A Progressive Learning Framework for Unconstrained Face Recognition

Zhenhua Chai, Shengxi Li, Huanhuan Meng, Shenqi Lai, Xiaoming Wei, Jianwei Zhang
Vison and Image Center (VIC) of Meituan
{chaizhenhua, lishengxi, menghuanhuan02, laishenqi, weixiaominng, zhangjianwei09}@meituan.com

Abstract

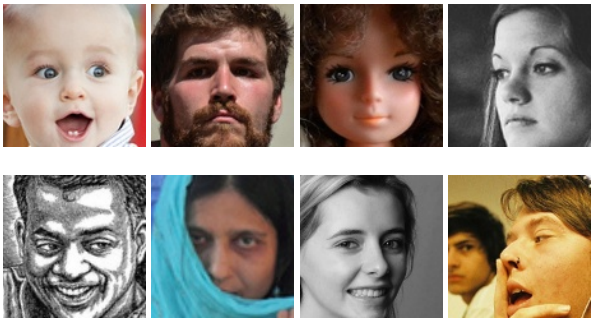
The carefully designed backbone network, the increase of training data and the improved training skills have boosted the performance of modern face recognition systems. However, in some deployment cases which aim at model compactness and energy efficiency, some of the existing systems may fail due to the high complexity. Lightweight Face Recognition Challenge is proposed in order to make some progress in this direction and establishes a new comprehensive benchmark. In this challenge, we have designed a light weight backbone architecture and all the parameters are trained in a progressive way. Finally we achieve the 5th in track 1 and the 4th in track 3.

1. Introduction

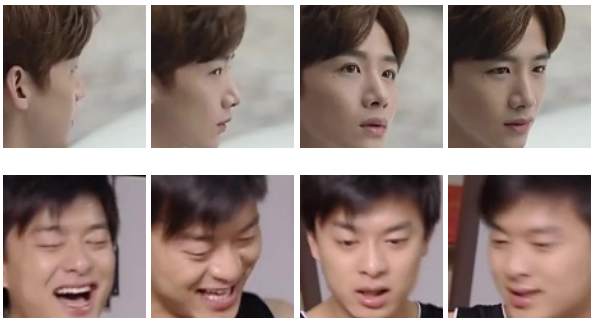
Face recognition is one of the most popular research topics in the field of computer vision, which has been studied by both academy and industry for several decades. Recently, benefited from the development of convolutional neural networks, great progress has been achieved for face recognition even in some unconstrained environments [16, 22, 10].

However, most of these benchmarks focus on the improvement of the accuracy while the model size and the runtime efficiency are neglected. Lightweight Face Recognition Challenge (LFRC)[5] is one of the first proposed to measure the performance in terms of both accuracy and the model complexity. In this way, although using deeper neural network with hundreds of layers and millions of parameters could achieve high accuracy, the computational cost will sometimes be beyond the requirement which will limit its use in some mobile or embedded applications. Some random selected sample images from this challenge are shown in Fig.1, and we can find that the applications of unconstrained face recognition under limited computational resources is still a challenge problem.

In the literature, as far as we can see there are mainly three directions to deal with lightweight face recognition problems: 1) the use of lightweight network structure; 2) the carefully designed loss function which will usually aim to reduce the intra class differences and at the same time increase the inter class differences; and 3) some fancy training skills in order to avoid the overfitting and enhance the model generalizability.



(a) Trillion-pairs



(b) iQIYI-VID

Figure 1: Sample images from testing datasets. The track 1 consists of still images with large pose, exaggerated expression, non-uniform lighting, occlusion and sketch. The track 3 contains a series of face images from consecutive video frames, on which the main challenges are large pose, motion blur, lighting and occlusions.

In this paper, we designed a lightweight network architecture based on channel pruning and network deepening, which achieved high performance on face recognition task. To enhance the model generalizability, we also pro-

posed an effective training process by warpping different loss functions in a progressive way. Extensive experiments on image-based test and video-based test demonstrate the effectiveness of our method.

The rest part of this paper will be organized as follows: in Section 2, we will briefly introduce the related works from the three directions mentioned above; in Section 3 we will give the details of our solution; in Section 4 the ablation study will be conducted and experimental results for each part of our solution will be compared; and finally in Section 5 the conclusion and our future plan will be made.

2. Related Work

2.1. Network Architecture

In this light weight face recognition challenge (LFRC), there is a strict limitation on both the model size (e.g. less than 20M) and the model complexity (e.g. less than 1G FLOPs). So the participants will be inclined to achieve a trade-off between accuracy and efficiency during model design, and actually this topic has been actively studied in the field of image classification [3]. For example, SqueezeNet [17] is one of the early works in this direction, which has proposed 1x1 convolution with squeeze and expand modules and the number of parameters will be reduced heavily. SqueezeNext [9] is its following work, which will shift the focus on reducing the number of operations (MAdds). In this way, not only the number of parameters will be reduced but also the inference latency will be substantially improved. After that MobileNet [14] and MobileNetV2 [24] are proposed to use novel structure for light weight image classification, where the former one uses depth-wise convolution and point-wise convolution to replace the vanilla convolution and the latter one uses linear bottleneck and inverted residual structure. Their performances are comparable with large VGG-16 [27] but with only one thirtieth of parameters and MAdds. There are another two important works named ShuffleNet [37] and ShuffleNetV2 [21] which utilize group convolution and channel shuffle operations to reduce the MAdds. IGCV1 [36], IGCV2 [34] and IGCV3 [28] propose to use interleaved group convolution (IGC) to further reduce the redundancy. Besides the manually designed architecture, there are some pioneer works (e.g. MobileNetV3 [13]) focusing on automatically Neural Architecture Search (NAS), and based on which the performance can still be improved while the model can still keep the relative low complexity.

However, there are mainly two obvious differences between image classification and face recognition. Firstly, there is usually an alignment preprocess step before face feature extraction while the model for image classification is required to be rotation invariants. Secondly, the face images even from different classes will share a similar structure while the inter class difference in image classification will be more obvious. Based on the analysis above, the model capacities for face recognition can be further explored. There are already some works proposed to design a light weight model for face recognition. For example, LightCNN [33] is presented to learn a compact embedding on the large-scale face data, and achieves 99.33% face verification accuracy on LFW with 12.6 million parameters. MobileFaceNet [1] is based on MobileNetV2 with reduced the expansion factor. Global Depth-wise Convolution (GDC) [1] is introduced to replace Global Average Pooling (GAP), and achieves better verification accuracy. MobiFace [8] is also based on MobileNetV2. They adopt fast downsampling strategy to reduce the size of feature map. Besides, ReLU and GAP are replaced by PReLU and Fully Connected Layer for better performance. More recently, a novel structure named MobileNetV3 [13] which introduced Squeeze-and-Excitation [15] (SE) module, Hard Swish activation function and Network Architecture Search technique exhibits even more promising results, which motivates us to apply this new structure in our model design.

2.2. Loss Function

Another important way to improve face recognition performance is the design of suitable loss functions. Some early works [33] treat face recognition as a classification problem, and Softmax loss with the identity labels will be used as a supervised signal. It has to be mentioned that the Softmax loss only considers the inter class differences while the intra class differences are ignored. In order to overcome this shortcoming, researchers will design an extra kind of loss function with Softmax loss to construct a joint supervision. In this way, the intra class difference will be reduced while the inter class difference will be enlarged. There are mainly two streams proposed for this direction. For example, in [32] the Center loss has been proposed which will explicitly impose extra loss term that penalizes the Euclidean distance between samples and their representative centers. In FaceNet [25], triplet loss has been introduced, where anchor based mining is implemented based on millions of images in the training set. Since the batch size is limited due to hardware, the results will reply heavily on the online hard example mining strategy and it becomes a little tricky to be implemented to obtain a good result.

The other main stream is to modify the original Softmax loss to angular space due to the fact that the features learned by Softmax loss have intrinsic angular distribution. SphereFace [18] can be viewed as a milestone work of this kind, which will model the features in angular space with weight normalization operation and introduce an angular margin to the decision boundary [20]. However, the margin used in [18] is multiplicative, which is a little hard to convergence during training process. Later, additive an-

gular margin based methods (e.g. ArcFace [19] CosFace [30] and AM-Softmax [29]) are proposed, which are relatively easier to train and can further boost the performance. Similar ideas are also presented in CosFace [30] and AM-Softmax [29] which enlarge the decision margin in the cosine manifold. More analysis and comparisons on angular margin based methods can be found in [19]. There are some more recent works, which will consider the inter-class separability. For instance, RegularFace [38] explicitly distances identities by penalizing the angle between an identity and its nearest neighbor, which will result in discriminative face representations. UniformFace [7] impose an equidistributed constraint by uniformly spreading the class centers on the manifold, so that the minimum distance between class centers can be maximized through complete exploitation of the feature space.

The losses introduced above have already exhibit good performance and may be complementary to each other. Therefore in our solution instead of designing one more loss function, we will propose a novel progressive learning framework, which will make use of all the introduced losses and train the network in a progressive manner. In this way, we believe that our model can avoid to be overfitting.

3. Proposed Method

3.1. Network Architecture

We use depthwise separable convolution in our basic block. Because compared with standard convolution, it can often save substantial parameters. The linear bottleneck (Fig.2 (a)) and inverted residual structure (Fig.2 (b)) from MobileNetV2 [24] are also designed to save parameters, which is composed of pointwise expansion convolution, depthwise convolution, pointwise linear convolution and residual connection. Pointwise expansion convolution increases input channels for depthwise convolution. Pointwise linear convolution has no activation layer to minimize the loss of the information. Residual connection can ensure stable optimization especially when network is deep. The squeeze and excitation block adaptively recalibrate channel-wise feature by explicitly modelling channel relationships, which can achieve significant performance improvement in modern architectures with slight computation cost. Meanwhile, PReLU [11] is used as activation function and the extra computation cost is almost negligible. So based on the structure mentioned above we design two kinds of basic blocks and details can be found in Fig.2 (c) and (d).

The network depth is one of the key factors to performance on vision task. General speaking, deeper network can capture richer information and get more remarkable performance. However, due to the vanishing gradient problem and gpu memory limit, deeper network is harder to optimize. Sometimes, with the growing of network depth, the performance could even decrease. The network width is also important [35], which has been studied in the literature [14, 24, 37, 21]. Wider network which has more parameters than the thin one is easier to train and usually can achieve better performance. However, too wider but shallow network will perform underachievement because of lacking high level semantic features. Low-level block has large feature map size. Increasing the depth and width will increase computation obviously. High-level block is related to large receptive filed and usually will contribute more on final performance. In consideration of the semantic information of high-level features, we prefer to increase the depth and width of high-level blocks, which have greater impact on results. The detailed structure can be found in Tab.1.

Input	Operator	t	c	n	s
112*112*3	Conv	-	24	1	2
56*56*24	Block	56	24	2	1
56*56*24	Block	116	56	1	2
28*28*56	Block	116	56	6	1
28*28*56	Block	248	124	1	2
14*14*124	Block	248	124	15	1
14*14*124	Block	512	256	1	2
7*7*256	Block	256	256	24	1
7*7*256	Conv	-	512	1	1
7*7*256	GDC	-	282	1	1

Table 1: Each line describes a sequence of 1 or more identical layers, repeated n times. All layers in the same sequence have the same number c of output channels. The first layer of each sequence has a stride s and all others use stride 1. All spatial convolutions use 3 3 kernels. The expansion channels t is the input channels of depthwise convolution.

3.2. Loss Function

Loss function also plays an important role on recognition performance. In our solution instead of designing a new loss, we propose to train the network in a progressive way.

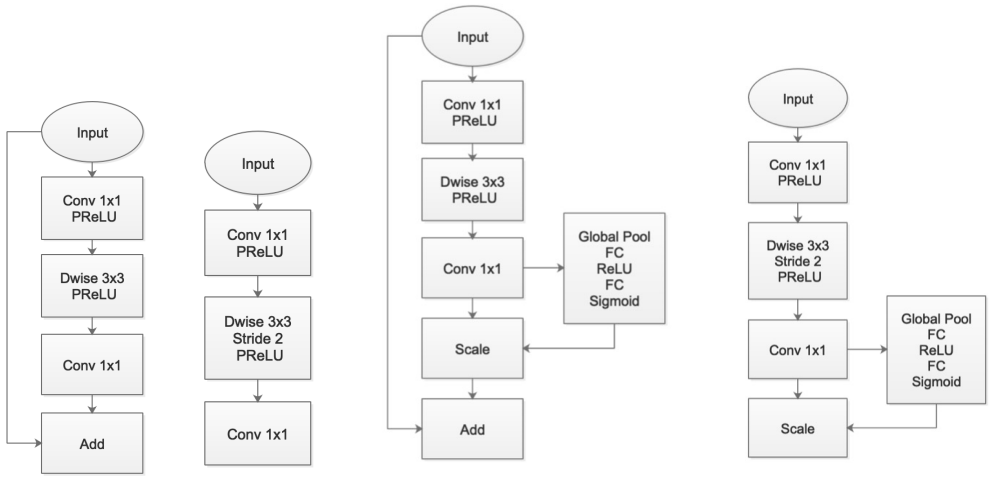
Firstly, the well designed backbone network introduced in previous section is trained with ArcFace loss [19] which can be formulated as follows:

$$L_1 = \frac{1}{N}\sum_i -\log\frac{e^{s\cdot\cos(\theta_{y_i}+m)}}{e^{s\cdot\cos(\theta_{y_i}+m)}+\sum_{j=1,j\neq y_i}^{n} e^{s\cdot\cos\theta_j}} \tag{1}$$

subject to:

$$W_j = \frac{W_j}{||W_j||}, x_i = \frac{x_i}{||x_i||}, \cos\theta_j = W_j^T x_i \tag{2}$$

ArcFace loss belongs the margin based loss and the trained model usually has good generalizability, which will



(a) (b) (c) (d)

Figure 2: (a) linear inverted residual bottleneck in MobileFaceNet; (b) linear inverted residual bottleneck when stride = 2 in MobileFaceNet; (c) linear bottleneck and inverted residual structure with SE module; (d) linear bottleneck and inverted residual structure with SE module when stride = 2.

be used as our base backbone. After convergence, we will add one of the following loss functions [32, 30, 38, 19, 29, 31, 18] at each time, and the process will be repeated till all the loss functions are added.

$$L_2 = \frac{1}{N}\sum_i -\log\frac{Z_1}{Z_1 + \sum_{j=1,j\neq y_i}^{n} Z_2} \tag{3}$$

subject to:

$$Z_1 = e^{s\cdot\cos(m_1\theta_{y_i}+m_2)-m_3}, Z_2 = e^{s\cdot\cos\theta_j} \tag{4}$$

$$L_3 = \frac{1}{N}\sum_i -\log\frac{Z_1}{Z_1 + \sum_{j=1,j\neq y_i}^{n} Z_3} \tag{5}$$

subject to:

$$Z_3 = e^{s\cdot(t-1)(\cos\theta_j+1)I_j} e^{s\cdot\cos\theta_j},$$
$$I_j = \begin{cases} 0, & \cos(m_1\theta_{y_i}+m_2) - m_3 - \cos\theta_j \geq 0 \\ 1, & \cos(m_1\theta_{y_i}+m_2) - m_3 - \cos\theta_j < 0 \end{cases} \tag{6}$$

$$L_{Intra} = \frac{1}{N}\sum_i \cos^{-1}\frac{W_{y_i}^T x_i}{||W_{y_i}^T||\cdot||x_{i||}} \tag{7}$$

$$L_{Inter} = \frac{1}{C}\sum_i \max_{j\neq i}\frac{W_i^T W_j}{||W_i^T||\cdot||W_{j||}} \tag{8}$$

$$L_{Ce} = \frac{1-\beta}{1-\beta^{n_y}}\log\frac{e^{z_y}}{\sum_{j=1}^{C} e^{z_j}} \tag{9}$$

Generally speaking, different losses may exhibit differently and could be complementary to each other. Our wrapped loss will usually perform better than using only one of them. Besides, the progressive learning makes the training easier to converge than training multitask loss directly.

4. Experiments

In order to show the effectiveness of the proposed framework, extensive experiments are conducted in different settings. Firstly, we design our light weight backbone network and train its parameters in manual search way with ArcFace[4] loss. After the training convergence, we use current best model as the warm up version and continue the learning process in a progressive manner by wrapping more loss functions. The process will be repeated till all the losses are involved. Finally, we test our models on Track 1 Trillion-pairs dataset and Track 3 iQIYI-VID dataset, and the final performances are evaluated and reported by the organizers.

4.1. Dataset

The officially provided training set is a cleaned version from the large scale MS1M [10] which contains 5.1M images of 93K identities. All the images are preprocessed according to RetinaFace[6], and only the cropped version sized of 112×112 is provided to public. Three popular datasets such as Labelled Faces in the Wild (LFW) [16], Celebrities in Frontal Profile (CFP) [26] and Age Database (AgeDB) [23] can be used as the validation set.

Table 2: The process of adjusting the network structure and the results on validation datasets for each step. C means the output channels, Num means the repeat times of SEResidual block. The right 3 columns are the accuracy on LFW, CFP-FP and AgeDB-30.

Step	layer name										Dim	FLOPs	Size	LFW	CFP-FP	AgeDB-30
	conv1		conv2		conv3		conv4		conv5							
	C	N	C	N	C	N	C	N	C	N						
0	64	1	64	2	128	8	256	16	512	4	256	933.3M	18M	0.9965	0.9797	0.9778
1	30	1	62	2	128	8	256	18	512	16	256	1G	18M	0.9977	0.9847	0.9798
2	28	1	60	2	128	8	256	16	512	19	256	1G	18M	0.9977	0.9854	0.9812
3	26	1	58	2	128	6	256	16	512	21	256	1G	19M	**0.9982**	0.9845	0.9802
4	24	1	56	2	124	6	256	16	512	23	256	998.4M	20M	0.9978	**0.9860**	0.9805
5	24	1	56	2	124	6	256	15	512	24	282	976.6M	20M	0.9977	0.9850	**0.9815**

LFW dataset collected from internet contains 13,233 images from 5749 identities, and a total of 6,000 image pairs are used to measure the performance in term of verification accuracy. The web-collected face images have large variations in pose, expression and illuminations. CFP dataset consists of 500 subjects, each with 10 frontal and 4 profile images. We take the most challenging subset CFP-FP to report the performance following [19]. AgeDB dataset is an in-the-wild dataset with large variations in pose, expression, illuminations and age. AgeDB contains 12,240 images of 440 distinct subjects. We use the most challenge subset from the four groups, AgeDB-30, to report the performance as well. We can also get a cropped version of the validation set from the organizers. For testing, two typical large-scale datasets are used, Trillion-pairs dataset for image test and iQIYI-VID for video test. Images in Trillion-pairs dataset come from ELFW and DELFW. ELFW contains 274K images from 5.7K identities. DELFW is the distractors for ELFW and contains 1.58M face images from Flickr. iQIYI-VID includes 200K videos of 10K identities, with each video extracted to frames at 8FPS. Besides, modification (e.g. re-alignment or resize) and data argumentation except Horizontal flipping on testing images are prohibited. This will force all the participants to pay more attentions on the network design and keep result comparison fair in this challenge. Finally we report our performance on Track 1 Trillion-pairs dataset and Track 3 iQIYI-VID dataset.

4.2. Training Details and Experimental Results

In the first stage, we adjust the network architecture within the limitations of computational complexity, model size and feature dimension, which are 1G FLOPs, 20M and 512 dim respectively.

We take MobileFaceNet as baseline and trace the accuracy on validation datasets to guide the adjustment of the network structure. Specially, inspired by MobileNetV3 we add SE-Block to the two basic blocks of MobileFaceNet for applying attention mechanism to feature maps. We name the new basic blocks SEResidual and SEDResidual. The computational complexity of the baseline is close to 1G FLOPs and the model size is close to 18M. To enhance the discrimination of our model, we mainly focus on expending the depth and width of higher blocks. In our solution, we increase the number of blocks in $Conv5\_x$ and the channels of feature map in deep layers. In order to avoid increasing the total computational complexity and model size, we decrease the corresponding item in shallow layers. We also take feature dimension into consideration and fix it to a suitable value through experiments. The process of adjusting the network structure is shown on Tab.2. Step 0 means the baseline, Step 1 to Step 5 adjust the depth and width progressively. What's more, step 5 also adjusts the feature dimension based on the previous work.

In the second stage, we keep the network architecture fixed and aim at exploring a better form of loss function. We use current best model as the warm up version and continue the learning process in a progressive manner by wrapping more loss functions. The details can be found in Tab.3. As we expected, model training in a loss wrapping way will further boost the performance and the training is much easier to convergence in comparison with directly multitask learning.

Table 3: Results on Trillion-pairs for loss function wrapping.

Methods	LFW	CFP-FP	AgeDG-30	ICCV19-challenge
ArcFace	99.767	98.500	98.150	86.666
Combined	98.800	98.371	98.017	86.877
Combined +svgs	99.800	98.371	98.017	87.181
Combined +svgs +Intra +Inter	99.800	98.343	98.183	87.195

In the final stage, we target on exploring better ways for data processing to boost the performance on large-scale image test benchmark and large-scale video test benchmark. After carefully analyzing the characteristics of training and

Table 4: Results on Trillion-pairs dataset using proposed progressive learning strategy.

Methods	LFW	CFP-FP	AgeDG-30	ICCV19-challenge
Combined +svs +Intra +Inter	99.800	98.343	98.183	87.195
Combined +Intra +Inter+CB	99.833	98.443	98.250	87.141
Combined +Intra +Inter+CB +batch id=4[12]	99.817	98.243	98.183	87.214
Combined +Intra +Inter+CB +batch id=4 + semi-hard samples mining	99.817	98.443	98.167	87.432

testing datasets, we designed a search strategy which includes cutting long tail identities (Fig.5), PK mining[12] and semi-hard samples mining. We conduct a series of experiments to examine each effect and match different data processing method for different benchmark.

According to our analysis, there are three kinds of noise in iQIYI-VID test dataset. We sample some of the videos shown as Fig.3. Obviously, the noise frames will affect the video feature and make it less representative on condition that we map the frame features to video feature by simply averaging. Specially, we use unsupervised clustering based on Union-Find for each video to remove the noise frames. We choose the relatively large cluster which contains the most elements as the cleaned frames of the video. In this way, a large number of outliers aforementioned can be filtered. Fig.4 shows the result of Union-Find cluster for video id 0144686. Extensive experiments have been conducted to determine the threshold of Union-Find cluster. The results are shown in Tab.5. Extensive experiments on iQIYI-VID test dataset demonstrate the effectiveness of the method.

Table 5: Results on iQIYI-VID dataset after Union-Find Cluster with different thresholds.

Model	Threshold	Result
baseline	0	0.57201
	0.3	0.57191
	0.4	**0.57442**
	0.5	0.57191
	0.6	0.56473
best model	0.3	0.57952
	0.4	**0.58207**
	0.5	0.57894

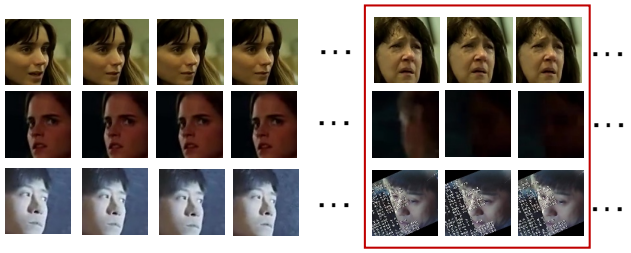


Figure 3: Examples for noise frames in iQIYI-VID test dataset. The first row shows the frames of 0144611. There are two indentities in the same video id. The second row shows the frames of 0144691. Some of the frames in the video id are not face images. The third row shows the frames of 0144610. Some of the frames in the video id suffer from heavy occlusion of faces. Noise frames are marked by red rectangle. Best viewed in color.

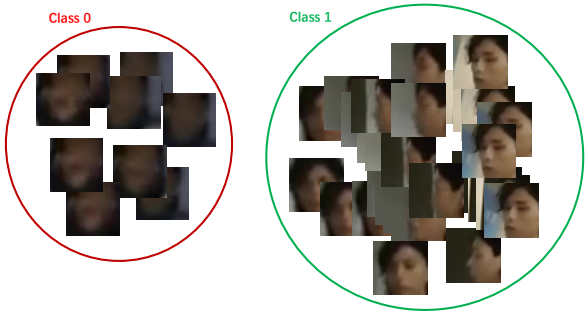


Figure 4: Sample result for Union-Find Cluster [2]. Frames in video id 0144610 are splitter by unsupervised cluster into two classes. noise class with few frames are marked by red circle. We map the frame features to video feature after noise removal. Best viewed in color.

5. Conclusion

In this paper, we have proposed a carefully designed backbone architecture for light weight face recognition. When the base backbone is ready, a novel loss progressive learning framework is used to further finetune the model. After cleaning the outlier of the training set, the generalizability of the model will be further enhanced. Finally, we have achieved the 5th in Track 1 and the 4th in Track 3.

(a) Image quality

(b) Wrong label

(c) Hard example

Figure 5: Sample images in the training set which will lead to bad results. Images in the same row are sampled from the same class. (a)Images with poor quality.(b)Images in the same class but come from different identities. (c)Images in the same class but come from different domain.

References

[1] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition - 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings*, pages 428–438, 2018.

[2] S. Conchon and J. Filliâtre. A persistent union-find data structure. In *Proceedings of the ACM Workshop on ML, 2007, Freiburg, Germany, October 5, 2007*, pages 37–46, 2007.

[3] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009.

[4] J. Deng, J. Guo, N. Xue, and S. 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.

[5] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, S. Shi, and S. Zafeiriou. Lightweight face recognition challenge. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[6] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

[7] Y. Duan, J. Lu, and J. Zhou. Uniformface: Learning deep equidistributed representation for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[8] C. N. Duong, K. G. Quach, N. Le, N. Nguyen, and K. Luu. Mobiface: A lightweight deep learning face recognition on mobile devices. *CoRR*, abs/1811.11080, 2018.

[9] A. Gholami, K. Kwon, B. Wu, Z. Tai, X. Yue, P. H. Jin, S. Zhao, and K. Keutzer. Squeezenext: Hardware-aware neural network design. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1638–1647, 2018.

[10] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, pages 87–102, 2016.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1026–1034, 2015.

[12] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[13] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam. Searching for mobilenetv3. *CoRR*, abs/1905.02244, 2019.

[14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[15] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7132–7141, 2018.

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[17] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360, 2016.

[18] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6738–6746.

[19] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6738–6746, 2017.

[20] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks. In *Proceedings of the*

33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, pages 507–516, 2016.

[21] N. Ma, X. Zhang, H. Zheng, and J. Sun. Shufflenet V2: practical guidelines for efficient CNN architecture design. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, pages 122–138, 2018.

[22] D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz. Megaface: A million faces for recognition at scale. *CoRR*, abs/1505.02108, 2015.

[23] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1997–2005, 2017.

[24] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4510–4520, 2018.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823, 2015.

[26] S. Sengupta, J. Chen, C. D. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, pages 1–9, 2016.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[28] K. Sun, M. Li, D. Liu, and J. Wang. IGCV3: interleaved low-rank group convolutions for efficient deep neural networks. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 101, 2018.

[29] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Process. Lett.*, 25(7):926–930, 2018.

[30] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5265–5274, 2018.

[31] X. Wang, S. Wang, S. Zhang, T. Fu, H. Shi, and T. Mei. Support vector guided softmax loss for face recognition. *CoRR*, abs/1812.11317, 2018.

[32] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, pages 499–515, 2016.

[33] X. Wu, R. He, Z. Sun, and T. Tan. A light CNN for deep face representation with noisy labels. *IEEE Trans. Information Forensics and Security*, 13(11):2884–2896, 2018.

[34] G. Xie, J. Wang, T. Zhang, J. Lai, R. Hong, and G. Qi. Interleaved structured sparse convolutional neural networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8847–8856, 2018.

[35] S. Zagoruyko and N. Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.

[36] T. Zhang, G. Qi, B. Xiao, and J. Wang. Interleaved group convolutions. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4383–4392, 2017.

[37] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6848–6856, 2018.

[38] K. Zhao, J. Xu, and M.-M. Cheng. Regularface: Deep face recognition via exclusive regularization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.