VGGFace2: A dataset for recognising faces across pose and age

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi and Andrew Zisserman
Visual Geometry Group, Department of Engineering Science, University of Oxford
{qiong,lishen,weidi,omkar,az}@robots.ox.ac.uk

***Abstract*—In this paper, we introduce a new large-scale face dataset named VGGFace2. The dataset contains 3.31 million images of 9131 subjects, with an average of 362.6 images for each subject. Images are downloaded from Google Image Search and have large variations in pose, age, illumination, ethnicity and profession (e.g. actors, athletes, politicians).**

The dataset was collected with three goals in mind: (i) to have both a large number of identities and also a large number of images for each identity; (ii) to cover a large range of pose, age and ethnicity; and (iii) to minimise the label noise. We describe how the dataset was collected, in particular the automated and manual filtering stages to ensure a high accuracy for the images of each identity.

To assess face recognition performance using the new dataset, we train ResNet-50 (with and without Squeeze-and-Excitation blocks) Convolutional Neural Networks on VGGFace2, on MS-Celeb-1M, and on their union, and show that training on VGGFace2 leads to improved recognition performance over pose and age. Finally, using the models trained on these datasets, we demonstrate state-of-the-art performance on the IJB-A and IJB-B face recognition benchmarks, exceeding the previous state-of-the-art by a large margin. The dataset and models are publicly available[1].

***Keywords*-face dataset; face recognition; convolutional neural networks**

I. Introduction

Concurrent with the rapid development of deep Convolutional Neural Networks (CNNs), there has been much recent effort in collecting large scale datasets to feed these data-hungry models. In general, recent datasets (see Table I) have explored the importance of intra- and inter-class variation. The former focuses on depth (many images of one subject) and the latter on breadth (many subjects with limited images per subject). However, none of these datasets was specifically designed to explore pose and age variation. We address that here by designing a dataset generation pipeline to explicitly collect images with a wide range of pose, age, illumination and ethnicity variations of human faces.

We make the following *four* contributions: first, we have collected a new large scale dataset, VGGFace2, for public release. It includes over nine thousand identities with between 80 and 800 images for each identity, and more than 3M images in total; second, a dataset generation pipeline is proposed that encourages pose and age diversity for each subject, and also involves multiple stages of automatic and manual filtering in order to minimise label noise; third, we provide template annotation for the test set to explicitly explore pose and age recognition performance; and, finally, we show that training deep CNNs on the new dataset substantially exceeds the state-of-the-art performance on the IJB benchmark datasets [13], [22]. In particular, we experiment with the recent Squeeze and Excitation network [9], and also investigate the benefits of first pre-training on a dataset with breadth (MS-Celeb-1M [7]) and then fine tuning on VGGFace2.

[1] http://www.robots.ox.ac.uk/~vgg/data/vgg_face2/

II. Dataset Review

In this section we briefly review the principal "in the wild" datasets that have appeared recently, inspired by the original Labelled Faces in the Wild (LFW) dataset [10] of 2007. This dataset had $5,749$ identities with $13,000$ images.

The CelebFaces+ dataset [20] was released in 2014, with $202,599$ images of $10,177$ celebrities. The CASIA-WebFace dataset [25] released the same year has $494,414$ images of $10,575$ people. The VGGFace dataset [16] released in 2015 has 2.6 million images covering $2,622$ people, making it amongst the largest publicly available datasets. The curated version, where label noise is removed by human annotators, has $800,000$ images with approximately 305 images per identity. Both the CASIA-WebFace and VGGFace datasets were released for training purposes only.

MegaFace dataset [12] was released in 2016 to evaluate face recognition methods with up to a million distractors in the gallery image set. It contains 4.7 million images of $672,057$ identities as the training set. However, an average of only 7 images per identity makes it restricted in its per identity face variation. In order to study the effect of pose and age variations in recognising faces, the MegaFace challenge [12] uses the subsets of FaceScrub [14] containing $4,000$ images from 80 identities and FG-NET [15] containing 975 images from 82 identities for evaluation.

Microsoft released the large Ms-Celeb-1M dataset [7] in 2016 with 10 million images from 100k celebrities for training and testing. This is a very useful dataset, and we employ it for pre-training in this paper. However, it has two limitations: (i) while it has the largest number of training images, the intra-identity variation is somewhat restricted because of an average of 81 images per person; (ii) images in the training set were directly retrieved from a search engine without manual filtering, and consequently there is

978-1-5386-2335-0/18/$31.00

Datasets	# of subjects	# of images	# of images per subject	manual identity labelling	pose	age	year
LFW [10]	5,749	13,233	1/2.3/530	-	-	-	2007
YTF [23]	1,595	3,425 videos	-	-	-	-	2011
CelebFaces+ [20]	10,177	202,599	19.9	-	-	-	2014
CASIA-WebFace [25]	10,575	494,414	2/46.8/804	-	-	-	2014
IJB-A [13]	500	5,712 images, 2,085 videos	11.4	-	-	-	2015
IJB-B [22]	1,845	11,754 images, 7,011 videos	36.2	-	-	-	2017
VGGFace [16]	2,622	2.6 M	1,000/1,000/1,000	-	-	Yes	2015
MegaFace [12]	690,572	4.7 M	3/7/2469	-	-	-	2016
MS-Celeb-1M [7]	100,000	10 M	100	-	-	-	2016
UMDFaces [5]	8,501	367,920	43.3	Yes	Yes	Yes	2016
UMDFaces-Videos [4]	3,107	22,075 videos	-	-	-	-	2017
VGGFace2 (**this paper**)	9,131	3.31 M	80/362.6/843	Yes	Yes	Yes	2018

Table I: Statistics for recent public face datasets. The three entries in the 'per subject' column are the minimum/average/maximum per subject.

label noise. The IARPA Janus Benchmark-A (IJB-A) [13] and Benchmark-B (IJB-B) [22] datasets were released as evaluation benchmarks (only test) for face detection, recognition and clustering in images and videos.

Unlike the above datasets which are geared towards image-based face recognition, the Youtube Face (YTF) [23] and UMDFaces-Videos [4] datasets aim to recognise faces in unconstrained videos. YTF contains 1,595 identities and 3,425 videos, whilst UMDFaces-Videos is larger with 3,107 identities and 22,075 videos (the identities are a subset of those in UMDFaces [5]).

Apart form these public datasets, Facebook and Google have large in-house datasets. For instance, Facebook [21] trained a face identification model using 500 million images of over 10 million subjects. The face recognition model by Google [18] was trained using 200 million images of 8 million identities.

III. An Overview of The VGGFace2

A. Dataset Statistics

The VGGFace2 dataset contains 3.31 million images from 9131 celebrities spanning a wide range of ethnicities, e.g. it includes more Chinese and Indian faces than VGGFace (though, the ethnic balance is still limited by the distribution of celebrities and public figures), and professions (e.g. politicians and athletes). The Images were downloaded from Google Image Search and show large variations in pose, age, lighting and background. The dataset is approximately gender-balanced, with 59.3% males, varying between 80 and 843 images for each identity, with 362.6 images on average. It includes human verified bounding boxes around faces, and five fiducial keypoints predicted by the model of [26]. In addition, pose (yaw, pitch and roll) and apparent age information are estimated by our pre-trained pose and age classifiers.

The dataset is divided into two splits: one for training having 8631 classes, and one for evaluation (test) with 500 classes.

B. Pose and Age Annotations

The VGGFace2 provides annotation to enable evaluation on two scenarios: face matching across different poses, and face matching across different ages.

Pose templates. A template here consists of five faces from the same subject with a consistent pose. This pose can be frontal, three-quarter or profile view. For a subset of 300 subjects of the evaluation set, two templates (5 images per template) are provided for each pose view. Consequently there are 1.8K templates with 9K images in total. An examples is shown in Figure 1 (left).

Age templates. A template here consists of five faces from the same subject with either an apparent age below 34 (deemed young), or 34 or above (deemed mature). These are provided for a subset of 100 subjects from the evaluation set with two templates for each age period, therefore, there are 400 templates with a total of 2K images. Examples are show in Figure 1 (right).

IV. Dataset Collection

In this section, we describe the dataset collection process, including: how a list of candidate identities was obtained; how candidate images were collected; and, how the dataset was cleaned up both automatically and manually. The process is summarised in Table II.

A. Stage 1: Obtaining and selecting a name list

We use a similar strategy to that proposed by [16]. The first stage is to find as many subjects as possible that have a sufficiently distinct set of images available, for example, celebrities and public figures (e.g. actors, politicians and athletes). An initial list of 500k public figures is obtained from the Freebase knowledge graph [2].

An annotator team is then used to remove identities from the candidate list that do not have sufficient distinct images. To this end, for each of the 500K names, 100 images are downloaded using Google Image Search and human annotators are instructed to retain subjects for which approximately 90% or more of the 100 images belong to a single identity. This removes candidates who do not have sufficient images or for which Google Image Search returns a mix of people for a single name. In this manner, we reduce the candidates to only 9244 names. Attribute information such as ethnicity and kinship is obtained from DBPedia [1].

Stage	Aim	Type	# of subject	total # of images	Annotation effort
1	Name list selection	M	500K	50.00 million	3 months
2	Image downloading	A	9244	12.94 million	-
3	Face detection	A	9244	7.31 million	-
4	Automatic filtering by classification	A	9244	6.99 million	-
5	Near duplicate removal	A	9244	5.45 million	-
6	Final automatic and manual filtering	A/M	9131	3.31 million	21days

Table II: Dataset statistics after each stage of processing in the collection pipeline.

Figure 1: VGGFace2 template examples. Left: pose templates from three different viewpoints (arranged by row) – frontal, three-quarter, profile. Right: age templates for two subjects for young and mature ages (arranged by row).

B. Stage 2: Obtaining images for each identity

We query in Google Image Search and download 1000 images for each subject. To obtain images with large pose and age variations, we then append the keyword 'sideview' and 'very young' to each name and download 200 images for each. This results in 1400 images for each identity.

C. Stage 3: Face detection

Faces are detected using the model provided by [26]. We use the hyper-parameters recommended in that work to favor a good trade-off between precision and recall. The face bounding box is then extended by a factor of 0.3 to include the whole head. Moreover, five facial landmarks are predicted by the same model.

D. Stage 4: Automatic filtering by classification

The aim of this stage is to remove outlier faces for each identity automatically. This is achieved by learning a classifier to identify the faces, and removing possible erroneous faces below a classification score. To this end, 1-vs-rest classifiers are trained to discriminate between the 9244 subjects. Specifically, faces from the top 100 retrieved images of each identity are used as positives, and the top 100 of all other identities are used as negative for training. The face descriptor features are obtained from the VGGFace [16] model. Then, the scores (between 0 and 1) from the trained model is used to sort images for each subject from most likely to least likely. By manually checking through images from a random 500 subjects, we choose a threshold of 0.5 and remove any faces below this.

E. Stage 5: Near duplicate removal

The downloaded images also contain exact or near duplicates due to the same images being found at different internet locations, or images differing only slightly in colour balance or JPEG artifacts for example. To alleviate this, duplicate images are removed by clustering VLAD descriptors for all images remaining at *stage* 4 and only retaining one image per cluster [3], [11].

F. Stage 6: Final automatic and manual filtering

At this point, two types of error may still remain: first, some classes still have outliers (i.e. images that do not belong to the person); and second, some classes contain a mixture of faces of more than one person, or they overlap with another class in the dataset. This stage addresses these two types of errors with a mix of manual and automated algorithms.

Detecting overlapped subjects. Subjects may overlap with other subjects. For instance, 'Will I Am' and 'William James Adams' in the candidate list refer to the same person. To detect confusions for each class, we randomly split the data for each class in half: half for training and the other for testing. Then, we train a ResNet-50 [8] and generate a confusion matrix by calculating top-1 error on the test samples. In this manner, we find 20 subjects confused with others. In this stage, we removed 19 noisy classes. In addition, we remove 94 subjects with samples less than 80 images, which results in a final list of 9131 identities.

Removing outlier images for a subject. The aim of this filtering, which is partly manual, is to achieve a purity

greater than 96%. We found that for some subjects, images with very high classifier scores at *stage* 4 can also be noisy. This happens when the downloaded images contain couples or band members who always appear together in public. In this case, the classifiers trained with these mixed examples at *stage* 4 tend to fail.

We retrain the model based on the current dataset, and for each identity the classifier score is used to divide the images into 3 sets: H (i.e. high score range [1, 0.95]), I (i.e. intermediate score range (0.95, 0.8]) and L (i.e. low score range (0.8, 0.5]). Human annotators clean up the images for each subject based on their scores, and the actions they carry out depends on whether the set H is noisy or not. If the set (H) contains several different people (noise) in a single identity folder, then set I and L (which have lower confidence scores), will undoubtedly be noisy as well, so all three sets are cleaned manually. In contrast, if set H is clean, then only set L (the lowest scores which is supposed to be the most noisy set) is cleaned up. After this, a new model is trained on the cleaned set H and L, and set I (intermediate scores, noise level is also intermediate) is then cleaned by model prediction. This procedure achieves very low label noise without requiring manual checking of every image.

G. Pose and age annotations

We train two networks to obtain the pose and age information for the dataset. To obtain head pose (roll, pitch, yaw), a 5-way classification ResNet-50 [8] is trained on the CASIA-WebFace dataset [25]. Then, this trained model is used to predict pose for all the images in the dataset.

Similarly, to estimate the apparent age, a 8-way classification ResNet-50 [8] is trained on IMDB-WIKI - 500k+ dataset [17]. Ages of faces are then predicted by this model.

V. EXPERIMENTS

In this section, we evaluate the quality of the VGGFace2 dataset by conducting a number of baseline experiments. We report the results on the VGGFace2 test set, and evaluate on the public benchmarks IJB-A [13] and IJB-B datasets [22]. The subjects in our training dataset are disjoint with the ones in IJB-A and IJB-B dataset. We also remove the overlap between MS-Celeb-1M and the two benchmarks when training the networks.

A. Experimental setup

Architecture. ResNet-50 [8] and SE-ResNet-50 [9] (SENet for short) are used as the backbone architectures for the comparison amongst training datasets. The Squeeze-and-Excitation (SE) blocks [9] adaptively recalibrate channel-wise feature responses by explicitly modelling channel relationships. They can be integrated with modern architectures, such as ResNet, and improve its representational power. This has been demonstrated for object and scene classification, with a Squeeze-and-Excitation network winning the ILSVRC 2017 classification competition.

The following experiments are developed under four settings: (a) networks are learned from scratch on VGGFace [16] (VF for short); (b) networks are learned from scratch on MS-Celeb-1M (MS1M for short) [7]; (c) networks are learned from scratch on VGGFace2 (VF2 for short); and, (d) networks are first pre-trained on MS1M, and then fine-tuned on VGGFace2 (VF2_ft for short).

Similarity computation. In all the experiments (i.e. for both verification and identification), we need to compute the similarity between subject templates. A template is represented by a single vector computed by aggregating the face descriptors of each face in the template set. In section V-B, the template vector is obtained by averaging the face descriptors of the images and SVM classifiers are used for identification. In sections V-C and V-D for IJB-A and IJB-B, where the template may contain both still images and video frames, we first compute the media vector (i.e. from images or video frames) by averaging the face descriptors in that media. A template vector is then generated by averaging the media vectors in that template, which is then L2 normalised. Cosine similarity is used to represent the similarity between two templates.

A face descriptor is obtained from the trained networks as follows: first the extended bounding box of the face is resized so that the shorter side is 256 pixels; then the centre 224×224 crop of the face image is used as input to the network. The face descriptor is extracted from from the layer adjacent to the classifier layer. This leads to a 2048 dimensional descriptor, which is then L2 normalised.

Training implementation details. All the networks are trained for classification using the soft-max loss function. During training, the extended bounding box of the face is resized so that the shorter side is 256 pixels, then a 224×224 pixels region is randomly cropped from each sample. The mean value of each channel is subtracted for each pixel.

Monochrome augmentation is used with a probability of 20% to reduce the over-fitting on colour images. Stochastic gradient descent is used with mini-batches of size 256, with a balancing-sampling strategy for each mini-batch due to the unbalanced training distributions. The initial learning rate is 0.1 for the models trained from scratch, and this is decreased twice with a factor of 10 when errors plateau. The weights of the models are initialised as described in [8]. The learning rate for model fine-tuning starts from 0.005 and decreases to 0.001.

B. Experiments on the new dataset

In this section, we evaluate ResNet-50 trained from scratch on the three datasets as described in the Sec. V-A, and VGGFace2 test set. We test identification performance and also similarity over pose and age, and validate the

capability of VGGFace2 to tackle pose and age variations.

Face identification. This scenario aims to predict, for a given test image, whose face it is. Specifically, for each of the 500 subjects in the evaluation set, 50 images are randomly chosen as the testing split and the remaining images are used as the training split. This training split is used to learn 1-vs-rest SVM classifiers for each subject. A top-1 classification error is then used to evaluate the performance of these classifiers on the test images. As shown in Table III, there is a significant improvement for the model trained on VGGFace2 rather than on VGGFace. This demonstrates the benefit of increasing data variation (e.g, subject number, pose and age variations) in the VGGFace2 training dataset. More importantly, models trained on VGGFace2 also achieve better result than that on MS1M even though it has tenfold more subjects and threefold more images, demonstrating the good quality of VGGFace2. In particular, the very low top-1 error of VGGFace2 provides evidence that there is very little label noise in the dataset – which is one of our design goals.

Training dataset	VGGFace	MS1M	VGGFace2
Top-1 error (%)	10.6	5.6	3.9

Table III: Identification performance (top-1 classification error) on the VGGFace2 test set for ResNet models trained on different datasets. A lower value is better.

Probing across pose. This test aims to assess how well templates match across three pose views: front, three-quarter and profile views. As described in section III-B, 300 subjects in the evaluation set are annotated with pose templates, and there are six templates for each subject: two each for front, three-quarter view and profile views.

These six templates are divided into two sets, one pose for each set, and a 3×3 similarity matrix is constructed between the two sets. Figure 3 visualises two example of these cosine similarity scores for front-to-profile templates.

Table IV compares the similarity matrix averaged over the 300 subjects. We can observe that (i) all the three models perform better when matching similar poses, i.e., front-to-front, three-quarter-to-three-quarter and profile-to-profile; and (ii) the performance drops when probing for different poses, e.g., front-to-three-quarter and front-to-profile, showing that recognition across poses is a much harder problem. Figure 2 shows histograms of similarity scores. It is evident that the mass of the VGGFace2 trained model is to the right of the MS1M and VGGFace trained models. This clearly demonstrates the benefit of training on a dataset with larger pose variation.

Probing across age. This test aims to assess how well templates match across age, for two ages ranges: young and mature ages. As described in section III-B, 100 subjects in the evaluation set are annotated with age templates, and there are four templates for each subject: two each for young and mature faces.

For each subject a 2×2 similarity matrix is computed, where an element is the cosine similarity between two templates. Figure 5 shows two examples of the young-to-mature templates, and their similarity scores.

Table V compares the similarity matrix averaged over the 100 subjects as the model changes. For all the three models, there is always a big drop in performance when matching across young and mature faces, which reveals that young-to-mature matching is substantially more challenging than young-to-young and mature-to-mature. Moreover, young-to-young matching is more difficult than mature-to-mature matching. Figure 4 illustrates the histograms of the young-to-mature template similarity scores.

Discussion. In the evaluation of pose and age protocols, models trained on VGGFace2 always achieve the highest similarity scores, and MS1M dataset the lowest. This can be explained by the fact that the MS1M dataset is designed to focus more on inter-class diversities, and this harms the matching performance across different pose and age, illustrating the value of VGGFace2 in having more intra-class diversities that cover large variations in pose and age.

C. Experiments on IJB-A

In this section, we compare the performance of the models trained on the different datasets on the public IARPA Janus Benchmark A (IJB-A dataset) [13].

The IJB-A dataset contains 5712 images and 2085 videos from 500 subjects, with an average of 11.4 images and 4.2 videos per subject. All images and videos are captured from unconstrained environment and show large variations in expression and image qualities. As a pre-processing, we detect the faces using MTCNN [26] to keep the cropping consistent between training and evaluation.

IJB-A provides ten-split evaluations with two standard protocols, namely, 1:1 face verification and 1:N face identification, where we directly extract the features from the models for the test sets and use cosine similarity score. For verification, the performance is reported using the true accept rates (TAR) vs. false positive rates (FAR) (i.e. receiver operating characteristics (ROC) curve). For identification, the performance is reported using the true positive identification rate (TPIR) vs. false positive identification rate (FPIR) (equivalent to a decision error trade-off (DET) curve) and the Rank-N (i.e. the cumulative match characteristic (CMC) curve). Table VI and Figure 6 presents the comparison results.

The effect of training set. We first investigate the effect of different training sets based on the same architecture

Training dataset	VGGFace			MS1M			VGGFace2		
	front	three-quarter	profile	front	three-quarter	profile	front	three-quarter	profile
front	0.5781	0.5679	0.4821	0.5661	0.5582	0.4715	0.6876	0.6821	0.6222
three-quarter	0.5706	0.5957	0.5345	0.5628	0.5766	0.5036	0.6859	0.6980	0.6481
profile	0.4859	0.5379	0.5682	0.4776	0.5064	0.5094	0.6264	0.6515	0.6488

Table IV: Face probing across poses. Similarity scores are evaluated across pose templates. A higher value is better.

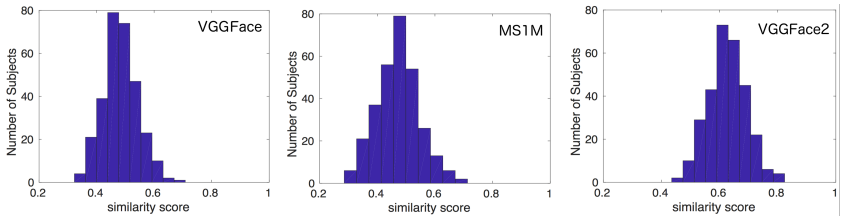

Figure 2: Histograms of similarity scores for front-to-profile matching for the models trained on different datasets.

Figure 3: Two example templates of front-to-profile matching. Left: the similarity scores produced by VGGFace, MS1M, VGGFace2 are 0.41, 0.35 and 0.59, respectively; Right: the scores are 0.41, 0.31 and 0.57, respectively.

Training dataset	VGGFace		MS1M		VGGFace2	
	young	mature	young	mature	young	mature
young	0.5231	0.4338	0.4983	0.4005	0.6256	0.5524
mature	0.4394	0.5518	0.4099	0.5276	0.5607	0.6637

Table V: Face probing across ages. Similarity scores are evaluated across age templates. A higher value is better.

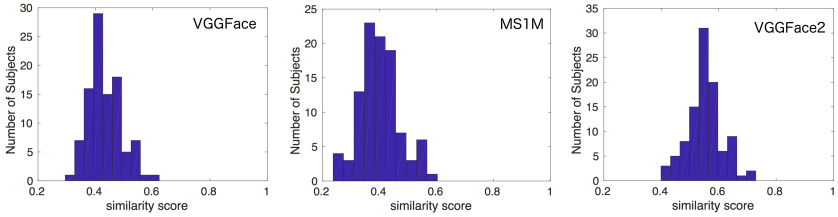

Figure 4: Histograms of similarity scores for young-to-mature matching for the models trained on different datasets.

Figure 5: Two example templates of young-to-mature matching. Left: the similarity scores produced by VGGFace, MS1M, VGGFace2 are 0.42, 0.30 and 0.58, respectively; Right: the scores are 0.43, 0.41 and 0.73, respectively.

ResNet-50 (Table VI), and start with networks trained from scratch. we can observe that the model trained on VGGFace2 outperforms the one trained on VGGFace by a large margin, even though VGGFace has a similar scale (2.6M images) it has fewer identities and pose/age variations (and more label noise). Moreover, the model of VGGFace2 is significantly superior to the one of MS1M which has 10 times subjects over our dataset. Specially, it achieve $\sim 4.4\%$ improvement over MS1M on FAR=0.001 for verification, $\sim 3.7\%$ on FPIR=0.01 and $\sim 1.5\%$ on Rank-1 for identification.

When comparing with the results of existing works, the model trained on VGGFace2 surpasses previously reported results on all metrics (best to our knowledge, reported on IJB-A 1:1 verification and 1:N identification protocols),

Training dataset	Arch.	1:1 Verification TAR			1:N Identification TPIR				
		FAR=0.001	FAR=0.01	FAR=0.1	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5	Rank-10
VGGFace [16]	ResNet-50	0.620 ± 0.043	0.834 ± 0.021	0.954 ± 0.005	0.454 ± 0.058	0.748 ± 0.024	0.925 ± 0.008	0.972 ± 0.005	0.983 ± 0.003
MS1M [7]	ResNet-50	0.851 ± 0.030	0.939 ± 0.013	0.980 ± 0.003	0.807 ± 0.041	0.920 ± 0.012	0.961 ± 0.006	0.982 ± 0.004	0.990 ± 0.002
VGGFace2	ResNet-50	0.895 ± 0.019	0.950 ± 0.005	0.980 ± 0.003	0.844 ± 0.035	0.924 ± 0.006	0.976 ± 0.004	0.992 ± 0.002	0.995 ± 0.001
VGGFace2_ft	ResNet-50	0.908 ± 0.017	0.957 ± 0.007	0.986 ± 0.002	0.861 ± 0.027	0.936 ± 0.007	0.978 ± 0.005	0.992 ± 0.003	0.995 ± 0.001
VGGFace2	SENet	0.904 ± 0.020	0.958 ± 0.004	0.985 ± 0.002	0.847 ± 0.051	0.930 ± 0.007	0.981 ± 0.003	$\mathbf{0.994 \pm 0.002}$	$\mathbf{0.996 \pm 0.001}$
VGGFace2_ft	SENet	$\mathbf{0.921 \pm 0.014}$	$\mathbf{0.968 \pm 0.006}$	$\mathbf{0.990 \pm 0.002}$	$\mathbf{0.883 \pm 0.038}$	$\mathbf{0.946 \pm 0.004}$	$\mathbf{0.982 \pm 0.004}$	0.993 ± 0.002	0.994 ± 0.001
Crosswhite *et al.* [6]	-	0.836 ± 0.027	0.939 ± 0.013	0.979 ± 0.004	0.774 ± 0.049	0.882 ± 0.016	0.928 ± 0.010	0.977 ± 0.004	0.986 ± 0.003
Sohn *et al.* [19]	-	0.649 ± 0.022	0.864 ± 0.007	0.970 ± 0.001	-	-	0.895 ± 0.003	0.957 ± 0.002	0.968 ± 0.002
Bansalit et al. [4]	-	$0.730^{\dagger}$	0.874	$0.960^{\dagger}$	-	-	-	-	-
Yang *et al.* [24]	-	0.881 ± 0.011	0.941 ± 0.008	0.978 ± 0.003	0.817 ± 0.041	0.917 ± 0.009	0.958 ± 0.005	0.980 ± 0.005	0.986 ± 0.003

Table VI: Performance evaluation on the IJB-A dataset. A higher value is better. The values with † are read from [4].

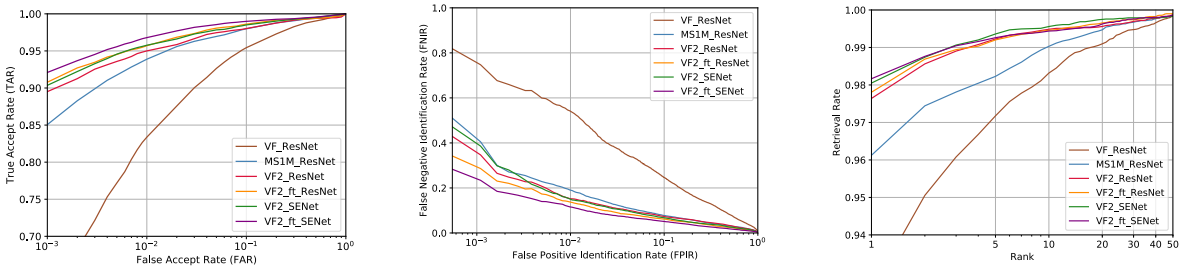

Figure 6: Results on the IJB-A dataset (average over 10 splits). Left: ROC (higher is better); Middle: DET (lower is better); Right: CMC (higher is better).

which further demonstrate the advantage of the VGGFace2 dataset. In addition, the generalisation power can be further improved by first training with MS1M and then fine-tuning with VGGFace2 (i.e. "VGGFace2_ft"), however, the difference is only 0.908 vs. 0.895.

Many existing datasets are constructed by following the assumption of the superiority of wider dataset (more identities) [7], [12], [25], where the huge number of subjects would increase the difficulty of model training. In contrast, VGGFace2 takes both aspects of breath (subject number) and depth (sample number per subject) into account, guaranteeing rich intra-variation and inter-diversity.

The effect of architectures. We next investigate the effect of architectures trained on VGGFace2 (Table VI). The comparison between ResNet-50 and SENet both learned from scratch reveals that SENet has a consistently superior performance on both verification and identification. More importantly, SENet trained from scratch achieves comparable results to the fine-turned ResNet-50 (i.e. first pre-trained on the MS1M dataset), demonstrating that the diversity of our dataset can be further exploited by an advanced network. In addition, the performance of SENet can be further improved by training on the two datasets VGGFace2 and MS1M, exploiting the different advantages that each offer.

D. Experiments on IJB-B

The IJB-B dataset is an extension of IJB-A, having $1,845$ subjects with 21.8K still images (including $11,754$ face and $10,044$ non-face) and 55K frames from $7,011$ videos. We evaluate the models on the standard 1:1 verification protocol (matching between the Mixed Media probes and two galleries) and 1:N identification protocol (1:N Mixed Media probes across two galleries).

We observe a similar behaviour to that of the IJB-A evaluation. For the comparison between different training sets (Table VII and Figure 7), the models trained on VGGFace2 significantly surpass the ones trained on MS1M, and the performance can be further improved by integrating the advantages of the two datasets. In addition, SENet's superiority over ResNet-50 is evident in both verification and identification with the two training settings (i.e. trained from scratch and fine-tuned). Moreover, we also compare to the results reported by others on the benchmark [22] (as shown in Table VII), and there is a considerable improvement over their performance for all measures.

VI. Conclusion

In this work, we have proposed a pipeline for collecting a high-quality dataset, VGGFace2, with a wide range of pose and age. Furthermore, we demonstrate that deep models (ResNet-50 and SENet) trained on VGGFace2, achieve state-of-the-art performance on the IJB-A and IJB-B benchmarks. The dataset and models are available at https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/.

Acknowledgment

We would like to thank Elancer and Momenta for their part in preparing the dataset[2]. This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract number 2014-14071600010. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either

[2]http://elancerits.com/ https://momenta.ai/

Training dataset	Arch.	1:1 Verification TAR				1:N Identification TPIR				
		FAR=$1E-5$	FAR=$1E-4$	FAR=$1E-3$	FAR=$1E-2$	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5	Rank-10
VGGFace [16]	ResNet-50	0.342	0.535	0.711	0.850	0.429 ± 0.024	0.635 ± 0.015	0.752 ± 0.038	0.843 ± 0.032	0.874 ± 0.026
MS1M [7]	ResNet-50	0.548	0.743	0.857	0.935	0.662 ± 0.036	0.810 ± 0.028	0.865 ± 0.053	0.917 ± 0.032	0.936 ± 0.024
VGGFace2	ResNet-50	0.647	0.784	0.878	0.938	0.701 ± 0.038	0.824 ± 0.034	0.886 ± 0.032	0.936 ± 0.019	0.953 ± 0.013
VGGFace2_ft	ResNet-50	0.671	0.804	0.891	0.947	0.702 ± 0.041	0.843 ± 0.032	0.894 ± 0.039	0.940 ± 0.022	0.954 ± 0.016
VGGFace2	SENet	0.671	0.800	0.888	0.949	0.706 ± 0.047	0.839 ± 0.035	0.901 ± 0.030	0.945 ± 0.016	0.958 ± 0.010
VGGFace2_ft	SENet	**0.705**	**0.831**	**0.908**	**0.956**	$\mathbf{0.743 \pm 0.037}$	$\mathbf{0.863 \pm 0.032}$	$\mathbf{0.902 \pm 0.036}$	$\mathbf{0.946 \pm 0.022}$	$\mathbf{0.959 \pm 0.015}$
Whitelam *et al.* [22]	-	0.350	0.540	0.700	0.840	0.420	0.640	0.790	0.850	0.900

Table VII: Performance evaluation on the IJB-B dataset. A higher value is better. The results of [22] are read from the curves reported in the paper. Note, [22] has a different evaluation for the verification protocol where pairs generated from different galleries are evaluated separately and averaged to get the final results.

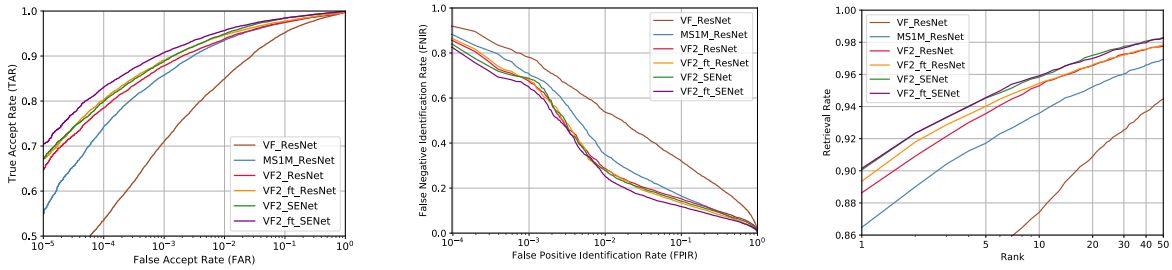

Figure 7: Results on the IJB-B dataset across gallery sets S1 and S2.
Left: ROC (higher is better); Middle: DET (lower is better); Right: CMC (higher is better).

REFERENCES

[1] Dbpedia. http://wiki.dbpedia.org/.

[2] Freebase. http://www.freebase.com/.

[3] R. Arandjelović and A. Zisserman. All about VLAD. In *Proc. CVPR*, 2013.

[4] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa. The do's and don'ts for cnn-based face verification. *arXiv preprint arXiv:1705.07426*, 2017.

[5] A. Bansal, A. Nanduri, C. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. *arXiv preprint arXiv:1611.01484*, 2016.

[6] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. In *Automatic Face & Gesture Recognition (FG)*, pages 1–8. IEEE, 2017.

[7] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *arXiv preprint arXiv:1607.08221*, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] J. Hu, L. Shen, and G. Sun. Squeeze-and-Excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

[10] 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, 2007.

[11] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *TPAMI*, 34(9):1704–1716, 2012.

[12] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, pages 4873–4882, 2016.

[13] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, pages 1931–1939, 2015.

[14] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *ICIP*, pages 343–347. IEEE, 2014.

[15] G. Panis and A. Lanitis. An overview of research activities in facial age estimation using the fg-net aging database. In *ECCV*, pages 737–750. Springer, 2014.

[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. BMVC.*, 2015.

[17] R. Rothe, R. Timofte, and L. Van Gool. Dex: Deep expectation of apparent age from a single image. In *CVPR Workshops*, pages 10–15, 2015.

[18] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[19] K. Sohn, S. Liu, G. Zhong, X. Yu, M.-H. Yang, and M. Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. *arXiv preprint arXiv:1708.02191*, 2017.

[20] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, pages 1891–1898, 2014.

[21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *CVPR*, pages 2746–2754, 2015.

[22] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, et al. Iarpa janus benchmark-b face dataset. In *CVPR Workshop on Biometrics*, 2017.

[23] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534. IEEE, 2011.

[24] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *CVPR*, pages 4362–4371, 2017.

[25] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.