Animal3D: A Comprehensive Dataset of 3D Animal Pose and Shape

Jiacong Xu[1] Yi Zhang[1] Jiawei Peng[1] Wufei Ma[1] Artur Jesslen[11] Pengliang Ji[3]
Qixin Hu[4] Jiehua Zhang[5] Qihao Liu[1] Jiahao Wang[1] Wei Ji[6] Chen Wang[7]
Xiaoding Yuan[1] Prakhar Kaushik[1] Guofeng Zhang[8] Jie Liu[9] Yushan Xie[2]
Yawen Cui[5] Alan Yuille[1] Adam Kortylewski[10,11]

[1]Johns Hopkins University [2]East China Normal University [3]Beihang University
[4]HUST [5]University of Oulu [6]University of Alberta [7]Tsinghua University [8]UCLA
[9]City University of Hong Kong [10]Max Planck Institute for Informatics [11]University of Freiburg

Xu, Jiacong, et al. "Animal3d: A comprehensive dataset of 3d animal pose and shape." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

Contributions

In summary, our main contributions are:

- We present Animal3D, the first benchmark for mammal animal 3D pose and shape estimation, with a diverse set of 40 mammal species, and high-quality annotations of 2D keypoints as well as 3D shape and pose parameters of the SMAL [50] model.
- We set up a set of baselines on Animal3D in various settings using state-of-the-art methods which demonstrates the versatility of the dataset.
- Our experimental results and in-depth analysis of the strengths and limitations of representative methods demonstrate the challenging nature of our benchmark.

	Animal3D (Ours)	Animal Pose[6]	Stanford Extra[2]	AP-10K [46]
Segmentation	✓	✗	✓	✗
3D Anno.	✓	✗	✗	✗
#Species	40	5	Dogs	54
#Keypoints	26	20	20	17
#Images	$3.4K$	$4K$	$8.1K$	$10K$

Table 1: Comparison of Animal3D with other animal datasets. Animal3D contains class labels of 40 species, 26 keypoints, and 3D pose and shape parameters from the SMAL model. Totally, there are 5.1k images are carefully annotated in Animal3D, but only 3.4k images are selected after 3-round inspection. The unselected images and annotations will also be published together with Animal3D.

Data Example

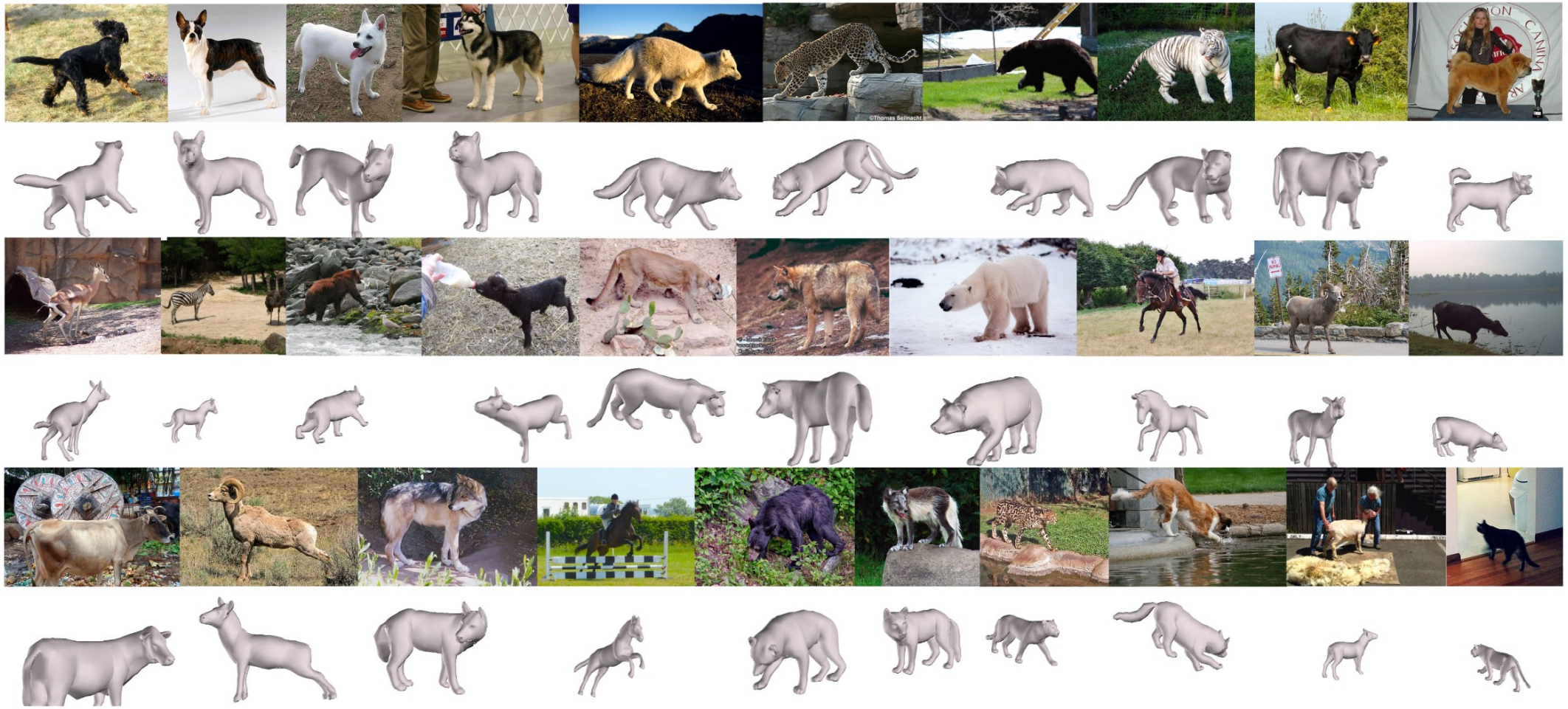


Figure 1: Samples from the proposed Animal3D dataset. Our dataset contains a diverse range of animal species with high-quality annotations of shape and pose parameters using the popular SMAL [50] model.

Data Annotation Pipeline

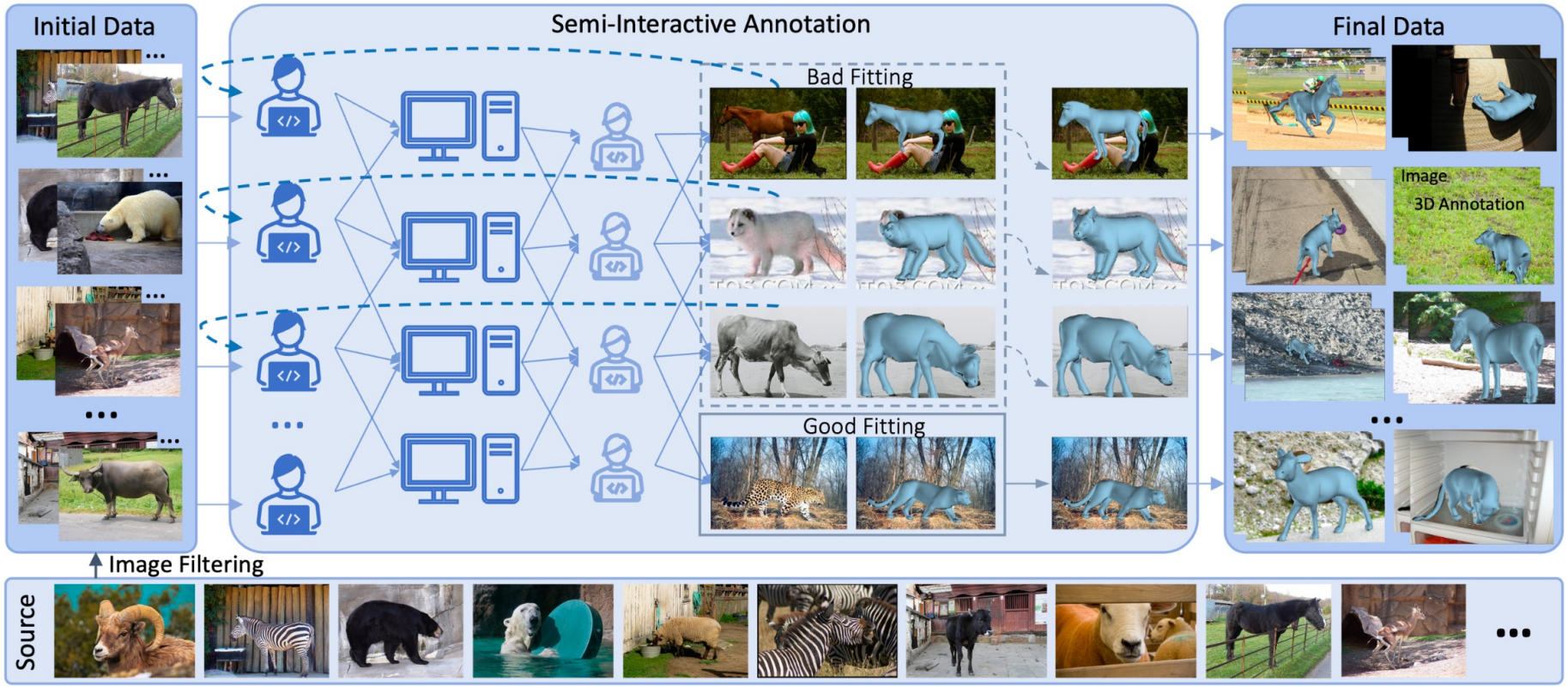


Figure 2: Data annotation pipeline for Animal3D. The process consists of three stages: Image Filtering, Semi-Interactive Annotation, and Data Integration. The data is sourced and filtered to obtain an initial set of images. During the Semi-Interactive Annotation, annotators submitted their annotation to the server to fit the SMAL model and render the results on the images. Then a set of inspectors examined the fitting results and send the bad-fitting images back to the annotator for revision. This process is repeated multiple times. Images that constantly lead to bad-fitting results are removed.

- Annotation Pipeline
  - Annotators
    - 2D keypoints
  - Inspectors:
    - fit 3D SMAL model
  - Semi-interactive
    - 3 rounds annotate-then-examine

Fitting SMAL to images

shape pose translation

SMAL model parameter: $\Pi(\beta, \alpha, t)$

Optimize: $$\mathcal{L}_{total}(\Pi, M) = \mathcal{L}_{kp}(\Pi, M) + \mathcal{L}_{silh}(\Pi, M) + \mathcal{L}_{shape}(\beta).$$

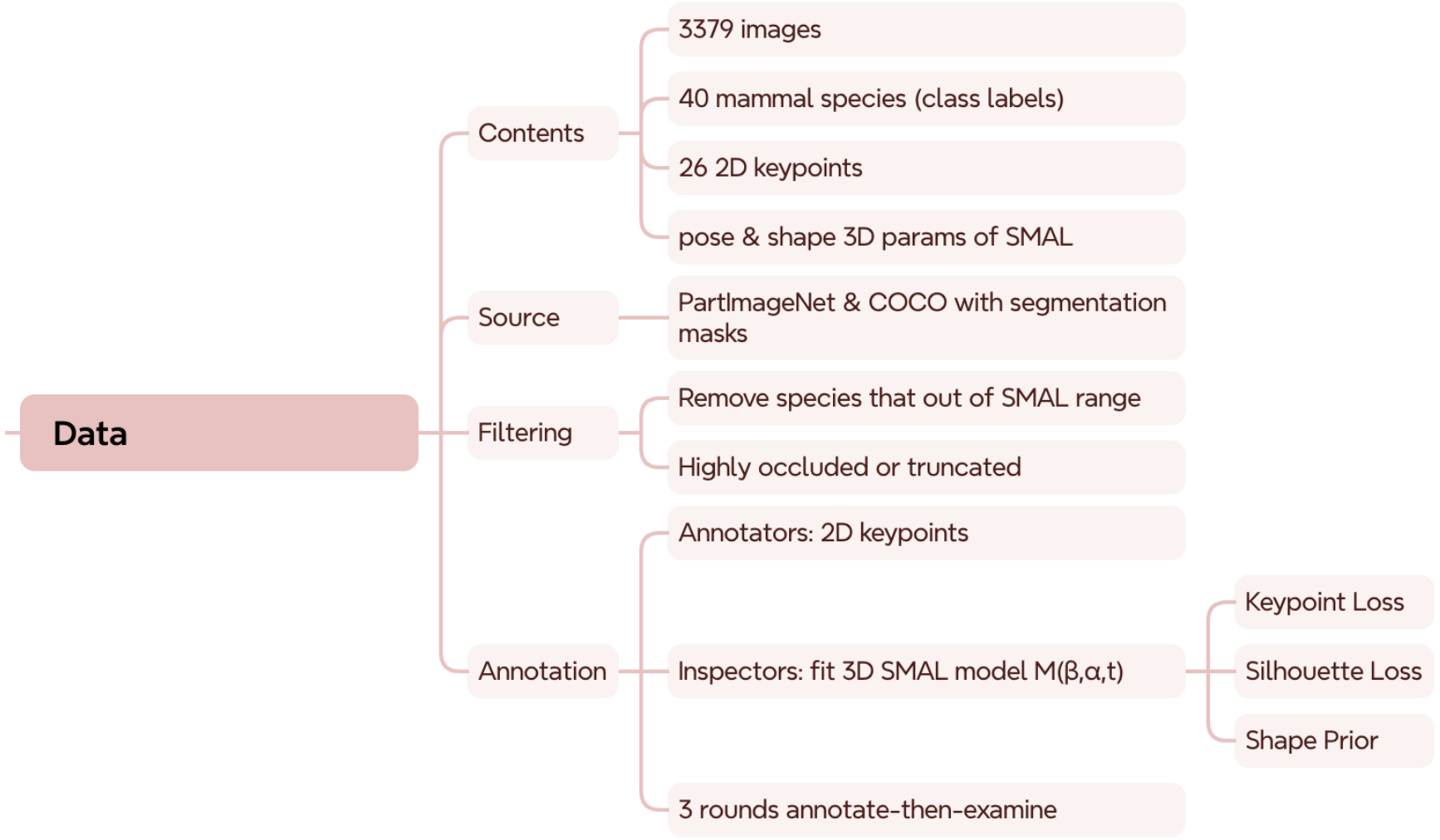
$$\mathcal{L}_{kp}(\Pi, M) = \sum_{i=1}^{26} \rho(|| \sum_{j=1}^{N^i} P(\mathbf{v}_j^i)/N^i - t_i||_2),$$

$$\mathcal{L}_{silh}(\Pi, M) = \sum_{x \in S} \mathcal{D}_{\bar{S}}(x) + \sum_{x \in S} \rho(\min_{\hat{x} \in S} ||x - \hat{x}||_2),$$

in the original work. We denote $P(\mathbf{v}_i)$ as the perspective projection of the the i'th mesh vertex into the image plane. Moreover, $P(M) = S$ is the projected mesh silhouette. To

Shape prior. We regularize the shape parameters β using a shape loss $\mathcal{L}_{shape}(\beta)$ using the PCA prior distribution. In paricular, the loss is defined to be the squared Mahalanobis distance defined using the PCA eigenvalues.

Dataset Summary



Baselines

Method	Supervised			Synthetic to Real			Human pre-trained		
	PA-MPJPE↓	S-MPJPE↓	PCK↑	PA-MPJPE↓	S-MPJPE↓	PCK↑	PA-MPJPE↓	S-MPJPE↓	PCK↑
HMR [16]	140.7	496.2	59.3	124.8	497.7	63.1	132.2	488.0	60.6
PARE [17]	134.8	443.9	79.1	127.2	392.3	83.7	130.7	374.9	85.6
WLDO [2]	128.8	502.1	60.1	123.9	484.0	65.1	-	-	-

Table 2: 3D pose and shape estimation results on the Animal3D dataset. We evaluate three representative baseline models, HMR, PARE and WLDO, in three settings: (1) Supervised on Animal3D data only, (2) Pre-training on synthetic data and fine-tuning on Animal3D, and (3) Pre-training on Human Pose Estimation datasets and fine-tuning on Animal3D. While pre-training improves results for all models, the final results are lower compared to object specific benchmarks for humans and dogs, hence indicating the difficulty of estimating 3D animal pose across species.

- 3059 Training & 320 test images

- For experiments with synthetic data, we pre-train the models for 100 epochs. For training on real data, all the models are trained for 1000 epochs (around 24k iterations).

[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 *ECCV*, 2020.
[16] 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/CVF conference on computer vision and pattern recognition, pages 7122–7131, 2018.
[17] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11127– 11137, 2021.

From Synthetic to Real

Inspired by [31], which utilize rendered images from CAD model of animals to boost the model performance on 2D tasks, we synthesize 45k images using SMALR [49] and select 40k for training and 5k for testing models, respectively, before fine-tuning them on Animal3D.

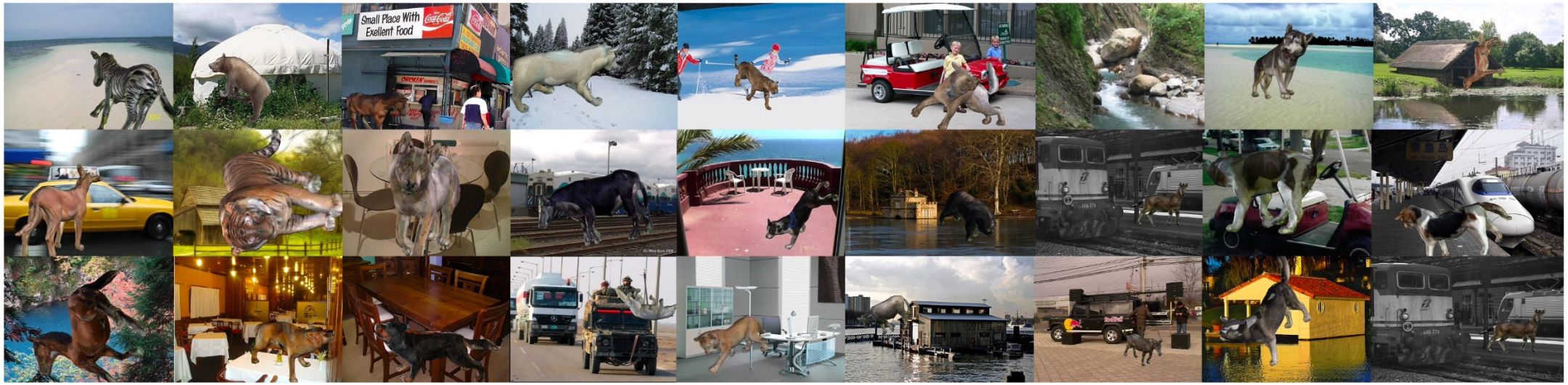


Figure 4: Example images from our synthetic dataset that is used for pre-training the animal pose estimation baselines. We simulate all species from the Animal3D dataset using the SMALR model in varying poses, shapes, and background images.

Qualitative Results