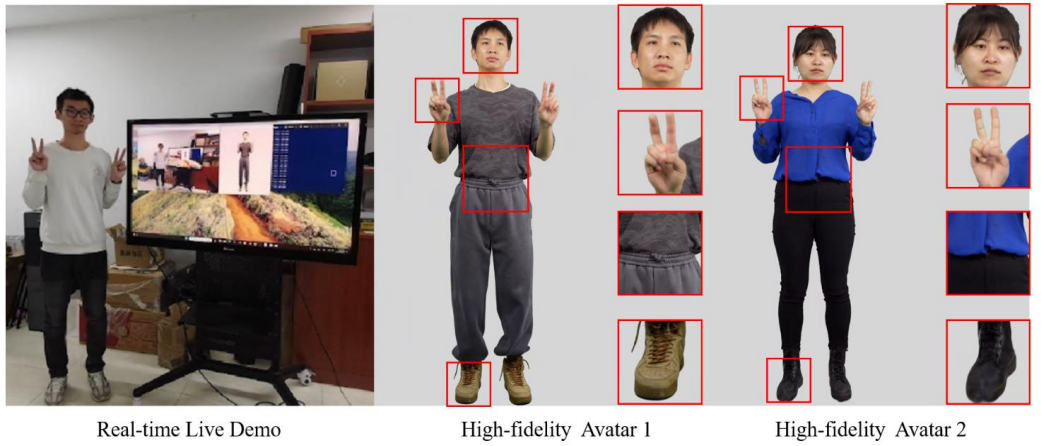
[1] Zhou et al. HDhuman: High-quality Human Novel-view Rendering from Sparse Views. IEEE TVCG 2023.
[2] Shao et al. Tensor4D : Efficient Neural 4D Decomposition for High-fidelity Dynamic Reconstruction and Rendering. CVPR 2023.
[3] Li et al. Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Human Avatar Modeling. CVPR 2024.

真实感数字人重建与生成：单目视频

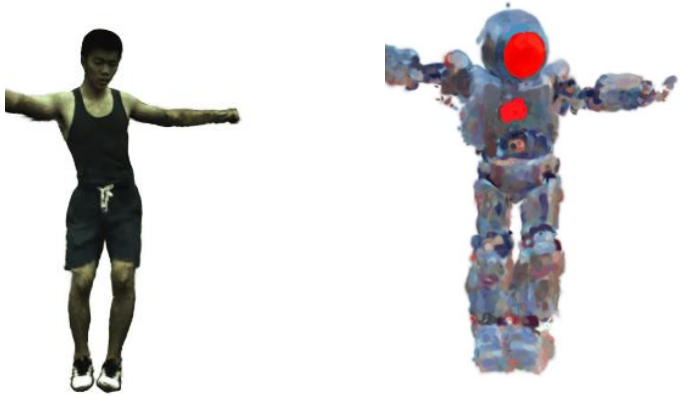
- 研究动机：单目视频输入下，无法获取准确的几何信息，导致难以建模大动作变化
- 解决思路：结合人体模板先验，同时建模姿态引导的非刚性人体形变
  - GaussianAvatar：通过UV位置图预测动态变化，联合优化姿态与纹理，解决单目姿态不准确问题
  - RAM-Avatar：提出双注意力模块和运动分布对齐模块，隐式建模非刚性运动，实现包括人脸人手在内的全身的动作驱动
  - GART：利用三维高斯结合可学习的蒙皮权重拟合人体的几何和外观，实现30s内的建模



GaussianAvatar



RAM-Avatar



GART

[1] Hu et al. GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians. CVPR 2024.
[2] Deng et al. RAM-Avatar: Real-time Photo-Realistic Avatar from Monocular Videos with Full-body Control. CVPR 2024.
[3] Lei et al. GART: Gaussian Articulated Template Models. CVPR 2024.

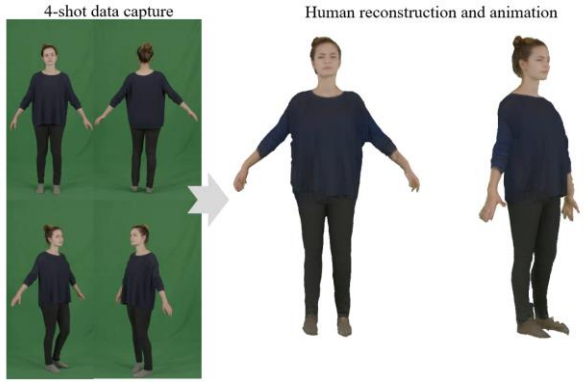
真实感数字人重建与生成：单图像 / Few Shot

- 研究动机：由于存在遮挡和不可见区域，现有方法容易得到模糊的结果且缺乏细节
- 解决思路：结合先验知识 / 设计合理的表示 / 利用预训练大模型
  - DINAR：结合神经纹理与SMPL-X模型，使用潜在扩散模型细化纹理，提高动画推断速度和鲁棒性
  - R²Human：设计Z-map新表示，结合隐式纹理场与显式神经渲染的优势，实现实时高质量纹理重建
  - TeCH：利用BLIP和DreamBooth微调大模型，结合DMTet表示，实现不可见区域的精细纹理重建
  - HAVE-FUN (Few Shot)：提出可驱动的DMTet表示，利用预训练Zero123，实现与图像一致的重建

DINAR [ICCV 2023]

R²Human [arXiv 2023]

TeCH [3DV 2024]



HAVE-FUN [CVPR 2024]

[1] Svitov et al. DINAR: Diffusion inpainting of neural textures for one-shot human avatars. ICCV 2023.
[2] Feng et al. R²Human: Real-Time 3D Human Appearance Rendering from a Single Image. arXiv 2023.
[3] Huang et al. TeCH: Text-guided reconstruction of lifelike clothed humans. 3DV 2024.
[4] Yang et al. HAVE-FUN: Human Avatar Reconstruction from Few-Shot Unconstrained Images. CVPR 2024.

真实感数字人重建与生成：衣服重建

- 研究动机：现有穿衣人体重建方法难以将服装与人体分离，从而无法实现换衣等应用
- 解决思路：利用人体先验和衣服相关的特征，建模人体运动和姿态对服装的形变
  - DGarment：结合人体先验和衣服特征，学习服装形变空间，实现单目视频下可驱动动态衣服重建
  - DiffAvatar：提出可微分的身体和服装协同优化的物理模拟方法，实现3D扫描数据高质量衣服重建
  - Neural-ABC：提出双层隐式场表示，建模人体和衣服低维隐空间，实现单张图像下可编辑衣服重建

DGarment [TCSVT 2023]

DiffAvatar [CVPR 2024]

Neural-ABC [TVCG 2024]

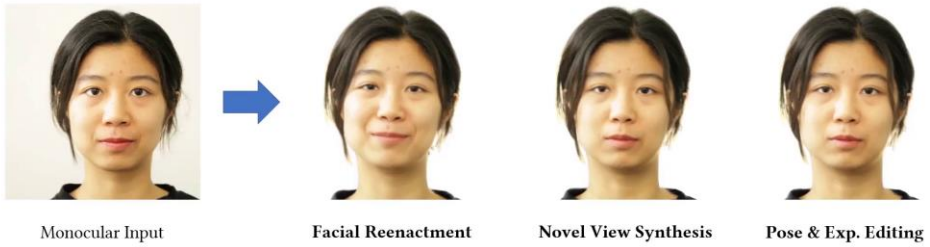
[1] Li et al. High-Quality Animatable Dynamic Garment Reconstruction from Monocular Videos. IEEE TCSVT 2023.

[2] Li et al. DiffAvatar: Simulation-Ready Garment Optimization with Differentiable Simulation. CVPR 2024.

[3] Chen et al. Neural-ABC: Neural Parametric Models for Articulated Body with Clothes. IEEE TVCG 2024.

真实感数字人重建与生成：人脸重建

- 研究动机：现有利用网格模板跟踪底层几何形状的方法无法学习复杂的几何细节，难以在稀疏视角甚至单视角下合成高保真的结果
- 解决思路：使用隐式函数学习头部的形状、表情或纹理表示，实现高保真重建
  - BakedAvatar：从单目视频中学习可变形的多层网格模型和纹理，实现实时渲染（甚至消费级设备）
  - HQ3DAvatar：从多视角视频中学习特征空间中的多分辨率哈希编码实现高质量渲染
  - Gaussian Head Avatar：从多视角视频中优化3D高斯函数和基于MLP的变形场来捕捉复杂表情



BakedAvatar [TOG 2023]

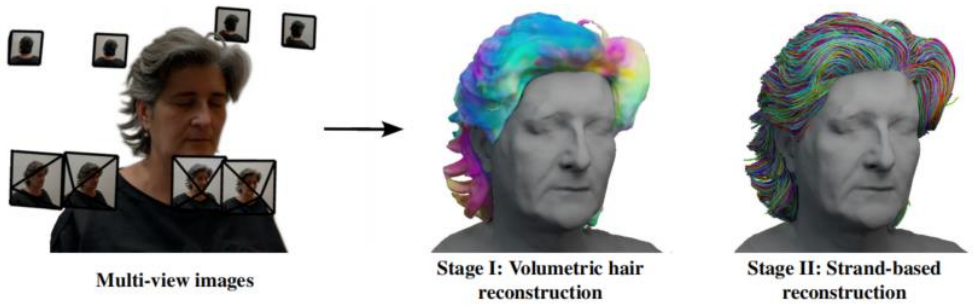
HQ3DAvatar [TOG 2024]

Gaussian Head Avatar [CVPR 2024]

[1] Duan et al. BakedAvatar: Baking Neural Fields for Real-Time Head Avatar Synthesis. ACM TOG 2023.
[2] Teotia et al. HQ3DAvatar: High Quality Implicit 3D Head Avatar. ACM TOG 2024.
[3] Xu et al. Gaussian Head Avatar: Ultra High-fidelity Head Avatar via Dynamic Gaussians. CVPR 2024.

真实感数字人重建与生成：头发重建

- **研究动机**：针对**复杂几何形状**的头发，现有方法依赖于**头发方向图**或**图像空间中的梯度，忽视**卷曲、光泽等**细节建模**，难以重建逼真数字头发
- **解决思路：头发股线的有效感知/表示 + 头发渲染器**
  - **Neural Haircut**：采用两阶段方法，即基于隐式体表示的粗糙头发重建 + 基于股线的精细头发，结合头发股线和发型先验，实现**单目视频或多视角图像**下的3D头发精确重建
  - **HairStep**：提出一种新的中间表示（包括股线谱和深度谱）以及对应的真实图像数据集，实现**单图像**下的3D头发真实建模
  - **GaussianHair**：利用一系列相连的三维高斯圆柱体表示每根头发股线，实现**多视角图像**下的头发的几何和外观重建，同时**支持头发的编辑、重光照和动态渲染**



Neural Haircut [ICCV 2023]

HairStep [CVPR 2023]

GaussianHair [arXiv 2024]

[1] Sklyarova et al. NeuralHaircut:Prior-GuidedStrand-BasedHairReconstruction. ICCV 2023.
[2] Zheng et al. HairStep: Transfer Synthetic to Real Using Strand and Depth Maps for Single-View 3D Hair Modeling. CVPR 2023.
[3] Luo et al. GaussianHair: Hair Modeling and Rendering with Light-aware Gaussians. arXiv 2024.

真实感数字人重建与生成：数字人生成

生成结果缺少3D结构一致性

3D数据集稀少且昂贵

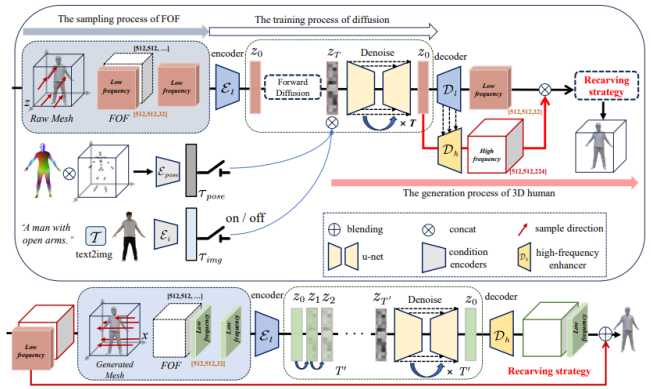
原生3D生成

2D升维生成

显式表达

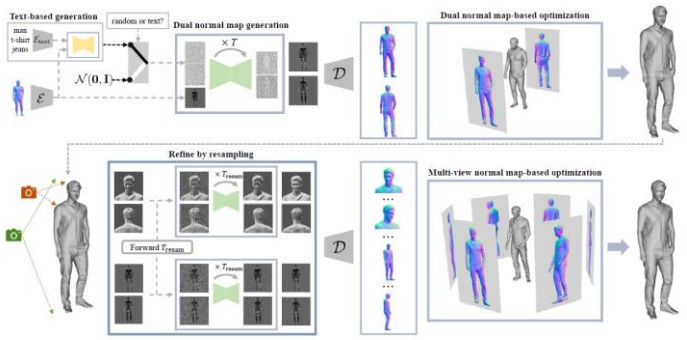
隐式表达

紧致高效表达，建模3D数据分布



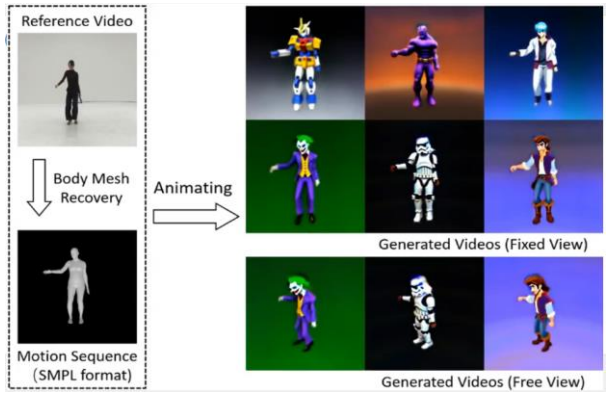
GETAvatar, Joint2Human…

充分利用2D先验知识



Chupa, Get3DHuman, HumanNorm…

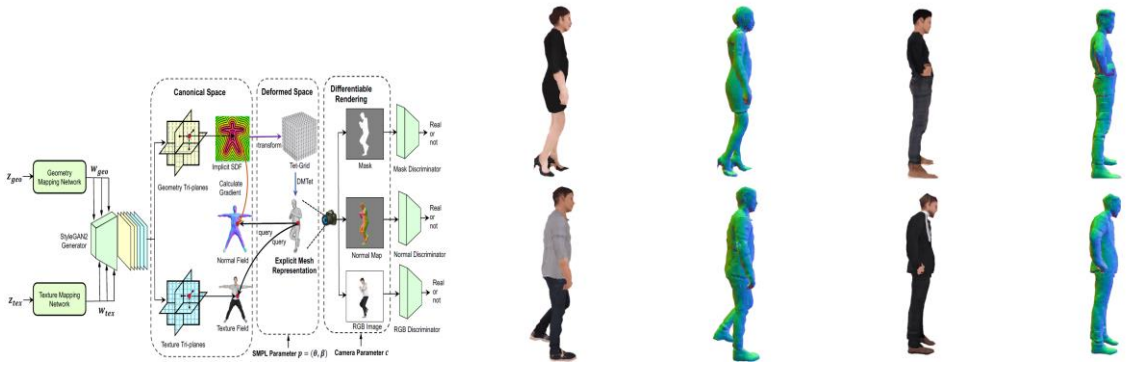
多条件约束，注入人体先验



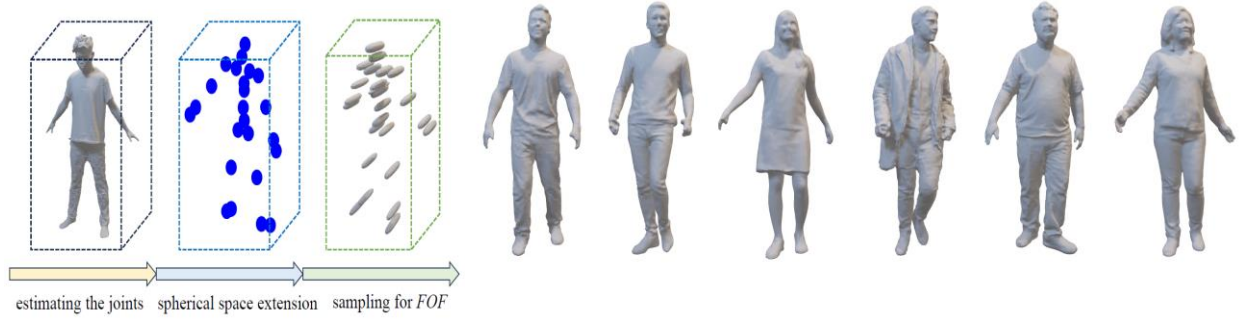
DreamHuman, AvatarVerse, DreamWaltz…

真实感数字人重建与生成：数字人生成（原生3D）

- 研究动机：由于3D数据集少，现有方法难以生成精细的几何结构
- 解决思路：（SMPL / Joints）+（Tri-plane / FOF）+（GAN / Diffusion）
  - GETAvatar：使用GAN采样生成几何三平面和纹理三平面，采用DMTet提取显式网格表示，利用可微分的光栅化得到高分辨率渲染图像
  - Joint2Human：提出一种新的关节点紧致球形编码表示，用于引导基于FOF的2D扩散模型来生成3D人体，实现了支持宽松衣服的多样化的几何生成



GETAvatar [ICCV 2023]



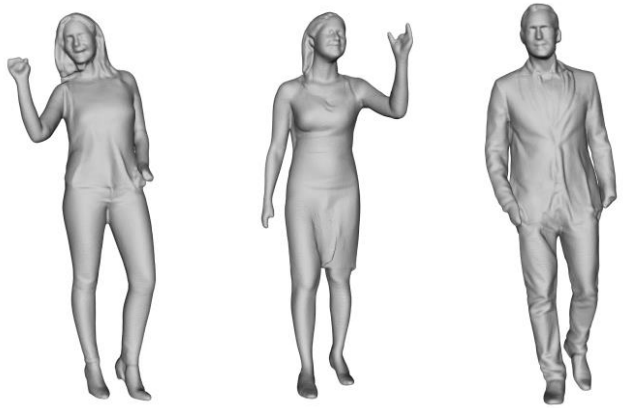
Joint2Human [CVPR 2024]

[1] Zhang et al. GETAvatar: Generative Textured Meshes for Animatable Human Avatars. ICCV 2023.

[2] Zhang et al. Joint2Human: High-quality 3D Human Generation via Compact Spherical Embedding of 3D Joints. CVPR 2024.

真实感数字人重建与生成：数字人生成（2D升维）

- 研究动机：由于缺少3D约束，现有方法难以生成视角一致的高质量数字人
- 解决思路：利用法线图提高几何质量
  - Chupa：通过扩散模型生成人体正反面的法线图，进而基于法线图和SMPL-X优化生成人体几何
  - Get3DHuman：充分利用2D人体生成和3D重建方法的模型先验，解耦几何和纹理，分别进行建模
  - HumanNorm：设计适应法线和深度的扩散模型、多阶段SDS损失来优化DMTet



Chupa [ICCV 2023]

Get3DHuman [ICCV 2023]

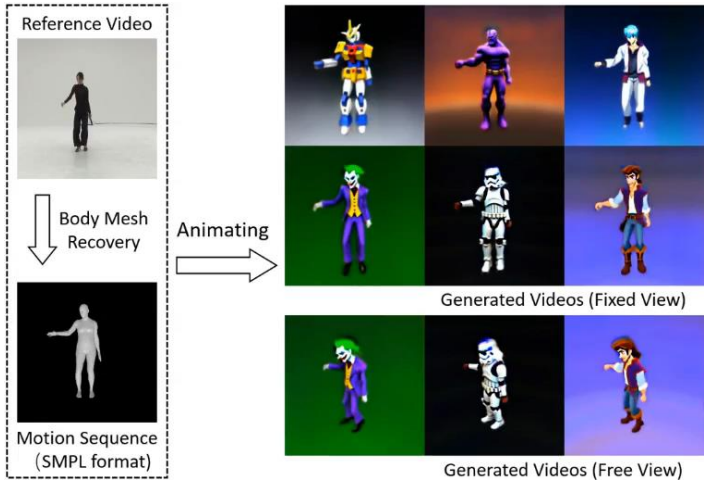
HumanNorm [CVPR 2024]

[1] Kim et al. Chupa: Carving 3D Clothed Humans from Skinned Shape Priors using 2D Diffusion Probabilistic Models. ICCV 2023.
[2] Xiong et al. Get3DHuman: Lifting StyleGAN-Human into a 3D Generative Model Using Pixel-Aligned Reconstruction Priors. ICCV 2023.
[3] Huang et al. HumanNorm: Learning Normal Diffusion Model for High-quality and Realistic 3D Human Generation. CVPR 2024.

真实感数字人重建与生成：数字人生成（2D升维）

- 研究动机：由于缺少3D约束，现有方法难以生成视角一致的高质量数字人
- 解决思路：使用人体形状、姿态和语义等先验，生成表征连续的高质量数字人
  - DreamHuman：采样姿态先验，并在优化过程中放大局部语义来细化结构，实现可驱动的生成
  - DreamWaltz：提出SMPL引导的3D一致的SDS损失，实现可驱动的复杂3D Avatar生成
  - AvatarVerse：提出DensePose引导的2D扩散模型，采用渐进式高分辨率合成策略，解决Janus问题

DreamHuman [NeurIPS 2023]



DreamWaltz [NeurIPS 2023]

AvatarVerse [AAAI 2024]

[1] Nikos et al. DreamHuman: Animatable 3D Avatars from Text. NeurIPS 2023.

[2] Huang et al. DreamWaltz: Make a Scene with Complex 3D Animatable Avatars. NeurIPS 2023.

[3] Zhang et al. AvatarVerse: High-quality & Stable 3D Avatar Creation from Text and Pose. AAAI 2024.

汇报提纲

一 真实感数字人重建与生成

二 运动捕捉与运动生成

三 推理与交互生成

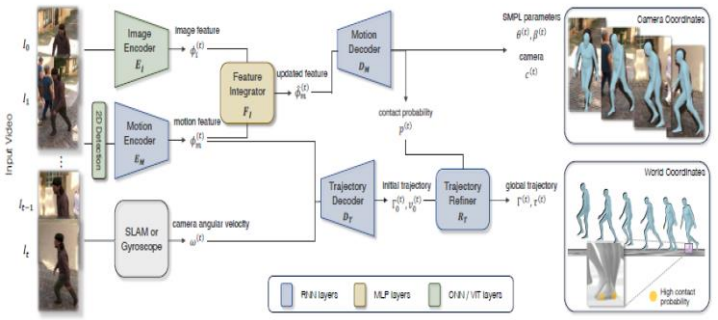
运动捕捉与运动生成：单人运动捕捉

现有局限和改进方向

单帧重建方法无法充分挖掘帧间的时间相关性，易造成抖动

动态重建

从视频中重建人体网格模型

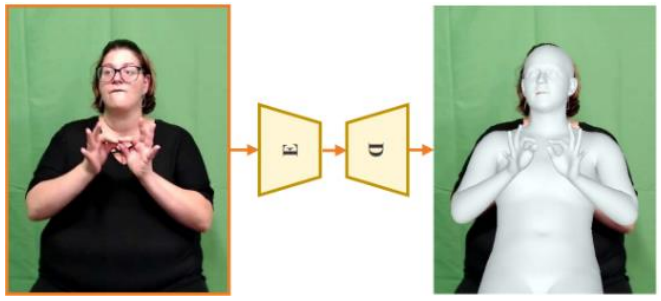


4Dhumans, WHAM …

仅身体部分的运动捕捉无法重建精细的手部和脸部姿态

全身重建

从图像中回归全身参数

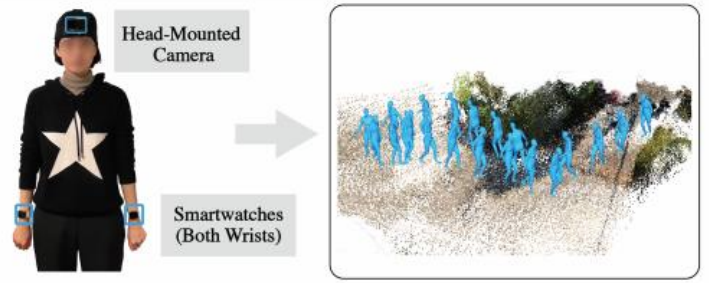


PyMAF-X, OSX, ProxyCap …

纯视觉方法无法有效处理严重遮挡，末端关节重建精度低

稀疏IMU

融合视觉和惯性信息

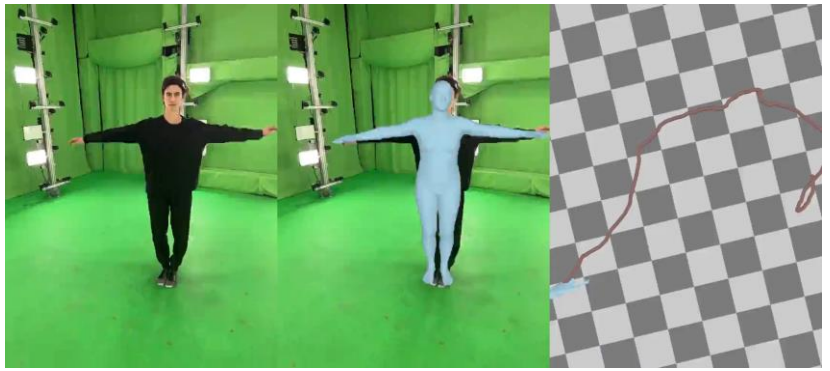


Robustcap, MocapEvery, HMD-Poser …

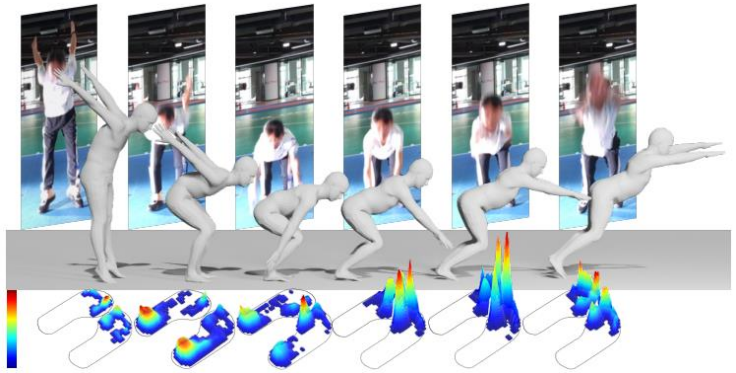
运动捕捉与运动生成：单人运动捕捉（动态重建）

- 研究动机：单人动作视频中含有复杂的人体姿态和背景信息，现有方法难以鲁棒重建视频中的人体姿态，且无法有效处理时间帧信息
- 解决思路：利用ViT大模型或设计时序网络来回归三维人体参数，从而增加模型鲁棒性
  - 4Dhumans：提出端到端的transformer架构来重建人体姿态与形状，并利用视频跟踪技术，实现复杂姿态的鲁棒动态重建
  - WHAM：使用AMASS数据训练运动编解码器，设计特征融合网络，实现移动相机下的全局重建
  - MMVP：提出包含视觉和压力信息的多模态动捕数据集，设计面向RGBD-P数据的动态SMPL拟合方法，实现基于单目视频的人体动作捕捉

4Dhumans [ICCV 2023]



WHAM [arXiv 2023]



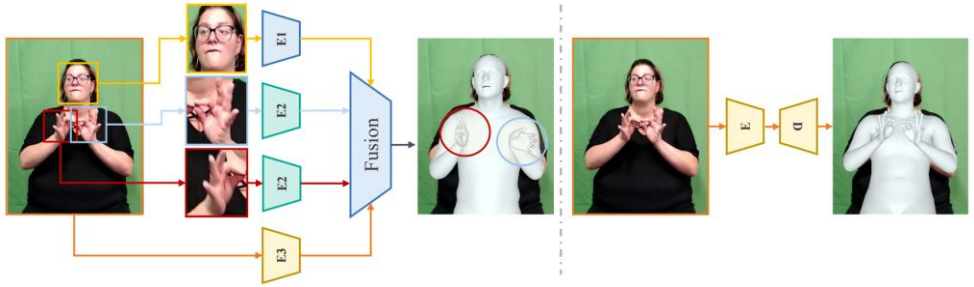
MMVP [CVPR 2024]

[1] Goel et al. Humans in 4D: Reconstructing and Tracking Humans with Transformers. ICCV 2023.
[2] Shin et al. WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion. arXiv 2023.
[3] Zhang et al. MMVP: A Multimodal MoCap Dataset with Vision and Pressure Sensors. CVPR 2024.

运动捕捉与运动生成：单人运动捕捉（全身重建）

- 研究动机：手部、面部等部位级特征占比小、姿态复杂，且简单拼接方法结果不自然
- 解决思路：设计拼合机制 / 设计单阶段网络 / 使用代理表示 + SMPL-X
  - PyMAF-X ：提出空域对齐注意力和自适应拼合机制，提高模型-图像的对齐精度和腕部拼接自然性
  - OSX：设计单阶段网络，编码器预测全局人体参数，解码器采用上采样裁剪方式提取高分辨率的人脸/人手特征，并利用关键点引导的变形注意力来精确估计人脸/人手参数
  - ProxyCap ：采用2D骨架序列作为代理表示，设计以人为中心的代理-运动学习机制，实现实时的世界空间下的人体运动捕捉

PyMAF-X [TPAMI 2023]



OSX [CVPR 2023]

ProxyCap [CVPR 2024]

[1] Zhang et al. PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images. IEEE TPAMI 2023.
[2] Lin et al. One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer. CVPR 2023.
[3] Zhang et al. ProxyCap: Real-time Monocular Full-body Capture in World Space via Sequential Proxy-to-Motion Learning. CVPR 2024.

运动捕捉与运动生成：单人运动捕捉（稀疏IMU）

- 研究动机：单人动作视频中存在复杂的背景和遮挡，现有方法无法有效处理遮挡并重建高精度的末端关节姿态
- 解决思路：融合视觉设备和IMU设备信息，增强运动捕捉的准确性
  - Robustcap：将稀疏IMU信号转换到相机坐标系中与图像信息融合，实现实时的人体运动捕捉
  - MocapEvery: 采用头戴式相机和2个智能手表，设计地面高度更新算法和基于流形的优化方法，实现室内室外的人体运动捕捉
  - HMD-Poser: 使用VR头显和稀疏IMU，设计轻量化的时空特征学习网络，实现实时运动捕捉与显示

Robustcap [SIGGRAPH ASIA 2023]

MocapEvery [CVPR 2024]

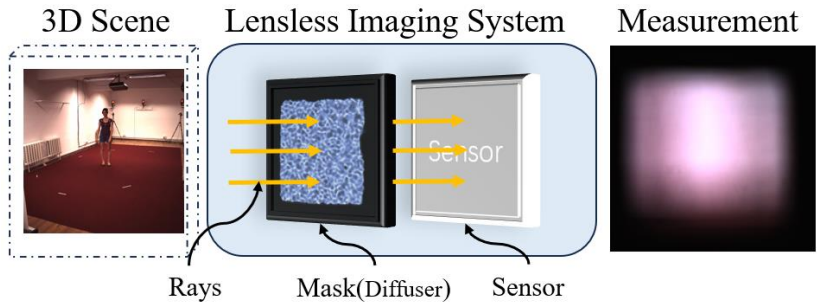
HMD-Poser [CVPR 2024]

[1] Pan et al. Fusing Monocular Images and Sparse IMU Signals for Real-time Human Motion Capture. SIGGRAPH ASIA 2023.
[2] Lee et al. Mocap Everyone Everywhere: Lightweight Motion Capture With Smartwatches and a Head-Mounted Camera. CVPR 2024.
[3] Dai et al. HMD-Poser: On-Device Real-time Human Motion Tracking from Scalable Sparse Observations. CVPR 2024.

单人运动捕捉：无透镜成像下的人体重建

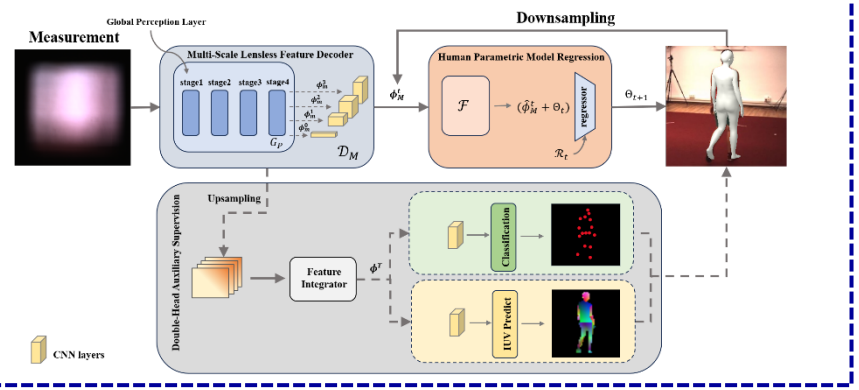
- 研究动机：无透镜成像下的人体三维姿态与形状估计不仅有利于保护隐私，而且由于设备体积小、结构简单，可用于军事等隐秘监测场景。

无透镜成像模型



- 物理优势：设备体积小、结构简单
- 成像优势：成像结果具有加密属性

解决方案



- 解决思路：
  - 多尺度无透镜特征解码器：对经过光学编码到全局的信息进行解码，并对特征进行有效提取
  - 双头辅助监督机制：通过额外的辅助监督提高人体姿态估计的精度，并且对肢体的末端进行有效的监督

无透镜成像数据（输入）

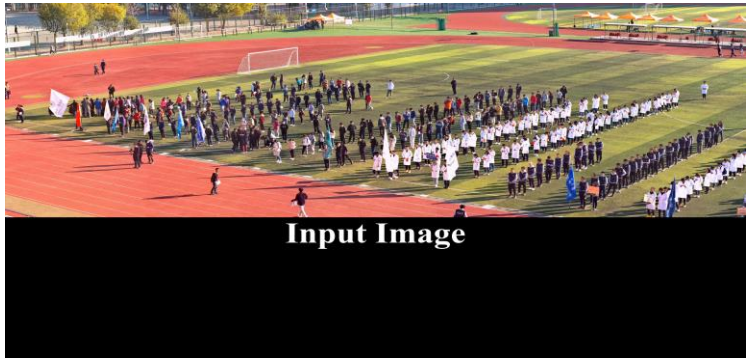
端到端网络

重建结果（输出）

[1] Ge et al. LPSNet: End-to-End Human Pose and Shape Estimation with Lensless Imaging. CVPR 2024.

运动捕捉与运动生成：多人运动捕捉

- 研究动机：仅采用检测-单人动捕的解决范式无法获得空间一致的、遮挡鲁棒的重建
- 解决思路：通过多对象表示直接回归多人位姿或通过全局优化实现空间一致的多人重建
  - Crowd3D：定义了人与场景虚拟交互点的新表示，进而建立了2D图像像素点与3D空间位置的映射关系，实现了数百人级别的全局三维位置、姿态与形状重建
  - Multi-HMR：采用ViT作为backbone，设计了人类预测头模块，实现多人在相机空间下的全身重建
  - AIOS：构建层级的查询和关联策略，渐进地估计全局与局部特征，实现单阶段的多人位置与全身姿态重建



Crowd3D [CVPR 2023]

Multi-HMR [arXiv 2024]

AiOS [CVPR 2024]

[1] Wen et al. Crowd3D: Towards hundreds of people reconstruction from a single image. CVPR 2023.
[2] Baradel et al. Multi-HMR: Multi-Person Whole-Body Human Mesh Recovery in a Single Shot. arXiv 2024.
[3] Sun et al. AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation. CVPR 2024.

运动捕捉与运动生成：运动生成

有条件的人体运动生成方法

文本 *A person paces from right to left.*

音频 

轨迹 文本提示

空间控制信息

生成模型

GAN

VAE

Normalizing Flow

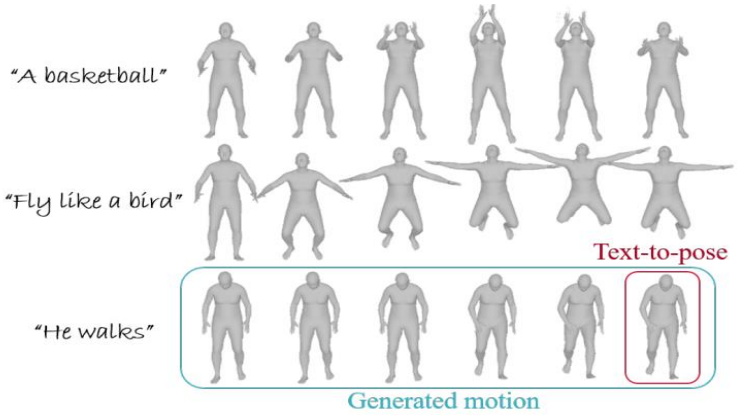
Diffusion Models

Motion Graph

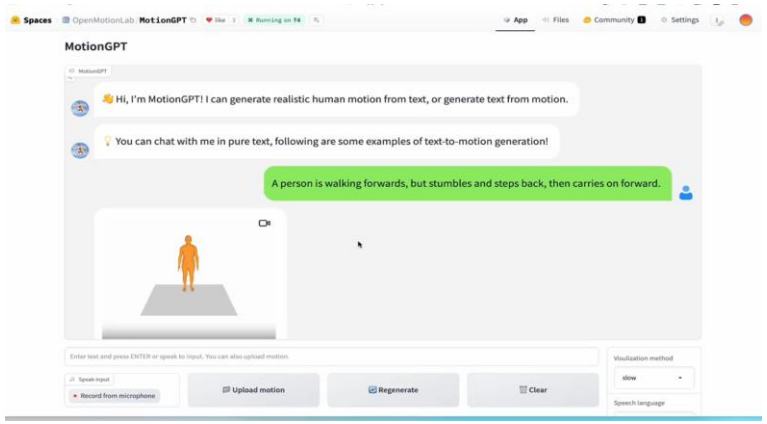
…

运动捕捉与运动生成：运动生成（文本驱动）

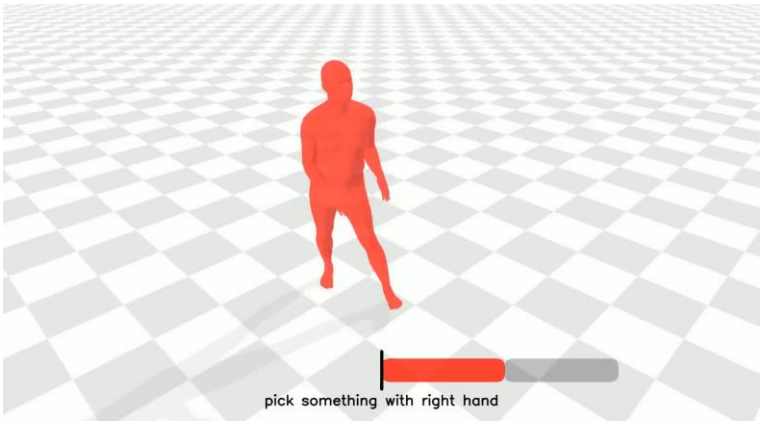
- 研究动机：仅用简单的文本编码与传统的生成模型相结合，难以保证动作与文本描述的语义一致性，也难以生成连续的长时序列动作
- 解决思路：利用预训练大模型来实现文本和动作的对齐，或改进采样方法与位置编码生成长时序列动作
  - OOHMG ：提出文本姿态对齐模型和预训练的运动生成器，能够为开放词汇文本生成动作序列
  - MotionGPT：构建基于微调LLM的人体通用运动生成器，可根据多模态输入生成连续的人体运动
  - FlowMDM：引入混合位置编码和姿态互注意力模块，实现无缝的长时序列动作生成



OOHMG [CVPR 2023]



MotionGPT [NeurIPS 2023]



FlowMDM [CVPR 2024]

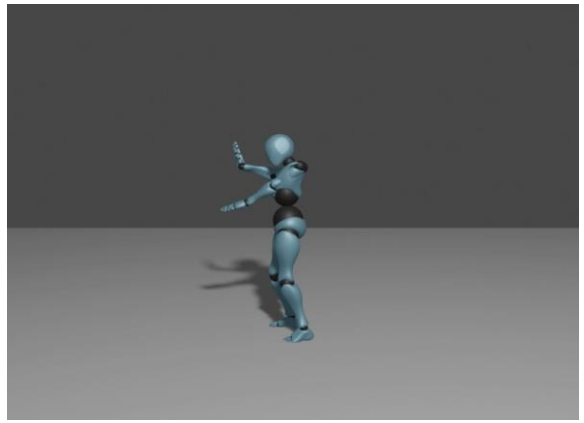
[1] Lin et al. Being Comes from Not-being: Open-vocabulary Text-to-Motion Generation with Wordless Training. CVPR 2023.
[2] Zhang et al. MotionGPT: Finetuned LLMs Are General-Purpose Motion Generators. NeurIPS 2023.
[3] Barquero et al. Seamless Human Motion Composition with Blended Positional Encodings. CVPR 2024.

运动捕捉与运动生成：运动生成（音频驱动）

- 研究动机：现有的音频驱动方法无法完全解决生成动作序列的多样性与风格化不足的问题，且无法生成长时动作序列
- 解决思路：引入组合模型或提取情绪特征以提高结果多样性；利用预训练大模型或训练异构数据提供风格化；改进自回归网络或用关键帧引导的并行局部扩散来生成长时序列
  - EMAGE：引入自注意力机制自适应地融合节奏和语义，并用组合的4个VQVAE来提高结果的多样性
  - GestureDiffuCLIP：基于CLIP，从多模态输入中提取风格与运动表示，生成风格化、有情绪的动作
  - Lodge：设计两阶段由粗到细的扩散模型框架，并行生成长时的舞蹈动作序列

EMAGE [CVPR 2024]

GestureDiffuCLIP [SIGGRAPH 2023]

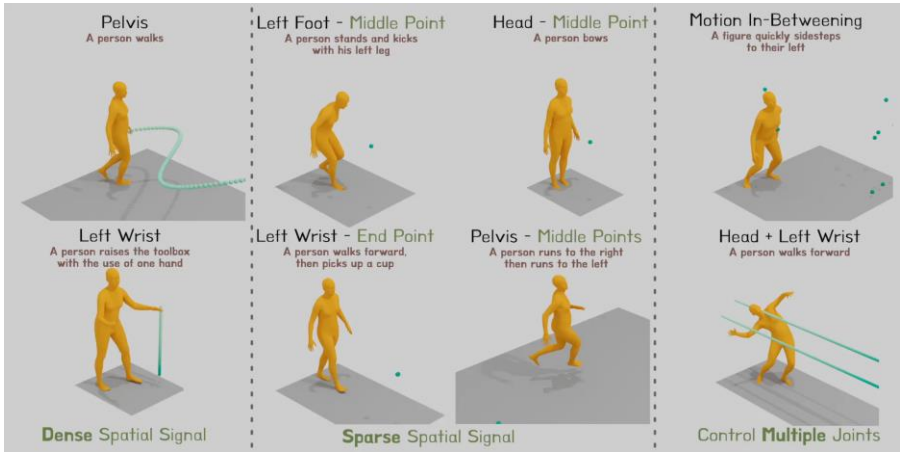


Lodge [CVPR 2024]

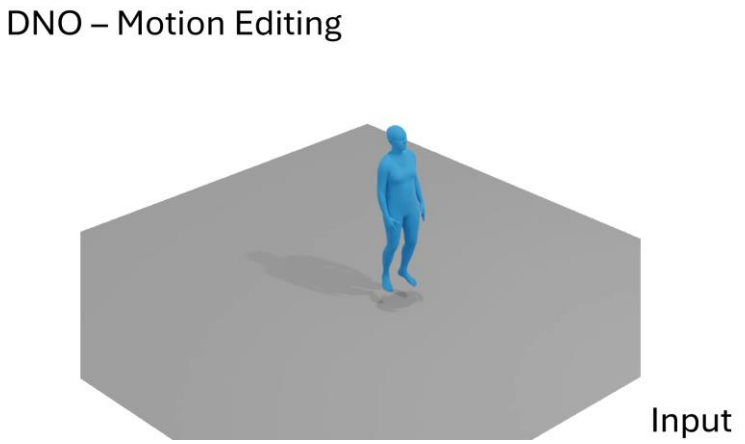
[1] Liu et al. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling. CVPR 2024.
[2] Ao et al. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. SIGGRAPH 2023.
[3] Li et al. Lodge: A Coarse to Fine Diffusion Network for Long Dance Generation Guided by the Characteristic Dance Primitives. CVPR 2024.

运动捕捉与运动生成：运动生成（轨迹驱动）

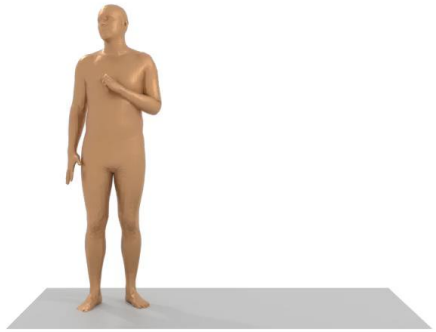
- 研究动机：语言条件运动生成模型无法轻松集成灵活的空间控制信号，也无法对关节位置实现细粒度控制
- 解决思路：使用精心设计的约束或调整扩散模型，实现灵活高效的细粒度关节与轨迹的控制和编辑
  - OmniControl：将空间和真实感引导整合进扩散生成模型，实现对任意时刻和关节位置的真实感控制
  - DNO：提出了扩散噪声优化方法，使扩散模型成为可编辑轨迹和关节位置的运动生成的先验
  - PriorMDM：引入扩散混合技术，通过混合微调模型，实现长时 / 两人 / 混合控制的运动生成



Omnicontrol [ICLR 2024]



DNO [CVPR 2024]



PriorMDM [ICLR 2024]

[1] Xie et al. Omnicontrol: Control any joint at any time for human motion generation. ICLR 2024.
[2] Karunratanakul et al. Optimizing diffusion noise can serve as universal motion priors. CVPR 2024.
[3] Shafir et al. Human motion diffusion as a generative prior. ICLR 2024.

汇报提纲

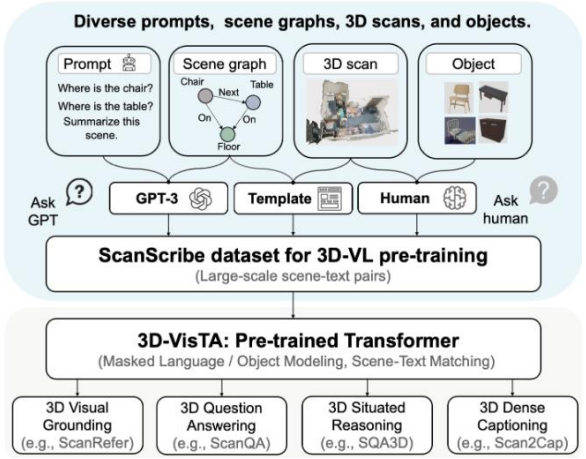
一 真实感数字人重建与生成

二 运动捕捉与运动生成

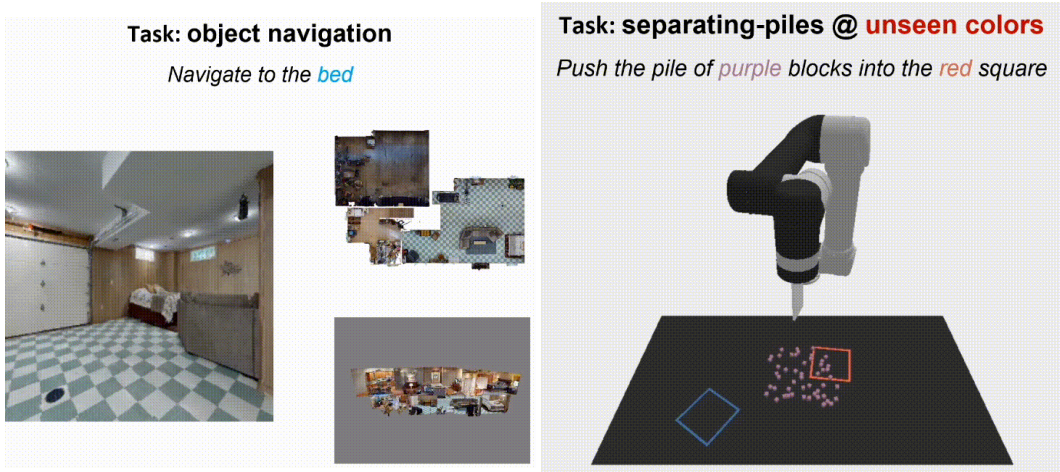
三 推理与交互生成

推理与交互生成：推理

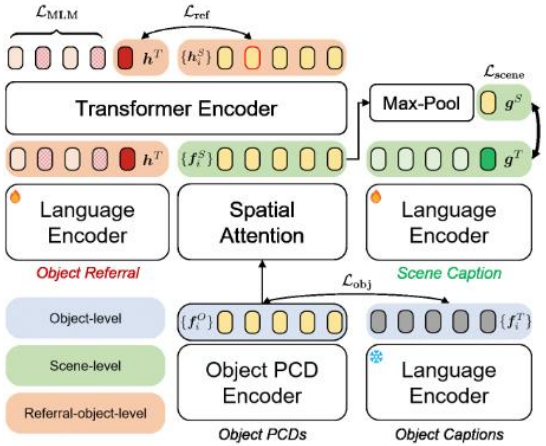
- 研究动机：数字人需要有理解与推理能力，但现有AI模型在3D环境中理解、推理和交互能力具有局限性，需要提高其通用性和执行多任务的能力
- 解决思路：大规模多模态数据集 + 预训练大模型 + 微调策略
  - 3D-VisTA：构建大规模3D场景-文本对数据集->自监督预训练策略->Finetune
  - LEO：将3D任务表述为序列预测问题，分阶段预训练强化agent的交互推理
  - SceneVerse：构建大规模数据集，通过多级场景-语言对比对齐训练提升模型的泛化能力



3D-VisTA [ICCV 2023]



LEO [ICLR2024]

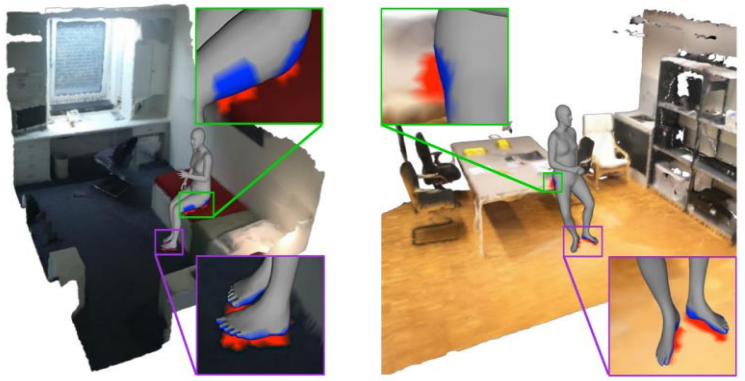


SceneVerse [arXiv 2024]

[1] Zhu et al. 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment. ICCV 2023.
[2] Huang et al. An Embodied Generalist Agent in 3D World. ICLR 2024.
[3] Jia et al. SceneVerse: Scaling 3D Vision-Language Learning for Grounded Scene Understanding. arXiv 2024.

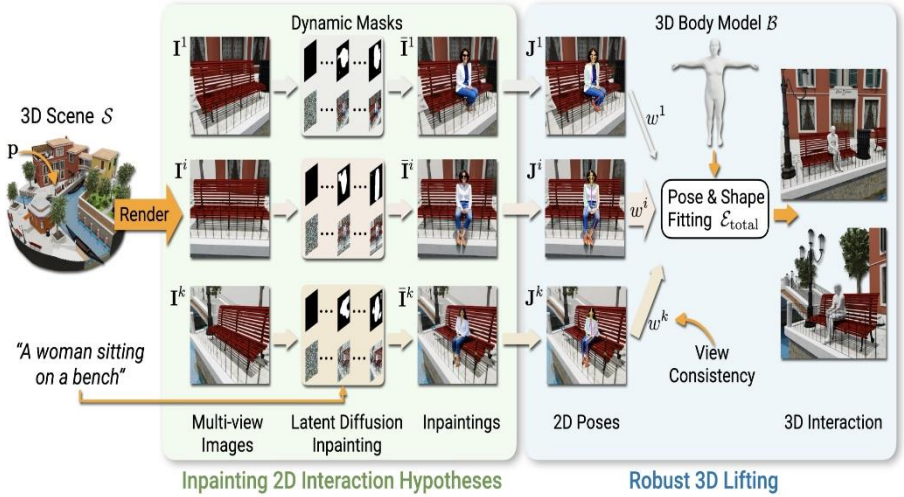
推理与交互生成：人与场景交互生成（静态）

- 研究动机：现有方法难以生成物理合理、自然可控、泛化性强的人与场景的交互
- 解决思路：条件约束 / 细粒度表示 / 2D大模型
  - PHIN：提出姿势引导的生成框架，设计几何对齐损失，生成与场景自然接触的3D人体
  - Narrator：提出联合全局与局部的场景图，以及身体动作部位级表示，实现符合自然语言描述习惯的细粒度控制的的人与场景交互生成和多人生成
  - GenZI：从2D视觉-语言大模型中蒸馏交互先验，设计inpainting和优化方法，实现零样本交互生成



PHIN [AAAI 2023]

Narrator [ICCV 2023]



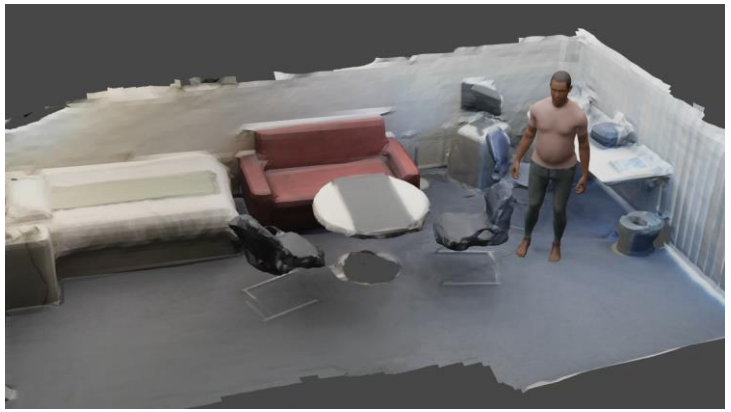
GenZI [CVPR 2024]

[1] Kim et al. Pose-Guided 3D Human Generation in Indoor Scene. AAAI 2023.
[2] Xuan et al. Narrator: Towards Natural Control of Human-Scene Interaction Generation via Relationship Reasoning. ICCV 2023.
[3] Li et al. GenZI: Zero-Shot 3D Human-Scene Interaction Generation. CVPR 2024.

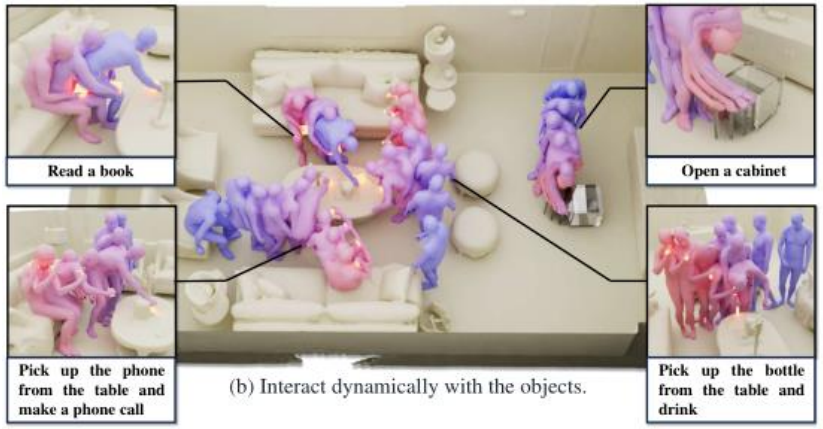
推理与交互生成：人与场景交互生成（动态）

- 研究动机：现有方法难以生成任意长度、真实多样的人与场景交互运动序列
- 解决思路：设置场景/目标感知策略、增加物理约束、建立人与场景交互数据集
  - SceneDiffuser：提出了支持场景感知的生成、基于物理的优化和目标引导的规划的统一模型
  - DIMOS：基于强化学习，设计场景和交互感知策略，可生成在复杂场景中真实且多样的交互动作
  - TRUMANS：建立了最大的HSI动捕数据集，提出了基于扩散模型的自回归方法，实时生成任意长度的人与场景交互序列

SceneDiffuser [CVPR 2023]



DIMOS [ICCV 2023]



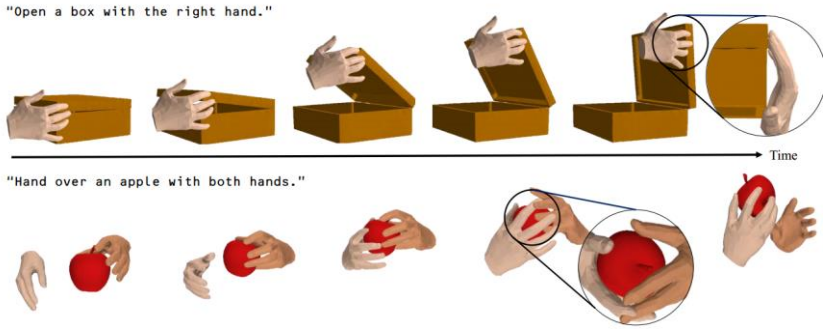
TRUMANS [CVPR 2024]

[1] Huang et al. Diffusion-based Generation, Optimization, and Planning in 3D Scenes. CVPR 2023.
[2] Zhao et al. Synthesizing diverse human motions in 3D indoor scenes. ICCV 2023.
[3] Jiang et al. Scaling Up Dynamic Human-Scene Interaction Modeling. CVPR 2024.

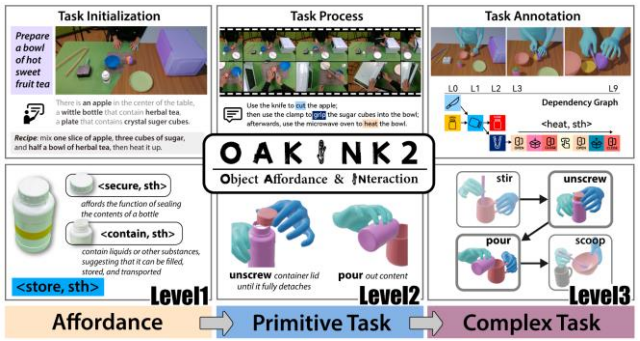
推理与交互生成：手物交互生成

- 研究动机：手-物体交互生成任务中涉及复杂的动作和手-物遮挡，现有方法无法有效生成物理上合理且语义上有意义的自然交互序列
- 解决思路：利用手-物接触信息、文本提示或构造大型手物交互数据集，来生成自然合理的手物交互结果
  - GeneOH Diffusion：提出以接触为中心的手-物交互表示以及域泛化去噪方案，支持多种下游应用
  - Text2HOI：将文本生成HOI的任务分解为接触生成与运动生成，实现物理上可信的交互生成
  - OakInk2：提出双手物体交互数据集，构建以物体为中心的三层级任务结构

GeneOH Diffusion [ICLR 2024]



Text2HOI [CVPR 2024]



OakInk2 [CVPR 2024]

[1] Liu et al. GeneOH Diffusion: Towards Generalizable Hand-Object Interaction Denoising via Denoising Diffusion. ICLR 2024.
[2] Cha et al. Text2HOI: Text-guided 3D Motion Generation for Hand-Object Interaction. CVPR 2024.
[3] Zhan et al. OakInk2: A Dataset of Bimanual Hands-Object Manipulation in Complex Task Completion. CVPR 2024.

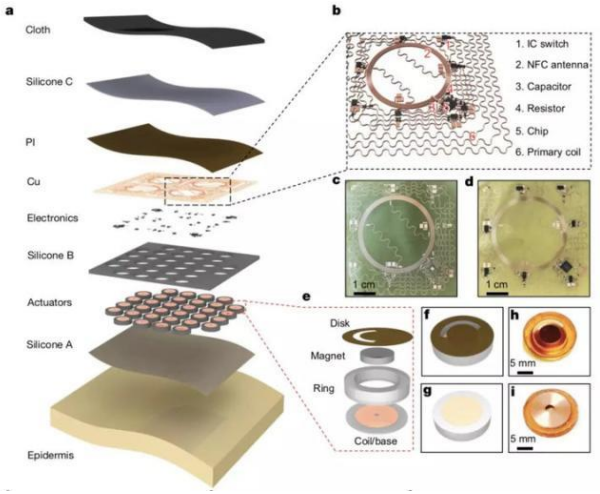
数字人研究的未来

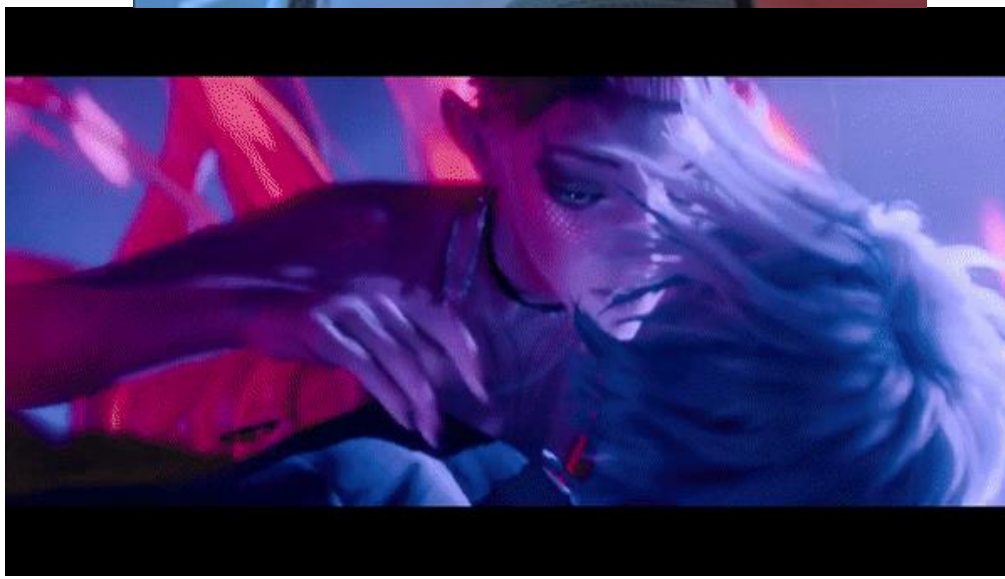
更明亮的“眼睛” | **更灵敏的“耳朵”** | **更灵巧的“手”**

- 让数字人有视听触觉



多模态感知与交互





数字人研究的未来

更聪明的“大脑”　　更精巧的“神经”

- 让数字人有感情

电影《Blade Runner (银翼杀手) 2049》2017

致谢

助理研究员 马健

博士生 温浩

博士生 张劲松

博士生 李雄政

博士生 于涛

博士生 黄敬

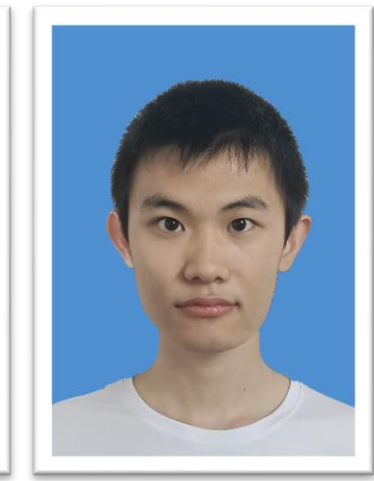
博士生 王毅

博士生 李金

硕士生 朱敏婕

硕士生 葛昊洋

硕士生 鲍贺浩



硕士生 杨源旺

硕士生 刘晓琳

硕士生 张木鑫

天津大学
Tianjin University

谢谢！