动态场景（人体）重建

Dynamic Scene (human) reconstruction

答辩人：李亚城　　导　师：黄天羽　　时间：2024/11/4

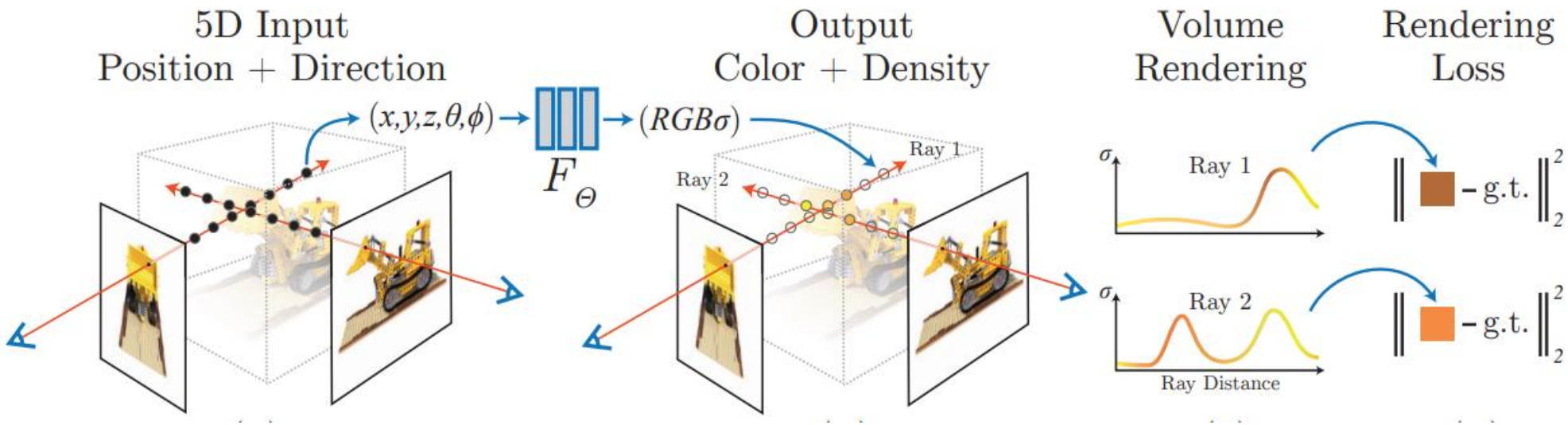
Contents

结构大纲

01

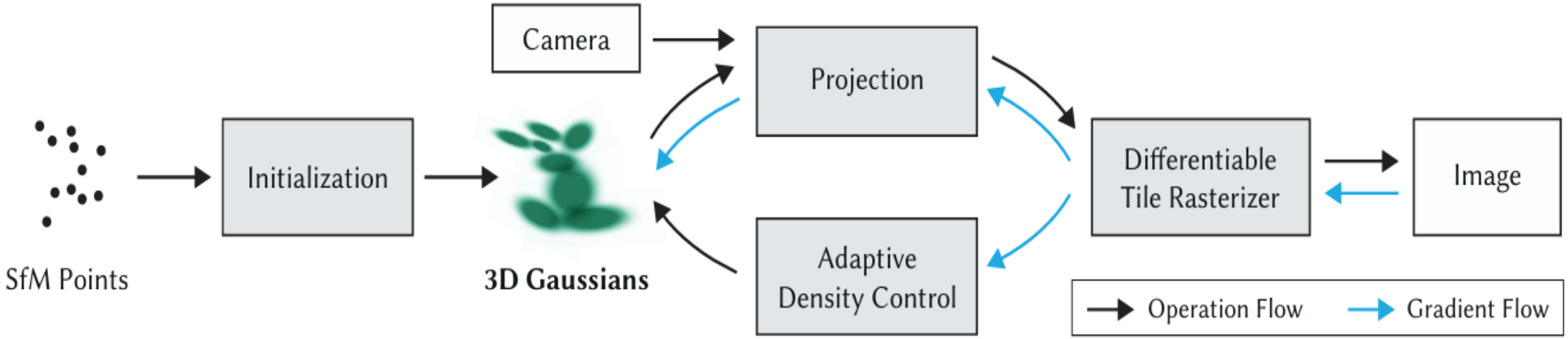
NeRF & 3DGS

全连接网络 隐式场景表示



$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t),\mathbf{d})dt\,, \text{ where } T(t) = \exp\left(-\int_{t_n}^{t}\sigma(\mathbf{r}(s))ds\right)$$

高斯椭球 显式场景表示



$$C = \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i \quad \text{with} \quad T_i = \exp\left(- \sum_{j=1}^{i-1} \sigma_j \delta_j \right)$$

02

GauHuman

GauHuman: Articulated Gaussian Splatting from Monocular Human Videos（CVPR 2024)

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

胡寿康

索尼AI科学家，研究方向包括**三维人体或物体重建**和生成、自动语音识别、自动机器学习

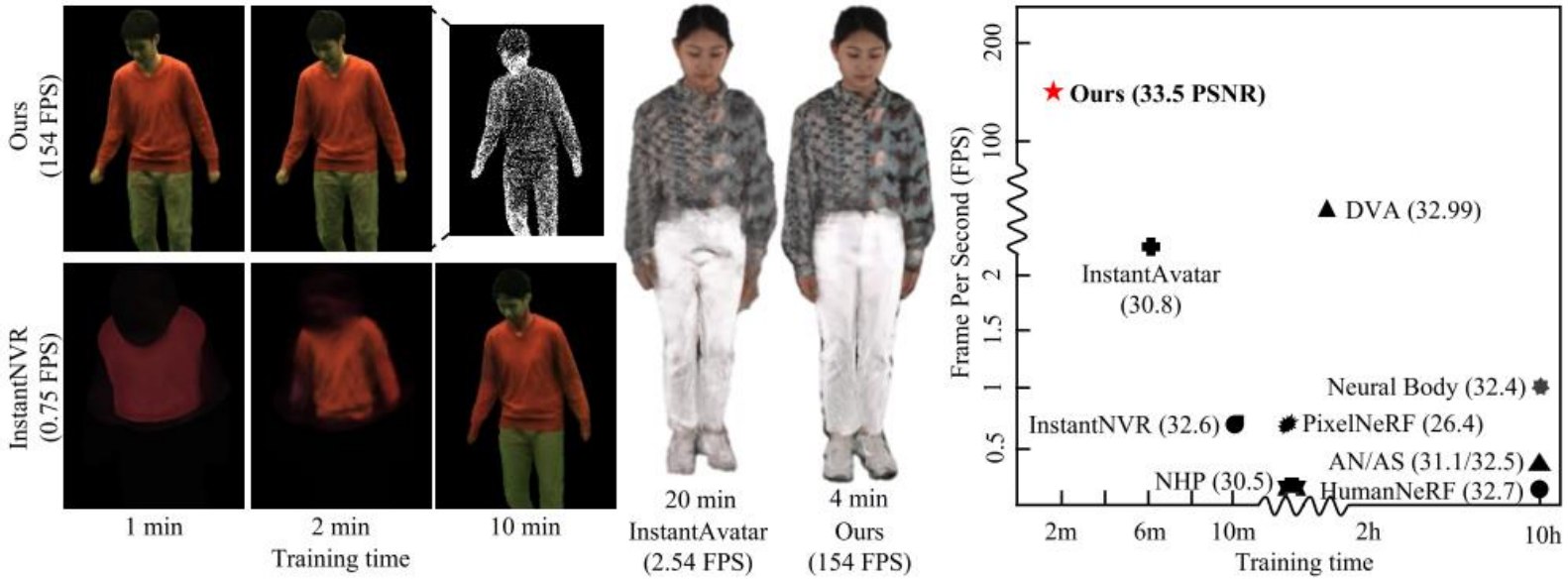
刘子纬

南洋理工MMLab助理教授，研究方向计算机视觉、计算机图形学，担任**CVPR** , ICCV, NeurIPS, ICLR, AAAI, WACV的区域主席

GauHuman 简述

GauHuman: Articulated Gaussian Splatting from Monocular Human Videos

Shoukang Hu Ziwei Liu
S-Lab, Nanyang Technological University
{shoukang.hu, ziwei.liu}@ntu.edu.sg

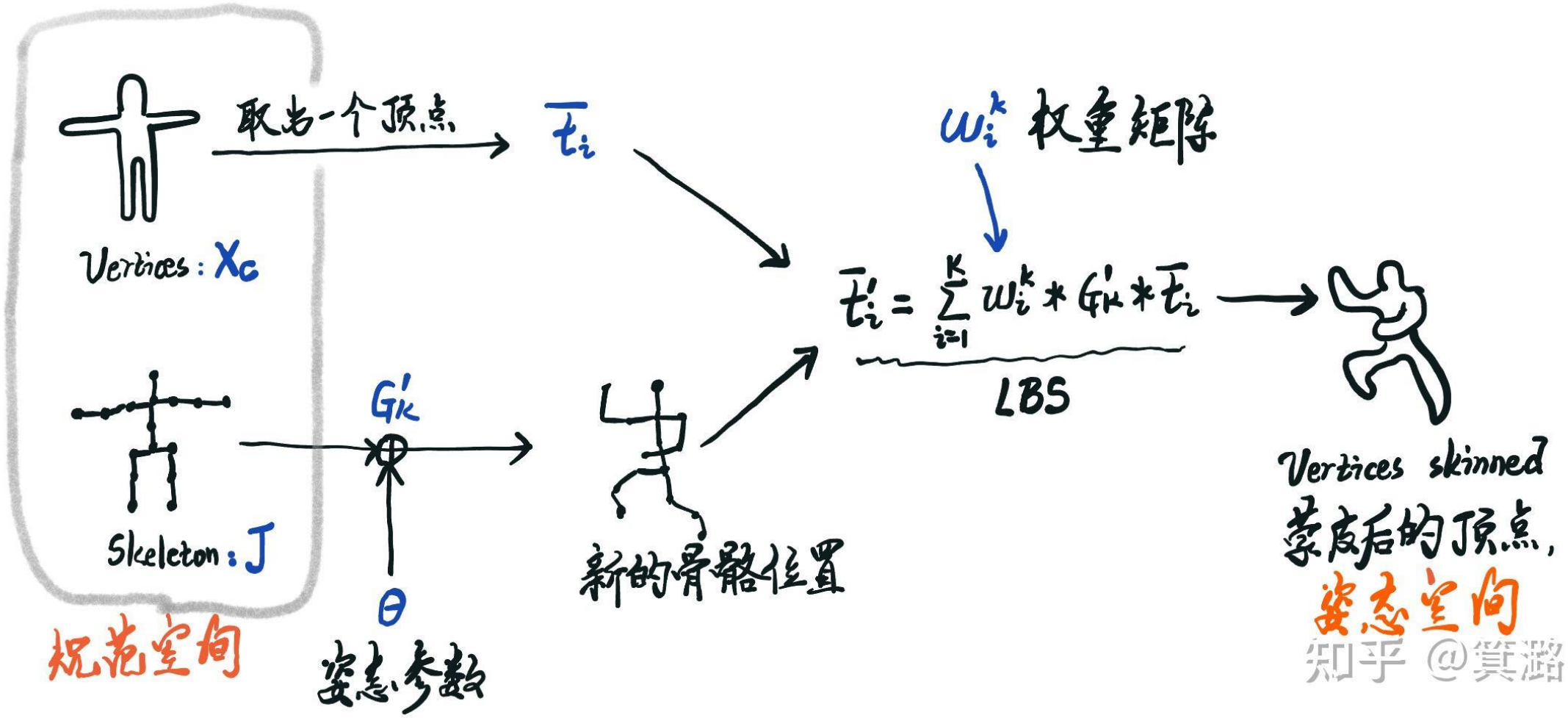


GauHuman采用高斯分布表示3D 人体模型，可在1~2分钟完成模型训练，并实现高达189 FPS的实时渲染结果。

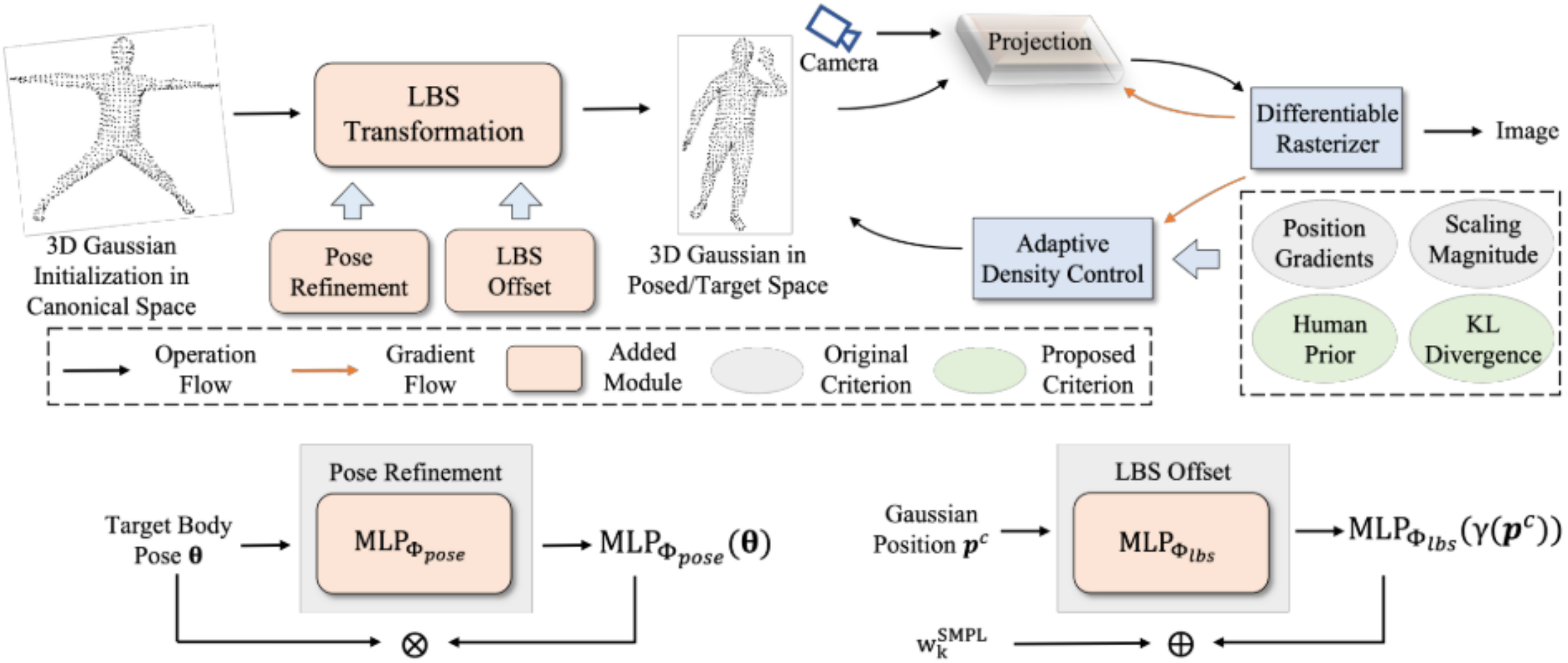
GauHuman 问题提出与实现思路

- 如何在Gaussian Splatting框架中正确整合来自稀疏视频的人类信息？如何有效优化Gaussian Splatting以实现快速训练?
- 可以在规范空间中建模3D人体，然后利用利用LBS变换将3D高斯从规范空间变换到姿势空间，但直接对3D高斯应用基于SMPL的LBS变换不可行——使用**神经网络**预测准确的LBS 转换
- 直接利用运动结构恢复（Sfm）生成的稀疏点云或随机点云来初始化3D高斯无法实现快速人体建模——使用**SMPL 等人类先验**进行 3D 高斯初始化
- 3D高斯自适应控制利用3D位置梯度信息来分裂和克隆高斯，会产生大量冗余高斯分布，影响优化过程和内存存储——根据**3D高斯距离**优化分裂克隆过程，并**利用人体先验修剪**控制高斯数量

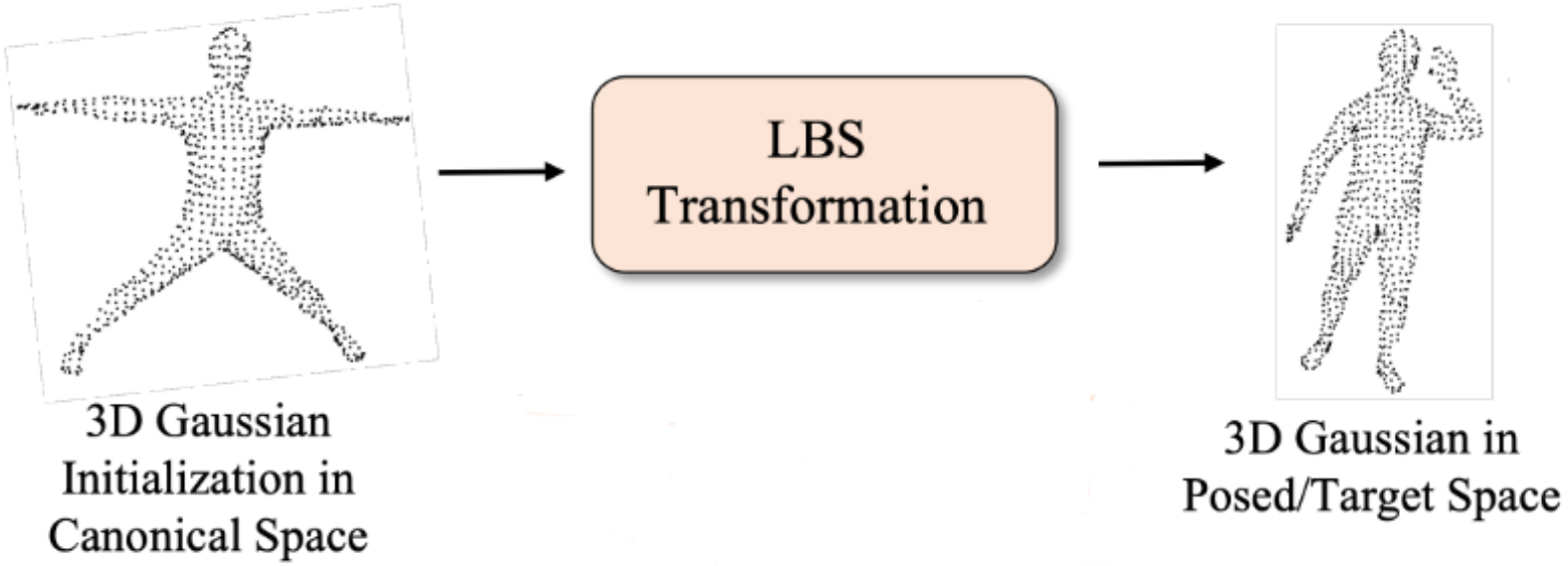
LBS（线性混合蒙皮）



GauHuman 具体实现



LBS转换



通过线性混合蒙皮 (LBS) 将3D 高斯从规范空间转换为姿势空间

使用估计的LBS变换矩阵旋转和平移每个3D高斯的位置和协方差：

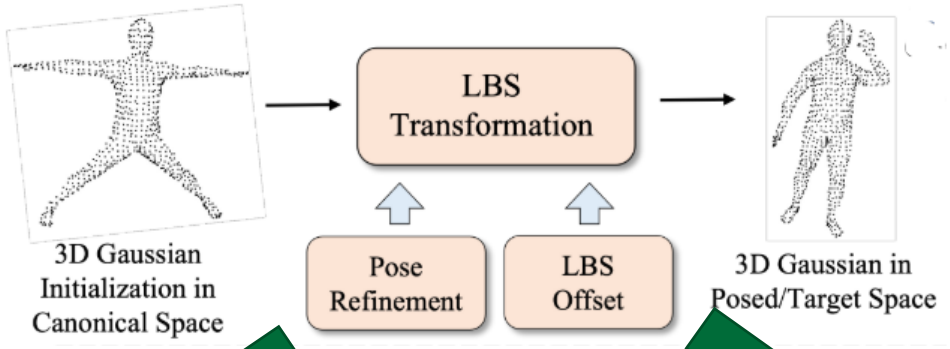
$$\boldsymbol{p}^t = \boldsymbol{G}(\boldsymbol{J}^t, \boldsymbol{\theta}^t)\boldsymbol{p}^c + \boldsymbol{b}(\boldsymbol{J}^t, \boldsymbol{\theta}^t, \boldsymbol{\beta}^t)$$

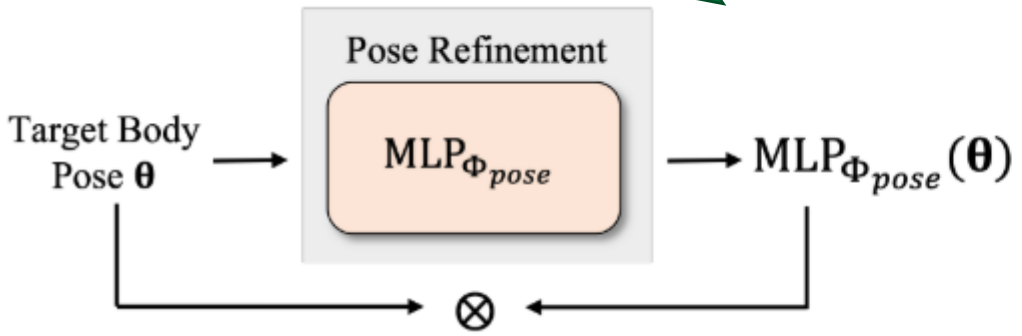
$$\boldsymbol{\Sigma}^t = \boldsymbol{G}(\boldsymbol{J}^t, \boldsymbol{\theta}^t)\boldsymbol{\Sigma}^c\boldsymbol{G}(\boldsymbol{J}^t, \boldsymbol{\theta}^t)^T,$$

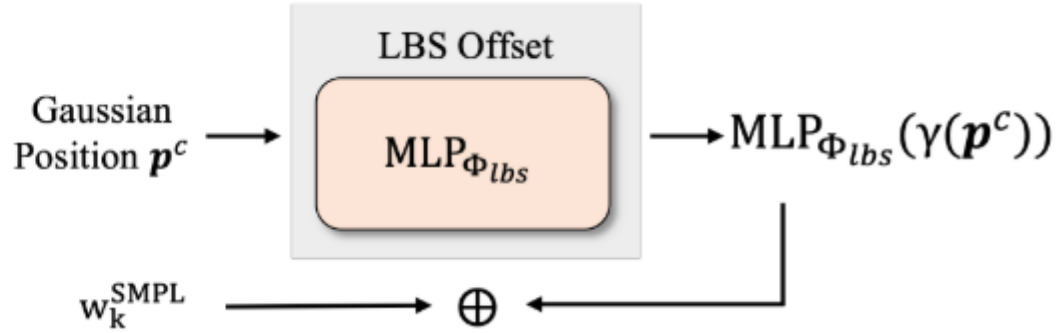
$$\boldsymbol{G}(\boldsymbol{J}^t, \boldsymbol{\theta}^t) = \sum_{k=1}^{K} w_k \boldsymbol{G_k}(\boldsymbol{J}^t, \boldsymbol{\theta}^t)$$

$$\boldsymbol{b}(\boldsymbol{J}^t, \boldsymbol{\theta}^t, \boldsymbol{\beta}^t) = \sum_{k=1}^{K} w_k \boldsymbol{b_k}(\boldsymbol{J}^t, \boldsymbol{\theta}^t, \boldsymbol{\beta}^t)$$

LBP权重预测和姿势细化



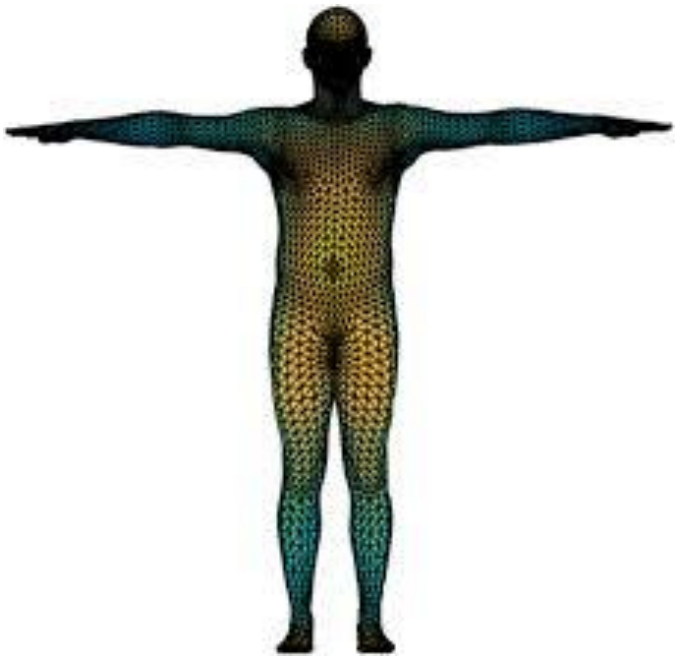


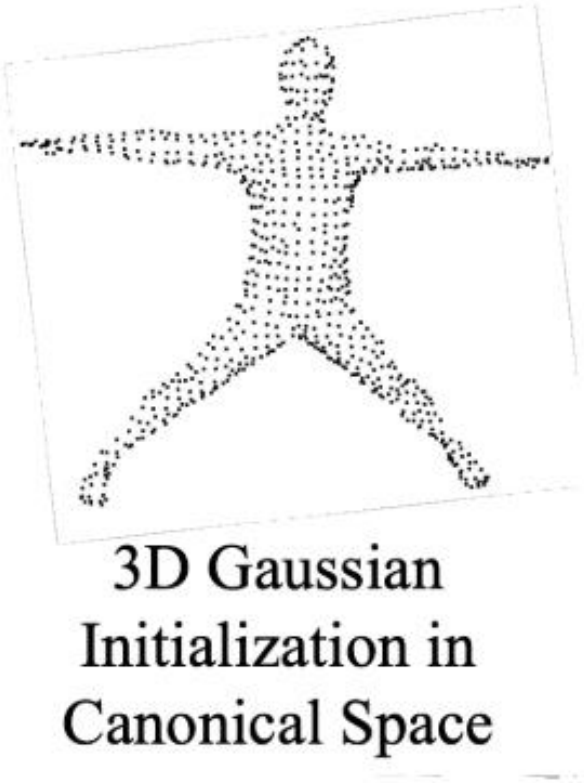


$$\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathrm{SMPL}} \otimes \mathrm{MLP}_{\Phi_{\mathrm{pose}}}(\boldsymbol{\theta})$$

$$w_k = \frac{e^{\log(w_k^{\mathrm{SMPL}}+10^{-8})+\mathrm{MLP}_{\Phi_{\mathrm{lbs}}}(\gamma(\boldsymbol{p}^c))[k]}}{\sum_{k=1}^{K} e^{\log(w_k^{\mathrm{SMPL}}+10^{-8})+\mathrm{MLP}_{\Phi_{\mathrm{lbs}}}(\gamma(\boldsymbol{p}^c))[k]}}$$

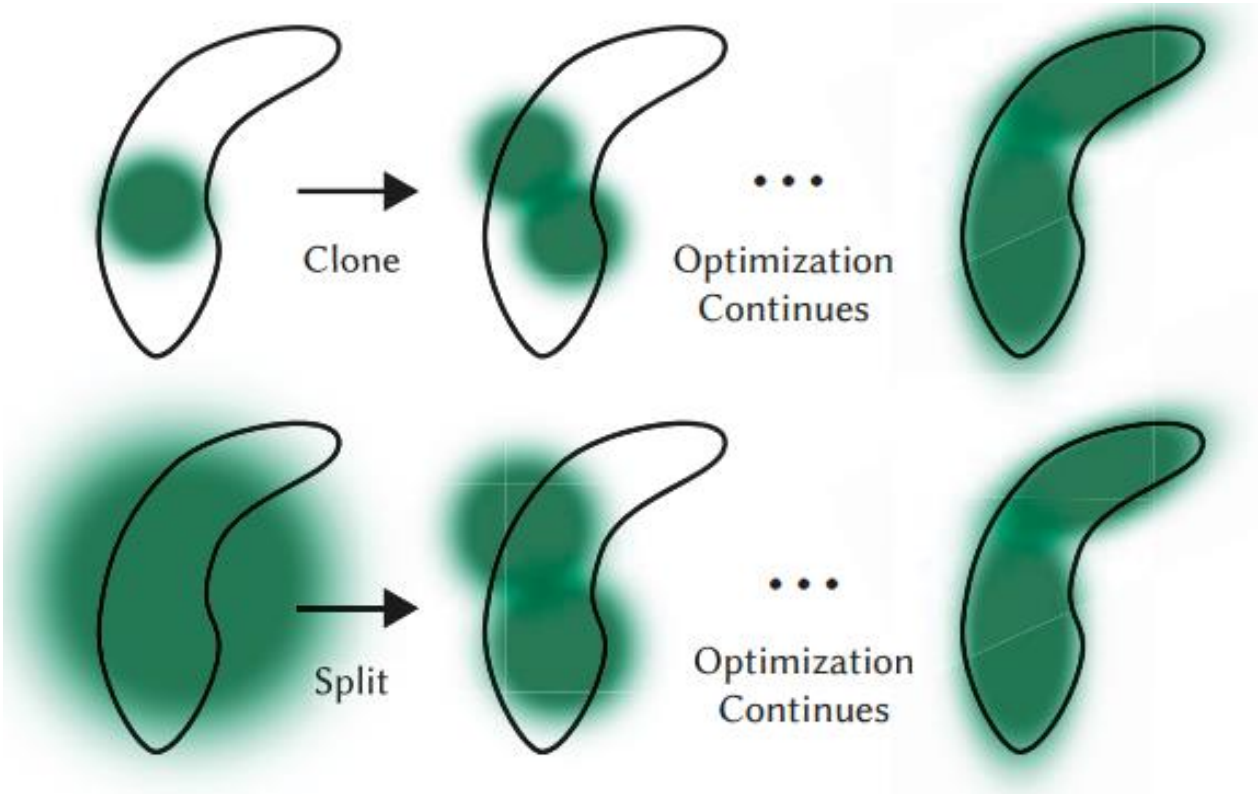
GauHuman快速优化之初始化





用SMPL顶点作为人类先验初始化3D高斯分布

GauHuman快速优化之分割&克隆



使用 KL散度作为 3D 高斯距离的度量，将3D高斯之间的距离纳入考虑范围：

$$KL(G(\boldsymbol{x}_0)|G(\boldsymbol{x}_1)) = \frac{1}{2}(tr(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) + \ln\frac{\det\boldsymbol{\Sigma}_1}{\det\boldsymbol{\Sigma}_2} + (\boldsymbol{p}_1-\boldsymbol{p}_0)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{p}_1-\boldsymbol{p}_0) - 3)$$

为了降低计算复杂度，使用K近邻算法识别其最接近的3D高斯，然后计算每对附近3D 高斯函数的 KL 散度，将时间复杂度从 $O(U^2)$ 降低到 $O(U)$

GauHuman快速优化之合并&修剪

- 为了将冗余的高斯进行合并，论文提出了一种**合并**操作：
  - 将**大位置梯度、小缩放幅度、KL散度小于 0.1**的高斯标记为冗余
  - 通过平均（位置、不透明度和 SH 系数）来合并两个高斯函数
  - 在初始化新高斯时将高斯的协方差（缩放）矩阵缩放 1.25 倍
- 在**修剪**操作中，结合SMPL顶点信息，修剪**远离 SMPL 顶点**的 3D 高斯

$$\mathcal{L} = \mathcal{L}_{color} + \lambda_1 \mathcal{L}_{mask} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{LPIPS},$$

实验结果-对比实验

Method	ZJU_MoCap					MonoCap				
	PSNR↑	SSIM↑	LPIPS*↓	Train	FPS	PSNR↑	SSIM↑	LPIPS*↓	Train	FPS
PixelNeRF [128]	24.71	0.892	121.86	1h†	1.20	26.43	0.960	43.98	1h†	0.75
NHP [55]	28.25	0.955	64.77	1h†	0.15	30.51	0.980	27.14	1h†	0.05
NB [78]	29.03	0.964	42.47	10h	1.48	32.36	0.986	16.70	10h	0.98
AN [77]	29.77	0.965	46.89	10h	1.11	31.07	0.985	19.47	10h	0.31
AS [79]	30.38	**0.975**	37.23	10h	0.40	32.48	**0.988**	**13.18**	10h	0.29
HumanNeRF [114]	30.66	0.969	33.38	10h	0.30	32.68	0.987	15.52	10h	0.08
DVA [84]	29.45	0.956	37.74	1.5h	16.5	32.99	0.983	15.83	1.5h	10.5
InstantNVR [30]	31.01	0.971	38.45	5m	2.20	32.61	**0.988**	16.68	10m	0.75
InstantAvatar [46]	29.73	0.938	68.41	3m	4.15	30.79	0.964	39.75	6m	2.54
GauHuman(Ours)	**31.34**	0.965	**30.51**	**1m**	**189**	**33.45**	0.985	13.35	**2m**	**154**

Training:	**1 min**	3 min	5 min	1.5 h	10 h	10 h	10 h	1h fine-tuning
Rendering:	**189 FPS**	4.15 FPS	2.20 FPS	16.5 FPS	0.30 FPS	0.40 FPS	1.48 FPS	0.15 FPS

实验结果-消融实验

Table 2. Quantitative Results of ablating pose refinement and LBS weight field modules. LPIPS* = 1000 × LPIPS.

	PSNR↑	SSIM↑	LPIPS*↓	Train	FPS
Ours (full model)	**28.08**	**0.886**	**103.3**	55s	189
Ours w/o pose refine	27.94	0.884	108.8	45s	192
Ours w/o LBS field	27.87	0.882	108.9	52s	190

Table 3. Quantitative Results of ablating 3D Gaussian initialization. #Gau denotes the number of 3D Gaussians at initialization.

	PSNR↑	SSIM↑	LPIPS*↓	Train
Articulated Init (w/ #Gau 6890)	**28.08**	**0.886**	**103.3**	55s
Random Init (w/ #Gau 6890)	27.98	0.882	108.9	5m
Random Init (w/ #Gau 30k)	**28.08**	0.884	104.3	30m

BS field | w/ random Init (#Gau 6890) | w/ random Init (#Gau 30000)

able 4. Quantitative Results of ablating our proposed KL-based lit/clone operations and a novel merge operation. #Gau denotes e final number of 3D Gaussians during optimization.

	PSNR↑	SSIM↑	LPIPS*↓	#Gau	Train
urs (full model)	**28.08**	**0.886**	**103.3**	**13162**	**55s**
/o KL split/clone	28.01	0.885	103.1	100k	1h
/o KL merge	**28.08**	**0.886**	103.6	14376	70s
/o articulated prune	**28.08**	**0.886**	103.4	13687	57s

03

一些4D GS方法

HiFi4G :High-Fidelity Human Performance Rendering via Compact Gaussian Splatting

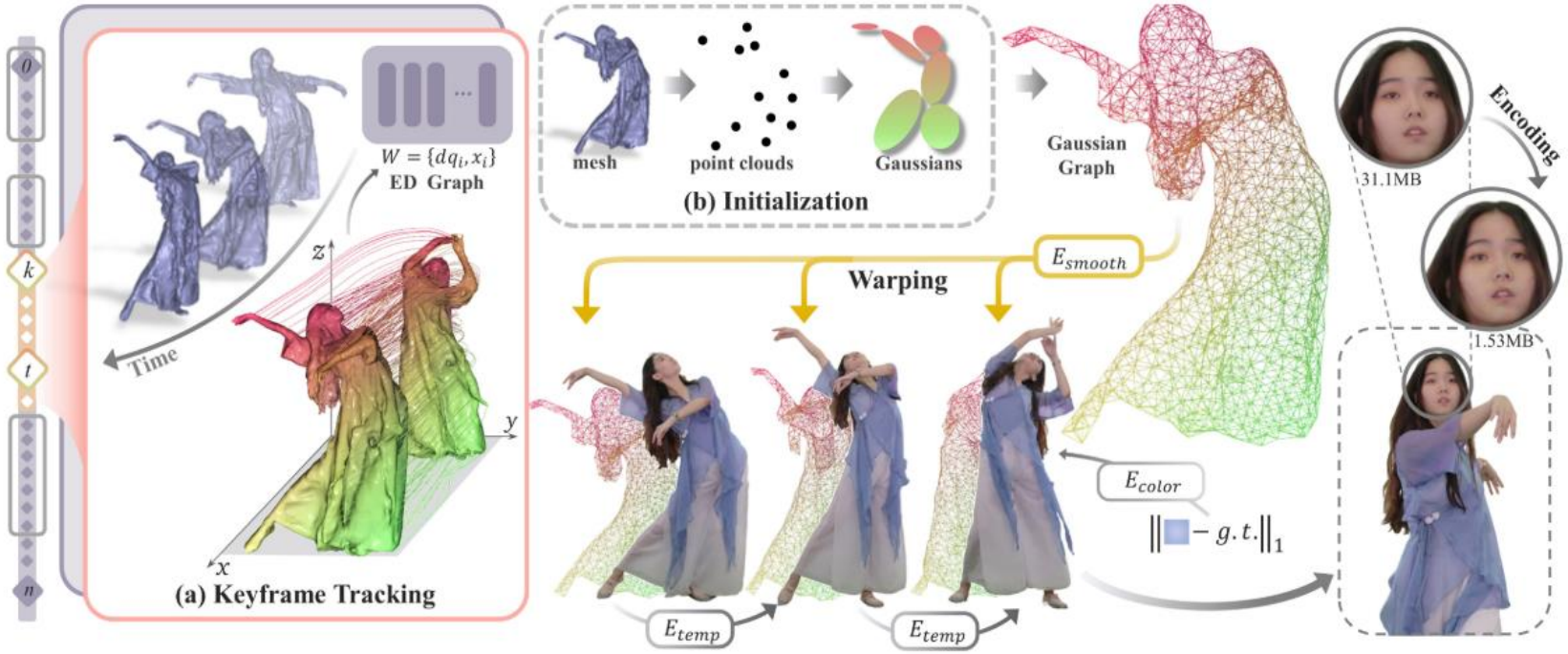
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

双图结构：

粗变形图根据关键帧构建网格序列，将运动参数转化为ED图

细粒度高斯图根据K近邻算法构建，初始化非关键帧时根据 ED 节点的运动插值将高斯图从关键帧扭曲到段内的其他帧

4DGS优化：直接扭曲细粒度高斯图会产生伪影，将高斯核的属性分为外观感知和运动感知两种，限制相邻帧外观属性变化，并设置运动属性变化平滑项



紧凑高斯表示：引入压缩方案-残差补偿、量化和熵编码，实现25倍压缩

认识了一个新指标VMAF

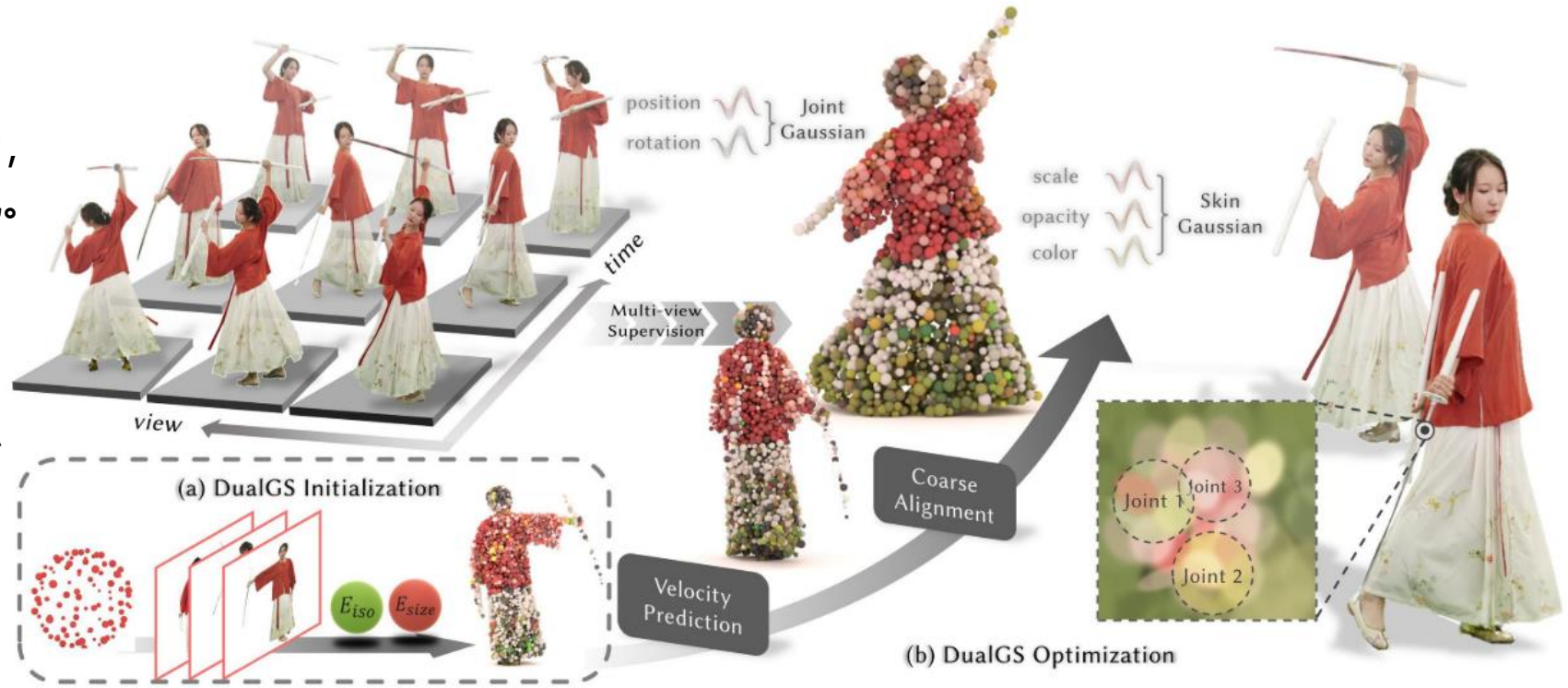
(CVPR 2024)

DualGaussian：Robust Dual Gaussian Splatting for Immersive Human-centric Volumetric Videos

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

双高斯表示：关节和皮肤
少量**运动感知联合高斯**来捕捉全局运动，
大量**外观感知皮肤高斯**来表达视觉外观。

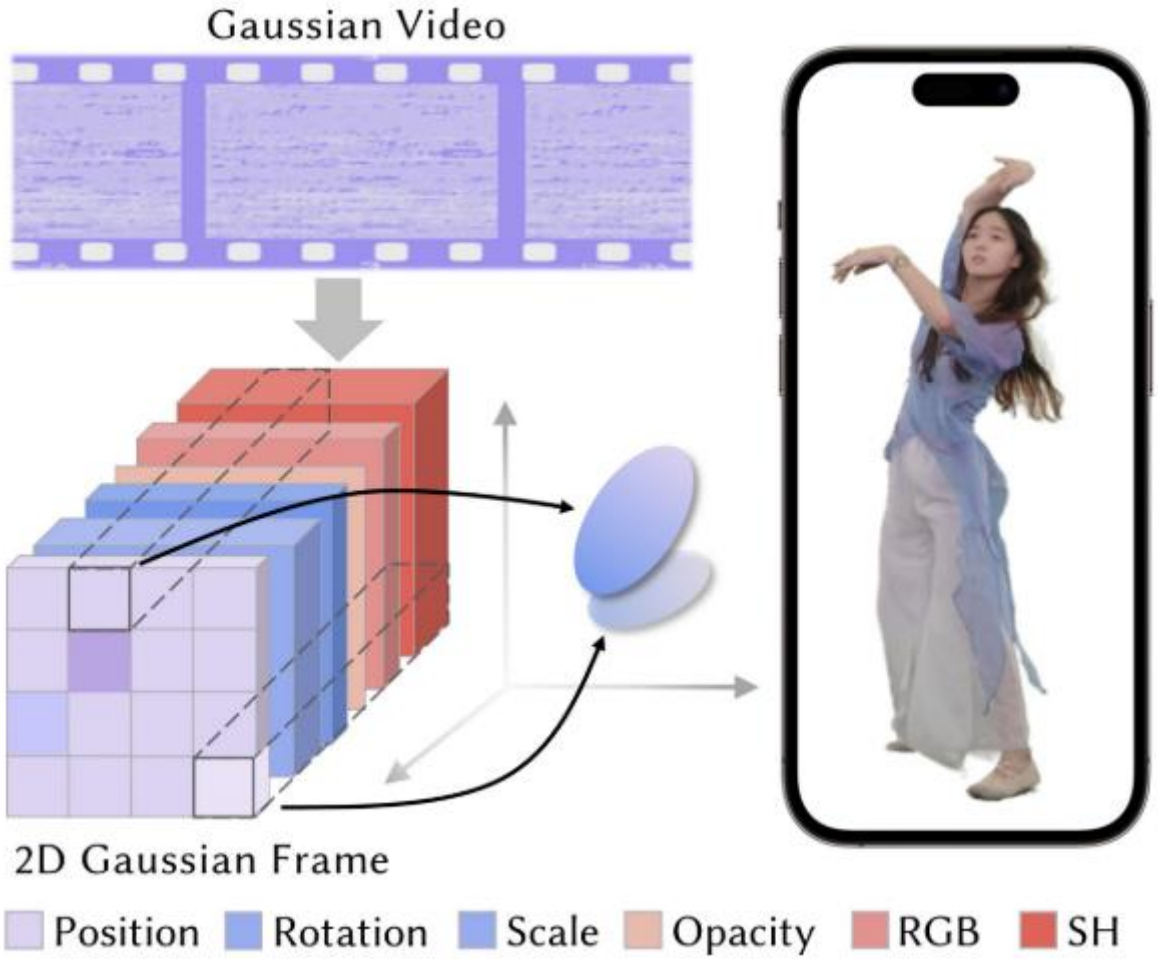
高斯初始化：高斯属性分两类，初始化后高斯数量不变。运动高斯随机初始化，在第一帧上训练；“皮肤”高斯则绑定到 k 最近的“关节”高斯进行训练。



高斯优化：固定 DualGS 的数量并优化高斯的运动以及皮肤高斯的外观。粗粒度专注优化运动，运动预测防止不合理运动，细粒度根据渲染损失优化关节高斯运动和皮肤高斯外观。渲染阶段使用关节高斯的运动对皮肤高斯的位置和旋转进行插值

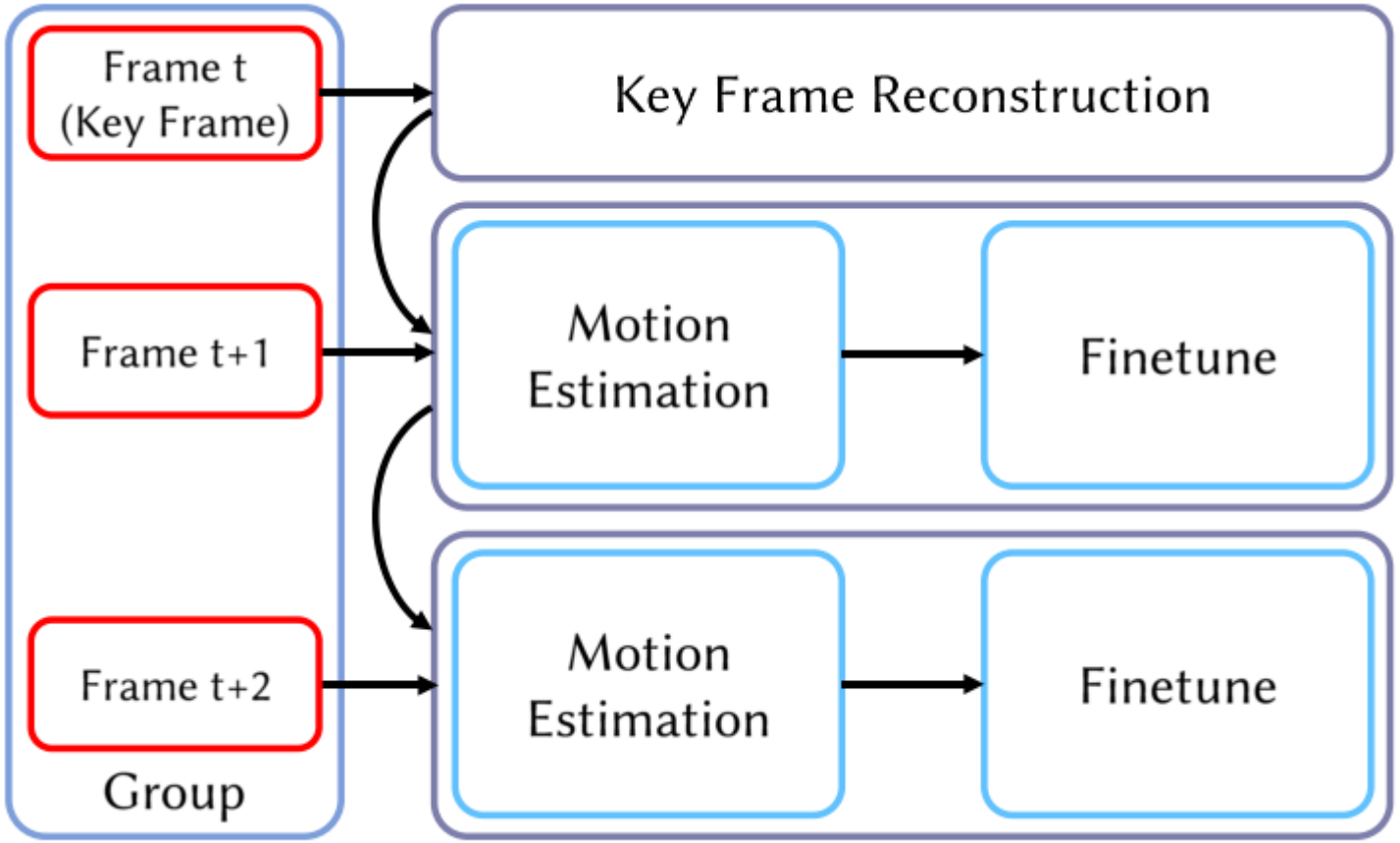
(SIGGRAPH ASIA 2024)

V3: Viewing Volumetric Videos on Mobiles via Streamable 2D Dynamic Gaussians



2D图表示：一种对动态场景的紧凑表示
高斯图的每个属性作为一个2D图，每个网格表示一帧的场景（本质上还是逐帧存储）

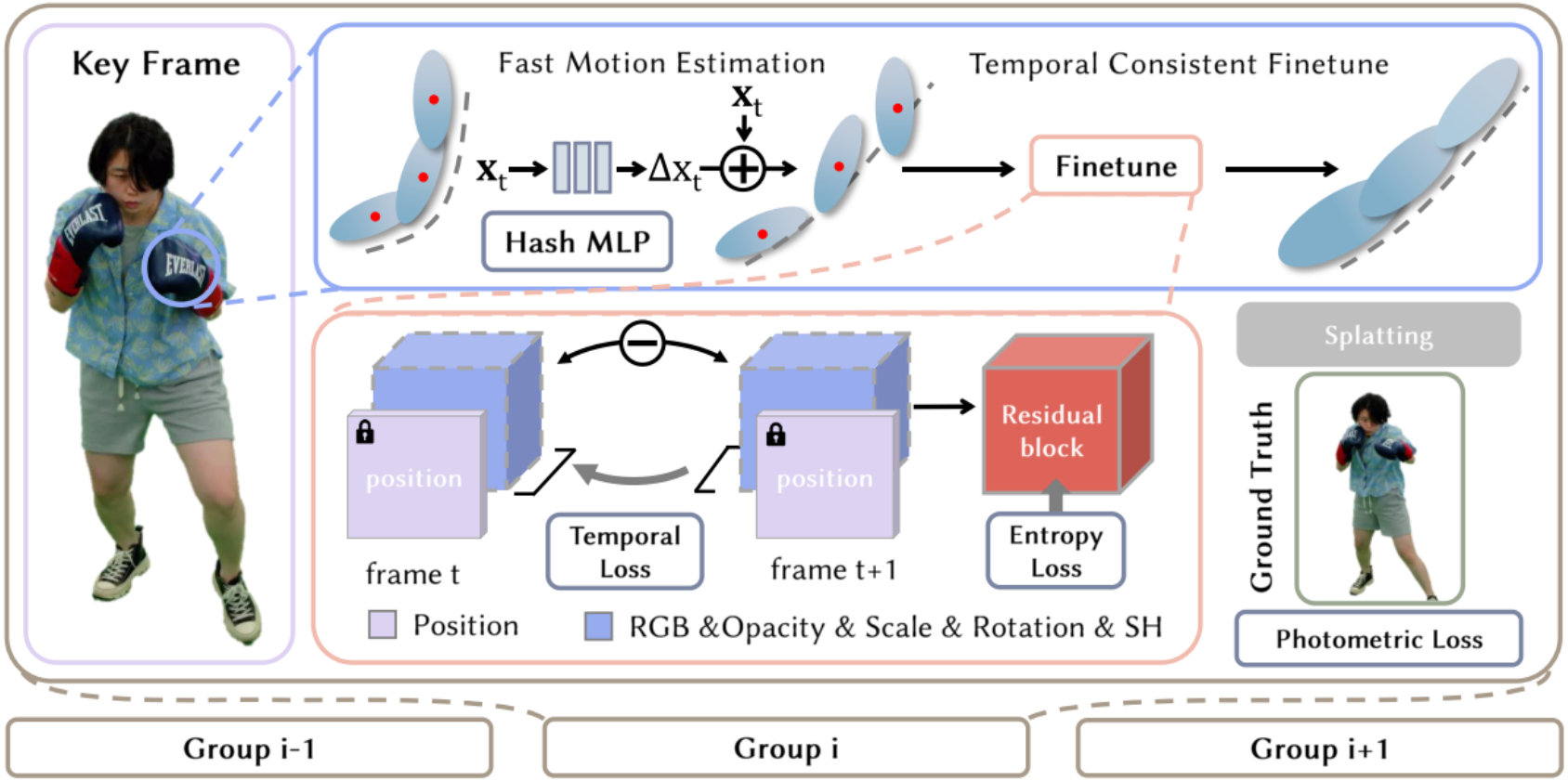
(SIGGRAPH ASIA 2024)



重建过程——考虑相邻帧之间高斯属性的连续性，避免逐帧重复训练

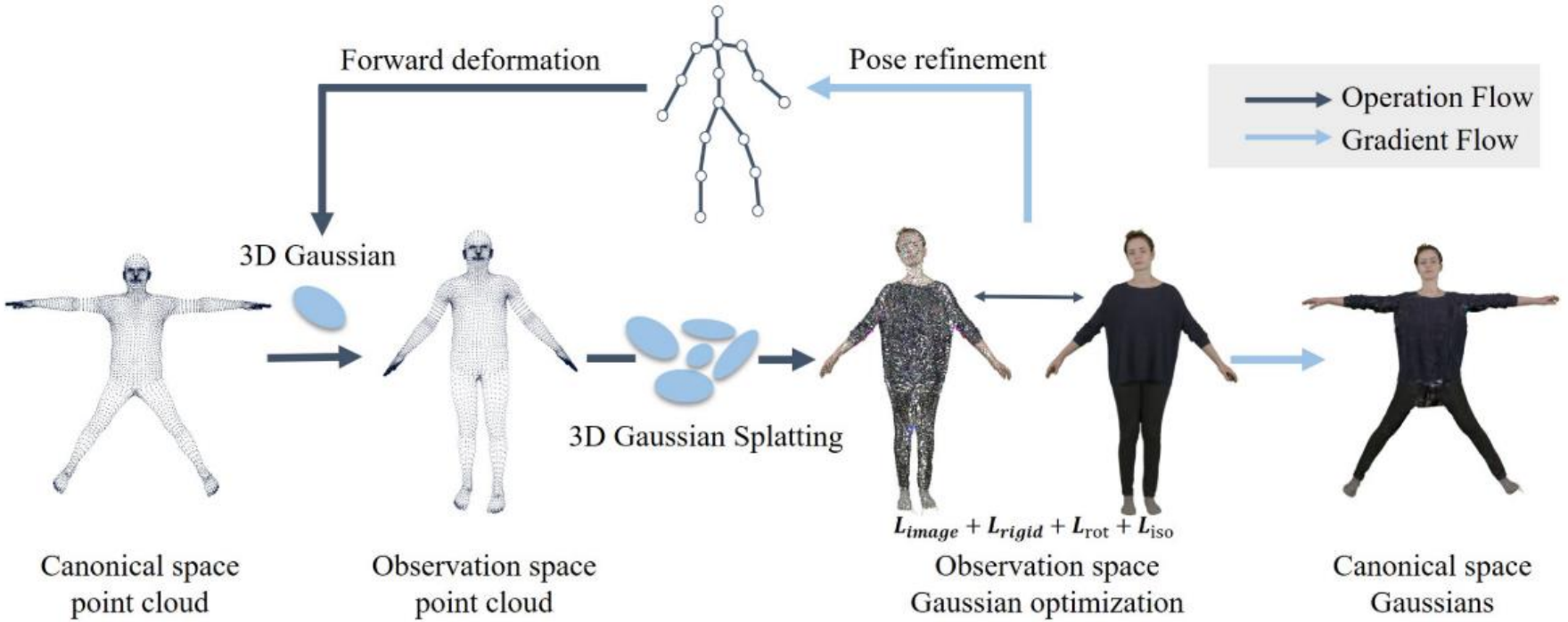
分组训练策略：20帧一组，取其中关键帧进行静态重建（并附带修剪），其余帧根据最后一帧的运动来扭曲模型,使用小型 MLP 的哈希编码来快速建模高斯随时间的顺序位置变化 ,保证帧与帧之间恒定高斯数量

两阶段训练策略：将运动和外观分开
第一阶段采用**哈希编码和浅层MLP** 来学习运动，然后通过剪枝减少高斯数量；
第二阶段使用**残差熵损失和时间损失**微调其他高斯属性以提高时间连续性



为了减少存储，**删除不透明度较低的点**来将高斯的数量控制在 100k 以下，根据不透明度对关键帧的图进行排序，修剪冗余点，并对属性进行微调——**其他论文表明剪枝最低不透明度的30%高斯不会影响场景质量**

GaussianBody: Clothed Human Reconstruction via 3d Gaussian Splatting



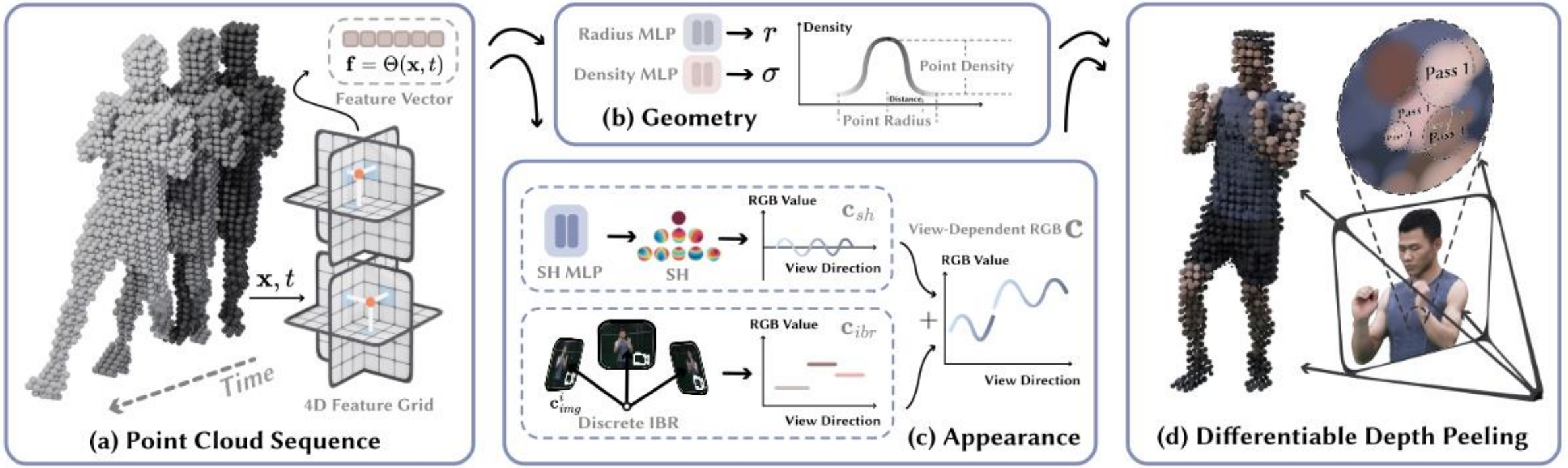
整体框架：将动态服装人体建模问题分解为**规范空间和运动空间**，定义模板3D高斯函数（用**SMPL模型**），利用位姿引导变形场将其转换为观测空间，记录每次姿态变换的梯度，并用于正则空间的反向优化

定义**尺度阈值**分割大高斯球，将SMPL位姿参数θ指定为优化参数（**定义相应损失**）

04

其他方法

4K4D: Real-Time 4D View Synthesis at 4K Resolution



(a)空间雕刻算法提取场景的粗点云：4D嵌入（**K-Plane**）通过定义六个平面给任意点分配特征f

(b)几何模型:动态场景几何通过学习每个点上的(**位置**、半径、密度)来表示，将f输入MLP中预测半径r和密度σ

(c)外观模型:图像混合技术(离散视图-C_{ibr})+球谐模型(连续视图-C_{sh})

(d)可微深度剥落:自定义K通道渲染器,使用硬件光栅化器将点云渲染到图像上,将点x_0的深度表示为t_0， 随后，在第k个渲染通道中，所有深度值t_k小于前一通道t_{k-1}记录深度的点都被丢弃，从而得到该像素的第k个最接近相机的点x_k，基于点$\{x_k|k = 1, \ldots, K\}$，使用体渲染来合成像素的颜色

(CVPR 2024)

4K4D-加速改进和实验结果

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

	PSNR ↑	SSIM ↑	LPIPS ↓	FPS
ENeRF [43]	28.108	0.972	0.056	6.011
IBRNet [85]	27.844	0.967	0.081	0.100
KPlanes [17]	27.452	0.952	0.118	0.640
Im4D [42]	28.991	0.973	0.062	15.360
Ours	31.173	0.976	0.055	203.610

（1）**推理前预先计算了点位置p、半径r、密度σ、SH系数s和颜色混合权值w_i**，并存储在主存，推理所需时间主要限制于深度剥离预测和球面谐波评估

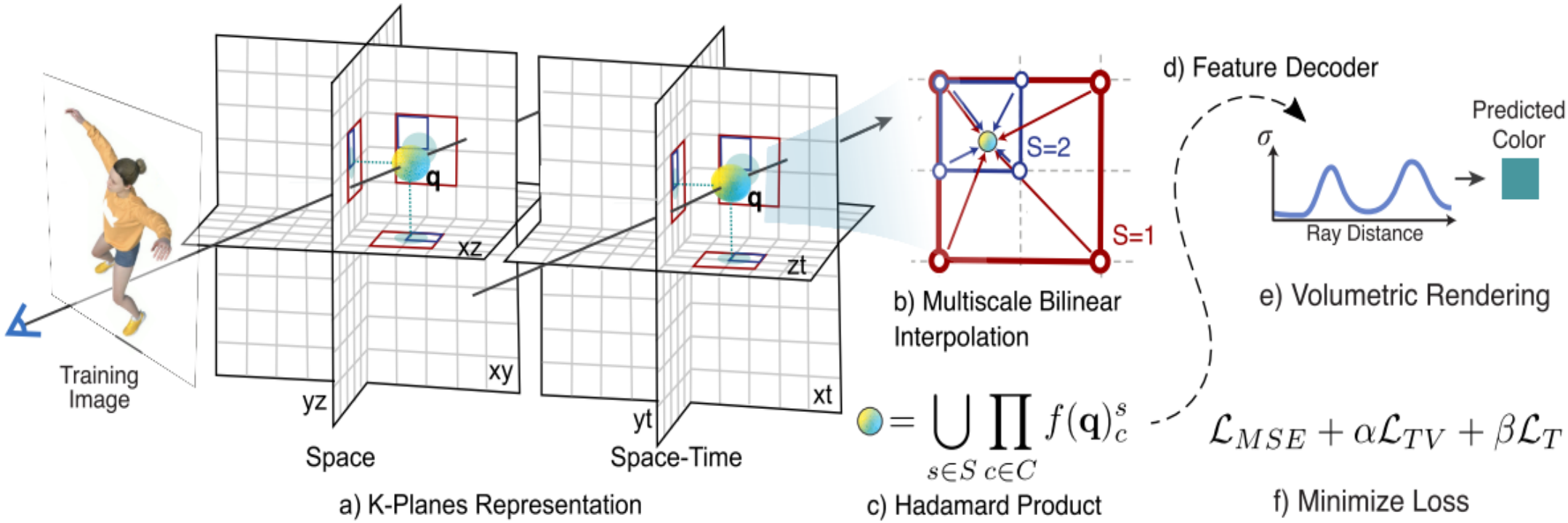
（2）将模型从32位浮点数转换为16位，高效内存访问

（3）可微深度剥离算法的渲染通道K从15次减少到12次，在没有视觉质量变化的情况下实现20帧/秒的加速

	Point Positions $\mathbf{p}$	4D Embedding Θ	MLPs and CNNs	Total Model Size	Encoded Video	Total Storage (*w/* Videos)
Storage	208.09 MB	16.77 MB	0.10 MB	224.96 MB	62.89 MB	287.86 MB
Storage / Frame	1.387 MB	0.112 MB	0.001 MB	1.500 MB	0.419 MB	1.919 MB

经过实验对比，**预计算**加速了**10倍**左右、**可微深度剥落（DDP）**比基于cuda的方法快**7倍**

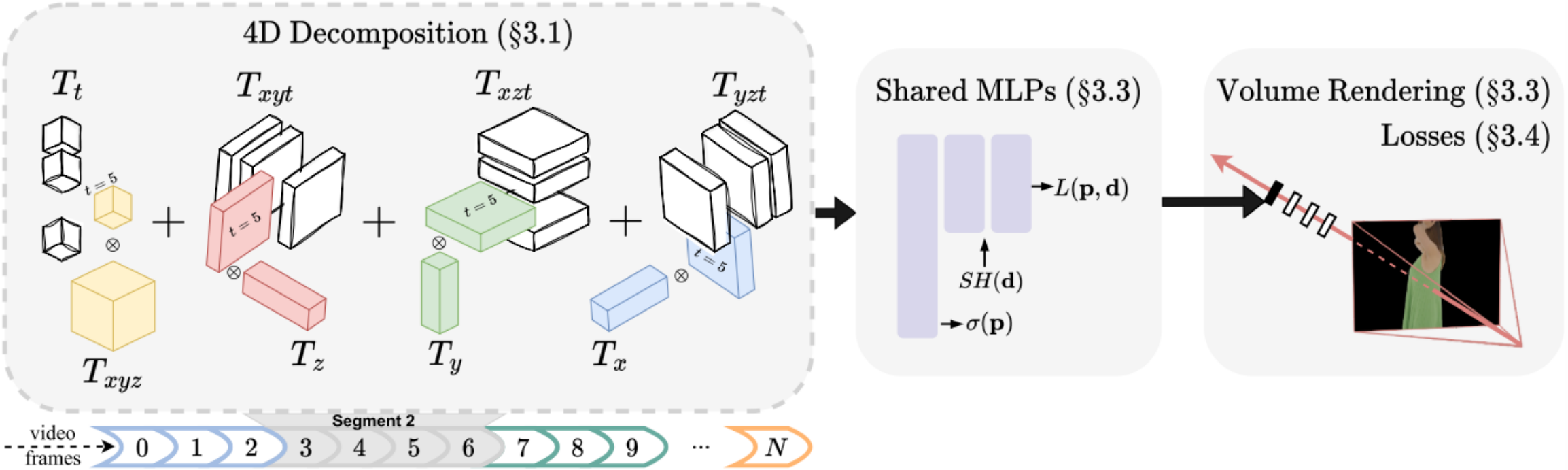
K-Planes: Explicit Radiance Fields in Space, Time, and Appearance



(a)将4D动态体积分解为**六个平面**，三个用于空间，三个用于时空变化
(b)对于任意四维点q = (x, y, z, t)的，将这个点**投影**到每个平面上，通过做多尺度**双线性插值**得到各平面的值
(c)插值值相乘，然后在**S尺度上连接**
(d)用小型MLP或显式线性解码器解码特征
(e)体绘制预测光线和密度 **(f)**在空间和时间上进行简单正则化，使重建损失最小化来优化模型

(CVPR 2023)

HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion



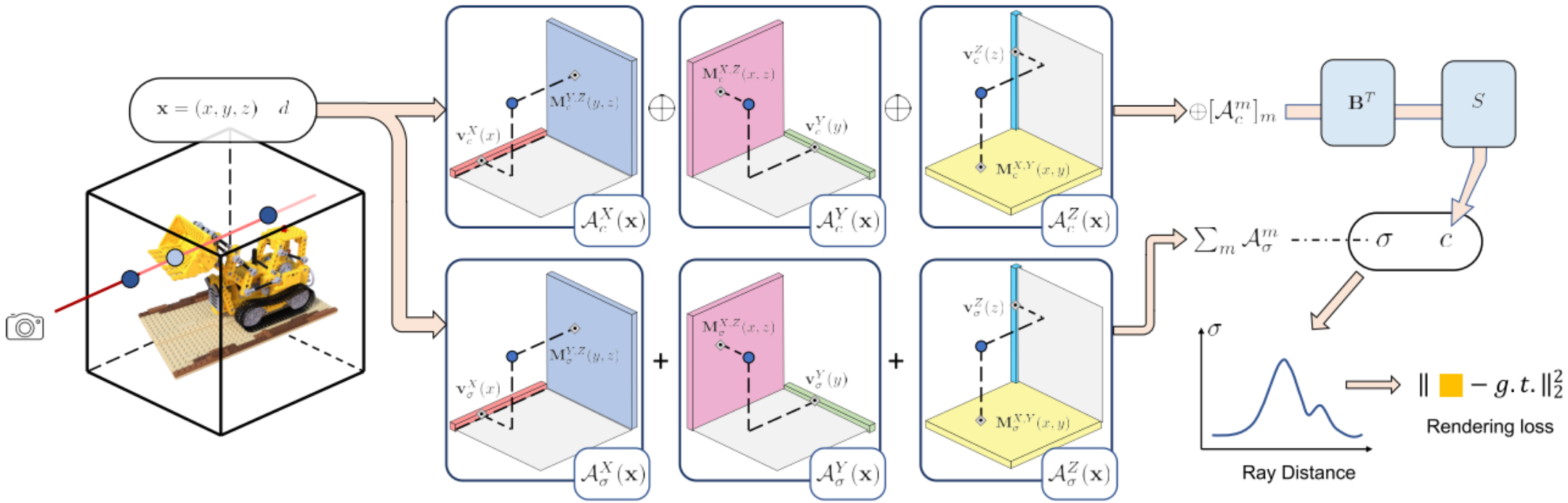
(a) 4D特征网格分解：使用四个3D网格和四个1D网格表示(TensoRF)

(b)自适应时间分区:考虑到使用单个4D段表示长序列不如使用多个固定大小的段 ,基于贪心算法计算累计空间占用，超过指定阈值生成一个新段（确保每个片段在4D空间中表示相似的体积）

(d)外观预测：使用共享的MLP预测体密度和颜色

(e)体渲染和优化

TensoRF：Tensorial Radiance Fields



将**4D场景张量分解**成多个紧凑的低秩张量分量

简要总结

- **GauHuman**：
  - 在规范空间中对高斯编码，并通过线性混合蒙皮（LBS）将3D高斯从规范空间转换到姿势空间，以此达成对场景每一帧的重建（也算是人体外观和运动分离）
  - 如何让人体先验（人体模型和运动动作）在初始化/优化过程中发挥更大的作用？
  - 能否把一些针对高斯本身的优化融入进去（比如特殊的自适应）？
  - 更多**4DGS**甚至**高斯数字人**相关论文….

请批评指正

Thanks for Your Attention