AnyFace: A Data-Centric Approach For Input-Agnostic Face Detection

Askat Kuzdeuov
Inst. of Smart Systems and AI
Nazarbayev University
Astana, Kazakhstan
askat.kuzdeuov@nu.edu.kz

Darina Koishigarina
Inst. of Smart Systems and AI
Nazarbayev University
Astana, Kazakhstan
darina.koishigarina@nu.edu.kz

Huseyin Atakan Varol
Dept. of Robotics and Mechatronics
Nazarbayev University
Astana, Kazakhstan
ahvarol@nu.edu.kz

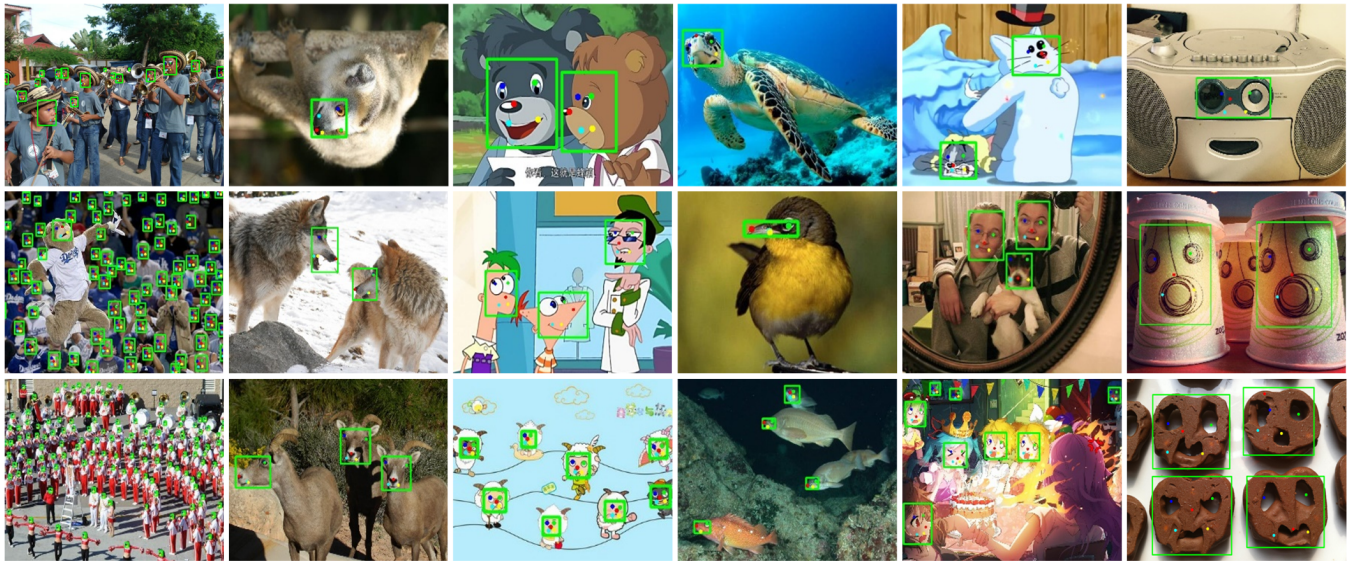

Fig. 1: Examples of detected faces and facial landmarks by the AnyFace model.

***Abstract*—Face detection is a mandatory step in many computer vision applications, such as face recognition, emotion recognition, age detection, virtual makeup, and vital sign monitoring. Thanks to advancements in deep learning and the introduction of annotated large-scale datasets, numerous applications have been developed for human faces. Recently, other domains, such as animals and cartoon characters, have started gaining attention but still lag far behind human faces. The biggest challenge is the limited number of annotated face datasets in these domains. The manual labeling of large-scale datasets is tedious and requires substantial human labor. In this regard, we present an input-agnostic face detector to ease the annotation of various face datasets. We propose a simple but effective data-centric approach instead of building a specific neural network architecture. Specifically, we trained a face detection model, YOLO5Face, on human, animal, and cartoon face datasets. The experiments show that the model can achieve accurate results in all domains. In addition, the model achieved decent results for animals and cartoon characters different from the ones in the training set. This implies that the model can extract agnostic facial features. We have made the source code and pre-trained models publicly available at https://github.com/IS2AI/AnyFace to stimulate research in these fields.**

***Index Terms*—face detection, input-agnostic models, data-centric AI, deep learning.**

I. Introduction

Over the past years, numerous computer vision applications have been developed for human faces, such as face recognition, facial expression recognition, virtual makeup, and vital sign monitoring. Although the field has achieved spectacular success for human faces, thanks to advances in deep learning and large-scale datasets, less attention is paid to other important domains, such as animals, cartoons, and artistic paintings. The main challenge is that there are only a limited number of annotated large-scale face datasets to train deep learning-based models.

Animal biometrics is one of the growing research fields concerning animal faces for animal identification, controlling animal diseases, production management, and ownership assignment. In particular, visual biometrics solutions are advantageous because previous methods of identifying animals have been invasive [1]. With the help of visual animal biometrics, it would be possible to develop monitoring systems for pets [2] in cities [3] paving the way to further research in animal re-identification and tracking. Animal faces can also be used to improve animal welfare by recognizing vital signs [4], [5].

Another domain in which faces and a plethora of face

variations are prevalent is cartoon faces. With the abundance of media with cartoon-style content, the demand for computer vision tools for solving various tasks is also rising. For instance, detecting faces and characters in comic books are useful for tagging comics for further analysis [6], [7]. Furthermore, face recognition for cartoon faces has its potential use in search engines [8], [9], in improving cartoon movie recommendation systems [10], and in detecting unauthorized use of copyrighted works [9]. Face recognition across various modalities has been implemented for multimedia facial analytics [11]. Similar to human faces, facial expression recognition for cartoon faces is used for parental control of cartoons for their children based on emotion [12], and in a professional environment, for animators to classify and label cartoon faces for future work [13].

Regardless of the application domain, face detection and facial landmark localization are essential steps in building many face applications. However, the existing face detection algorithms are domain-specific. Therefore, building a face detection model for a new domain requires the collection and annotation of new data. This process is time-consuming and requires human labor. In this regard, we propose an input-agnostic face detector to facilitate the annotation of large-scale face datasets. Instead of building a dedicated neural network architecture, we leverage the power of datasets from different domains (i.e., human, animal, cartoon) to train one of the existing face detection models. Our experiment results show that the model can learn agnostic facial features as it generalizes animals, cartoon characters, and artistic paintings unseen during the training and validation steps. To the best of our knowledge, it is the first input-agnostic face detection model. We have made the source code and pre-trained models publicly available to facilitate research in this field.

II. Related Work

The early face detection methods used feature extractors, such as Haar cascades [14] and Histogram of Oriented Gradients [15], to train classic machine learning algorithms. Nowadays, deep learning based end-to-end models (e.g., MTCNN [16], RetinaFace [17], YOLO5Face [18], TinaFace [19]) are dominating this field thanks to the advancements in training deep neural networks for image classification [20], [21], annotated large-scale face datasets (e.g., Wider Face [22], FDDB [23]), and modern graphical processing units (GPUs). However, a high degree of variability in scale, head pose, facial expressions, blurring, and illumination are still challenging problems. Most of these are related to traditional visual cameras that operate in the visible light spectrum. In this regard, new types of cameras, such as depth, thermal, and neuromorphic, can be employed to mitigate the disadvantages of the visual cameras. However, these cameras are not widely used compared to visual cameras. Therefore, a limited number of face detection models are available for these cameras. For instance, the YOLOv5 model was used to implement a thermal face detection model [24]. The authors collected 9,982 thermal images and manually annotated 16,509 faces to train the model. The dataset was collected in controlled and uncontrolled environments. Similarly, a face detection and tracking algorithm based on the dynamics of eye blinks for event-based cameras was proposed in [25]. The authors collected a dataset of 50 recordings in indoor and outdoor environments to evaluate the algorithm.

There are only a few works in the literature about animal face detection. For instance, a RetinaNet [26] object detection model was used to perform multi-view cattle face detection in housing farms [27]. The authors collected and manually annotated 3,000 images of 85 cattle to train the model. A Faster R-CNN [28] object detection model was trained to detect small-scale dog faces [29]. The Viola-Jones object detection framework was used to develop a sheep face detector [30]. A pretrained version of YOLOv3 [31] was modified and tuned for mouse face detection [32]. The authors collected and annotated 2,222 images to train the model. Similarly, a YOLOv3 was trained to detect sheep faces [33]. The authors collected and labeled 1,958 sheep face images. Recently, a large-scale animal face dataset, AnimalWeb, was introduced [34]. The dataset contains 22,451 faces of 334 animal species captured in the wild condition, with each face annotated with nine facial landmarks. As a baseline, the authors trained a Faster R-CNN object detection model for animal face detection.

A Faster R-CNN object detection model was adapted for comic character face detection [35]. For this purpose, the authors constructed and annotated a new dataset consisting of 3,375 comic pages. An MTCNN [36] face detection model was employed to develop a cartoon face detection model [37]. The authors used the IIIT-CFW dataset [38], which contains 8,928 annotated images of cartoon faces of 100 global public figures. A Manga FaceNet neural network architecture was proposed to detect manga characters' faces [39]. The model was trained on manually annotated images of 3,760 frontal and 1,110 side-view faces. A large-scale annotated cartoon face dataset, iCartoonFace [40], was developed for cartoon face detection and face recognition tasks. As a baseline, the authors trained a RetinaFace model on 50,000 images (91,163 faces) and tested it on 10,000 images (18,647 faces).

As can be seen, the existing face detection methods are domain-specific. The main challenge of developing animal face detection models is the limited number of annotated large-scale datasets. Moreover, there are more than a million species inhabiting the planet. Collecting and manually annotating a face dataset for each species is impractical. The same applies to cartoon faces. There are a large number of different characters, and this number is growing rapidly. Thus, there is a clear need for an input-agnostic face detection model which can be applied to many domains.

III. Method

We propose a simple but effective data-centric method to develop an input-agnostic face detector. We assume that the existing deep learning-based object/face detection models are advanced enough to learn general facial features. Therefore, we focus on experimenting with face datasets from various domains. Our idea is to provide the model with the "right"

composition of datasets so that it learns to extract agnostic facial features.

We used faces of humans (visual and thermal), animals, and cartoon characters to develop the input-agnostic model. We employed these domains for two reasons. The first reason is that each domain provides an annotated large-scale face dataset (see Table I). The second reason is that they represent different facial features. To visually illustrate this, we extracted facial embeddings from randomly selected face images using a ResNet-50 model (pre-trained on the ImageNet dataset). Then, we projected the high-dimensional facial features into two dimensions using the t-SNE algorithm [41]. The t-SNE plot shows some hidden cues about each domain (see Fig. 2). For instance, animal faces are clustered by species. This means that each species can be considered a separate domain. Human faces are divided into visual and thermal domains. Human faces are also clustered by appearance. In the thermal domain, human faces with glasses are separated from those without glasses.

A. Datasets

Wider Face [22] is a benchmark face dataset for human faces in the visual domain. The dataset contains 32,203 images and 393,703 labeled faces. AnimalWeb is a large-scale hierarchical dataset of annotated animal faces [34]. The dataset contains 22,451 annotated faces of 350 various species from 21 animal order taxonomies. The animal faces were annotated with nine key-point facial landmarks. iCartoonFace is a benchmark dataset for face detection and recognition of cartoon characters [40]. The dataset consists of 389,678 images of 5,013 cartoon characters. The dataset contains 60,000 images (109,807 annotated faces) for the face detection task. TFW dataset [24] comprises 9,982 thermal images with 16,509 annotated faces. The dataset was collected in controlled indoor and uncontrolled outdoor conditions.

We split each dataset into training, validation, and test sets. The training set was used to train the model, while the validation set was used to tune the hyperparameters. The Wider Face dataset contains 12,880 images (159,424 faces) in the training set, 3,226 images (39,798 faces) in the validation set, and 16,097 images (194,571 faces) in the test set. Labels are available for the training and validation sets. However, results on the test set are obtained by sending the predictions to the authors of the dataset. The TFW dataset contains 6,558 images (10,801 faces) in the training set, 764 images (1,081 faces) in the validation set, and 2,660 images (4,627 faces) in the test set. Annotations are available for all sets. The AnimalWeb dataset comes without split into training, validation, and test sets. Therefore, we split the dataset such that each set contains different animal species. As a result, the training set consisted of 13,265 images (15,730 faces), the validation set had 2,062 images (2,374 faces), and the test set contained 3,752 images (4,347 faces). Also, the facial bounding boxes are not available in the AnimalWeb dataset. Therefore, we generated bounding boxes using the coordinates of facial landmarks. The iCartoonFace dataset provided 50,000 images (91,160 faces) for training and 10,000 images (18,647 faces) for testing. We used 45,000 images (81,579 faces) for model training and the remaining 5,000 images (9,581 faces) for model validation. In total, our training set consisted of 77,703 images (267,534 faces), the validation set had 11,052 images (52,744 faces), and the test set contained 32,509 images (222,192 faces).

Dataset	Images	Faces
Wider Face [22]	32,203	393,703
AnimalWeb [34]	19,079	22,451
iCartoonFace [40]	60,000	109,807
TFW [24]	9,982	16,509

TABLE I: Statistics for the datasets

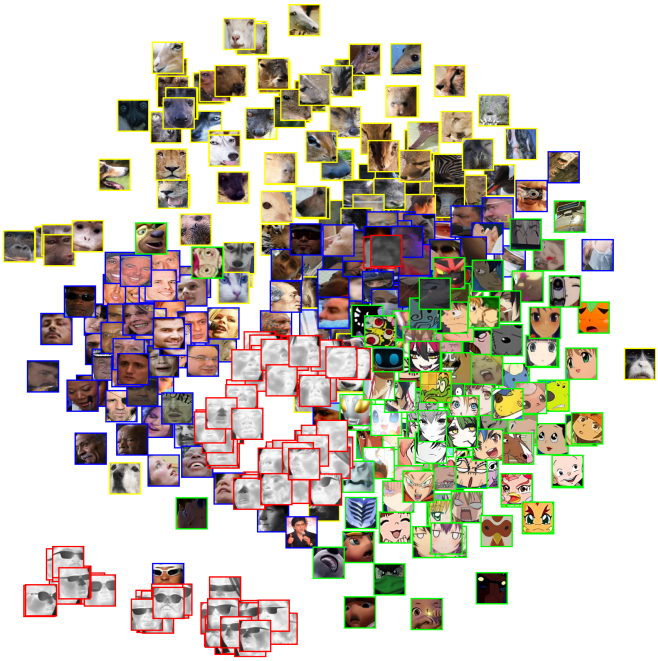

Fig. 2: t-SNE visualisation of facial embeddings generated using the ResNet-50 model.

B. Face Detection Model

We used a YOLO5Face [18] face detection model. It is based on the YOLOv5 object detector [42], but modified specifically for face detection. The model has an additional regression output for five facial landmarks. In addition, the model offers nano, small, medium, and large architectures. As a backbone network, the nano models use ShuffleNetv2, while others use CSPNet. In addition, each model can be trained with an additional P6 output block to improve the detection of large faces. For more information, an interested reader is referred to the original paper.

In our case, we used the default hyperparameters to train the models. Namely, a stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01 was used to minimize the loss functions. The models were trained for 350 epochs with a batch size of 32. We used default augmentation

Model	Params (M)	Wider Face (visual)			TFW (thermal)		AnimalWeb	iCartoonFace
		Easy	Medium	Hard	Indoor	Outdoor		
YOLOv5n	1.72	92.1	89.6	76.7	**100**	98.16	94.66	86.04
YOLOv5n6	2.54	93.5	90.7	76.3	**100**	98.36	95.17	88.13
YOLOv5s	7.07	93.8	91.7	79.9	**100**	98.56	95.25	87.60
YOLOv5s6	12.38	94.5	92.2	79.7	**100**	98.72	95.70	89.24
YOLOv5m	21.06	95.2	93.4	83.2	**100**	98.79	95.80	89.85
YOLOv5m6	35.48	95.9	93.8	82.8	**100**	99.09	**96.28**	90.24
YOLOv5l	46.62	95.8	94.2	**84.9**	**100**	99.19	96.17	90.31
YOLOv5l6	76.67	**96.1**	**94.4**	84.1	**100**	**99.20**	96.26	**90.61**

TABLE II: Average precision (AP) scores of YOLO5Face models on the validation set ($IoU = 0.5$).

settings, such as image translation, scaling, shearing, horizontal flipping, and mosaic. The nano, small, and medium models were trained on a single A100-SXM4-40GB GPU while the training of large models was distributed on two GPUs.

C. Experiments

We conducted three experiments. In Experiment 1, we combined the Wider Face, TFW, AnimalWeb, and iCartoonFace datasets and trained the YOLO5Face models. In total, we trained and evaluated eight models: nano, nano-P6, small, small-P6, medium, medium-P6, large, and large-P6. Our goal was to study the generalizability of the models of various sizes to the four domains. As a result of this experiment, we also selected the most accurate model for the next experiment.

In Experiment 2, to find an optimal combination that would provide the most accurate results, the best model from the previous experiment was trained on different combinations of the domains. As can be seen in Table I, Wider Face was the largest of the datasets considered. Thus, it was supplemented with different combinations of other domains.

In Experiment 3, we tested the model trained in the previous experiment on previously unseen images of animals and cartoon characters, as well as artistic paintings. The experiment aimed to verify whether the model could extract general facial features to become an input-agnostic model. In addition, we compared our results on Wider Face, TFW, AnimalWeb, and iCartoonFace with the results published in the literature.

IV. Results and Discussion

We used the average precision (AP) metric at an intersection over the union (IoU) threshold of 0.5 to evaluate the accuracy of the predicted facial bounding boxes. We resized the images in the validation and test sets before evaluation. The longest side was set to 640 pixels, while the shortest side was resized, with the original aspect ratio maintained. In the Experiments 1 and 2, to find the best model, we evaluated the models only on the validation set. We then used the test set to evaluate the best-performing model.

A. Training all Models on all Data

In this experiment, we trained all architectures of the YOLO5Face model on the combined dataset. The results on the validation set are shown in Table II. The nano model with 1.72 M parameters yielded scores of 92.1%, 89.6%, and 76.7% AP on the easy, medium, and hard sets of the Wider Face dataset, respectively. The results on the TFW dataset were 100% AP on the indoor set and 98.16% AP on the outdoor set. The model showed 94.66% AP on the AnimalWeb dataset and 86.04% AP on the iCartoonFace dataset.

The nano model with the P6 output block, YOLOv5n6, improved accuracy on the easy and medium sets by 1.4% and 1.1%, respectively. However, the accuracy on the hard set dropped by 0.4%, as it contained many small faces, while the P6 output block focused on improving the detection of large faces. In addition, the YOLOv5n6 increased the AP scores for the outdoor set of TFW, AnimalWeb, and iCartoonFace by 0.2%, 0.51%, and 2.09%. Similar trends were observed for the small, medium, and large models.

The results also showed that model capacity increased accuracy across all domains. The large model with the P6 output block (YOLOv5l6) produced 96.1% AP, 94.4% AP, and 84.1% AP on the easy, medium, and hard sets, respectively. For TFW, the model achieved 100% AP on the indoor set and 99.2% AP on the outdoor set. The model showed 96.26% AP on AnimalWeb and 90.61% AP on iCartoonFace. We used the YOLOv5l6 model in our further experiments.

B. Training the Best Model on Different Subsets

In this part, we trained the YOLOv5l6 model by merging Wider Face with the combinations of the other datasets. The results for different cases are given in Table III.

1) Wider Face.: We trained the model only on the Wider Face dataset. In this case, the model achieved 96.2% AP, 94.5% AP, and 84.9% AP on the easy, medium, and hard sets, respectively. Although the model did not see the TFW dataset during training, it achieved 100% AP on the indoor set and 94.13% AP on the outdoor set. However, the model failed on the AnimalWeb dataset, yielding only 3.72% AP. This seems to point toward a large discrepancy between the human and animal face domains. The model also produced a low accuracy score on the iCartoonFace dataset (20.9%).

2) Wider Face + TFW.: In case of including the TFW dataset, the AP on the easy, medium, and hard sets improved by 0.1%, 0.2%, and 0.4%, respectively. For TFW, the AP on the outdoor set increased by a significant 4.39%. The AP score on AnimalWeb and iCartoonFace also improved—although insignificantly, since Wider Face and TFW are human face datasets.

3) Wider Face + AnimalWeb.: In this case, the AP on the AnimalWeb dataset increased significantly, reaching 96.54%.

Case	Training dataset				Wider Face (visual)			TFW (thermal)		AnimalWeb	iCartoonFace
	WF	TFW	AW	iCF	Easy	Medium	Hard	Indoor	Outdoor		
1	+	–	–	–	96.20	94.50	84.90	**100.00**	94.13	3.72	20.90
2	+	+	–	–	96.30	94.70	85.30	**100.00**	98.52	4.04	24.84
3	+	–	+	–	96.20	94.70	**85.60**	**100.00**	95.48	**96.54**	27.44
4	+	–	–	+	**96.50**	**94.80**	84.80	**100.00**	97.16	51.92	90.67
5	+	+	+	–	96.20	94.70	85.30	**100.00**	98.81	96.53	28.46
6	+	+	–	+	96.40	94.60	84.70	**100.00**	99.01	50.63	90.47
7	+	–	+	+	96.10	94.50	84.70	**100.00**	97.31	96.40	**91.01**
8	+	+	+	+	96.10	94.40	84.10	**100.00**	**99.20**	96.26	90.61

TABLE III: Average precision (AP) scores of YOLOv5l6 model on the validation set ($IoU = 0.5$). The model was trained on different combinations of domains (WF: Wider Face, AW: AnimalWeb, iCF: iCartoonFace.)

The accuracy score on Wider Face, TFW, and iCartoonFace also improved, compared to Case 1. However, accuracy on the iCartoonFace dataset was still low (27.44%). This suggests that a large number of human and animal faces in a dataset may not be a sufficient condition for achieving a good result on cartoon faces.

4) Wider Face + iCartoonFace.: The use of cartoon faces improved the AP up to 90.67% on the iCartoonFace dataset. Compared to Case 1, the AP on the easy, medium, and hard sets improved by 0.3%, 0.2%, and 0.7%, respectively. The AP also improved by 3.3% for the outdoor set of TFW. Most surprisingly, the model achieved 51.92% AP on the AnimalWeb dataset, which was significantly higher than in Cases 1 and 2. The reason might be that the iCartoonFace dataset contains animal characters.

5) Wider Face + TFW + AnimalWeb.: The model achieved 96.2%, 94.7%, and 85.3% for the easy, medium, and hard sets, respectively. For TFW, it yielded 98.81% AP for the outdoor data. The AP on AnimalWeb was 96.53%. These improvements were negligible compared to Cases 2 and 3. Although we combined three datasets, the AP on the iCartoonFace dataset was only 28.46%.

6) Wider Face + TFW + iCartoonFace.: In this scenario, the results on the Wider Face were similar to those of the previous cases (96.40%, 94.60%, and 84.70%). However, the AP on the outdoor set of TFW improved noticeably (99.01%) compared to the above cases. The AP for iCartoonFace was 90.47% while the AP for the AnimalWeb dataset is only 50.63%.

7) Wider Face + AnimalWeb + iCartoonFace.: In the case of excluding TFW from the training set, the AP scores on the outdoor set decreased from 99.01% to 97.31%. The AP for Wider Face and AnimalWeb was similar to the APs in the previous cases. The accuracy on cartoon faces witnessed a slight increase to 91.01%.

8) Wider Face + TFW + AnimalWeb + iCartoonFace.: This case illustrates the necessity of using all domains to obtain accurate results in each domain. It is also indicative of the different facial features that human, animal, and cartoon faces represent. In further experiments, we used this YOLOv5l6 model trained on the combined datasets. To avoid confusion with the original model, it will hereafter be referred to as AnyFace.

Wider Face	Easy	95.4
	Medium	93.9
	Hard	84.0
TFW	Indoor	100
	Outdoor	99.47
AnimalWeb		93.59
iCartoonFace		91.65

TABLE IV: Average precision scores of the AnyFace model on the test set ($IoU = 0.5$).

C. Results on the Test Set

The results of the AnyFace model on the test set are in Table IV. The model achieved 95.4%, 93.9%, and 84.0% AP scores on the easy, medium, and hard sets of Wider Face, respectively. For TFW, the model showed 100% and 99.47% AP scores on the indoor and outdoor sets, respectively. The model also produced 93.59% AP for AnimalWeb and 91.65% AP for iCartoonFace. The results were similar to the validation set results, which means that the model did not overfit.

The results of the other human face detection models on the test set of the Wider Face dataset are shown in Fig. 3. Usually, authors augment the test set to obtain multiple predictions per image. The predictions are then fused to obtain the final prediction. The original YOLO5Face model [18] showed 94.9%, 94.3%, and 90.1% on the easy, medium, and hard sets, respectively, as shown in Fig. 3. We used the default augmentation settings used in the YOLO5Face model [18]. The scores on the easy, medium and hard sets were 95.2%, 94.7%, and 90.5%, respectively. There are several models with better results as shown in Fig. 3, but the main advantage of our model is its ability to detect faces from various domains.

Regarding the TFW dataset, the authors reported 100% for the indoor set and 97.2% AP for the outdoor set as their best results [24]. In our case, the AnyFace model achieved 100% for the indoor set and 99.2% for the outdoor set without test set augmentation. For iCartoonFace, the authors obtained 92.4% AP using the RetinaFace model [40]. In our case, the AnyFace model showed 90.61% AP. The difference could be attributed to the difference in models. As we mentioned earlier, the AnimalWeb dataset comes without being split into training, validation, and test sets. The authors used 80% of the data to train the Faster-RCNN model for the face detection task. The model achieved a mean AP of 63.6% on the 20% of dataset. In our case, the model showed 96.26% AP.

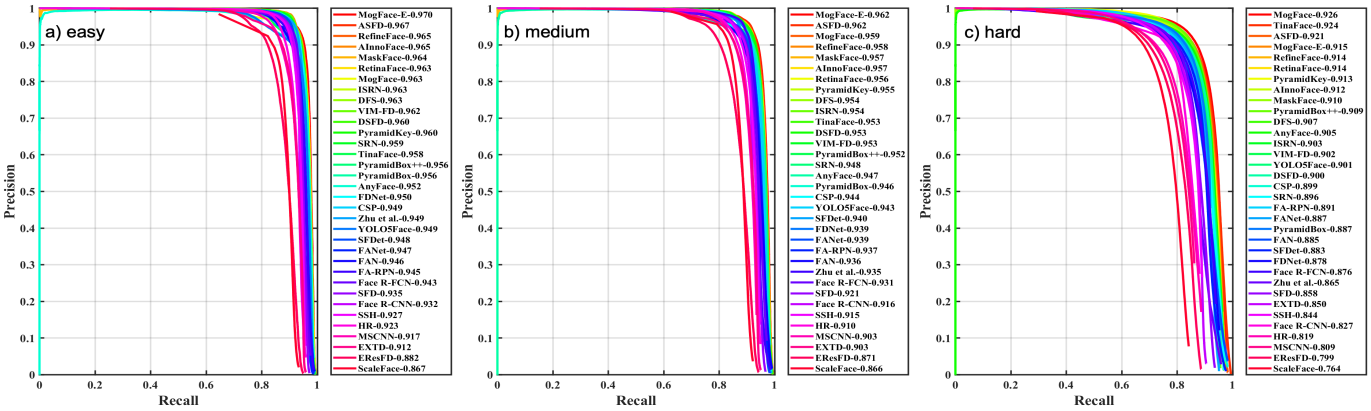

Fig. 3: Comparison of the AnyFace model with the other face detection models on the test set of the Wider Face dataset.

D. Testing on External Datasets

We tested the AnyFace model on external datasets to evaluate its performance on previously unseen images of animals, cartoon characters, and artistic paintings. The Oxford-IIIT PET dataset [43] contains images of 37 different breeds of cats and dogs with annotations for the head region. The CUB-200-2011 dataset [44] contains 11,788 images of 200 bird species with annotations for 15-part locations. The Labeled Fishes in the Wild dataset [45] contains images of different fish species. Annotations for facial bounding boxes are not available. The Sea Turtle Face Detection [46] dataset contains 2,000 images of annotated turtle faces. The Artistic-Faces dataset [47] consists of 160 artistic portraits of 16 different artists with annotated facial landmarks. The MetFaces dataset [48] contains 1,336 artistic images without annotations. The Anime Faces dataset [49] contains 6,641 images of annotated anime faces. The Tom & Jerry Detection dataset [50] consists of 343 images of the famous cat and mouse characters with annotated facial bounding boxes.

These datasets vary in size, and only a few have facial bounding boxes. Therefore, we randomly selected 160 images from each dataset to annotate with the AnyFace model. To evaluate the annotations, we manually checked the annotated bounding boxes and registered true positive, false positive, and false negative samples. We then computed the precision, recall, and accuracy scores for each dataset. The results are listed in Table V. For comparison, we included the results for the YOLO5Face model [18] trained only on the Wider Face dataset.

The Oxford-IIIT PET, CUB-200-2011, Fishes in the Wild, and Sea Turtle Faces datasets contain images of real animals. The AnimalWeb dataset, used to train the AnyFace model, has animals resembling cats and dogs. However, it does not have fish, birds, and turtles. These species were unknown to the model. The YOLO5Face model achieved 91.67% of precision, 34.38% of recall, and 33.33% of accuracy on the Oxford-IIIT PET dataset. However, the model failed on bird, fish, and turtle faces. In contrast, the AnyFace model showed 97.56% of precision, 100.0% of recall, and 97.56% of accuracy on the Oxford-IIIT PET dataset. Moreover, it yielded 88.17% of precision, 93.13% of recall, and 82.78% accuracy on the CUB-200-2011 dataset. It also achieved 97.95% of precision, 89.38% of recall, and 87.73% of accuracy on the Fishes in the Wild dataset. For the Sea Turtle Face dataset, the model showed 77.58% of precision, 76.19% of recall, and 62.44% of accuracy. These results demonstrate that the AnyFace model can detect previously unseen animal faces.

The Artistic-Faces and MetFaces datasets contain images of artistic paintings with human faces. The YOLO5Face model produced 100.0% of precision, 98.13% of recall, and 98.13% of accuracy in the Artistic-Faces dataset. Similarly, it achieved 99.5% of precision, 97.55% of recall, and 97.07% of accuracy on the MetFaces dataset. Such high performance can be attributed to the fact that the model was trained on the Wider Face dataset. In comparison, the AnyFace model yielded 97.56% of precision, 100.0% of recall, and 97.56% of accuracy on the Artistic-Faces dataset. On the MetFaces dataset, the model achieved 99.02% of precision, 99.02% of recall, and 98.06% of accuracy. These results illustrate that the AnyFace model can accurately detect human faces in different domains.

Anime Faces and Tom & Jerry are cartoon face datasets. The YOLO5Face model achieved 96.18% of precision, 57.27% of recall, and 56% of accuracy on the Anime Faces dataset. However, the model was unable to detect a single face in the Tom & Jerry dataset. In contrast, the AnyFace model produced 91.85% of precision, 97.27% of recall, and 89.54% of accuracy on the Anime Faces dataset. For the Tom & Jerry dataset, the model showed 86.53% of precision, 98.82% of recall, and 85.64% accuracy. The results show that the AnyFace can detect the faces of unknown cartoon characters.

In Fig. 1, we provide visual examples of detected bounding boxes and facial landmarks for images from various domains, illustrating that the AnyFace model can detect faces of different scales, poses, illuminations, and expressions. Remarkably, the model also recognizes objects that superficially resemble a face (the last column in Fig. 1). In psychology, this phenomenon is called face pareidolia. It is a compelling illusion of seeing fake faces in everyday objects and is driven by a face-

Dataset	YOLO5Face			AnyFace (our model)		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy
Oxford-IIIT PET	91.67	34.38	33.33	97.56	100.00	97.56
CUB-200-2011	0.00	0.00	0.00	88.17	93.13	82.78
Fishes in the Wild	0.00	0.00	0.00	97.95	89.38	87.73
Sea Turtle Faces	6.67	0.60	0.55	77.58	76.19	62.44
Artistic-Faces	100.00	98.13	98.13	97.56	100.00	97.56
MetFaces	99.50	97.55	97.07	99.02	99.02	98.06
Anime Faces	96.18	57.27	56.00	91.85	97.27	89.54
Tom & Jerry Detection	0.00	0.00	0.00	86.53	98.82	85.64

TABLE V: Precision, recall, and accuracy of the AnyFace and YOLO5Face models with a confidence threshold of 0.02 on external datasets

detection mechanism that we share with other species [51]. The AnyFace model could experience the same phenomenon because of cartoon faces in the training set.

V. Conclusion

We introduce a simple but effective method of building an input-agnostic face detector to facilitate the annotation process. Instead of developing a specific neural network architecture, we exploit the power of various facial datasets, such as human faces in the visual and thermal domain, animal faces, and cartoon faces. We employed the YOLO5Face face detection model, which provides nano, small, medium, and large models. We trained the models on the Wider Face, TFW, AnimalWeb, and iCartoonFace datasets. We have made the source code and pre-trained models publicly available to promote research in this field.

The experiment results in Table II illustrate that increasing the capacity of models yields more accurate results for all datasets. The limitation is that an increase in the capacity leads to an increase in the inference time. We chose the most accurate model, YOLOv5l6, and named it AnyFace for further experiments. We experimented with the AnyFace model by training it on different combinations of domains. The experiment results in Table III demonstrate that accurate results are obtained only when all domains are used to train the model. The results also prove that the datasets under consideration represent different facial features. The results on the test sets in Table IV show that the model does not overfit the training data.

We evaluated the AnyFace model on external datasets to test its performance on previously unseen images of animals and cartoon characters, as well as artistic paintings. The AnimalWeb dataset, used in the training set, contains images of various cat and dog species. However, the dataset does not have species of birds, fishes, and turtles. As a result, the AnyFace model 'sees' these animals for the first time. That said, the accuracy, precision, and recall scores of the model suggest that it can be generalized to these datasets. In addition, the AnyFace model achieved accurate results on artistic, anime, and cartoon faces. The YOLO5Face model was able to achieve accurate results on human-like faces (i.e., art, anime), but failed to detect animal faces.

Our work has a few limitations. For instance, the training set in imbalanced. In our future work, we will balance the dataset using style-transfer-based augmentation methods. Also, we noticed that the model outputs more false-positive predictions for sea animals than for land animals. The reason for this could be that underwater environments are much more challenging than terrestrial environments. Moreover, marine animals were not presented in the training set. We will test the model more on sea animals in our future work.

References

[1] S. Kumar, S. Tiwari, and S. K. Singh, "Face recognition of cattle: Can it be done?" *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, vol. 86, no. 2, p. 137–148, Jun 2016.

[2] S. Kumar and S. K. Singh, "Monitoring of pet animal in smart cities using animal biometrics," *Future Generation Computer Systems*, vol. 83, p. 553–563, Jun 2018.

[3] P. C. Ravoor and S. T.S.B., "Deep learning methods for multi-species animal re-identification and tracking – a survey," *Computer Science Review*, vol. 38, p. 100289, Nov 2020.

[4] P. H. Andersen, S. Broomé, M. Rashid, J. Lundblad, K. Ask *et al.*, "Towards machine recognition of facial expressions of pain in horses," *Animals*, vol. 11, no. 6, p. 1643, Jun 2021.

[5] M. F. Hansen, E. M. Baxter, K. M. D. Rutherford, A. Futro, M. L. Smith, and L. N. Smith, "Towards facial expression recognition for on-farm welfare assessment in pigs," *Agriculture*, vol. 11, no. 9, p. 847, Sep 2021.

[6] N.-V. Nguyen, C. Rigaud, and J.-C. Burie, "Comic characters detection using deep learning," in *Proceedings of the IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Kyoto: IEEE, Nov 2017, p. 41–46. [Online]. Available: http://ieeexplore.ieee.org/document/8270235/

[7] W.-T. Chu and W.-W. Li, "Manga FaceNet: Face detection in manga based on deep neural network," in *Proc. of the ACM on International Conference on Multimedia Retrieval*. Bucharest Romania: ACM, Jun 2017, p. 412–415. [Online]. Available: https://dl.acm.org/doi/10.1145/3078971.3079031

[8] S. Jha, N. Agarwal, and S. Agarwal, "Bringing cartoons to life: Towards improved cartoon face detection and recognition systems," *arXiv preprint arXiv:1804.01753*, 2018. [Online]. Available: https://arxiv.org/abs/1804.01753

[9] K. Takayama, H. Johan, and T. Nishita, "Face detection and face recognition of cartoon characters using feature extraction," *Image, Electronics and Visual Computing Workshop*, p. 5, 2012.

[10] Y. Li, L. Lao, Z. Cui, S. Shan, and J. Yang, "Graph jigsaw learning for cartoon face recognition," *IEEE Transactions on Image Processing*, vol. 31, p. 3961–3972, 2022.

[11] W. Zheng, L. Yan, F.-Y. Wang, and C. Gou, "Learning from the past: Meta-continual learning with knowledge embedding for jointly sketch, cartoon, and caricature face recognition," in *Proceedings of the ACM International Conference on Multimedia*. Seattle WA USA: ACM, Oct 2020, p. 736–743. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413892

[12] N. Jain, V. Gupta, S. Shubham, A. Madan, A. Chaudhary, and K. C. Santosh, "Understanding cartoon emotion using integrated deep neural network on large dataset," *Neural Computing and Applications*, p. 21481–21501, Apr 2021. [Online]. Available: https://link.springer.com/10.1007/s00521-021-06003-9

[13] Q. Cao, W. Zhang, and Y. Zhu, "Deep learning-based classification of the polar emotions of "moe"-style cartoon pictures," *Tsinghua Science and Technology*, vol. 26, no. 3, p. 275–286, Jun 2021.

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2001, pp. I–I.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.

[16] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[17] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5202–5211.

[18] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: Why reinventing a face detector," *arXiv preprint arXiv:2105.12931*, 2021.

[19] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, "TinaFace: Strong but simple baseline for face detection," *ArXiv*, vol. abs/2011.13183, 2020.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[22] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5525–5533.

[23] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[24] A. Kuzdeuov, D. Aubakirova, D. Koishigarina, and H. A. Varol, "TFW: Annotated thermal faces in the wild dataset," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2084–2094, 2022.

[25] G. Lenz, S.-H. Ieng, and R. Benosman, "Event-based face detection and tracking using the dynamics of eye blinks," *Frontiers in Neuroscience*, vol. 14, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2020.00587

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[27] B. Xu, W. Wang, L. Guo, G. Chen, Y. Wang, W. Zhang, and Y. Li, "Evaluation of deep learning for automatic multi-view face detection in cattle," *Agriculture*, vol. 11, no. 11, 2021. [Online]. Available: https://www.mdpi.com/2077-0472/11/11/1062

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[29] L. Mu, Z. Shen, J. Liu, and J. Gao, "Small scale dog face detection using improved Faster RCNN," in *Proc. of the International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*, 2022, pp. 573–579.

[30] Y. Lu, M. Mahmoud, and P. Robinson, "Estimating sheep pain level using facial action unit detection," in *Proc. of the IEEE International Conference on Automatic Face & Gesture Recognition*, 2017, pp. 394–399.

[31] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018. [Online]. Available: https://arxiv.org/abs/1804.02767

[32] A. Vidal, S. Jha, S. Hassler, T. Price, and C. Busso, "Face detection and grimace scale prediction of white furred mice," *Machine Learning with Applications*, vol. 8, p. 100312, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827022000330

[33] S. Song, T. Liu, H. Wang, B. Hasi, C. Yuan, F. Gao, and H. Shi, "Using pruning-based YOLOv3 deep learning algorithm for accurate detection of sheep face," *Animals*, vol. 12, no. 11, 2022. [Online]. Available: https://www.mdpi.com/2076-2615/12/11/1465

[34] M. H. Khan, J. McDonagh, S. Khan, M. Shahabuddin, A. Arora, F. S. Khan *et al.*, "AnimalWeb: A large-scale hierarchical dataset of annotated animal faces," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[35] X. Qin, Y. Zhou, Z. He, Y. Wang, and Z. Tang, "A Faster R-CNN based method for comic characters face detection," in *Proc. of the IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 1074–1080.

[36] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[37] S. Jha, N. Agarwal, and S. Agarwal, "Towards improved cartoon face detection and recognition systems," *ArXiv*, vol. abs/1804.01753, 2018.

[38] A. Mishra, S. N. Rai, A. Mishra, and C. V. Jawahar, "IIIT-CFW: A benchmark database of cartoon faces in the wild," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2016, pp. 35–47.

[39] W.-T. Chu and W.-W. Li, "Manga face detection based on deep neural networks fusing global and local information," *Pattern Recognition*, vol. 86, pp. 62–72, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320318303066

[40] Y. Zheng, Y. Zhao, M. Ren, H. Yan, X. Lu, J. Liu, and J. Li, "Cartoon face recognition: A benchmark dataset," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 2264–2272.

[41] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[42] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon *et al.*, "ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference," Feb. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6222936

[43] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3498–3505.

[44] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[45] G. Cutter, K. Stierhoff, and J. Zeng, "Automated detection of rockfish in unconstrained underwater videos using Haar cascades and a new image dataset: Labeled fishes in the wild," in *Proc. of the IEEE Winter Applications and Computer Vision Workshops*, 2015, pp. 57–62.

[46] S. Ghose, "Sea turtle face detection dataset," 2021, last accessed on 2022-06-25: "https://www.kaggle.com/datasets/smaranjitghose/sea-turtle-face-detection".

[47] J. Yaniv, Y. Newman, and A. Shamir, "The face of art: Landmark detection and geometric style in portraits," in *ACM Transactions on Graphics (Proceedings SIGGRAPH)*, vol. 38, 2019, pp. 60:1–60:15.

[48] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 104–12 114. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf

[49] X. Zhou, "Anime faces," 2018, last accessed on 2022-06-25: "https://github.com/qhgz2013/anime-face-detector".

[50] V. Kumar, "Tom & Jerry detection," 2020, last accessed on 2022-06-25: "https://www.kaggle.com/datasets/vijayjoyz/tom-jerry-detection".

[51] J. Taubert, S. G. Wardle, M. Flessert, D. A. Leopold, and L. G. Ungerleider, "Face pareidolia in the rhesus monkey," *Current Biology*, vol. 27, no. 16, pp. 2505–2509.e2, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982217308126