Supplementary Material of Zero-shot Pose Transfer for Unrigged Stylized 3D Characters

Jiashun Wang[1] Xueting Li[2] Sifei Liu[2] Shalini De Mello[2]
Orazio Gallo[2] Xiaolong Wang[3] Jan Kautz[2]
[1]Carnegie Mellon University [2]NVIDIA [3]UC San Diego

In this supplementary, we introduce more details about the evaluation data curation procedure, the implementation of our method and the baseline methods, more qualitative results and the limitations of our method.

1. Evaluation Data Curation

Mixamo. Because the preprocessed Mixamo [1] testing sequences used in [5] are not publicly available, we follow the instructions in [5] and download the testing data from the Mixamo website [1]. In [5], 20 stylized characters and 28 motion sequences are used for evaluation. Among the 20 characters, the "liam" character is not publicly available on the Mixamo website, thus we evaluate our method and the baselines on the other 19 stylized characters. Moreover, some evaluation motions (e.g., "Teeter") include more than one motion sequence on the Mixamo website with the same name. However, it is not public information as to what exact sequences were used for evaluation in the prior work [5]. Thus, we download all motion sequences with the same name and randomly pick one for evaluation. Given a character in rest pose and the desired pose, we use the linear blend skinning algorithm to obtain the ground truth deformed mesh. We then compare the prediction from each method with the ground truth mesh by computing the PMD and ELS scores as discussed in Sec.4.3 in the main paper. For a fair comparison, all poses in the evaluation motion sequences are not used during training. All methods are evaluated using these collected testing pairs.

MGN. We follow NBS [3] and download the MGN dataset[1], which includes 96 clothed human characters. We use the same evaluation set (i.e., the last 16 human characters) as in NBS. To obtain the ground truth deformed characters, we sample 200 poses (unseen during training) and deform each of the 16 clothed characters using the MultiGarment Net [2].

Pose code extraction from Mixamo characters. To obtain target poses from the Mixamo motion sequences, we apply a similar fitting procedure introduced in [4]. We optimize the SMPL parameters to minimize the L2 distance between the SMPL joints and the Mixamo joints. Different from [4], we also add a constraint to minimize the Chamfer distance between the SMPL shape vertices and the Mixamo shape vertices. Similarly as [7], we directly optimize the pose code in the VPoser's [6] latent space, instead of the parameters in SMPL. We fit the SMPL shape to the "marker man" character in Mixamo to get all the testing poses.

2. Implementation Details

Shape code computation. We use an off-the-shelf method[2] that computes occupancy with "virtual laser scans" and does not require a watertight mesh. We sample 10,000 points in a unit space, which takes **2.35s** on average. Then, we use the occupancy of each query point as supervision to optimize the shape code. We run 2,000 iterations with a batch size of 2,000 to get the shape code, which takes **3.41s** on average. For each character, we only compute its shape code **once** and use it to transfer poses from different motion sequences. All the time cost reported in this supplementary was measured on a laptop with I7-11700h and a RTX 3060.

Detailed test-time training (TTT) procedure. Following the inference procedure in [5], TTT takes a stylized character in T-pose, and a source human character in T-pose and target pose as inputs. TTT finetunes the pose module to perform two tasks: a) the T-pose stylized character is deformed to the target pose, while being constrained by the self-supervised volume-preserving loss L_v. b) the source human character in T-pose is deformed to the target pose, while being supervised by the ground truth human character in the target pose (L_{dr}). TTT further refines the results' smoothness and resemblance to driving poses. L_{dr} helps the pose module understand and generalize to the target pose, rather than enforcing that the human and stylized character have similar offsets. TTT is carried out for each pair of stylized character and target pose. It is highly efficient and only requires fine-tuning the pose module for 20 iterations, which takes **18ms** without batching. We can

[1]https://github.com/bharat-b7/MultiGarmentNetwork

[2]https://github.com/marian42/mesh_to_sdf

speed it up to **12ms** for each pair with a batch size of 8.

3. Baseline Methods Implementation

NBS [3]. We evaluate NBS using its publicly available code and pre-trained model[3]. NBS [3] takes the SMPL pose parameters as input, thus we feed the optimized SMPL parameters discussed above to NBS.

SPT [5]. To evaluate both SPT(full) and SPT on human-like stylized characters, we use the publicly available code[4] and pre-trained models generously provided by the authors. For the quadruped category, we train and evaluate the SPT model using its public code on the dataset discussed in Sec.4.1 in the main paper. Specifically, we utilize the SMAL model [10] to produce motion pairs, including an animal mesh in rest pose and the desired pose. We also supervise SPT with the ground truth skinning weights from SMAL. Note that our model is trained and evaluated using the same quadruped dataset as SPT.

4. Visualization

We provide more visualizations, including qualitative comparisons (Fig. 1), deformation results by using source poses from in-the-wild videos for both human-like (Fig. 2 and Fig. 3) and quadupeds (Fig. 4). To obtain the pose code from a video frame, we apply PyMAF [9] for human and BARC [8] for quadupeds. We provide more visualizations in the supplementary video.

5. Limitation

Although our approach exhibits good generalization performance for bipedal and quadrupedal characters, modeling other categories whose poses are not being studied well remains difficult. Additionally, our method is unable to solve the articulation of hands and just treats them as rigid parts.

[3] https://github.com/PeizhuoLi/neural-blend-shapes
[4] https://github.com/zycliao/skeleton-free-pose-transfer

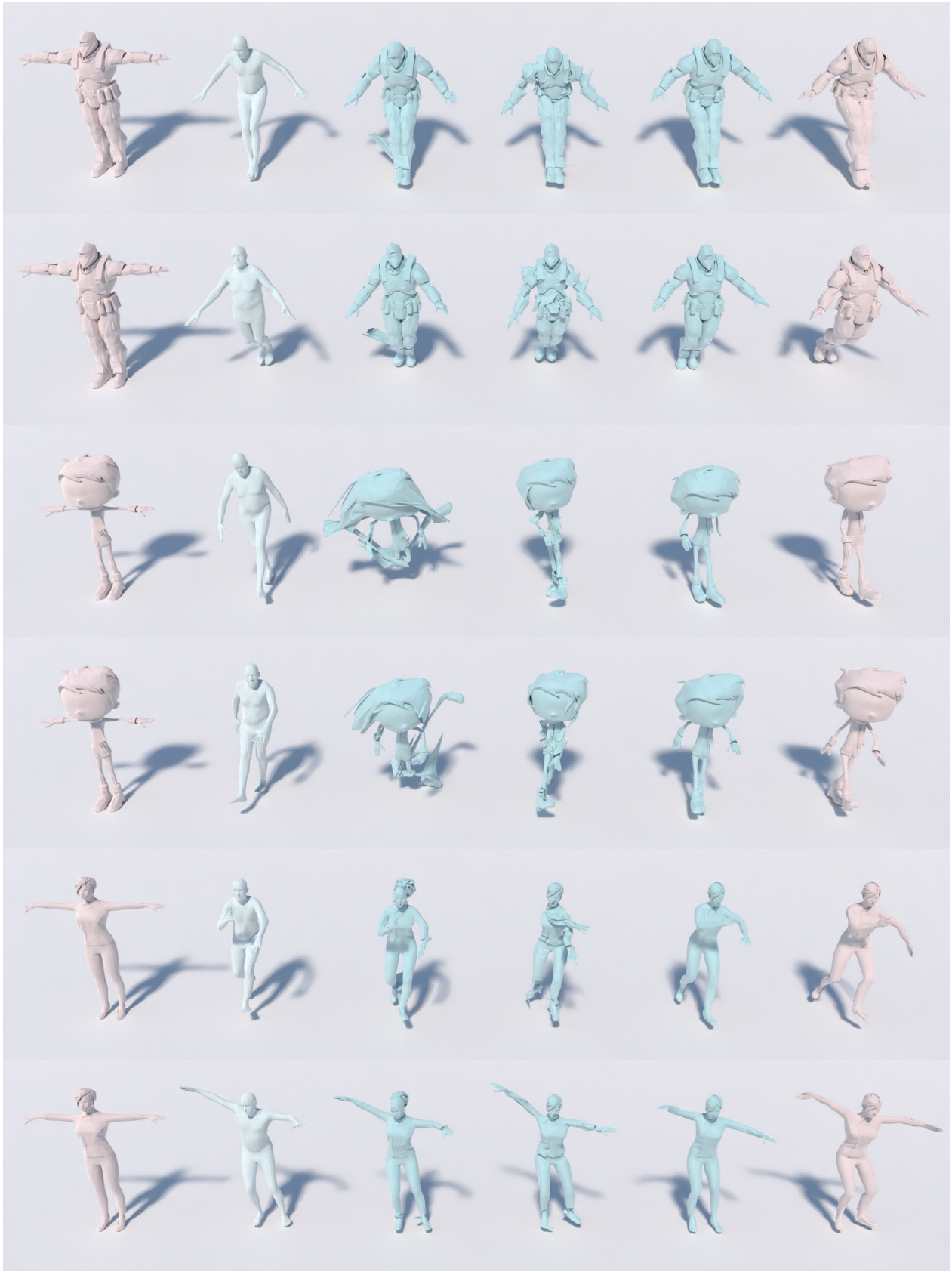

Figure 1. **Qualitative comparisons on Mixamo [1].**

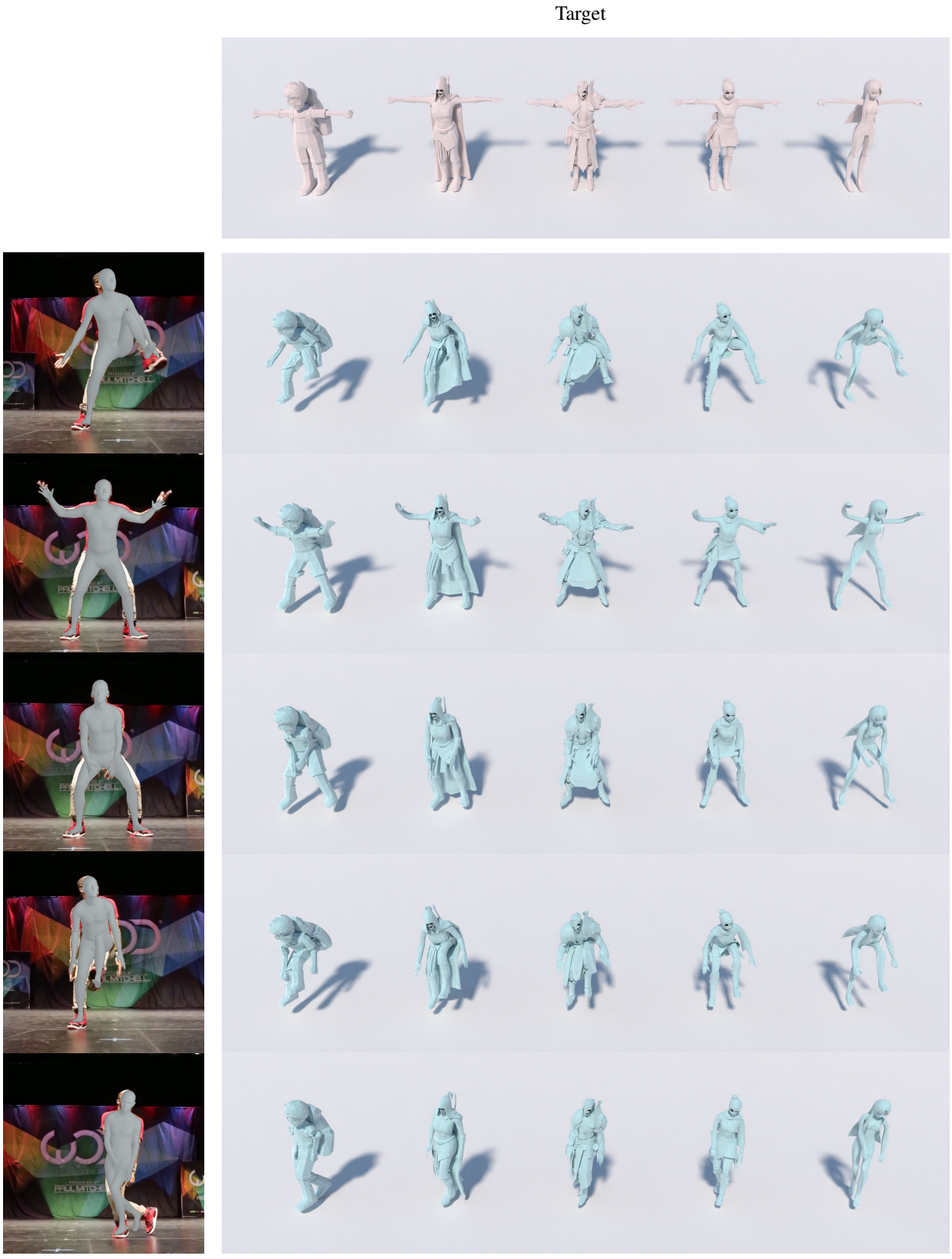

Figure 2. **Transferring poses from in-the-wild videos to stylized characters.**

Figure 3. **Transferring poses from in-the-wild videos to stylized characters.**

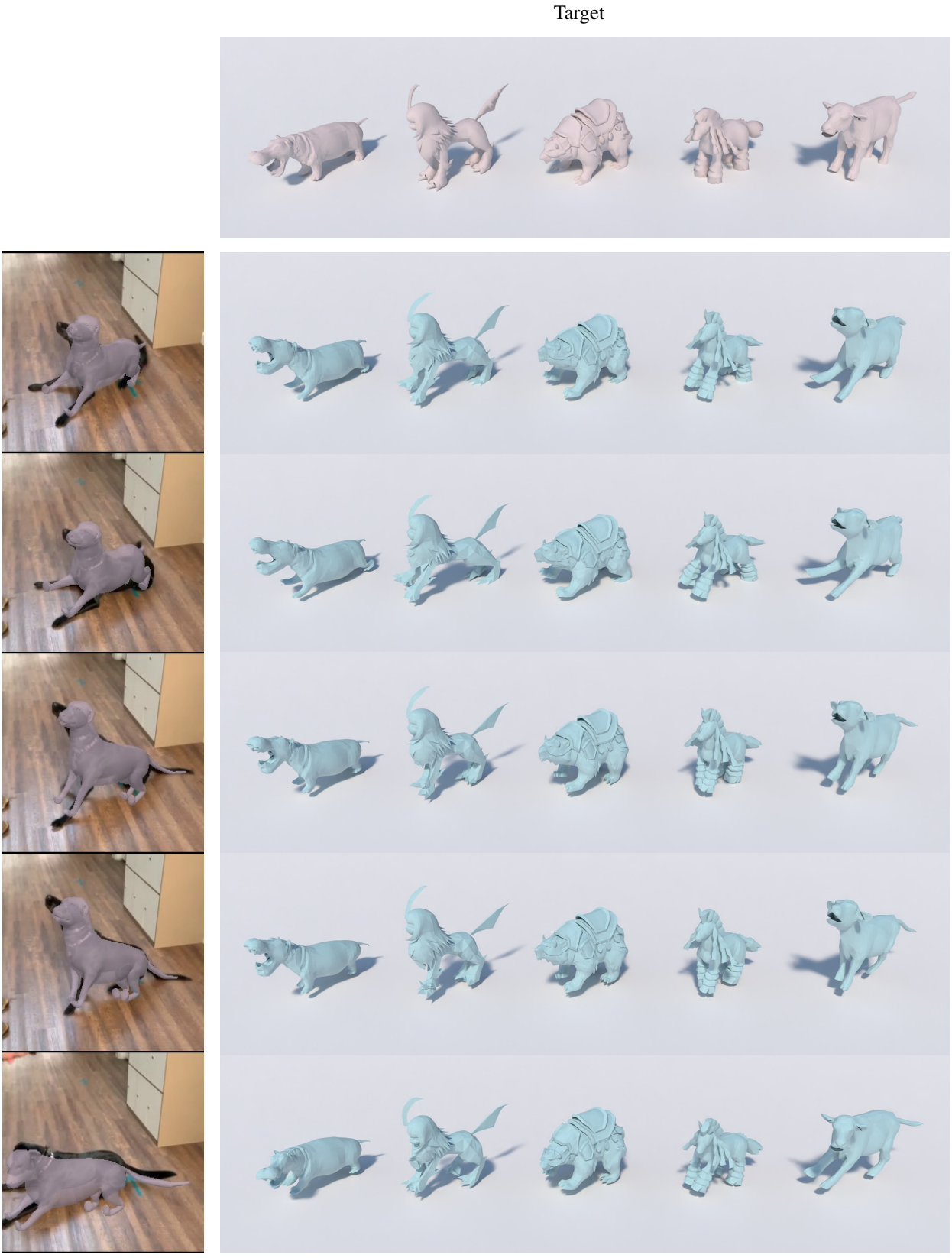

Figure 4. **Transferring animal poses from in-the-wild videos to stylized quadrupedal characters.**

References

[1] Mixamo. http://www.mixamo.com/. Accessed on November 09th, 2022. 1, 3

[2] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3D people from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[3] Peizhuo Li, Kfir Aberman, Rana Hanocka, Libin Liu, Olga Sorkine-Hornung, and Baoquan Chen. Learning skeletal articulations with neural blend shapes. In *ACM Transactions on Graphics (SIGGRAPH)*, 2021. 1, 2, 3

[4] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[5] Zhouyingcheng Liao, Jimei Yang, Jun Saito, Gerard Pons-Moll, and Yang Zhou. Skeleton-free pose transfer for stylized 3D characters. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3

[6] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[7] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3D human motion model for robust pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[8] Nadine Rüegg, Silvia Zuffi, Konrad Schindler, and Michael J Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[9] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[10] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3D Menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2