2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)

Class-Balanced Training for Deep Face Recognition

Yaobin Zhang, Weihong Deng*
Beijing University of Posts and Telecommunications
{zhangyaobin, whdeng}@bupt.edu.cn

Abstract

The performance of deep face recognition depends heavily on the training data. Recently, larger and larger datasets have been developed for the training of deep models. However, most face recognition training sets suffer from the class imbalance problem, and most studies ignore the benefit of optimizing dataset structures. In this paper, we study how class-balanced training can promote face recognition performance. A medium-scale face recognition training set BUPT-CBFace is built by exploring the optimal data structure from massive data. This publicly available dataset is characterized by the uniformly distributed sample size per class, as well as the balance between the number of classes and the number of samples in one class. Experimental results show that deep models trained with BUPT-CBFace can not only achieve comparable results to larger-scale datasets such as MS-Celeb-1M but also alleviate the problem of recognition bias.

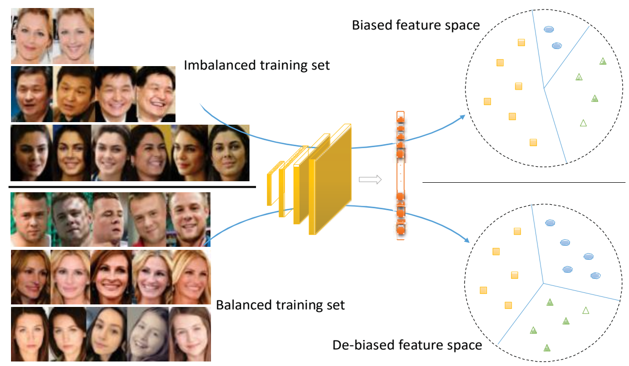


Figure 1: Upper: imbalanced training set leads to unequal feature space. Lower: balanced training set leads to equal feature space, improving recognition accuracy and fairness.

1. Introduction

In recent years, face recognition technology is becoming more mature and applicable. A lot of public face recognition training sets [5, 13, 31, 33, 46] are developed to meet the needs of training deep models. The recognition performance on public benchmarks such as LFW [18] are also becoming saturated. However, the class imbalance problem [2, 3, 14, 15, 20] remains a bottleneck in the field of deep face recognition, which means, the number of samples in majority classes is much more than that in minority classes in the training sets. The imbalanced data distribution is characterized by the long tail distribution [28, 51]: a few classes have many face images as the "head" data, and most classes have fewer face images as a long "tail".

Developing a face recognition system using imbalanced training sets, which is a common practice, can really impair the representation ability of the model. First, the recognition accuracy is affected. As shown in the upper part of Figure 1, if the model is trained with class-imbalanced training sets, the volume of different classes in the feature space is unequal. The majority classes occupy bigger spaces so that when the model is applied, samples with similar distribution to the minority classes have a greater chance of being misidentified. In contrast, if the number of samples in each class in the training set is the same, as shown in the lower part of Figure 1, the model can reserve equal volume space for different identities. Second, fairness is affected. Due to the limitation of data collection methods, different populations have different probability of appearing in the dataset. For example, most face recognition training sets are composed of celebrities [13, 36, 46], so that the proportion of men is much larger than that of women, the proportion of Americans is much larger than that of Africans, and the proportion of elderly and infants is seriously insufficient. As a result, women, Africans, the elderly and infants have less chance to be well learned by the model, leading to the bias in face recognition. Some bias-related researches [19, 40] prove the existence of this kind of misidentification and unfairness. We firmly believe that in face recognition, everyone should be treated equally, and the unfairness can be alleviated with balanced training data.

Besides the class imbalance problem, the data structure of the dataset is also worth studying. Zhou *et al.* [52] prove that when training a small portion of a large dataset, using

2160-7516/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPRW50498.2020.00420

the “head” part can reach a better recognition result on the LFW [18] than randomly sampling classes. However, as the number of selected classes increases, “head” data begins to suffer from the long tail problem, resulting in performance degradation, although the number of images and identities for training is indeed increasing. This suggests that feeding a large amount of data to the model does not necessarily lead to better training results. Carefully selected classes and well-designed sample distributions also play vital roles in the effectiveness of face recognition.

In this paper, we study the impact of dataset structure on deep face recognition, and especially observe the phenomenon produced by class-balanced training. Extensive experiments are performed to compare face recognition performance on the long-tailed and uniformly distributed training data, showing that the long tail phenomenon is likely to be one of the important factors that restrict the performance of a dataset. In addition, the issues of class selection and balance between the number of classes and the number of samples per class are also well studied experimentally. Finally, in light of the experimental observations, an optimized training set BUPT-CBFace is built for efficient deep face recognition. As shown in Figure 2, BUPT-CBFace is a class-balanced face dataset, which is constructed by searching optimal data structure for face recognition.

Based on state-of-the-art ResNet [16] architecture and ArcFace loss [9], compared to the widely-used CASIA-WebFace [46] dataset, training deep models using BUPT-CBFace of the same size can improve the accuracy on LFW [18], RFW [40] and IJB-C [29] by a large margin, and reach state-of-the-art performance on MegaFace challenge 1 [21] under the small protocol with 79.57% identification accuracy and 95.20% verification accuracy. Moreover, BUPT-CBFace even outperforms the large-scale face dataset MS-Celeb-1M [10, 13], exceeding it by 2.10% on the average accuracy of five verification sets with *eight times fewer* training images. To encourage more class balance researches, the BUPT-CBFace dataset is made publicly available at http://whdeng.cn.

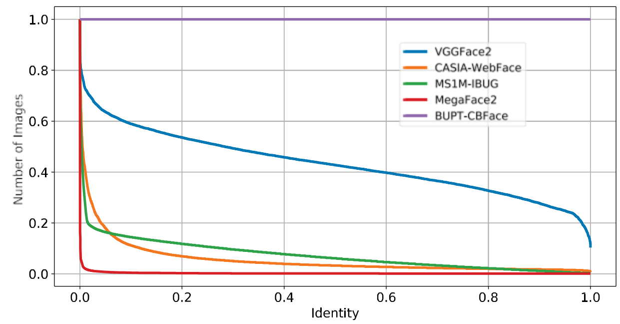


Figure 2: Sample distributions of widely-used long-tailed datasets and BUPT-CBFace. The two axes are *normalized*.

2. Related Work

Class Imbalance Problem In recent years, a lot of work [2, 15, 20] has been devoted to addressing the problem of imbalanced training samples in deep learning. In terms of algorithm, UP [12] imposes a penalty on the norm of weight vectors so that minority classes can have comparable feature space volume with majority classes. Wu *et al*. [44] propose a center invariant loss that aligns the feature centers of the minority classes to the majority. Fair loss [24] uses reinforcement learning to balance different classes. Zhong *et al*. [51] train the head data and tail data separately to reduce the long tail effect. Ring loss [50] applies soft feature normalization to augment standard loss functions.

Some other work improves in terms of data, such as data resampling and data augmentation. SMOTE [6] combines over-sampling the minority classes and under-sampling the majority classes to achieve better classifier performance. BalanceCascade [26] trains the learners sequentially, where in each step, the majority class examples that are correctly classified by the currently trained learners are removed from further consideration. OOB and UOB [41, 42] build an ensemble model overcoming class imbalance in real-time through resampling and time-decayed metrics. Lin *et al*. [23] use a clustering technique during the data preprocessing step for data undersampling. REPAIR [22] learns weights for different classes to re-sample data to remove representation bias. However, in the field of deep face recognition, no attempt has been made to directly establish a class-balanced training set. In this paper, we try to explore the gains of training a face recognition model in the case of absolute fairness in terms of the number of samples between all classes.

Face Recognition Datasets Large-scale face recognition training datasets are critical to recognition performance. CASIA-WebFace [46] is the first large-scale dataset for efficient deep face recognition. VGGFace2 [5], MS-Celeb-1M [13] and MegaFace2 [31] provide over one million training images, pushing the face recognition benchmark performance to a new level. However, existing large-scale datasets are usually composed of in-the-wild face images collected from the web rather than in the laboratory, which makes them suffer from the imbalance of classes. Figure 2 shows the normalized identity distribution of four widely-used training datasets, *i.e*., CASIA-WebFace [46], MS1M-IBUG [10] (cleaned from MS-Celeb-1M [13]), MegaFace2 [31] and VGGFace2 [5]. The curves are drawn by arranging all classes according to the number of their images in descending order. The long tail problem of MegaFace2 [31] is the most serious. VGGFace2 [5] handles this problem better, but it only contains 9,131 classes.

Unfortunately, previous studies mostly keep the natu-

Dataset	# of photos	# of subjects	STD
MillionCelebs	18.8M	636.2K	-
MS1M-IBUG	3.8M	84.2K	-
CASIA-WebFace	494.4K	10.6K	-
Long-tail	500.0K	10.0K	65.5
Uniform	500.0K	10.0K	0.0
BUPT-CBFace	500.0K	41.6K	0.0

Table 1: Face datasets used in the experiments for training recognition models. The MillionCelebs dataset [47] is used to extract subsets of 500k images under different conditions.

ral distribution of the web-collected datasets for deep face recognition model training, and the impact of the dataset structure has not been well studied. One possible reason is that the data size is too small to select partial data for effective training, so researchers tend to use all available data. However, it is meaningful to explore whether the recognition model can benefit from better data distribution. It is possible that, by adjusting sample distribution and selecting classes, medium-scale datasets also achieve comparable training effects of a larger-scale one. Besides, such an "efficient" dataset may benefit the training of a lightweight model that is important for industrial applications.

3. How Does Class Balance Help Training?

In this section, we explore the effects of class imbalance and data structure distributions on face recognition performance through experiments. Specifically, we hope to answer the following three questions:

1. Can a uniformly distributed dataset with balanced classes lead to better recognition performance?
2. Does the class imbalance contribute to recognition biases such as racial bias and gender bias?
3. Can training classes be deliberately selected to improve recognition performance?

For a fair comparison, we study these issues by training deep models with training sets of same level data size as CASIA-WebFace [46]. The training sets are built by extracting samples from MillionCelebs [47], which is a well-cleaned long-tailed face dataset with abundant images and identities so that it is suitable for extracting such subsets for specific studies. Table 1 shows the information of related datasets. For data preprocessing, we use MTCNN [45] face detector to localize five landmarks, then align and crop the images to 112×112 face warps. The images are normalized by subtracting 127.5 and being divided by 128. In training, all input images are horizontally flipped with probability 0.5 for data augmentation. All experiments in this paper are implemented by MXNet [8].

3.1. Experimental Setup

Evaluation Metrics Face recognition performance is evaluated on 10-fold verification sets LFW [18], CALFW [49], CPLFW [48], CFP [35] and AgeDB [30]. The RFW [40] benchmark is used to test model performance on four kinds of races so that the degree of algorithm fairness can be measured by the standard deviation (STD) of the four races. Moreover, the MegaFace Challenge1 [21] evaluates face recognition performance under one million distractors, and the IJB-C [29] benchmark evaluates template-wise face recognition performance. CMC curves and Rank-1 are adopted to evaluate face identification performance, while ROC curves and TPR at given FPR are adopted to evaluate face verification performance.

CNN Architecture and Loss Function Many CNN architectures [7, 16, 17] and loss functions [9, 38, 43] are developed to promote the face recognition ability. In this paper, ResNet-X [16] and MobileNetV2 [34] are deployed to test data performance at different network scale. ResNet-X refers to a ResNet [16] architecture with X layers. For measuring training loss, the cross-entropy Softmax loss L_S and large-margin ArcFace loss [9] L_A are used:

$$L_S = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{1}$$

where $x_i \in \mathbb{R}^d$ denotes the deep feature of the i-th sample, y_i denotes the label of x_i. W is the weight matrix and b is the bias term. N and n is batch size and class number. For simplicity, we fix $b = 0$ as in many works [9, 12, 25, 37].

$$L_A = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{2}$$

where θ_j is the angle between W_j and x_i, m is the angular margin that aims to enlarge the gradient towards the class prototypes, and s is the scale of l_2 normalized feature vectors. m and s are set 0.5 and 64.

Training All experiments are performed on two NVIDIA GTX 1080Ti GPUs with batch size 256. The initial Stochastic Gradient Descent (SGD) learning rate is set 0.1, then is divided by 10 three times when the loss plateaus. The hyper-parameters weight decay and momentum are set 0.0005 and 0.9, respectively.

3.2. Face Recognition Accuracy

To study the class imbalance issue of existing face recognition training sets, we build a synthetic set called "Long-tail" by simulating their long tail distribution. Specifically, "Long-tail" is extracted from a big dataset in the following three steps:

Architecture	Loss	Dataset	LFW	CALFW	CPLFW	CFP	AgeDB	Avg.	U. - L.
ResNet-18	Softmax	Long-tail	98.67	86.70	78.98	92.31	90.53	89.44	0.20
		Uniform	98.78	87.18	79.55	92.27	90.40	**89.64**	
	ArcFace	Long-tail	99.47	92.80	83.58	93.76	94.97	92.92	0.21
		Uniform	99.45	93.07	84.42	93.56	95.13	**93.13**	
MobileNetV2	Softmax	Long-tail	98.40	85.93	76.57	90.81	89.60	88.26	0.24
		Uniform	98.60	86.37	76.85	91.10	89.57	**88.50**	
	ArcFace	Long-tail	98.90	90.35	81.43	92.24	92.65	91.11	0.13
		Uniform	99.15	90.83	81.18	91.69	93.33	**91.24**	

Table 2: Face recognition accuracy (%) of the "Long-tail" and the "Uniform" with different architectures and loss functions. "Avg." means average accuracy on the 5 test sets. "U. - L." means how the "Uniform" surpasses "Long-tail" on average.

1. **Simulate the long tail curve.** Simulate the long tail shape of a dataset D and scale its distribution to i identities, with a total number of m images. Then this long tail curve can be expressed using a discrete function $S(k), k = 1, 2, \cdots, i$, where

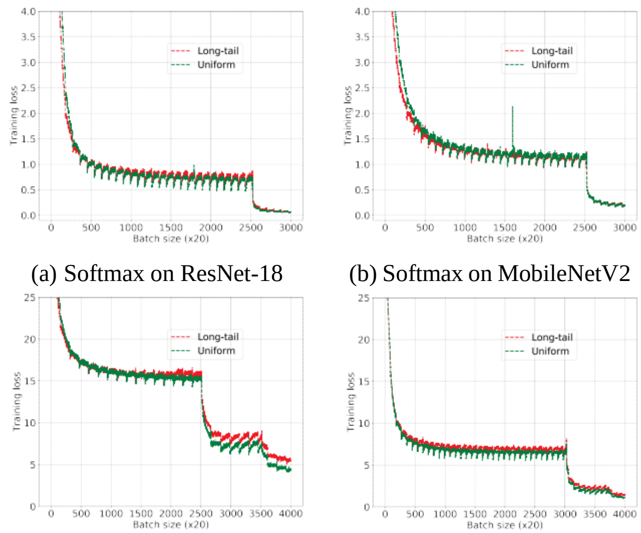
$$\sum_{k=1}^{i} S(k) = m \tag{3}$$

2. **Determine source distribution.** In a big dataset B, intercept its head identities with the number of face images greater than n.

3. **Extract subset.** Randomly select i identities from the intercepted head part, rearrange them from 1 to i. For identity k, randomly select $S(k)$ images to generate the subset.

To construct the "Long-tail" dataset, we set $i = 10,000$, $m = 500,000$, and $n = 90$ to ensure that every class has enough images to choose from. B and D refers to MillionCelebs [47] and CASIA-WebFace [46], respectively. A comparative "Uniform" dataset is also constructed by using the *same* classes and data size as the "Long-tail" but each class has 50 randomly selected images. By controlling the variables, we ensure that the accuracy differences between the experimental results of the two datasets depend only on whether the classes are balanced. Table 2 compares performances of deep models on five validation sets by different architectures and loss functions. It is observed that class-balanced training data can effectively enhance face recognition performance on average for all tested architectures and loss functions. For example, when training a MobileNetV2 [34] model with Softmax as loss function, Uniform outperforms Long-tail on four out of the five test sets and increases the mean accuracy by 0.24%. When the ResNet [16] architecture or ArcFace loss [9] is used, the class-balanced dataset also achieves higher accuracy.

Figure 3 shows the loss decreasing curves of Long-tail (red) and Uniform (green). It is observed that the Softmax loss of an imbalanced training set converges faster at the beginning. This is because the majority classes can quickly converge due to its large number of training samples, but this does not help improve the final training effect. On the other hand, the ArcFace loss [9] of Uniform can decrease lower than Long-tail. This shows that balanced classes are easier to fit into the large margin feature space, so the model performance is also improved as expected.



(a) Softmax on ResNet-18

(b) Softmax on MobileNetV2

(c) ArcFace on ResNet-18

(d) ArcFace on MobileNetV2

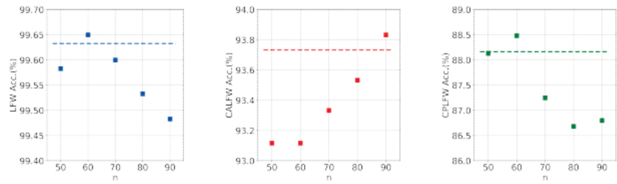
Figure 3: Comparison of loss curves. Softmax loss of a class-imbalanced dataset decreases earlier. ArcFace loss [9] of a class-balanced dataset decreases lower.

3.3. Bias in Recognition

In Section 1, it is analyzed that imbalanced training sets hinder the fairness among people of different races and genders, resulting in bias in face recognition. The Long-tail and Uniform sets are helpful to explore existence of such bias.

#	Dataset	Caucasian	African	Indian	Asian	STD
1	Long-tail	92.17	82.25	88.47	85.02	3.72
	Uniform	**93.73**	**83.97**	**88.83**	**86.23**	**3.63**
2	Long-tail	90.42	78.43	85.77	83.18	4.33
	Uniform	**90.57**	**79.05**	**86.02**	**83.50**	**4.17**

Table 3: Performance on the RFW benchmark of ResNet-18 [16] (#1) and MobileNetV2 [34] (#2) trained with ArcFace loss [9]. The class-balanced training set can reach **higher accuracy** (%) on all races with **lower standard deviation**. Therefore, the fairness of race is guaranteed.



(a) LFW [18] (b) CALFW [49] (c) CPLFW [48]

Figure 4: The recognition accuracy on three verification sets with the variety of n. The dashed lines represent the results of model combination of $n = 60$ and $n = 90$.

RFW RFW [40] is a benchmark for measuring face recognition accuracy on four kinds of races, *i.e.*, Caucasian, Asian, Indian and African, which can be used to test the bias problem in face recognition. Table 3 reports the results on RFW [40] of ResNet-18 [16] (#1) and MobileNetV2 [34] (#2) models trained by Long-tail and Uniform with ArcFace loss [9]. It is observed that in the two comparative experiments, Uniform not only performs better than Long-tail on any race but also has a smaller standard deviation in the accuracy of the four races, which means, the difference between recognition accuracy for the four races is even smaller. It is worth noting that we achieve this improvement by only adapting the sample distribution to a uniform distribution without using any race-related information to deliberately select the classes. This confirms that the class-balanced training is of great benefit to the fairness of deep face recognition.

3.4. Class Selection

For a class-balanced training set, there is still great optimization potential. For example, the composition of the classes in the dataset can be carefully designed to better fit the spatial distribution of the human face. When constructing the Uniform dataset, n is the main variable controlling the choice of classes. It is observed that with the change of n, although the generated datasets are in the same shape and size, the training effects are totally different. As is shown in Figure 4, training ResNet-34 [16] with ArcFace [9] as loss function, LFW [18] and CPLFW [48] peak at $n = 60$, but CALFW [49] peaks at $n = 90$.

Noted that a big n only considers majority classes while a small n can consider more minority classes, this phenomenon indicates that majority classes perform better on the "cross-pose" recognition task, and adding a certain proportion of the minority classes can improve the performance on "cross-age" recognition task. Considering the data collection process, the majority classes are often composed of famous people, who have more pictures on the web, so the collecting recall is lower, and the photos after his fame will be collected first, which means there is more cross-pose information. On the contrary, the minority classes are collected with high recall, including his pictures of different ages, so the "cross-age" performance is improved. This interesting observation gives guidance on the selection of training data. According to the application scenario, there should be different emphasis on the majority or minority. For comprehensive optimization, it is necessary to have a compromise or deploy model combination. The dashed lines in Figure 4 show one possible model combination attempt: we simply concatenate the output features of the $n = 60$ and $n = 90$ models, then the balance of the recognition accuracy on different tasks is reached.

4. BUPT-CBFace: Class-Balanced Training

Following previous observations, a novel face recognition training set BUPT-CBFace is constructed to help convenient yet effective deep face recognition models training.

4.1. Balance Between Breadth and Depth

There are many studies [1, 4, 39] that discuss whether the training set should have more classes or more images in one class, but their answers are not the same. We define two parameters for a class-balanced dataset:

Breadth The number of identities.

Depth The number of images per identity.

Keeping the data distribution and data size unchanged, we can observe how the variation of breadth and depth affect training. To this end, we set $n = 60$ and select seven kinds of setups (breadth from 5,000 to 62,500) to build training sets, in which the identities and images are still randomly selected. Table 4 shows the recognition accuracy of training ResNet-34 [16] with ArcFace loss [9] and these datasets. Figure 5 draws the average verification accuracy and final training loss vary with breadth. It is observed that when the data size remains constant, the variety of dataset shape plays an important role in training. Starting from 5,000, each increase in breadth brings a significant accuracy enhancement. However, the excessive number of

Breadth	Depth	LFW	CALFW	CPLFW	Avg.
5,000	100	99.28	91.72	84.85	91.95
10,000	50	99.65	93.12	88.48	93.75
20,000	25	99.57	94.15	89.22	94.31
31,250	16	**99.68**	94.18	89.58	94.48
41,667	12	99.63	**94.60**	89.90	**94.71**
50,000	10	99.50	94.20	**90.30**	94.67
62,500	8	99.58	94.50	89.75	94.61

Table 4: At the same data scale (500k images), proper breadth and depth of a class-balanced dataset can significantly improve the recognition accuracy (%).

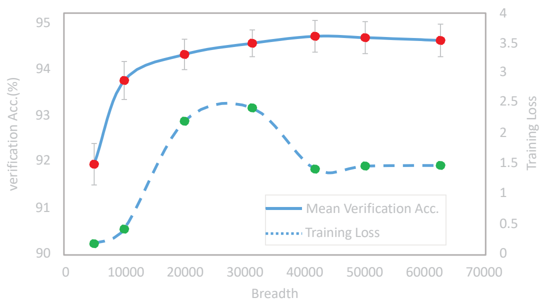


Figure 5: The variations of mean verification accuracy and final training loss with dataset breadth.

classes also leads to insufficient depth, which inhibits the training effect. This shows that there is a demand for both the number of images and the number of identities in the deep model learning, and the performance limit lies in the side of the shortboard. Finally, the average recognition accuracy peaks at the breadth of around 40,000 to 50,000.

The final loss curve in Figure 5 is also intriguing. As is observed, the loss keeps very small when the breadth is less than 15,000, which means that a small number of classes is easy to fit. When the breadth is greater than 15,000, the feature space is gradually saturated so the loss increases. However, when the data breadth reaches 30,000 or more, the loss falls back to a medium level at around 1.5 because the number of images in one class is smaller, so that they are easier to fit into the feature space. It confirms that deep learning can gain from depth and breadth, separately.

Comprehensive consideration, we regard the dataset with 41,667 classes and 12 images per class as the BUPT-CBFace dataset. Figure 6 shows images of five classes in BUPT-CBFace. In addition to its balanced classes, it also strikes a balance between depth and breadth. In recognition tasks, BUPT-CBFace not only considers the balance between cross-age and cross-pose recognition but also reduces recognition bias to certain extent. Due to its small size and good recognition performance, BUPT-CBFace can

Figure 6: Images of five classes in BUPT-CBFace. There are twelve images in each class with rich facial information such as poses, lighting and expressions.

be easily trained on a single NVIDIA GTX 1080Ti GPU to achieve the same level results as large-scale parallel training like training on the MS-Celeb-1M [13] dataset.

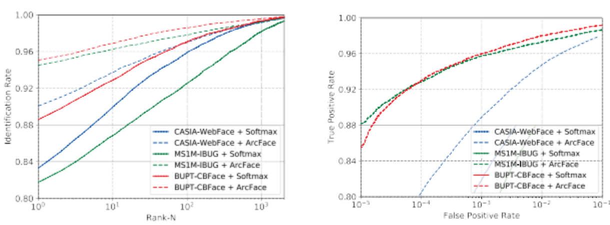
4.2. Evaluation Results

We evaluate the benchmark performance of BUPT-CBFace comparing with the other two public training sets CASIA-WebFace [46] and MS1M-IBUG [10] under the same training environments. Table 5 reports face recognition accuracy of the ResNet-50 [16] models trained with Softmax or ArcFace loss [9]. BUPT-CBFace reaches the highest accuracy on three of the five verification sets, even more than MS1M-IBUG [13] that has nearly eight times more face images of it. Especially on the cross-pose test set CPLFW [48] and CFP [35], BUPT-CBFace surpasses MS1M-IBUG [13] by 5.75% and 5.27% with ArcFace loss [9], which means that it contains a large amount of pose-related information. BUPT-CBFace also obtains the highest average accuracy of the 5 verification sets, surpassing MS1M-IBUG [13] by 2.10% to reach 95.60%.

IJB-C The IJB-C benchmark [29] tests template-wise face recognition performance. Training ResNet-50 [16] with Softmax or ArcFace loss [9], the verification TPR at 1e-4 FPR and identification Rank-1 on IJB-C [29] are reported in Table 5. BUPT-CBFace reaches higher accuracy than CASIA-WebFace [46] and MS1M-IBUG [13] on all tests. Trained with ArcFace loss [9], BUPT-CBFace reaches 93.95% identification accuracy and 92.99% verification accuracy. Figure 7 shows the corresponding CMC and ROC curves. In Figure 7a, BUPT-CBFace has the highest Rank-N accuracy for any N in all comparisons, which shows its strong identification ability. In Figure 7b, when trained with ArcFace loss [9], MS1M-IBUG [13] can reach higher TPR at 1e-5 FPR. This shows that when the requirements for identifying negative pairs become stricter, the number of training samples becomes more important. However, in or-

Loss	Training Dataset	Size(M)	LFW	CALFW	CPLFW	CFP	AgeDB	Avg.	IJB-C	
									Id.	Ver.
Softmax	CASIA-WebFace	0.5	98.77	86.38	80.88	92.37	88.83	89.45	79.82	69.23
	MS1M-IBUG	**3.8**	98.97	90.92	79.98	87.46	92.38	89.94	79.05	56.53
	BUPT-CBFace	0.5	99.05	89.67	83.32	92.93	90.47	**91.09**	**85.73**	**81.21**
Arcface	CASIA-WebFace	0.5	99.52	92.55	87.17	95.33	95.20	93.95	88.05	80.44
	MS1M-IBUG	**3.8**	99.62	94.85	84.95	90.97	97.13	93.50	93.54	92.86
	BUPT-CBFace	0.5	99.65	94.80	90.70	96.24	96.60	**95.60**	**93.95**	**92.99**

Table 5: Face recognition accuracy (%) of different datasets with ResNet-50 [16] as backbone and Softmax or ArcFace [9] as loss function. Training with BUPT-CBFace can obtain a better performance than other two datasets with smaller data size.



(a) CMC curves (b) ROC curves

Figure 7: Identification CMC curves and verification ROC curves on the IJB-C [29] benchmark.

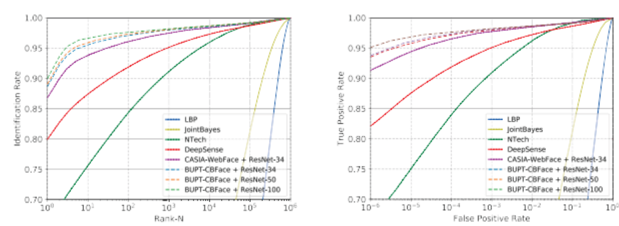
dinary scenes, a medium-scale class-balanced training set is more suitable for face recognition tasks.

MegaFace MegaFace challenge 1 [21] evaluates face recognition performance under one million distractors. It measures TPR at 1e-6 FPR for verification and Rank-1 retrieval performance for identification. Adopting Face-Scrub [32] as probe set, Table 6 shows BUPT-CBFace and comparative methods on the official leaderboard under the "small" protocol. Corresponding CMC and ROC curves of the highest official published methods are drawn in Figure 8. BUPT-CBFace and CASIA-WebFace [46] trained with ArcFace loss [9] are included for comparison. Training the same ResNet-34 [16] architecture with ArcFace loss [9], BUPT-CBFace exceeds CASIA-WebFace [46] by 2.13% identification accuracy and 2.03% verification accuracy. When training ResNet-100 [16] architecture with ArcFace loss [9], BUPT-CBFace reaches state-of-the-art performance on both face identification and verification tests under small protocol, outperforming CVTE V2 by 1.25% identification accuracy and 0.78% verification accuracy.

RFW In Section 3.3, it is proved that a class-balanced training set can obtain higher accuracy and lower recognition bias for different races. Table 7 compares training results of CASIA-WebFace [46] and BUPT-CBFace on RFW [40]. For fairness, MS1M-IBUG [13] is excluded for comparison because RFW [40] is a subset of MS-Celeb-1M

Methods	Id.	Ver.	Protocol
DeepSense	70.98	82.85	small
SphereFace [25]	75.77	90.05	small
FaceAll V2	76.66	77.61	small
GRCCV	77.68	74.89	small
FUDAN	77.98	79.20	small
CVTE V2	78.32	94.42	small
CASIA-WebFace + ResNet-34	76.22	91.48	small
BUPT-CBFace + ResNet-34	78.35	93.45	small
BUPT-CBFace + ResNet-50	78.75	93.81	small
BUPT-CBFace + ResNet-100	**79.57**	**95.20**	small

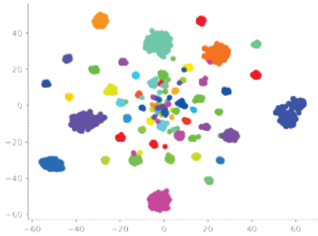
Table 6: FaceScrub [32] results (%) of the MegaFace challenge 1 [21] under small protocol. BUPT-CBFace reaches state-of-the-art performance on the official leaderboard.



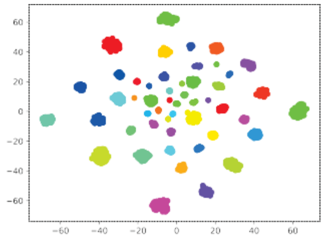
(a) CMC curves (b) ROC curves

Figure 8: Identification CMC curves and verification ROC curves of all official published methods under the MegaFace challenge 1 [21] small protocol.

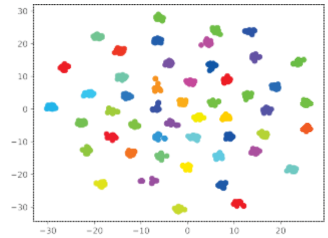
and the identity duplication can cause serious interference. It is observed that the accuracy of BUPT-CBFace in all races greatly exceeds that of CASIA-WebFace [46]. For example, the ArcFace [9] model trained by BUPT-CBFace are 6.25% higher on the worst performed Asian faces, and 2.82% higher on the best performed Caucasian faces. Therefore, differences in accuracy between races are also reduced. The standard deviation of different races decreases to 1.61 from



(a) CASIA-WebFace [46]



(b) MS1M-IBUG [13]



(c) BUPT-CBFace

Figure 9: Visualization of randomly selected 50 classes of three datasets on t-sne [27] feature space.

#	Dataset	Caucasian	African	Indian	Asian	STD
1	CASIA-WebFace	87.65	76.38	80.98	76.73	4.54
	BUPT-CBFace	**89.98**	**81.93**	**85.30**	**83.38**	**3.04**
2	CASIA-WebFace	94.43	88.53	89.85	86.88	2.81
	BUPT-CBFace	**97.25**	**93.53**	**94.87**	**93.13**	**1.61**

Table 7: Face recognition accuracy (%) and standard deviation on the RFW [40] benchmark of ResNet-50 [16] trained with Softmax loss (#1) and Arcface loss [9] (#2).

Datasets	Angle (Mean)	Norm (STD)
CASIA-WebFace	15.29	0.13
MS1M-IBUG	17.34	4.96
BUPT-CBFace	**14.01**	**0.03**

Table 8: Statistics of weight matrix of ResNet-50 [16] models trained with ArcFace loss [9] and different datasets. "Angle (Mean)" refers to the mean of angles between W_j and the corresponding embedding feature center. "Norm (STD)" refers to the standard deviation of $|W_j|$.

2.81 of CASIA-WebFace [46], so that the recognition bias problem is greatly alleviated.

4.3. Analysis and Discussion

Weight Matrix As many studies [9, 12] show, the weight matrix W in Equation 1 can reflect the training quality of the model. Table 8 shows the mean of the angle between W_j and the corresponding embedding feature center and standard deviant of $|W_j|$ for all classes of three datasets. First, the angle between W_j and centers of feature embeddings x_i of samples belong to class j shows how the training samples are fitted to the model. In the model trained with BUPT-CBFace, the mean angle is 1.18° smaller than CASIA-WebFace [46] and 3.33° smaller than MS1M-IBUG [13], which means the output feature embeddings of training classes are closer to W_j and therefore more representative, and the model converges better on the training set. On the other hand, a majority class j usually leads to a larger weight vector norm $|W_j|$, while a minority class usually leads to a smaller weight vector norm. In this case, if the vectors are not l_2 normalized, the decision boundary is shifted towards the smaller-norm classes (see analysis in [11] and [12]). When training with the class-balanced BUPT-CBFace, weight vector norms $|W_j|$ have very small standard deviation 0.03, which is 4.93 smaller than that of MS1M-IBUG [13] and 0.10 smaller than that of CASIA-WebFace [46]. Therefore, even if no additional constraints are added on the norms of weight vectors, the norms of different classes in BUPT-CBFace tend to be more consistent.

Visualization In Figure 9, we visualize the feature distributions of randomly selected 50 classes from three training sets, where each class is represented by one color. The ResNet-50 [16] models with ArcFace loss [9] are used to extract deep features, and t-sne [27] is used to generate visual embeddings. It is observed that both CASIA-WebFace [46] and MS1M-IBUG [13] have extremely uneven sample spaces. On the one hand, the majority classes occupy a large volume of space, on the other hand, the minority classes are squeezed closer and difficult to separate. So the class imbalance causes biases in the recognition effect between the majority and the minority. In contrast, the spacial volumes of different classes in BUPT-CBFace are basically equal, so the recognition fairness is guaranteed.

5. Conclusion

In this paper, we study the impact of class balance and data structures on deep face recognition. A class-balanced face recognition training set BUPT-CBFace is built by carefully adjusting data shapes and classes. BUPT-CBFace has a significant recognition performance and fairness improvement compared to long-tailed datasets of the same scale. Moreover, BUPT-CBFace can be easily trained on a single NVIDIA GTX 1080Ti GPU to achieve the same level results as large-scale parallel training, which is very friendly to many institutes. BUPT-CBFace is publicly available as an alternative option to the existing long-tailed datasets.

References

[1] Ankan Bansal, Carlos Castillo, Rajeev Ranjan, and Rama Chellappa. The do's and don'ts for cnn-based face verification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2545–2554, 2017.

[2] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[3] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.

[4] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. Celeb-500k: A large training dataset for face recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2406–2410. IEEE, 2018.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 67–74. IEEE, 2018.

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[7] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018.

[8] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

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

[10] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 60–68, 2017.

[11] Zhengming Ding, Yandong Guo, Lei Zhang, and Yun Fu. Generative one-shot face recognition. *arXiv preprint arXiv:1910.04860*, 2019.

[12] Yandong Guo and Lei Zhang. One-shot face recognition by promoting underrepresented classes. *arXiv preprint arXiv:1707.05574*, 2017.

[13] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[14] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[15] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[16] 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.

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[18] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 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*, 2008.

[19] Isabelle Hupont and Carles Fernández. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019.

[20] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

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

[22] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019.

[23] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409:17–26, 2017.

[24] Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10052–10061, 2019.

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[26] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[28] Iacopo Masi, Anh Tuan Tran, 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.

[29] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018.

[30] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *Computer Vision and Pattern Recognition Workshops*, pages 1997–2005, 2017.

[31] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3406–3415. IEEE, 2017.

[32] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*, pages 343–347. IEEE, 2014.

[33] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015.

[34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

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

[36] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *European Conference on Computer Vision*, pages 780–795. Springer, 2018.

[37] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

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

[39] Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.

[40] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 692–702, 2019.

[41] Shuo Wang, Leandro L Minku, and Xin Yao. A learning framework for online class imbalance learning. In *2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL)*, pages 36–45. IEEE, 2013.

[42] Shuo Wang, Leandro L Minku, and Xin Yao. Resampling-based ensemble methods for online class imbalance learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(5):1356–1368, 2014.

[43] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[44] Yue Wu, Hongfu Liu, Jun Li, and Yun Fu. Deep face recognition with center invariant loss. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 408–414, 2017.

[45] Jia Xiang and Gengming Zhu. Joint face detection and facial expression recognition with mtcnn. In *Information Science and Control Engineering (ICISCE), 2017 4th International Conference on*, pages 424–427. IEEE, 2017.

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

[47] Yaobin Zhang, Weihong Deng, Mei Wang, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. Global-local gcn: Large-scale label noise cleansing for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[48] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5, 2018.

[49] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[50] Yutong Zheng, Dipan K Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5089–5097, 2018.

[51] Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Unequal-training for deep face recognition with long-tailed noisy data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7812–7821, 2019.

[52] Erjin Zhou, Zhimin Cao, and Qi Yin. Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv preprint arXiv:1501.04690*, 2015.