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@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

Abstract

We propose an end-to-end unified 3D mesh recovery of humans and quadruped animals trained in a weakly-supervised way. Unlike recent work focusing on a single target class only, we aim to recover 3D mesh of broader classes with a single multi-task model. However, there exists no dataset that can directly enable multi-task learning due to the absence of both human and animal annotations for a single object, *e.g.*, a human image does not have animal pose annotations; thus, we have to devise a new way to exploit heterogeneous datasets. To make the unstable disjoint multi-task learning jointly trainable, we propose to exploit the morphological similarity between humans and animals, motivated by *animal exercise* where humans imitate animal poses. We realize the morphological similarity by semantic correspondences, called sub-keypoint, which enables joint training of human and animal mesh regression branches. Besides, we propose class-sensitive regularization methods to avoid a mean-shape bias and to improve the distinctiveness across multi-classes. Our method performs favorably against recent uni-modal models on various human and animal datasets while being far more compact.

1 Introduction

Considering the fact that countless species of animals are present in nature, it is challenging to develop a machine that understands the behavior of general animals. To resolve this challenge, the humans' capability of imitating animals, *i.e.*, *animal exercise*, can be a crucial clue to build such a machine. Concretely, humans can express and mimic animals by fully utilizing their body parts similar to those of animals during animal exercise, which is capable thanks to the morphological similarity between humans and animals. Inspired by this morphological similarity, we tackle a unified multi-task model that estimates the 3D volumetric meshes of deformable objects from a single image (see Fig. 1), including humans and a few quadruped animals, called *Deformable Objects Mesh Recovery*, *DeMR*.

A major challenge to develop such a unified model is the requirement of heterogeneous datasets of humans and animals that have incompatible label formats, *i.e.*, disjoint label set setup [13, 18]. In other words, human images do not have animal keypoint annotations or *vice versa*. It is known that directly deploying multi-task learning with disjoint label sets by a naïve approach yields performance degradation compared to uni-modal models [13, 18]. We tackle this challenge by leveraging a few sub-keypoints, which is a subset of full body keypoints morphologically shared across heterogeneous classes, *e.g.*, {left wrist: left front paw}. We enforce the model to at least fit morphologically corresponding keypoints, such that it can implicitly learn such similarity among different classes. This enables stable multi-task learning by jointly computing the losses of each multi-task branch even without compatible annotations across all the classes. Lastly, despite this mutually beneficial development, since our model mainly focuses on learning shared knowledge, it could be biased to predict mean-like shapes for different species. We mitigate it by introducing class-sensitive regularizations: class-selective shape prior loss and class-specific batch normalization layers.

Figure 1: *DeMR* is a unified model that reconstructs meshes of humans and animals.[1]

On top of these developments, our proposed model can reconstruct realistic and distinctive 3D shapes of different quadruped animals as well as those of humans. As a favorable by-product, our training scheme can mitigate data scarcity issues that exist in the animal keypoint annotated datasets. Besides, our unified model benefits shared knowledge of multiple heterogeneous classes, whereby the model can preserve comparable or better performance than each of the uni-modal models without increasing the model size. We show our model's 3D reconstruction performance with competing methods in human and animal domains [3, 10]. We summarize our main contributions as follows:

- We present *DeMR*, a unified neural regression model that can directly reconstruct 3D pose and shape of various classes (humans and quadruped animals) from a single image.
- We suggest a data-efficient way to deal with cumbersome training with human and animal datasets with incompatible labels. We tackle it by introducing joint multi-task learning using sub-keypoint that leverages morphological similarity among heterogeneous classes.
- We propose novel methods to tackle the challenges due to the domain gaps among heterogeneous classes: class-selective shape prior loss and class-specific normalization layer.

2 Related Work

Our task is related to the multi-class 3D full-body shape and pose estimation. The 3D volumetric recovery has been mainly studied using parametric body models, such as SMPL [22] and SMAL [31], called model-based approaches. We briefly review these lines of researches.

Monocular 3D Human Reconstruction. Recent advances in body models and model-based 3D human reconstruction have enabled accurate 3D human mesh reconstruction from a single image. SMPL provides a compact representation of volumetric 3D body mesh with interpretable parameterization of shape, pose, and camera parameters. SMPLify [4] introduced the first fully automatic approach to fit the SMPL model onto an image via an optimization pipeline. Later, HMR [10] introduced a learning-based SMPL parameter regressor

[1] We performed *DeMR* for each cropped image and then composited them for visualization.

using Convolutional Neural Network (CNN). HMR directly regresses mesh parameters using extracted body features from CNN. Compared to the optimization-based method, the CNN-based method enables efficient estimation and opens versatile applications. Recently, numerous other approaches, *e.g.*, optimization in-the-loop, self attention-based vertex regression [15, 19], have achieved significant advances by focusing only on humans. Our *DeMR* leverages HMR as a backbone to obtain shared body features across heterogeneous classes, extending beyond the human class.

Monocular 3D Animal Reconstruction. Analogous to the human parametric model, the 3D animal body model, *i.e.*, SMAL [31], has boosted the development of model-based 3D animal reconstruction. Optimization-based approaches [2, 32] and learning-based methods [3, 33] were suggested to infer 3D animal body meshes onto a single image. However, despite these advances, 3D animal reconstruction has been less spotlighted than that of humans. The reason stems from insufficient animal shape annotations and limited animal class coverage for 3D motion capture datasets. Such limitations have narrowed down the research scope of the previous works to only cover one or few animal classes. We leverage the morphological similarity across different animal and human classes. This enables a unified model to be well-supervised in data-scarce regimes of a few classes, which deals with the limitation of the prior works. In contrast to recent works that focus on 3D mesh reconstruction of either only a single object class at a time or a similar class of animals, *i.e.*, uni-modal model, our method is the first learning-based method that can directly estimate 3D meshes of multiple heterogeneous classes, including humans and quadruped animals, and do not require any test time annotations.

Pose Estimation of Multiple Animal Classes. Aside from 3D reconstruction tasks of a single object class, there have been a few attempts to build a unified model of multiple object classes but with 2D or 2.5D pose representations. Cao *et al*. [5] showed physical similarity among humans and animals can be learned by estimating 2D keypoints with cross-domain adaptation. With 2.5D representation, recent work [29] showed that physically analogous 2.5D pose representation of humans can be extended to primates, *e.g.*, gorilla and chimpanzee. Recently, functional maps [25] opened up an opportunity to extend the 2.5D pose representation of humans to broader animal classes [24]. Although these works have limited scopes of pose representations, they have proved the existence of morphological similarity among heterogeneous classes. We leverage this insight to build our unified model but tackle a more challenging regime, *i.e.*, 3D pose and shape with a unified model.

3 *DeMR*: Deformable Objects Mesh Recovery

In this section, we first define the term, *deformable objects*, and review body representations for deformable objects in Sec. 3.1. Next, we present the overall structure of our method in Sec. 3.2. Then, we propose novel methods to enable learning *morphological similarity* and class-specific regularizations in Sec. 3.3.

3.1 Preliminary

We consider four deformable object classes that frequently appear in the wild nature with distinct body poses and shapes, and the classes that have a reasonable amount of data available for training and evaluation. Accordingly, we consider *Human, Dog, Horse* and *Cow* as our target deformable object classes.

Body Representation of Deformable Objects. We adopt parametric 3D body models to model 3D bodies of deformable objects that we consider. SMPL [22] represents a human

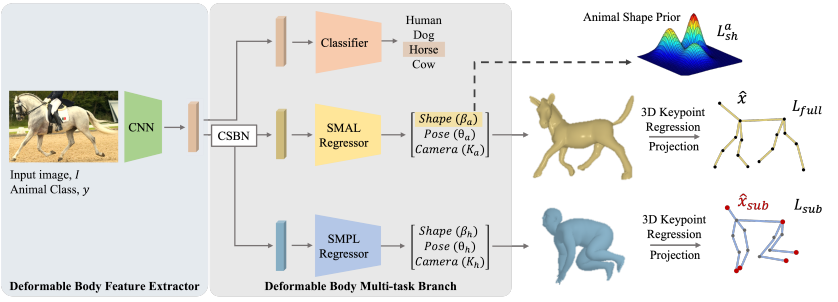


Figure 2: ***DeMR*'s Architecture**. During training, given an input image, the model regresses SMPL and SMAL mesh parameters, and class probability in a multi-task manner. At test time, *DeMR* can (i) *regress* both SMPL and SMAL parameters regardless of input object class, and (ii) *select* which parameter it should use to construct appropriate mesh, according to the classification result.

body with the 3D mesh vertices $M_h \in \mathbb{R}^{6890\times3}$ by a differentiable function $\mathcal{M}_h(\theta_h, \beta_h)$ that takes the human pose parameters $\theta_h \in \mathbb{R}^{24\times3}$ in angle-axis representation and the human shape parameters $\beta_h \in \mathbb{R}^{10}$ as input. Correspondingly, as the quadruped animal model, we use SMAL [31] that represents the 3D mesh vertices $M_a \in \mathbb{R}^{3889\times3}$ by $\mathcal{M}_a(\theta_a, \beta_a)$ that takes animal pose parameters $\theta_a \in \mathbb{R}^{35\times3}$ and animal shape parameters $\beta_a \in \mathbb{R}^{20}$ as input.

3.2 *DeMR*'s Architecture and Loss Functions

The overall architecture of *DeMR* mainly consists of two parts: *Deformable body feature extractor* and *Deformable body multi-task branch* (refer to Fig. 2).

Deformable Body Feature Extractor. It is a shared CNN that takes a single center-cropped deformable object image as an input and extracts a single feature vector $\mathbf{f} \in \mathbb{R}^{2048}$. We use the same backbone network with the previous SMPL regression works [10, 15]. We guide this shared network to learn shared body morphology across heterogeneous classes.

Deformable Body Multi-task Branch. The multi-task branches take the extracted feature vector as input and perform parameter estimation for 3D body reconstruction and classification in a multi-task manner. Each task is conducted via respective fully connected layers. The extracted feature vector encodes appearance information so that the model can perform object classification. This classification allows the model to select which mesh parameters (or branch) to be used at test time. The other multi-task branches perform class-specific mesh parameter estimation from a shared feature. Note that, since multi-class datasets are mixed in a batch during training, a statistical gap between humans and animals exists. Thus, we propose class-sensitive regularizations to mitigate the gap and facilitate the multi-task branches to learn the class-specific residual information, described later.

Loss Function. Our model is trained with 2D keypoint annotations mainly, and silhouette masks are used if available, *i.e.*, no 3D annotation is used. Our total loss L is defined as:

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

where L^*_{full} are full-keypoint reconstruction loss terms, L^*_{sub} are sub-keypoint reconstruction loss terms, L_{sil} is a silhouette reconstruction loss term, and L^a_{sh} is a class-selective shape prior loss. Besides, $\{\lambda_*\}$ denotes the balance parameters for each loss term, and α a human/animal indicator; *i.e.*, $\alpha = 1$ for human data, $\alpha = 0$ otherwise. Details about hyper-parameters can be found in the supplementary material.

Following recent works [9, 10, 15], we use the full 2D keypoint reconstruction loss as the primary training loss. Once we reconstruct SMPL and SMAL meshes from the estimated

mesh parameters, θ_h, β_h, θ_a and β_a and the respective camera parameters, K_h and K_a, we can regress 3D keypoints $\mathbf{J}_h$ and $\mathbf{J}_a$ by the linear combination of each subset of the mesh vertices. By orthographic projection of the estimated 3D keypoints, as $\hat{\mathbf{x}}_h = \Pi(\mathbf{J}_h; K_h)$ and $\hat{\mathbf{x}}_a = \Pi(\mathbf{J}_a; K_a)$, where $\Pi(\cdot;\cdot)$ is an orthographic projection, we can re-project 3D keypoints to 2D keypoints.

Then, given the ground truth 2D keypoints, $\mathbf{x}_{\{h,a\}}$, the full-keypoint reconstruction loss can be computed as the Euclidean distance between $\mathbf{x}_{\{h,a\}}$ and $\hat{\mathbf{x}}_{\{h,a\}}$:

$$L^h_{full} = \frac{1}{N_h}\sum_i v_{h_i}\|\mathbf{x}_{h_i} - \hat{\mathbf{x}}_{h_i}\|_2^2, \quad L^a_{full} = \frac{1}{N_a}\sum_i v_{a_i}\|\mathbf{x}_{a_i} - \hat{\mathbf{x}}_{a_i}\|_2^2, \tag{2}$$

where $N_{\{h,a\}}$ are the respective numbers of visible keypoints of humans and animals, $\mathbf{x}_{\{h_i,a_i\}}$ are the i-th human and animal ground truth 2D keypoints, respectively, and $v_{\{h_i,a_i\}}$ are the visibility for each ground truth keypoint (1 if visible, 0 otherwise). Besides, we use the silhouette loss for animal classes, since animals show distinctive body shape difference across the classes. Given the animal binary mask annotation S_a and the estimated binary mask $\hat{S}_a$, the silhouette loss can be computed by the negative IoU [12, 21] as $L_{sil} = 1 - \frac{\|\hat{S}_a \otimes S_a\|_1}{\|\hat{S}_a \oplus S_a - \hat{S}_a \otimes S_a\|_1}$, where $\oplus$ and $\otimes$ denote the respective element-wise sum and product.

Given an image of a class, *e.g.*, human, the described modules and losses are only partially available due to the absence of annotations of another branch, *e.g.*, animals; thus, the other branch is not jointly trainable. Hence, we introduce our novel components of *DeMR* that relax original disjoint multi-task problem into a jointly trainable scheme.

3.3 Morphological Similarity based Multi-task Learning

Our *DeMR* becomes jointly trainable under the following hypothesis: the *morphological similarity* across different classes of deformable objects provides additional knowledge and it can mitigate the difficult optimization in training. We first define the morphological similarity such that the effectiveness of hypothesis is realized, and then present the techniques that can induce more distinctive outputs specific for each class branch.

Morphological Similarity. Our observation is that physical corresponding body parts exist among different species, such as arms and legs. We use the keypoint of *Eyes, Nose, Top-of-Limb (TL), Middle-of-Limb (ML), Bottom-of-Limb (BL)* on the arms, the legs, and the head, called sub-keypoints (see Fig. 3). We force the model to reconstruct an object class with a body model of a different class, *i.e.*, mimicking a pose like the animal exercise. We induce it by minimizing the difference between the supervision of sub-keypoints of a class and the estimated sub-keypoint of the other class. Thereby, the model learns morphological similarity among heterogeneous classes. We empirically select N_{sub} keypoints from the ground truth full 2D keypoints and define them as sub-keypoints, *i.e.*, $\mathbf{x}^{sub}_{\{h,a\}} \subset \mathbf{x}_{\{h,a\}}$ (refer to the supplementary materials for details).

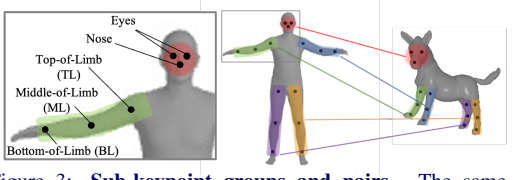


Figure 3: **Sub-keypoint groups and pairs.** The same colored regions correspond to morphologically similar ones across heterogeneous classes.

Sub-keypoint based Multi-task Learning. Introducing the sub-keypoint concept allows joint learning of multi-task branches in the incompatible labeled set setup. Specifically, given the example with a horse in an input image in Fig. 2, the model tries to estimate both SMPL and SMAL parameters in each branch. However, since human keypoint annotations do not exist for a horse image, the model cannot compute the full keypoint loss L^h_{full}, and discards the

predictions from the human branch. This wastes that branch, and leads to the forgetting phenomenon during training [18], resulting in unstable multi-task learning. Our sub-keypoint reconstruction loss for inducing morphological similarity can resolve this limitation of the naïve multi-task learning. Since the loss enforces the estimated sub-keypoints of the human SMPL mesh to follow the morphologically corresponding sub-keypoint ground truth of a horse; thus, even without human labels for the horse image, we can compute the loss for the human mesh branch as well and get gradients during the backpropagation, enabling stable joint training of both branches. This is *vice versa* for a human image.

The computation of the sub-keypoint reconstruction loss is identical to that of the full-keypoint loss, except it uses the N_{sub} morphologically corresponding sub-keypoints (we set N_{sub}=10). Given the ground truth sub-keypoints $\mathbf{x}^{sub}_{\{h,a\}} \subset \mathbf{x}_{\{h,a\}}$, estimated sub-keypoints $\hat{\mathbf{x}}^{sub}_{\{h,a\}} \subset \hat{\mathbf{x}}_{\{h,a\}}$ and the visibility $v^{sub}_{\{h,a\}}$, the sub-keypoint loss can be computed as follows:

$$L^{h}_{sub}=\frac{1}{N_{sub}}\sum_{i=0}^{N_{sub}}\|v^{sub}_{h_i}(\mathbf{x}^{sub}_{h_i}-\hat{\mathbf{x}}^{sub}_{a_i})\|_2^2, \quad L^{a}_{sub}=\frac{1}{N_{sub}}\sum_{i=0}^{N_{sub}}\|v^{sub}_{a_i}(\mathbf{x}^{sub}_{a_i}-\hat{\mathbf{x}}^{sub}_{h_i})\|_2^2. \tag{3}$$

The effect of sub-keypoint loss is shown in mesh estimation results in Fig. 2. Since the model is supervised with the ground truth sub-keypoint of a horse, it eventually estimates a human SMPL mesh that mimics the pose and shape of the horse.

Class-selective Shape Prior Loss. As shown in the prior works [3, 15], regularizing the estimated shape parameters to follow the natural distribution of each class is essential in achieving realistic bodies of deformable objects. Since different animal classes show distinctive shape parameter distributions, we extend the uni-modal shape prior to the *class-selective shape prior loss* to adaptively regularize shape parameters according to the target animal class. Given a multivariate Gaussian shape prior distribution of the animal class, c, with a mean $\mu^{c}_{\beta_a}$ and a covariance $\Sigma^{c}_{\beta_a}$, the shape prior loss can be computed by the Mahalanobis distance for the predicted animal shape parameters, β_a, as:

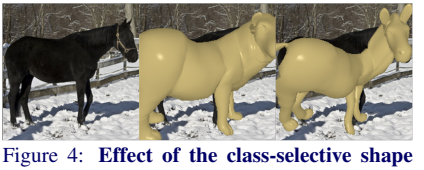


Figure 4: **Effect of the class-selective shape prior.** The models trained with uni-modal (middle) and class-selective shape prior (right).

$$L^{a}_{sh}=\sum_{c=0}^{N_{animals}}\mathbb{1}[y=c]\cdot(\beta_a-\mu^{c}_{\beta_a})^{\top}(\Sigma^{c}_{\beta_a})^{-1}(\beta_a-\mu^{c}_{\beta_a}), \tag{4}$$

where y denotes the ground truth animal class label, $\mathbb{1}[\cdot]$ the indicator function that returns 1 for true and 0 otherwise. Since we aim to estimate realistic 3D mesh of three different animal classes which have class-wise different shape prior means and covariance matrices, the proposed class-selective shape prior loss L^{a}_{sh} can reflect class-specific statistics of shape given the ground truth animal class label y in training time. The effect of class-selective shape prior loss is shown in Fig. 4, where the model trained with the uni-modal prior estimates mean-shaped animal mesh, while the model with our prior estimates a more horse-like shape.

Class-specific Batch Normalization (CSBN). As mentioned, there are statistical gaps between humans and animals in shared feature vectors obtained from the feature extractor. This is detrimental for multi-task branches because the mismatch is likely to be propagated to the mesh parameter output layer. To take into account the statistical gap, we adopt two different batch normalization layers in the front of each branch, denoted CSBN. CSBN enables the model to learn class-specific residual information and statistical gaps among heterogeneous class data while allowing the model to share all other model parameters, *i.e.*, CNN backbone. There is a similar work called domain-specific batch normalization [6], but different from ours in that it switches the statistics at a single path according to an input domain label but ours is deployed for adapting statistics for independent branching.

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 [10]	27M	Human	169.50	66.50	81.30	-	-
WLDO [8]	95M	Dog	-	-	-	78.80	41.05
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 (2cls, naïve)	27M+5M	Human, Dog	160.14	81.13	**68.75**	73.87	-
+ *Sub* + *CSBN*			**156.69**	**80.28**	70.64	**74.29**	-

Table 1: Quantitative evaluation on various human/animal datasets.

	PA-MPJPE [mm] ↓
HMR [10]	81.3
Kanazawa *et al.* [11]	72.6
Kolotouros *et al.* [16]	70.2
Ours	**69.85**

(a)

	PA-MPJPE [mm] ↓
Lassner *et al.* [21]	93.9
Pavlakos *et al.* [23]	75.9
HMR [10]	66.5
Ours	79.7

(b)

Table 2: (a) Evaluation on 3DPW dataset. (b) Evaluation on Human3.6M dataset P2.

4 Experiments

In this section, we evaluate our proposed *DeMR* in several aspects. Specifically, we first introduce datasets and evaluation metrics for training and evaluation in Sec. 4.1. We then evaluate our method by comparing it with the previous uni-modal methods in Sec. 4.2. Lastly, we further analyze the effects of the proposed key components in Sec. 4.3.

4.1 Experiment Setup

Unlike human datasets that have diverse annotations, including 2D, 3D keypoints, or segmentation mask annotations, supervised datasets themselves are rare for animals; moreover, real 3D keypoint annotations are extremely difficult to obtain and far rarer for animals. Considering practical usages, we focus on the weakly-supervised case with 2D keypoint datasets for both humans and animals. Concretely, we follow the evaluation setup same with [8, 15]. We use LSP [9], LSP-extended [9], MS COCO [20], and MPII [1] for training the human domain, and use Stanford Extra [8] and Animal Pose [5] for training the animal domain. For evaluation, we employ Human3.6M [7], MPI-INF-3DHP [23], 3DPW [30] as the human domain evaluation sets, and Stanford Extra and Animal Pose for the animal domain evaluation sets. We report the Procrustes-aligned mean per joint position error (PA-MPJPE), the mean per joint position error (MPJPE) and the Percentage of Correct Keypoints (PCK).

Note that even though the model is trained with 2D keypoint datasets in both human and animal domains, we evaluate *DeMR*'s 3D recovery performance in the human domain to report our method's favorable performance with the competing methods.

4.2 Evaluation

Competing Methods. Since our model is the first approach that addresses the 3D mesh reconstruction of both humans and animals using a unified model, no prior works are directly comparable. We first compare with the base neural models that we mainly refer to and are the closest comparators, *i.e.*, HMR [10] for humans and WLDO [8] for dog breeds. Then, we compare our unified method with more recent uni-modal models for humans.

Quantitative Evaluation. Tables 1 and 2 provide detailed comparisons to other methods on several 3D human datasets. While our base structure is the same with HMR [10], *DeMR* show improved performances on MPI-INF-3DHP and 3DPW, but has lower performance

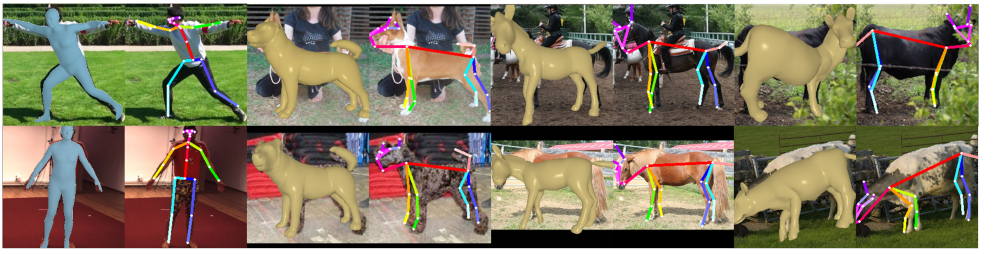


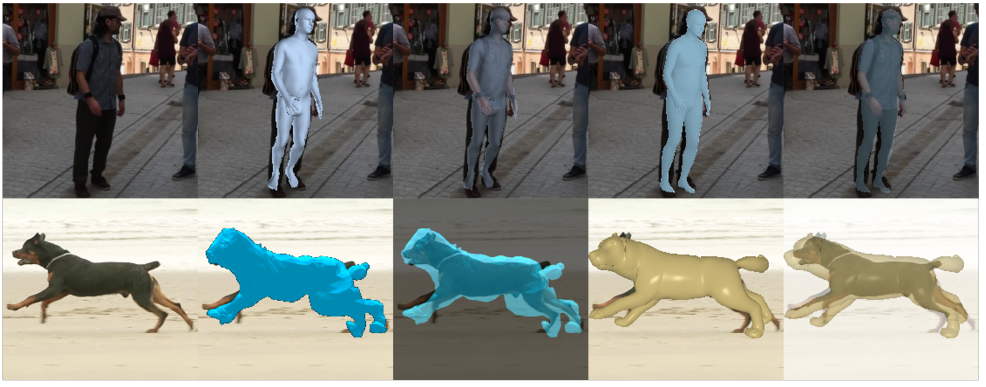
Figure 5: **Qualitative results on multiple datasets.** Each column shows the result of predicted mesh and predicted keypoints alternatively for human, dog, horse, and cow classes. The images are from 3DPW, Human3.6M (human), Stanford Extra (dog) and Animal Pose dataset (horse, and cow).

on Human3.6M. We postulate that a huge domain gap between our training datasets of humans and animals and Human3.6M led to performance degradation. Remind that *DeMR* is trained with 2D outdoor human and animal datasets [1, 8, 9, 20]. The MPI-INF-3DHP and 3DPW datasets, where our method performs better, mainly consist of outdoor human motion sequences similar to our training datasets. On the other hand, Human3.6M is captured in controlled indoor scenes, which shows totally different characteristics, including backgrounds, human poses, and light conditions. Particularly, *DeMR* leverages animal exercise, whereby our method may be more favorable for unusual poses rather than indoor poses.

We further compare with other competing methods in Table 2, where our model favorably performs against the uni-modal approaches specifically designed for humans. Note that our model performs better than Kanazawa *et al*. [11], which is a video-based multi-frame approach to deal with temporally consistent human motions, *i.e*., unfair to ours. The result supports our hypothesis that the morphological similarity among humans and animals gains to obtain complementary information and powerful reconstruction in a single unified model.

Regarding the animal classes, WLDO is the strong competing method because it is designed and tuned for dog breeds and is evaluated on Stanford Extra that is a dog-only dataset. In Table 1, our *DeMR* shows slightly lower performance than that of WLDO, but this is still favorable performance in that, from HMR, our method adds up a new capability to deal with more diverse animals with a small increase of the model size over that of HMR. *DeMR* is three times smaller than the model size of WLDO, and can deal with three additional animals, *i.e*., humans, horses, and cows. This result implies that the morphological similarity among different classes helps our model cover more classes while preserving the uni-modal model's performance with much compact networks.

For Animal Pose, we use the horse and cow categories as the test set. Although WLDO also evaluates the model on Animal Pose, they only use the dog category as a test set since WLDO only covers dogs. Note that there is no prior work that directly measures PCK on diverse animal classes as ours. Therefore, we mainly compare our full model with our simpler versions as an ablation study, *i.e*., ours with/without CSBN and with/without the sub-keypoint loss, which will be discussed in Sec. 4.3. We additionally report the WLDO's performance measured on other Animal Pose classes. Since WLDO has the most similar structure with ours that reconstructs 3D animal meshes, and can be tested on other quadruped animals, we additionally test WLDO on other proximal animal classes for broader evaluation. It is not surprising that *DeMR* shows about 10% higher PCK than WLDO, since WLDO did not see horse and cow images in training time. The result shows *DeMR*'s ability to cover more animal classes comparably even with smaller network capacity. Also, by the fact that our two-class model has better performance than our four-class one, it shows that there exists a trade-off relationship according to the number of animal classes to be dealt with.



(a) (b) (c) (d) (e)

Figure 6: **Qualitative comparison with competing methods.** (a) input image, (b) mesh prediction and (c) mesh silhouette of competing methods (top: HMR [10], bottom: WLDO [3]), (d) *DeMR*'s mesh prediction and (e) *DeMR*'s mesh silhouette for benchmarks, respectively.

Compactness. As shown in Table 1, the main advantage of *DeMR* is its competitive performance to uni-modal models while having much compact architecture in terms of the scale of network parameters. We highlight that our model has a smaller number of network parameters (27M+5M) compared to the sum of independent networks (27M+95M) of HMR and WLDO while covering 5 times broader classes of objects than HMR and WLDO.

Qualitative Results. In Fig. 5, we visualize *DeMR*'s 3D mesh estimation for humans and animals and estimated joint keypoints regressed from the estimated meshes. It shows plausible results for every target class and distinctive shapes for different species. Figure 6 shows the comparison results of *DeMR* to its competing methods, HMR and WLDO. Qualitative improvements can be found in parts, such as the limbs of the human and the dog. More qualitative results can be found in supplementary materials.

4.3 Ablation study

We ablate *DeMR* with different settings of the loss and the architecture. We define our full model as the one that uses CSBN and 10 sub-keypoint pairs in loss computation. Also, the naïve multi-task model is the one that uses no CSBN and no sub-keypoint pairs. Figure 7 shows the predicted mesh comparison across ablated models. In detail, the mesh predicted by the naïve model shows the poorest quality, while adding sub-keypoint and CSBN in training significantly improves the reconstruction quality.

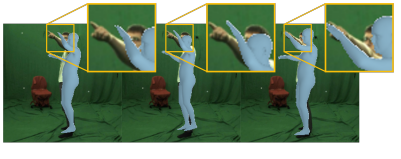


Figure 7: **Qualitative ablation.** Predicted mesh results of the models. Naïve (left), naïve with sub-keypoint (mid), and full (right).

Effects of L^*_{sub}. We evaluate how many sub-keypoints we need to enable the model to learn the morphological similarity. Table 3 shows PCK evaluated on Stanford Extra [3] according to the different numbers of sub-keypoints used in L^*_{sub}. We found that the model using 10 sub-keypoints achieves the best performance on animal reconstruction. Using 15 sub-

# of Subkpt	Labels of Sub-Keypoints: Eyes	BL	TL	ML	Nose	Class: Dog PCK [%] ↑
15	✓	✓	✓	✓	✓	73.16
10	✓	✓	✓			**73.23**
6	✓	✓				72.14
0						71.89

(a)

(b)

Table 3: **Quantitative ablation.** (a) PCK measured on various sub-keypoint numbers. (b) Sub-keypoint groups denoted by colors.

keypoint pairs in training turns out disturbing, while using 6 sub-keypoint pairs lacks the morphological information to be learned properly. The model that uses no sub-keypoint in training shows the lowest performance. Simply adding sub-keypoint loss on the naïve multi-task model improves the reconstruction performance on heterogeneous classes. More analysis and discussion can be found in the supplementary material.

Effects of CSBN. In Table 1, we compare our full model with the model without CSBN. Adding CSBN enhances the reconstruction performance on almost every dataset except Animal Pose. All the other datasets cover one target class, while Animal Pose consists of multiple animal classes. We analyze this result as the limitation stemming from the simple deployment of CSBN only in front of the human and animal branches. Thus, CSBN deals with the inter-class statistical difference only, *i.e.*, humans and animals, and restricts the statistics to be similar across quadruped animals by a single CSBN. We postulate that a fine-grained extension of CSBN like DSBN at the animal branch may further improve the performance for multiple animals.

Sample Efficiency. We evaluate the sample efficiency of *DeMR* in extremely low training data scenarios. Suppose one has a limited amount of dog data, *e.g.*, about 300 images, and wants to reconstruct 3D dogs reasonably. By virtue of our disjoint multi-task learning of *DeMR* we can leverage other species' data, including humans and quadruped animals. Table 4 shows the results on several low-data regimes. While the model trained with dog data only has poor performance, the model trained with both human and dog datasets starts to have plausible PCK of 40.83%. To fully exploit the morphological similarity, simply adding other quadruped animal data in training improves the performance by 7.2%.

	Number of datasets available				Class: Dog
Dataset #	Human	Dog	Horse	Cow	PCK [%] ↑
0.6K	0.3K	0.3K	-	-	40.83
1.2K	0.3K	0.3K	0.3K	0.3K	**48.07**
1.0K	0.3K	0.1K	0.3K	0.3K	38.50
1.2K	0.5K	0.1K	0.3K	0.3K	**43.64**

Table 4: **Quantitative results.** Sample efficiency experiments on Stanford Extra.

In the case of a smaller number of training data, *e.g.*, 100 dog images, PCK significantly decreases by 9.57% even with other class data. While not being as much power as adding additional data of the same quadruped animals, adding 200 more non-quadruped animal data, *i.e.*, human data, in training improves the performance. This shows our method of learning the morphological similarity is effective even in low-data scenarios; thus, sample efficient.

5 Conclusion

We present *DeMR*, a compact and unified 3D mesh recovery of deformable objects. To deal with heterogeneous classes, we identify the challenge of disjoint multi-task learning, and relax it to be jointly trainable by introducing morphological similarity. We embody it by proposing the concept of sub-keypoint and class-aware regularizations. Our *DeMR* encompasses broad classes with competitive performance compared to the uni-modal models despite its compactness and sample-efficiency of our model. *DeMR*'s next attention is on reconstructing broader animal classes with extreme morphological characteristics, such as elephants and giraffes. We believe that adopting implicit body representation [28] or neural-parametric model-based body reconstruction [26] along with our proposed sub-keypoint would be promising future directions. Moreover, extending *DeMR* to temporally coherent mesh regression [14] would be mandatory future research direction to enable comprehensive motion analysis of deformable objects.

Acknowledgment. This work was carried out as a part of the deformable object recognition technology research project supported by the Agency for Defense Development, Korea, and by the Defense Acquisition Program Administration, Korea (UD200025ID). This work was partially supported by the NIA grant and the IITP grant funded by the Korea government (MSIT) (Artificial Intelligence Graduate School Program; [No. 2019-0-01906, POSTECH], [No. 2020-0-01336, UNIST]). The GPU resource is supported by a study on the "HPC Support" project, supported by the MSIT and NIPA.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] 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.

[3] 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.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2016.

[5] 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.

[6] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] 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.

[8] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*, 2010.

[9] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[10] 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.

[11] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, Youngjin Yoon, and In So Kweon. Disjoint multi-task learning between heterogeneous human-centric tasks. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2018.

[14] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] 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.

[16] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(12):2935–2947, 2017.

[19] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[21] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 34(6):248, 2015.

[23] 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.

[24] Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[25] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: A flexible representation of maps between shapes. *ACM Transactions on Graphics (SIGGRAPH)*, 31(4):1–11, 2012.

[26] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[27] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] Artsiom Sanakoyeu, Vasil Khalidov, Maureen McCarthy, Andrea Vedaldi, and Natalia Neverova. Transferring dense pose to proximal animal classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[30] 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.

[31] Silvia Zuffi, Angjoo Kanazawa, David 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.

[32] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[33] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J. Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *IEEE International Conference on Computer Vision (ICCV)*, 2019.