MoCaNet: Motion Retargeting in-the-wild via Canonicalization Networks

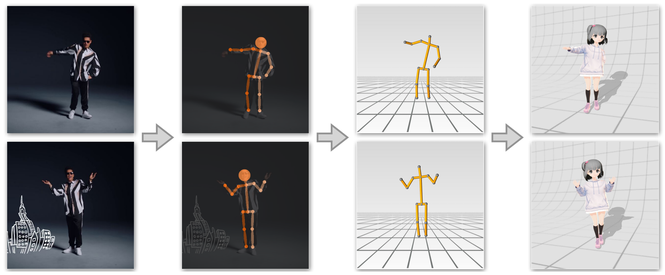
Wentao Zhu,[*1,2] Zhuoqian Yang,[*3] Ziang Di,[4] Wayne Wu,[+2,3] Yizhou Wang,[1] Chen Change Loy[5]

[1]School of Computer Science, Peking University [2]Shanghai AI Laboratory [3]SenseTime Research [4]Southeast University [5]S-Lab, Nanyang Technological University

AAAI-22

Problem

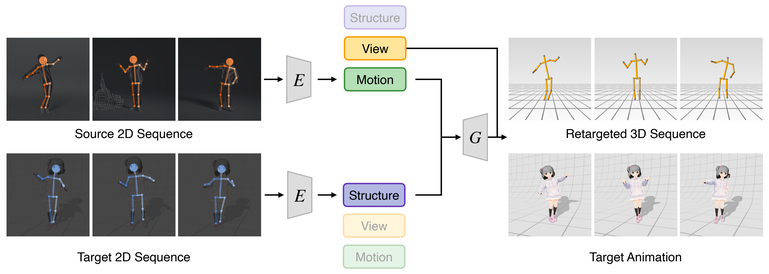
3D motion retargeting aims at transferring one character's motion to a virtual 3D avatar. In this work, we are interested in retargeting body motion from a character in a 2D monocular video to a 3D character without using any motion capture system or 3D reconstruction procedure.



Motion is extracted from the video clip and retargeted to the virtual avatar. Our method extracts 2D skeleton sequences rather than 3D sequences from in-the-wild videos.

Approach

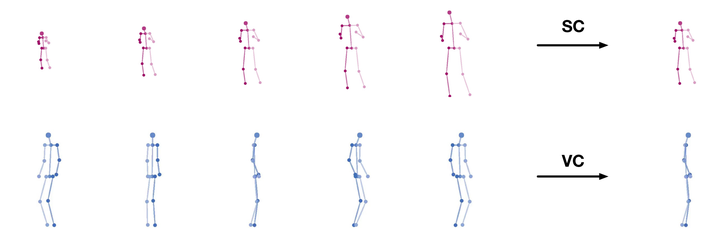
We propose an end-to-end learnable model, *Canonicalization Networks*, which allows us to bypass the error-prone direct 3D pose estimation and use an off-the-shelf 2D pose estimator to extract geometrical information that is more robust and reliable.



The encoder, E, decomposes the input skeleton sequence into three latent codes. The decoder, G, takes the motion code from the source character and the structure code from the target character, then cast the decoded 3D skeleton sequence to the source view angle. This yields the retargeted 3D sequence, which could be used to animate the target character.

Canonicalization Operation

Canonicalization aims at eliminating variations in a specific domain.



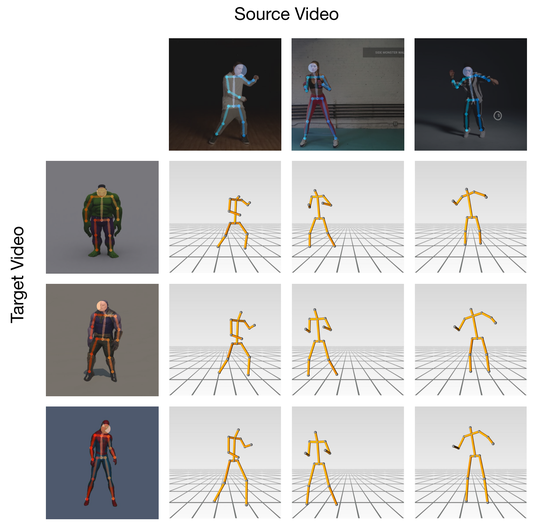
- *Structure Canonicalization* yields skeleton sequences with a uniform body structure, while the motion and other features are preserved.
- *View Canonicalization*. provides skeleton sequences with the same view angle, casting different sequences to a uniform view

Training the Canonicalization Networks

- Based on the canonicalization operations, we formulate a set of novel self-supervised canonicalization losses, which randomly perturb a targeted factor (structure or view angle subspaces) and apply canonicalization in that space for both the original sequence and the manipulated sequence.
- The model can be trained on an extensive collection of human pose sequences extracted from Internet videos **without any annotation**, which notably increases the robustness and generalization of the model.
- Please refer to the paper for details.

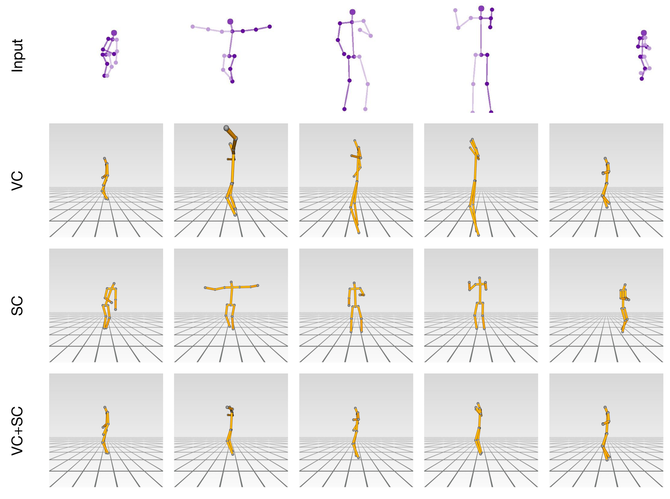
Results: 3D Motion Retargeting in-the-Wild

The network transfers the motion from source to target and produces 3D results.



Results: Canonicalization Effects

By canonicalization training, the model learns to disentangle the three latent factors without external supervision.



The first row is the input 2D skeleton sequences. The following rows show the 3D results after applying view canonicalization, structure canonicalization, and both.

Results: Quantitative Comparision

We calculate the 2D-to-3D motion retargeting error on a synthetic animation dataset called Mixamo because it has ground-truths available. Test characters are shown below.

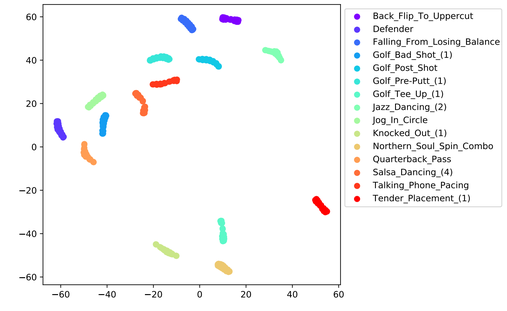
The table shows the 3D Mean Square Error (MSE $\times 10^{-2}$).

	3DGT+IK	Ours	Ours (wild)	Transmomo[4]+IK	LCN[1]+IK	P2M[2]+IK	VP3D[3]+IK
A ↔ PH	0.049	0.799	0.792	1.246	1.986	2.822	3.198
A↔ SG	0.042	0.948	0.945	1.668	1.973	2.504	5.309
A ↔ TY	0.016	0.853	0.832	1.497	1.769	1.910	4.044
PH ↔ SG	0.131	0.955	0.951	2.057	2.235	3.182	6.390
PH ↔ TY	0.064	0.874	0.853	1.837	2.135	2.544	5.471
SG ↔ TY	0.024	0.927	0.912	2.166	2.141	2.313	7.542
OVERALL	0.056	0.891	0.878	1.750	2.046	2.552	5.329

- 3DGT+IK: ground-truth 3D motion is retargeted with IK.
- Ours(wild) is trained on Solo-Dancer, an **unannotated** set of web-crawled videos which has only 30% of Mixamo training set video length. The result demonstrates the strength of training on in-the-wild data with higher motion diversity.

Results: Motion Clustering and Retrieval

The 3D skeleton sequence after both structure and view canonicalization contains unmixed motion information, independent of body structure and view angle. Therefore, it can be used as a direct, disentangled, and interpretable motion representation. We examine the practicability of using the dual-canonicalized skeleton sequence as a distilled motion representation.



	ARI	AMI	Homogeneity	Completeness	V-Measure
TransMoMo	0.241	0.620	0.657	0.756	0.703
Canonical	**0.347**	**0.707**	**0.750**	**0.808**	**0.778**

References

[1] Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Learning character-agnostic motion for motion retargeting in 2d. *ACM Trans. Graph.*, 38(4):75:1–75:14, 2019.

[2] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision (ECCV)*, 2020.

[3] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019.

[4] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. Transmomo: Invariance-driven unsupervised video motion retargeting. In *CVPR*, 2020.

* These authors contributed equally. + Correspondence author.