Learning Relative Depth Guidance for Human Pose Transfer

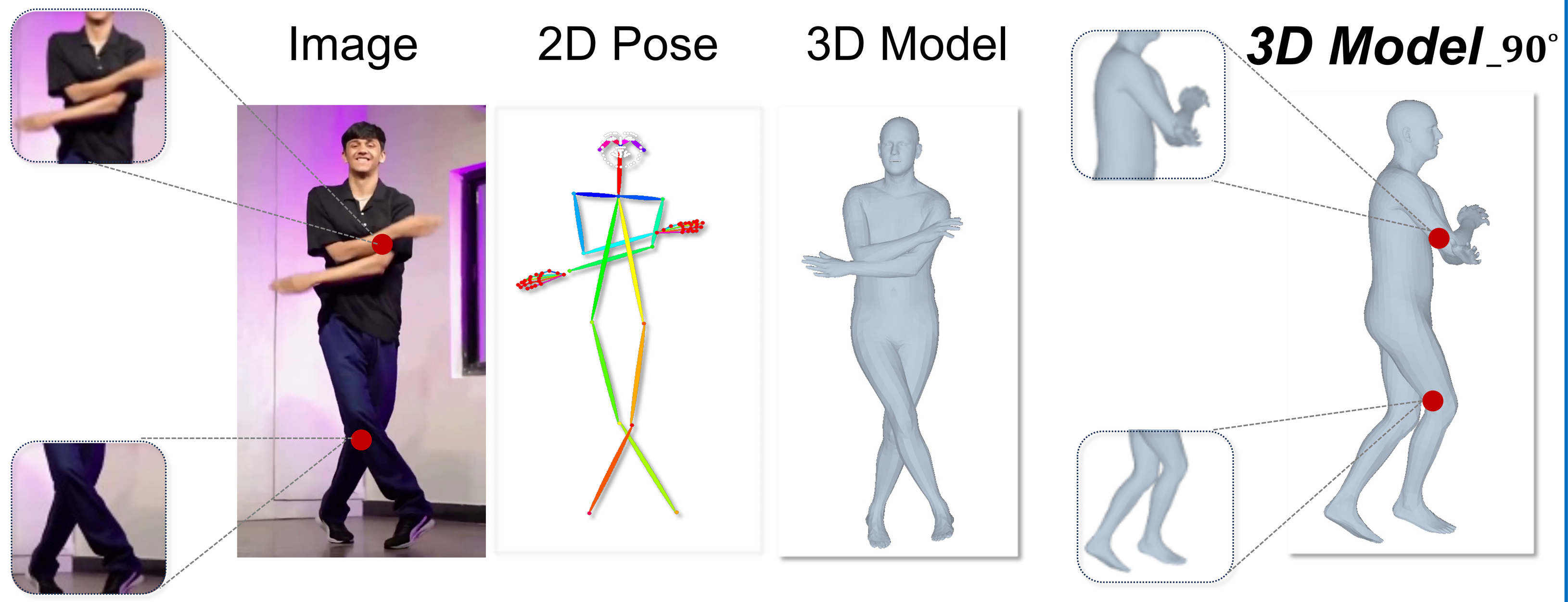
Lihan Zhang and Ciyu Ruan

TBSI, Tsinghua University, China

Abstract

- **Goal:** Propose a 3D parametric model-driven network for precise human pose transfer.
- **Limitations:** Existing methods project 3D parameters to 2D or combine 3D models with rendering, not fully utilizing 3D information.
- **Problem:** Depth ambiguity in human pose transfer leads to sub-optimal results.
- **Solution:** Integrate 3D parametric models within the diffusion framework to control pose and eliminate depth ambiguity.

Introduction



Core Problem: Depth ambiguity

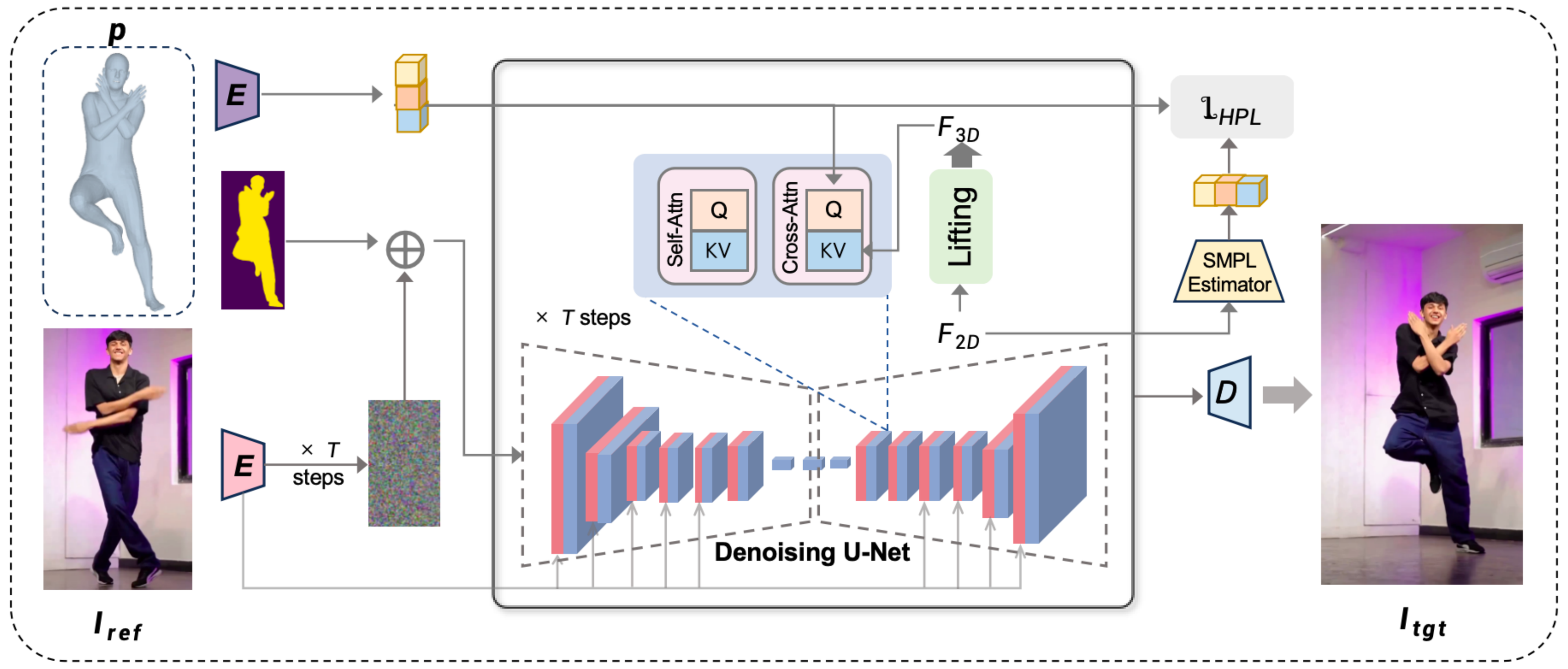
Input:

A source image of a person and a target pose represented by 3D mesh.

Output:

A generated image of the person in the source image, transformed to adopt the target pose.

Methods



- Stable Diffusion: $\mathcal{L}_{SD} = E_{z_t,c,\epsilon,t}(||\epsilon - \epsilon_\theta(z_t,c,t)||_2^2)$
- 3D Feature Lifting: $F_{Med} = \phi(MLP(F_{2D}))$

$$F_{3D} = LN(MLP(F_{Med}))$$

$$\mathcal{L}_{HPL} = ||[\theta,\beta] - [\tilde{\theta},\tilde{\beta}]||_2$$

- 3D Control Guidance: $\theta, \beta, C = \mathcal{H}(I_{ref})$

$$F_p = \mathcal{F}_{proj}(\mathcal{F}_{Fourier}(p))$$

$$F_Q = \phi_Q(F_P), F_K = \phi_K(F_{3D}), F_V = \phi_V(F_{3D})$$

$$F_{Attn} = softmax(\frac{F_Q F_K^T}{\sqrt{d}}) \cdot F_V$$

- Training Objective:

$$\mathcal{L}_{Overall} = E_{z_t,c_{id},c_{pose},\epsilon,t}||\epsilon - \epsilon_\theta(z_t,c_{id},c_{pose})||_2 + \mathcal{L}_{HPL}$$

Quantitative Result

- Quantitative comparison with several state-of-the-art models.

Dataset	Method	Venue	FID↓	LPIPS↓	SSIM↑	PSNR↑
Deepfashion (256 x 176)	PATN [46]	CVPR 19'	20.728	0.2533	0.6714	-
	ADGAN [21]	CVPR 20'	14.540	0.2255	0.6735	-
	GFLA [29]	CVPR 20'	9.827	0.1878	0.7082	-
	PISE [41]	CVPR 21'	11.518	0.2244	0.6537	-
	DPTN [43]	CVPR 22'	17.419	0.2093	0.6975	17.811
	NTED [27]	CVPR 22'	8.517	0.1770	0.7156	17.740
	CASD [44]	ECCV 22'	13.137	0.1781	0.7224	17.880
	PIDM [1]	CVPR 23'	6.812	0.2006	0.6621	15.630
	PoCoLD [6]	ICCV 23'	8.067	0.1642	0.7310	-
	CFLD [19]	CVPR 24'	6.804	0.1519	0.7378	18.235
	CFLD* [19]	CVPR 24'	6.713	0.1704	0.7235	17.626
	Ours	-	**7.110**	**0.1582**	**0.7201**	**17.702**

- Ablation experiments on individual modules

Modules	FID↓	LPIPS↓	SSIM↑	PSNR↑
Baseline	7.841	0.4121	0.5192	11.494
+ DecoupAttn	6.178	0.3542	0.5655	12.834
+ 2D Prior	8.129	0.1852	0.7079	17.051
+ 3DCG	**7.110**	**0.1582**	**0.7201**	**17.702**

Qualitative Result

Future Work

- The appearance details such as hands and face need to be improved.
- The model need to be improved under the training of larger dataset of various poses.
- Our current effects still require the assistance of a 2D mask to be achieved, and in the future we will need to achieve true 3D control.

References

[1] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-Resolution Image Synthesis with Latent Diffusion Models[C/OL]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA. 2022..

[2] CHEONG S, MUSTAFA A, GILBERT A. UPGPT: Universal Diffusion Model for Person Image Generation, Editing and Pose Transfer[J]. 2023.