Cartoon Face Recognition: A Benchmark Dataset

Yi Zheng[1], Yifan Zhao[2], Mengyuan Ren[1], He Yan[1], Xiangju Lu[1,*], Junhui Liu[1], Jia Li[2,3,*]
[1]iQIYI, Inc
[2]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[3]Peng Cheng Laboratory, Shenzhen, China

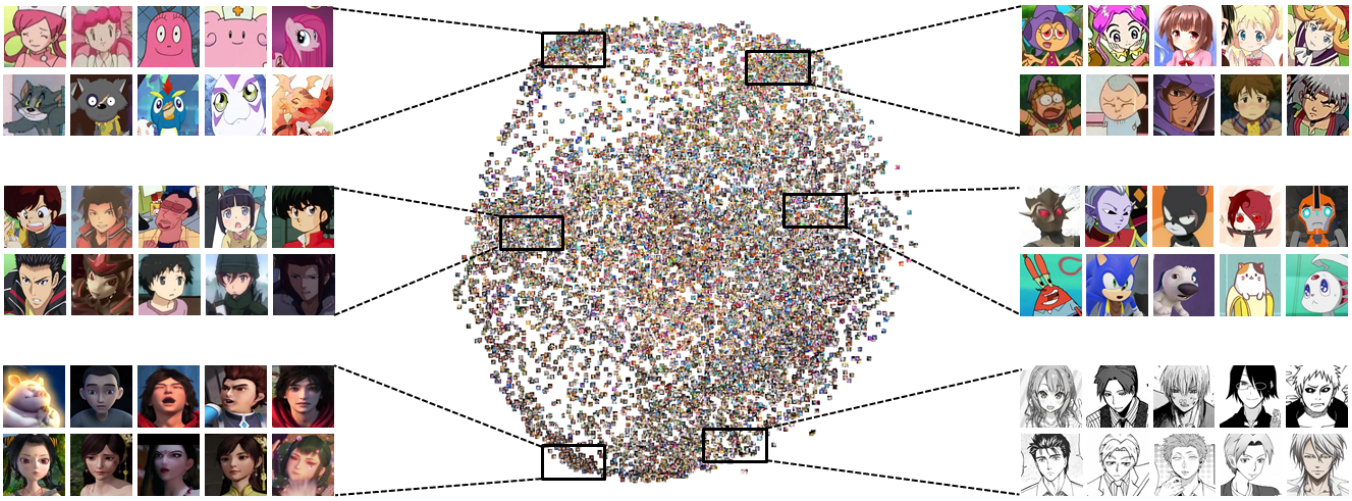

Figure 1: Illustration of iCartoonFace embedding. The proposed dataset consists of diverse data sources for face recognition and detection task. Dataset has been publicly available for promoting subsequent researches.

ABSTRACT

Recent years have witnessed increasing attention in cartoon media, powered by the strong demands of industrial applications. As the first step to understand this media, cartoon face recognition is a crucial but less-explored task with few datasets proposed. In this work, we first present a new challenging benchmark dataset, consisting of 389,678 images of 5,013 cartoon characters annotated with identity, bounding box, pose, and other auxiliary attributes. The dataset, named iCartoonFace, is currently the largest-scale, high-quality, rich-annotated, and spanning multiple occurrences in the field of image recognition, including near-duplications, occlusions, and appearance changes. In addition, we provide two types of annotations for cartoon media, *i.e.*, face recognition, and face detection, with the help of a semi-automatic labeling algorithm. To further investigate this challenging dataset, we propose a multi-task domain adaptation approach that jointly utilizes the human and cartoon domain knowledge with three discriminative regularizations. We hence perform a benchmark analysis of the proposed dataset and verify the superiority of the proposed approach in the cartoon face recognition task. The dataset is available at https://iqiyi.cn/icartoonface.

* Correspondence should be addressed to Xiangju Lu and Jia Li
(E-mail: luxiangju@qiyi.com, jiali@buaa.edu.cn).

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413726

CCS CONCEPTS

• **Computing methodologies** → **Object recognition**.

KEYWORDS

cartoon face, recognition, domain adaptation

ACM Reference Format:
Yi Zheng[1], Yifan Zhao[2], Mengyuan Ren[1], He Yan[1], Xiangju Lu[1,*], Junhui Liu[1], Jia Li[2,3,*] . 2020. Cartoon Face Recognition: A Benchmark Dataset. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413726

1 INTRODUCTION

What helps one recognize a face? Despite the faces in real-world images, cartoon face is a vital part to understand and interact with the virtual world. Accurately recognizing these cartoon characters is an essential prerequisite for many vision applications, such as automatic editing, filming, advertisement recommendation, and computer-aided modeling [4, 6]. With the proposals of large datasets [13, 32], deep models on human faces have achieved transpersonal accuracies, which greatly surpasses the conventional hand-crafted methods. For example, ArcFace [5] reached a precision of 99.83% on LFW benchmark and the best accuracy on

Table 1: Summary of existing datasets related to cartoon recognition.

Dataset	Type	#images	#identities	Artistic style	Face Anno.	Atrribute Anno.
Klare *et al.*. [17]	Caricature recognition	392	196	✗	✗	✗
Abaci *et al.* [1]	Caricature recognition	400	200	✗	✓	✗
WebCaricature [14]	Caricature recognition	12,016	252	✗	✓	Facial Landmark
IIIT-CFW [21]	Caricature recognition	8,928	100	✗	✗	Pose (1D), Age
Manga109 [7]	Cartoon detection&retrival	21,142	-	Unified	✓	✗
iCartoonFace	Cartoon recognition&detection	**389,678**	**5,013**	Unified	✓	Pose (3D), Gender

MegaFace [15] have also reached 99.39%. However, the performance gap is mainly achieved by the utilization of tremendous manual labeling datasets, which is extremely deficient in the cartoon media.

Over the years, cartoon media have shown its strong correlations with the real-world knowledge. Artists create and imagine the cartoon characters based on real-world abstractions and thus the created faces share a lot of similarities with human faces. To answer the aforementioned question, two natural problems raise our concerns: 1) what is the desirable need in cartoon dataset? 2) what is the relationship between human faces and cartoon ones?

In this less-explored recognition of virtual media, few datasets [2, 7, 14, 21] have been proposed for specific purposes, which can be roughly grouped into two categories. The first category is the caricature dataset, which is fundamentally based on real human identities. These caricature images share strong similarities with the human portrait but exaggerate certain specific facial features. For example, WebCaricature [14] built a large photograph-caricature dataset consisting of 6,042 caricatures and 5,974 photographs from 252 persons. IIIT-CFW [21] established a challenging dataset of 8,928 annotated unconstrained cartoon faces of 100 international celebrities. Both datasets share lots of similarities with the real-world human faces but with the variation of artistic styles. This would lead to a severe recognition problem, when regarding the same figure drawn by different artists as one class. The second category focuses on the cartoon recognition task, with very few datasets proposed. Datasets of this category do not rely on real celebrities or actors, but follow the basic rules in constructing a face. In these virtual media and cartoon videos, most characters show exaggerated or humorous facial expressions, which brings us new challenges in recognizing the same identity. For example, Manga109 [7] proposed a dataset for cartoon image retrieval and face detection, consisting of 21,142 images from 109 Japanese comic books. Besides its lack of face recognition, this dataset mainly collected limited images and restricted in the Japanese style comics, which may not satisfy the demand of large-scale training data for deep learning approaches. In addition, the cartoon dataset is required to contain substantial complex scenarios, thus is available and robust for the industrial application. To concretely meet the first concern, a high-quality, representative, large-scale dataset for cartoon face is in high demand.

It is notable that even the most unrealistic cartoon faces are created with *anthropomorphism*, indicating the correlation between the virtual media and realistic human images. Hence, for the second concern, we would like to examine how useful the human face can help the cartoon ones, including the recognition task and detection task. Based on this thought, we develop semi-automatic annotation procedures that make use of existing human faces as embedding knowledge, which serves as preliminary detectors and classifiers for the labeling process. On the other hand, the existing human faces could also help the recognition of cartoon faces, which serves as a teacher network and transfers the domain knowledge to the cartoon domains. Toward this end, a high-quality benchmark dataset and a learning approach for cartoon face recognition are needed to be proposed.

In this paper, we present iCartoonFace: a high-quality, large-scale, rich annotated benchmark dataset for cartoon face recognition. The iCartoonFace dataset consists of 389,678 images of 5,013 cartoon persons from public websites and online videos. In addition, we also provide 60,000 images of 109,810 faces with bounding boxes to form the detection dataset The iCartoonFace dataset also provides auxiliary information, including 3d pose (yaw, pitch and roll angles), collection source, album, and gender (shown in Tab. 1). This dataset spans multiple occurrences in the cartoon recognition challenges, including near-duplication, inter-class diversity, illumination conditions, and appearance changes. Besides this challenging benchmark, we thus set out to propose a multi-task domain adaptation approach to transfer the real-world faces to the cartoon media, which jointly regularizes the embedding space with three cues, *i.e.*, the classification constraint, unknown identity constraint and cross-domain constraint. Experimental evidence demonstrates the challenges in the face recognition benchmark and the improvement potential in utilizing real-world data.

Main contributions of this work can be summarized as: 1) We present a high-quality, large-scale, challenging benchmark dataset for cartoon face recognition and face detection, which consists of 389,678 images of 5,013 cartoon persons. We thus develop a semi-automatic assembling strategy to collect these diverse images from 1,302 albums. 2) We propose a multi-task domain adaptation framework to investigate the potential of transferring knowledge from human domains and cartoon domains. 3) We perform a benchmark analysis of the state of the arts and open-source for the subsequent researches. Further experiments demonstrate the effectiveness of the proposed method and the challenge of the benchmark.

2 RELATED WORK

2.1 Datasets

Existing works [9, 13] pay many efforts in the real-world human faces. For example, Labeled Faces in the Wild (LFW) [13] database of face images is designed for studying the problem of unconstrained face recognition. The database contains more than 13,000 images of

5,749 characters. MegaFace [15] is a large scale face database of over 1 million faces photos from 690k persons. CASIA-WebFace [32] contains more than 10k identities and about 500k images for unlimited face recognition. CAS-PEAL [8] contains more than 30k images of 1,040 persons for constrained face recognition, mainly includes gestures, expressions, lighting variations.

There are also some datasets [1, 17] related to cartoon and caricature recognition. For example, IIIT-CFW [21] contains 8,928 annotated cartoon faces of 100 international celebrities. IIIT-CFW can be used for a wide spectrum of problems due to the fact that it contains detailed annotations such as type of cartoon, pose, expression, age group, and *etc.* WebCaricature [14] is a large photograph-caricature dataset consisting of 6,042 caricatures and 5,974 photographs from 252 persons collected from the web. For each image in the dataset, 17 labeled facial landmarks are provided. All two cartoon datasets are created from caricature. Manga109 [7] is a dataset of a variety of 109 Japanese comic books and created for detection. Danbooru [2] simply gathered a large-scale collection of over 970k images of 70k identities. However, there are two main problems existing in this dataset. First, the annotations are roughly collected noisy labels, which are not annotated by human annotators. Even the best performed model [11] can only reach the 37.3% accuracy. Second, the same identity in this dataset may be created by different artists. Thus a high-quality, manual labeling dataset for cartoon face recognition is in high demand. Thus we defined the two different types of artistic style in Tab. 1: *Unified* indicates that characters are created in the same style or by the same artist.

2.2 Face Recognition

Face recognition can be seen as a sub-problem of image classification. Numerous models [10–12, 30] are proposed for solving image classification problem and get impressive results. For instance, ResNet [10] presented a residual learning framework to ease the training of networks that are substantially deeper than those used previously. DenseNet [12] connected each layer to every other layer in a feed-forward fashion. SENet [11] re-calibrated channel-wise feature responses by explicitly modeling inter-dependencies between channels.

Despite these network architectures, advanced face recognition algorithms [19, 20, 26, 27, 29, 36] have also been proposed. For example, Wang *et al.* [26] proposed a hyper-sphere embedding strategy to enhance face verification efficiencies. Taigman *et al.* [25] presented a framework for recognizing the faces in the wild. SphereFace [19] presented a novel loss that enables convolutional neural networks (CNNs) to learn angular discriminative features. CosFace [27] maximized the decision margin in the cosine space. Arcface [5] maximized the decision margin in the angular space. However, these algorithms only study this problem in the image and face recognition in the real world. For object detection, focal loss [18] adopts a re-weighting scheme to address the class imbalance problem, which is also widely used in the face recognition task. Some work [28, 36] presented an improved softmax loss to regularize the features in normalized space. In addition, Yang *et al.* [31] proposed to embed the landmark information in a deep network to help the recognition progress.

2.3 Multi-task Learning

Zhang *et al.* [34] made a detailed survey of general multitask learning (MTL). More specifically related to our work, [35] solved a convex optimization problem to estimate task relationships, while [22] analyzed the weighted sum loss algorithm and its applications in the online, active, and transductive scenarios. Moreover, [16] proposed probabilistic models through the construction of a task covariance matrix or estimate the multitask likelihood from a deep Bayes model. In the face recognition field, HyperFace [23] proposed to classify a given image region as face or non-face, estimate the head pose, locate face landmarks, and recognize gender in one network. Different from these aforementioned approaches, in this paper, we designed to propose a multi-task domain adaptation framework to regularize the embedding vectors with reasonable decision borders.

3 DATASET CONSTRUCTION AND ANALYSIS

3.1 Semi-automatic Assembling Process

To reduce the labeling burden, we develop a semi-automatic algorithm to collect and annotate the iCartoonFace dataset (See Fig. 3). Our framework can be conducted in the following three stages: 1) hierarchical data collection; 2) data filtering process; 3) Q/A manual annotation.

Hierarchical Data Collection. The iCartoonFace dataset is collected by hierarchical manners (from cartoon album names to cartoon person names, and finally to cartoon person images). We first form a cartoon album name list regarding the rank[1]. Secondly, we obtain the main characters from the internet based on the album name list. Hence a list of cartoon persons and corresponding albums could be obtained. Thirdly, we download publicly available images from multimodal media, including public images, comic books, and video sources. In this manner, millions of images with noisy labels are obtained for subsequent data filtering process.

Guided Data Filtering. In practice, there are tremendous irrelevant or duplicate data in downloaded images, which brings us a great challenge to select valid data, especially without any prior knowledge. Hence we resort to the manual labeled human faces, which provide coarse filtering for the useless samples. In other words, we resort to two existing human knowledge to help the data filtering process, *i.e.*, face detection filtering, and face recognition filtering.

Hence, we build a data filtering process that consists of a face detection branch and a feature extraction branch. In the face detection part, we firstly manually labeled 60,000 images with 109,810 cartoon face bounding box as our cartoon face detection dataset, of which 80% was used as the trainset and 20% was used as the testset. We mix this part of data with the real human faces, resulting in a final 200,000 images with 500,000 bounding boxes to enhance the detection part. We adopt RetinaNet [18] as our detection backbone and achieved 89% mAP in 0.5 IoU on the test set. In the feature extraction part, we initially use Arcface model [5] pre-trained on existing human face recognition datasets as our feature extraction model. With the increase of cartoon datasets, the model can be jointly trained and the performances are improved steadily. The

[1] https://en.wikipedia.org/

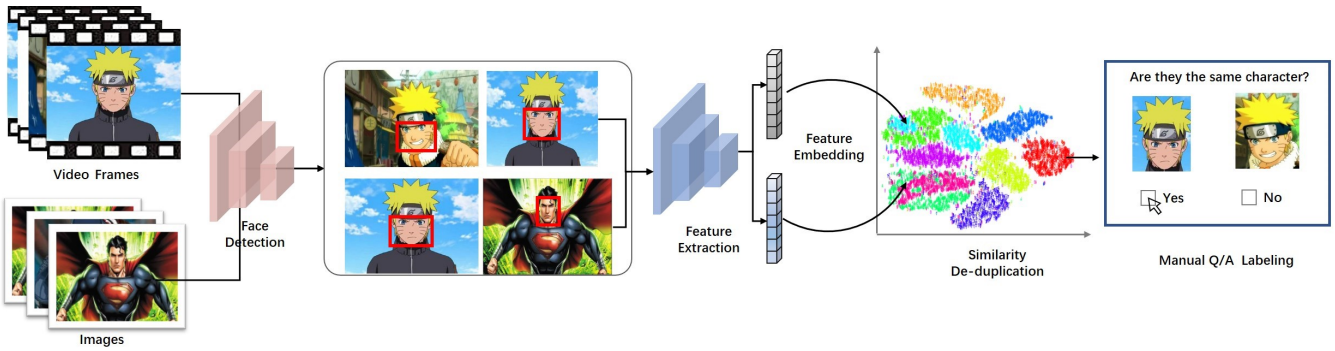

Figure 2: The overview of semi-automatic assembling process. Instead of labeling IDs, human annotators only need to answer the T/F questions.

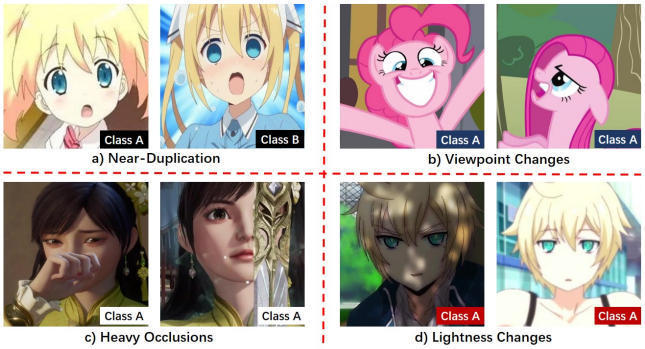

Figure 3: Four representative challenges in iCartoonFace: near-duplication, viewpoint changes, heavy occlusions, lightness changes.

success of the filtering process in turn verifies the correlations of the real-world data and virtual ones.

Q/A Manual Annotation. We developed a Q/A system to manually annotate the identity information of cartoon faces. In the annotation page (see Fig. 2), one part shows a reference image, and the other part displays the image to be labeled. The annotators needed to determine whether each new image shares the same identity with the reference image. The reference image is an identity picture provided by the expert based on the cartoon album name and cartoon person name to which the cartoon person belongs. In our dataset, 5,013 reference pictures are included, meaning that there is one probe for each identity.

3.2 Dataset Statistics

Large scale. The iCartoonFace dataset consists of 389,678 images of 5,013 cartoon persons coming from 1,302 cartoon albums. To our knowledge, this is currently the largest manual annotated image dataset for cartoon face recognition. As shown in Fig. 4 (a), the dataset cartoon persons are widely distributed in Japan, China, Europe and America.

In the wild/long-tailed. The dataset is created naturally with a long-tailed distribution in Fig. 4 (f). 50% of the identities own less than 30 images, while some of the cases even own about 500 images.

High quality. After the dataset was manually labeled, we perform a cross-checking method, and the re-checking error rate is guaranteed to be less than 5%. Fig. 4 (c) shows that the resolution of images are more than 100×100 and more than 65% of them are larger than 200×200. The sharpness of images is calculated by the Laplacian metric, and the values of most samples are more than 100 as shown in Fig. 4 (e), to ensure the clearness and sharpness of image boundaries.

Rich attributes. It provides information such as the face bounding box, identity, region, pose, and gender in each image. The statistical information about pose and gender is shown in Fig. 4 (b) and (d). Random 10,000 samples are selected and annotated with 3D pose information, *i.e.*. yaw, pitch and roll angles. 66% samples have small angles less than 30 degrees, and about 8% samples have large angles with more than 60 degrees up to 90 degrees.

3.3 Challenges and Tasks

Challenges. We visualize four representative challenges of cartoon data in the wild (Fig. 3): a) *Near-Duplication*: two images with different IDs are in a very similar appearance. This motivates the algorithms to be aware of subtle local differences. b) *Viewpoint Changes*: viewpoint changes of the same character brings us new challenges in recognizing one person. c) *Heavy Occlusions*: the faces might be occluded by other objects of the scenarios. Algorithms need to extract the most discriminative features for recognition. d) *Lightness Changes*: there are other kinds of variations, including the lightness and resolution changes. All these cues force the algorithms to be robustness for different scenarios.

Face recognition. Following the pioneer human face dataset [15], we proposed an identification task to benchmark the performance of the cartoon face recognition algorithms. In the identification task setting, given a probe photo, and a gallery containing at least one image of the same cartoon person, the algorithm needs to rank orders all images in the gallery based on similarity to the probe. Specifically, the probe set includes N cartoon persons and each cartoon person has M images. The algorithm then tests each of the M images per cartoon person by adding it to the gallery of distractors

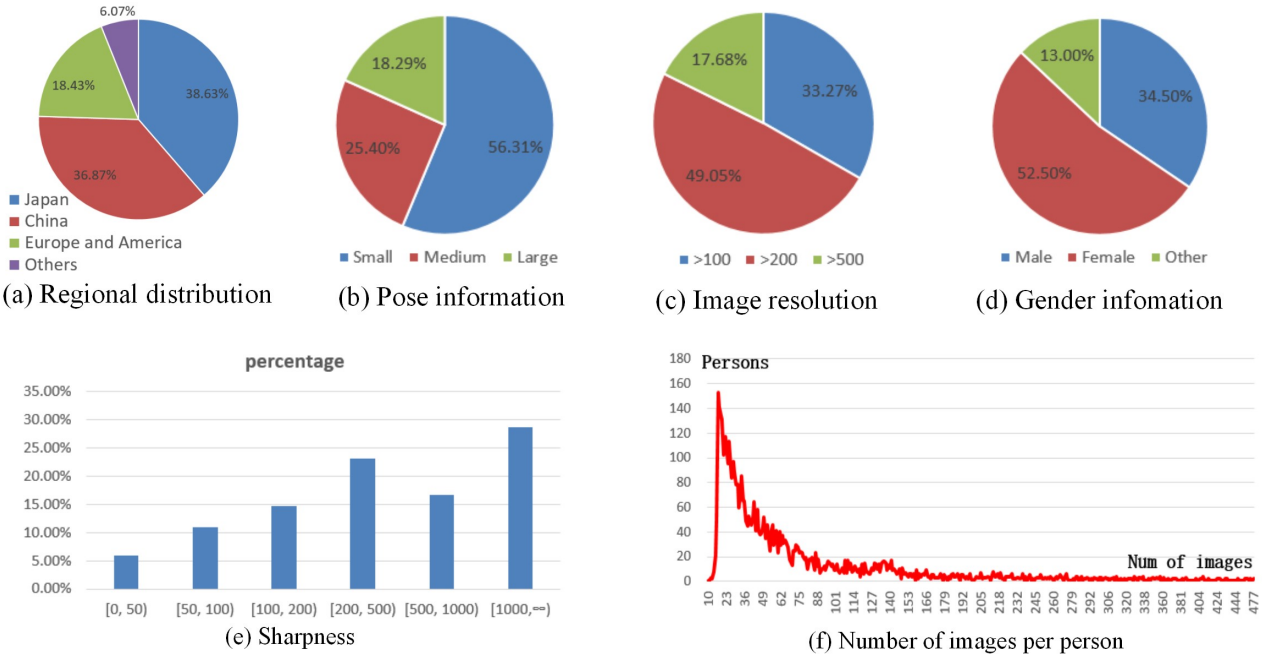

Figure 4: iCartoonFace dataset statistics. We present region distributions of identities, pose annotations, image resolution distribution, sharpness and number of samples per identity.

and use each of the other $M-1$ photos as a probe. Results are presented with the identification rate of rank-K. For the identification test, a gallery set with 2,500 images with 2,500 different persons is created. To ensure a fair identification, the gallery set is created by the persons that their identity does not appear in the training set and probe set. The identification probe set contains 20,000 images from 2,000 persons (1,200 of person identity which in training set and 800 of person identity which not in training set). The number of samples for each person ranges from 5 to 17.

Face detection. To construct a high-quality dataset, we select 50,000 images of 91,163 faces as training set and 10,000 images of 18,647 faces as testing set. The face detection annotations are manually labeled, undergone multiple quality checks to ensure that the error rate is less than 5%. More than 60% of faces have a resolution greater than 50×50 in the training set We use mAP (mean Average Precision) as evaluation protocol, *i.e.*, $mAP = \frac{\sum_i^C AP_i}{C}$. Remarkably, this is also the largest cartoon face detection dataset to date. The annotated samples can be found in Fig. 5 and the details can be found in supplementary.

Figure 5: Detection annotations. The bounding boxes are viewed in red.

4 A BASELINE APPROACH

4.1 Problem Formulation

As mentioned above, the cartoon faces are derived from abstractions of human faces, thus share a lot of similar structures and common knowledge. In theory, our basic idea is to utilize the domain data from real-world human faces, helping to partition classification hyper-planes in cartoon domain.

To achieve this, we design three meaningful cues: 1) We design a Siamese network to conduct the training procedure and optimize the classification loss of two domains per batch. As shown in Fig. 6, these two networks are built in a weight-sharing manner and regularized with additional training data from the human domain. 2) As the labeled data is in the one-hot form, when re-partition the classification hyper-planes, the classification loss would force the predicted distribution to be a one-hot vector, *e.g.*,$[0, ..., 0, 1]$. This optimization would lead to severe overfitting, especially when testing on unseen identities. To solve this issue, we adopt the unknown identity loss to smooth the predicted distributions with the help of data from other domains. That is to say, when smoothing the

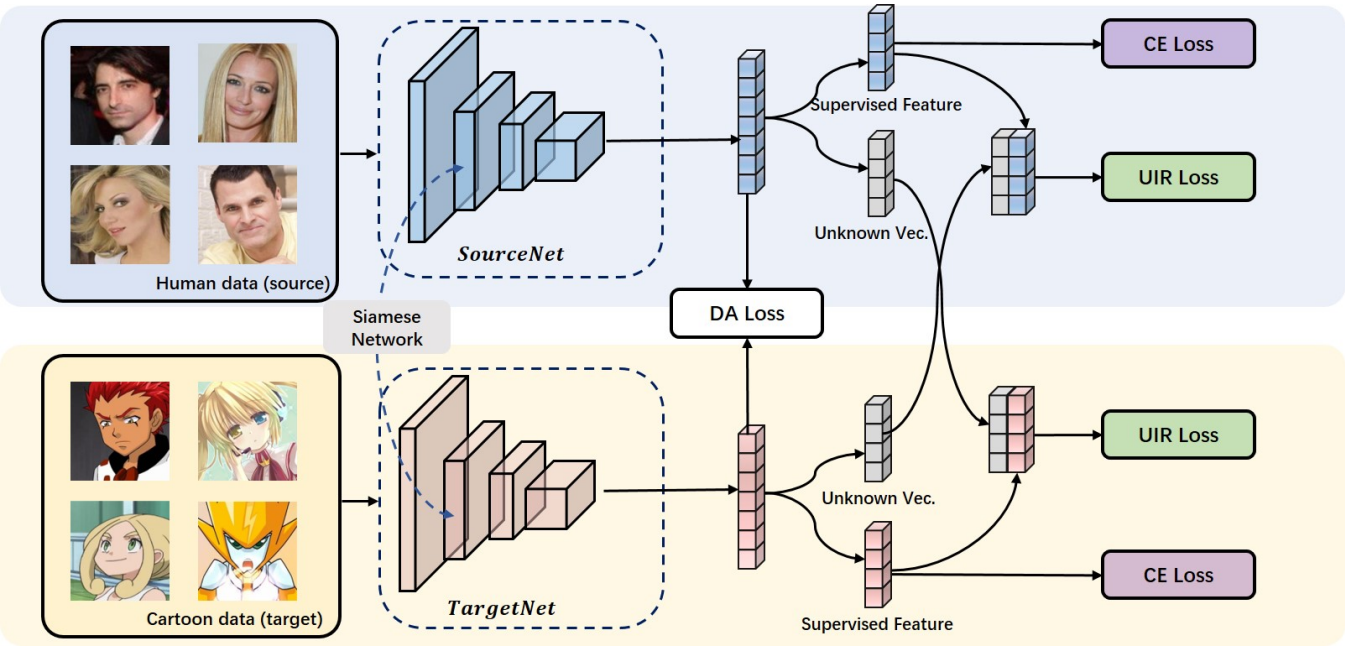

Figure 6: The overview of the proposed approach. Our framework is composed of three important constraints: the classifier of two domains, the unknown identity rejection and domain adaption adversarial loss.

distribution of cartoon faces, we adopt human faces as unlabeled data for this optimization. 3) Inspired by the gradient reversal layer, we develop a domain classifier in an adversarial learning manner. A general classifier for both domains should not show specific domain features in the high-level layers, *i.e.*, whatever the input image is from, the distribution of final prediction should be similar. Our motif is to fool the classifier to do not be aware of its domain and to find a generalized classification hyper-planes.

Let $\mathbb{H}$ represent human face domains and $\mathbb{C}$ be cartoon face domains. We thus put the source training data $\mathcal{D}_h \sim \mathbb{H}$ and target data $\mathcal{D}_c \sim \mathbb{C}$ in one batch to conduct this training procedure. Let $\mathbf{T} = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ represents a typical triplet sampled from these two domains. In this triplet, $\mathbf{x}$ denotes the input image, $\mathbf{y}$ denotes the identity label of image $\mathbf{x}$, and $\mathbf{z}$ indicates whether the image belongs to a cartoon or a human face. When the input image is a cartoon face $\mathbb{C}$, $\mathbf{z}$ is set as 1 and otherwise 0. With the given input, our aim is to predict a generalized distribution $\mathbf{p}$ of both domains with the three aforementioned cues.

4.2 Learning objective

As shown in Fig. 6, with an input image $\mathbf{x}$ forward-propagated through the network, three discriminative classifiers can be obtained. Considering that these two networks are designed in a weight-sharing trend, in the testing phase, we do not introduce any additional computation cost. With these classifiers, the total loss function $\mathcal{L}$ is composed of three parts: classification loss $\mathcal{L}_{cls}$, unknown identity rejection loss $\mathcal{L}_{uir}$, and domain transfer loss $\mathcal{L}_{da}$.

Classification loss. To obtain a discriminative feature extractor, we adopt the classification loss to regularize both cartoon and human face classifiers, represented as $\mathcal{L}^c_{cls}$ and $\mathcal{L}^h_{cls}$, respectively. For $\forall(\mathbf{x}_c, \mathbf{y}_c) \in \mathcal{D}_c, (\mathbf{x}_h, \mathbf{y}_h) \in \mathcal{D}_h$, a typical classification loss can be presented as:

$$\begin{aligned}\mathcal{L}_{cls} &= \mathcal{L}^c_{cls}(\mathbf{x}_c, \mathbf{y}_c) + \mathcal{L}^h_{cls}(\mathbf{x}_h, \mathbf{y}_h), \\ &= -\sum_i \mathbf{y}^c_i \log(\mathbf{p}^c_i) - \sum_j \mathbf{y}^h_j \log(\mathbf{p}^h_j).\end{aligned} \tag{1}$$

Unknown identity rejection loss. The UIR loss $\mathcal{L}_{uir}$ aims to find a feature re-projection with unsupervised regularization between different domains. Inspired by [33], we develop the unknown identity rejection loss $\mathcal{L}^c_{uir}$, for each batch sample $\mathcal{B}^c$, taking 1/4 human face data $\mathbf{x}_h$ as unlabeled data and 3/4 of $\mathbf{x}_c$ as known ones. $\mathcal{L}^h_{uir}$ is defined respectively. These two combined losses can be expressed as:

$$\begin{aligned}\mathcal{L}_{uir} &= \mathcal{L}^c_{uir}(\mathbf{x}) + \mathcal{L}^h_{uir}(\mathbf{x}), \\ &= -\sum_{\mathbf{p}\in\mathcal{B}^c} \log(\frac{\mathbf{p}_i}{\sum_j \mathbf{p}_j}) - \sum_{\mathbf{p}'\in\mathcal{B}^h} \log(\frac{\mathbf{p}'_i}{\sum_j \mathbf{p}'_j}), \\ &\forall(\mathbf{x}_c) \in \mathcal{D}_c, (\mathbf{x}_h) \in \mathcal{D}_h.\end{aligned} \tag{2}$$

Domain adaption loss. To reduce the domain gap between $\mathbb{H}$ and $\mathbb{C}$, we adopt the reciprocal of binary softmax loss to constrain the domain correlation between the cartoon and the human face dataset. The final loss function has the form:

$$\mathcal{L}_{da} = \Sigma_{\mathbb{H},\mathbb{C}} \frac{1}{-\mathbf{z_i}\log(\mathbf{D}(\mathbf{x_i})) - (1-\mathbf{z_i})\log(\mathbf{D}(\mathbf{x_i}))}, \tag{3}$$

where $\mathbf{D}(\cdot)$ denotes the network prediction of the domain adaption classifier. In other words, if the smaller the loss, the better generalization of two domains $\mathbb{H}$ and $\mathbb{C}$. The total loss $\mathcal{L}_{sum}$ function can

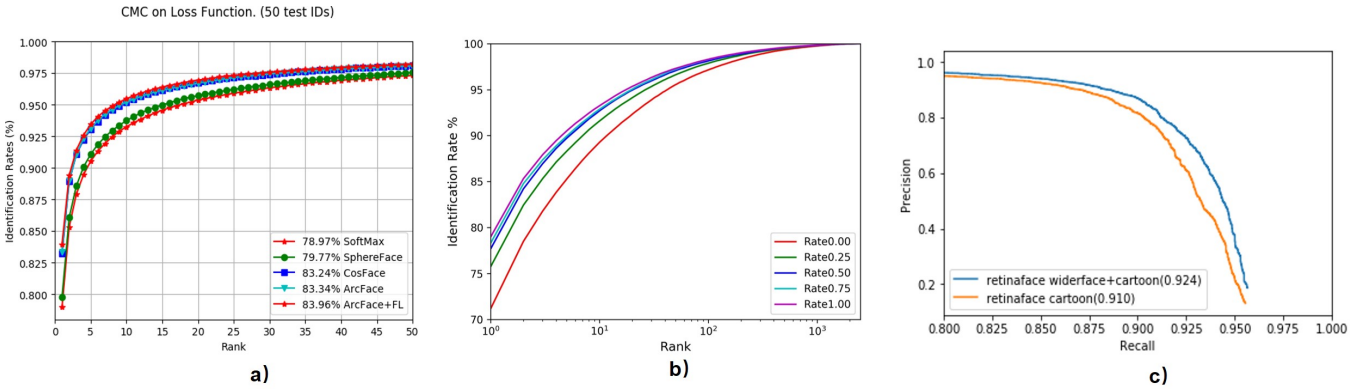

Figure 7: Visualization results. a) CMC curve of five state-of-the-art recognition algorithms. b) CMC curve of recognition with different enlarged rates. c) Precision-Recall on detection dataset for data filtering.

be formally presented as:

$$\mathcal{L}_{sum} = \alpha \cdot \mathcal{L}_{cls} + \beta \cdot \mathcal{L}_{uir} + \gamma \cdot \mathcal{L}_{da}. \quad (4)$$

In the above expression, α, β, and γ are the corresponding proportional weights of the three loss functions, respectively.

5 EXPERIMENTS

5.1 Experimental Protocol

Datasets. We adopt the CASIA-WebFace [32] as the human face dataset. In this dataset, we randomly selected 168,266 images of 5,000 people as the human domain data. Besides this dataset, we make use of the proposed iCartoonFace as the cartoon training set, consisting of 389,678 images of 5,013 identities. The test set is composed of 20,000 images of 2,000 identities, accompanied with 2,500 distracters.

Implementation. To make a fair benchmarking, we uniformly employ DenseNet-169 [12] as our backbone network for all the classifiers, *i.e.*, SoftMax, SphereFace [19], CosFace [27], ArcFace [5], Focal loss [18]. The training images are uniformly resized to 256 × 256 in this paper. We first adopt the softmax-loss model on the cartoon dataset as the pretrained model to initialize our training. Based on this pretrained model, we then use the proposed three losses to jointly train on the cartoon and human face datasets. Especially, the cartoon face regions in all experiments are enlarged as 2 times width and 2 times length. We use stochastic gradient descent (SGD) as the optimization function. The initial learning rate is set as 0.1, and every 20 epochs the learning rate decay as 0.1 times of the previous epoch. We empirically set $\alpha = 1, \beta = 0.1, \gamma = 10$ to balance three weights in the same order of magnitudes.

5.2 Comparisons and Relations

In Tab. 2, we exhibit several widely-used face recognition dataset, *i.e.*, CFP [24], LFW [13] and MegaFace [15]. It is notable that LFW and CFP datasets are composed of a limited number of images with 13k and 7k images. This limited scale makes state-of-the-art recognition algorithms [5, 18, 19, 27] undifferentiated performance. (v) indicates the accuracy on verification set, otherwise identification set. With the proposal of large-scale dataset [15], various algorithms achieved a rapid development even on million-level datasets, indicating the aligned frontal human faces can be basically solved

Table 2: Accuracy (%) of state-of-the-art methods on public datasets. Results on human faces are reported by [29].

	Human			Cartoon
	LFW(v)	CFP(v)	MegaFace	iCartoonFace
#images	13k	7k	1,000k	389k
SoftMax	99.59	94.04	93.94	78.97
SphereFace	99.65	94.22	94.18	79.77
CosFace	99.71	95.68	97.69	83.24
ArcFace	99.76	95.28	97.28	83.34
F-ArcFace	99.71	95.62	97.51	83.96

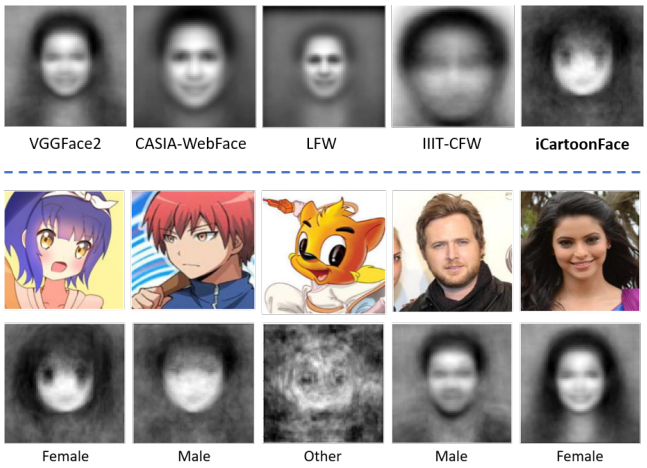

Figure 8: Top: Meanface of three human face dataset and two cartoon dataset. Bottom: representative samples and mean faces in our dataset. The last two human faces are calculated by VGGFace2 [3].

Table 3: Accuracy (%) on state-of-the-art models. The best performance is viewed in bold.

Models	Rank@1	Rank@5	Rank@10
SoftMax	78.97	90.51	93.24
SphereFace	79.77	91.07	93.74
CosFace	83.24	93.05	95.19
ArcFace	83.34	93.12	95.20
F-ArcFace	83.96	93.43	95.46
Ours	**84.34**	**93.52**	**96.14**

Table 4: Ablations of the proposed approach. Our full model reaches the highest performance.

$\mathcal{L}_{cls}$	$\mathcal{L}_{uir}$	$\mathcal{L}_{da}$	Rank@1	Rank@5	Rank@10
✓			81.46	91.94	94.30
✓	✓		81.67	92.15	94.50
✓	✓	✓	**84.34**	**93.52**	**96.14**

by the existing techniques. However, in the domain of cartoon face, although notable progress can be achieved by the latest techniques [5, 18], there is still a performance gap when comparing to the human domain. For example, the ArcFace performs over 95% on all human datasets while only 83.34% on cartoon faces.

Starting from another perspective, we visualize the mean face of different types of datasets in Fig. 8. The first three are from the human faces, which are carefully aligned with notable features. For the last two faces of cartoon set, the facial landmarks are thus inconspicuous which brings us new challenges, including the gender issue on the bottom.

5.3 Performance Analysis

Which is the best algorithm for cartoon face? To fairly evaluate these algorithms, we integrate the same backbone with five algorithms: softmax, SphereFace [19], CosFace [27], ArcFace [5], Focal loss [18] with Arcface (F-Arcface). Compared to the basic softmax, [5, 19, 27] improves performance steadily, with over 5% in rank1. In addition, we visualize the CMC curve in Fig. 7, it can be found that F-Arcface shows a leading performance among the state-of-the-art methods, including the low-rank rate (*e.g.*, rank@1) and higher rate (rank@50). This verifies that the improvements of the algorithm are consistent in all stages for identification.

Can human face knowledge transfer to cartoon face? To evaluate the effectiveness of our different proposed loss functions, we visualize the cartoon recognition task in Tab. 4. It can be found that the performance of joint cartoon and human face recognition dataset training directly will be worse, compared to only cartoon dataset is used for training in Tab. 3. After adding the unknown identity rejection loss and selecting the better hyperparameter β, the performance will be slightly improved. Finally, a domain transfer loss is added to make the face domain and cartoon domain unified, which can greatly improve the performance of cartoon face recognition.

Table 5: Recognition performance with different enlarged rates of facial region.

Enlarged Rate	Rank@1	Rank@5	Rank@10
0.00	71.15	85.26	89.26
0.25	75.73	88.30	91.58
0.50	77.67	89.72	92.69
0.75	78.29	89.95	92.86
1.00	**78.97**	**90.51**	**93.24**

Despite the recognition task, we also utilize the human domain data for the detection task. We validate the effectiveness of our semi-automatic labeling procedure, which can be found in Fig. 7 c). The first trained cartoon face model reaches an accuracy of 91.0%. After including the WiderFace dataset of human beings, the final performance reaches 92.4%, which helps the annotation process.

Is context useful for cartoon face recognition? Cartoon face is the main discriminative part in identifying one character. However, separating from human faces based on cartoon faces may not be enough to distinguish different cartoon persons in some cases. To explore this, we compare the effects of different expansion proportions by $[0.0, 1.0]$ to introduce more context information, such as haircut and decorations. In this setting, we use the same DenseNet-169 as the backbone network and softmax losses. From Tab. 5, it can be easily obtained that more contextual information can be beneficial to the identification task. This verifies that in the cartoon person recognition task, the face region plays the most important role in identifying a cartoon person. And the CMC curve in Fig. 7 b) also indicates the higher contextual information is included, the higher performance can be achieved.

6 CONCLUSION

In this paper, we take a closer look into the cartoon media by establishing a benchmark dataset, namely iCartoonFace. The dataset shows many meaningful features, including high-quality, large-scale, in-the-wild, and with rich annotations. Beyond these issues, we thus designed to investigate the challenges and present two typical tasks, *i.e.*, face recognition, and face detection. Under this setting, we present a multi-task domain adaptation solution to utilize the human data to cartoon media. Experimental evidences verify the potential of multi-domain exploitation and analyze the data benchmark. We believe that the research on cartoon face recognition would bring more attractive researches and broad industrial applications.

ACKNOWLEDGMENTS

We would like to thank Song Shi and his team for the organization of competitions, Yan Fu and He Chen team for the help of data annotations, Chenwei Yang team for providing computing resources, and Lingyun Xiao, Ke Chen, Xiang Xia *et al.* for developing and designing competition websites. This work was supported in part by the National Natural Science Foundation of China (No. 61922006 and No. 61532003), the Beijing Nova Program (No. Z181100006218063), and iQIYI Inc. (iTP19-2500067).

REFERENCES

[1] Bahri Abaci and Tayfun Akgul. 2015. Matching caricatures to photographs. *Signal, Image and Video Processing* 9, 1 (2015), 295–303.

[2] Anonymous, Danbooru community, Gwern Branwen, and Aaron Gokaslan. 2019. Danbooru2018: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset. https://www.gwern.net/Danbooru2018. https://www.gwern.net/Danbooru2018 Accessed: DATE.

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

[4] Hong Chen, Nan-Ning Zheng, Lin Liang, Yan Li, Ying-Qing Xu, and Heung-Yeung Shum. 2002. PicToon: a personalized image-based cartoon system. In *Proceedings of the tenth ACM international conference on Multimedia*. 171–178.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4690–4699.

[6] Michael Elad, J-L Starck, Philippe Querre, and David L Donoho. 2005. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Applied and Computational Harmonic Analysis* 19, 3 (2005), 340–358.

[7] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2016. Manga109 dataset and creation of metadata. In *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*. ACM, 2.

[8] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. 2007. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38, 1 (2007), 149–161.

[9] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*. 87–102.

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

[11] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7132–7141.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4700–4708.

[13] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

[14] Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. 2017. WebCaricature: a benchmark for caricature face recognition. *arXiv preprint arXiv:1703.03230* (2017).

[15] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4873–4882.

[16] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7482–7491.

[17] Brendan F Klare, Serhat S Bucak, Anil K Jain, and Tayfun Akgul. 2012. Towards automated caricature recognition. In *2012 5th IAPR International Conference on Biometrics (ICB)*. 139–146.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2980–2988.

[19] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 212–220.

[20] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan. 2016. Pose-aware face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4838–4846.

[21] Ashutosh Mishra, Shyam Nandan Rai, Anand Mishra, and CV Jawahar. 2016. IIIT-CFW: a benchmark database of cartoon faces in the wild. In *European Conference on Computer Vision (ECCV)*. Springer, 35–47.

[22] Anastasia Pentina and Christoph H Lampert. 2017. Multi-task learning with labeled and unlabeled tasks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 2807–2816.

[23] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. 2017. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence* 41, 1 (2017), 121–135.

[24] Soumyadip Sengupta, Jun Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. 2016. Frontal to profile face verification in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9.

[25] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1701–1708.

[26] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. 2017. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*. 1041–1049.

[27] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5265–5274.

[28] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. 2017. Deep metric learning with angular loss. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2593–2601.

[29] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. 2020. Mis-classified Vector Guided Softmax Loss for Face Recognition. *AAAI Conference on Artificial Intelligence (AAAI)* (2020).

[30] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1492–1500.

[31] Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. 2020. FAN-Face: a Simple Orthogonal Improvement to Deep Face Recognition. *AAAI Conference on Artificial Intelligence (AAAI)* (2020).

[32] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).

[33] Haiming Yu, Yin Fan, Keyu Chen, He Yan, Xiangju Lu, Junhui Liu, and Danming Xie. 2019. Unknown Identity Rejection Loss: Utilizing Unlabeled Data for Face Recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*. 0–0.

[34] Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114* (2017).

[35] Yu Zhang and Dit-Yan Yeung. 2012. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536* (2012).

[36] Yutong Zheng, Dipan K Pal, and Marios Savvides. 2018. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5089–5097.