Direct Reward Fine-Tuning on Poses for Single Image to 3D Human in the Wild

Seoul National University

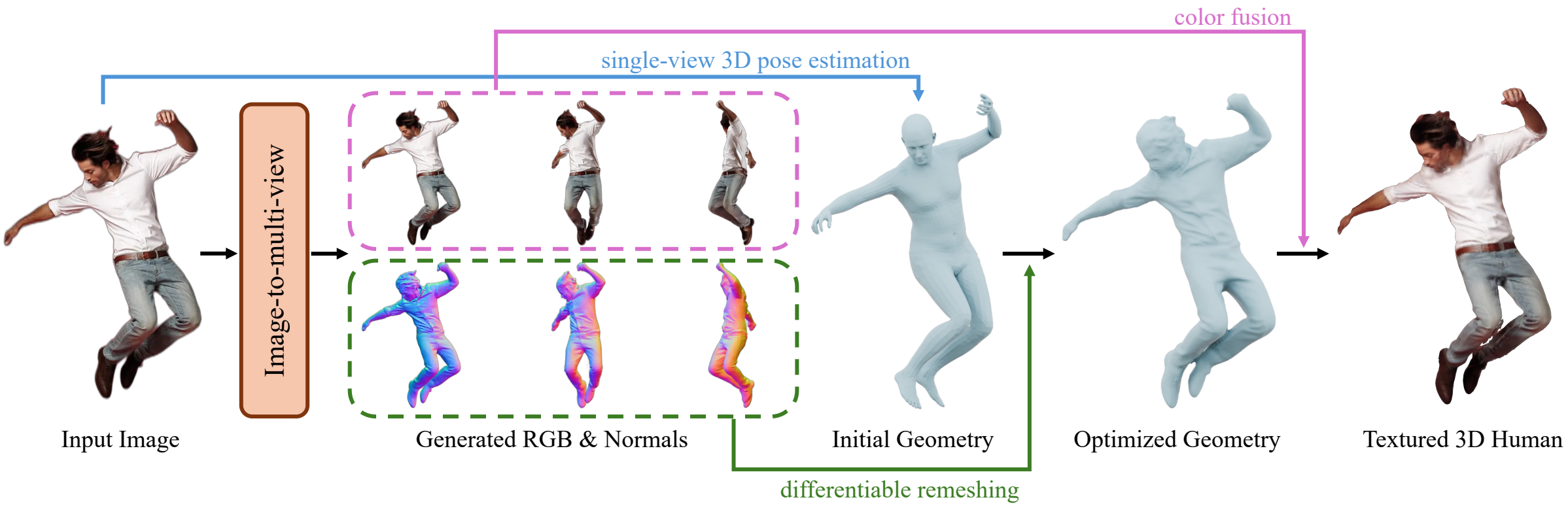
Seunguk Do

Minwoo Huh

Joonghyuk Shin

Jaesik Park

Diffusion-based Single-view 3D Human Reconstruction



Li, Peng, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi et al.
"Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing." CVPR. 2025.

Limitation of Current Diffusion-based Approaches

- Current diffusion-based 3D human reconstruction methods **struggle with dynamic poses**
- Root cause: **limited scale and diversity of poses** in publicly available 3D human scan datasets
  - Lack of diverse, or extreme body configurations and limited scale of Existing datasets (e.g., THuman2.1[1], CustomHumans[2])

Input Image

Reconstructed 3D Human

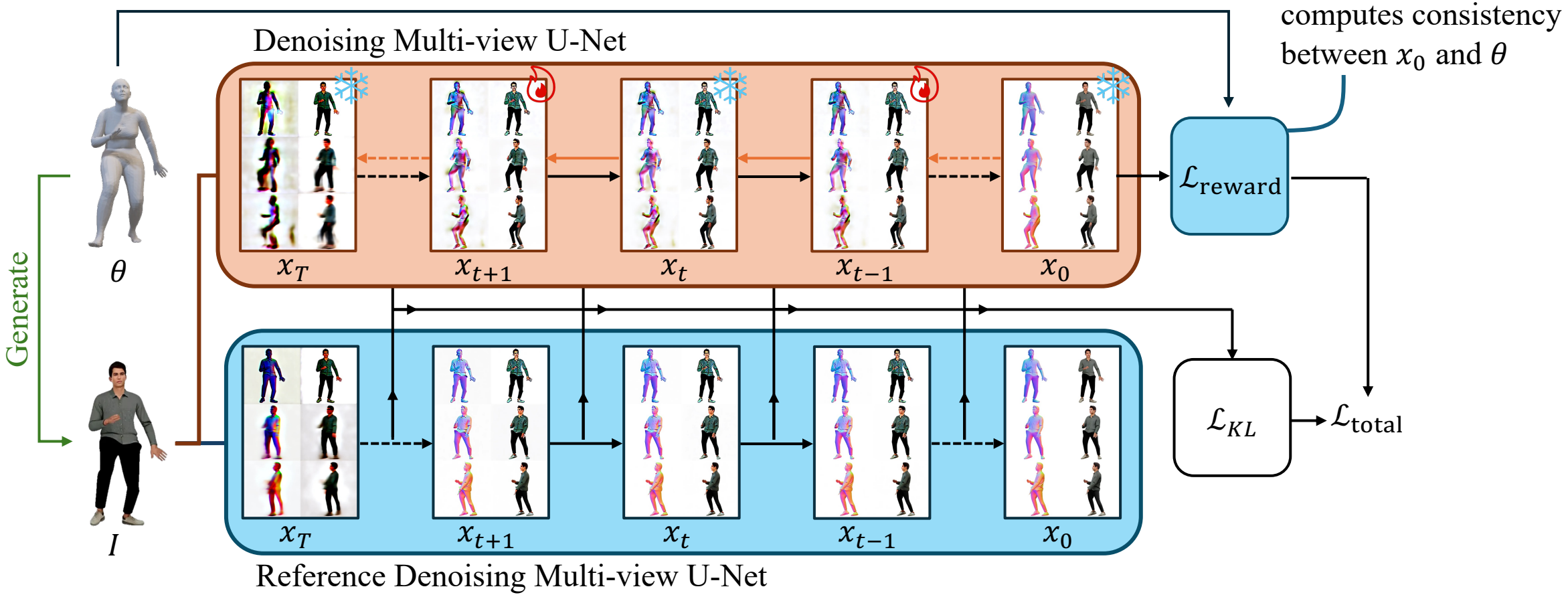
Input Image

Reconstructed 3D Human

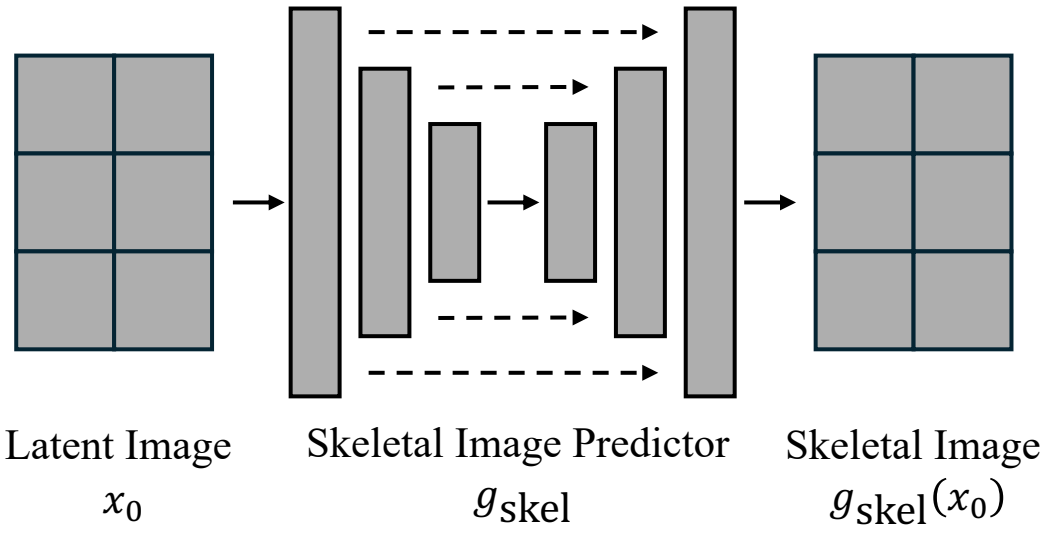
[1] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu.
Function4d:Real-time human volumetric capture from very sparse consumer rgbd sensors. CVPR, 2021

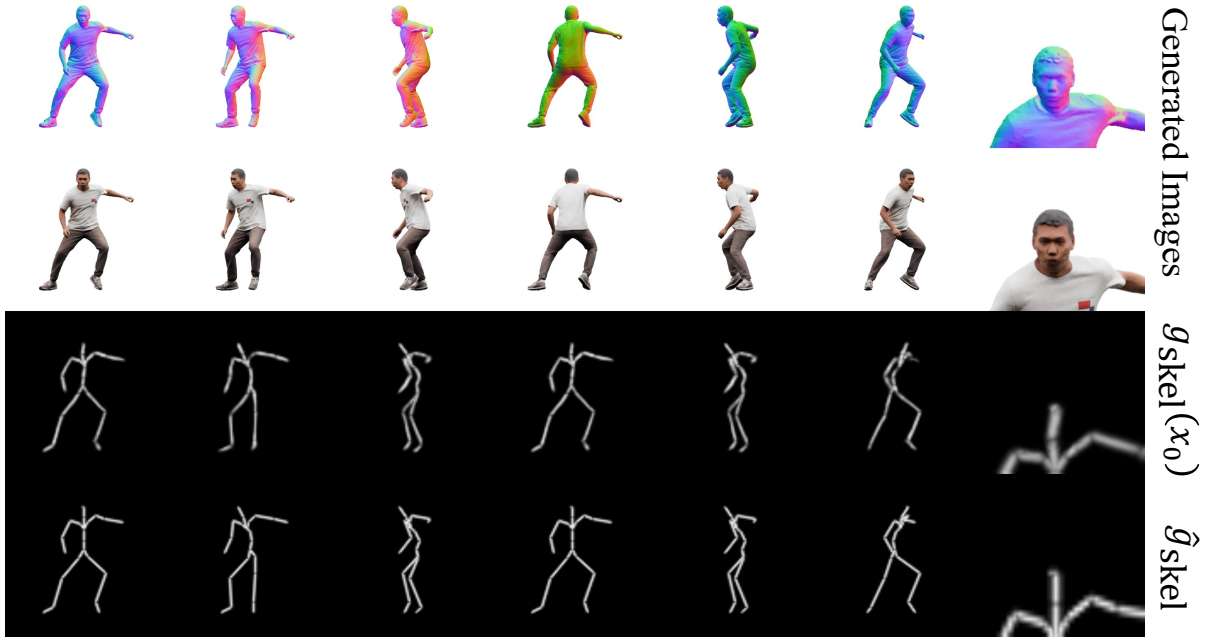
[2] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans., CVPR, 2023

DrPose

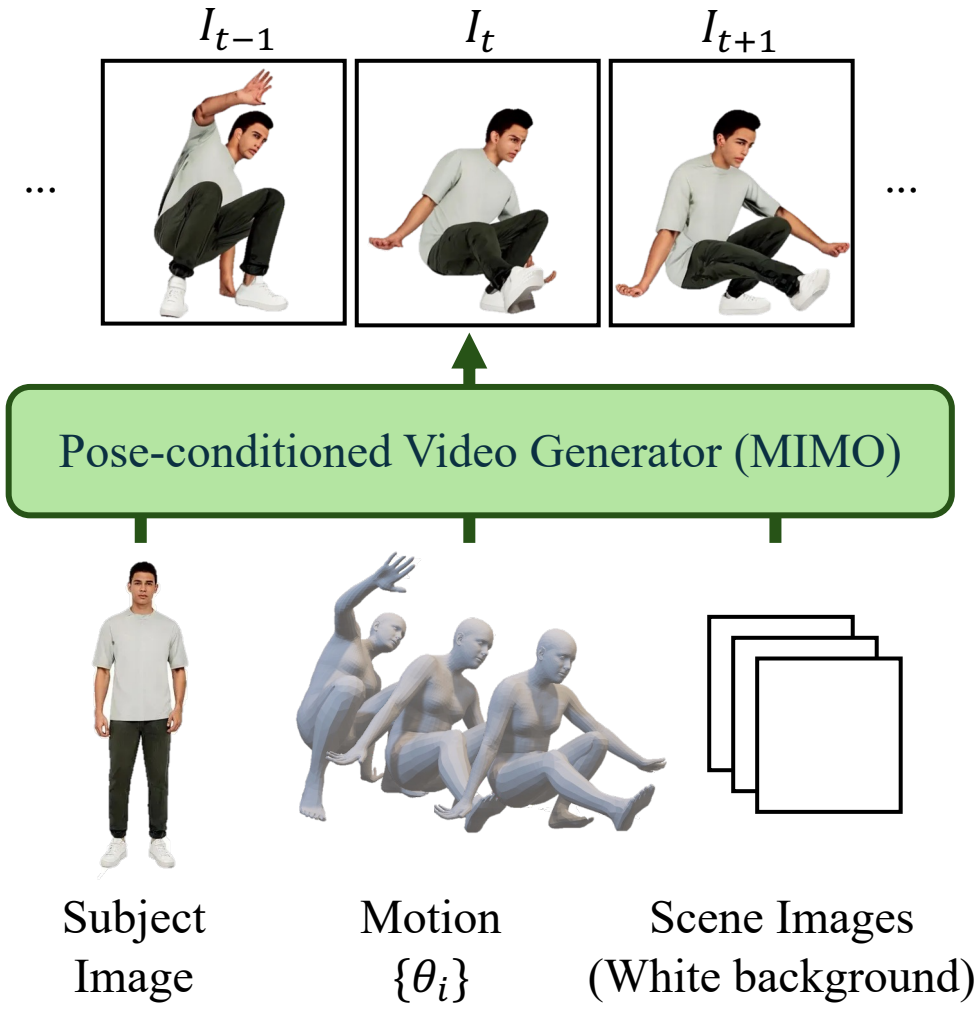

Denoising Multi-view U-Net
computes consistency between x_0 and θ
θ
Generate
I
x_T
x_{t+1}
x_t
x_{t-1}
x_0
$\mathcal{L}_{\text{reward}}$
$\mathcal{L}_{KL}$
$\mathcal{L}_{\text{total}}$
Reference Denoising Multi-view U-Net
trainable module
pretrained module

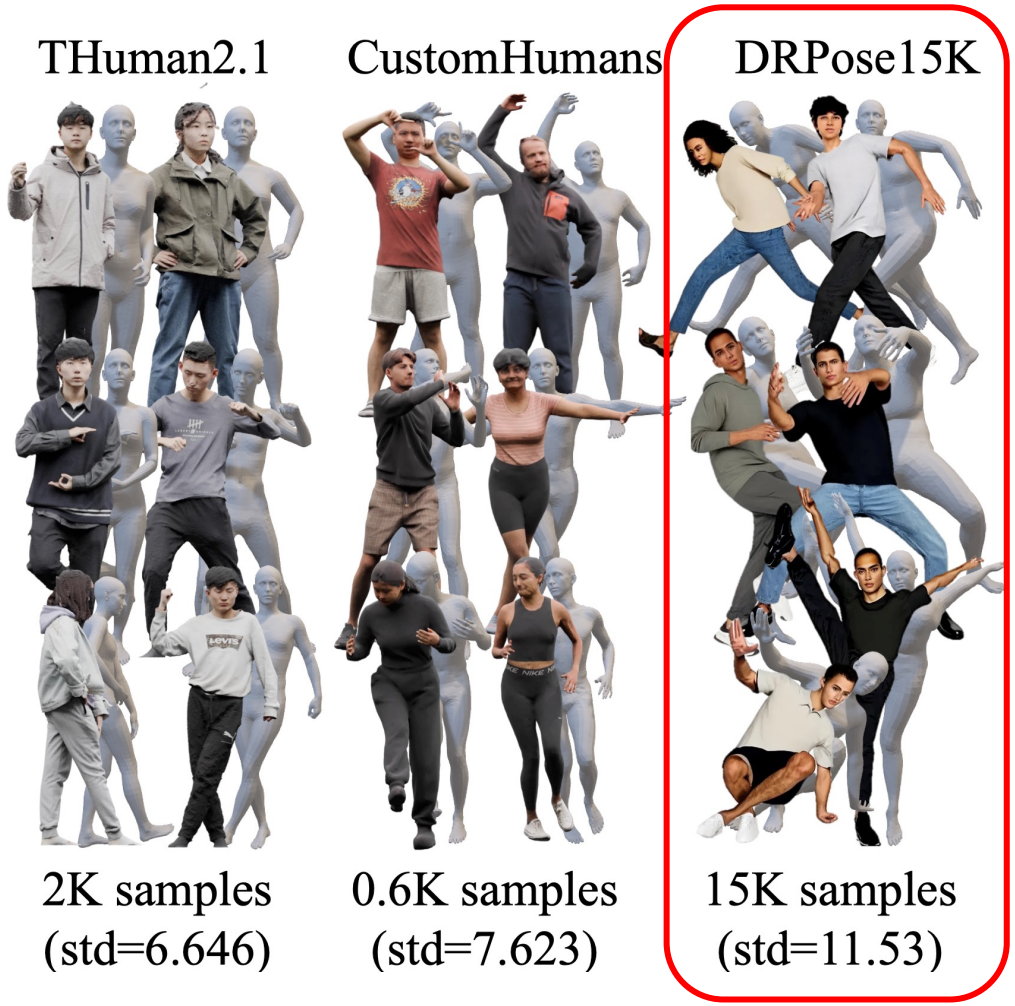
Reward Loss: $\mathcal{L}_{\text{reward}} = \mathbb{E}\left(||g_{\text{skel}}(x_0) - \hat{g}_{\text{skel}}||\right)$





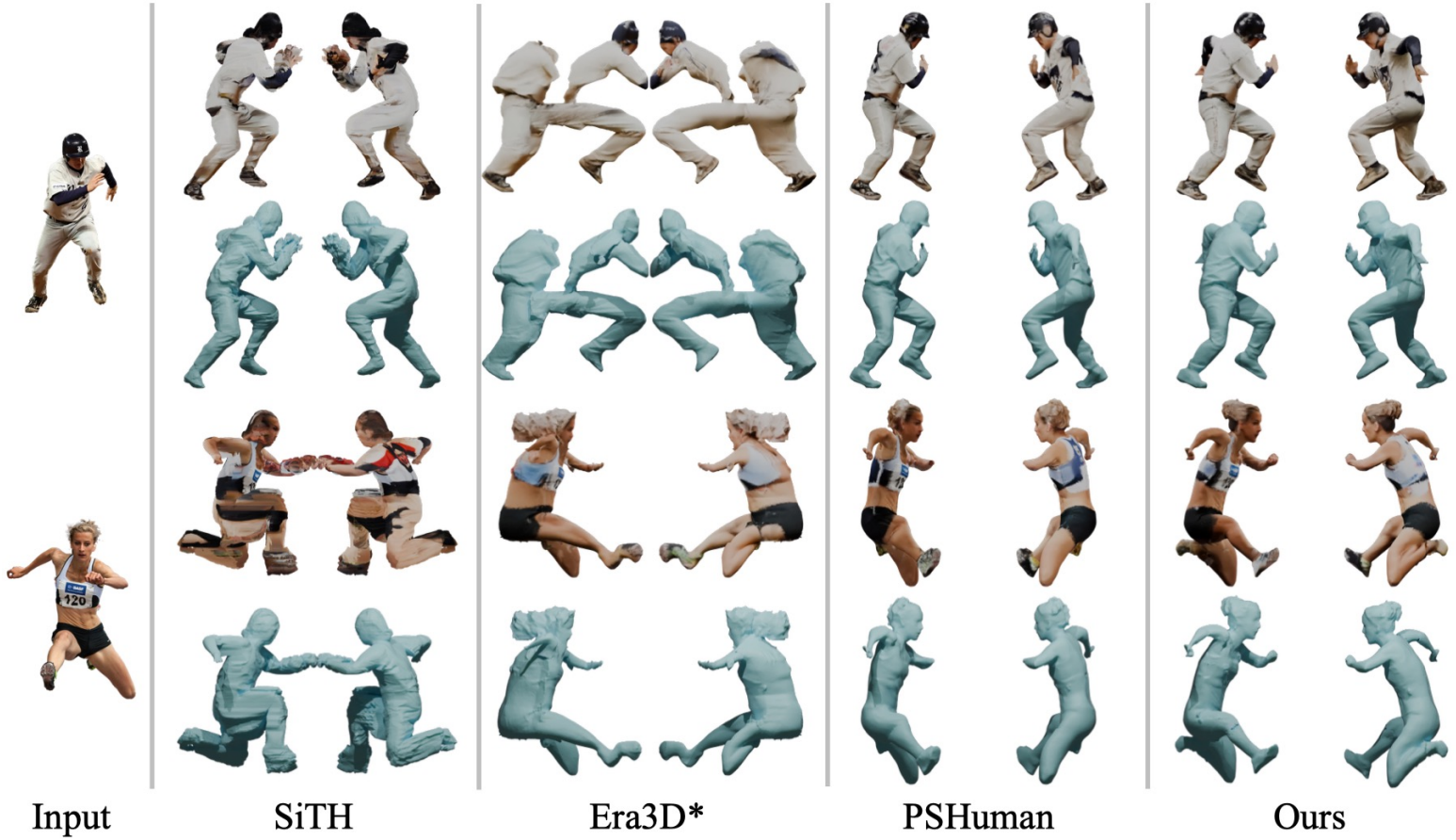
Construction of DrPose15 from Motion Data





Men, Yifang, Yuan Yao, Miaomiao Cui, and Liefeng Bo. "Mimo: Controllable character video synthesis with spatial decomposed modeling." *CVPR*. 2025.

Qualitative Results



Era3D* is a finetuned Era3D model on human datasets (THuman2.1 and CustomHumans).

Quantitative Results

	THuman2.1-test			CustomHumans-test			MixamoRP		
Method	CD↓	NC↑	f-Score↑	CD↓	NC↑	f-Score↑	CD↓	NC↑	f-Score↑
ECON	101.6465	0.6311	8.5244	126.1430	0.6205	6.4700	166.5384	0.5705	5.2173
SiTH	63.3041	0.6790	14.9221	71.9378	0.6713	12.7957	158.2729	0.5685	6.5176
H3D	75.8328	0.5959	12.2189	94.0864	0.5872	10.5563	149.2832	0.5400	7.2219
Era3D*	55.4071	0.6976	16.2928	63.1260	0.6914	14.1348	150.0118	0.5916	7.3487
PSHuman	52.9643	0.7194	18.6308	52.2187	0.7272	18.5624	137.2814	0.5876	8.2065
Ours (Era3D)	41.1770	0.7265	20.7102	44.3811	0.7310	20.1153	126.0622	0.5887	8.3071
Ours	42.0529	0.7252	19.6811	44.1326	0.7336	18.9600	126.5312	0.5998	8.8185

	THuman2.1-test			CustomHumans-test			MixamoRP		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SiTH	17.7473	0.8829	0.1371	18.7462	0.8599	0.1533	15.6096	0.8118	0.2182
H3D	17.3290	0.8656	0.1561	17.6244	0.8328	0.1882	15.2037	0.7915	0.2440
Era3D*	17.7662	0.8826	0.1410	18.6322	0.8581	0.1584	17.5337	0.8623	0.1519
PSHuman	18.3880	0.8922	0.1286	18.9082	0.8612	0.1538	17.5931	0.8638	0.1499
Ours (Era3D)	20.3563	0.9050	0.1134	18.9286	0.8644	0.1508	17.5064	0.8662	0.1474
Ours	20.8594	0.9078	0.1063	19.1887	0.8638	0.1476	17.6632	0.8646	0.1465

Era3D* is a finetuned Era3D model on human datasets (THuman2.1 and CustomHumans).

Contributions

- We propose **DrPose**, a novel **post-training** algorithm for aligning image-to-multi-view images to **diverse human poses**.
- We construct **DrPose15K**, a dataset comprising 15K diverse human poses paired with generated single-view images.
- We demonstrate our method achieves **consistent improvements** across three benchmarks.