The Devil of Face Recognition is in the Noise

Fei Wang *[1][0000−0002−1024−5867], Liren Chen *[2][0000−0003−0113−5233], Cheng Li[1][0000−0002−0892−4705], Shiyao Huang[1][0000−0002−5198−2492], Yanjie Chen[1][0000−0003−1918−6776], Chen Qian[1][0000−0002−8761−5563], and Chen Change Loy[3][0000−0001−5345−1591]

[1] SenseTime Research
[2] University of California San Diego
[3] Nanyang Technological University
{wangfei, chengli, huangshiyao, chenyanjie, qianchen}@sensetime.com, lic002@eng.ucsd.edu, ccloy@ieee.org

Abstract. The growing scale of face recognition datasets empowers us to train strong convolutional networks for face recognition. While a variety of architectures and loss functions have been devised, we still have a limited understanding of the source and consequence of label noise inherent in existing datasets. We make the following contributions: 1) We contribute cleaned subsets of popular face databases, *i.e.*, MegaFace and MS-Celeb-1M datasets, and build a new large-scale noise-controlled IMDb-Face dataset. 2) With the original datasets and cleaned subsets, we profile and analyze label noise properties of MegaFace and MS-Celeb-1M. We show that a few orders more samples are needed to achieve the same accuracy yielded by a clean subset. 3) We study the association between different types of noise, *i.e.*, label flips and outliers, with the accuracy of face recognition models. 4) We investigate ways to improve data cleanliness, including a comprehensive user study on the influence of data labeling strategies to annotation accuracy. The IMDb-Face dataset has been released on https://github.com/fwang91/IMDb-Face.

1 Introduction

Datasets are pivotal to the development of face recognition. From the early FERET dataset [16] to the more recent LFW [7], MegaFace [8, 13], and MS-Celeb-1M [5], face recognition datasets play a main role in driving the development of new techniques. The datasets not only become more diverse, the scale of data is also growing tremendously. For instance, MS-Celeb-1M [5] contains around 10M images for 100K celebrities, far exceeding FERET [16] that only has 14,126 images from 1,199 individuals. Large-scale datasets together with the emergence of deep learning have led to the immense success of face recognition in recent years.

Large-scale datasets are inevitably affected by label noise. The problem is pervasive since well-annotated datasets in large-scale are prohibitively expensive

* = equal contribution

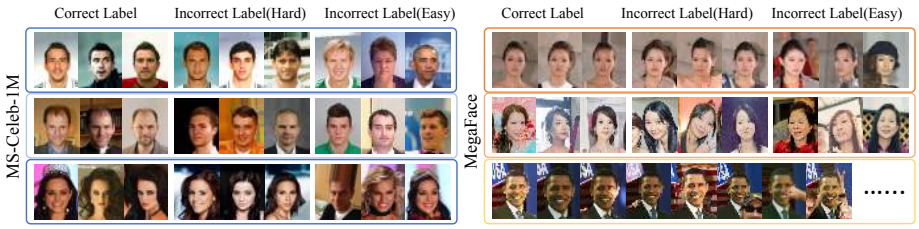

Fig. 1. Label noises in MegaFace [13] and MS-Celeb-1M [5]. Each row depicts images that are labeled with the same identity. Some incorrect labels are easy while many of them are hard.

and time-consuming to collect. That motivates researchers to resort to cheap but imperfect alternatives. A common method is to query celebrities' images by their names on search engines, and subsequently clean the labels with automatic or semi-automatic approaches [15, 11, 4]. Other methods introduce clustering with constraints on social photo sharing sites. The aforementioned methods offer a viable way to scale the training samples conveniently but also bring label noises that adversely affect the training and performance of a model. We show some samples with label noises in Figure 1. As can be seen, MegaFace [13] and MS-Celeb-1M [5] consist considerable incorrect identity labels. Some noisy labels are easy to remove while many of them are hard to be cleaned. In MegaFace, there are a number of redundant images too (shown in the last row).

The first goal of this paper is to develop an understanding of the source of label noise and its consequences towards face recognition by deep convolutional neural networks (CNN) [19, 18, 23, 6, 1, 26]. We seek answers to questions like: How many noisy samples are needed to achieve an effect tantamount to clean data? What is the relationship between noise and final performance? What is the best strategy to annotate face identities? A better understanding of the aforementioned questions would help us to design a better data collection and cleaning strategy, avoid pitfalls in training, and formulate stronger algorithms to cope with real-world problems. To facilitate our research, we manually clean subsets of two most popular face recognition databases, namely, MegaFace [13] and MS-Celeb-1M [5]. We observe that a model trained with only 32% of MegaFace or 20% of MS-Celeb-1M cleaned subsets, can already achieve comparable performance with models that are trained on the respective full dataset. The experiments suggest that a few orders more samples are needed for face recognition model training if noisy samples are used.

The second goal of our study is to build a clean face recognition dataset for the community. The dataset could help training better models and facilitate further understanding of the relationship between noise and face recognition performance. To this end, we build a clean dataset called **IMDb-Face**. The dataset consists of 1.7M images of 59K celebrities collected from movie screenshots and

Table 1. Various face recognition datasets.

Dataset	#Identities	#Images	Source	Cleaned?	Availablity
LFW [7]	5K	13K	Search Engine	Automatic Detection	Public
CelebFaces [19, 20]	10K	202K	Search Engine	Manually Cleaned	Public
VGG-Face [15]	2.6K	2.5M	Search Engine	Semi-automated Clean	Public
CASIA-WebFace [25]	10k	0.5M	IMDb	Automatic Clean	Public
MS-Celeb-1M(v1) [5]	100k	10M	Search Engine	None	Public
MegaFace [13]	670K	4.7M	Flickr	Automatic Cleaned	Public
Facebook [21]	4k	4.4M	–	–	Private
Google [18]	8M	200M	–	–	Private
IMDb-Face	**59K**	**1.7M**	**IMDb**	**Manually Cleaned**	**Public**

posters from the IMDb website[1]. Due to the nature of the data source, the images exhibit large variations in scale, pose, lighting, and occlusion. We carefully clean the dataset and simulate corruption by injecting noise on the training labels. The experiments show that the accuracy of face recognition decreases rapidly and nonlinearly with the increase of label noises. In particular, we confirm the common belief that the performance of face recognition is more sensitive towards label flips (example has erroneously been given the label of another class within the dataset) than outliers (image does not belong to any of the classes under consideration, but mistakenly has one of their labels). We also conduct an interesting experiment to analyze the reliability of different ways of annotating a face recognition dataset. We found that label accuracy correlates with time spent on annotation. The study helps us to find the source of erroneous labels and thereafter design better strategies to balance annotation cost and accuracy.

We hope that this paper could shed lights on the influences of data noise to the face recognition task, and point to potential labelling strategies to mitigate some of the problems. We contribute the new data **IMDb-Face** with the community. It could serve as a relatively clean data to facilitate future studies of noises in large-scale face recognition. It can also be used as a training data source to boost the performance of existing methods, as we will show in the experiments.

2 How Noisy is Existing Data?

We first introduce some popular datasets used in face recognition study and then approximate their respective signal-to-noise ratio.

2.1 Face Recognition Datasets

Table 2.1 provides a summary of representative datasets used in face recognition research.

[1] www.IMDb.com

LFW: Labeled Faces in the Wild (LFW) [7] is perhaps the most popular dataset to date for benchmarking face recognition approaches. The database consists of 13,000 facial images of 1,680 celebrities. Images are collected from Yahoo News by running the Viola-Jones face detector. Limited by the detector, most of the faces in LFW is frontal. The dataset is considered sufficiently clean despite some incorrectly labeled matched pairs are reported. Errata of LFW are provided in http://vis-www.cs.umass.edu/lfw/.

CelebFaces: CelebFaces [19, 20] is one of the early face recognition training databases that are made publicly available. Its first version contains 5,436 celebrities and 87,628 images, and it was upgraded to 10,177 identities and 202,599 images in a year later. Images in CelebFaces were collected from search engines and manually cleaned by workers.

VGG-Face: VGG-Face [15] contains 2,622 identities and 2.6M photos. More than 2,000 images per celebrity were downloaded from search engines. The authors treat the top 50 images as positive samples and train a linear SVM to select the top 1,000 faces. To avoid extensive manual annotation, the dataset was 'block-wise' verified, *i.e.*, ranked images of each identity are displayed in blocks and annotators are asked to validate blocks as a whole. In this study we did not focus on VGG-Face [15] since it should have the similar 'search-engine bias' problem with MS-Celeb-1M [5].

CASIA-WebFace: The images in CASIA-WebFace [25] were collected from IMDb website. The dataset contains 500K photos of 10K celebrities and it is semi-automatically cleaned via tag-constrained similarity clustering. The authors start with each celebrity's main photo and those photos that contain only one face. Then faces are gradually added to the dataset constrained by feature similarity and name tag. CASIA-WebFace uses the same source as the proposed IMDb-Face dataset. However, limited by the feature and clustering steps, CASIA-WebFace may fail to recall many challenging faces.

MS-Celeb-1M: MS-Celeb-1M [5] contains 100K celebrities who are selected from the 1M celebrity list in terms of their popularities. Public search engines are then leveraged to provide approximately 100 images for each celebrity, resulting in about 10M web images. The data is deliberately left uncleaned for several reasons. Specifically, collecting a dataset of this scale requires tremendous efforts in cleaning the dataset. Perhaps more importantly, leaving the data in this form encourages researchers to devise new learning methods that can naturally deal with the inherent noises.

MegaFace: Kemelmacher-Shlizerman *et al.* [13] clean massive number of images published on Flickr by proposing algorithms to cluster and filter face data from the YFCC100M dataset. For each user's albums, the authors merge face pairs with a distance closer than β times of average distance. Clusters that contain more than three faces are kept. Then they drop 'garbage' groups and clean potential outliers in each group. A total of 672K identities and 4.7M images were collected. MegaFace2 avoids 'search-engine' bias as in VGG-Face [15] and MS-Celeb-1M [5]. However, we found this cluster-based approach introduces new bias. MegaFace prefers small groups with highly duplicated images, *e.g.*, face

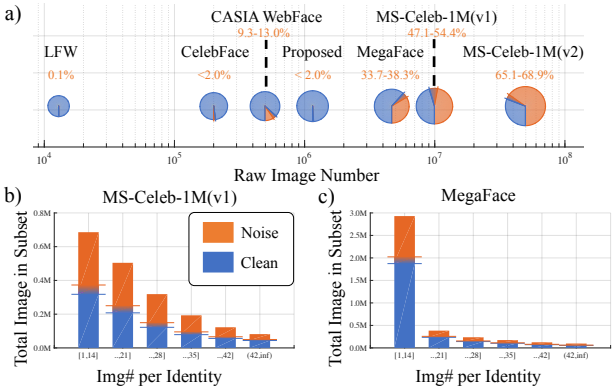

Fig. 2. (a) A visualization of size and estimated noise percentage of datasets. (b) Noise distribution of MS-Celeb-1M(v1) [5]. (c) Noise distribution of MegaFace [13]. The two horizontal lines in each bar represent the lower- and upper-bounds of noise, respectively. See Sec. 2.2 for details.

captured from the same video. Limited by the base model for clustering, considerable groups in MegaFace contain noises, or sometimes mess up multiple people in the same group.

2.2 An Approximation of Signal-to-Noise Ratio

Owing to the source of data and cleaning strategies, existing large-scale datasets invariably contain label noises. In this study, we aim to profile the noise distribution in existing datasets. Our analysis may provide a hint to future research on how one should exploit the distribution of these data.

It is infeasible to obtain the exact number of these noises due to the scale of the datasets. We bypass this difficulty by randomly selecting a subset of a dataset and manually categorize them into three groups – 'correct identity assigned', 'doubtful', and 'wrong identity assigned'. We select a subset of 2.7M images from MegaFace [13] and 3.7M images from MS-Celeb-1M [5]. For CASIA-WebFace [25] and CelebFaces [19, 20], we sampled 30 identities to estimate their signal-to-noise ratio. The final statistics are visualized in Figure 2(a). Due to the difficulty in estimating the exact ratio, we approximate an upper and a lower bound of noisy data during the estimation. The lower-bound is more optimistic considering doubtful labels as clean data. The upper-bound is more pessimistic considering all doubtful cases as badly labeled. We provide more details on the estimations in the supplementary material. As observed in Figure 2(a), the noise percentage increases dramatically along the scale of data. This is not surprising given the difficulty in data annotation. It is noteworthy that the proposed IMDb-Face pushes the envelope of large-scale data with a very high signal-to-noise ratio (noise is under 10% of the full data).

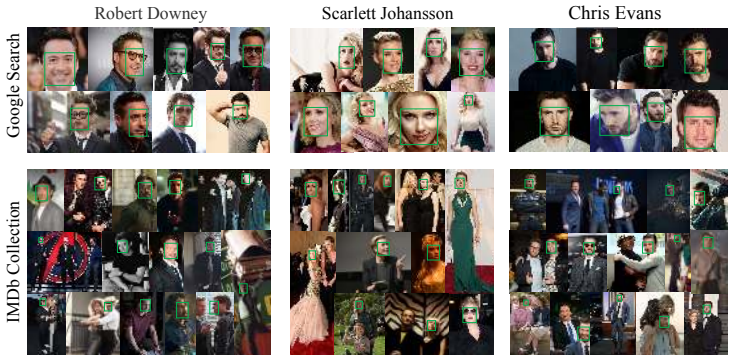

Fig. 3. The second row depicts the raw data from the IMDb website. As a comparison, we show the images of the same identity queried from the Google search engine in the first row.

We investigate further the noise distribution of the two largest public datasets to date, MS-Celeb-1M [5] and MegaFace [13]. We first categorize identities in a dataset based on their number of images. A total of six groups/bins are established. We then plot a histogram showing the signal-to-noise ratio of each bin along the noise lower- and upper-bounds. As can be seen in Figure 2(b,c), both datasets exhibit a long-tailed distribution, *i.e.*, most identities have very few images. This phenomenon is especially obvious on the MegaFace [13] dataset since it uses automatically formed clusters for determining identities, therefore, the same identity may be distributed in different clusters. Noises across all groups in MegaFace [13] are less in comparison to MS-Celeb-1M [5]. However, we found that many images in the clean portion of MegaFace [13] are duplicated images. In Sec. 4.2, we will perform experiments on the MegaFace and MS-Celeb-1M datasets to quantify the effect of noise on the face recognition task.

3 Building a Noise-Controlled Face Dataset

As shown in the previous section, face recognition datasets that are more than a million scale typically have a noise ratio higher than 30%. How about building a large scale noise controlled face dataset? It can be used to train better face recognition algorithms. More importantly, it can be used to further understand the relationship between noise and face recognition performance. To this end, we seek not only a cleaner and more diverse source to collect face data, but also an effective way to label the data.

3.1 Celebrity Faces from IMDb

Search engines are important sources from which we can quickly construct a large-scale dataset. The widely used ImageNet [3] was built by querying images from Google Image. Most of the face recognition datasets were built in the

same way (except MegaFace [13]). While querying from search engines offers the convenience of data collection, it also introduces data bias. Search engines usually operate in a high-precision regime [2]. Observing the queried images in Figure 3, they tend to have a simple background with sufficient illumination, and the subjects are often in a near frontal posture. These data, to a certain extent, are more restricted than those we could observe in reality, *e.g.*, faces in videos (IJB-A [9] and YTF [24]) and selfie photos (millions of distractors in MegaFace). Another pitfall in crawling images from search engines is the low recall rate. We performed a simple analysis and found that on average the recall rate is only 40% for the first 200 photos we query for a particular name.

In this study, we turn our data collection source to the IMDb website. IMDb is more structured. It includes a diverse range of photos under each celebrity's profile, including official photos, lifestyle photos, and movie snapshots. Movie snapshots, we believe, provide essential data samples for training a robust face recognition model. Those screenshots are rarely returned by querying a search engine. In addition, the recall rate is much higher (90% on average) when we query a name on IMDb. This is much higher than 40% from search engines. The IMDb website lists about 300K celebrities who have official and gallery photos. By clawing IMDb dataset, we finally collected and cleaned 1.7M raw images from 59K celebrities.

3.2 Data Distribution

Figure 4-a presents the distribution of yaw angle in our dataset compared with MS-Celeb-1M and MegaFace. Figures 4-c, -d and -e present the age, gender and race distributions. As can be observed, images in IMDb-Face exhibit larger pose variations, and they also show diversity in age, gender and race.

3.3 How Good can Human Label Identity?

The data downloaded from IMDb are noisy as multiple celebrities may co-exist on the same image. We still need to clean the dataset before it can be used for training. We take this opportunity to study how human annotators would clean a face data. The study will help us to identify the source of noise during annotation and design a better data cleaning strategy for the full dataset.

For the purpose of the user study, we extract a small subset of 30 identities from the IMDb raw data. We carefully select three images with confirmed identity serving as gallery images. The remaining images of these 30 identities are treated as query images. To make the user study more challenging and statistically more meaningful, we inject 20% outliers to the query set. Next, we prepare three annotation schemes as follows. The interface of each scheme is depicted in Figure 5.

Scheme I - Draw the box: We present the target person to a volunteer by showing the three gallery faces. We then show a query image selected from the query set. The image may contain multiple persons. If the target appears in the query image, the volunteer is asked to draw a bounding box on the target. The

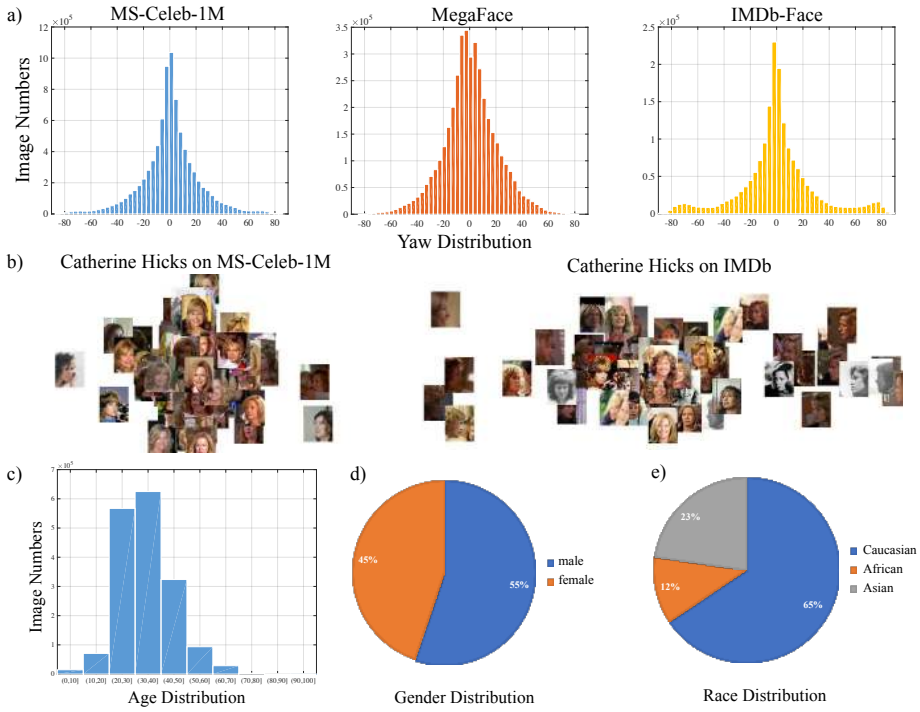

Fig. 4. a) Comparing the distribution of yaw angle of images in the proposed dataset against MS-Celeb-1M and MegaFace. b) A qualitative sample from the proposed IMDb-Face and MS-Celeb-1M. c) Age distribution of images in IMDb-Face. d) Gender distribution of identities in IMDb-Face. e) Race distribution of identities in IMDb-Face.

volunteer can either confirm the selection or assign a 'doubt' flag on the box if he/she is not confident about the choice. 'No target' is selected when he/she cannot find the target person.

Scheme II - Choose 1 in 3: Similar to Scheme I, we present the target person to a volunteer by showing the gallery images. We then randomly sample three faces detected from the query set, from which the volunteer will select a single image as the target face. We ensure that all query faces have the same gender as the target person. Again, the volunteer can choose a 'doubt' flag if he/she is not confident about the selection or choose 'no target' at all.

Scheme III - Yes or No: Binary query is perhaps be the most natural and popular way to clean a face recognition set. We first rank all faces based on their similarity to probe faces in the gallery, and then ask a volunteer to make a choice if each belongs to the target person. The volunteer is allowed to answer 'doubt'.

Which scheme to choose?: Before we can quantify the effectiveness of different schemes, we first need to generate the ground truth of these 30 identities. We use a 'consensus' approach. Specifically, each of the aforementioned schemes was conducted on three different volunteers. We ensure that each query face

Fig. 5. Interfaces for user study: (a) Scheme I - volunteers were asked to draw a box on the target's face. (b) Scheme II - given three query faces, volunteers were asked to select the face that belongs to the target person. (c) Scheme III - volunteers were asked to select the face that belongs to the target.

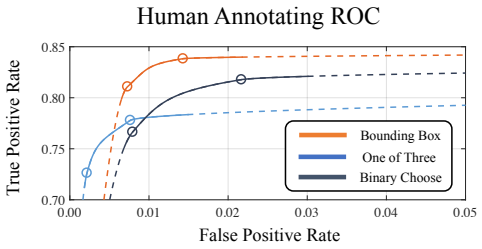

Fig. 6. A ROC comparison between three different annotating schemes; volunteers were allowed to select 'doubt' so two data points can be obtained depending if we count doubt data as positive or negative.

was annotated nine times across the three schemes. If four of the annotations consistently point to the same identity, we assign the query face to the targeted identity. With this ground truth, we can measure the effectiveness of each annotation scheme.

Figure 6 shows the Receiver operating characteristic (ROC) curve of each of the three schemes[2]. Scheme I achieves the highest F_1 score. It recalls more than 90% faces with under 10% false positive samples. Finding a face and drawing a box seems to make annotators more focused on finding the right face. Scheme II provides a high true positive rate when the false positive is low. The existence of distractors forces annotators to work harder to match the faces. Scheme III yields the worse true positive rate when the false positive is low. This is not surprising since this task is much easier than Schemes I and II. The annotators tend to make mistakes given this relaxing task, especially after a prolonged annotation process. We observe an interesting phenomenon: *the longer a volunteer spends on annotating a sample, the more accurate the annotation is.* With full speed in one hour, each volunteer can draw 180-300 faces in Scheme I, or finish around 600 selections in Scheme II, or answer over 1000 binary questions in Scheme

[2] We should emphasize that the curves in Figure 6 are different from actual human's performance on verifying arbitrary face pairs. This is because in our study the faces from a query set are very likely to belong to the same person. The ROC thus represents human's accuracy on 'verifying face pairs that likely belong to the same identity'.

III. We believe the most reliable way to clean a face recognition dataset is to leverage both Schemes I and II to achieve a high precision and recall. Limited by our budget, we only conducted Scheme I to clean the IMDb-Face dataset.

During the cleaning of the IMDb-Face, since multiple identities may co-exist on the same image, first we annotated gallery images to make sure the queried identity. The gallery images come from the official gallery provided by the IMDb website, which most of these official gallery images contain the true identity. We ask volunteers to look through the 10 gallery images back and forth and draw bounding box of the face that occurs most frequently. Then, annotators label the rest of the queried images guided by the three largest labeled faces as galleries. For identities having fewer than three gallery images, their queried images may have too much noise. To save labor, we did not annotate their images.

It took 50 annotators one month to clean the IMDb-Face dataset. Finally, we obtained 1.7M clean facial images from 2M raw images. We believe that the cleaning is of high quality. We estimate the noise level of IMBb-Face as the product of approximated noise level in the IMDb raw data ($2.7 \pm 4.5\%$) and the false positive rate (8.7%) of Scheme I. The noise level is controlled under 2%. The quality of IMDb-Face is validated in our experiments.

4 Experiments

We divide our experiments into a few sections. First, we conduct ablation studies by simulating noise on our proposed dataset. The studies help us to observe the deterioration of performance in the presence of increasing noise, or when a fixed amount of clean data is diluted with noise. Second, we perform experiments on two existing datasets to further demonstrate the effect of noise. Third, we examine the effectiveness of our dataset by comparing it to other datasets with the same training condition. Finally, we compare the model trained on our dataset with other state-of-the-arts. Next, we describe the experimental setting.

Evaluation Metric: We report rank-1 identification accuracy on the Megaface benchmark [8]. It is a very challenging task to evaluate the performance of face recognition methods at the million scale of distractors. The MegaFace benchmark consists of one gallery set and one probe set. The gallery set contains more than 1 million images and the probe set consists of two existing datasets: Facescrub [14] and FGNet. We use Facescrub [14] as MegaFace probe dataset in our experiments. Verification performance of MegaFace (reported as TPR at FPR= 10^{-6}) is included in the supplementary material due to page limit. We also test LFW [7] and YTF [24] in Section 4.4.

Architecture: To better examine the effect of noise, we use the same architecture in all experiments. After a comparison among ResNet-50, ResNet-101 and Attention-56 [22], we finally choose Attention-56 that achieves a good balance between computation and accuracy. As a reference, the model converges on a database with 80 hours on an 8-GPU server with a batch-size of 256. The output of Attention-56 is a 256-dimensional feature for each input image. We use cosine similarity to compute scores between image pairs.

Pre-processing: We cropped and aligned faces, then rigidly transferred them onto a mean shape. Then we resized the cropped image into 224×256, and subtracted them with the mean value in each RGB channel.
Loss: We apply three losses: SoftMax [20], Center Loss [23] and A-Softmax [12]. Our implementation is based on the public implementation of these losses:
Softmax: Softmax loss is the most commonly used loss, either for model initialization or establishing a baseline.
Center Loss: Wen *et al.* [23] propose center loss, which minimizes the intra-class distance to enhance features' discriminative power. The authors jointly trained CNN with the center loss and the softmax loss.
A-Softmax: Liu *et al.* [12] formulate A-Softmax to explicitly enforce the angle margin between different identities. The weight vector of each category was restricted on a hypersphere.

4.1 Investigating the Effect of Noise on IMDb-Face

The proposed IMDb-Face dataset enables us to investigate the effect of noise. There are two common types of noise in large-scale face recognition datasets: 1) *label flips*: example has erroneously been given the label of another class within the dataset 2) *outliers*: image does not belong to any of the classes under consideration, but mistakenly has one of their labels. Sometimes even non-faces may be mistakenly included. To simulate the first type of noise, we randomly perturb faces into incorrect categories. For the second type, we randomly replace faces in IMDb-Face with images from MegaFace.

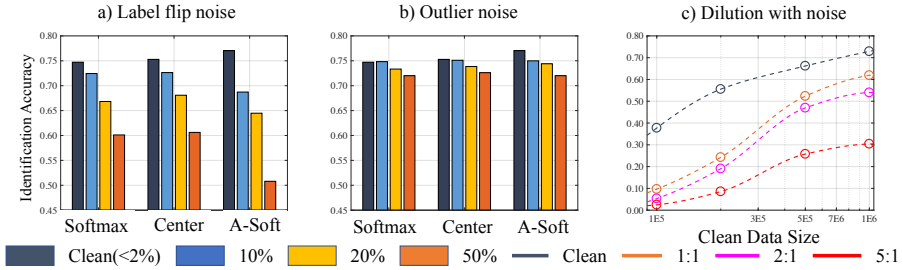

Fig. 7. 1:1M rank-1 identification results on MegaFace benchmark: (a) introducing label flips to IMDb-Face, (b) introducing outliers to IMDb-Face, and (c) fixing the size of clean data and dilute it with different ratios of label flips.

Here we perform two experiments: 1) We gradually contaminate our dataset with different types of noise. We gradually increase the noise in our dataset by 10%, 20% and 50%. 2) We fix the size of clean data and 'dilute' it with label flips. We do not use ensemble models in these experiments.

Figure 7(a) and (b) summarize the results of our first experiment. 1) Label flips severely deteriorate the performance of a model, more so than outliers. 2)

Table 2. Noisy data vs. Clean data. The results are obtained from rank-1 identification test on the MegaFace benchmark [8]. Abbreviation MSV1 = MS-Celeb-1M(v1).

Dataset	#Iden.	#Imgs.	MegaFace Rank-1(%)		
			Softmax	Center	A-softmax
MSV1-raw	96k	8.6M	71.70	73.82	73.99
-sampled	46k	3.7M	66.15	69.81	70.56
-clean	46k	1.76M	70.66	73.15	73.53
MegaFace-raw	670k	4.7M	64.32	64.71	66.95
-sampled	270k	2.7M	59.68	62.55	63.12
-clean	270k	1.5M	62.86	67.64	68.88

A-Softmax, which used to achieve a better result on a clean dataset, becomes worse than Center loss and Softmax in the high-noise region. 3) Outliers seem to have a less abrupt effect on the performance across all losses, matching the observation in [10] and [17].

The second experiment was inspired by a recent work from Rolnick *et al.* [17]. They found that if a dataset contains sufficient clean data, a deep learning model can still be properly trained on it when the data is diluted by a large amount of noise. They show that a model can still achieve a feasible accuracy on CIFAR-10, even the ratio of noise to clean data is increased to $20:1$. Can we transfer their conclusion to face recognition? Here we sample four subsets from IMDb-Face with $1E5$, $2E5$, $5E5$ and $1E6$ images. And we dilute them with an equal number, double, and five times of label flip noise. Figure 7(c) shows that a large performance gap still exists against the completely clean baseline, even we maintain the same number of clean data. We conjecture two reasons that cleanliness of data still plays a key role in face recognition: 1) current dataset, even it is clean, still far from sufficient to address the challenging face recognition problem thus noise matters. 2) Noise is more lethal on a 10,000-class problem than on a 10-class problem.

4.2 The Effect of Noise on MegaFace and MS-Celeb-1M

To further demonstrate the effect of noise, we perform experiments on two public datasets: MegaFace and MS-Celeb-1M. In order to quantify the effect of noise on the face recognition, we sampled subsets from the two datasets and manually cleaned them. This provides us with a noisy sampled subset and a clean subset for each dataset. For a fair comparison, the noisy subset was sampled to have the same distribution of image numbers to identities as the original dataset. Also, we control the scale of noisy subsets to make sure the scales for each clean subset are nearly the same. Because of the large size of the sampled subsets, we have chosen the third labeling method mentioned in Sec. 3.3, which is the fastest.

Three different losses, namely, SoftMax, Center Loss and A-Softmax, are respectively applied to the original datasets, sampled, and cleaned subsets. Table 2 summarizes the results on the MegaFace recognition challenge [8]. The effect of

clean datasets is tremendous. By comparing the results between cleaned datasets and sampled datasets, the average improvement of accuracy is as large as 4.14%. The accuracies on clean subsets even surpass those on raw datasets, which are 4 times larger on average. The results suggest the effectiveness of reducing noise for large-scale datasets. As the mater of fact, the result of this experiment is part of our motivation to collect IMDb-Face dataset.

It is worth pointing out that recent metric learning based methods such as A-Softmax [12] and Center-loss [23] also benefit from learning on clean datasets, although they already perform much better than Softmax [20]. As shown in Table 2, the improvements of accuracy on MegaFace using A-Softmax and Center-loss are over 5%. The results suggest that reducing dataset noise is still helpful, especially when metric learning is performed. Reducing noisy samples could help an algorithm focuses more on hard examples learning, rather than picking up meaningless noises.

4.3 Comparing IMDb-Face with other Face Datasets

In the third experiment, we wish to show the competitiveness of IMDb-Face against several well-established face recognition training datasets including: 1) CelebFaces [19, 20], 2) CASIA-WebFace [25], 3) MS-Celeb-1M(v1) [5], and 4) MegaFace [13]. The data size of the two latter datasets is a few times larger than the proposed IMDb-Face. Note that MS-Celeb-1M has a larger subset(v2), containing 900,000 identities. Limited by our computational resources we did not conduct experiments on it. We do not use ensemble models in this experiment. Table 3 summarizes the results of using different datasets as the training source across three losses. We observed that the proposed noise-controlled IMDb-Face dataset is competitive as a training source despite its smaller size, validating the effectiveness of the IMDb data source and the cleanliness of IMDb-Face.

4.4 Comparisons with State-of-the-Arts

We are interested to compare the performance of model trained on IMDb-Face with state-of-the-arts. Evaluation is conducted on MegaFace [8], LFW [7], and

Table 3. Comparative results on using different face recognition datasets for training. Rank-1 identification accuracy on MegaFace benchmark is reported.

Dataset	#Iden.	#Imgs.	Rank-1 (%)		
			Softmax	Center Loss	A-Softmax
CelebFaces	10k	0.20M	36.15	42.54	43.72
CASIA-WebFace	10.5k	0.49M	65.17	68.09	70.89
MS-Celeb-1M(V1)	96k	8.6M	71.70	73.82	73.99
MegaFace	670k	4.7M	64.32	64.71	66.95
IMDbFace	59k	1.7M	**74.75**	**79.41**	**84.06**

Table 4. Comparisons with state-of-the-arts methods on LFW, MegaFace and YTF benchmarks.

Method, Dataset	LFW	Mega(Ident.)	YTF
Vocord-deep V3†, Private	-	**91.76**	-
YouTu Lab†, Private	-	83.29	-
DeepSense V2†, Private	-	81.23	-
Marginal Loss♯ [4] MS-Celeb-1M	99.48	80.278	95.98
SphereFace [12],CASIA-WebFace	99.42	75.77	95.00
Center Loss [23],CASIA-WebFace	99.28	65.24	94.90
A-Softmax♯, MS-Celeb-1M	99.58	73.99	97.45
A-Softmax♯, IMDb-Face	**99.79**	**84.06**	**97.67**

† Commercial, have not been published
♯ Single Model

YTF [24] following the standard protocol. For LFW [7] we compute equals error rate (EER). For YTF [24] we report accuracy for recognition. To highlight the effect of training data, we do not adopt model ensemble. The comparative results are shown in Table 4. Our single model trained on IMDb-Face (A-Softmax♯, IMDb-Face) achieves a state-of-the-art performance on LFW, MegaFace, and YTF against published methods. It is noteworthy that the performance of our final model is also comparable to a few private methods on MegaFace.

5 Conclusion

Beyond existing efforts of developing sophisticated losses and CNN architectures, our study has investigated the problem of face recognition from the data perspective. Specifically, we developed an understanding of the source of label noise and its consequences. We also collected a new large-scale data from IMDb website, which is naturally a cleaner and wilder source than search engines. Through user studies, we have discovered an effective yet accurate way to clean our data. Extensive experiments have demonstrated that both data source and cleaning effectively improve the accuracy of face recognition. As a result of our study, we have presented a noise-controlled IMDb-Face dataset, and a state-of-the-art model trained on it. A clean dataset is important as the face recognition community has been looking for large-scale clean datasets for two practical reasons: 1) to better study the training performance of contemporary deep networks as a function of noise level in data. Without a clean dataset, one cannot induce controllable noise to support a systematic study. 2) to benchmark large-scale automatic data cleaning methods. Although one can use the final performance of a deep network as a yardstick, this measure can be affected by many uncontrollable factors, *e.g.*, network hyperparameters setting. A clean and large-scale dataset enables unbiased analysis.

References

1. Cao, K., Rong, Y., Li, C., Tang, X., Loy, C.C.: Pose-robust face recognition via deep residual equivariant mapping. In: CVPR (2018)
2. Chen, X., Shrivastava, A., Gupta, A.: Enriching visual knowledge bases via object discovery and segmentation. In: CVPR (2014)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
4. Deng, J., Zhou, Y., Zafeiriou, S.: Marginal loss for deep face recognition. In: CVPRW (2017)
5. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In: ECCV (2016)
6. Huang, C., Li, Y., Loy, C.C., Tang, X.: Deep imbalanced learning for face recognition and attribute prediction. arXiv preprint arXiv:1806.00194 (2018)
7. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst (2007)
8. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: CVPR (2016)
9. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: CVPR (2015)
10. Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., Philbin, J., Fei-Fei, L.: The unreasonable effectiveness of noisy data for fine-grained recognition. In: ECCV (2016)
11. Li, J., Zhao, J., Zhao, F., Liu, H., Li, J., Shen, S., Feng, J., Sim, T.: Robust face recognition with deep multi-view representation learning. In: ACMMM (2016)
12. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: CVPR (2017)
13. Nech, A., Kemelmacher-Shlizerman, I.: Level playing field for million scale face recognition. In: CVPR (2017)
14. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: ICIP (2014)
15. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC (2015)
16. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The feret database and evaluation procedure for face-recognition algorithms. Image and vision computing (1998)
17. Rolnick, D., Veit, A., Belongie, S., Shavit, N.: Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694 (2017)
18. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)
19. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: NIPS (2014)
20. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: CVPR (2014)
21. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR (2014)
22. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR (2017)

23. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV (2016)
24. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: CVPR (2011)
25. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
26. Zhan, X., Liu, Z., Yan, J., Lin, D., Loy, C.C.: Consensus-driven propagation in massive unlabeled data for face recognition. In: ECCV (2018)