

BITE: Beyond Priors for Improved Three-D Dog Pose Estimation

Supplementary Material

Nadine Rüegg^{1,2}, Shashank Tripathi², Konrad Schindler¹, Michael J. Black², and Silvia Zuffi³

¹ETH Zürich, Switzerland

²Max Planck Institute for Intelligent Systems, Tübingen, Germany

³IMATI-CNR, Milan, Italy

1. 2D Reprojection Errors

For lack of a 3D test set, previous work on 3D dog shape and pose estimation has most commonly been evaluated in terms of 2D reprojection scores, namely intersection over union (IOU) and percentage of correct keypoints (PCK). While we maintain that those metrics are unable to capture the 3D quality of a predicted shape (see discussion in the main paper), we report them for the sake of completeness. Table S.1 shows a comparison between our method and related work, and Tab. S.2 reports scores for all ablated versions of our BITE. We outperform previous state-of-the-art methods by a considerable margin: our full method reaches an IOU of 85.5 and a PCK of 86.3, compared to 81.6 IOU and 83.4 PCK for the closest competitor (CTF). The threshold for a correct keypoint in PCK was set to 0.15, as in [3].

	3D errors [cm]		rep. scores	
	s→m	m→s	IOU	PCK
WLDO [1]	2.65	7.55	74.2	78.8
CTF [2]	2.59	6.17	81.6	83.4
BARC [3]	2.40	3.93	75.1	82.8
BITE (ours)	2.07	3.15	79.4	84.8
BITE ttopt (ours)	2.03	2.84	85.5	86.3

Table S.1. **Comparison to SOTA.** Scan-to-mesh and mesh-to-scan distances on our novel 3D test set, and IOU as well as PCK scores on the Stanford Extra test set.

2. 3D Evaluation

This section shows visual examples of input images from our test set and elaborates on the calculation of scan-to-mesh and mesh-to-scan distances.

	3D errors [cm]		rep. scores	
	s→m	m→s	IOU	PCK
BARC+ w/o gc	2.32	3.92	76.1	82.6
BARC+ with gc	2.13	3.48	75.3	80.7
BITE w/o gc	2.30	4.16	80.5	85.6
BITE w/o sampler	2.09	3.31	78.4	83.6
BITE with shape in BARC+	2.12	3.17	79.6	85.2
BITE with shape in ref	2.29	4.24	79.8	85.7
BITE (ours)	2.07	3.15	79.4	84.8
BITE ttopt (ours)	2.03	2.84	85.5	86.3

Table S.2. **Ablation study.** Scan-to-mesh and mesh-to-scan distances on our novel 3D test set, and IOU as well as PCK scores on the Stanford Extra test set.

2.1. Input Images

In Figure S.1 we provide an overview of the textured 3D scans of real dogs that make up our 3D test dataset. For each scan, the figure displays a randomly selected rendering among the 7 available frontal viewpoints, cropped according to the BITE preprocessing pipeline.

2.2. Scan-to-Mesh and Mesh-to-Scan Distance Calculation

To evaluate results on our new 3D test set, we follow the scheme proposed by [4]. First, a rigid alignment (rotation, translation, and scaling) between the prediction and the ground truth scan is performed based on a few easy-to-locate, labeled keypoints such as for example toes and eyes. To further refine the alignment, we continue with a rigid alignment that minimizes the absolute distance between each scan (*i.e.*, ground truth) vertex and the nearest point in the predicted mesh. Given this alignment, we calculate two error measures for each image:

- **Scan-to-mesh distance:** The scan-to-mesh distance



Figure S.1. *3D test dataset*. Scans of real dogs were textured and rendered. For each scan we show one out of 7 renderings from different frontal viewpoints. The renderings are cropped according to the BITE preprocessing pipeline.

between the ground truth scan and the reconstructed mesh is calculated as the absolute distance between each scan vertex and the nearest point on the predicted mesh.

- **Mesh-to-scan distance:** Because the predicted mesh has SMAL topology and thus vertices are denser on head and paw regions than on other body parts, we calculate also a mesh-to-scan distance based on a uniformly re-sampled mesh surface, obtained through Voronoi clustering on the mesh of the mean shape in standard t-pose, as described in Section 3.3. The mesh-to-scan distance is the average distance between each of those vertices and the closest point on the ground truth scan.

3. Stability

Our work on ground contact enforces a simple, pervasive mechanical constraint. One effect of this constraint is that body parts that are likely to touch the ground actually touch it, consequently our predicted 3D dogs are more likely to be stable. To illustrate the stability of our predictions, we take a random set of resulting 3D dogs in supposedly stable poses (i.e. not running or jumping) and place them in a “Bullet” physics simulation. If stable, the animals should remain standing under the force of gravity. Figure S.2 illustrates 100 dogs as predicted by our network. To the left, we show for each dog the orientation w.r.t. the floor, as predicted by BITE-ttopt. The right image depicts the dogs by the end of the physics simulation. We find that 93 out of 100 dogs remain standing, indicating that most of our predictions satisfy stability under the influence of gravity.

4. Failure Cases

The most common failure cases of BITE are illustrated in Fig. S.3. They concern images with occlusions, dogs viewed from the back, body parts outside the image borders, rare and complicated poses, as well as images with multiple dogs.

5. D-SMAL

In the following, we provide further details about the learning procedure of the D-SMAL model.

Method. We use 39 toy figurines representing animals in the canine family, thus adding 34 toy scans of dogs to the those of the original SMAL training set, which contained two dogs, a fox, a wolf and a hyena from the canine group. The new set of toys includes different breeds, covering many popular ones, with variation in size, type of ears and fur length. Information about the breeds in the training set is provided in Fig. S.5, where the breed label is placed below each 3D scan, in gray. In D-SMAL we also consider a finer segmentation of the SMAL template, adding two more body parts to the SMAL model skeleton: left and right ear. The resulting skeleton is compatible with SMAL, as the two additional parts are leaves in the kinematic tree, and their relative pose can be set to zero for backward compatibility. We follow the SMAL learning pipeline [5], namely we first register all the toys with the GLoSS algorithm, then we place the toys in a reference pose and learn the shape space with Principal Component Analysis (PCA) over the pose-normalized training set.

3D landmarks. To guide the GLoSS registration, we manually annotated all the toy scans with 38 surface landmarks, extending the 24 landmarks used for the original SMAL model. This was necessary to reduce the alignment errors when fitting the GLoSS template to the toy scans. For several breeds, especially those with long fur, the registration with a smaller set of landmarks was prone to errors. The manually annotated 3D landmarks are: nose tip, lower lip, eye point toward the nose, eye point toward the ear, ear tip, neck back, belly, back midpoint, shoulder, armpit, wrist, feet, hip, knee, tail begin, tail middle, tail tip, pelvis, middle chest, neck base, between ears, tail 1/8, between shoulder blades, back of the hip, cheek.

Registration. We follow the GLoSS registration procedure and augment it where necessary. We add extra side information, namely a flag that indicates whether the dog has floppy ears. In that case, we set the ears’ 3D rotation in the GLoSS model template accordingly. We also indicate if the dog has the mouth closed. If that is the case, vertices on the inside of the mouth are ignored in the mesh-to-scan loss, furthermore we also define a lip-matching loss. This is applied during both main steps of the GLoSS registration pipeline (see [5]), model-based and model-free registration.

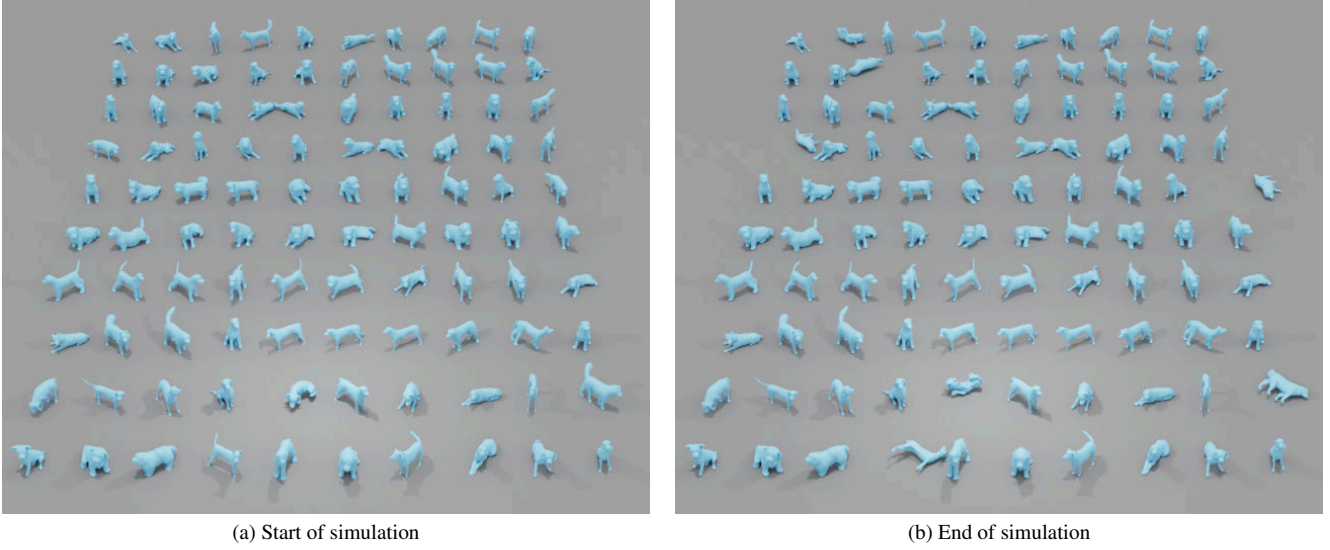


Figure S.2. *"Bullet" physics simulation.* Initialization and final output of a bullet physics simulation for 100 predicted dogs from BITE-ttopt.

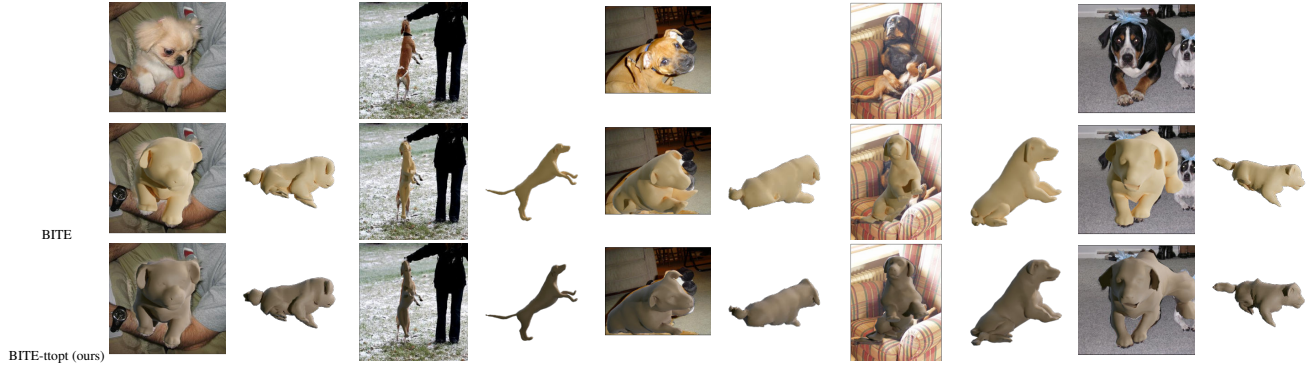


Figure S.3. *Failure Cases.* From top to bottom: input image, BITE and BITE-ttopt.

Finding a good set of hyper-parameters for the GLoSS energy took some effort. The final set was selected after an extensive search, in such a way that the same hyper-parameters could be kept for all samples in the training set.

We had to discard one case with extremely long fur, the Shih-Tzu in Figure S.7, since the fur hides the legs completely and their registration cannot be determined. Figure S.5 illustrates the 3D scans, in gray; and the outcome of the registration, with different colors for different body parts.

Unposing. Given GLoSS estimates of the scan 3D pose, we reverse the forward kinematics to obtain all registered samples in a reference pose (commonly referred to as *T-pose*, borrowing the term from human models). This is illustrated in Figure S.6. Since we observed a few cases with unrealistic limb proportions, where the back legs were too long, we further improve the registered instances by scaling the front and back legs to the same length, as explained in

the paper. On the final set thus obtained, we perform PCA to learn the D-SMAL model. A visualization of the PCA shape space is presented in Figure 2 in the main paper, with the D-SMAL template and shape variations of $\pm 3\sigma$ along the leading principal directions.

6. Qualitative Results

6.1. BARC+ vs. BARC

In this analysis, we aim to visually demonstrate the impact of replacing the SMAL model with our new D-SMAL model. To that end, we train the BARC network twice, to obtain two variants: one that uses SMAL as in the original paper, termed BARC; and on that uses D-SMAL as in our initialisation, termed BARC+. Figure S.4 compares their predictions for a number of test images. One can clearly see that the shape improves, and breed-specific features are

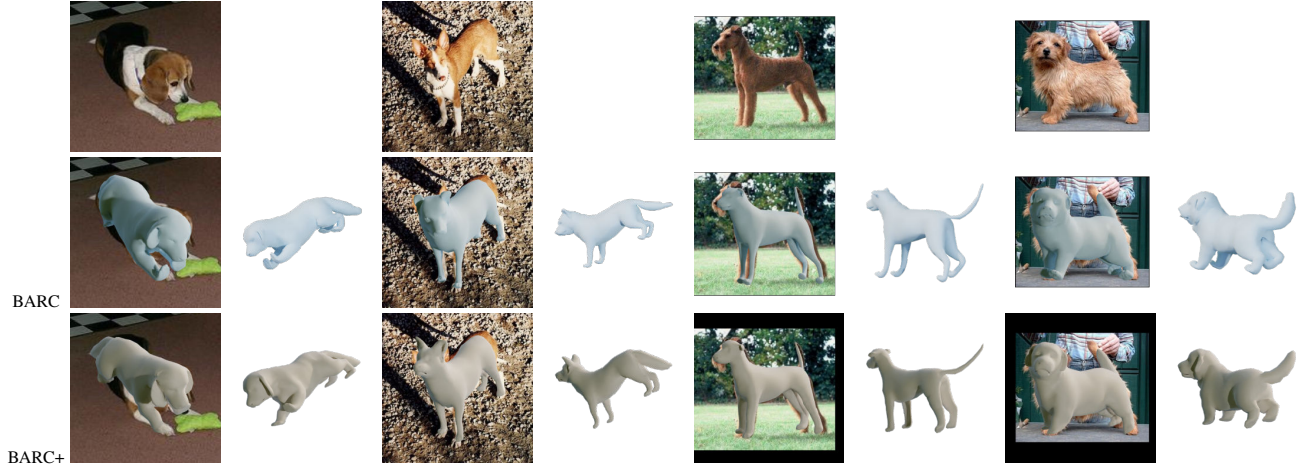


Figure S.4. *Shape Improvements*. Top row: input image, Middle: BARC, bottom row: BARC+. Note that these are not full BITE results. The improvements illustrate the effect of the enhanced D-SMAL dog model in isolation.

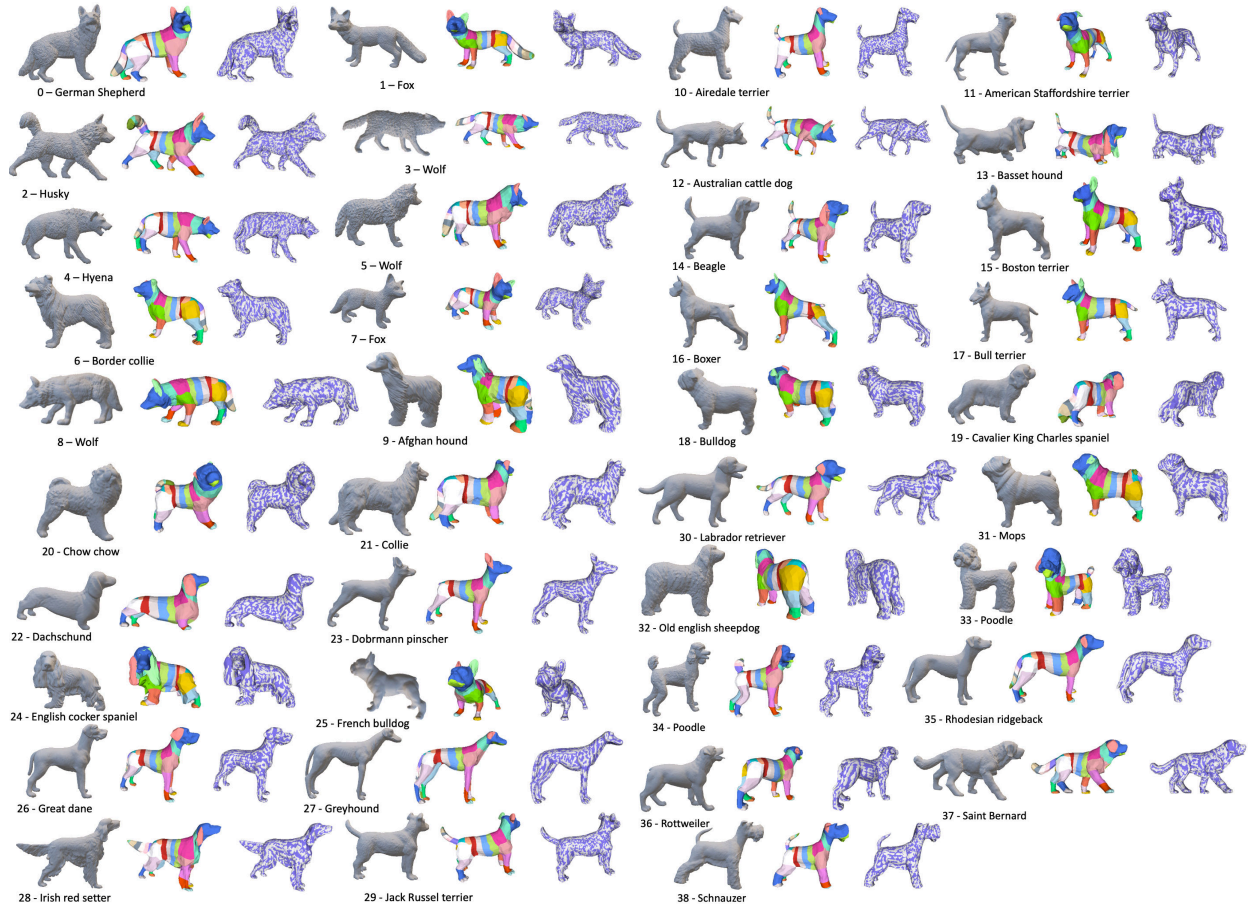


Figure S.5. **Registration**. 3D scans (gray), the registration result with colors denoting body parts, and the result overlaid on the scan (purple and white, respectively).

recovered more faithfully when using D-SMAL.

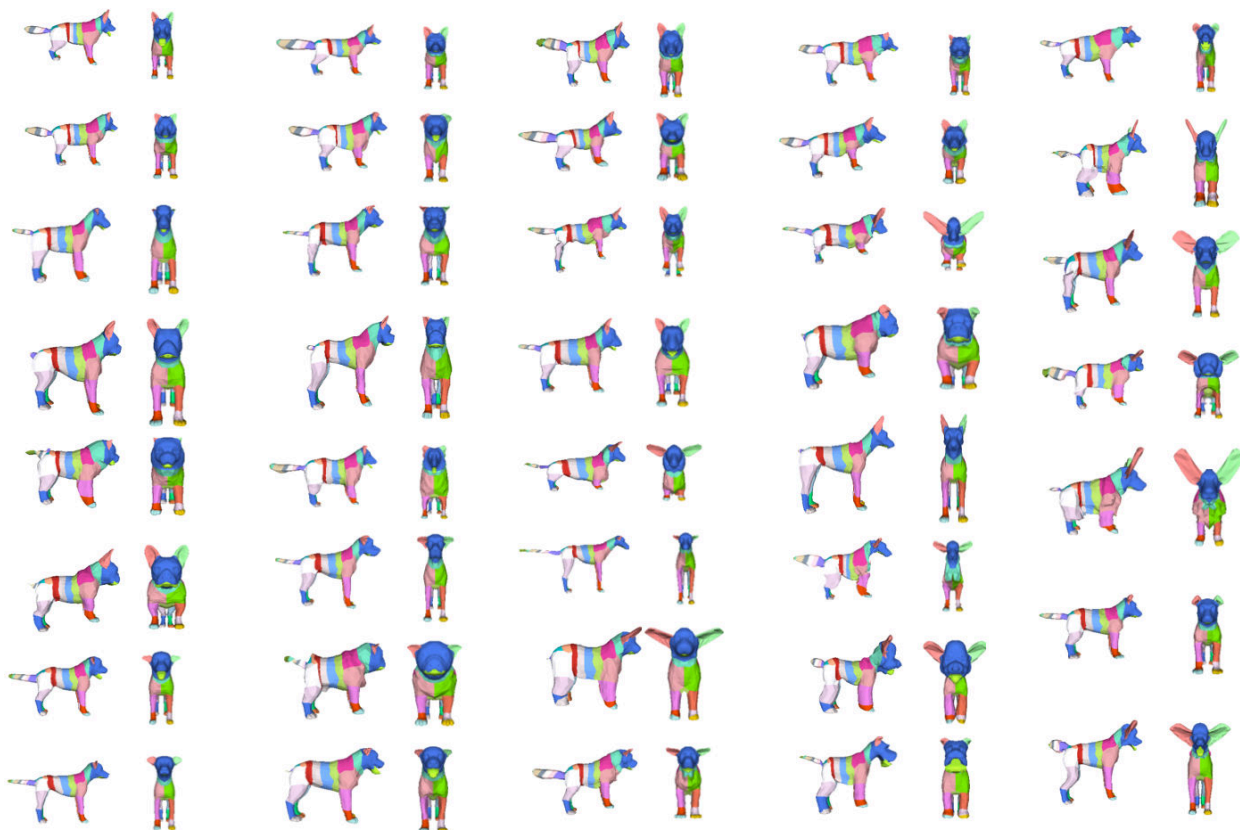


Figure S.6. Result of pose-normalization.
All toys are in a reference pose.



Figure S.7. **3D scan of a toy representing a Shih-Tzu.** This example was not included in the D-SMAL training, due to the difficulty of defining correspondences on the legs.

6.2. Comparison to Non-Parametric Methods

In recent times, novel non-parametric approaches, such as BANMo and ViSER, have emerged. Unlike BITE, which functions on single images, these methods require multiple images of the same dog (usually in the form of video data) to establish a non-parametric model of the dog's 3D shape. To give the reader an impression of the quality of BITE-ttopt predictions in comparison to the non-parametric models, we showcase several outcomes for both BITE-ttopt and BANMo in Figure S.8. For the BANMo results, the reconstruction of each dog has been optimized over 11 videos. BITE-ttopt, on the other hand, produces a faithful 3D shape

and pose from a single image and does not require access to an extensive set of videos of the subject.

6.3. Random Selection of Results

Figures S.9, S.10, S.11, S.12, S.13, S.14, S.15, S.16, S.17, S.18, S.19 and S.20 show a random selection of Stanford Extra test images with results from WLDO (who left the dogs out [1]), CTF (coarse-to-fine animal pose and shape estimation [2]), BARC [3], BITE and BITE-ttopt.

References

- [1] 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. 1, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
- [2] Chen Li and Gim Hee Lee. Coarse-to-fine animal pose and shape estimation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
- [3] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *Computer*

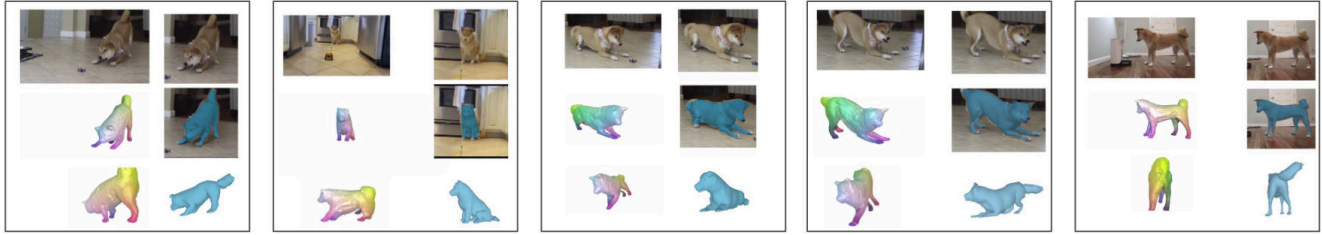


Figure S.8. *BANMo* and *BITE-ttopt* results. Visualized are five example frames from *BANMo* videos, each with *BANMo* results in the left column and *BITE-ttopt* results in the right column. From top to bottom, we show the input image, the predicted mesh from the same viewpoint, and a rotated view of it

Vision and Pattern Recognition Conference (CVPR), 2022. [1](#), [5](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#)

- [4] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019. [1](#)
- [5] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2017. [2](#)



Figure S.9. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.



Figure S.10. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.



Figure S.11. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.



Figure S.12. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.



Figure S.13. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.



Figure S.14. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.

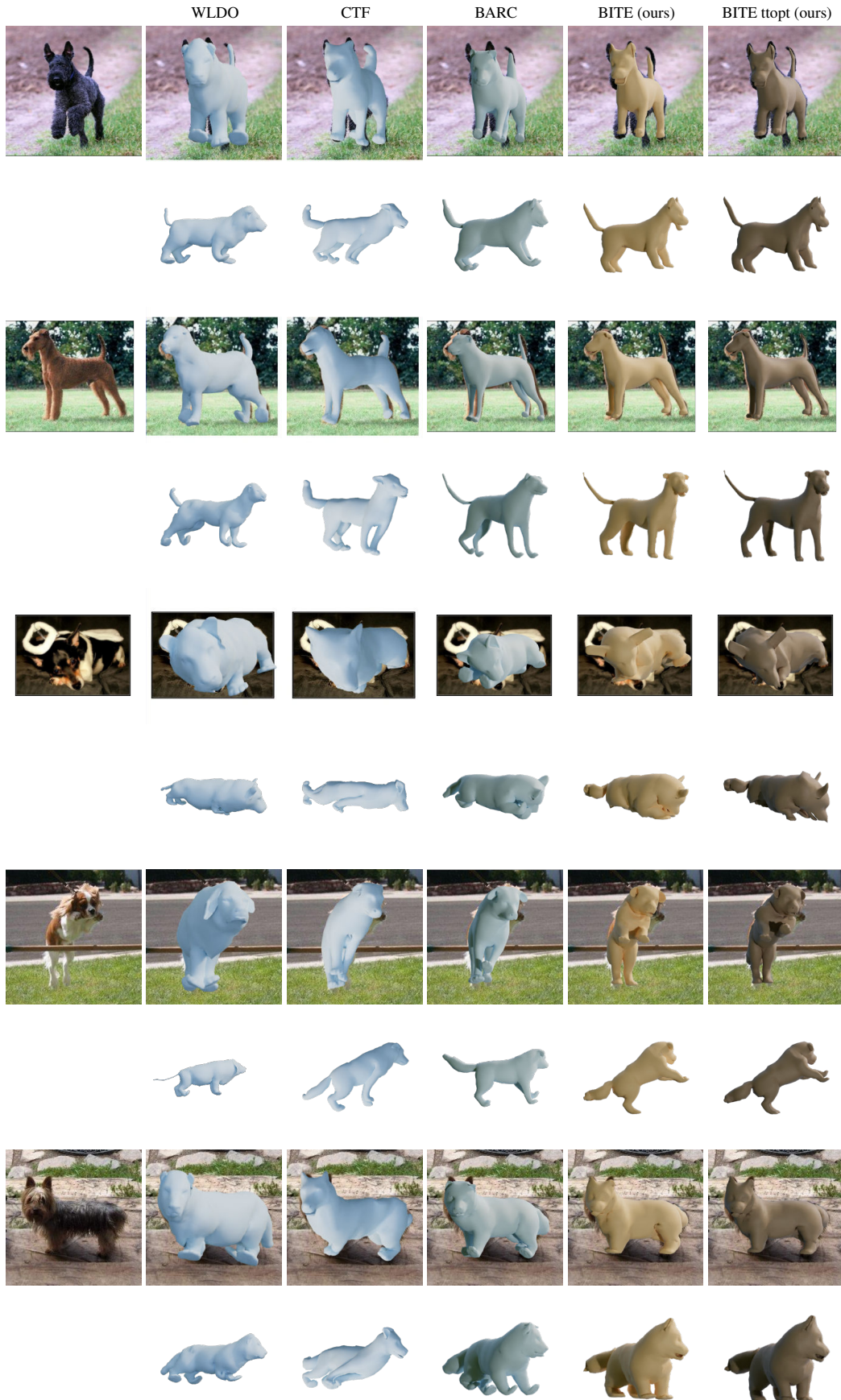


Figure S.15. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.



Figure S.16. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.



Figure S.17. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.



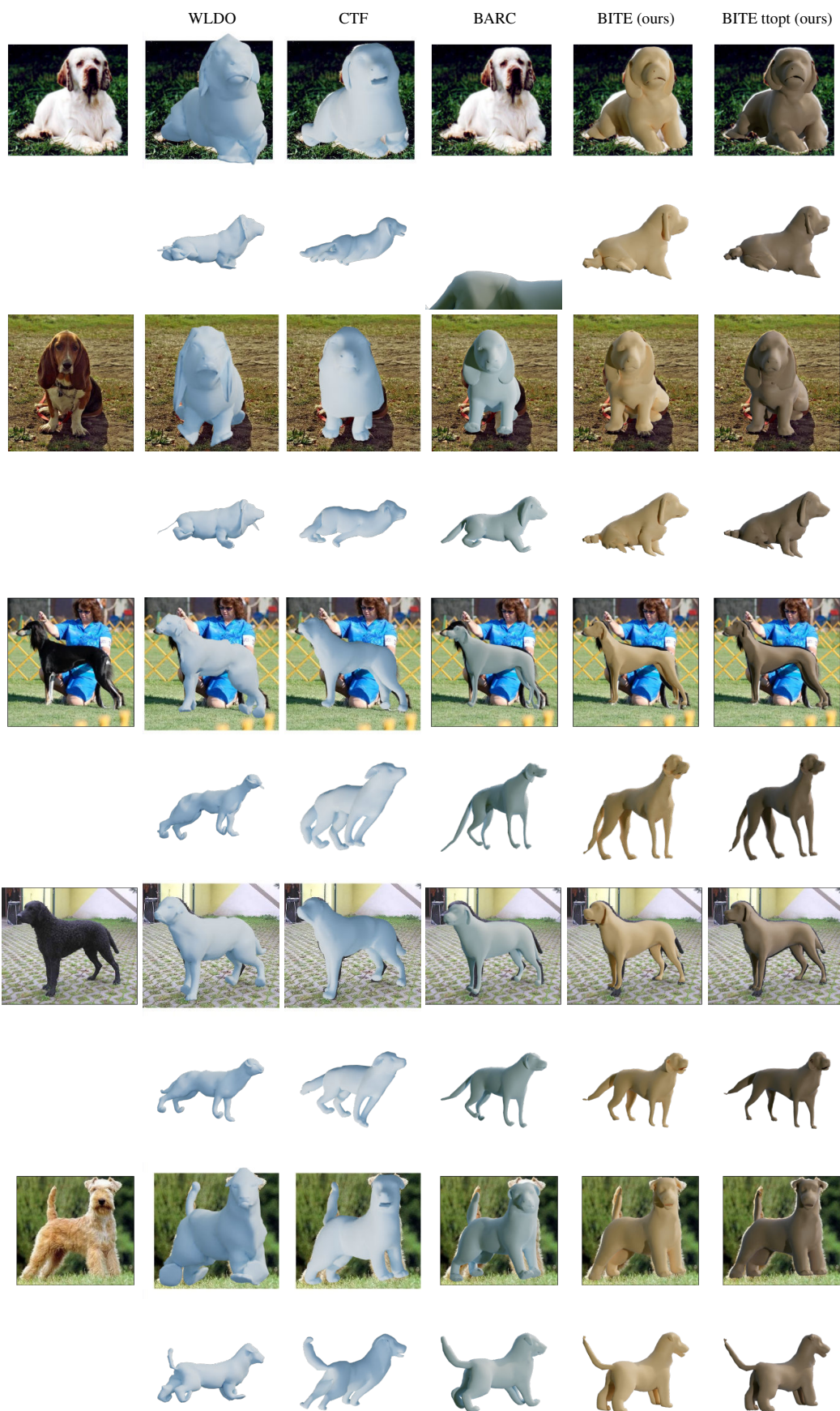


Figure S.19. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.



Figure S.20. *Qualitative results for random test images.* From left to right: input image, WLDO [1], CTF [2], BARC [3], BITE and BITE-ttopt.