Supplementary Materials for Weakly-supervised 3D Pose Transfer with Keypoints

Jinnan Chen Chen Li Gim Hee Lee
Department of Computer Science, National University of Singapore
{jinnan.c, lichen}@u.nus.edu gimhee.lee@nus.edu.sg

We provide more details on the network architecture and more qualitative results for different datasets.

More details on the Network Architecture. The detailed network architecture is shown in Tab. 1. Our network contains 4 learnable parts: keypoints detector, twist predictor, skinning predictor, and refinement network. For the keypoints detector, we use a Pointnet [2] architecture with MLPs. The twist prediction network shares the same network architecture as the keypoints detector with the only difference in the output dimension. For the skinning weights predictor, we first calculate the Euclidean distance between each vertex and 24 detected keypoints as the point-wise feature (33 for animals). This point-wise feature is then concatenated with the original point 3D coordinates and fed into several 1D convolutional layers. The number of dimensions for the output is the number of bones, which equals to number of keypoints minus one. Finally, a softmax function is applied to make sure that the skinning weights for each vertex sum up to 1. For the refinement network, we first extract features from the source and the coarse mesh using several 1D convolutional layers. We then use the ElaINResnetBlock [4] as the basic block to gradually update the features. Finally, one 1D convolutional layer is applied to predict the deformation for each vertex. The deformation is directly added to the coarse mesh to obtain the final refined mesh.

The Definition of the Keypoints, Bones, and Kinematic Tree. For SMPL-based datasets (NPT and FAUST), we use the joint defined in SMPL [3] as our keypoints, as shown in Fig. 1. For the Mixamo [1] Dataset, we select 24 joints which are semantically similar to SMPL joints. For the SMAL Dataset, we use officially defined 33 joints as keypoints. The bone centers are defined as the mid-points of any two connected joints. The kinematics tree is defined based on SMPL [3] and SMAL [5].

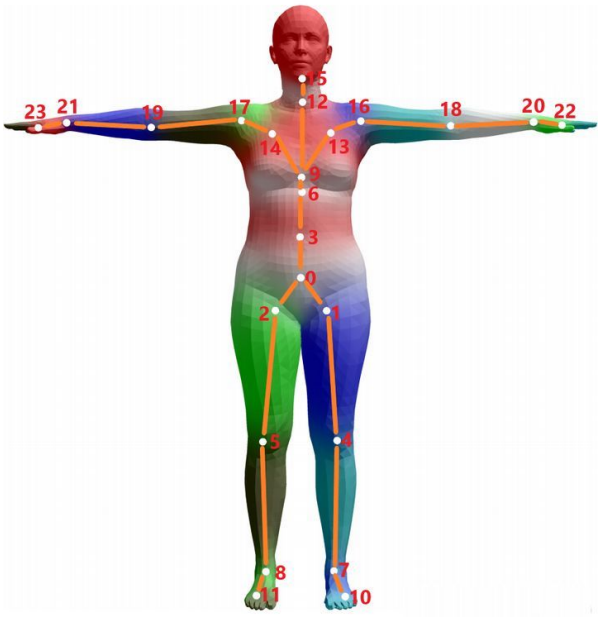

Figure 1. SMPL Joints definition, the figure is from SMPL [3].

More Qualitative Results We show more qualitative results on the NPT Dataset, FAUST Dataset, Mixamo Dataset and SMAL Dataset in Fig. 2, Fig. 3 Fig. 4 and Fig. 5 respectively. We can see that our model is able to transfer the pose successfully on both the template-based and non-template-based datasets.

References

[1] Adobe. Mixamo. https://www.mixamo.com, 2022.

[2] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. In *International Conference on Computer Graphics and Interactive Techniques*, 2015.

Model	Input	Module	Operation	Outputs shape	Outputs index
Keypoint detector	Source Point cloud	Feature extractor	Conv1D(3,64), Relu	B × 64 × N	(1)
	(1)		Conv1D(64,128), Relu	B × 128 × N	(2)
	(2)		Conv1D(128,256)	B × 256 × N	(3)
	(3)		Maxpooling	B × 256	(4)
	(4)	MLP deformer	Linear(256,256), LeakyRelu	B × 256 × N	(5)
	(5)		Linear(256,512), LeakyRelu	B × 256 × N	(6)
	(6)		Linear(512,256), LeakyRelu	B × 256 × N	(7)
	(7)		Linear(256,24 × 3), Reshape(B ×3× 24)	B ×3× 24	(8)
Twist predictor	Source Point cloud	Feature extractor	Conv1D(3,64), Relu	B × 64 × N	(9)
	(9)		Conv1D(64,128), Relu	B × 128 × N	(10)
	(10)		Conv1D(128,256)	B × 256 × N	(11)
	(11)		Maxpooling	B × 256	(12)
	(12)	MLP deformer	Linear(256,256), LeakyRelu	B × 256 × N	(13)
	(13)		Linear(256,512), LeakyRelu	B × 256 × N	(14)
	(14)		Linear(512,256), LeakyRelu	B × 256 × N	(15)
	(15)		Linear(256,23)	B × 23	(16)
Skinning predictor	Source Point cloud	Feature computing	Distance computing, Concatenate	B × 27× N	(17)
	(17)	Feature converter	Conv1D(27,64), Relu	B × 64 × N	(18)
	(18)		Conv1D(64,128), Relu	B × 128 × N	(19)
	(19)		Conv1D(128,1024)	B × 1024 × N	(20)
	(20)	Feature decoder	Conv1D(1024,512), LeakyRelu	B × 512×N	(21)
	(21)		Conv1D(512,23), LeakyRelu	B × 23×N	(22)
	(22)		Exponentiation, Softmax	B × 23×N	(23)
Refinement network	Source Point cloud	Source feature extractor	Conv1D(3,64), Relu	B × 64 × N	(24)
	(24)		Conv1D(64,128), Relu	B × 128 × N	(25)
	(25)		Conv1D(128,256)	B × 256 × N	(26)
	Coarse Point cloud	Coarse feature extractor	Conv1D(3,256,3,1,1)	B × 256 × N	(27)
	(27)		Conv1D(256,256)	B × 256 × N	(28)
	(28), (26))	ElaINResnetBlock (1)	Re-normalization	B × 256 × N	(29)
	(29)		Conv1D(256,128)	B × 128 × N	(30)
	(30), (26)	ElaINResnetBlock (2)	Re-normalization	B × 128 × N	(31)
	(31)		Conv1D(128,64)	B × 64 × N	(32)
	(32), (26)	ElaINResnetBlock (3)	Re-normalization	B × 64 × N	(33)
	(33)		Conv1D(64,3)	B × 3 × N	(34)

Table 1. The detailed network architecture: the numbers in Conv1D represent the input channels, output channels, kernel size, stride, and padding, respectively. The kernel size is set to 1, stride to 1, and padding to 0 if not specified. The number of keypoints and bones in the table are used for human-shape dataset, which can be changed to the number of those in the animal dataset.

[4] Chaoyue Song, Jiacheng Wei, Ruibo Li, Fayao Liu, and Guosheng Lin. 3d pose transfer with correspondence learning and mesh refinement. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[5] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Figure 2. More qualitative results on the NPT Dataset: from left to right are the source, target, our result, and ground truth.

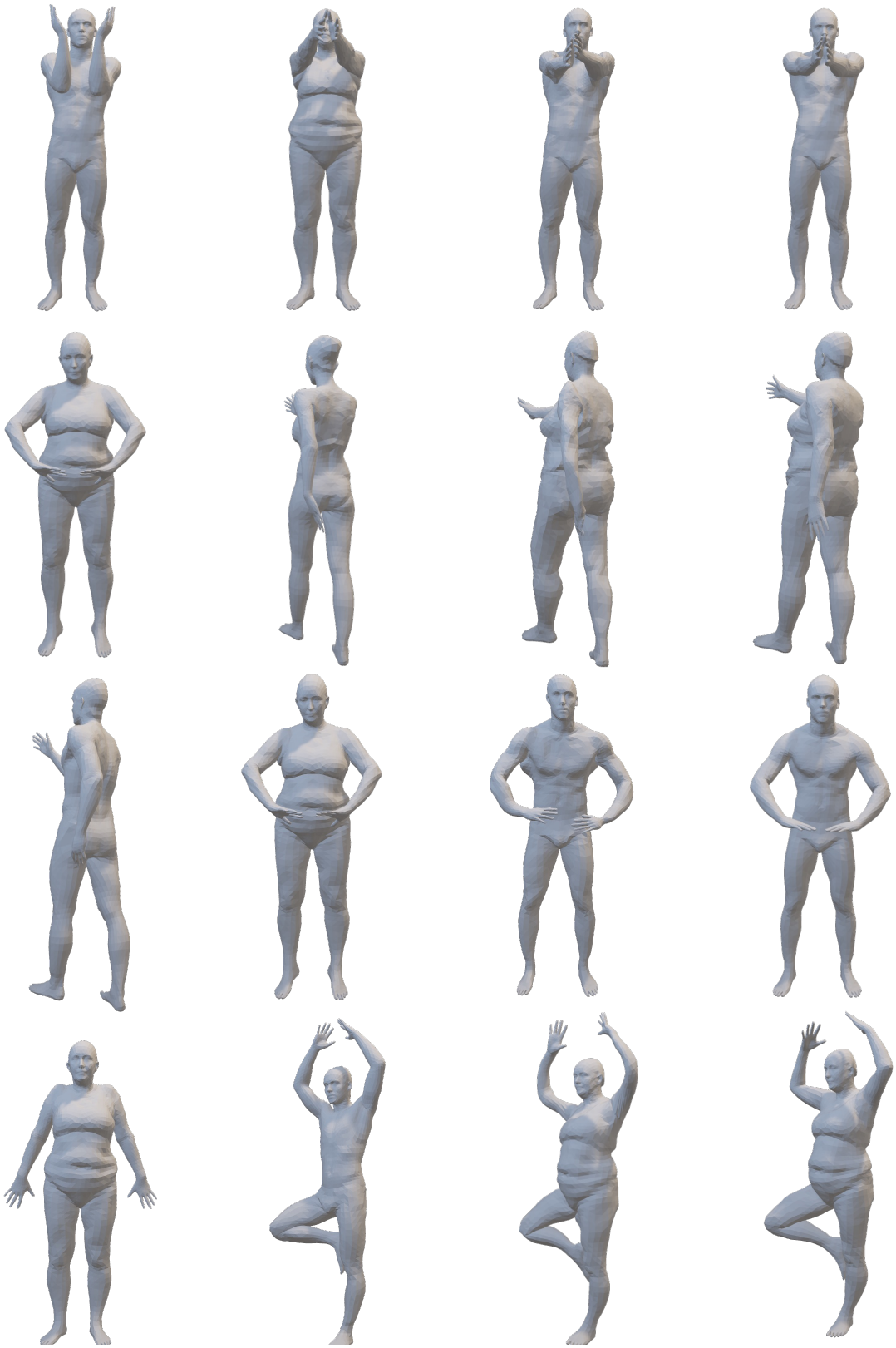

Figure 3. More qualitative results on the FAUST Dataset: from left to right are the source, target, our result, and ground truth.

Figure 4. More qualitative results on the Mixamo Dataset: from left to right are the source, target, our result, and ground truth.

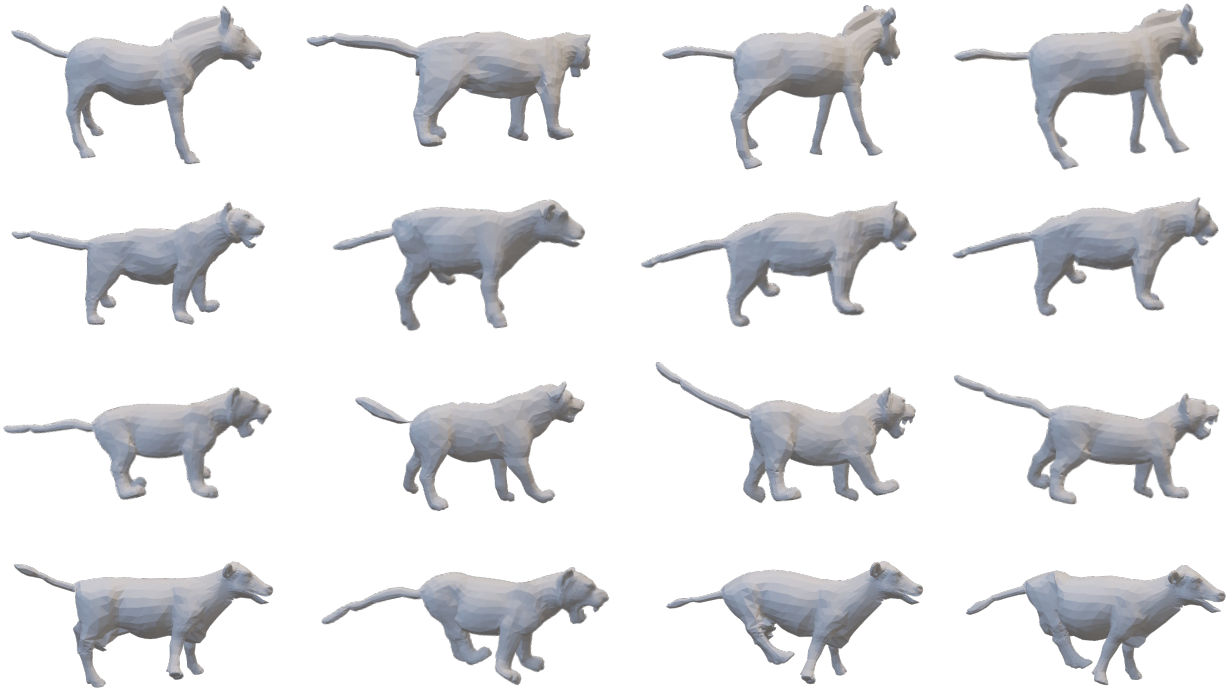

Figure 5. More qualitative results on the SMAL Dataset: from left to right are the source, target, our result, and ground truth.