A. Supporting figures and experiments

A.1. AP based Evaluation: Selected sigma values

The correctly selected metric is crucial for model comparison and results evaluation. The commonly used metric for 2D pose estimation is AP based on OKS similarity. The metric requires a per-keypoint standard deviation value that controls the required precision of a given keypoint. Based on our observations, adjusted sigma values allow for a better evaluation of the 2D pose estimation models. We provide the comparison of proposed and original (Animal-Pose) values in Table 1.

Name	σ_{AP}	σ_{ours}
left-eye	0.025	0.018
right-eye	0.025	0.018
left-earbase	0.026	0.025
right-earbase	0.035	0.025
nose	0.035	0.020
throat	0.100	0.040
tailbase	0.100	0.040
withers	0.100	0.040
L-F-elbow	0.107	0.043
R-F-elbow	0.107	0.043
L-B-knee	0.107	0.043
R-B-knee	0.107	0.043
L-F-wrist	0.087	0.030
R-F-wrist	0.087	0.030
L-B-ankle	0.087	0.030
R-B-ankle	0.087	0.030
L-F-paw	0.089	0.032
R-F-paw	0.089	0.032
L-B-paw	0.089	0.032
R-B-paw	0.089	0.032

Table 1. **Comparison of proposed and Animal-Pose standard deviations (σ).** In column σ_{AP} are values proposed for the Animal-Pose dataset by MMPose. In column σ_{ours} are ours adjusted values.

A.2. Fine-tuning with synthetic data

We measured the performance of HRNet-w32 trained from the ImageNet checkpoint on a combination of the Animal-Pose dataset and synthetic data. We trained the models with different numbers of synthetic data. Furthermore, we measured the effect of augmentations and hyperparameters on performance. Initially, we trained models for 210 epochs with a learning rate $5e^{-4}$. Based on the sweep of the hyperparameters we adjusted the number of epochs to 100 and the learning rate to $1e^{-4}$. We provide a details comparison of models trained with augmentations and adjusted hyperparameters in Table 2.

Augment.	H. Param.	Synthetic	AP	$AP_{0.75}$	$AP_{0.85}$	$PCK_{0.05}$
		0	0.425	0.347	0.008	0.707
		10	0.440	0.401	0.019	0.715
		100	0.465	0.459	0.035	0.726
		1000	0.464	0.453	0.019	0.728
		10000	0.475	0.488	0.044	0.737
✓		0	0.465	0.427	0.017	0.728
✓		10	0.451	0.400	0.018	0.725
✓		100	0.509	0.541	0.057	0.752
✓		1000	0.516	0.554	0.102	0.764
✓		10000	0.494	0.514	0.052	0.754
	✓	0	0.382	0.244	0.012	0.688
	✓	10	0.425	0.338	0.013	0.706
	✓	100	0.449	0.401	0.022	0.721
	✓	1000	0.486	0.504	0.052	0.740
	✓	10000	0.499	0.524	0.048	0.758
✓	✓	0	0.426	0.312	0.015	0.712
✓	✓	10	0.441	0.367	0.016	0.717
✓	✓	100	0.502	0.514	0.055	0.751
✓	✓	1000	0.523	0.569	0.077	0.763
✓	✓	10000	0.469	0.444	0.034	0.746

Table 2. **Fine-tuning with synthetic data.** Detailed results of models trained from ImageNet checkpoint on a combination of Animal-Pose and synthetic data. Column **H. Param.** indicates whether adjusted hyperparameters were used.

A.3. Synthetic data generation method

Scenes: For the generation of synthetic data, we created 4 distinct 3D scenes, each scene in 5 variants. Different variants are introduced to increase the diversity of generated data. All the variants of one scene have the same composition but differ in used assets (i.e., trees, grass, snow, rocks). Figure 1 shows all scenes and their variants.

Augmentations: We did not simulate different lighting conditions or apply any post-processing transformation to the synthetic images during generation. Instead, we evaluated the performance of the models with different augmentations. We provide an example of different augmentations in Figure 2.

A.4. Qualitative evaluation

We provide qualitative results of several models in Figure 3. The zero-shot model and model trained only on synthetic data predict often incorrectly leg keypoints and tale base keypoint. Predictions of models trained on real data are close to the ground truth labels.

Figure 1. **Four proposed environments.** Each scene is created in five variants to create more dynamic environments. The composition of the scene stays the same but assets (i.e., trees, rocks, grass, snow, logs) change.

Figure 2. **Examples of augmentations.** Each column of images shows one of the 5 used augmentations. Starting from the left, the augmentations used are noise, Gaussian blur, brightness and contrast adjustment, HSV shift and conversion to grayscale.

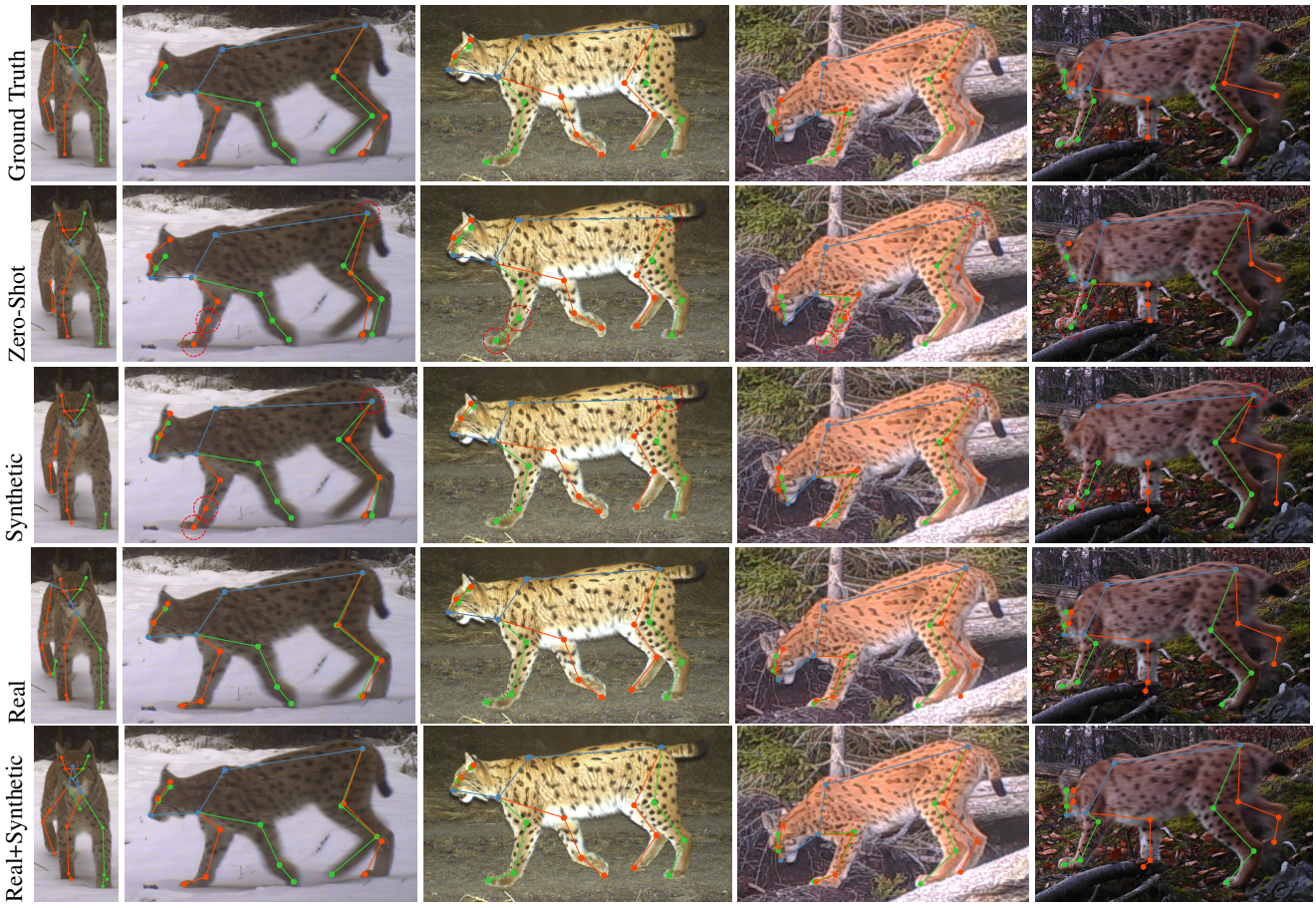

Figure 3. **Qualitative comparison of predictions.** The zero-shot model consistently predicted incorrectly the tale base, the tip of the paws, and the front wrists (marked in the images with a red circle). The model trained on synthetic data also predicted incorrectly tale base however precision of leg keypoints improved. Both models trained with real data predict keypoints close to the ground truth labels. For better readability, we provide only the part of the image with the animal and not the full image.