Supplementary Material: Unified 3D Mesh Recovery of Humans and Animals by Learning Animal Exercise

Kim Youwang[1]
youwang.kim@postech.ac.kr
Kim Ji-Yeon[2]
jyeonkim@postech.ac.kr
Kyungdon Joo[4,5]
kyungdon@unist.ac.kr
Tae-Hyun Oh[1,2,3]
taehyun.oh@postech.ac.kr

[1] Dept. of Electrical Eng. POSTECH, Korea
[2] Dept. of Convergence IT Eng. POSTECH, Korea
[3] Grad. School of Artificial Intelligence POSTECH, Korea
[4] Artificial Intelligence Grad. School UNIST, Korea
[5] Dept. of Comp. Science and Eng. UNIST, Korea

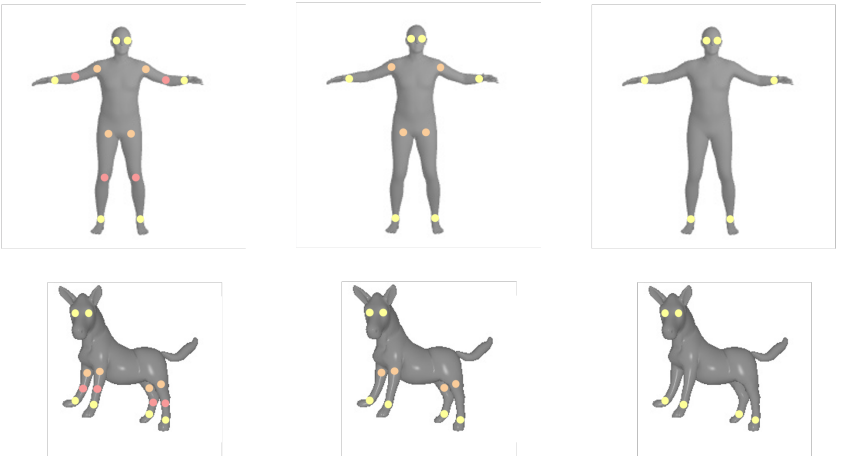
The goal of this supplementary material is to provide additional contents and details that are not included in the main paper due to space constraints. In Sec. 1, we provide additional analysis on our proposed *sub-keypoint*. In Sec. 2, we quantitatively evaluate *DeMR* with increased model capacity, showing promising further performance improvement. In Sec. 3, we qualitatively compare our method with the animal domain competing method, WLDO on dog. In Sec. 4, we describe our training details, including hyper-parameters and training hardware specifications. Lastly, in Sec. 5, we describe the datasets we used to train and evaluate *DeMR*, and evaluation metrics to measure the performance.

1 Analysis on the Effects of the Sub-keypoints

In this section, we explain the different settings of the number of sub-keypoints and discuss why we set the 10 sub-keypoints setting as our full model in training time. Moreover, we show the effectiveness of sub-keypoints with heterogeneous mesh prediction result.

1.1 Sub-keypoint Selection

The concept of sub-keypoints that we propose in the main manuscript is defined as a subset of full body keypoints using morphological similarity between humans and animals. Since most of the semantic correspondences are found on their specific body parts, *e.g.*, {left arm: left front leg}, we use *Eyes, Nose, Top-of-Limb (TL), Middle-of-Limb (ML)*, and *Bottom-of-Limb (BL)* as sub-keypoints for both humans and animals. The main concern is that the number of sub-keypoints can be interpreted as the amount of morphological similarity that



(a) (b) (c)

Figure 1: **Sub-keypoint settings**. (a) 15 sub-keypoints setting, including {*Eyes, Nose, TL, ML, BL*}. (b) 10 sub-keypoints setting (Ours), including {*Eyes, Nose, TL, BL*}. (c) 6 sub-keypoints setting, including {*Eyes, BL*}. Best viewed in color.

we consider in training time. Therefore, we set different settings of sub-keypoints (see Fig. 1) and analyze them.

For the 15 sub-keypoints setting, we choose the maximum number of corresponding keypoints, including *Eyes, Nose, TL, ML,* and *BL*. Even though we fully consider available sub-keypoints, fitting all of them rather disturbs training because humans and animals have different rotation angle range on their legs. Since animals can bend their legs in the opposite direction compared to humans so that poses taken by animals can be different with that of humans. Therefore, we exclude *Middle-of-Limb* for 10 sub-keypoints setting according to the priority order in morphological similarity. The setting without *ML* can more accurately represent morphological similarity between humans and animals and also show best performance on animal reconstruction. On the other hand, the setting for 6 sub-keypoints considers only *Eyes* and *BL* that is the minimal option of selecting keypoints. This minimum setting is not sufficient for the model to learn proper similarity. Based on these observations, we found 10 sub-keypoints setting is the most appropriate to embody morphological similarity among heterogeneous classes. In addition, developing the method to automatically select the proper sub-keypoint will be a promising future research direction.

1.2 Effectiveness of Sub-keypoint

In the main manuscript, we discuss the effectiveness of using sub-keypoint in training time. The concept of sub-keypoint not only models the morphological similarity between heterogeneous classes but also relaxes the disjoint multi-task learning problem into a jointly trainable multi-task setting. By using sub-keypoint, both human and animal keypoints are supervised by their counterparts, and the results of supervision are shown in Fig. 2. As an animal image is taken

Figure 2: **Heterogeneous mesh prediction in test time**. Similarly-posed meshes for both human and animal show the effect of sub-keypoint.

Models	# of Params.	Trained Object Classes	MPJPE [mm] ↓	PA-MPJPE [mm] ↓		PCK [%] ↑	
			MPI-INF-3DHP (Human)	Human3.6M-P2 (Human)	3DPW (Human)	Stanford Extra (Dog)	Animal Pose (Horse, Cow)
HMR [8]	27M	Human	169.50	66.50	81.30	-	-
WLDO [2]	95M	Dog	-	-	-	78.8	-
Ours (4cls, naïve)	27M+5M	Human, Dog, Horse, Cow	154.57	81.01	70.74	72.79	51.14
+ *Sub*			155.34	82.42	69.94	72.90	52.72
+ *Sub* + *CSBN (full)*			**140.76**	**79.70**	**69.85**	73.23	50.09
Ours (4cls, 3FC)	**27M+6M**	Human, Dog, Horse, Cow	-	-	-	**73.97**	**54.56**

Table 1: Quantitative evaluation results of *DeMR* on various human/animal datasets. The last row shows the reconstruction performance of our method, when trained with a larger model.

as the input of the model, a human mesh who is imitating the animal can be predicted by guidance of morphological similarity. We observe that the introduction of sub-keypoint has a great influence on training the model, although the sub-keypoint loss accounts small percentage of a total loss; thus, such weak supervision is enough for *DeMR* to learn morphological similarity. The demo video for the results of heterogeneous mesh prediction on BADJA dataset is provided as a supplementary video.

2 Additional Quantitative Results

In this section, we report our model's quantitative performance when trained with a larger model in terms of the capacity. In detail, we additionally increased the number of the parameters in the *deformable body multi-task branch* by adding one more FC layer. Since the SMAL mesh regression branch should cover three animal classes, increasing the branch's capacity would be helpful.

We trained the model with increased parameter size (27M+6M) for 300 epochs and evaluated PCK of both Stanford Extra dataset [2, 7] for dog and Animal Pose for horse and cow (see Table 1). *DeMR* showed improved performances for both datasets. Especially, the performance for Animal Pose increased about 3.5% compared to the main paper's full model. In this work, we focused on building a unified model with novel ideas such as sub-keypoint rather than achieving just a higher performance. However, although *DeMR* is effective and compact in terms of the parameter numbers, Table 1 tells one can achieve higher numerical performance when compromising with a larger capacity.

3 Additional Qualitative Results

In this section, we provide additional qualitative evaluation of our model. Mainly, we compare *DeMR*'s 3D reconstruction performance with the animal uni-modal competing method, WLDO [2]. For this qualitative comparison, we used BADJA dataset [1], which is a benchmark for animals that contains multiple animal classes in it. Then, we provide additional mesh prediction results for various human and animal datasets [2, 3, 9, 10] (see Fig. 5).

BADJA: rs_dog (seen class). First, we compare *DeMR*'s 3D reconstruction performance for dog images with WLDO. Although WLDO is an expert on dog mesh reconstruction and *DeMR* is made to cover multiple classes, *DeMR* shows qualitatively better mesh predictions for dog images (see Fig. 3-(Top)). WLDO shows bumpy mesh surfaces and not dog-like shaped mesh, despite its dog-specific shape regularizations. On the other hand, our method shows a more realistic and smooth mesh shape. In addition, the lower right leg of the predicted mesh from WLDO is in front of the lower left leg, which is inaccurate (see original

Figure 3: **Comparison with WLDO**. Each column shows the input image, WLDO's mesh prediction (blue), WLDO's mesh silhouette, *DeMR*'s mesh prediction (yellow), *DeMR*'s mesh silhouette for BADJA: rs_dog sequence (Top row), and for BADJA: impala0 sequence (Bottom row).

image). However, *DeMR* shows more accurate pose reconstruction than WLDO. Finally, *DeMR* shows a better mesh silhouette for the dog. Silhouette for lower legs or the neck is more accurate in our method than WLDO. This result shows the *DeMR*'s effectiveness in terms of the number of model parameters achieved by the introduction of sub-keypoint.

BADJA: impala (unseen class). In addition, we test both *DeMR* and WLDO whether they can deal with unseen classes during training time, *e.g.*, impala; thus, we test their generalizability. Since WLDO is not trained with other animal classes, it ends up predicting randomly-shaped mesh, although it tried to estimate approximate pose for the impala. However, thanks to the *class-selective shape prior loss* proposed in our method, although *DeMR* have not seen the class, impala, during training time, our method can reconstruct more impala-like shaped mesh and pose just from a single image. This result shows the powerful generalization of our method, while proving the effectiveness of the proposed class-specific shape prior loss.

4 Training Details

We provide training details, including hyper-parameters, training hardware specifications, and training time. *DeMR* requires four main loss functions in order to learn morphological similarity across heterogeneous classes. Recall that the total training objective is given as follows:

$$L = \alpha(\lambda_f^h L_{full}^h + \lambda_s L_{sub}^h) + (1-\alpha)(\lambda_f^a L_{full}^a + \lambda_s L_{sub}^a) + \lambda_{sil} L_{sil} + \lambda_{sh}^a L_{sh}^a, \quad (1)$$

The weights for the losses in Eq. (1) were set to λ_f^h=15, λ_f^a=5, λ_s=0.1, λ_{sil}=0.0075, λ_{sh}^a=10^{-5}, respectively. We additionally used pose prior losses for both humans and animals, shape prior loss for humans to further regularize the model to predict natural bodies. Those auxiliary losses were directly adopted from previous human and animal uni-modal models [2, 8]. Our model was trained using the AdamW optimizer with a batch size of 64, and the learning rate set to 5×10^{-5}. In these conditions, overall training required 200 epochs to converge and took about 40 hours with one NVIDIA TITAN RTX GPU.

5 Datasets and Evaluation Metrics

In this section, we explain datasets we used in test time. In addition, we introduce three different evaluation metrics to quantitatively measure our method's favorable performance on other competing methods.

5.1 Datasets

We use 2D and 3D keypoint annotated datasets in test time to evaluate our method's 3D reconstruction performance. The evaluation datasets we used are as follows:

HUMAN3.6M [5]. It is a large-scale video dataset with 3D human keypoint annotations captured in the indoor environment. We used 'Protocol 2' of the dataset following the previous human competing method [6].

MPI-INF-3DHP [9]. It is a dataset for 3D human body pose estimation captured mostly in indoor environments, with a small number of outdoor samples. We used the test splits for the evaluation.

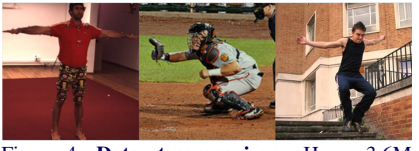


Figure 4: **Dataset comparison.** Human3.6M (left) shows distinct image characteristics, such as limited human poses in controlled indoor scenes, compared to the training datasets, COCO (middle) and LSP-extended (right), respectively.

3DPW [10]. It is a human video dataset captured in outdoor settings. 3D annotations are obtained using video and IMU sensors. We use the test split to evaluate our model.

StanfordExtra [2, 7]. It is a large-scale dataset for dogs with 2D keypoints and silhouette annotations. It covers 120 breeds of dogs with diverse shapes and poses. Along with keypoint location, binary segmentation masks are used to compute the silhouette loss and give a threshold to the PCK metric.

Animal Pose Dataset [3]. It is a 2D keypoint annotated animal dataset covering 5 kinds of animals: dog, cat, cow, horse, and sheep. We selected 2 classes, *i.e.*, horse and cow, to train and evaluate the model.

5.2 Evaluation Metrics

For the evaluation, three different metrics, *i.e.*, *PCK*, *MPJPE*, *PA-MPJPE* are used. The unit of PCK metric is in percentage (%) while the others are in millimeter (mm).

PCK. Percentage of Correct Keypoint (PCK) is a metric that evaluates the quality of keypoint prediction. It is the ratio of the predicted keypoints that are within the threshold distance from the ground truth keypoints. The threshold is normalized by silhouette area; details can be found in [2]. Since animal datasets do not contain 3D keypoint annotations, we measured PCK by re-projecting predicted 3D keypoints of animals into 2D.

MPJPE. Mean per joint position error (MPJPE) is a metric that measures the 3D keypoint reconstruction error. It computes the mean of the Euclidean distances between the predicted keypoints and the ground truth keypoints. We use MPJPE to evaluate our model for the human datasets with 3D ground truth annotations.

PA-MPJPE. PA-MPJPE also measures the 3D keypoint reconstruction error, but after the rigid alignments of reconstructed keypoints to the ground truth keypoints using Procrustes Analysis [4]. Thus, PA-MPJPE is effective when computing the 3D keypoint error regardless of global misalignments. We used PA-MPJPE to evaluate our model on the 3D human datasets.

Figure 5: **Additional Qualitative Results**.The first two columns show the input images of humans and SMPL mesh prediction results. The last two columns show the input images of animals and SMAL mesh prediction results.

References

[1] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and SMAL: Recovering the shape and motion of animals from video. In *Asia Conference on Computer Vision (ACCV)*, 2018.

[2] Benjamin Biggs, Ollie Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In *European Conference on Computer Vision (ECCV)*, 2020.

[3] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Yu-Wing Tai, and Cewu Lu. Cross-domain adaptation for animal pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[4] J. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[5] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36 (7):1325–1339, 2014.

[6] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization : Stanford dogs. In *First Workshop on Fine-Grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[8] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[9] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*, 2017.

[10] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018.