AniMer: Animal Pose and Shape Estimation Using Family Aware Transformer (Supplementary Material)

Jin Lyu[1,*], Tianyi Zhu[2,*], Yi Gu[3], Li Lin[1,4], Pujin Cheng[1,4], Yebin Liu[5], Xiaoying Tang[1,†], Liang An[5,†]
[1]Southern University of Science and Technology
[2]China Mobile Research Institute
[3]The Hong Kong University of Science and Technology
[4]The University of Hong Kong [5]Tsinghua University

1. Datasets

Animal Pose dataset. The Animal Pose dataset [3] includes five categories: dog, cat, cow, horse and sheep, comprising a total of over 6,000 instances across more than 4,000 images. Each animal instance in Animal Pose dataset is annotated with 20 keypoints.

APT-36k dataset. The APT-36k dataset [10] contains 36000 images covering 30 different animal species from different scenes. There are typically 17 keypoints labeled for each animal instance.

AwA Pose dataset. The AwA Pose dataset [1] is introduced for 2D quadruped animal pose estimation. AwA contains 10064 images of 35 quadruped animal species and each image is annotated with 39 keypoints.

Stanford Extra dataset. The Stanford Extra dataset [2] consists of 20,580 images and covers 120 dog breeds. Each image is annotated with 20 2D keypoints and silhouette.

Zebra synthetic dataset. The Zebra synthetic dataset [12] consists of 12850 images. Each image is randomly generated that differs in background, shape, pose, camera, and appearance.

Animal Kingdom dataset. The Animal Kingdom dataset [7] includes a diverse range of animal species. We only use 8 major animal classes of pose estimation dataset to evaluate our method.

Animal3D dataset. Animal3D dataset [9] contains a total of 3379 images, which are classified into 40 classes. Each image is annotated with SMAL [11] parameters, 2d keypoints, 3d keypoints and masks.

CtrlAni3D dataset. Our dataset is annotated in the same style as Animal3D dataset. More details about our dataset can be found in Sec. 2.

For all datasets, we filter out images of animals not included in SMAL [11], such as elephants. We then aggregate all the aforementioned datasets (excluding Animal Kingdom) for training, assigning different sampling weights to each dataset based on its type and size, as shown in Tab. S1.

Table S1. **Full dataset statistics for training.**

Dataset	Number	Ratio	Training Sample Weight
Animal3D	3065	7.4%	1
CtrlAni3D	8277	20.0%	0.5
Animal Pose	1680	4.0%	0.15
AwA-Pose	2884	7.0%	0.15
Zebra Synthetic	12850	31.1%	0.05
Stanford Extra	7689	18.6%	0.15
APT-36K	4887	11.8%	0.15
Total	41332	100%	-

2. More Details about CtrlAni3D

Each image in the CtrlAni3D dataset is annotated with SMAL parameters, including $\beta \in \mathbb{R}^{41}$, $\theta \in \mathbb{R}^{35\times3}$ (expressed by axis angle), and $\gamma \in \mathbb{R}^3$. Additionally, similar to Animal3D [9], CtrlAni3D provides annotations for 26 3D keypoints and their corresponding 2D keypoints. The visibility of 2D keypoints is determined by comparing the depth d_k of the keypoint with the depth d_p at the corresponding pixel location. Specifically, visibility is set to 1 when $d_k \leq d_p$; otherwise, it is set to 0.

During the generation of CtrlAni3D dataset, we prompt the ControlNet using common names instead of scientific names of animals. However, to better indicate the position our CtrlAni3D dataset in the animal taxonomy, we list the most relevant scientific names of used animal species in Tab. S2. During the image generation process, we require human annotators to filter out misaligned results, as described in main text Sec.4. Such misaligned results, denoted as "Failure cases", are illustrated in Fig. S1.

In addition, COCO backgrounds are used only when SAM2 achieves satisfactory segmentation quality. Therefore, increasing the ratio of COCO backgrounds is equivalent to

* Equal contribution.
† Corresponding author.

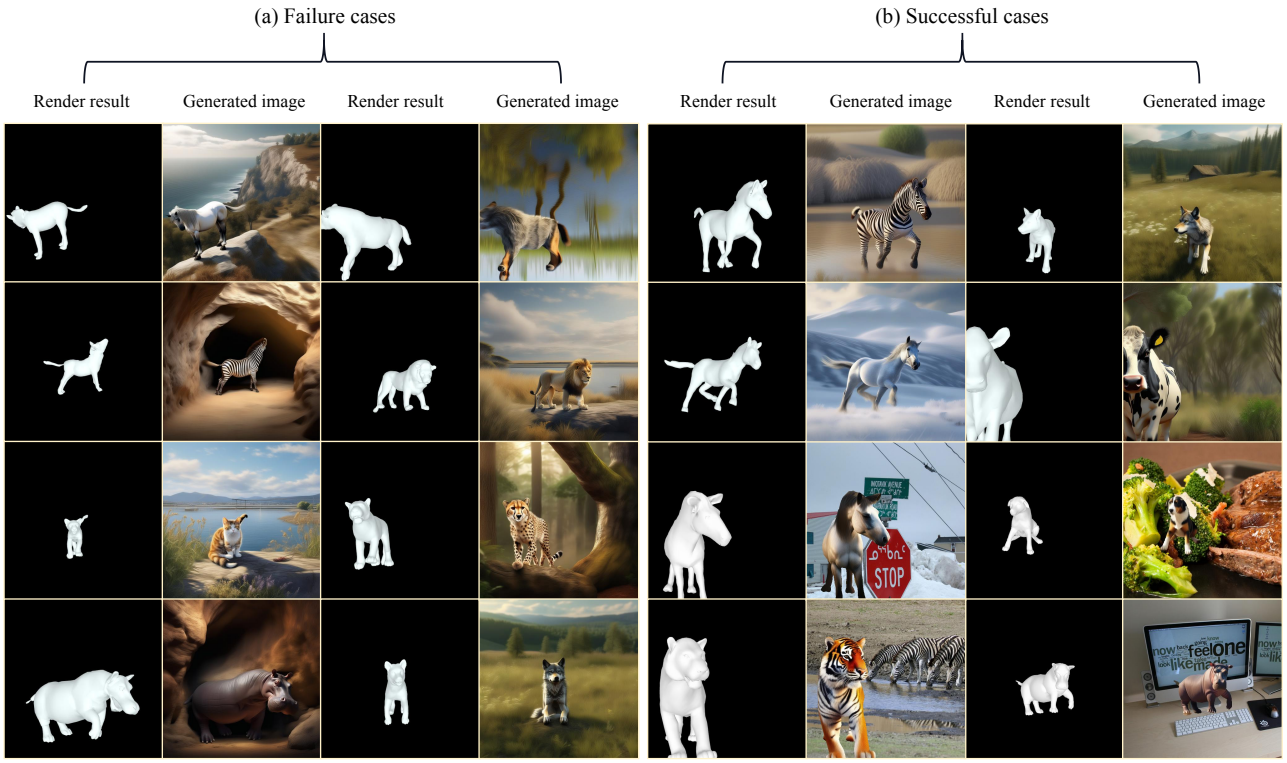

Figure S1. **CtrlAni3D failure cases and successful cases.** (a) Failure cases. There are two main cases of failure: (1) At times, ControlNet may struggle to generate mesh-aligned poses (first row and second row). (2) Additionally, ControlNet may not effectively generate the intricate details of the animal body (third row and fourth row). (b) Successful cases. The backgrounds of the first and second rows are generated by ControlNet, while the backgrounds of the third and fourth rows are sourced from the COCO dataset.

lowering the IoU threshold, which results in a decline in data quality. This may negatively impacts model training.

Table S2. **Scientific names of used animal species in CtrlAni3D.** The image counts of each species are listed at the right column.

Family	Species	Prompt Commands	Count
Felidae	*Felis catus*	Cat	80
	Panthera leo	Lion	630
	Acinonyx jubatus	Cheetah	299
	Panthera tigris	Tiger	280
Canidae	*Canis lupus familiaris*	Dog	2976
	Canis lupus	Wolf	413
Equidae	*Equus ferus caballus*	Horse	2228
	Equus zebra	Zebra	1460
Bovidae	*Bos taurus*	Cow	890
Hippopotamidae	*Hippopotamus amphibius*	Hippo	455
Total			9711

3. Comparison between CtrlAni3D and Animal3D

CtrlAni3D and Animal3D are both based on SMAL. As a result, both datasets encompass five animal families, as presented in Tab. S2. Animal3D includes more subcategories (e.g., bighorn) compared to CtrlAni3D, which has a greater number of entries. However, there are some animal species that SMAL doesn't express very well (in the second row of Fig. S2). In addition, Animal3D requires manual 2D annotations for fitting, which introduces a degree of error, as shown in the first row of Fig. S2. CtrlAni3D ensures the quality of data through cycle consistency and manual filtering.

Figure S2. **Some bad cases in Animal3D.**

4. More results and analysis.

Effect of CtrlAni3D for 3D pose estimation. To demonstrate the improvements of CtrlAni3D for 3D pose estima-

tion, we report the results of the 3D metrics on Animal3D, as shown in Tab. S3. We can observe that training with CtrlAni3D enhances performance on Animal3D.

Table S3. **Effect of including CtrlAni3D in training.** We evaluate the performance of 3D pose estimation on two models. For the first model, we do not use CtrlAni3D during training, while for the second model, we incorporate CtrlAni3D into the training process.

	PA-MPJPE ↓	PA-MPVPE ↓
AniMer(no CtrlAni3D)	82.6	88.4
AniMer(with CtrlAni3D)	**80.4**	**85.7**

Domain gap between CtrlAni3D and real-world data. Although CtrlAni3D can improve the model performance on the real-world data, there still exist a certain domain gap. Tab. S4 shows a certain domain gap between Animal3D and CtrlAni3D. However, the comparable results between AniMer(A3D) and AniMer(C3D) on the Animal Kingdom dataset indicate a similar generalization ability on in-the-wild data between Animal3D and CtrlAni3D. This is why we aggregate many datasets for full training. The performance gain shown in Tab.2 (in main text) further validates the effectiveness of CtrlAni3D to assist in generalization.

Table S4. **The generalizability of CtrlAni3D.** AniMer(A3D) trains only on Animal3D, AniMer(C3D) trains only on CtrlAni3D.

Method	Animal3D		CtrlAni3D		Animal Kingdom	
	PCK@HTH ↑	AUC ↑	PCK@HTH ↑	AUC ↑	PCK@HTH ↑	AUC ↑
AniMer(A3D)	87.0	86.0	89.7	89.9	78.0	78.6
AniMer(C3D)	83.8	81.9	93.5	95.0	77.8	80.3

Ablations on different encoder and decoder. To emphasize the significance of the ViT encoder and the SMAL transformer decoder, we substitute the ViT encoder with a ResNet-152 encoder and the SMAL transformer decoder with an MLP decoder, respectively. The results are presented in Tab. S5.

Table S5. **Ablations on different encoder and decoder.** AniMer-b: use ResNet-152 as encoder. AniMer-e: use MLP as decoder.

Method	Animal3D		CtrlAni3D		Animal Kingdom	
	PA-J ↓	PA-V ↓	PA-J ↓	PA-V ↓	AUC ↑	PCK@0.1 ↑
AniMer-b	115.5	128.7	117.0	129.4	68.9	10.2
AniMer-e	83.9	89.2	55.8	60.9	81.9	31.6
AniMer	**80.4**	**85.7**	**44.1**	**47.6**	**82.9**	**34.9**

More discussion about contrastive learning. The contrastive loss directly impacts the feature tokens $\mathbf{F}$, which in turn indirectly impacts the feature vectors $\boldsymbol{f}$ and aligns features to model the global structure, capturing family differences. This ensures that the final output shape aligns more closely with the category of the input image. Compared with contrastive learning, $\mathcal{L}_{\text{cls}}$ ("w $\mathcal{L}_{\text{cls}}$" in Tab. S6) focuses solely on optimizing classification accuracy, which may not necessarily improve geometric parameter regression. Moreover, contrastive learning facilitates a more compact intra-class distribution and a more separable inter-class distribution in the feature space [6], thereby enhancing the model's capability for few-shot learning. In Tab. S6, we report the results for various animals. $\mathcal{L}_{\text{con}}$ can improves performance for animals with limited training samples (e.g., boars are less than one percent of the training set).

Table S6. **Evaluation of some species in A3D.** PA-J: PA-MPJPE, PA-V: PA-MPVPE.

species	w $\mathcal{L}_{\text{cls}}$		w/o $\mathcal{L}_{\text{con}}$		w $\mathcal{L}_{\text{con}}$	
	PA-J	PA-V	PA-J	PA-V	PA-J	PA-V
dog	74.9	81.3	72.1	76.7	**71.1**	**75.1**
zebra	66.3	68.3	60.4	63.7	**60.6**	**62.7**
horse	78.5	86.6	77.4	86.5	**75.9**	**84.1**
cat	**129.4**	**132.0**	134.5	136.1	131.2	132.8
cow	83.0	86.0	80.3	84.8	**78.1**	**83.2**
sheep	83.9	**88.0**	83.8	91.1	**80.1**	88.5
bear	79.4	80.0	**76.5**	80.5	76.8	**79.3**
boar	126.5	158.7	119.1	150.5	**115.9**	**142.6**

Ablation study on two stage training. We compare to "AniMer-c", which trains the AniMer model for one stage. Both models are trained using the same batchsize and training steps. The two-stage training makes the model training more stable. By increasing the training steps or tuning the other hyperparameters, one-stage training may achieve comparable results. Quantitative results are shown in Tab. S7, and qualitative results are shown in Fig. S3.

The effect of different setting. Similar to the findings of [4], our model exhibits varying performance metrics (PA-MPJPE(↓) on Animal3D: $87-78$, PCK@0.15(↑) on Animal Kingdom: $0.5-0.6$) under different settings (e.g., varying hyperparameters, different devices).

5. More Qualitative Results

We provide qualitative results from the Animal Kingdom dataset in Fig. S4. For each case, we display the input image and the output results, which include both a front view rendering and a side view rendering. It can be observed that AniMer performs well even in challenging conditions such as motion blur (the second sample in the first row), unusual lighting (the first sample in the third row), partial occlusion (the first sample in the fourth row), and truncation (the second sample in the fifth row).

6. Failure Cases and Discussion

We provide failure cases in Fig. S5. Although AniMer demonstrates strong robustness, it can fail in certain scenarios. For example, large-scale occlusion (first row), extreme poses (second row) and excessively blurred images (third row) can lead to large reconstruction errors.

Table S7. **Quantitative comparisons on Animal3D, CtrlAni3D and AnimalKingdom datasets. Bold** numbers indicate the best values. P@H, P@0.1, P@0.15, PAJ, and PAV represent PCK@HTH, PCK@0.1, PCK@0.15, PA-MPJPE, and PA-MPVPE, respectively.

Dataset	Animal3D				CtrlAni3D				Animal Kingdom			
Metric	AUC↑	P@H↑	PAJ↓	PAV↓	AUC↑	P@H↑	PAJ↓	PAV↓	AUC↑	P@H↑	P@0.1↑	P@0.15↑
AniMer-c	87.2	86.3	85.9	90.4	91.7	93.4	59.5	64.2	80.6	80.4	28.6	47.5
AniMer	**88.9**	**89.5**	**80.4**	**85.7**	**93.8**	**95.4**	**44.1**	**47.6**	**82.9**	**83.7**	**34.9**	**54.7**

Figure S3. **More qualitative results on Animal3D and CtrlAni3D dataset.** We compare our results with HMR [5], WLDO [2], AniMer-a (ResNet152 backbone), AniMer-b (no pretraining), AniMer-c(train only one stage) and HMR2.0 [4].

Our framework is based on SMAL, which is suitable for most quadrupedal animals. However, animals such as mice, fish, and birds cannot be represented using SMAL. As a result, we plan to adapt AniMer to accommodate a broader range of animal species in the future.

In addition, with the advent of large-scale synthetic

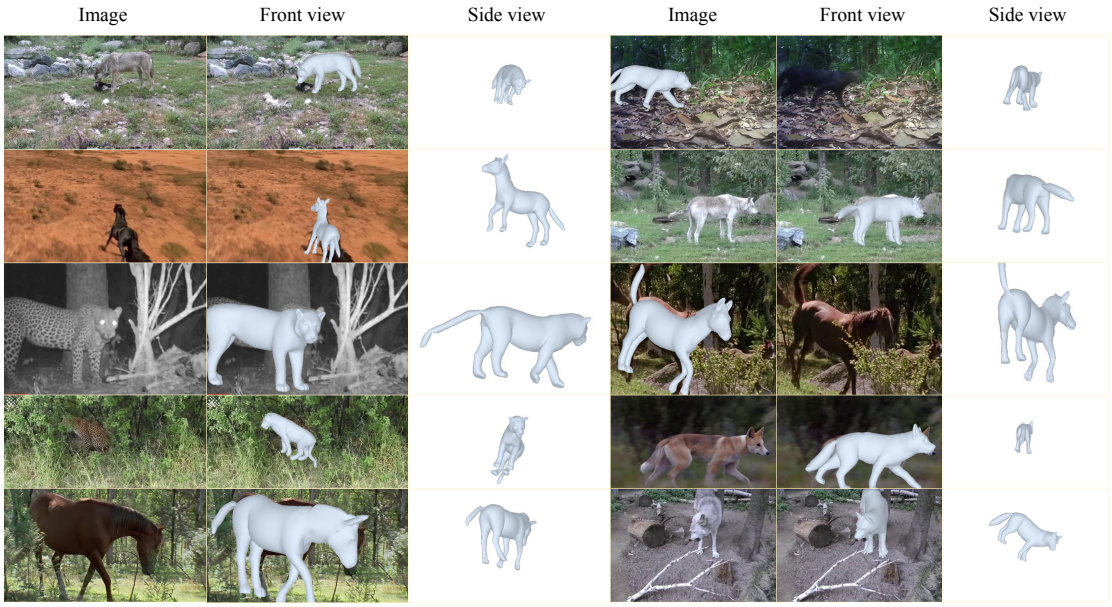

Figure S4. **Results on the Animal Kingdom dataset.**

datasets (e.g., GenZoo [8]), we will further explore the performance of dataset scaling and contrastive learning on these extensive datasets.

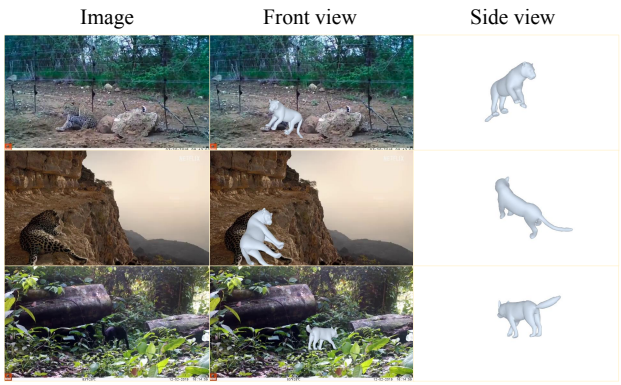

Figure S5. **Failure cases.**

References

[1] Prianka Banik, Lin Li, and Xishuang Dong. A novel dataset for keypoint detection of quadruped animals from images. *arXiv preprint arXiv:2108.13958*, 2021. 1

[2] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out?: 3D animal reconstruction with expectation maximization in the loop. In *ECCV*, 2020. 1, 4

[3] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9498–9507, 2019. 1

[4] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 3, 4

[5] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 4

[6] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3

[7] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19023–19034, 2022. 1

[8] Tomasz Niewiadomski, Anastasios Yiannakidis, Hanz Cuevas-Velasquez, Soubhik Sanyal, Michael J Black, Silvia Zuffi, and Peter Kulits. Generative zoo. *arXiv preprint arXiv:2412.08101*, 2024. 5

[9] Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9099–9109, 2023. 1

[10] Yuxiang Yang, Junjie Yang, Yufei Xu, Jing Zhang, Long Lan, and Dacheng Tao. Apt-36k: A large-scale benchmark for

animal pose estimation and tracking. *Advances in Neural Information Processing Systems*, 35:17301–17313, 2022. 1

[11] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6365–6373, 2017. 1

[12] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5359–5368, 2019. 1