IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 August 2022, accepted 1 September 2022, date of publication 12 September 2022, date of current version 21 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3206040

RESEARCH ARTICLE

Exploiting the Tail Data for Long-Tailed Face Recognition

SONG GUO[1], RUJIE LIU[1], MENGJIAO WANG[1], MENG ZHANG[1], SHIJIE NIE[1], SEPTIANA LINA[2], AND NARISHIGE ABE[2]

[1]Fujitsu Research and Development Center Company Ltd., Beijing 100022, China
[2]Fujitsu Laboratories Ltd., Kawasaki, Kanagawa 211-8588, Japan

Corresponding author: Song Guo (guosong@fujitsu.com)

ABSTRACT Long-tailed distribution generally exists in large-scale face datasets, which poses challenges for learning discriminative feature in face recognition. Although a few works conduct preliminary research on this problem, the value of the tail data is still underestimated. This paper addresses the long-tailed problem from the perspective of maximally exploiting the tail data. We propose a Joint Alternating Training (JAT) framework to learn discriminative feature from both the long-tailed data and the tail data by using alternating training strategy. JAT consists of two branches: 1) the long-tailed data branch is adopted to learn the universal discrimination information from the whole long-tailed data with instance-balanced sampling. 2) the tail data branch is designed to exploit the discriminative information in the tail data with class-balanced sampling. To compensate the insufficient samples and lack of intra-class variations, we apply data augmentation (DA) to the tail data. We further propose margin-based mixup (MarginMix) for data augmentation, which can deal with the nonlinearity of margin-based softmax loss and stabilize the training process in mixup. Furthermore, we obtain the best combination of strategies (i.e., JAT+DA+ MarginMix) for long-tailed face recognition, which can maximally exploit the discriminative information in the tail data while retaining the universal discrimination learned from the long-tailed data. Extensive experiments on 8 face datasets demonstrate that our proposed methods and combination of strategies can effectively address the long-tailed problem in face recognition.

INDEX TERMS Face recognition, convolutional neural network, long-tailed distribution, margin softmax loss, data augmentation.

I. INTRODUCTION

Deep face recognition has made significant development in recent years. Besides the evolution of network architecture and a variety of loss functions [5], [10], [11], the growing of face datasets has greatly promoted the development of face recognition. It has been proved that larger training dataset can enhance the performance of face model, because the model can learn more discriminate feature when more identities are provided [1], [3]. Therefore, much effort has been put into building large-scale face datasets recently [1], [2], [3].

Most large-scale face datasets in real-world exhibit a long-tailed distribution, in which a small number of identities account for most of the samples (the head data), while many other identities only have relatively few face images (the tail data). A common problem in training on long-tailed dataset is that the head identities are properly trained, but the tail identities are under-represented. Consequently, this will bring difficulty in learning feature with good representation and generalization ability for face recognition. According to [5], the model trained on a part of the long-tailed dataset (remove 20% or 50% of the tail data) obtains higher accuracy than that learned on the whole dataset. On the other hand, if too much tail data (70% or more) is discarded, the model performance will drop. This preliminary research reveals the fact that the tail data is a double-edged sword, i.e., if used properly, it can boost the performance of the trained model; otherwise, it will bring negative effect in learning discriminative feature.

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

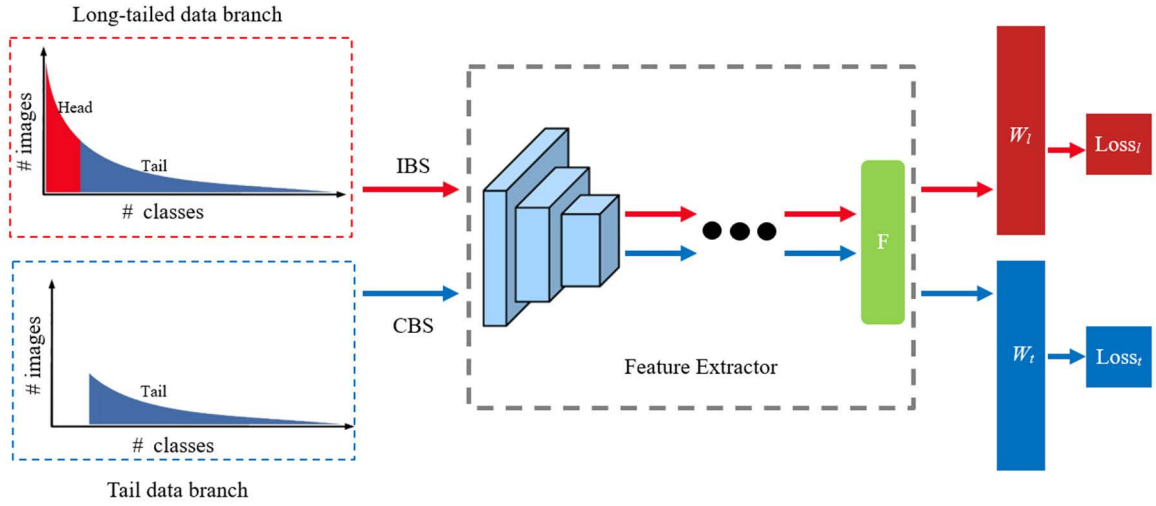


FIGURE 1. Framework of the proposed JAT framework. It consists of two data branches: 1) the long-tailed data branch is used to learn the universal information from the whole long-tailed dataset with instance-balanced sampling. 2) the tail data branch is designed to maximally exploit the discriminative information in the tail data by using class-balanced sampling. The alternating training strategy is adopted to learn more discriminative feature from the two data branches jointly.

Therefore, how to make the best of the tail data is a key issue towards training better face recognition model.

The long-tailed distribution is a common issue in many visual recognition tasks. Compared with other tasks, the long-tailed face recognition is featured in three aspects: (1) The long-tailed face dataset is much larger in scale, e.g., WebFace260M [1] contains 4M identities in contrast to 1K classes in ImageNet-LT [12]. (2) The imbalance ratio (the ratio of image number between the largest class and the smallest class) of face datasets is much higher, e.g., the imbalance ratio is up to 1,500 in MegaFace Challenge 2 (MF2) [4] while it is only 256 in ImageNet-LT. (3) Face recognition is generally an open-set problem, so it is essentially a metric learning problem instead of a classification problem. Due to these characteristics, it is usually infeasible to directly apply the existing solutions in long-tailed visual recognition to face recognition.

We address the long-tailed face recognition problem from the perspective of maximally exploiting the tail data in this paper. A Joint Alternating Training framework is proposed to learn more discriminative feature from both the long-tailed data and the tail data by using alternating training strategy. As shown in Fig. 1, there are two data branches in our JAT framework. The long-tailed data branch is responsible to learn the universal information from the whole long-tailed dataset. The tail data branch is designed to exploit the discriminative information in the tail data, which cannot be discovered in the regular training on the long-tailed data due to the domination of the head data. Moreover, different sampling strategies are adopted for different data branches. Specifically, instance-balanced sampling (IBS), i.e., each training sample is selected once in one epoch with equal probability, is adopted for the long-tailed data branch. The class-balanced sampling (CBS), i.e., each class is selected with equal probability, is used for the tail data branch. By using the alternating training strategy, JAT can effectively exploit the discriminative information in the tail data while retaining the universal discrimination learned from the whole long-tailed dataset.

Due to the insufficient number of samples, the intra-class variance in the tail data is usually limited, which leads to unreasonable squeeze of the tail classes in feature space. To alleviate this problem, we propose MarginMix for data augmentation, which can deal with the nonlinearity of margin-based softmax loss and stabilize the training process in mixup. MarginMix applies different loss functions on different parts of the linearly mixed label for the mixed sample. More specifically, MarginMix adopts margin-based softmax loss for the label with larger mixing coefficient and traditional softmax for the label with smaller coefficient. To further enlarge the intra-class variance in the tail data, a larger mixing coefficient is assigned to the sample with label in minority class. Furthermore, the MarginMix is combined with other data augmentation techniques to generate face images with more intra-class variations.

The major contributions of this work are the followings:

1) We address the long-tailed problem in face recognition from the perspective of maximally exploiting the tail data. We are the first to explore how to maximally exploit the long-tailed data and the tail data jointly in the deep face recognition literature.

2) We propose a JAT framework to learn discriminative feature from both the long-tailed data and the tail data jointly. JAT aims to exploit the discriminative information in the tail data while retaining the universal discrimination learned from the whole long-tailed dataset, which is achieved by alternating training on the long-tailed data branch and the tail data branch.

3) We propose MarginMix to deal with the nonlinearity of margin-based softmax loss and stabilize the training process in mixup. Combined with other data augmentation techniques, MarginMix can generate face images with much more intra-class variations for the tail data.

4) We further combine different strategies together to obtain the best combination for long-tailed face recognition. The experimental results validate the effectiveness of our proposed methods and combination of strategies in dealing with the long-tailed problem in face recognition.

II. RELATED WORKS

A. LONG-TAILED FACE RECOGNITION

Most methods in face recognition focus on designing various loss functions or building larger and larger training datasets, whereas few works pay attention to the long-tailed problem. By reducing the intra-class variations and enlarging the inter-class discrepancy simultaneously, Range loss [5] can relieve the long-tail effect in face recognition. To balance the distributions of different identities in the feature space, a feature transfer framework is proposed to generate samples in feature level for the tail identities by transferring the intra-class variance from head identities [6]. Fair Loss [8] and AdaptiveFace [7] are designed to learn adaptive margins for different classes according to their number of samples. By enhancing the intra-class compactness of tail data, they can improve the generalization capability of the learned deep features. An unequal training strategy is proposed to deal with the head data and the tail data separately, in order to learn the intra-class variations and inter-class discriminative information from the head classes and the tail classes respectively [9]. This strategy focuses on the difference of the head and tail data but ignores the universal information in the whole dataset. Although these methods can obtain performance improvement for long-tailed face recognition, they often suffer from high complexity or sensitivity to hyper-parameters in the training process. Furthermore, the discriminative information in the tail data is not fully exploited in the existing methods.

B. LONG-TAILED VISUAL RECOGNITION

Visual recognition on the challenging long-tailed dataset has been comprehensively studied in the literature. Traditional approaches employ the class re-balancing strategies, such as data re-sampling (e.g., over-sampling for the tail classes, under-sampling for the head classes) and cost-sensitive re-weighting (e.g., assign variant weights to different classes or samples) [17], [19]. Recent methods explore other learning paradigms for long-tailed recognition, such as transfer learning [6], meta-learning [13], metric learning [7], [8], two-stage training [14], [15] and self-supervised learning [16], etc. In addition to these learning methods, data augmentation approaches are also widely used to deal with the long-tailed recognition problem. As a representative data augmentation technique, mixup [20] shows its ability to improve the generalization and robustness of the trained model for long-tailed visual recognition. Although these approaches can achieve accuracy improvements on public long-tailed visual datasets, e.g., CIFAR-10-LT [17], CIFAR-100-LT [17], iNaturalist 2018 [18] and ImageNet-LT [12], etc., their feasibilities in face recognition have not been well studied. In this paper, we make preliminary attempt to apply simple tricks (e.g., data re-sampling, mixup), which are commonly used and hyper-parameters insensitive, to long-tailed face recognition.

III. METHODLOGY

We first describe our proposed JAT framework in detail, and then we introduce MarginMix for data augmentation.

A. JOINT ALTERNATING TRAINING

We propose JAT framework to learn discriminative feature from both the long-tailed data and the tail data jointly. When training model on long-tailed dataset, the head classes are properly trained, but the tail classes are inadequately trained due to the limited number of samples. Therefore, we design an additional branch to exploit the hidden discriminative information in the tail data in our framework. JAT follows the training paradigm of multi-task learning, where the long-tailed data branch and the tail data branch are designed to learn feature representation from the long-tailed data and the tail data with alternating training strategy. As shown in Fig. 1, two branches share the same base model and weights for deep feature learning. The base model is followed by a fully connected layer (i.e., classifier) in each branch, which maps the deep feature into respective label space. The classification loss for each branch is calculated respectively in the training process. We describe the two data branches and the alternating training process in detail.

1) LONG-TAILED DATA BRANCH

this branch is introduced to learn the universal information from the whole long-tailed dataset. According to [14], [15], the model trained on the original long-tailed dataset with instance-balanced sampling can learn more discriminative and generalizable feature compared with other data re-sampling strategies, so the instance-balanced sampling strategy is employed in this branch. By maintaining the characteristics of original distributions, the universal discriminative information is well learned and retained in the long-tailed data branch.

2) TAIL DATA BRANCH

This branch is designed to discover and exploit the hidden discriminative information in tail data. The tail data is defined as the samples in the tail identities, which have limited number of samples but account for a significant portion in the whole dataset. After removing the head data, the distribution of the tail data becomes much more balanced than that of the long-tailed data, so class-balanced sampling strategy is used in this branch. To compensate the insufficient training samples and lack of intra-class variance of the tail data, face data augmentation is further applied. Equipped with

class-balanced sampling and data augmentation, the discriminative information in the tail data can be maximally exploited.

Algorithm 1 Joint alternating training algorithm

Input: the long-tailed data $\boldsymbol{D_l}$, the tail data $\boldsymbol{D_t}$
Output: the parameters of base model θ, the classifiers' weights of the long-tailed data branch $\boldsymbol{W_l}$ and the tail data branch $\boldsymbol{W}_t$
Require:instance-balanced sampling IBS(), class-balanced sampling CBS($\cdot$), the step ratio between the long-tailed data branch and the tail data branch S_1: S_2, the maximum training epochs E_{max}, the classification loss function CL($\cdot$)
Alternating Training:
for epoch=1 to E_{max}:
  for s_1 =1 to S_1:
    $\text{Batch}_\text{l}(x_l, y_l) \leftarrow \text{IBS}(\boldsymbol{D_l})$
    Feature extraction: $f_l \leftarrow F_{cnn}(\text{Batch}_\text{l}(x_l), \theta)$
    Compute loss: $L_l \leftarrow \text{CL}(\text{softmax}(\boldsymbol{W}_l^\mathbf{T} f_l), y_l)$
    Update model parameters (θ, $\boldsymbol{W_l}$)
  end for
    for s_t =1 to S_2:
    $\text{Batch}_\text{t}(x_t, y_t) \leftarrow \text{CBS}(\boldsymbol{D_t})$
    Feature extraction: $f_t \leftarrow F_{cnn}(\text{Batch}_\text{t}(x_t), \theta)$
    Compute loss: $L_t \leftarrow \text{CL}(\text{softmax}(\boldsymbol{W}_t^\mathbf{T} f_t), y_t)$
    Update model parameters (θ, $\boldsymbol{W_t}$)
  end for
end for

3) JOINT ALTERNATING TRAINING

this strategy is designed to control the training process on two data branches. The overview of the joint alternating training procedure is described in Algorithm 1. For each data branch, a batch of training samples is constructed respectively by using the instance-balanced sampling in the long-tailed data and the class-balanced sampling in the tail data. Batch of samples from different data branches are alternately feed to the base model to acquire their feature vectors, which then pass through the corresponding classifier to calculate the classification loss. The parameters of the base model and corresponding classifier are updated according to the gradients of the loss. Specifically, the parameters of the base model are updated by both of the two branches, but the weights of each classifier are only updated in its own data branch.

Since the scale of the long-tailed data is larger than that of the tail data, we set a step ratio to balance the training process (i.e., update frequency of the parameters) of the two branches, and thus the two data branches are trained alternately by predefined step ratio. We set a larger step for the long-tailed data branch and a smaller step for the tail data branch in implementation, so that the number of training epochs in two branches are almost synchronous. By setting different step ratio, we can adjust the emphasis of the two data branches in training the base model. JAT can learn more discriminative feature by enhancing the model's learning ability towards the tail data while retaining the universal discrimination learned from the whole long-tailed dataset.

B. MARGIN-BASED MIXUP

As mentioned above, insufficient training samples will lead to small intra-class variance in the tail data. To alleviate this problem, data augmentation techniques, including mixup and generic data augmentation methods are adopted to generate samples with more variations for the tail data.

Let (x, y) denote a sample and its label in the training dataset. Based on the assumption that linear interpolations of samples should be labelled by the linear interpolations of their associated labels, a mixed sample $(\tilde{x},\tilde{y})$ is generated by

$$\tilde{x} = \lambda x_i + (1-\lambda)\, x_j \tag{1}$$
$$\tilde{y} = \lambda y_i + (1-\lambda)\, y_j \tag{2}$$

where $\lambda \in [0, 1]$ is the mixing coefficient. The pair of samples (x_i, y_i) and (x_j, y_j) are drawn from the training dataset. Accordingly, the loss of the mixed sample is calculated by the linear weighted summation of two losses on label y_i and y_j, i.e., L_{y_i} and L_{y_j}

$$L_{\tilde{y}} = \lambda L_{y_i} + (1-\lambda)\, L_{y_j} \tag{3}$$

However, this linear assumption does not hold good in the nonlinear margin-based softmax loss function. In margin-based softmax (e.g., ArcFace [10]), margin penalty is added on the target label to learn more discriminative feature by enforcing the intra-class compactness and the inter-class discrepancy. Therefore, the linearly generated label $\tilde{y}$ does not accurately describe the probability of classes that the mixed image $\tilde{x}$ belongs to. Furthermore, margins are added on both the label of y_i and y_j in mixup training, and thus the marginal softmax loss for the mixed sample will be much larger, especially for the label with smaller mixing coefficient. This will cause an unstable training of the neural network and bring difficulty in the convergence of model.

We propose MarginMix to deal with the nonlinearity of margin-based softmax loss and stabilize the training process. The key idea in MarginMix is to apply different loss functions on different parts of the linearly mixed label. Specifically, the margin-based softmax loss is adopt for the label with larger mixing coefficient and the traditional softmax is used for the label with smaller coefficient. Take ArcFace for example, the loss in MarginMix can be formulated as:

$$L_{y_i} = \frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum_{k=1,k\neq y_i}^{n} e^{scos\theta_k}} \tag{4}$$

$$L_{y_j} = \frac{1}{N}\sum_{j=1}^{N} log\frac{e^{scos\theta_{y_j}}}{\sum_{k=1}^{n} e^{scos\theta_k}} \tag{5}$$

$$L_{all} = \lambda L_{y_i} + (1-\lambda)\, L_{y_j} \tag{6}$$

where $\lambda \geq 0.5$ is the mixing coefficient. The coefficient λ is usually sampled from a beta distribution in practice. In the

TABLE 1. The detailed information of the training datasets. The imbalance ratio is the ratio of sample number between the largest class and the smallest class. The percentages of identities and images in tail data are given in parentheses.

Dataset	Long-tailed data		#Images per identity	Imbalance ratio	Tail data	
	#Identities	#Images			#Identities	#Images
ID30K	30K	2M	67	665 (1329/2)	23K (76%)	419K (21%)
Glint360K	360K	17M	47	443 (1329/3)	268K (74%)	4.3M (25%)

case that λ is smaller than 0.5, we simply use $1-\lambda$ to replace λ to ensure $\lambda \geq 0.5$, which is formulated as:

$$\lambda = \begin{cases} \lambda & if\, \lambda \geq 0.5 \\ 1-\lambda & if\, \lambda < 0.5 \end{cases} \tag{7}$$

To further enlarge the intra-class variations of the tail data, we always assign the larger mixing coefficient λ to the sample in minority class, i.e., sample (x_i, y_i) comes from a relatively minority class compared with (x_j, y_j). This strategy is consistent with the observation in AdaptiveFace [7] and Fair Loss [8] that a minority class generally requires a larger margin to enhance its intra-class compactness, so as to improve the generalization of the trained model. Moreover, MarginMix can be combined with other data augmentation techniques to generate face images with larger intra-class variations for the tail class.

IV. EXPERIMENTS

A. DATASETS AND SETTING

1) TRAINING DATASETS

We employ Glint360K [2] as our training dataset. Glint360K contains more than 17M images from 360K subjects with an imbalance ratio of 443, so it is a typical large-scale long-tailed dataset which is suitable for our research. To validate the effectiveness of the proposed methods, we construct a long-tailed training dataset ID30K from a subset of Glint360K. ID30K consists of 2M images of 30K identities with an imbalance ratio of 665.

In JAT training, we need to split the tail data out of the whole long-tailed dataset. In our implementation, the tail identities are defined as those whose number of samples are less than the average number of samples in the whole dataset. Under this definition, the ratios of tail identities and tail images in ID30K are 76% and 21% respectively, while these two ratios in Glint360K are 74% and 25% respectively. The detailed information of the two long-tailed training datasets is listed in Table 1.

2) TESTING DATASETS

We evaluate the face verification accuracy on five face datasets, e.g., LFW [21], CALFW [22], CPLFW [23], AgeDB [24] and CFP-FP [25]. LFW is widely used for performance evaluation on unconstrained scenario. CPLFW and CFP-FP contain large-pose variance, while CALFW and AgeDB contain cross-age variance. Ten-fold verification sets are used to test the face verification accuracy on these five datasets.

In addition, the model performance is also evaluated on three large-scale benchmark datasets, i.e., MegaFace [26], IJB-B [27] and IJB-C [28]. The MegaFace [26] contains two testing protocols, i.e., face verification and face identification with 1M distractors. The IJB-B and IJB-C benchmarks evaluate face verification on mixed-media, i.e., still image vs video template.

We adopt the refined version of MegaFace [10] for fair comparison, which adopts FaceScrub as the probe set. It contains 1M photos of 690K subjects in the gallery set and 100K images of 530 unique subjects in the probe set.

The IJB-B [27] dataset includes 1,845 identities with 55K video frames and 22K still images, which provides 10K genuine and 8M impostor matches. The IJB-C dataset [28] consists of 3,531 individuals with 118K video frames and 31K still images, including 20K genuine matches and 16M impostor matches.

3) DATA AUGMENTATION

As introduced previously, several generic data augmentation techniques are adopted for face image augmentation, i.e., color jittering, occlusion, blur, horizontally flip, and grey level transformation. These augmentation techniques are selected with a probability of 0.25 successively, so a combination of different techniques may be applied on a single sample. The max number of augmentations applied on a single image is limited to 3 in our experiments to avoid the augmented image drifting far from the original image. MarginMix is applied with a probability of 0.5, and the mixing coefficient λ is drawn from a beta distribution with $\alpha = \beta = 0.2$ in all our experiment. We further use (7) to ensure $\lambda \geq 0.5$. All of these data augmentation methods are only applied in the tail data.

4) IMPLEMENTATION

We adopt ResNet50 and ResNet100 as our backbone network and use the ArcFace [10] as our loss function. We set the angular margin m at 0.5 and the feature scale s at 64 for ArcFace. The Stochastic Gradient Descent optimizer is employed, and the learning rate starts from 0.1 with fixed momentum of 0.9 and weight decay of 5e-4. On ID30K, we divide the learning rate by 10 at 10, 16, 22 epochs and finish the training process at 25 epochs. On Glint360K, the learning rate is divided at 8, 12, 16, 20 epochs and the training process is finished at 22 epochs.

TABLE 2. Face verification accuracy (%) with different training strategy on ID30K.

Strategy	LFW	CFP-FP	AgeDB	CPLFW	CALFW
Baseline	99.55±0.31	97.20±0.90	97.68±0.80	95.73±1.35	91.30±1.39
DA	**99.70±0.27**	97.04±1.14	97.83±0.74	95.72±1.23	91.28±1.27
JAT	99.63±0.27	**97.54±0.83**	97.88±0.72	**95.85±1.15**	91.80±1.37
MarginMix	99.68±0.23	97.34±0.72	**97.98±0.79**	95.62±1.15	**91.92±1.52**

TABLE 3. Performance comparisons of different combinations of training strategies on various benchmarks. 1:1 verification accuracy (%) is reported on the LFW, CFP-FP, AgeDB, CPLFW, CALFW datasets. Identification and verification evaluation on MegaFace dataset. "Id." refers to the rank-1 face identification accuracy (%) with 1M distractors, and "Ver." refers to the face verification (%) TAR@FAR=1e-6.

Model	Verification Accuracy					MegaFace	
	LFW	CFP-FP	AgeDB	CPLFW	CALFW	Id.	Ver.
Baseline	99.55±0.31	97.20±0.90	97.68±0.80	95.73±1.35	91.30±1.39	95.42	96.96
JAT	99.63±0.27	97.54±0.83	97.88±0.72	95.85±1.15	91.80±1.37	96.57	97.54
JAT+DA	99.82±0.26	98.06±0.57	97.92±0.79	95.88±1.12	92.05±1.33	96.85	98.32
JAT+MarginMix	99.73±0.23	98.04±0.71	97.89±0.83	95.88±1.10	92.33±1.25	96.78	98.16
JAT+DA+MarginMix	**99.83±0.19**	**98.11±0.84**	**98.02±0.77**	**95.90±1.12**	**92.53±1.08**	**97.05**	**98.42**

B. ABLATION STUDY ON ID30K

In this section, all the models are trained using ResNet50 network and ArcFace loss function. The face verification accuracy is tested on the five datasets with ten-fold verification sets. Furthermore, face identification and verification evaluations are performed on MegaFace with 1M distractors.

1) EFFECTIVE STRATEGY

We apply three different strategies independently in the training process, i.e., data augmentation (DA), JAT and MarginMix. We firstly train a model on the original long-tailed ID30K dataset as a baseline, which is trained without using any strategy. In DA strategy, we only apply generic data augmentation on the tail data with instance-balanced sampling. The face verification accuracies of these strategies are shown in Table 2.

Compared with the baseline, we can see that: (1) Apply DA on the tail data alone cannot gain effective performance improvement in most of the test sets. This can be attributed to the instance-balanced sampling strategy, in which the tail classes are still inadequately trained even by using data augmentation. (2) JAT can achieve higher accuracy on all the five benchmarks than the baseline, which proves that JAT can learn more discriminative feature from both the long-tailed data and the tail data together. (3) Except for CPLFW, MarginMix performs better than the baseline, and it can obtain comparable improvement with JAT. (4) The accuracy improvement is not so significant by applying single strategy alone, e.g., JAT only brings 0.1-0.5% accuracy increase on these test sets.

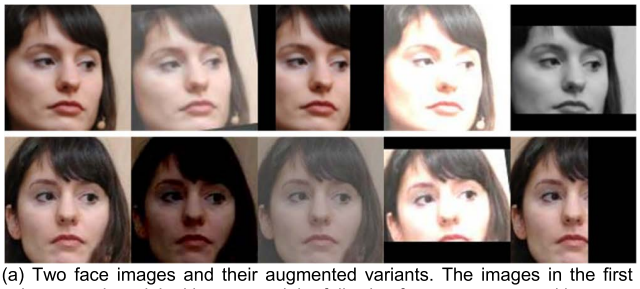
2) EFFECTIVE STRATEGY COMBINATION

Based on the JAT framework, we further combine the other two strategies to search the best solution for long-tailed face recognition. In this section, three different combinations, i.e., JAT+DA, JAT+ MarginMix and JAT+DA+ MarginMix, are applied in the training of face models. The performance comparisons on various benchmarks are shown in Table 3.

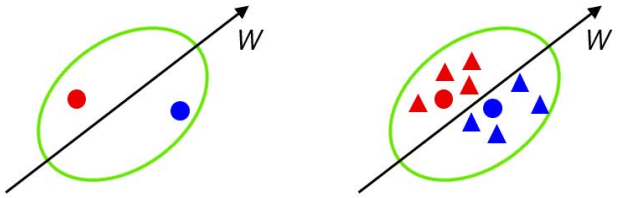
From the results, we can see that all the three combinations outperform the baseline and the model using single JAT strategy. The combination of JAT+DA+ MarginMix (JDM) achieves the highest accuracy on all the benchmarks, which demonstrates its effectiveness in dealing with the long-tailed problem in face recognition. More importantly, the accuracy improvement in JDM is not trivial anymore. For instance, the verification accuracy increases by about 1% on CFP-FP and CALFW, and the identification and verification accuracy on MegaFace increases by about 1.5%. When JAT, DA and MarginMix is incrementally applied, the accuracies are steadily improved on all the test sets, i.e., JAT+DA is better than JAT and JAT+DA+ MarginMix is even better, which demonstrates that there is negligible conflict between the three strategies.

Another informative observation is that data augmentation does steadily improve model performance when combined with the JAT framework, and this observation holds for both generic DA and MarginMix. This is because the tail data is particularly enhanced in both sample number and intra-class variance in the framework of JAT, which helps to exploit more discriminative information.

We randomly select a tail identity in ID30K for illustration, which only contains two face images. We show these two images with their augmented variants in both image space and feature space in Fig. 2. In Fig. 2(a), we can see that data augmentation can effectively enlarge the intra-class variance for the tail class. We visualize the image features by projecting them onto 2D space using t-SNE [29] in Figure 2(b). The features of the original images in the left are extracted from the baseline model, and the features in the



(a) Two face images and their augmented variants. The images in the first column are the original images and the following four are augmented images.



(b) Face images in the feature space. The circles represent the original images, and the triangles represent the augmented images (red for images in the first row, and blue for images in the second row). *W* denotes the prototype which represents the center of the class.

FIGURE 2. Example of images and their augmented variants from a randomly selected tail identity. The original and augmented face images in image space (a) and feature space (b).

TABLE 4. Face identification and verification evaluation on MegaFace. "Id." refers to the rank-1 face identification accuracy with 1M distractors, and "Ver." refers to the face verification TAR@FAR=1e-6.

Model	MegaFace	
	Id.	Ver.
Glint360K (CosFace, r=1.0) [2]	99.13	98.98
Glint360K (CosFace, r=0.1) [2]	98.94	99.10
Glint360K (ArcFace)	99.15	98.86
Glint360K (JDM)	**99.19**	**99.12**

right are extracted from our JAT+DA+ MarginMix model. The distance between the red circle and blue circle becomessmaller in the right of Fig 2(b), therefore, our model can learn more compact feature representations for the tail class. This observation is consistent with the conclusions in [6], [7], and [8] that enhancing the intra-class compactness of tail classes can lean more discriminative feature and improve the generalization ability of the model.

C. EXPERIMENTS ON GLINT360K

To further validate the effectiveness of the combination of strategies in JDM on larger scale training dataset, we conduct experiments on Glint360K using ResNet100 as backbone architecture. The models are tested on three large-scale test datasets, i.e., MegaFace [26], IJB-B [27] and IJB-C [28]. Firstly, we train a face model on Glint360K dataset as a baseline, and further compare our method with the state-of-the-art methods in [2].

1) RESULTS ON MegaFace

As shown in Table 4, our method achieves the best performance under both verification and identification protocols, achieving the accuracy of 99.19% and 99.10% respectively. The result demonstrates that the combination of JDM can effectively address the long-tailed problem in face recognition on highly imbalanced and large-scale training dataset.

TABLE 5. Face verification TAR (@FAR=1e-5 and 1e-4) on the IJB-B and IJB-C benchmarks.

Model	IJBB		IJBC	
	1e-5	1e-4	1e-5	1e-4
Glint360K (CosFace, r=1.0) [2]	-	96.10	-	97.30
Glint360K (CosFace, r=0.1) [2]	-	96.10	-	97.20
Glint360K (ArcFace)	92.28	95.88	95.23	97.20
Glint360K (JDM)	**93.33**	**96.15**	**95.81**	**97.45**

2) RESULTS ON IJB

Following the testing protocol in [10], we adopt the feature norm and the face detection score to reweight the face within each template. We show the TAR@FAR=1e-5 and 1e-4 of different methods in Table 5. We can find that our method outperforms the other methods on both IJB-B and IJB-C benchmarks. Compared with the baseline trained with Arcface, the combination of JDM gains accuracy improvement of 1.05%, 0.27% on IJB-B and 0.58%, 0.25% on IJB-C at TAR@FAR=1e-5, 1e-4 respectively.

V. CONCLUSION

In this paper, we address the long-tailed face recognition problem from the perspective of maximally exploiting the tail data in long-tailed dataset. We propose a JAT framework to learn more discriminative feature by alternating training on the long-tailed data and the tail data with different sampling strategies. We further propose MarginMix to deal with the nonlinearity of margin-based softmax loss in mixup training, which is further combined with other DA techniques to generate face images with more variations for the tail data. Furthermore, we obtain the best combination of strategies, i.e., JAT+DA+ MarginMix, for long-tailed face recognition, which can maximally exploit the discriminative information in the tail data while retaining the universal discrimination learned from the long-tailed dataset. Extensive experiments demonstrate that our proposed methods and combination of strategies can learn more discriminative deep feature on long-tailed face datasets.

REFERENCES

[1] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, and J. Zhou, "WebFace260M: A benchmark unveiling the power of million-scale deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashvile, TN, USA, Jun. 2021, pp. 10492–10502.

[2] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, and Y. Fu, "Partial FC: Training 10 million identities on a single machine," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 1445–1449.

[3] J. Cao, Y. Li, and Z. Zhang, "Celeb-500K: A large training dataset for face recognition," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 2406–2410.
[4] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 7044–7053.
[5] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy Oct. 2017, pp. 5409–5418.
[6] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5704–5713.
[7] H. Liu, X. Zhu, Z. Lei, and S. Z. Li, "AdaptiveFace: Adaptive margin and sampling for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 11947–11956.
[8] B. Liu, W. Deng, Y. Zhong, M. Wang, J. Hu, X. Tao, and Y. Huang, "Fair Loss: Margin-aware reinforcement learning for deep face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 10052–10061.
[9] Y. Zhong, W. Deng, M. Wang, J. Hu, J. Peng, X. Tao, and Y. Huang, "Unequal-training for deep face recognition with long-tailed noisy data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7812–7821.
[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4690–4699.
[11] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashvile, TN, USA, Jun. 2021, pp. 14225–14234.
[12] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2537–2546.
[13] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng, "Meta-Weight-Net: Learning an explicit mapping for sample weighting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1919–1930.
[14] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Sep. 2022, pp. 1–16.
[15] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9719–9728.
[16] Y. Yang and Z. Xu, "Rethinking the value of labels for improving class-imbalanced learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19290–19301.
[17] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9268–9277.
[18] iNatrualist. *The iNatrualist Competition Dataset*. Accessed: Apr. 10, 2018. [Online]. Available: https://github.com/visipedia/inat_comp/tree/master/2018
[19] Y. Zhang, X. S. Wei, B. Zhou, and J. Wu, "Bag of tricks for long-tailed visual recognition with deep convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3447–3455.
[20] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Feb. 2018, pp. 1–10.
[21] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, Oct. 2008, pp. 1–10.
[22] T. Zheng, W. Deng, and J. Hu, "Cross-Age LFW: A database for studying Cross-age face recognition in unconstrained environments," 2017, *arXiv:1708.08197*.
[23] T. Zheng and W. Deng, "Cross-Pose LFW: A database for studying cross-pose face recognition in unconstrained environments," Beijing Univ. Posts Telecommun., Beijing, China, Tech. Rep. 5:7, 2018.
[24] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 51–59.
[25] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Placid, NY, USA, Mar. 2016, pp. 1–9.
[26] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 Million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4873–4882.
[27] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother, "IARPA Janus benchmark-B face dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 90–98.
[28] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus benchmark-C: Face dataset and protocol," in *Proc. Int. Conf. Biometrics (ICB)*, Gold Coast, QLD, Australia, Feb. 2018, pp. 158–165.
[29] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

SONG GUO received the B.S. degree in biomedical engineering and the Ph.D. degree in signal and information processing from Beijing Jiaotong University, in 2007 and 2018, respectively.

Since then, he has been working at Fujitsu Research and Development Center Company Ltd., Beijing, China. His research interests include pattern recognition, image processing, and machine learning.

RUJIE LIU received the B.S., M.S., and Ph.D. degrees in electronic engineering from Beijing Jiaotong University, in 1995, 1998, and 2001, respectively.

Since then, he has been working as a Researcher at Fujitsu Research and Development Center Company Ltd., Beijing, China. He has published more than 40 papers and tens of inventions. His research interests include the areas of AI, pattern recognition, and image processing.

MENGJIAO WANG received the Ph.D. degree in biomedical engineering from Tsinghua University, Beijing, China, in 2014.

Since 2016, he has been a Researcher at Fujitsu Research and Development Center Company Ltd., Beijing. His research interest includes face detection and recognition.

MENG ZHANG received the M.Sc. degree from the Beijing University of Technology, China, in 2016. He is currently pursuing the Ph.D. degree in intelligent systems with the Graduate School of Informatics, Nagoya University, Japan. Since then, he has been working as a Researcher at Fujitsu Research and Development Center Company Ltd., Beijing, China. His research interests include image processing, pattern recognition, face recognition, and deep learning.

SHIJIE NIE received the Ph.D. degree from the Graduate University for Advanced Studies (SOKENDAI), Japan. He is currently a Computer Vision Researcher at Fujitsu Research and Development Center Company Ltd., China. His current research interests include biometric and deep learning.

SEPTIANA LINA received the Bachelor of Engineering (B.Eng.) degree in electronic engineering from Satya Wacana Christian University, Indonesia, in 2007, the Master of Science (M.Sc.) degree in electrical engineering and computer science from Chung Yuan Christian University, Taiwan, in 2013, and the Doctor of Engineering (D.Eng.) degree in information and communication engineering from the Tokyo Institute of Technology, Japan, in 2020.

From 2007 to 2020, she was worked as a Researcher in the university and an information-communication industry with a research interest specifically in image analysis, computer vision, and artificial intelligence. Since 2020, she has been a Researcher at Fujitsu Laboratories Ltd. Her current research interests include biometric field and its related technology.

NARISHIGE ABE received the B.S. degree in engineering from Osaka City University, in 2005, and the M.S. degree in information science from Osaka University, in 2007. Since 2007, he has been working at Fujitsu Laboratories Ltd. He was also a Visiting Scholar at Stanford University (2013–2014). He received the OHM Technology Award, in 2017. His research interests include image processing, machine learning, and biometric authentication algorithm.