Learning Articulated Shape with Keypoint Pseudo-labels from Web Images

Anastasis Stathopoulos
Rutgers University

Georgios Pavlakos
UC Berkeley

Ligong Han
Rutgers University

Dimitris Metaxas
Rutgers University

In this Supplementary Material we provide additional details that were not included in the main manuscript due to space constraints. In Section A we provide additional implementation details for our experiments. In Section B we offer additional evaluation on hard viewpoints (front & back view). Finally, we present qualitative results including results on random test samples and failure cases in Section C.

A. Implementation details

Pose estimation networks. In all of our experiments, we use *SimpleBaselines* [8] with an ImageNet pretrained ResNet-18 [3] backbone as the primary 2D pose estimation network h_ϕ. For criterion CF-CM we train Stacked HourGlass [7] with 8 stacks for the auxiliary 2D keypoint detector g_ψ. Both networks use input images of size 256 × 256 and predict K heatmaps of size 64 × 64. Random rotations (+/- 30 degrees) and scaling (0.75-1.25) are used as data augmentation. We train both models using Adam [5] optimizer with learning rate 1×10^{-4} for 50K iterations with mini-batches of size 32.

CMR. For experiments in Section 4.1 of the main manuscript, we train CMR [4] with the same hyperparameters as in [4]. We train using Adam [5] optimizer with learning rate 1×10^{-4} for 100K iterations with mini-batches of size 32. In a preprocessing step, CMR uses SfM on keypoints to initialize the template shape T and acquire a camera estimate for each training instance. We use only the keypoints from $\mathcal{S}$ to initialize T. During bundle adjustment, we use the confidence estimate of each keypoint to weight its contribution to the total reprojection error.

ACSM. For experiments in Section 4.2 of the main manuscript, we train ACSM [6]. We only train the network predicting camera poses and articulations from ACSM. We train for 70K iterations with mini-batches of size 12 using Adam [5] optimizer with learning rate 1×10^{-4}. We denote training ACSM in the above way as ACSM-ours.

B. Evaluation on hard viewpoints

In Table 4, we show results from evaluation on hard viewpoints (front & back view) in Pascal [2]. Those viewpoints are hard for the following reasons: 1) they are not frequent in the training sets (most animals are shown from side views), 2) the instances in those views contain a high degree of self-occlusion. Similar analysis is not applicable to Animal Pose [1] since most instances in that dataset are shown from side views. Results from Table 4 suggest that using keypoint pseudo-labels does not only improve 3D reconstruction performance in the mean case. The results are consistent with those in Table 1 & 2 of the main document, suggesting that consistency-based methods are more effective in our setting.

		Horse		Cow		Sheep	
		AUC (↑)	err$_R$ (↓)	AUC (↑)	err$_R$ (↓)	AUC (↑)	err$_R$ (↓)
	ACSM (Mask) [6]	26.7	49.1	19.3	102.2	14.2	94.7
	ACSM (KP+Mask) [6]	24.3	106.4	-	-	-	-
	ACSM-ours	45.0	35.9	41.7	38.5	42.3	38.5
	ACSM-ours + KP-all	46.3	39.4	39.0	42.0	39.9	49.8
$N = 1K$	ACSM-ours + KP-conf	47.2	37.5	41.6	40.6	44.1	45.1
	ACSM-ours + CF-MT	47.7	37.0	43.2	44.4	44.5	36.3
	ACSM-ours + CF-CM	47.6	36.6	43.8	42.7	**46.3**	37.1
	ACSM-ours + CF-CM^2	47.5	36.6	40.3	**36.8**	41.9	**32.6**
$N = 3K$	ACSM-ours + KP-conf	46.2	36.4	43.6	42.0	42.4	41.6
	ACSM-ours + CF-MT	**48.3**	35.3	45.7	44.8	44.7	40.5
	ACSM-ours + CF-CM	47.8	36.1	41.8	41.8	44.2	36.5
	ACSM-ours + CF-CM^2	47.7	**33.5**	**46.7**	39.6	45.5	33.5

Table 4. **Evaluation on hard viewpoints in Pascal.** We report the AUC and camera rotation error err$_R$ (in degrees) averaged over images with hard viewpoints (front & back view). N is the number of selected images from the web.

C. Additional qualitative results

Birds with CMR. In Figures 7, 8 & 9 we compare the predictions of CMR trained with: i) 300 labeled images; ii) the same labeled images and additional keypoint pseudo-labels, using random test samples from CUB. For each input image, the first 2 columns show the predicted shape and texture from the inferred camera viewpoint, while the last 2 columns are novel viewpoints of the textured mesh. We observe that the use of keypoint pseudo-labels during training significantly improve the 3D reconstruction quality. Training with pseudo-labels enables the model to capture some deformations that the fully-supervised model misses (*e.g.* open wings in the first row of Figure 9). These results clearly indicate the merit of using keypoint pseudo-labels with CMR. Finally, we visualize some failure cases of the proposed method in Figure 15. In those cases even the CMR model supervised with all the 6K images from the training set struggles.

Quadrupeds with ACSM. In Figures 10 & 11 we show qualitative comparisons between all the methods used for predicting the 3D shape of quadrupeds. For each input image, we show the predicted 3D mesh from the inferred camera view (first row) and a novel view (second row). We show results with $N = 3K$ samples from $\mathcal{U}$ for KP-conf, CF-MT, CF-CM and CF-CM^2. From Figures 10 & 11 we observe that the quality of the predicted 3D shapes is consistent with the quantitative evaluation conducted in the main maniscript (Tables 1, 2, 3). First, we observe that ACSM-ours (trained only with 150 images with keypoint-labels) achieves more accurate reconstructions than ACSM for all categories. A failure mode of ACSM is erroneous camera pose prediction. With ACSM-ours camera poses are improved, but some articulations are not well captured by the model. Using supervision from all web images (KP-all) increases errors in camera poses and results in unnatural articulations (see the prediction for sheep in Figure 10). KP-conf improves the quality of the predicted shapes compared to KP-all, but still results in unnatural articulations in some cases (see the second giraffe's neck at Figure 11). Finally, consistency-based filtering can lead in more accurate camera pose and articulation prediction than other alternatives. For instance, the articulation of the last bear in Figure 11 is only captured by CF-CM and CF-CM^2.

In Figures 12 & 13 we visualize the recovered 3D shape for models trained with data selected using consistency-filtering criteria. For each input image, we show the recovered 3D shape from the predicted and a novel view. As also captured by the quantitative evaluation in our main paper, we observe that the quality of the recovered 3D shape is in some cases higher for CF-CM^2 compared to other alternatives.

In Figure 14, we visualize some failure cases. We also include the prediction of the original ACSM model. Common failure modes include the inability to capture certain articulations (top row). These articulations are impossible to be captured by ACSM since it models articulations as rigid transformations of some pre-defined parts. Another failure mode is erroneous camera pose prediction for hard viewpoints (second row). ACSM fails worse in those cases (see unnatural head prediction in top row of Figure 14).

D. Sample web images from Flickr

In Figure 16 we show 8 random samples per object category from the unlabeled web images. Most images are not suitable from training 3D shape prediction models and should be filtered out.

References

[1] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *CVPR*, pages 9498–9507, 2019. 1

[2] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016. 1

[4] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, pages 371–386, 2018. 1

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[6] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, pages 452–461, 2020. 1

[7] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016. 1

[8] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018. 1

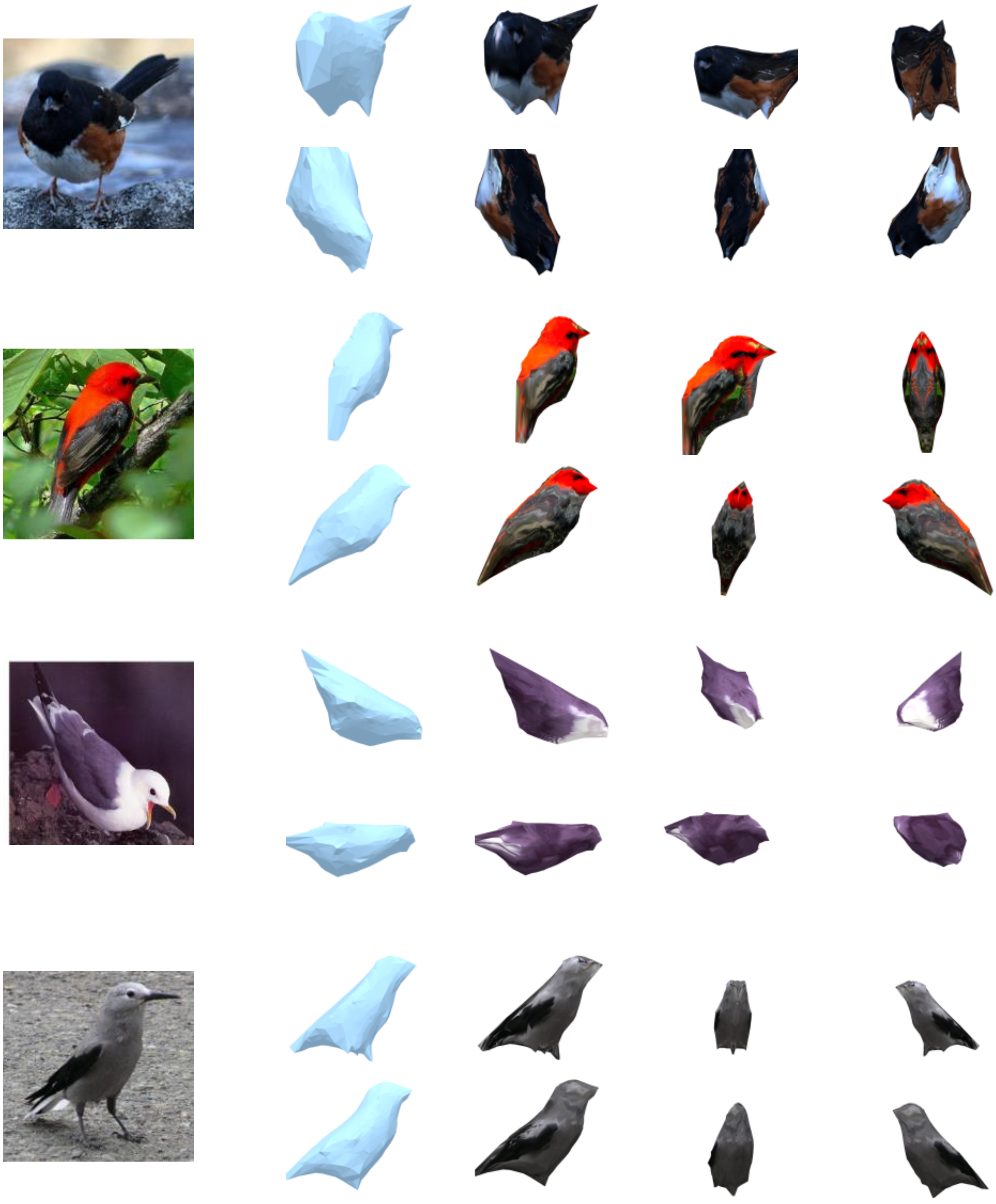

Figure 7. **Qualitative results on random samples CUB.** For each sample, we compare CMR trained with (first row) and without (second row) keypoint pseudo-labels. The first 2 columns show the predicted shape and texture from the inferred camera viewpoint. The last 2 columns are novel viewpoints of the textured mesh.

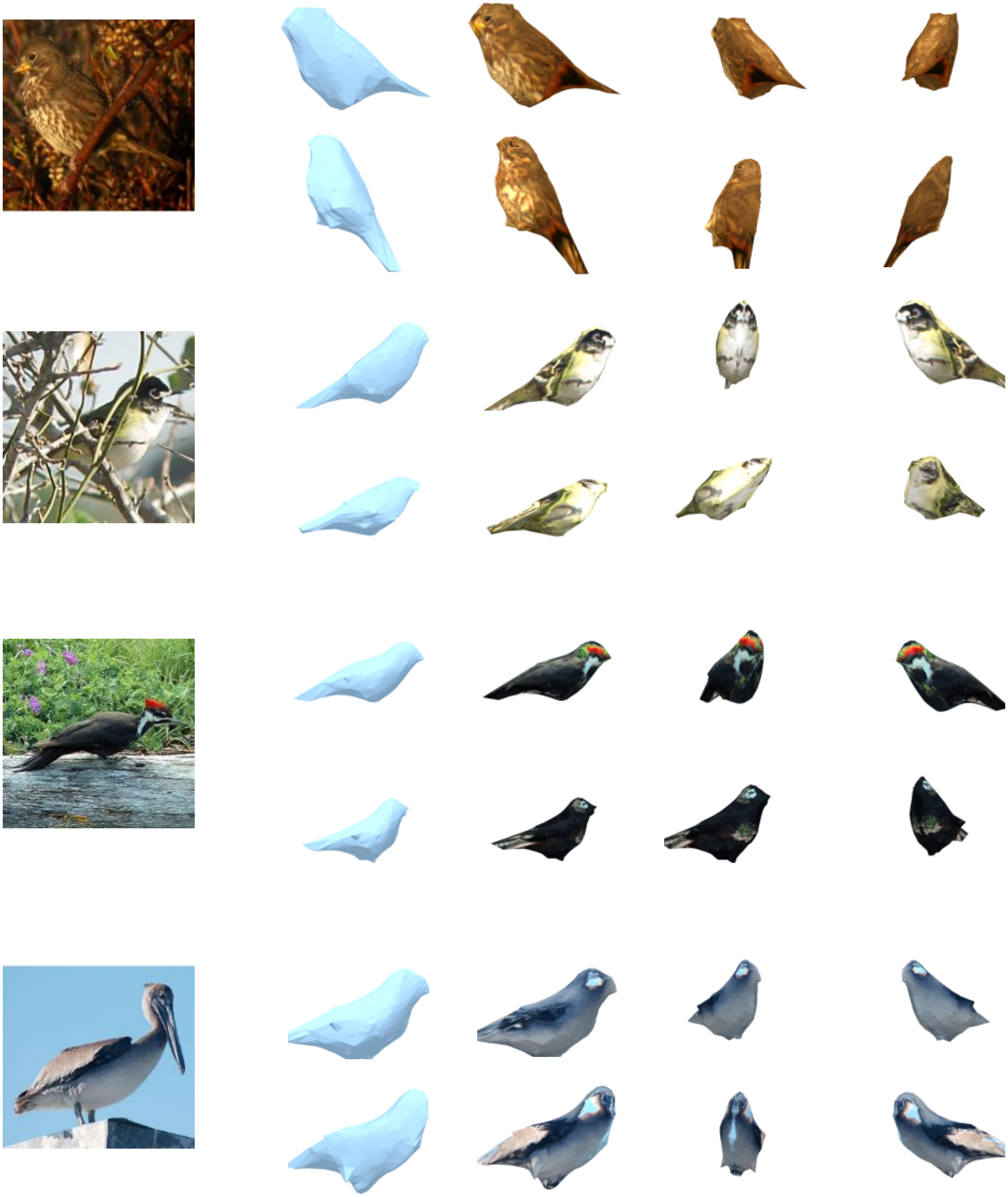

Figure 8. **Qualitative results on random samples from CUB.** For each sample, we compare CMR trained with (first row) and without (second row) keypoint pseudo-labels. The first 2 columns show the predicted shape and texture from the inferred camera viewpoint. The last 2 columns are novel viewpoints of the textured mesh.

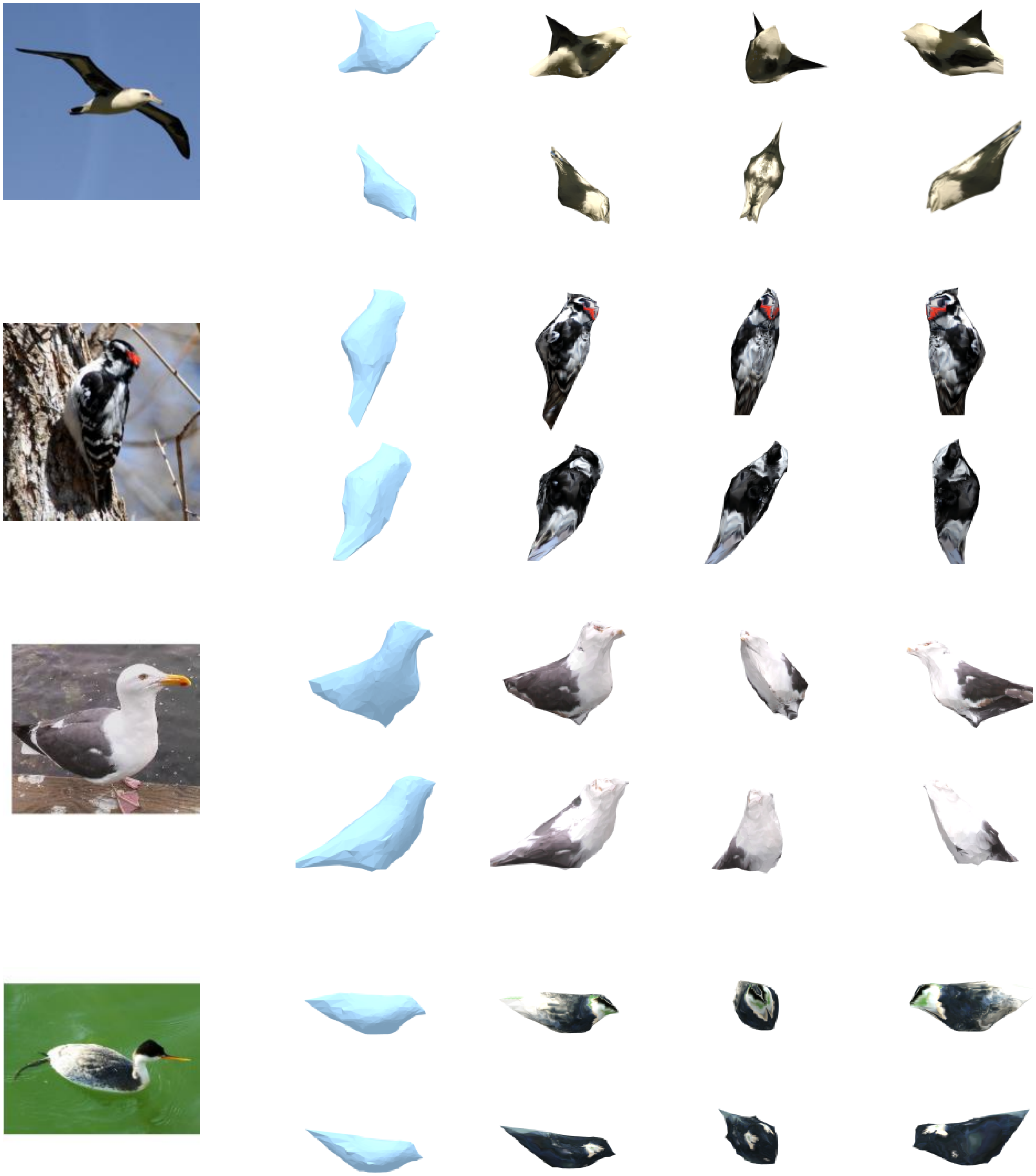

Figure 9. **Qualitative results on random samples from CUB.** For each sample, we compare CMR trained with (first row) and without (second row) keypoint pseudo-labels. The first 2 columns show the predicted shape and texture from the inferred camera viewpoint. The last 2 columns are novel viewpoints of the textured mesh.

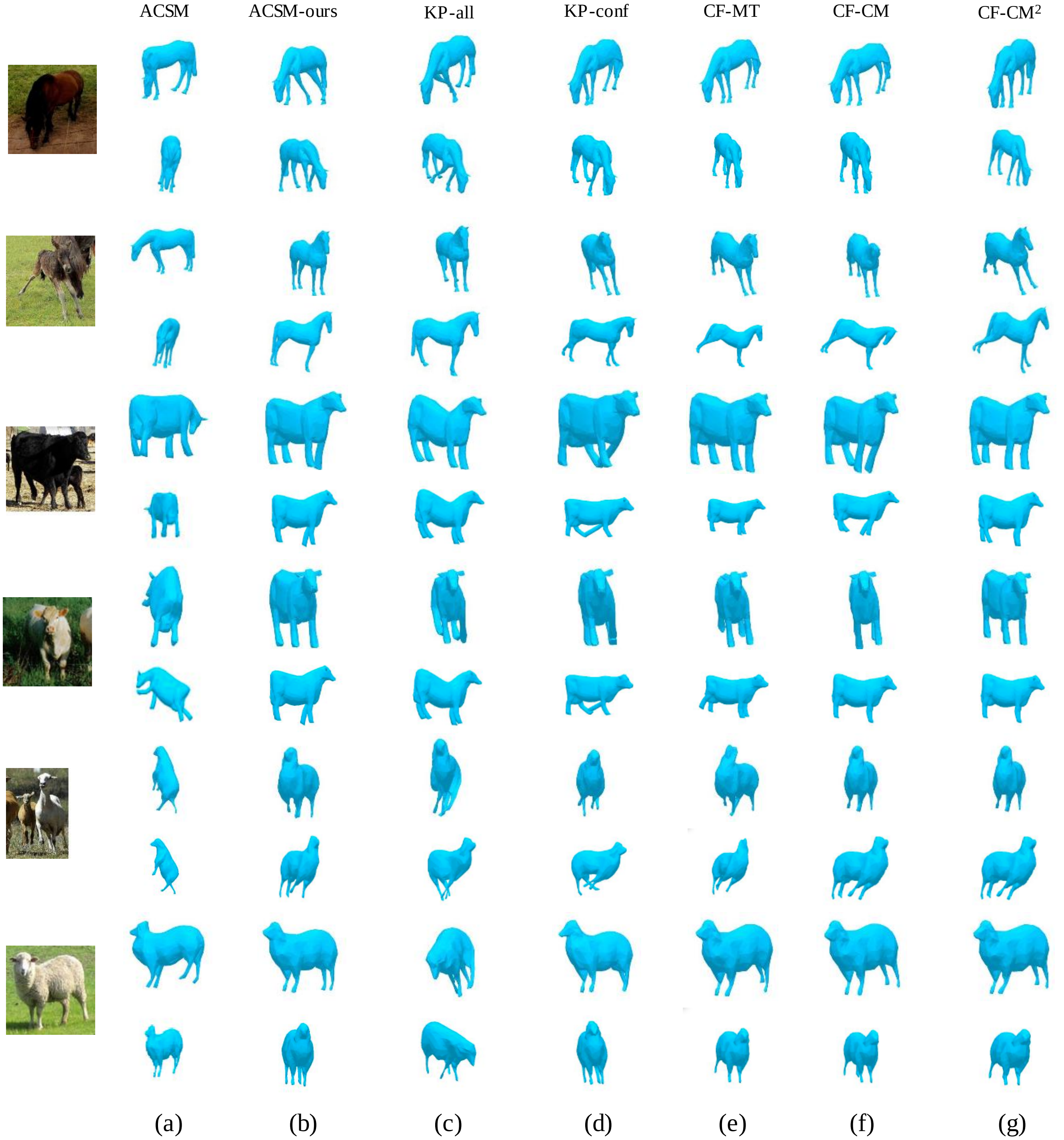

Figure 10. **Qualitative results for quadrupeds.** Qualitative comparisons between all methods with images from Pascal's test set. For each image, we show the articulated shape from the inferred camera viewpoint (top row) and a side view (bottom row).

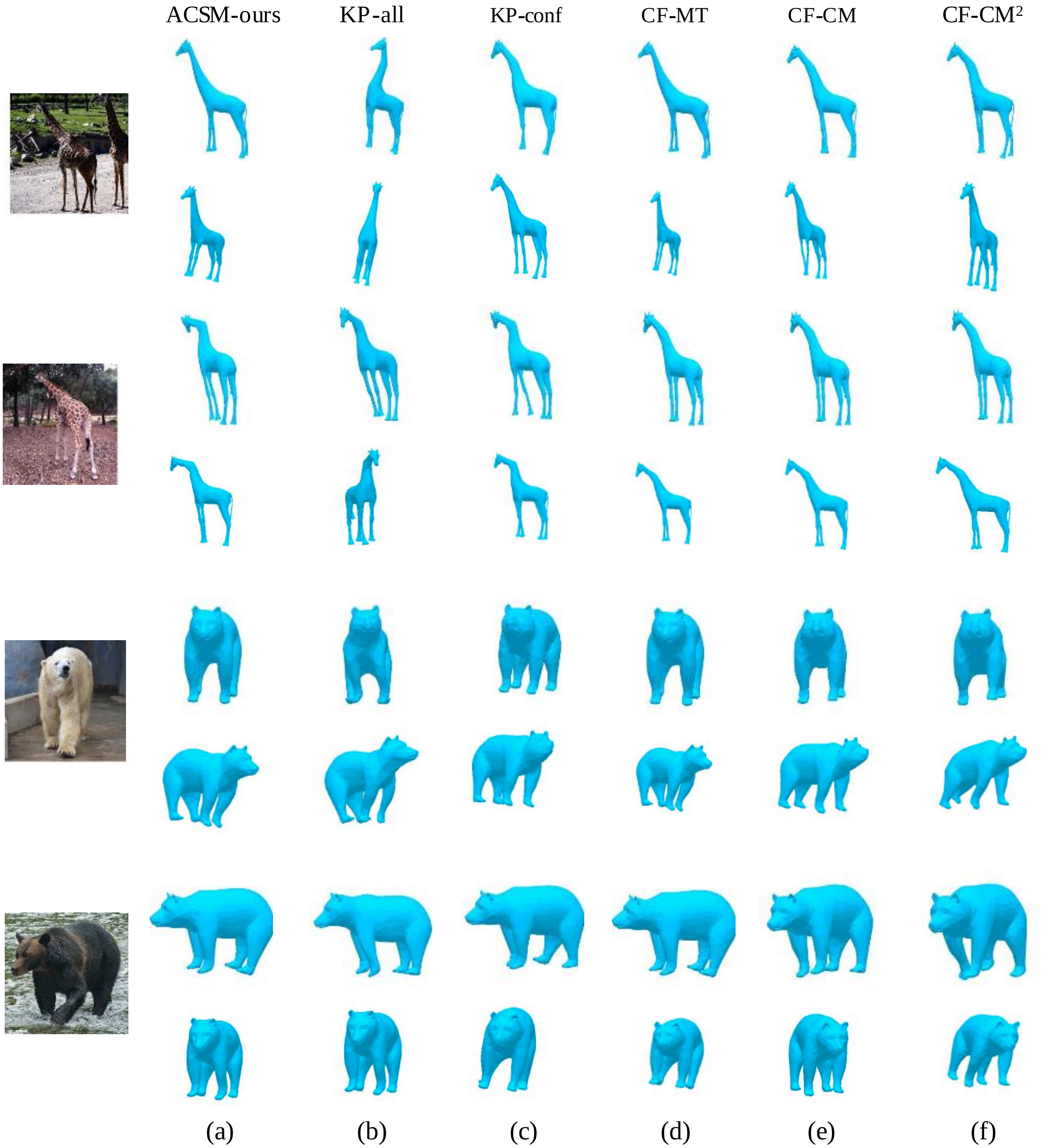

Figure 11. **Qualitative results for quadrupeds.** Qualitative comparisons between all methods with images from COCO. For each image, we show the articulated shape from the inferred camera viewpoint (top row) and a side view (bottom row).

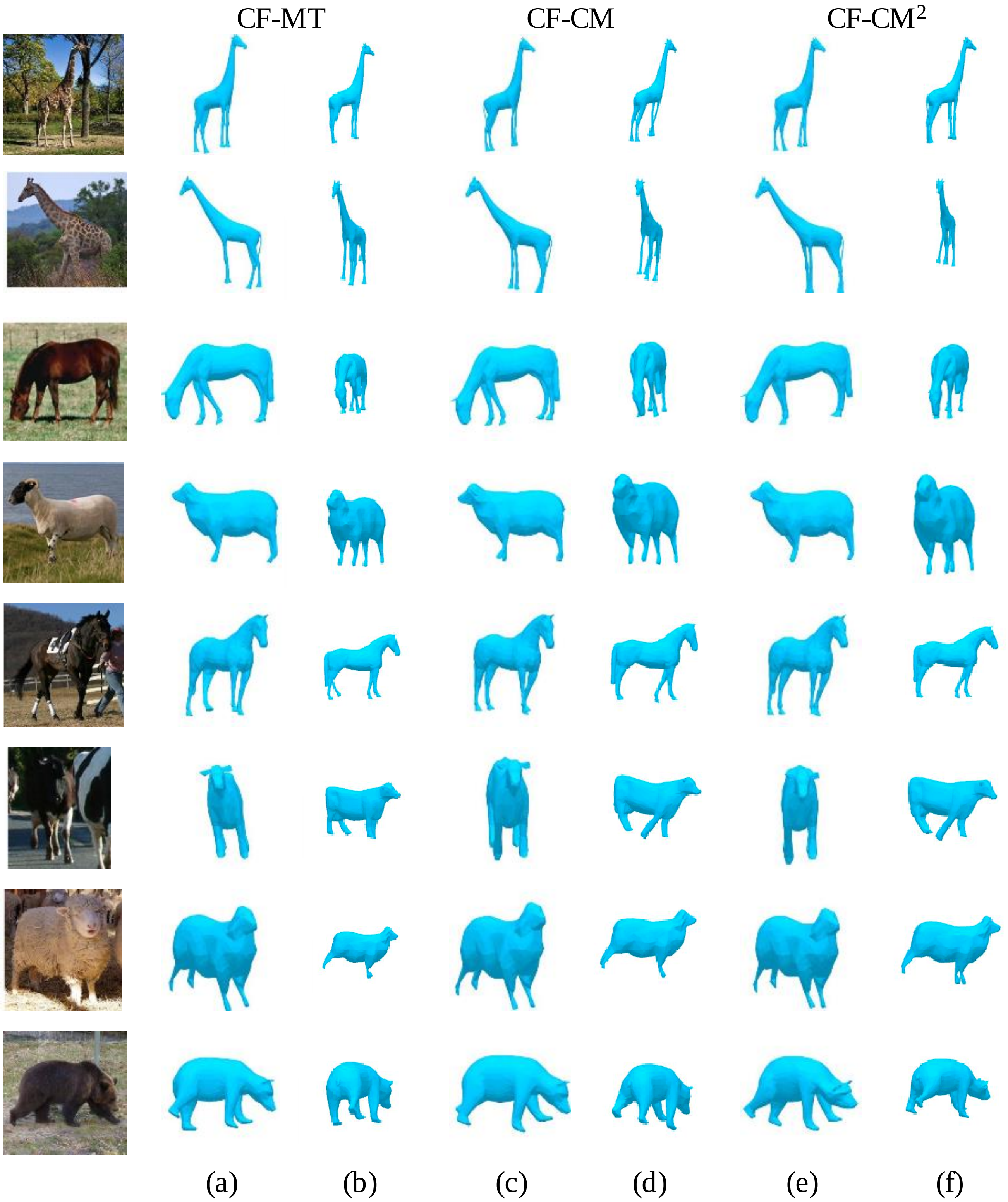

Figure 12. **Qualitative results on random samples for quadrupeds.** We visualize the recovered 3D shape from the predicted (a, c, e) and a novel view (b, d, f) for models trained with data selected from consistency-based criteria.

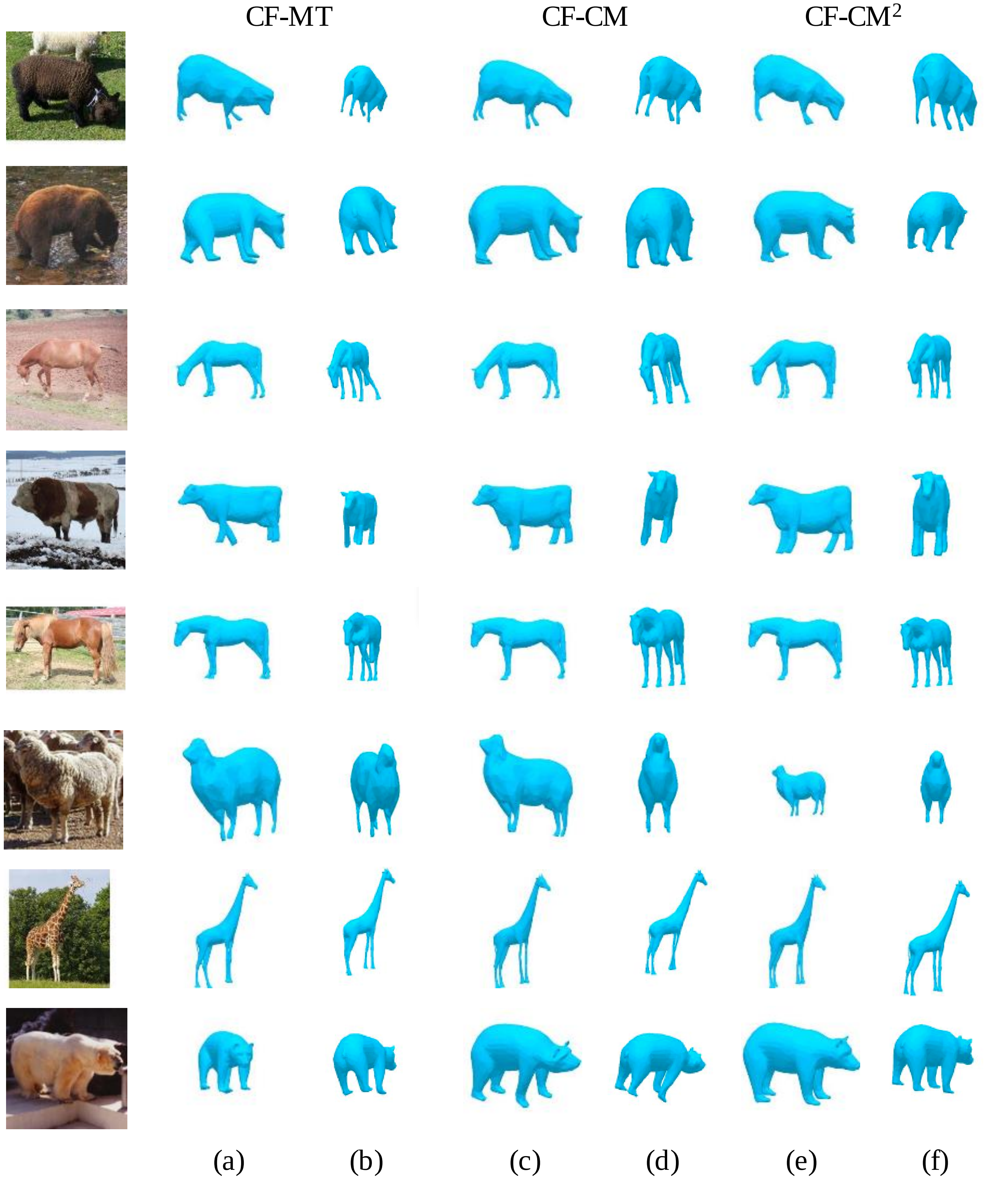

Figure 13. **Qualitative results on random samples for quadrupeds.** We visualize the recovered 3D shape from the predicted (a, c, e) and a novel view (b, d, f) for models trained with data selected from consistency-based criteria.

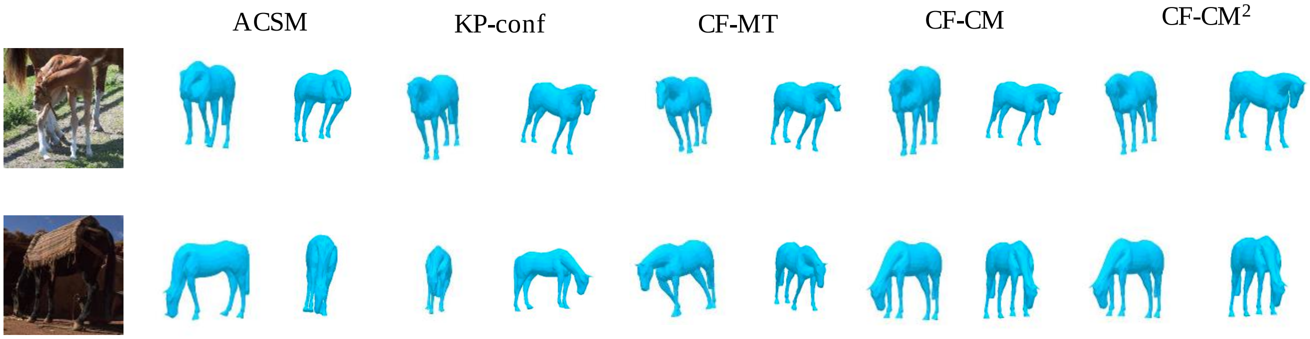

Figure 14. **Failure cases for quadrupeds.** For each sample, we show the predictions from the inferred (left column) and a novel (right column) view. Common failure modes include errors in articulations (top row) that the ACSM model is not possible to capture by design, and erroneous camera pose prediction for hard viewpoints.

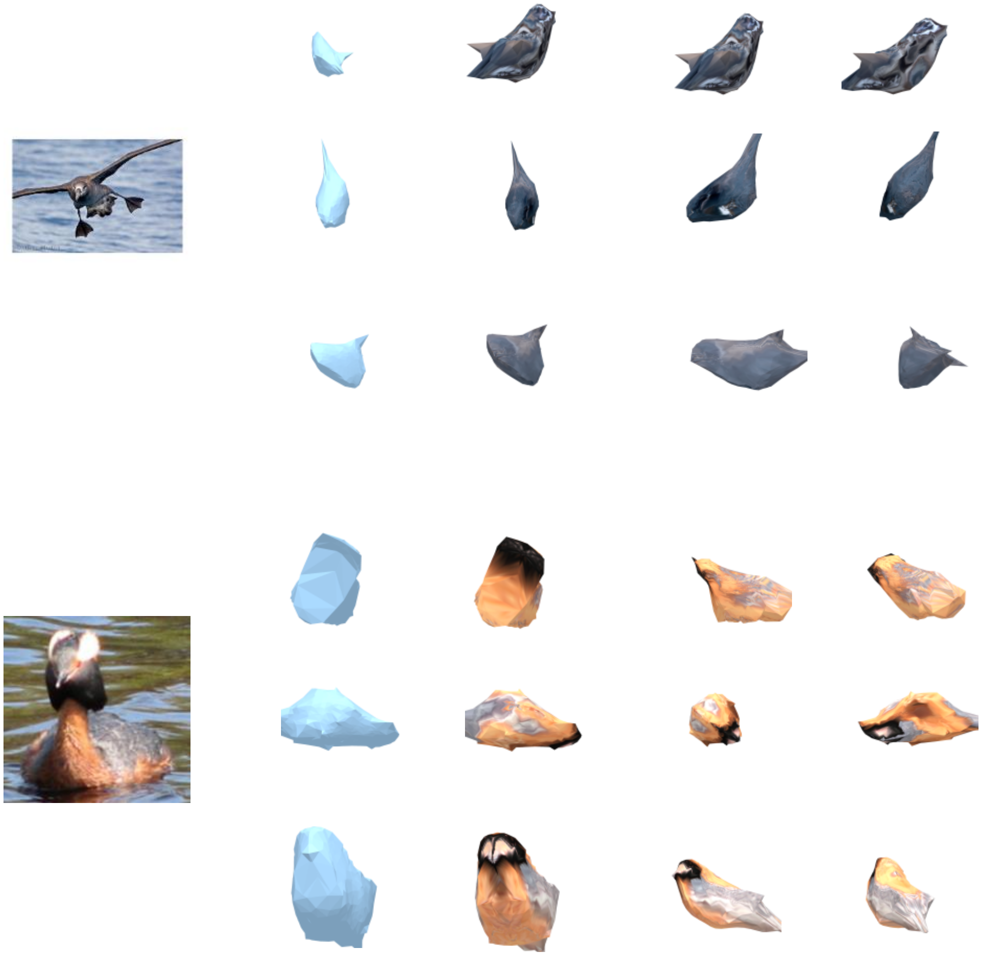

Figure 15. **Failure cases for birds.** We compare CMR trained with (first row) and without (second row) keypoint pseudo-labels. For a reference, we also show the fully-supervised model trained with 6K mask and keypoints annotations (third row). The first 2 columns show the predicted shape and texture from the inferred camera viewpoint. The last 2 columns are novel viewpoints of the textured mesh. We can see that even the fully-supervised model struggels in those cases.

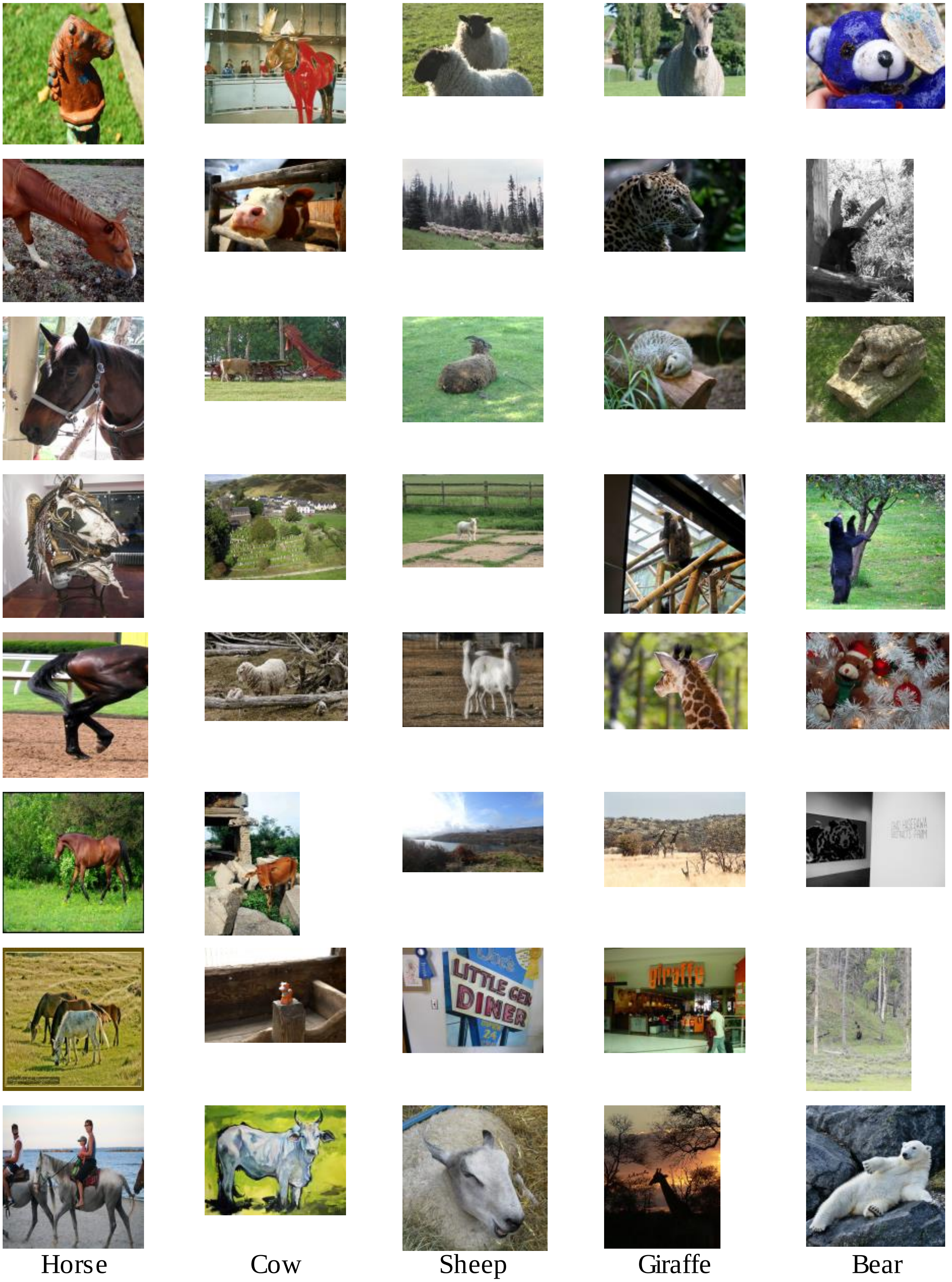

Figure 16. We randomly sample 8 images per object category from the unlabeled web images to stress the necessity for an effective data selection mechanism in our setting.