AP-10K: A Benchmark for Animal Pose Estimation in the Wild

Hang Yu[1*], **Yufei Xu**[2*], **Jing Zhang**[2†], **Wei Zhao**[1†], **Ziyu Guan**[1], **Dacheng Tao**[3,2]

[1]Xidian University, China, [2]The University of Sydney, Australia, [3]JD Explore Academy, China

Abstract

Accurate animal pose estimation is an essential step towards understanding animal behavior, and can potentially benefit many downstream applications, such as wildlife conservation. Previous works of animal pose estimation only focus on specific animals while ignoring the diversity of animal species, limiting their generalization ability. In this paper, we propose AP-10K, the first large-scale benchmark for mammal animal pose estimation, to facilitate research in animal pose estimation. AP-10K consists of 10,015 images collected and filtered from 23 animal families and 54 species following the taxonomic rank and high-quality keypoint annotations labeled and checked manually. Based on AP-10K, we benchmark representative pose estimation models on the following three tracks: (1) supervised learning for animal pose estimation, (2) cross-domain transfer learning from human pose estimation to animal pose estimation, and (3) intra- and inter-family domain generalization for unseen animals. The experimental results provide sound empirical evidence on the superiority of learning from diverse animals species in terms of both accuracy and generalization ability. It opens new directions for facilitating future research in animal pose estimation. AP-10k is publicly available at https://github.com/AlexTheBad/AP10K[3].

1 Introduction

Pose estimation, referring to identifying the category and location of a series of body keypoints, is a fundamental computer vision task, which is also useful in many practical applications, such as activity recognition[37, 5], behavior understanding [4], human-object interaction[46, 34]. While many research efforts have been made for human pose estimation [45, 42] in both large-scale benchmarks as well as advanced algorithms, fewer research works are focusing on animal pose estimation. One of the main reasons is that it is very challenging to collect and annotate a large-scale benchmark dataset for animal pose estimation, like MPII dataset [2] and COCO dataset [31] for human pose estimation, especially considering the fact that there are many species of animals and it requires some domain knowledge to perform annotation. Nevertheless, animal pose estimation is of great significance in both research purposes and practical applications such as zoology and wildlife conservation [20, 15]. Therefore, it is necessary to pave the way for research in this field by establishing a large-scale benchmark covering many different animal species.

Previous works for animal pose estimation focus on specific animal species, *e.g.*, horse [33], zebra [18], macaque [28], fly [38], and tiger [30]. While they make contributions in advancing the research for animal pose estimation regarding both datasets and algorithms, they suffer from the

[*]Equal contribution. The work was done during the first authors' internship at JD Explore Academy.

[†]Corresponding author

[3]The code will also be integrated into mmpose.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

Figure 1: A glance at diverse animal species in AP-10K. Figures are best viewed in color.

poor generalization ability to unseen animal species, limiting their practical significance. Some other works [6, 18] try to mitigate this issue by taking several animal species into consideration and construct new datasets covering multiple animal species. However, the number of animal species is still very limited, *e.g*., only five species in [6]. In addition, these datasets are not organized following the taxonomic rank, and there is no animal family and species structure. Consequently, the following important research questions remain unclear, *i.e*., 1) how about the performance of different representative human pose models on the animal pose estimation task? 2) will the representation ability of a deep model benefit from training on a large-scale dataset with diverse species? 3) how about the impact of pretraining, *e.g*., on the ImageNet dataset [16] or human pose estimation dataset [31], in the context of the large-scale of dataset with diverse species? 4) how about the intra- and inter-family generalization ability of a model trained using data from specific species or family?

In this paper, we make an attempt to answer these questions by collecting the first large-scale benchmark AP-10K for general mammal pose estimation. It consists of 10,015 images collected and filtered from 23 animal families and 54 species following the taxonomic rank, where the keypoints of all animal instances on each image are manually labeled and carefully double-checked. Specifically, various animal images are collected from existing publicly available datasets, and a cleaning process is carried out to remove the replicated images. Then, they are organized following the taxonomic rank. Finally, thirteen annotators are recruited to carefully annotate the bounding boxes for all animal instances in each image and their body keypoints, as well as the background category of each image, following the COCO [31] annotation style. In addition to the labeled 10,015 images, AP-10K also contains about 50k animal images organized following the taxonomic rank but without keypoint annotations, which can be used for animal pose estimation at the settings of semi-supervised learning [14] and self-supervised learning [21, 8, 9].

The diversity in the animal species and the family-species organization of AP-10K provides a possibility to study the aforementioned research questions. To this end, we benchmark representative pose estimation models [10, 24, 44] on the following three tracks, *i.e*., (1) supervised learning for animal pose estimation (SL track), (2) cross-domain transfer learning between human pose estimation and animal pose estimation (CD-TL track), and (3) intra- and inter-family domain generalization for unseen animals (DG track). In the SL track, we study the representation ability of different models at the settings of random weight initialization or initialization with ImageNet pretrained weights. In the CD-TL track, we investigate whether pretraining on a human pose estimation dataset like COCO benefits the animal pose estimation on AP-10K. In the DG track, we first study the intra-family domain generalization, where the model is trained using some species belonging to a specific family and tested on other species from the same family. Then, we also study the inter-family domain generalization, where the model is trained using all species from a certain family and tested on species from a different family. The detailed experiment settings and results are presented in Section 4. In summary, the experimental results provide sound empirical evidence on the superiority of learning from diverse animals species in terms of both accuracy and generalization ability.

The main contribution of the paper is twofold. (1) We establish the first large-scale benchmark AP-10K for mammal animal pose estimation, which contains 10,015 images from 23 families and 54

species following the taxonomic rank, along with high-quality keypoint annotations. (2) Based on AP-10K, we study several challenging and open research questions in this area by benchmarking representative pose estimation models on the SL track, the CD-TL track, and the DG track. The results show sound empirical evidence on the superiority of learning from diverse animals species.

2 Related work

2.1 Human pose estimation

Human pose estimation is an active research area in computer vision. In general, human pose estimation methods can be categorized as bottom-up methods [35] and top-down ones [44, 41, 32]. While the former ones are usually fast, the latter ones always achieve the top entries in representative benchmarks due to their high accuracy. Well-known large-scale human pose estimation benchmarks are COCO [31] and MPII [2], while some new datasets have been established recently, regarding either crowd scenes like CrowdPose [29] or occlusion scenes like OCHuman [47]. These datasets have significantly advanced the research in this area. Models [44, 45, 42] trained using these datasets have been proven effective for human pose estimation, owing to the rich diversity of posture, illumination, scale, occlusion, and the number of person instances per image.

2.2 Animal pose estimation

Animal pose estimation has attracted increasing attention in the research community. Typical human pose estimation models can be applied to animals since there is no difference in modeling, *i.e*., except for the number of defined keypoints. It is the labeled dataset that matters in the era of deep learning. In the past few years, some datasets have been established to facilitate research in animal pose estimation [33, 18, 30]. However, most of them focus on specific animal species, *e.g*., horse [33], zebra [18], macaque [28], fly [38], and tiger [30], and are limited in diversity of postures, textures, and habitats. Recently, the Animal Pose Dataset is introduced in [6], which has 5 animal species. In addition to its limited number of species, it is not organized following the taxonomic rank, *i.e*., there is no structure of animal families and species, making it impossible to study the intra- and inter-family generalization problem and other ones mentioned above. In contrast to human pose datasets, which have only one species, the very familiar human, collecting and annotating animal keypoints is more challenging since there are many different species of animals belonging to different families, and specific biological knowledge is required to distinguish the keypoints of different animals. While challenging and costly, it is very necessary to establish a large-scale benchmark to facilitate research in this field.

2.3 Transfer learning

Transfer learning refers to transferring the learned knowledge from a source task to a target task, which is related to source one. Usually, the source domain has a large scale of labeled data while there is only a limited number of labeled data in the target domain. The prevalent transfer learning methods follow the "pretraining and finetuning" route. It has been proven effective in many computer vision tasks, including image classification [27], object detection [23], semantic segmentation [7], as well as human pose estimation [41]. While the ImageNet dataset is widely used as the source domain for transfer learning in different computer vision tasks, including animal pose estimation, some works also investigate the effectiveness of transferring knowledge from the human pose estimation task to the animal pose estimation task [6]. However, it is unexplored and unclear whether transferring from ImageNet or existing human pose datasets is still effective or not, when a large-scale animal pose dataset is available. A related research is [22], which shows that a longer training schedule on a large-scale dataset for a downstream task (*i.e*., object detection) can catch up the performance gain from pretraining on ImageNet. In this paper, we study this problem based on our proposed AP-10K dataset and show that pretraining on ImageNet is not necessary for animal pose estimation when a longer training schedule is used.

3 Dataset

3.1 Data collection and organization

Data collection

As we have discussed, the limited number of animal species in existing animal pose datasets makes it difficult to comprehensively evaluate the performance of animal pose estimation models, considering the diversity in real-world animal species. To facilitate research in this area, there is a need to collect a large-scale animal pose estimation dataset with many different animal species. To this end, we propose the AP-10K dataset. First, we resort to existing publicly available datasets focusing on animals [43, 17, 48, 40, 39, 1, 3, 25]. Although these datasets are designed for the animal classification task or animal detection task, they provide abundant animal images from different fine-grained species. Then, we collect them as a whole and remove the repeated images or mislabeled images, including a coarse stage by aHash [11] and a refinement stage by manual double-check. In this way, we obtain all the candidate animal images with high-quality species category labels for our AP-10K dataset, *i.e.*, 59,658 images in total.

Data organization We re-organize and re-label the images following the taxonomic rank, *i.e.*, family and species. Instead of using the exact biology definition of family-genus-species, we do not distinguish genus and species for simplicity. Such a classification takes advantage of a biological prior about evolution, where animals belonging to the same family/species have a similar texture, appearance, and pose distributions. Besides, when evaluating the model generalization to unseen animals, ignoring biological relationships among seen and unseen animals will lead to highly biased results. The hierarchical categorization in our AP-10K can provide a possibility to evaluate intra and inter-family model generalization in a fair setting.

Keypoint	Definition	Keypoint	Definition
1	Left Eye	10	Right Elbow
2	Right Eye	11	Right Front Paw
3	Nose	12	Left Hip
4	Neck	13	Left Knee
5	Root of Tail	14	Left Back Paw
6	Left Shoulder	15	Right Hip
7	Left Elbow	16	Right Knee
8	Left Front Paw	17	Right Back Paw
9	Right Shoulder		

Table 1: Definition of animal keypoints.

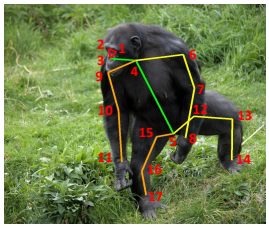

Figure 2: The keypoints of a Chimpanzee.

3.2 Data annotation

To obtain high-quality animal pose annotations for each image, we recruited 13 well-trained annotators and asked them to annotate all the keypoints of each animal they can distinguished. Three rounds of cross-checking and correction are then carried out to improve the annotation quality. Finally, it took about three months to complete the whole annotation process, where 10,015 images were labeled. Similar to the keypoints defined for representing human pose, 17 keypoints are defined to represent animal pose, including two eyes, one nose, one neck, one tail, two shoulders, two elbows, two knees, two hips, and four paws, as listed in Table 1. A visual example of the keypoint annotations of a Chimpanzee is shown in Figure 2. Besides, We also labeled 8 background type for all posed images, *i.e.*, grass or savanna, forest or shrub, mud or rock, snowfield, zoo or human habitation, swamp or riverside, desert or gobi and mugshot. The annotations are saved in line with the COCO format to facilitate further research in animal pose estimation by reusing common training and evaluation tools developed in the human pose estimation community. The AP-10K is split into three disjoint subsets, *i.e.*, train, validation, and test sets, at the ratio of 7:1:2 per animal species. It is also noteworthy that in addition to the 10,015 images with keypoint annotations, we also include the remaining 49,643 images with their family and species labels in our AP-10K to facilitate future research, *e.g.*, semi-supervised learning and self-supervised learning for animal pose estimation.

3.3 Statistics of the AP-10K dataset

Overview of AP-10K As shown in Table 2, AP-10K dataset covers 23 families and 54 species of animals, which is much richer in the diversity of animal species than previous datasets. Besides, AP-10K contains much more labeled images and instances than the other datasets, *i.e.*, 10,015 images

Table 2: Comparison of different animal pose datasets.

dataset	species	family	labeled image	unlabeled image	keypoint	instance
Animal-Pose Dataset [6]	5	N/A	4,666	0	20	6,117
Horses-10 [33]	1	N/A	8,110	0	22	8,110
ATRW [30]	1	N/A	8,076	0	15	9,496
AP-10K	**54**	**23**	**10,015**	**50k**	**17**	**13,028**

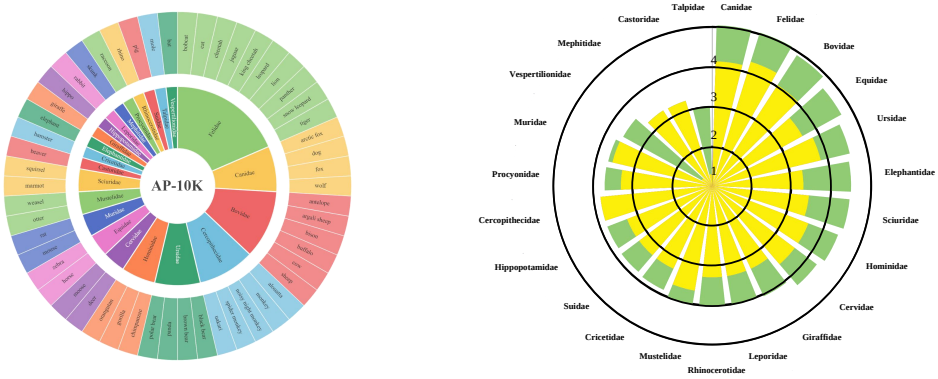

Figure 3: (a) The organization of family and species in AP-10K. (b) The distributions (in the logarithmic scale) of the number of labeled (yellow) and whole (green) images in each family.

and 13,028 instances in total, resulting in a more complex animal pose distribution. Besides, we also provide extra 50k images without keypoint annotations but family and species labels in our AP-10K. Thanks to the taxonomic rank-based organization of AP-10K, these unlabeled data can be exploited for further research, *e.g.*, semi-supervised learning and self-supervised learning for animal pose estimation, which is not applicable in other datasets.

Long-tail property As shown in Figure 3, the number of images in each family of AP-10K has a long-tail distribution, which reflects the true distribution of animals in the wild due to the commonness or rarity of the animals in some extent. For example, there are 1,913 labeled images (9,277 unlabeled images) for common families like Felidae and only 200 labeled images (312 unlabeled images) for rare families like Procyonidae. Similar properties can be found in CityScape [13] and LVIS [19] datasets. On the other hand, the number of species per family varies considerably. For example, the most common family in AP-10K, *i.e.*, Felidae, contains 10 more species than the rarest family, *i.e.*, Castoridae, which contains only 1 species. The property of long-tail distribution makes AP-10K be a challenging benchmark for mammal animal pose estimation. Specifically, it can also be used to study few-shot learning for animal pose estimation.

4 Experiment

4.1 Implementation details

We benchmark several representative pose estimation frameworks [36, 44, 41] with different backbone networks [22, 41] on the proposed AP-10K dataset based on the MMPose [12] codebase. A single NVIDIA Tesla V100 GPU with 16GB memory is used during both the training and testing for all tracks, *i.e.*, the SL Track, CD-TL Track, and DG Track. The detailed settings for each track are presented in the following part. We adopt the mean average precision (mAP) as the primary evaluation metric in all tracks, following the human and animal pose estimation literature [41, 42, 44, 6].

4.2 Supervised learning track

Table 3: The evaluation results (mAP) of different models on the validation set of the SL Track.

	HRNet-w32 [41]	HRNet-w48 [41]	ResNet50 [24]	ResNet101 [24]	Hourglass [36]
w/o pretraining	$0.703_{\pm 0.002}$	$0.713_{\pm 0.002}$	$0.646_{\pm 0.001}$	$0.667_{\pm 0.002}$	$0.686_{\pm 0.006}$
w/ pretraining	$0.738_{\pm 0.006}$	$0.744_{\pm 0.004}$	$0.699_{\pm 0.004}$	$0.698_{\pm 0.002}$	$0.729_{\pm 0.001}$

Table 4: The evaluation results (mAP) of HRNet-w32 [41] on the validation set of the SL track.

Training epochs	210	420	630
Training from scratch	$0.703_{\pm 0.002}$	$0.721_{\pm 0.005}$	$0.725_{\pm 0.003}$
Pretraining on ImageNet	$0.738_{\pm 0.005}$	$0.738_{\pm 0.005}$	$0.735_{\pm 0.002}$

Settings The SL track aims to evaluate the representation ability of different human pose models, including Hourglass [36], SimpleBaseline [44] with ResNet50 [24] and ResNet101 [24] as backbones, and HRNet [41] for the animal pose estimation task. Hourglass utilizes multi-stage structure with skip connections for human pose estimation. SimpleBaseline adopts an encoder-decoder structure to regress the keypoints while HRNet keeps both high- and low-level features for more accurate pose estimation. We use Adam [26] as the optimizer and train these representative models for 210 epochs with a batch size of 64. The initial learning rate is set to 5e-4, and we adopt the step-wise learning rate schedule which reduces the learning rate at the 170 and 200 epoch, respectively. The training image of each instance is cropped and resized to 256 $\times$ 256, with random rotation, flip, and scale jitter as data augmentation. The AP-10K dataset is randomly split into the disjoint train, validation, and test sets three times following the same ratio 7:1:2, where we perform the experiment three times accordingly. We consider two popular training settings, *i.e*., with and without ImageNet [16] pretraining, respectively. ***Pretraining on ImageNet***: We follow the common practice in human pose estimation literature [44, 41] by initializing the backbone network with the pretrained weights on ImageNet and then finetuning the whole network on the AP-10K training set. ***Training from scratch on AP-10K***: We randomly initialize the network weights and directly train them on the AP-10K training set. We also adopt longer training schedules, *i.e*., 2$\times$ for 420 epochs and 3$\times$ for 630 epochs at these two settings. By comparing the performance gap between these two settings, we can investigate the impact of pretraining on ImageNet, especially when a longer training schedule is used for training from scratch.

Results on the SL Track The results are summarized in Table 3. We have the following observations. Firstly, with the increase of network complexity, the performance of both SimpleBaseline [44] and HRNet [41] methods is generally improved, especially at the setting of training from scratch. It is reasonable since the representation ability of each model mainly depends on the representation ability of the backbone network. Secondly, advanced network architecture like HRNet outperforms the encoder-decoder architecture in [44] and hourglass architecture in [36] due to its high-resolution feature representation ability. Thirdly, pretraining on the ImageNet leads to better performance for all the methods, which is attributed to the strong generalization ability of the pretrained network. By making a trade-off between model complexity and performance, we choose HRNet-w32 as the default model in the following experiments.

When adopting a longer training schedule, the performance gap between pretraining on ImageNet and training from scratch has been reduced as shown in Table 4, which is similar to the observation in [22] for generic object detection. In other word, pretraining helps accelerate the convergence speed since it can provide a good starting point of the network weights in the parameter space. However, when a longer training schedule is used, the network can also learn a good feature representation from the abundant training data. It again confirms the value of the proposed AP-10K.

4.3 Cross-domain transfer learning track

Table 5: The evaluation results (mAP) of HRNet-w32 [41] on the validation set of the CD-TL track.

epoch	AP	$AP_{.5}$	$AP_{.75}$	AP_M	AP_L
20	$0.606_{\pm 0.004}$	$0.906_{\pm 0.005}$	$0.635_{\pm 0.006}$	$0.501_{\pm 0.037}$	$0.610_{\pm 0.003}$
30	$0.642_{\pm 0.002}$	$0.921_{\pm 0.010}$	$0.680_{\pm 0.002}$	$0.521_{\pm 0.044}$	$0.645_{\pm 0.002}$
40	$0.667_{\pm 0.003}$	$0.934_{\pm 0.004}$	$0.714_{\pm 0.007}$	$0.547_{\pm 0.059}$	$0.671_{\pm 0.003}$
210	$0.753_{\pm 0.005}$	$0.962_{\pm 0.002}$	$0.827_{\pm 0.003}$	$0.616_{\pm 0.031}$	$0.756_{\pm 0.004}$

Settings Since four-foot animals usually share some similar keypoints as humans, *e.g*., two eyes, one nose, and one neck, the feature representation learned from a human pose estimation task may also be beneficial for animal pose estimation, especially on a small-scale dataset as shown in [6]. However, it is unclear the impact of such cross-domain transfer learning on a large-scale animal pose dataset with diverse species, *e.g*., our AP-10K. To answer this question, we use the pretrained

HRNet-w32 [41] on the COCO human pose dataset [31] to initialize the network weights and then finetune the model for different epochs, *i.e.*, 20, 30, 40, and 210 epochs, respectively. We follow the same setting in the SL track except that the learning rate is fixed as 1e-4.

Results on the CD-TL Track The results are summarized in Table 5, $AP_{.5/75}$ means using 0.5 and 0.75 as the threshold when computing the average precision and $AP_{M/L}$ represents the average precision on medium and large scale objects, respectively, as in COCO [31]. It can be concluded that the models transferred from human pose estimation do not perform well on the animal pose estimation task when the finetuning schedule is short, *e.g.*, 0.606 mAP (20 epochs), 0.642 mAP (30 epochs), and 0.667 mAP (40 epochs) compared with 0.738 mAP in Table 3. We suspect the performance gap is attributed to the domain gap between these two domains, *e.g.*, the instances in human pose estimation always wear clothes while animals are always covered by different types of fur. Nevertheless, training for a longer schedule (*e.g.*, 210 epochs) can help to close the gap and even improve the performance further, compared with pretraining on ImageNet, *e.g.*, 0.753 v.s. 0.738. implying that the domain gap between classification and animal pose estimation is larger than that between the two pose estimation tasks.

4.4 Intra- and inter-family domain generalization track

4.4.1 Intra-family domain generalization

Settings We choose the three largest families, *i.e.*, Bovidae, Canidae, and Felidae, to conduct the Intra-family DG experiments. For each family, we randomly select one species as the test set and use all other species as the train set. We follow the same setting as the pretraining on ImageNet in the SL track except that we train the models for more epochs to ensure that the same amount of training images are seen as in the SL Track.

Results on the Intra-family DG Track The results are summarized in Table 6, Table 7, and Table 8. The diagonal scores denote the results when testing on unseen species at different settings (as shown in the first column). Besides, we calculate the average evaluation result on each seen species at different settings and show them at the bottom of each column, respectively. We have the following observations. Firstly, when testing on unseen species, the model can still obtain a good performance although it is inferior to that when testing on seen species. It is reasonable since the species within the same family share some common features while having unique characteristics too. It is noteworthy that the evaluation result on Dog is worse than others since there are more images of dogs than fox and wolf. Besides, dogs varies widely in appearance owning to human's cultivation for family pets. Similar phenomena can also be observed in Table 8, *e.g.*, Cat. Secondly, the performance on each seen species is not as good as the corresponding performance when using all the species in AP-10K for training, *e.g.*, 0.663 mAP v.s. 0.761 mAP (as shown in Table S2 in the supplementary.) for Sheep. It implies that using the amount of training data but from more diverse species helps the model learning better feature representation, *i.e.*, confirming the value of our AP-10K again.

Table 6: Intra-family DG results (mAP) of HRNet-w32 [41] on the test set of the Bovidae family. Bov.=Bovidae, Ant.=Antelope, A.S.=Argali Sheep, Bis.=Bison, Buf.=Buffalo, She.=Sheep.

Train / Test	***Bov.***/Ant.	***Bov.***/A.S.	***Bov.***/Bis.	***Bov.***/Buf.	***Bov.***/Cow	***Bov.***/She.	**Average**
Antelope	$\mathbf{0.607_{\pm 0.010}}$	$0.742_{\pm 0.013}$	$0.775_{\pm 0.004}$	$0.842_{\pm 0.005}$	$0.729_{\pm 0.002}$	$0.853_{\pm 0.002}$	$0.788_{\pm 0.051}$
A.S.	$0.836_{\pm 0.016}$	$\mathbf{0.655_{\pm 0.015}}$	$0.805_{\pm 0.022}$	$0.840_{\pm 0.007}$	$0.725_{\pm 0.027}$	$0.697_{\pm 0.008}$	$0.781_{\pm 0.059}$
Bison	$0.731_{\pm 0.017}$	$0.646_{\pm 0.009}$	$\mathbf{0.530_{\pm 0.006}}$	$0.605_{\pm 0.006}$	$0.616_{\pm 0.009}$	$0.693_{\pm 0.014}$	$0.658_{\pm 0.047}$
Buffalo	$0.783_{\pm 0.010}$	$0.748_{\pm 0.031}$	$0.726_{\pm 0.017}$	$\mathbf{0.658_{\pm 0.004}}$	$0.794_{\pm 0.022}$	$0.750_{\pm 0.008}$	$0.760_{\pm 0.025}$
Cow	$0.597_{\pm 0.011}$	$0.691_{\pm 0.004}$	$0.740_{\pm 0.007}$	$0.732_{\pm 0.009}$	$\mathbf{0.586_{\pm 0.006}}$	$0.683_{\pm 0.002}$	$0.689_{\pm 0.051}$
Sheep	$0.707_{\pm 0.012}$	$0.607_{\pm 0.006}$	$0.681_{\pm 0.004}$	$0.676_{\pm 0.007}$	$0.645_{\pm 0.005}$	$\mathbf{0.520_{\pm 0.001}}$	$0.663_{\pm 0.034}$

Table 7: Intra-family DG results (mAP) of HRNet-w32 [41] on the test set of the Canidae family.

Train / Test	***Can.***/Dog	***Can.***/Fox	***Can.***/Wolf	**Average**
Dog	$0.224_{\pm 0.011}$	$0.699_{\pm 0.009}$	$0.699_{\pm 0.003}$	$0.699_{\pm 0.000}$
Fox	$0.614_{\pm 0.013}$	$0.627_{\pm 0.005}$	$0.732_{\pm 0.013}$	$0.673_{\pm 0.059}$
Wolf	$0.663_{\pm 0.024}$	$0.694_{\pm 0.013}$	$0.633_{\pm 0.006}$	$0.679_{\pm 0.016}$

Table 8: Intra-family DG results (mAP) of HRNet-w32 [41] on the test set of the Felidae family. Fel.=Felidae, Bob.=Bobcat, Che.=Cheetah, Jag.=Jaguar, K.C.=King Cheetah, Leo.=Leopard, Lio.=Lion, Pan.=Panther, S.L.=Snow Leopard, Tig.=Tiger.

Test \ Train	*Fel.*/Bob.	*Fel.*/Cat	*Fel.*/Che.	*Fel.*/Jag.	*Fel.*/K.C.	*Fel.*/Leo.	*Fel.*/Lio.	*Fel.*/Pan.	*Fel.*/S.L.	*Fel.*/Tig.	**Average**
Bob.	**0.631 ±0.005**	0.714 ±0.016	0.664 ±0.004	0.674 ±0.013	0.673 ±0.013	0.663 ±0.006	0.691 ±0.016	0.623 ±0.004	0.669 ±0.005	0.713 ±0.008	0.676 ±0.026
Cat	0.638 ±0.002	**0.332 ±0.004**	0.625 ±0.018	0.552 ±0.010	0.629 ±0.007	0.641 ±0.009	0.601 ±0.004	0.609 ±0.010	0.582 ±0.014	0.608 ±0.007	0.609 ±0.027
Che.	0.715 ±0.002	0.716 ±0.012	**0.660 ±0.003**	0.762 ±0.013	0.731 ±0.014	0.747 ±0.010	0.734 ±0.021	0.790 ±0.008	0.713 ±0.008	0.662 ±0.008	0.730 ±0.034
Jag.	0.757 ±0.005	0.770 ±0.017	0.754 ±0.006	**0.704 ±0.008**	0.750 ±0.004	0.759 ±0.012	0.798 ±0.013	0.724 ±0.008	0.756 ±0.011	0.734 ±0.005	0.756 ±0.020
K.C.	0.961 ±0.008	0.804 ±0.035	0.692 ±0.042	0.771 ±0.028	**0.779 ±0.010**	0.958 ±0.008	0.713 ±0.017	0.924 ±0.026	0.864 ±0.033	0.838 ±0.016	0.836 ±0.094
Leo.	0.730 ±0.005	0.697 ±0.007	0.766 ±0.014	0.741 ±0.006	0.682 ±0.005	**0.686 ±0.009**	0.700 ±0.012	0.705 ±0.012	0.775 ±0.010	0.744 ±0.004	0.727 ±0.031
Lio.	0.623 ±0.016	0.582 ±0.023	0.639 ±0.012	0.694 ±0.010	0.688 ±0.002	0.690 ±0.018	**0.528 ±0.002**	0.638 ±0.007	0.630 ±0.011	0.625 ±0.024	0.645 ±0.036
Pan.	0.705 ±0.020	0.722 ±0.011	0.718 ±0.020	0.720 ±0.023	0.727 ±0.013	0.785 ±0.014	0.763 ±0.026	**0.511 ±0.014**	0.719 ±0.004	0.684 ±0.018	0.727 ±0.028
S.L.	0.792 ±0.011	0.776 ±0.008	0.810 ±0.018	0.779 ±0.019	0.790 ±0.024	0.818 ±0.004	0.821 ±0.009	0.760 ±0.015	**0.724 ±0.010**	0.855 ±0.012	0.800 ±0.027
Tig.	0.754 ±0.008	0.741 ±0.018	0.751 ±0.012	0.715 ±0.015	0.768 ±0.021	0.753 ±0.015	0.797 ±0.005	0.848 ±0.023	0.744 ±0.011	**0.675 ±0.007**	0.763 ±0.036

4.4.2 Inter-family domain generalization

Settings In this track, we select one family, *i.e.*, Bovidae (Bov.), for training while using several other families, *i.e.*, Cervidae (Cerv.), Equidae (Equ.), and Hominidae (Hom.), for testing. We follow the same training setting as the pretraining on ImageNet in the SL track.

Results on the inter-family DG Track The results are summarized in Table 9. We calculate the average score of all species in each family at the top row in each part (*i.e.*, train and test). As can be seen, the model trained on the Bovidae family not only performs well on the species belonging to it but also on species belonging to the Cervidae family. In addition, its performance drops a little on the Equidae family while generalizing poorly on the Hominidae family. It is reasonable as the Bovidae family and the Cervidae family belong to the same order, *i.e.*, Artiodactyla, and the biological proximity implies the similarity of pose distribution in these two families. For the Equidae family belonging to the same Euungulata clade but in adifferent order as Bovidae, their posture distributions may have a small difference, which explains the performance drop. For the Hominidae family belonging to a different clade, *i.e.*, Primates clade, the learned knowledge about body keypoints from the Bovidae family is not suitable anymore. As shown in the last column in Table 9, where the model is trained on the Cercopithecidae family, it can generalize to Hominidae well since they belong to the same order, *i.e.*, in line with the observation from the first column.

Table 9: Inter-family DG results (mAP) of HRNet-w32 [41] on the different families' test set. Bov. = Bovidae, Ant. = Antelope, A.S. = Argali Sheep, Bis. = Bison, Buf. = Buffalo, She. = Sheep, Cer. = Cervidae, Der. = Deer, Moo. = Moose, Equ. = Equidae, Hor. = Horse, Zeb. = Zebra, Hom. = Hominidae, Chi. = Chihuahua, Gor. = Gorilla, Cerc. = Cercopithecidae, Alo. = Alouatta, Mon. = Monkey, N.N.M. = Noisy Night Monkey, S.M. = Spider Monkey, Uak. = Uakari.

	Bov.	$0.782_{\pm 0.002}$	Bov.	$0.782_{\pm 0.002}$	Bov.	$0.782_{\pm 0.002}$	Cerc.	$0.695_{\pm 0.007}$
	Ant.	$0.856_{\pm 0.001}$	Ant.	$0.856_{\pm 0.001}$	Ant.	$0.856_{\pm 0.001}$	Alo.	$0.697_{\pm 0.020}$
	A.S.	$0.887_{\pm 0.006}$	A.S.	$0.887_{\pm 0.006}$	A.S.	$0.887_{\pm 0.006}$	Mon.	$0.725_{\pm 0.013}$
train	Bis.	$0.643_{\pm 0.005}$	Bis.	$0.643_{\pm 0.005}$	Bis.	$0.643_{\pm 0.005}$	N.N.M.	$0.750_{\pm 0.027}$
	Buf.	$0.815_{\pm 0.004}$	Buf.	$0.815_{\pm 0.004}$	Buf.	$0.815_{\pm 0.004}$	S.M.	$0.581_{\pm 0.008}$
	Cow	$0.737_{\pm 0.004}$	Cow	$0.737_{\pm 0.004}$	Cow	$0.737_{\pm 0.004}$	Uak.	$0.720_{\pm 0.009}$
	She.	$0.754_{\pm 0.005}$	She.	$0.754_{\pm 0.005}$	She.	$0.754_{\pm 0.005}$		
	Cer.	$0.641_{\pm 0.007}$	Equ.	$0.468_{\pm 0.019}$	Hom.	$0.015_{\pm 0.001}$	Hom.	$0.446_{\pm 0.007}$
test	Der.	$0.724_{\pm 0.004}$	Hor.	$0.618_{\pm 0.005}$	Chi.	$0.005_{\pm 0.000}$	Chi.	$0.446_{\pm 0.011}$
	Moo.	$0.558_{\pm 0.010}$	Zeb.	$0.319_{\pm 0.035}$	Gor.	$0.026_{\pm 0.003}$	Gor.	$0.445_{\pm 0.011}$

4.4.3 Inter-family transfer learning and few-shot learning

Settings We further evaluate the ability of models about inter-family generalization under a normal transfer learning setting and a few-shot learning setting. We follow the setting in Section 4.4.2 in these experiments, *i.e.*, we pretrain the model using the Bovidae family and test on the species from the Cervidae, Equidae, and Hominidae families. ***Transfer learning setting***: We randomly select 140 images from each species as training set and use the others as test set. The models are then finetuned with each species's training images and evaluated with the test images. We use initial learning rate 1e-5 with linear weight decay schedule and Adam [26] optimizer for 35 epochs during the training. ***Few-shot learning setting***: Further, while keeping the test set unchanged, we select 20 images randomly in the training set of each species to finetune the models at the few-shot learning setting. The models are trained for 50 epochs with fixed learning rate 1e-5 and Adam [26] optimizer.

Inter-family transfer learning and few-shot learning results The results are summarized in Table 10. 'Generalization' means directly test the pretrained models on each species' test set, as in Section 4.4.2. With only 20 images for finetuning, the models' performance on the other species increase quickly, especially for the Zebra species, which have similar pose with the Bovidae family but different textures. With more images for finetuning, *i.e.*, in the transfer learning setting, the performance further improves. *e.g.*, the models' performance on the Chimpanzee species increase from 0.022 to 0.550. Such observation demonstrates that with more training data do help to improve the performance on new categories.

Table 10: Inter-family's generalization, few-shot (20-shot), and transfer learning results (mAP).

Species	Setting	Performance	Species	Setting	Performance
Deer	Generalization	$0.723_{\pm 0.036}$	Moose	Generalization	$0.587_{\pm 0.025}$
	Few-Shot	$0.742_{\pm 0.034}$		Few-Shot	$0.648_{\pm 0.025}$
	Transfer	$0.751_{\pm 0.024}$		Transfer	$0.726_{\pm 0.011}$
Horse	Generalization	$0.592_{\pm 0.047}$	Zebra	Generalization	$0.324_{\pm 0.021}$
	Few-Shot	$0.635_{\pm 0.034}$		Few-Shot	$0.480_{\pm 0.029}$
	Transfer	$0.718_{\pm 0.023}$		Transfer	$0.708_{\pm 0.024}$
Chimpanzee	Generalization	$0.009_{\pm 0.006}$	Gorilla	Generalization	$0.017_{\pm 0.006}$
	Few-Shot	$0.022_{\pm 0.010}$		Few-Shot	$0.144_{\pm 0.121}$
	Transfer	$0.550_{\pm 0.032}$		Transfer	$0.662_{\pm 0.039}$

4.4.4 Cross animal pose dataset evaluation

Table 11: Results of HRNet-w32 for cross animal pose dataset evaluation. Direct Test: evaluating the pretrained model from the source dataset on the target test set directly. Finetune&Test: finetuning the pretrained model from the source dataset on the target dataset and then testing on the target test set. Train&Test: training and testing the model on the target dataset.

	Direct Test (mAP)	Finetune&Test (mAP)	Train&Test (mAP)
Animal-Pose Dataset [6] $\rightarrow$ AP-10K	0.424	0.722	0.727
AP-10K $\rightarrow$ Animal-Pose Dataset [6]	0.913	0.935	0.932

In this part, we evaluate the generalization ability of models trained on different animal pose datasets, including the Animal Pose dataset [6] and our AP-10K dataset. The results are shown in Table 11. To be fair, we only compute the mAP on the 17 common keypoints shared by these two datasets. As can be seen, the model trained on the AP-10K dataset can generalize well on the Animal Pose dataset [6] but not vice versa. This is because the Animal Pose dataset [6] only has 5 animal species, which is far from enough to generalize well to other animal species. On the contrary, AP-10K has 60 animal species and can generalize well to animals, including those in the Animal Pose dataset [6]. Some visual results are presented in Figure 4. It is exciting to find that the model trained on our AP-10K can even predict the keypoints that are not annotated by annotators, *e.g.*, the right eye of the bison and panda or the throat of the uakari and gorilla. Such phenomena imply that the model has learned good knowledge of animal body from the diverse species in our AP-10K dataset.

5 Discussion

Based on the experimental results on the SL track, the CD-TL track, and the DG track, we make a step forward towards understanding the research questions posed in Section 1 and obtain empirical

Figure 4: Some qualitative results of HRNet-w32 trained on the Animal Pose dataset [6] (the first row) and our AP-10K dataet (the second row). The ground truth poses are shown in the last row.

evidence to support the value of the proposed AP-10K dataset. The results also suggest that more efforts could be made to improve the efficiency of cross-domain transfer learning, address the domain gap issue between different intra- and inter-family species, as well as deal with the rare species in the long-tail distribution of animals. In addition to the study, AP-10K can also enable further research in different contexts. Firstly, since there are about 50k animal images without keypoint annotations while having family and species category labels, it is promising to study semi-supervised learning and self-supervised learning for animal pose estimation. Secondly, few-shot learning can also be an interesting direction, since images in AP-10K have a long tail distribution. How to deal with the rare species and improve the poses estimation performance remains underexplored. Thirdly, the study of inter-family or inter-species domain generalization deserves more effort, especially for those families or species that are not in the same clade or order. Although AP-10K is the largest dataset in this area, it is $10\times$ smaller than those for human pose estimation, *e.g.*, COCO [31], we plan to increase its volume as well as species diversity in the future.

6 Conclusion

In this paper, we establish the first large-scale dataset for mammal pose estimation, *i.e.*, AP-10K. It facilitates the study of new research questions due to its rich diversity in posture, scale, occlusion, and species of animals and organization structure following the taxonomic rank. We benchmark representative pose estimation methods on AP-10K and benefit from the empirical results to understand the representation ability of different models, the impact of pretraining on ImageNet or human pose dataset, the benefit of using diverse animal species for training, as well as intra- and inter-family model generalization. We hope AP-10K can pave the way for the follow-up study in this area.

Broad Impacts AP-10K can potentially benefit the study of animal behavior understanding, zoology, and wildlife conservation. Nevertheless, it covers much fewer species than those in the real world, the generalization ability of trained models on it should be paid careful attention to.

Acknowledgement The creation of the dataset is founded by the Innovation Capability Support Program of Shaanxi under the grant of Program No.2021TD-05 and the National Natural Science Foundation of China under the grant of No.62133012, No.61936006. Mr. Yufei Xu and Dr. Jing Zhang are supported by the ARC project FL-170100117.

References

[1] C. Alessio. Animals 10. https://www.kaggle.com/alessiocorrado99/animals10, 2019.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[3] antoreepjana. Iucn animals dataset. https://www.kaggle.com/antoreepjana/iucn-animals-dataset, 2021.

[4] A. Arac, P. Zhao, B. H. Dobkin, S. T. Carmichael, and P. Golshani. Deepbehavior: A deep learning toolbox for automated analysis of animal and human behavior imaging data. *Frontiers in systems neuroscience*, 13:20, 2019.

[5] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–478, 2018.

[6] J. Cao, H. Tang, H.-S. Fang, X. Shen, C. Lu, and Y.-W. Tai. Cross-domain adaptation for animal pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[9] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[10] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020.

[11] B. P. CLi Weng. A secure perceptual hash algorithm for image content authentication. 2011.

[12] M. Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[14] A. M. Dai and Q. V. Le. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28:3079–3087, 2015.

[15] K. Davies. Keep the directive that protects research animals. *Nature*, 521(7550):7–7, 2015.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] B. Ferreira. African wildlife. https://www.kaggle.com/biancaferreira/african-wildlife, 2021.

[18] J. M. Graving, D. Chae, H. Naik, L. Li, B. Koger, B. R. Costelloe, and I. D. Couzin. Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife*, 8:e47994, 2019.

[19] A. Gupta, P. Dollar, and R. Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019.

[20] E. J. Harding, E. S. Paul, and M. Mendl. Animal behaviour: Cognitive bias and affective state. *Nature*, 427(6972):312–312, 2004.

[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[22] K. He, R. Girshick, and P. Dollar. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] S. Jamil. Endangered animals. https://www.kaggle.com/sonain/endangered-animals, 2020.

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[28] R. Labuguen, J. Matsumoto, S. Negrete, H. Nishimaru, H. Nishijo, M. Takada, Y. Go, K.-i. Inoue, and T. Shibata. Macaquepose: A novel in the wildmacaque monkey pose dataset for markerless motion capture. *bioRxiv*, 2020.

[29] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018.

[30] S. Li, J. Li, H. Tang, R. Qian, and W. Lin. Atrw: A benchmark for amur tiger re-identification in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2590–2598, 2020.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[32] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng. Improving convolutional networks with self-calibrated convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10096–10105, 2020.

[33] A. Mathis, T. Biasi, S. Schneider, M. Yuksekgonul, B. Rogers, M. Bethge, and M. W. Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1859–1868, 2021.

[34] O. Mazhar, S. Ramdani, B. Navarro, R. Passama, and A. Cherubini. Towards real-time physical human-robot interaction using skeleton information and hand gestures. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–6. IEEE, 2018.

[35] A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in neural information processing systems*, pages 2277–2287, 2017.

[36] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[37] B. Ni, T. Li, and X. Yang. Learning semantic-aligned action representation. *IEEE transactions on neural networks and learning systems*, 29(8):3715–3725, 2017.

[38] T. D. Pereira, D. E. Aldarondo, L. Willmore, M. Kislin, S. S.-H. Wang, M. Murthy, and J. W. Shaevitz. Fast animal pose estimation using deep neural networks. *Nature methods*, 16(1):117–125, 2019.

[39] A. Saxena. Animal image dataset(dog, cat and panda). `https://www.kaggle.com/ashishsaxena2209/animal-image-datasetdog-cat-and-panda`, 2019.

[40] Y. V. Trivedi. Animals 5. `https://www.kaggle.com/ytrivedi1/animals-5`, 2020.

[41] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[42] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019.

[43] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learninga comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[44] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[45] J. Zhang, Z. Chen, and D. Tao. Towards high performance human keypoint detection. *International Journal of Computer Vision*, pages 1–24, 2021.

[46] J. Zhang and D. Tao. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal*, 8(10):7789–7817, 2020.

[47] S.-H. Zhang, R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang, and S.-M. Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 889–898, 2019.

[48] E. ahovi. wild cats. `https://www.kaggle.com/enisahovi/cats-projekat-4`, 2020.