AnimalWeb: A Large-Scale Hierarchical Dataset of Annotated Animal Faces

Muhammad Haris Khan[1], John McDonagh[2], Salman Khan[1], Muhammad Shahabuddin[4]
Aditya Arora[1], Fahad Shahbaz Khan[1], Ling Shao[1], Georgios Tzimiropoulos[3]
[1]Inception Institute of Artificial Intelligence, UAE [2]University of Nottingham, UK
[3]Queen Mary University of London, UK [4]Comsats University Islamabad, Pakistan
{muhammad.haris,salman.khan,fahad.khan,ling.shao}@inceptioniai.org, shahab.pk05@gmail.com
john.mcdonagh@nottingham.ac.uk, g.tzimiropoulos@qmul.ac.uk

Abstract

Several studies show that animal needs are often expressed through their faces. Though remarkable progress has been made towards the automatic understanding of human faces, this has not been the case with animal faces. There exists significant room for algorithmic advances that could realize automatic systems for interpreting animal faces. Besides scientific value, resulting technology will foster better and cheaper animal care.

We believe the underlying research progress is mainly obstructed by the lack of an adequately annotated dataset of animal faces, covering a wide spectrum of animal species. To this end, we introduce a large-scale, hierarchical annotated dataset of animal faces, featuring 22.4K faces from 350 diverse species and 21 animal orders across biological taxonomy. These faces are captured 'in-the-wild' conditions and are consistently annotated with 9 landmarks on key facial features. The dataset is structured and scalable by design; its development underwent four systematic stages involving rigorous, overall effort of over 6K man-hours. We benchmark it for face alignment using the existing art under two new problem settings. Results showcase its challenging nature, unique attributes and present definite prospects for novel, adaptive, and generalized face-oriented CV algorithms. Further benchmarking the dataset across face detection and fine-grained recognition tasks demonstrates its multi-task applications and room for improvement. The dataset is available at: https://fdmaproject.wordpress.com/.

1. Introduction

Animals are a fundamental part of our world. Their needs are often expressed through faces which, if understood properly, can help us improve the well-being of animals in labs, farms and homes. Behavioural and neurophysiological studies have shown that mammalian brains can interpret social signals on fellow animal's faces and have developed specialized skills to process facial features. Therefore, the study of animal faces is of prime importance.

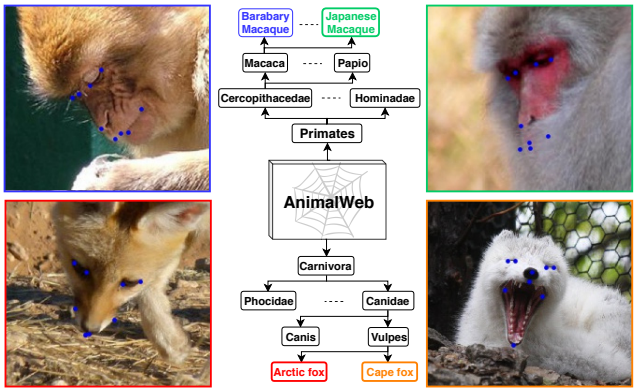

Figure 1: *AnimalWeb:* We introduce a large-scale, hierarchical dataset of annotated animal faces featuring diverse species while covering a broader spectrum of animal biological taxonomy. It exhibits unique challenges e.g., large biodiversity in species, high variations in pose, scale, appearance, and backgrounds. Further, it offers unique attributes like class imbalance (CI), multi-task applications (MTA), and zero-shot face alignment (ZFA). Facial landmarks shown in blue and the images belong to classes with identical color in the hierarchy.

Facial landmarks can help us better understand animals and foster their well-being via deciphering their facial expressions. Facial expressions reflect the internal emotions and psychological state of an animal being. As an example, animals with different anatomical structure (such as mice, horses, rabbits and sheep), show a similar grimace expression when in pain i.e., tighten eyes and mouth, flatten cheeks and unusual ear postures. Understanding abnormal animal expressions and behaviours with visual imagery is a much cheaper and quicker alternative to clinical examinations and vital signs monitoring.

Encouraging indicators show that such powerful tech-

nologies could indeed be possible, e.g., fearful cows widen their eyes and flatten their ears [19], horses close eyes in depression [10], sheep positions its ears backward when facing unpleasant situations [2], and rats ear change colors and shape when in joy [9]. Furthermore, large-scale annotated datasets of animal faces can help advance the animal psychology understanding. For example, for non-primate animals, the scientific understanding of animal expressions is generally limited to the development of only pain coding systems [13]. However, other expressions could be equally important to understand e.g., sadness, boredom, hunger, anger and fear.

We believe the research progress towards automatic understanding of animal facial behaviour is largely hindered by the lack of sufficiently annotated animal faces (Tab. 1), covering a wide spectrum of animal species. In comparison, significant progress has been made towards automatic understanding and interpretation of human faces [40, 5, 35, 34, 3, 21, 38], while animal face analysis is largely unexplored in vision community [41, 25]. There is a plenty of room for new algorithms and a pressing need to develop computational tools capable of understanding animal facial behavior. To this end, we introduce a large-scale, hierarchical dataset of annotated animal faces, termed AnimalWeb, featuring diverse species while covering a broader spectrum of animal biological taxonomy. Every image has been labelled with the genus-species terminology. Fig. 1 provides a holistic overview of the dataset key features.

Contributions: To our knowledge, we build and annotate the largest animal faces dataset captured under altogether in-the-wild conditions. It encompasses 21 different orders and within order explores various families and genuses. This diverse coverage results in 350 different animal species and a total count of 22.4K animal faces. Each face is consistently annotated with 9 fiducial landmarks on key facial components (e.g., eyes and mouth). Finally, the dataset design and development followed four systematic stages involving an overall, rigorous effort of over 6K man-hours by experts and trained volunteers.

We benchmark AnimalWeb for face alignment with the state-of-the-art (SOTA) human face alignment algorithms [3, 39]. Results show that it is challenging for them particularly due to biodiversity, species imbalance, and adverse in-the-wild conditions (e.g., extreme poses). We further validate this by reporting results from various analysis, including pose-wise and face sizes. We show the capability of our dataset for testing under two novel problem settings: few-shot and zero-shot face alignment. Further, we demonstrate related applications possible with this dataset: animal face detection and fine-grained species recognition. Our results show that it 1) is a strong experimental base for algorithmic advances, and 2) will facilitate the development of novel, adaptive, and generalized face-oriented algorithms.

2. Related Datasets

This section briefly overviews existing human and animal face alignment benchmarks.

Human Face Alignment. Since the seminal work of Active Appearance Models (AAMs) [6], various 2D datasets featuring human face landmark annotations have been proposed. Among these, the prominent ones are XM2VTS [22], BioID [16], FRGC [23], and Multi-PIE [12]. These datasets were collected under constrained environments with limited expression, frontal pose, and normal lighting variations. Following them, few datasets were proposed with faces showing occlusions and other variations such as COFW [4, 11] and AFW [44].

300W [29] is a popular dataset amongst several others in human face alignment, and has been widely adopted both by scientific community and industry [34, 40, 26, 43]. It was developed for the 300W competition held in conjunction with ICCV 2013. 300W benchmark originated from LFPW [1], AFW [44], IBUG [29], and 300W private [28] datasets. In total, it provides 4,350 images with faces annotated using the 68 landmark frontal face markup scheme. To promote face tracking research, 300VW [30] is introduced featuring 114 videos. Such datasets paced research progress towards human face alignment in challenging conditions.

Recently, efforts are directed to manifest greater range of variations. For instance, Annotated Facial Landmarks in the wild (AFLW) [18] proposed a collection of 25K annotated human faces with up to 21 landmarks. It, however, excluded locations of invisible landmarks. Zhu *et al*. [43] provided manual annotations for invisible landmarks, but there are no landmark annotations along the face contour. Along similar lines, Zhu *et al*. [44] developed a large scale training dataset by synthesizing profile views from 300W dataset using a 3D Morphable Model (3DMM). Though it could serve as a large training set, the synthesized profile faces have artifacts that can hurt fitting accuracy. Jeni *et al*. [15] introduced a dataset in an ECCV 2016 competition, comprising photographed images in controlled conditions or synthetically produced images.

Lately, Menpo benchmark [8] was released in competitions held along ICCV 2017. It contains 2D and 3D landmarks annotations and exhibits large variations in pose, expression, illumination and occlusions. Faces are also classified into semi-frontal and profile based on their orientation and annotated accordingly. Menpo-2D contains 7,576 and 7,281 annotated training and testing images, respectively.

Animal Face Alignment. Despite scientific value, pressing need and direct impact on animal healthcare, only little attention has been paid in developing an annotated dataset of animal faces [41, 25]. Although datasets such as ImageNet [8] and iNaturalist [36] offer reasonable species variety, they are targeted at image-level classification and region-level detection tasks. The two animal face alignment

Figure 2: Some representative examples from randomly chosen species in AnimalWeb. Animal faces tend to exhibit large variations in pose, scale, appearance and expressions.

Dataset	Target Face	Faces	Points
Multi-PIE [12] (semi-frontal)	Human	6665	68
Multi-PIE [12] (profile)	Human	1400	39
AFLW [18]	Human	25,993	21
COFW [4]	Human	1007	29
COFW [11]	Human	507	68
300 W[29, 28]	Human	3837	68
Menpo 2D [8] (semi-frontal)	Human	10,993	68
Menpo 2D [8] (profile)	Human	3852	39
AFLW2000-3D [44]	Human	2000	68
300W-LP [44](synthetic)	Human	61,225	68
Sheep faces [41]	Animal	600	8
Horse faces [25]	Animal	3717	8
AnimalWeb (Ours)	Animal	22,451	9

Table 1: Comparison between AnimalWeb and various popular face alignment datasets. AnimalWeb is bigger (in terms of faces offered) than 80% of the datasets targeted at human face alignment. Further, the existing efforts on animal face datasets are limited to only single species. This work targets a big gap in this area by building a large-scale annotated animal faces dataset.

datasets were reported in [41] and [25]. Yang *et al*. [41] collected 600 sheep faces and annotated them with 8 fiducial landmarks. Similarly, Rashid *et al*. [25] reported a collection of 3717 horse faces with points marked around 8 facial features. These datasets are severely limited in terms of biodiversity, size, and range of possible real-world conditions. To our knowledge, the proposed dataset is a first large-scale, hierarchical collection of annotated animal faces with 9 landmarks, possessing real-world properties (e.g., large poses) and unique attributes e.g., species imbalance, multi-task applications, and zero-shot face alignment.

3. AnimalWeb Properties

In this section, we highlight some of the unique aspects of the newly introduced dataset (Fig. 2).

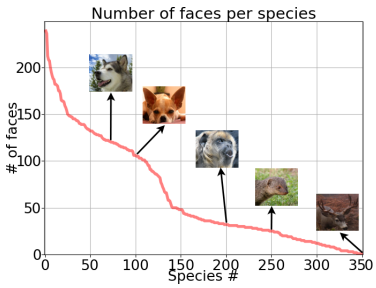

Figure 3: Distribution of faces per species in AnimalWeb. We see that 29% of the total species contain 65% of the total faces. The dataset shows the natural occurrence patterns of different species.

Scale. The proposed dataset is offering a large-scale and diverse coverage of annotated animal faces. It contains 22.4K annotated faces, offering 350 different animal species with variable number of animal faces in each species. Fig. 3 shows the distribution of faces per species. We see that 29% of the total species contain 65% of the total faces. Also, the maximum and minimum number of faces per species are 239 and 1, respectively. Both these statistics highlight the large imbalance between species and high variability in the instance count for different species. This marks the conformity with the real-world where different species are observed with varying frequencies.

Tab. 1 compares AnimalWeb and various popular datasets for face alignment. AnimalWeb is bigger (in face count) compared to 80% of datasets targeted at human face alignment. Importantly, very little or rather no attention is subjected towards constructing annotated animal faces dataset mimicking real-world properties, and the existing ones are limited to only single species.

Diversity. Robust computational tools aimed at detecting/tracking animal facial behaviour in open environments are difficult to realize without observations that can exhibit real-world scenarios as much as possible. We therefore aim at ensuring diversity along two important dimensions, **(1)**

imaging variations in scale, pose, expression, and occlusion, **(2)** species coverage in the animal biological taxonomy. Fig. 2 shows some example variations captured in the dataset. We observe that animal faces exhibit great pose variations and their faces are captured from very different angles (e.g., top view) that are quite unlikely for human faces. In addition, animal faces can show great range of pose and scale variations.

Fig. 4 (top row) reveals that faces in AnimalWeb exhibits much greater range of shape deformations. Each image is obtained by warping all possible ground truth shapes to a reference shape, thereby removing similarity transformations. Fig. 4 (bottom row) attempts to demonstrate image diversification in AnimalWeb and other datasets. We observe that it comprises more diversified images than other commonly available human face alignment datasets. To gauge scale diversity, we plot the distribution of normalized face sizes for AnimalWeb in Fig. 5 and popular human face alignment datasets. AnimalWeb offers 32% more range of small face sizes (< 0.2) in comparison to competing datasets for human face alignment.

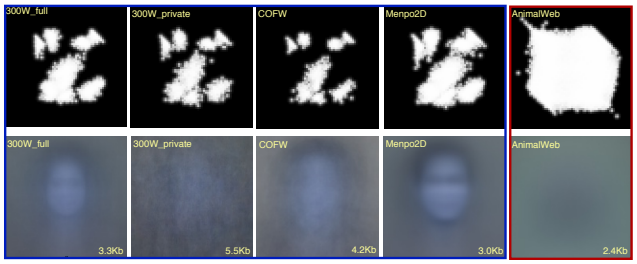

Figure 4: Top: AnimalWeb covers significantly larger deformations. Bottom: It offers more diversity - large variability in appearances, viewpoints, poses, clutter and occlusions resulting in the blurriest mean image with the smallest lossless JPG file size.

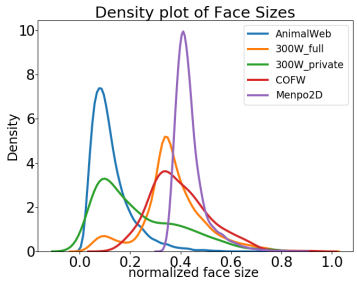

Figure 5: Face sizes distribution in AnimalWeb and popular human face alignment datasets. AnimalWeb offers 32% more range of small face sizes (< 0.2) in comparison to competing datasets.

Fig. 6 provides a miniature view of the hierarchical nature, illustrating diversity in AnimalWeb. Primates and Carnivora orders have been shown with randomly chosen 8 and 5 families alongside a few genuses. We observe that it exhibits hierarchical structure with variable number of children nodes for each parent node. We refer to Tab. 2 for the count of families, genuses, species, and faces in top 5 orders (ranked by face count).

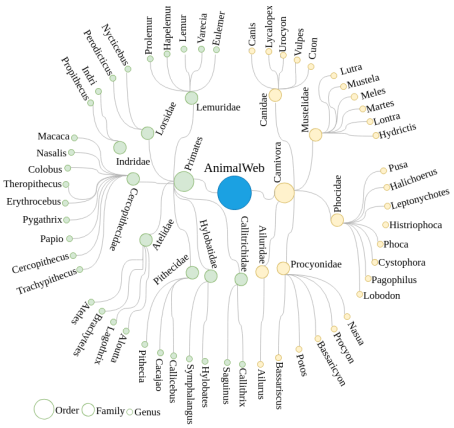

Figure 6: A miniature glimpse of the hierarchical nature of AnimalWeb. Primates and Carnivora orders have been shown with a few families and respective genuses.

4. Constructing AnimalWeb

This section details four key steps followed towards the construction of AnimalWeb (see Fig. 7). They include image collection, workflow development, facial point annotation, and annotation refinement.

4.1. Image Collection

We first developed a taxonomic framework to realise a structured, scalable dataset design followed by a detailed collection protocol to ensure real-world conditions before starting image collection process.

Taxonomic Framework Development. A simple, hierarchical tree-like data structure is designed following the well established biological animal classification. The prime motivation is to carry out image collection - the next step - in a structured and principled way. Further, this methodology enables recording various statistics e.g., image count at different nodes of the tree.

Data Collection Protocol. Starting from animal kingdom we restricted ourselves to vertebrates group (phylum), and further within vertebrates to Mammalia class. We wanted those animals whose faces exhibit roughly regular and identifiable face structure. Some excluded animal examples are insects and worms that possibly violate this condition. Given these restrictions, 21 orders were shortlisted for collection task. Scientific names of top 5 orders in terms of face count are reported in Tab. 2.

Order	Families	Genuses	Species	Faces
Carnivora	11	57	144	8281
Artiodactyla	7	42	55	4546
Primates	12	30	59	3468
Rodentia	11	19	19	1521
Sphenisciformes	1	5	10	1516

Table 2: Top 5 orders in terms of face count covered in AnimalWeb. For each order we show the number of families, genuses, species, and faces. There are a total of 21 orders and each order explores on average 3 families, 8 genuses, and 1024 faces.

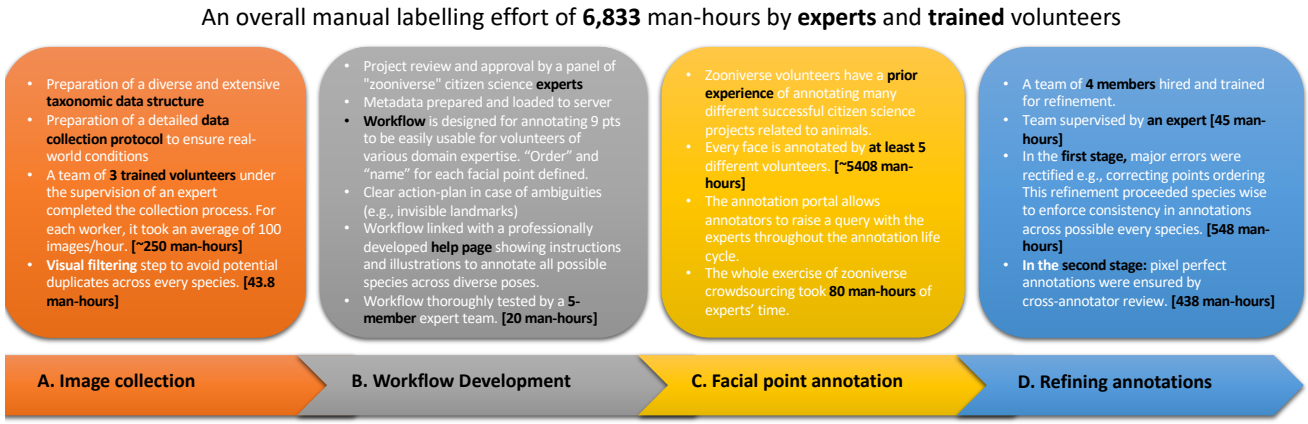

Figure 7: Four systematic stages in AnimalWeb development with details and man-hours involved. Zoom-in for details.

Finally, we set the bound for number of images to be collected per genus-species between 200-250. This would increase the chances of valuable collection effort to be spent in exploring the different possible species - improving biodiversity - rather than heavily populating a few (commonly seen). With this constraint, we ended up with an average of 65 animal faces per species.

Image Source. The Internet is the only source used for collecting images for this dataset. Other large-scale computer vision datasets such as ImageNet [7] and MS COCO [20] have also relied on this source to achieve the same. Specifically, we choose Flickr[1], which is a large image hosting website, to search first, then select, and finally download relevant animal faces.

Collection. We use both common and scientific names of animal species from the taxonomic framework (described earlier) to query images. Selection is primarily based on capturing various in-the-wild conditions e.g. various face poses. A team of 3 trained volunteers completed the image collection process under the supervision of an expert. For each worker, it took an average of 100 images per hour amounting to a total of ∼250 man-hours. After download, we collected around 25K candidate images. Finally, a visual filtering step helped removing potential duplicates across species in 43.8 man-hours.

4.2. Workflow Development

Annotating faces can unarguably be the most important, labour-intensive and thus a difficult step towards this dataset construction. To actualize this, we leveraged the great volunteers resource from a large citizen science web portal, called Zooniverse [2]. It is home to many successful citizen science projects. We underwent the following stages to accomplish successful project launch through this portal.

Project Review. This is the *first* stage and it involves project design and review. The project is only launched once it gets reviewed by Zooniverse experts panel whom main selection criterion revolves around gauging the impact of a research project.

Workflow design and development. Upon clearing review process, in the *second* phase, the relevant image metadata is uploaded to the server and an annotator interface (a.k.a workflow) is developed. The workflow is first designed for annotating points and is then thoroughly verified. Two major quality checks are 1) its ease of use for a large volunteer group, bearing different domain expertise, and 2) its fitness towards the key project deliverables. In our case, the workflow defines 'order' and 'name' for each facial point. Further, it also comprises a clear action-plan in case of ambiguities (e.g., invisible landmarks) by linking a professionally developed help page. It shows instructions and illustrations to annotate points across all possible species across diverse poses. Lastly, our workflow is thoroughly tested by a 5-member team of experts and it took 20 man-hours of effort.

9 pts. markup scheme. The annotator interface in our case required annotators to adhere to the 9 landmarks markup scheme as shown in Fig. 8. We believe that 9 landmarks provide good trade-off between annotation effort and facial features coverage.

4.3. Facial Point Annotation

After workflow development, the project is exposed to a big pool of Zooniverse volunteers for annotating facial landmarks. These volunteers have a prior experience of annotating many different successful citizen science projects related to animals. Every face is annotated by at least 5 different volunteers and this equals a labour-intensive effort of ∼5408 man-hours in total. Multiple annotations of a single face improves the likelihood of recovering annotated points closer to the actual location of facial landmarks, provided more than half of these multiple annotations qualify this assumption. To this end, we choose to take median value of multiple annotations of a single face.

The annotation portal allows annotators to raise a query

[1] https://www.flickr.com/
[2] https://www.zooniverse.org/

Figure 8: Nine landmarks markup scheme used for annotation of faces in AnimalWeb. The markup scheme covers major facial features around key face components (eyes, nose, and lips) while keeping the total landmark count low.

with the experts throughout the annotation life cycle. This also helps in removing many different annotation ambiguities for other volunteers as well who might experience the same later in time. The whole exercise of Zooniverse crowdsourcing took 80 man-hours of experts' time.

4.4. Refining Annotations

Annotations performed by zooniverse volunteers can be inaccurate and missing for some facial points. Further they could be inconsistent, and unordered. Unordered point annotations result if, for instance, left eye landmark is swapped with right eye. Above mentioned errors are in some sense justifiable since point annotations on animal faces, captured in real-world settings, is a complicated task.

We hired a small team of 4 trained volunteers for refinement. It had to perform manual corrections and was also supervised by an expert. The refinement completed in two passes listed below.

Refinement Passes. In the first pass, major errors were rectified e.g., correcting points ordering. This refinement proceeded species-wise to enforce consistency in annotations across every possible species in the dataset. A total of 548 man-hours were spent in the first pass. In the second pass, pixel perfect annotations were ensured by cross-annotator review in 438 man-hours of effort. For instance, the refinements on the portion of the dataset done by some member in the first pass is now reviewed and refined by another member of the team.

5. Benchmarking AnimalWeb

We extensively benchmark AnimalWeb for face alignment task. In addition, we demonstrate multi-task applications by demonstrating experimental results for face detection and fine-grained image recognition.

5.1. Animal Facial Point Localization

We select the state-of-the-art (SOTA) method in 2D human face alignment for evaluating AnimalWeb. Specifically, we take Hourglass (HG) deep learning based architecture; it has shown excellent results on a range of challenging 2D face alignment datasets [3, 32] and competitions [39].

Datasets and Evaluation Protocols. We use 300W-public, 300W-private, AFLW2000-3D, and COFW for comparison as they are the most challenging ones and are publicly available. 300W-public contains 3148 training images and 689 testing images. 300W-private comprises 600 images for testing only. We only use COFW for testing purposes; its testing set contains 507 images. Similarly, AFLW2000-3D is used for testing only after training on 300WLP dataset.

We use Normalized Mean Error (NME) as the face alignment evaluation metric,

$$\text{NME} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L}\left(\frac{\parallel xi^{'}(l) - xi^{g}(l) \parallel}{d_i}\right).$$

It calculates the Euclidean distance between the predicted and the ground truth point locations and normalizes by d_i. We choose ground truth face bounding box size as d_i, as other measures such as Interocular distance could be biased for profile faces [24]. In addition to NME, we report results using Cumulative Error Distribution (CED) curves, Area Under Curve (AUC) @0.08 (NME) error, and Failure Rate (FR) @0.08 (NME) error.

Training Details. For all our experiments, we use the settings described below to train HG networks both for human datasets and AnimalWeb. Note, these are similar settings as described in [32, 39] to obtain top performances on 2D face alignment datasets. We set the initial learning rate to 10^{-4} and used a mini-batch of 10. During the process, we divide the learning rate by 5, 2, and 2 at 30, 60, and 90 epochs, respectively, for training a total of 110 epochs. We also applied random augmentation: rotation (from -30^o to 30^o), color jittering, scale noise (from 0.75 to 1.25). All networks were trained using RMSprop [33].

Evaluation Settings. AnimalWeb is assessed under two different settings. The first randomly takes 80% images for training and the rest 20% for testing purposes from each species [3]. We call it '*Known species evaluation*' or so-called '*few-shot face alignment*' since during training the network sees examples from every species expected upon testing phase. The second setting randomly divides all species into 80% for training and 20% for testing. We term it as '*Unknown species evaluation*' or so-called '*zero-shot face Alignment*' (ZFA) as the species encountered in testing phase are not available during training. Unknown species evaluation is, perhaps, more akin to real-world settings than its counterpart. It is likely for a deployed facial behaviour monitoring system to experience some species that were unavailable at training. It is also more challenging than first as facial appearance of species during testing can be quite different to the ones available at training time.

Known Species Evaluation. Tab. 3 reveals comparison between AnimalWeb and various human face alignment benchmarks, when stacking 2 and 3 modules of HG network. Human face alignment results are shown both in

[3] For validation, we recommend using 10% data from the training set.

Datasets	9 pts.		68 pts.	
	HG-2	HG-3	HG-2	HG-3
300W(common)	1.21/84.8/0.18	1.19/85.0/0.00	1.26/84.1/0.00	1.25/84.2/0.00
300W(full)	1.42/82.1/0.14	1.40/82.4/0.00	1.41/82.2/0.00	1.40/82.3/0.00
300W(challenging)	2.28/71.4/0.00	2.25/71.7/0.00	2.03/74.5/0.00	2.01/74.8/0.00
300W(private)	2.26/72.2/0.66	2.31/72.4/1.16	1.82/77.5/0.50	1.77/77.8/0.16
AFLW2000-3D	3.27/60.8/3.27	3.23/61.3/2.75	2.73/66.5/0.50	2.71/66.9/0.55
COFW	3.43/60.0/3.74	3.26/61.3/3.55	2.66/67.2/1.97	2.60/68.2/1.57
AnimalWeb (Known)	5.22/46.8/16.4	5.12/47.4/16.3	-	-
AnimalWeb (Unknown)	6.14/41.5/22.0	5.96/42.9/20.7	-	-

Table 3: Accuracy comparison between the AnimalWeb and 6 different human face alignment benchmarks when stacking 2 and 3 modules of HG network. We show human face alignment results both in terms of 68 pts. and 9 pts. Format for each table entry is: NME error/AUC@0.08 (NME) error/FailureRate@0.08 (NME) error. All results are in %.

terms of 68 pts. and 9 pts. For fair comparison, the 9 pts. chosen on human faces are the same as for animal faces. Further, 9 pts. results correspond to the model trained with 9 pts. on human faces. We see a considerable gap (NME difference) between all the results for human face alignment datasets and AnimalWeb. For instance, the NME difference between COFW tested using HG-2 network is ~ 1 unit with AnimalWeb under the known species evaluation protocol. We observe a similar trend in the CED curves displayed in Fig. 9. Performance of COFW dataset, the most challenging among human faces, is 15% higher across the whole spectrum of pt-pt-error. Finally, we display some example fittings under known species evaluation settings in the first row of Fig. 10. We see that the existing art struggles under adverse in-the-wild situations exhibited in AnimalWeb.

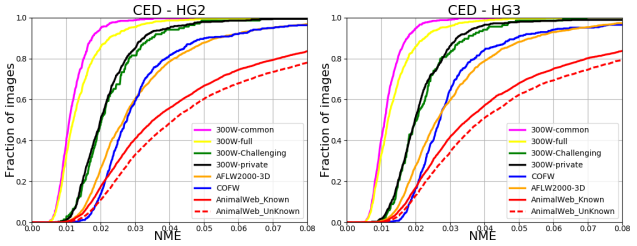

Figure 9: Comparison between AnimalWeb and popular face alignment datasets using HG-2&3 networks.

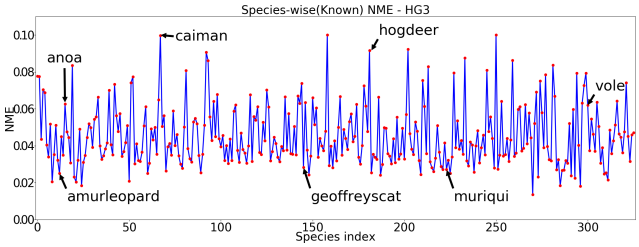

Figure 12: Specie-wise results for AnimalWeb under Known Species settings. Zoom-in for details.

Fig. 12 depicts species-wise testing results for AnimalWeb. For each species, we average results along the number of instances present in it. We observe poorer performance for some species compared to others. This is possibly due to large intra-species variations coupled with the scarcity of enough training instances relative to others. For instance, *hogdeer* species has only 20 training samples compared to *amurleopard* species populated with 91 training examples. Next, we report pose-wise results based on yaw angle in Tab. 4. We can observe that AnimalWeb is challenging for large poses. The performance drops as we move towards the either end of (shown) yaw angle spectrum from $[-45^o, 45^o]$ range. Further, Tab. 5 shows results under different face sizes. We observe room for improvement across a wide range of face sizes.

Unknown Species Evaluation. Here, we report results under unknown species settings. Note, we randomly choose 80% of the species for training and the rest 20% for testing. Tab. 3 draws comparison between unknown species settings and its counterpart. As expected, accuracy is lower for unknown case versus the known case. For example, HG-2 displays ~ 1 unit poor performance under unknown case in comparison to known. Animal faces display much larger inter-species variations between some species. For example, *adeliepenguins* and *giantpandas* whom face appearances are radically different (Fig. 10). Bottom row of Fig. 10 displays example fittings under this setting. We see that the fitting quality is low for frontal poses; the face appearance of species seen during training could be very different to ones testing species.

Low accuracy of existing methods under unknown species present opportunities for the development of 'zero-shot face alignment algorithms' that are robust to unseen facial appearance patterns. For instance, new methods that can better leverage similarities across seen species to perform satisfactorily under unknown species.

5.2. Animal Face Detection

We evaluate the performance of animal face detection using a Faster R-CNN [27] baseline. Our ground-truth is a tightly enclosed face bounding box for each animal face, that is obtained by fitting the annotated facial landmarks. We first evaluate our performance on the face localization task. We compare our dataset with one of the most challenging human face detection dataset WIDER Face [42] in terms of Precision-Recall curve (Fig. 11). Note that WIDER Face is a large-scale dataset with $393,703$ face instances in 32K images and introduces three protocols for evaluation namely 'easy', 'medium' and 'hard' with the increasing level of difficulty. The performance on our dataset lies close to that of medium curve of WIDER Face, which shows that there exists a reasonable margin of improvement for animal face detection. We also compute overall class-wise detec-

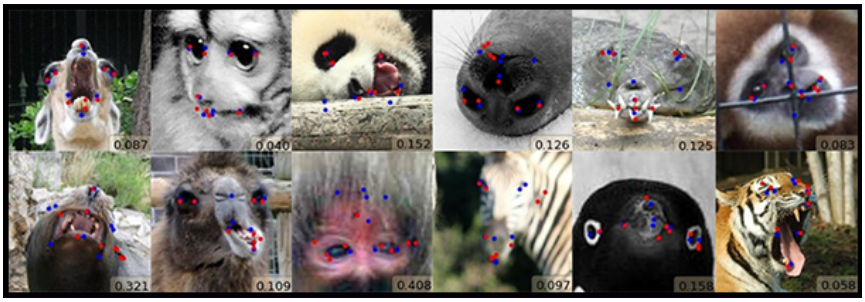

Figure 10: Example landmark fittings from AnimalWeb. Top row: fittings under known species evaluation. Bottom row: fittings under unknown species evaluation. Red points denote fittings results of HG-3 and blue points are the ground truths.

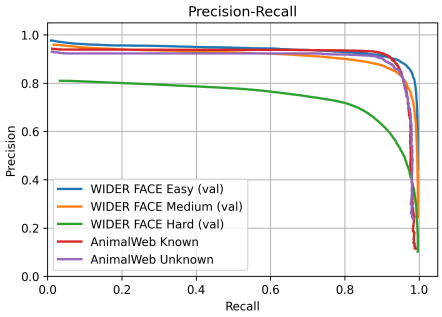

Figure 11: Precision-recall curve for AnimalWeb settings and WIDER Face datasets.

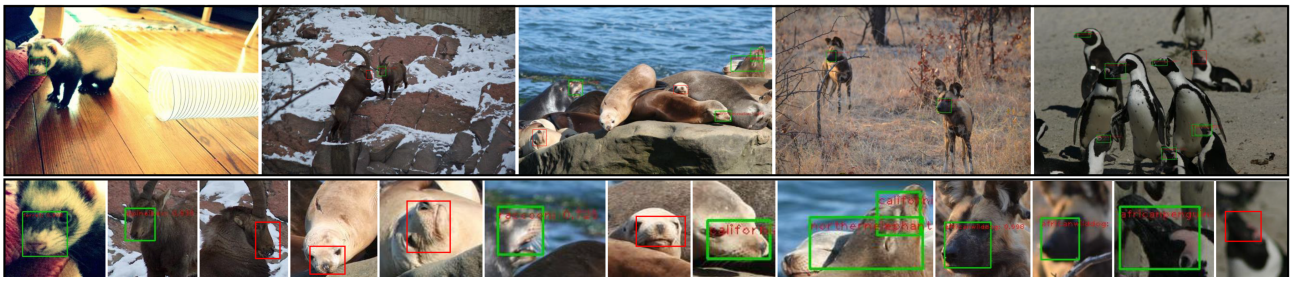

Figure 13: Example face detections from AnimalWeb. Green/red boxes denote true/missed detections from Faster-RCNN [27] baseline.

Yaw	-90^o	[-90^o,-45^o]	[-45^o,45^o]	[45^o,90^o]	90^o
Faces	584	993	1092	991	689
NME	6.75	5.02	3.31	4.99	6.94

Table 4: Pose-wise NME(%) based on yaw-angles with HG-3 under Known species settings of AnimalWeb.

Face size	[0,0.16]	[0.16,0.32]	[0.32,0.48]
Faces	3388	817	129
NME	5.29	4.41	4.73

Table 5: NME(%) w.r.t face size distribution with HG-3 under Known species settings of AnimalWeb. Face sizes are normalized by the corresponding image sizes.

tion scores where the Faster R-CNN model achieves a mAP of 0.727. Some qualitative examples of our animal face detector are shown in Fig. 13.

5.3. Fine-grained species recognition

Since our dataset is labeled with fine-grained species, one supplementary task of interest is the fine-grained classification. We evaluate the recognition performance on our dataset by applying Residual Networks [14] with varying depths (18, 34, 50 and 101). Results are reported in Tab. 6. We can observe a gradual boost in top-1 accuracy as the network capacity is increased. Our dataset shows a similar difficulty level in comparison to other fine-grained datasets of comparable scale, e.g., CUB-200-2011 [37] and Stanford Dogs [17] with 200 and 120 classes, respectively. A ResNet50 baseline on CUB-200 and Stanford Dogs achieve

Network	ResNet18	ResNet34	ResNet50	ResNet101
Accuracy	78.46	81.51	83.09	84.23

Table 6: Fine-grained recognition accuracy on AnimalWeb. Top-1 accuracies (in %) are reported using four ResNet variants [14].

an accuracy of 81.7% and 81.1% [31], while the same network achieves an accuracy of 83.09% on AnimalWeb.

6. Conclusion

We introduce a large-scale, hierarchical dataset, named AnimalWeb, of annotated animal faces. It features 22.4K faces from 350 diverse animal species while exploring 21 different orders. Each face is consistently annotated with 9 landmarks around key facial features. Benchmarking AnimalWeb under two novel settings for face alignment, employing current SOTA method, reveals its challenging nature. We observe that SOTA methods for human face alignment relatively underperform for animal faces. This highlights the need for specialized and robust algorithms to analyze animal faces. We also show the applications of the dataset for face detection and fine-grained recognition. Our results show that it is a promising experimental base for algorithmic advances.

Acknowledgments This work was supported by the EPSRC project EP/M02153X/1 Facial Deformable Models of Animals. Further, it uses data generated via the Zooniverse.org platform, funded by Google Global Impact Award and Alfred P. Sloan Foundation.

References

[1] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2930–2940, 2013. 2

[2] Alain Boissy, Arnaud Aubert, Lara Désiré, Lucile Greiveldinger, Eric Delval, Isabelle Veissier, et al. Cognitive sciences to relate ear postures to emotions in sheep. *Animal Welfare*, 20(1):47, 2011. 2

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 2, 6

[4] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520, 2013. 2, 3

[5] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 2

[6] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. In *European Conference on Computer Vision*, pages 484–498. Springer, 1998. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern recognition*, pages 248–255. Ieee, 2009. 5

[8] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *International Journal of Computer Vision*, pages 1–26, 2018. 2, 3

[9] Kathryn Finlayson, Jessica Frances Lampe, Sara Hintze, Hanno Würbel, and Luca Melotti. Facial indicators of positive emotions in rats. *PloS one*, 11(11):e0166446, 2016. 2

[10] Carole Fureix, Patrick Jego, Séverine Henry, Léa Lansade, and Martine Hausberger. Towards an ethological animal model of depression? a study on horses. *PLoS One*, 7(6):e39280, 2012. 2

[11] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015. 2, 3

[12] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 2, 3

[13] M.J. Guesgen, N.J. Beausoleil, M. Leach, E.O. Minot, M. Stewart, and K.J. Stafford. Coding and quantification of a facial expression for pain in lambs. *Behavioural Processes*, 132:49 – 56, 2016. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 770–778, 2016. 8

[15] László A Jeni, Sergey Tulyakov, Lijun Yin, Nicu Sebe, and Jeffrey F Cohn. The first 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision*, pages 511–520. Springer, 2016. 2

[16] Oliver Jesorsky, Klaus J Kirchberg, and Robert W Frischholz. Robust face detection using the hausdorff distance. In *International Conference on audio-and video-based biometric person authentication*, pages 90–95. Springer, 2001. 2

[17] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011. 8

[18] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international Conference on Computer Vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011. 2, 3

[19] T Kutzer, M Steilen, L Gygax, and B Wechsler. Habituation of dairy heifers to milking routine—effects on human avoidance distance, behavior, and cardiac activity during milking. *Journal of Dairy Science*, 98(8):5241–5251, 2015. 2

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[21] Iacopo Masi, Anh Tun Trn, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, pages 579–596. Springer, 2016. 2

[22] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. 1999. 2

[23] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer society Conference on Computer Vision and Pattern recognition (CVPR'05)*, volume 1, pages 947–954. IEEE, 2005. 2

[24] Deva Ramanan and Xiangxin Zhu. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern recognition*, pages 2879–2886. IEEE, 2012. 6

[25] Maheen Rashid, Xiuye Gu, and Yong Jae Lee. Interspecies knowledge transfer for facial keypoint detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6894–6903, 2017. 2, 3

[26] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014. 2

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 7, 8

[28] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision computing*, 47:3–18, 2016. 2, 3

[29] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 2, 3

[30] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 50–58, 2015. 2

[31] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–821, 2018. 8

[32] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris Metaxas. Quantized densely connected u-nets for efficient landmark localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 339–354, 2018. 6

[33] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. In *COURSERA: Neural networks for Machine learning*, page 4(2), 2012. 6

[34] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016. 2

[35] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, 2015. 2

[36] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018. 2

[37] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 8

[38] Dayong Wang, Charles Otto, and Anil K Jain. Face search at scale. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1122–1136, 2017. 2

[39] Pengfei Xiong, Guoqing Li, and Yuhang Sun. Combining local and global features for 3d face tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2529–2536, 2017. 2, 6

[40] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013. 2

[41] Heng Yang, Renqiao Zhang, and Peter Robinson. Human and sheep facial landmarks localisation by triplet interpolated features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016. 2, 3

[42] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 5525–5533, 2016. 7

[43] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3409–3417, 2016. 2

[44] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016. 2, 3