Virtual Pose Coach: A Motion-Retargeting Approach for Pose Training

Supplementary Material

1. Training Details

The motion data is normalized using the z-score method. We set the frame length t to 64 and employed the Adam optimizer with $\beta_1 = 0.9,\ \beta_2 = 0.999$ and a learning rate of 2.0e-4. The model was trained with a batch size of 64 for 4000 epochs.

2. Experiments Details

2.1. Brief Introduction to Baselines

NKN and PMnet utilize joint positions as their input. Specifically, NKN is tailored for intra-structural motion retargeting and employs two RNNs combined with a Forward Kinematics (FK) approach. It leverages cycle consistency and adversarial training to facilitate unsupervised learning, allowing for natural motion retargeting without the need for paired data from different skeleton structures. PMnet operates as an unsupervised motion retargeting framework that learns both frame-by-frame poses and the overall movement of a character. This framework independently addresses two components: one for the individual pose at every frame and the other for the character's total movement. It employs a pose encoder and mapping networks for pose retargeting, while a movement regressor and normalization process ensure realistic motions, regardless of character size. To enable the comparison under the cross-structural setting, we followed [11] by modifying the methods of NKN and PMnet. We refer to these modified versions as NKN* and PMnet*. On the other hand, SAN and PAN function with joint rotations as input. SAN introduces "skeleton-aware" differentiable convolution, pooling, and unpooling to facilitate cross-structural retargeting, considering the hierarchical structure and joint connectivity. It connects edges to transform various skeletons into a unified primal skeleton and embeds motions within a consolidated latent space. Meanwhile, PAN improves its ability to capture spatial motion features through an attention mechanism. By treating body segments as fundamental units, PAN dynamically predicts the importance of each joint within each segment, achieving state-of-the-art performance in cross-structural motion retargeting.

2.2. Evaluations Details

The Mixamo dataset [13], created by SAN [2], comprises 29 unique humanoid characters. These characters are categorized into two groups based on their skeletal structures. Group A features six additional joints compared to Group B, which are located in the limbs, torso, and head, as illustrated in Fig. 1. Group A consists of 24 characters, with 20 allocated for training and 4 for testing. On the other hand, Group B includes 5 characters, with 4 designated for training and 1 for testing.

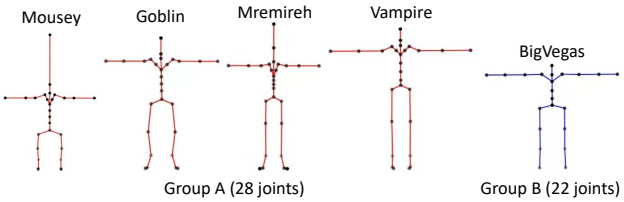

Figure 1. The testing characters can be grouped into two types based on their skeletal structures. Mousey, Goblin, Mremireh, and Vampire belong to Group A, while BigVegas is in Group B.

Fig. 2 presents the qualitative results comparing our method with PAN. Fig. 2 (a) displays the results for intra-structural retargeting, while Fig. 2 (b) illustrates the results for cross-structural retargeting.

2.3. Ablation Study Details

The ablation study examines five characters, as shown in Fig. 1. For intra-structural retargeting (see Tab. 2), we evaluate the following retargeting tasks: $G \leftrightarrow Mo$, $G \leftrightarrow Mr$, $G \leftrightarrow V$, $Mo \leftrightarrow Mr$, $Mo \leftrightarrow V$, and $Mr \leftrightarrow V$. Here, $\leftrightarrow$ indicates that retargeting occurs in both directions. For cross-structural retargeting (see Tab. 1), we investigate the following retargeting tasks: $B \rightarrow G$, $B \rightarrow Mo$, $B \rightarrow Mr$, and $B \rightarrow V$. In this context, $\rightarrow$ denotes the direction of retargeting.

Table 1. Ablation study on cross-structural retargeting.

(a) global root position errors ($\times 10^{-3}$).

Method	B→G	B→Mo	B→Mr	B→V	Overall
parent-based	**7.423**	**3.851**	**6.073**	**9.233**	**6.645**
root-based	59.672	29.943	47.159	78.224	53.750
use L'_{rec}	120.349	50.99	102.472	155.341	107.288
use L'_{cyc}	72.061	34.970	59.279	94.448	65.190

(b) local joint position errors ($\times 10^{-3}$).

Method	B→G	B→Mo	B→Mr	B→V	Overall
parent-based	**1.629**	1.373	1.084	**1.431**	**1.380**
root-based	2.641	1.138	1.058	2.339	1.794
use L'_{rec}	3.730	3.007	2.543	4.014	3.323
use L'_{cyc}	2.568	**0.953**	**0.984**	2.073	1.644

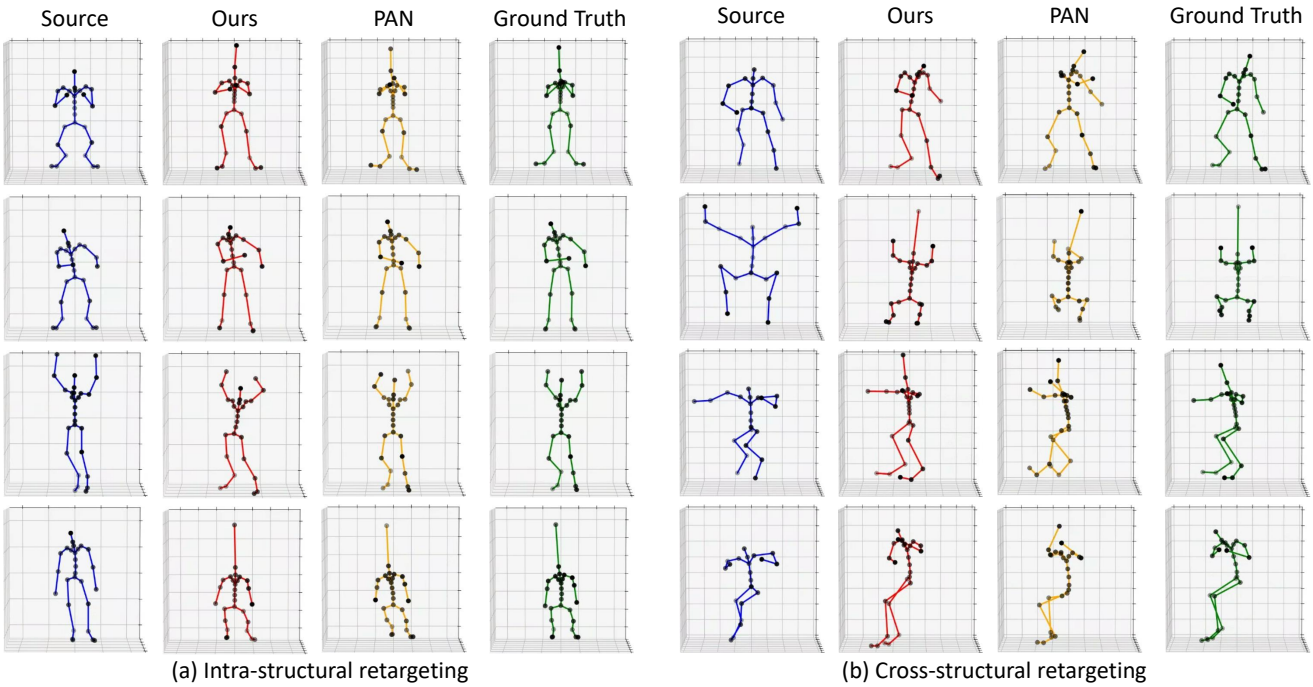

Figure 2. The visual comparison of retargeting results between PAN and our approach, encompassing both intra-structural and cross-structural conditions.

Table 2. Ablation study on intra-structural retargeting.

(a) global root position errors ($\times 10^{-3}$).

Method	G↔Mo	G↔Mr	G↔V	Mo↔Mr	Mo↔V	Mr↔V	Overall
parent-based	0.245	0.542	0.224	0.665	0.512	0.282	0.412
root-based	0.315	0.487	0.709	0.583	0.939	0.592	0.604
use L'_{rec}	**0.213**	**0.369**	**0.155**	0.527	**0.461**	**0.119**	**0.307**
use L'_{cyc}	0.219	0.410	0.200	**0.519**	0.480	0.147	0.329

(b) local joint position errors ($\times 10^{-3}$).

Method	G↔Mo	G↔Mr	G↔V	Mo↔Mr	Mo↔V	Mr↔V	Overall
parent-based	**1.737**	**1.62**	**2.178**	1.165	1.609	1.375	**1.614**
root-based	1.915	1.831	2.34	1.101	1.808	1.584	1.763
use L'_{rec}	1.953	1.783	2.484	0.999	1.727	**1.364**	1.718
use L'_{cyc}	1.739	1.753	2.489	**0.892**	**1.511**	1.431	1.636