Improving Face Recognition by Clustering Unlabeled Faces in the Wild

Aruni RoyChowdhury[1⋆], Xiang Yu[2], Kihyuk Sohn[2⋆⋆], Erik Learned-Miller[1], and Manmohan Chandraker[2,3]

[1] University of Massachusetts, Amherst
[2] NEC Labs America
[3] University of California, San Diego

Abstract. While deep face recognition has benefited significantly from large-scale labeled data, current research is focused on leveraging unlabeled data to further boost performance, reducing the cost of human annotation. Prior work has mostly been in controlled settings, where the labeled and unlabeled data sets have no overlapping identities by construction. This is not realistic in large-scale face recognition, where one must contend with such overlaps, the frequency of which increases with the volume of data. Ignoring identity overlap leads to significant labeling noise, as data from the same identity is split into multiple clusters. To address this, we propose a novel identity separation method based on extreme value theory. It is formulated as an out-of-distribution detection algorithm, and greatly reduces the problems caused by overlapping-identity label noise. Considering cluster assignments as pseudo-labels, we must also overcome the labeling noise from clustering errors. We propose a modulation of the cosine loss, where the modulation weights correspond to an estimate of clustering uncertainty. Extensive experiments on both controlled and real settings demonstrate our method's consistent improvements over supervised baselines, e.g., 11.6% improvement on IJB-A verification.

1 Introduction

Deep face recognition has achieved impressive performance, benefiting from large-scale labeled data. Examples include DeepFace [38], which uses 4M labeled faces for training and FaceNet [32], which is trained on 200M labeled faces. Further improvements in recognition performance using traditional supervised learning may require tremendous annotation efforts to increase the labeled dataset volume, which is impractical, labor intensive and does not scale well. Therefore, exploiting unlabeled data to augment the labeled data, i.e., *semi-supervised learning*, is an attractive alternative. Preliminary work on generating pseudo-labels by clustering unlabeled faces has been shown to be effective in improving performance under controlled settings [37, 42, 44].

⋆ Now at Amazon, work done prior to joining, while interning at NEC Labs America.
⋆⋆ Now at Google, work done prior to joining.

Fig. 1: Given a face recognition model trained on labeled faces ($\mathcal{X}^{\mathcal{L}}$), we wish to cluster unlabeled data ($\mathcal{X}^{\mathcal{U}}$) for additional training samples to further improve recognition performance. Key challenges include ***overlapping identities*** between labeled and unlabeled data (*George W Bush* images present in both $\mathcal{X}^{\mathcal{L}}$ and $\mathcal{X}^{\mathcal{U}}$) as well as noisy training labels arising from ***incorrect cluster assignments*** (a picture of *George Bush Sr.* is erroneously assigned to a cluster of *George W Bush* images).

However, although learning from unlabeled data is a mature area and theoretically attractive, face recognition as a field has yet to adopt such methods in practical and realistic settings. There are several obstacles to directly applying such techniques that are peculiar to the setting of large-scale face recognition. First, there is a common assumption of semi-supervised face recognition methods is that there is *no class or identity overlap* between the unlabeled and the labeled data. This seemingly mild assumption, however, violates the basic premise of semi-supervised learning – that nothing is known about the labels of the unlabeled set. Thus, either practitioners must manually verify this property, meaning that the data is no longer *"truly" unlabeled*, or proceed under the assumption that identities are disjoint between labeled and unlabeled training datasets, which inevitably introduces labeling noise. When such overlapping identities are in fact present (Fig. 1), a significant price is paid in terms of performance, as demonstrated empirically in this work. A further practical concern in face recognition is the availability of massive labeled face datasets. Most current work using unlabeled faces focus on improving the performance of models trained with *limited labeled data* [42,44], and it is unclear if there are any benefits from using unlabeled datasets when baseline face recognition models are trained on *large-scale labeled data*.

In this paper, we present recipes for exploiting unlabeled data to further improve the performance of fully supervised state-of-the-art face recognition models, which are mostly trained on large-scale labeled datasets. We demonstrate that learning from unlabeled faces is indeed a practical avenue for improving deep face recognition, also addressing important practical challenges in the process – accounting for overlapping identities between labeled and unlabeled data, and ameliorating the effect of noisy labels when training on pseudo-labeled data.

We begin with Face-GCN [42], a graph convolutional neural network (GCN) based face clustering method, to obtain pseudo-labels on unlabeled faces. To deal with the overlapping identity problem, we observe that the distribution of classification confidence on overlapping and disjoint identities is different – since our initial face feature is provided by a recognition engine trained on known identities, the confidence score of the overlapping identity images should be higher than those of non-overlapping identity images, as visualized in Fig. 3. Based on this observation, we approach the problem as *out-of-distribution detection* [9, 18, 20], and propose to parameterize the distribution of confidence scores as a mixture of Weibulls, motivated by extreme value theory. This results in an unsupervised procedure to separate overlapping identity samples from unlabeled data on-the-fly.

After resolving the overlapping identity caused label noise, the *systematic label noise* from the clustering algorithm remains, which is another prime cause for deteriorating performance in face recognition [40]. Instead of an additional complicated pruning step to discard noisy samples, *e.g.* as done in [42], we deal with the label noise during the re-training loop using the joint data of both labeled and clustered faces, by introducing a simple clustering uncertainty based modulation of the training loss to reduce the effect of erroneous gradients caused by the noisy labeled data. This effectively smoothes the re-training procedure and has shown clear performance gains in our experiments. Our contributions are summarized as the following:

– To our knowledge, we are the first to tackle the practical issue of overlapping identities between labeled and unlabeled face data during clustering, formulated as an out-of-distribution detection.
– We successfully demonstrate that jointly leveraging large scale unlabeled data along with labeled data in a semi-supervised fashion can indeed significantly improve over supervised face recognition performance, i.e., substantial gains over a supervised CosFace [41] model across multiple public benchmarks.
– We introduce a simple and scalable uncertainty-modulated training loss into the semi-supervised learning setup, which is designed to compensate for the label noise introduced by the clustering procedure on unlabeled data.
– We provide extensive and ablative insights on both controlled and real-world settings, serving as a recipe for the semi-supervised face recognition or other large scale recognition problems.

2 Related Work

Face Clustering: Jain [12] provides a survey on classic clustering techniques. Most recent approaches [14, 21, 22, 26, 36] work on face features extracted from supervisedly-trained recognition engines. "Consensus-driven propagation" (CDP) [44] assigns pseudo-labels to unlabeled faces by forming a graph over the unlabeled samples. An ensemble of various network architectures provides multiple views of the unlabeled data, and an aggregation module decides on positive and negative pairs. Face-GCN [42] formulates the face clustering problem into a regression for cluster proposal purity, which can be fully supervised. Re-training the recognition engine with the clustered "pseudo-identities" and the original data improves performance, however, CDP [44] and Face-GCN re-training assumes the "pseudo-identities" and the original identities have no overlap, which does not always hold true. Meanwhile, their investigation stays in a controlled within-distribution setting using the MS-Celeb-1M dataset [7], which is far from realistic. In contrast,

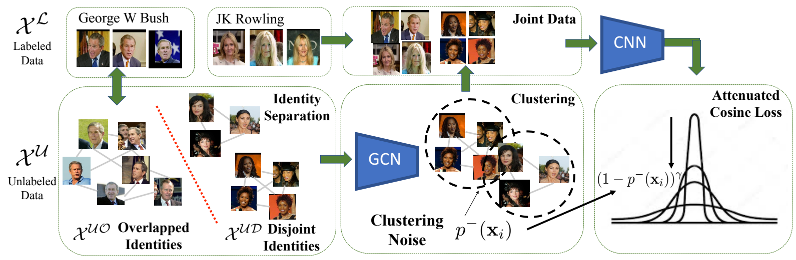


Fig. 2: Our approach trains a deep neural network [41] jointly on labeled faces $\mathcal{X}^{\mathcal{L}}$ and unlabeled faces $\mathcal{X}^{\mathcal{U}}$. Unlabeled samples with ***overlapping*** and ***disjoint*** identities w.r.t. $\mathcal{X}^{\mathcal{L}}$ are separated into $\mathcal{X}^{\mathcal{UO}}$ and $\mathcal{X}^{\mathcal{UD}}$, respectively (Sec. 3.1). The unlabeled faces in $\mathcal{X}^{\mathcal{UD}}$ are clustered using a ***graph conv-net*** or ***GCN*** (Sec. 3.2). Estimates of ***cluster uncertainty*** $p^-(\mathbf{x}_i)$ are used to modulate the cosine loss during re-training (Sec. 3.3).

we demonstrate that these considerations are crucial to achieve gains for practical face recognition with truly large-scale labeled and unlabeled datasets.

Out-Of-Distribution Detection: Extreme value distributions have been used in calibrating classification scores [29], classifier meta-analysis [31], open set recognition robust to adversarial images [2] and as normalization for score fusion from multiple biometric models [30], which is quite different from our problem. Recent approaches to out-of-distribution detection utilize the confidence of the predicted posteriors [9, 20], while Lee *et al.* [18] use Mahalanobis distance-based classification along with gradient-based input perturbations. [18] outperforms the others, but does not scale to our setting – estimating per-subject covariance matrices is not feasible for the typical long-tailed class distribution in face recognition datasets.

Learning with Label Noise: Label-noise [24] has a significant effect on the performance of the face embeddings obtained from face recognition models trained on large datasets, as extensively studied in Wang *et al.* [40]. Indeed, even large scale human-annotated face datasets such as the well-known MS 1 Million (MS-1M) are shown to have some incorrect labeling, and gains in recognition performance can be attained by cleaning up the labeling [40]. Applying label-noise modeling to our problem of large-scale face recognition has its challenges – the labeled and unlabeled datasets are class-disjoint, a situation not considered by earlier methods [10, 19, 27]; having ∼100k identities, typically long-tailed, make learning a label-transition matrix challenging [10, 27]; label-noise from clustering pseudo-labels is typically structured and quickly memorized by a deep network, unlike the uniform-noise experiments in [1, 39, 45]. Our unsupervised label-noise estimation does not require a clean labeled dataset to learn a training curriculum unlike [13, 28], and can thus be applied out-of-the-box.

3 Learning from Unlabeled Faces

Formally, let us consider samples $\mathcal{X} = \{\mathbf{x}_i\}_{i\in[n]}$, divided into two parts: $\mathcal{X}^{\mathcal{L}}$ and $\mathcal{X}^{\mathcal{U}}$ of sizes l and u respectively. Now $\mathcal{X}^{\mathcal{L}} := \{\mathbf{x}_1, ..., \mathbf{x}_l\}$ consist of faces that are provided with identity labels $\mathcal{Y}^{\mathcal{L}} := \{y_1, ..., y_l\}$, while we do not know the identities of the unlabeled faces $\mathcal{X}^{\mathcal{U}} := \{\mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$. Our approach aims to improve the performance of a supervised face recognition model, trained on ($\mathcal{X}^{\mathcal{L}}$, $\mathcal{Y}^{\mathcal{L}}$), by first clustering the unlabeled faces $\mathcal{X}^{\mathcal{U}}$, then re-training on both labeled and unlabeled faces, using the cluster assignments on $\mathcal{X}^{\mathcal{U}}$ as pseudo-labels. Fig. 2 visually summarizes the steps – **(1)** train a supervised face recognition model on ($\mathcal{X}^{\mathcal{L}}$, $\mathcal{Y}^{\mathcal{L}}$); **(2)** separate the samples in $\mathcal{X}^{\mathcal{U}}$ having overlapping identities with the labeled training set; **(3)** cluster the disjoint-identity unlabeled faces; **(4)** learn an unsupervised model for the likelihood of incorrect cluster assignments on the pseudo-labeled data; **(5)** re-train the face recognition model on labeled and pseudo-labeled faces, modulating the training loss for pseudo-labeled samples using the estimated clustering uncertainty. In this section, we first describe the separation of overlapping identity samples from unlabeled data (Sec. 3.1), followed by an overview of the face clustering procedure (Sec. 3.2) and finally re-training the recognition model with an estimate of clustering uncertainty (Sec. 3.3).

3.1 Separating Overlapping Identities

Overlapping identities. We typically have no control over the gathering of the unlabeled data $\mathcal{X}^{\mathcal{U}}$, so the same subject S may exist in labeled data (thus, be a class on which the baseline face recognition engine is trained) and also within our unlabeled dataset, *i.e.* $\mathcal{X}^{\mathcal{U}} = \mathcal{X}^{\mathcal{UO}} \cup \mathcal{X}^{\mathcal{UD}}$, where $\mathcal{X}^{\mathcal{UO}}$ and $\mathcal{X}^{\mathcal{UD}}$ denote the identity overlapped and identity disjoint subsets of $\mathcal{X}^{\mathcal{U}}$. By default, the clustering will assign images of subject S in the unlabeled data as a new category. In this case, upon re-training with the additional pseudo-labeled data, the network will incorrectly learn to classify images of subject S into *two* categories. This is an important issue, since overlapping subjects can occur naturally in datasets collected from the Internet or recorded through passively mounted cameras, which to our knowledge has not been directly addressed by most recent pseudo-labeling methods [37, 42, 44].

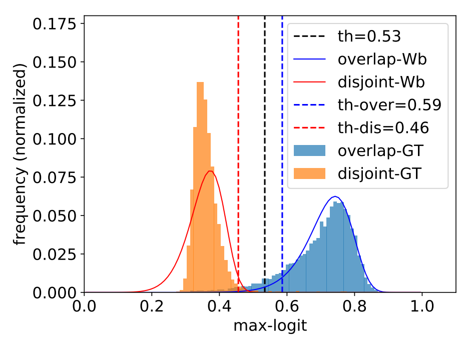


Fig. 3: Extreme value theory (EVT) provides a principled way of setting thresholds on the max-logits z_i for $\mathbf{x}_i \in \mathcal{X}^{\mathcal{U}}$ to separate *disjoint* and *overlapping* identities (red and blue vertical lines). An initial threshold is determined by Otsu's method (black vertical line). This plot uses splits from the MS-Celeb-1M dataset [7].

Out-of-distribution detection. The problem of separating unlabeled data into samples of disjoint and overlapping classes (w.r.t. the classes in the labeled data) can be regarded as an "out-of-distribution" detection problem. The intuition is that unlabeled samples with overlapping identities will have higher confidence scores from a face recognition engine, as the same labeled data is used to train the recognition model [9]. Therefore, we search for thresholds on the recognition confidence scores that can separate disjoint and overlapping identity samples. Note, since the softmax operation over several thousand categories can result in small values due to normalization, we use the *maximum logit* $\boldsymbol{z}_i$ for each sample $\mathbf{x}_i \in \mathcal{X}^{\mathcal{U}}$ as its confidence score. Since the $\boldsymbol{z}_i$ are the maxima over a large number of classes, we can draw upon results from extreme value theory (EVT) which state that the limiting distribution of the maxima of i.i.d random variables belongs to either the Gumbel, Fréchet or Weibull family [4]. Specifically, we model the $\boldsymbol{z}_i$ using the Weibull distribution,

$$f(\boldsymbol{z}_i; \lambda, k) = \begin{cases} \frac{k}{\lambda}\left(\frac{\boldsymbol{z}_i}{\lambda}\right)^{k-1} e^{-(\boldsymbol{z}_i/\lambda)^k} & \boldsymbol{z}_i \geq 0, \\ 0 & \boldsymbol{z}_i < 0, \end{cases} \tag{1}$$

where $k > 0$ and $\lambda > 0$ denote the shape and scale parameters, respectively. We use Otsu's method [25] to obtain an initial threshold on the $\boldsymbol{z}_i$ values, then fit a two-component mixture of Weibulls, modeling the *identity-overlapping* and *identity-disjoint* sets $\mathcal{X}^{\mathcal{UO}}$ and $\mathcal{X}^{\mathcal{UD}}$, respectively. Selecting values corresponding to 95% confidence under each Weibull model provides thresholds for deciding if $\mathbf{x}_i \in \mathcal{X}^{\mathcal{UO}}$ or $\mathbf{x}_i \in \mathcal{X}^{\mathcal{UD}}$ with high confidence; we reject samples that fall outside of this interval. This approach does not require setting any hyper-parameters a priori, and can be applied to any new unlabeled dataset.

3.2 Clustering Faces with GCN

We use Face-GCN [42] to assign pseudo-labels for unlabeled faces in $\mathcal{X}^{\mathcal{UD}}$, which leverages a graph convolutional network (GCN) [15] for large-scale face clustering. We provide a brief overview of the approach for completeness. Based on features extracted from a pre-trained face recognition engine, a nearest-neighbor graph is constructed over all samples. By setting various thresholds on the edge weights of this graph, a set of connected components or cluster proposals are generated. During training, the aim is to regress the precision and recall of the cluster proposals arising from a single ground truth identity, motivated by object detection frameworks [8]. Since the proposals are generated based on labeled data, the Face-GCN is trained in a fully supervised way, unlike regular GCN training, which are typically trained with a classification loss, either for each node or an input graph as a whole. During testing, a "de-overlap" procedure uses predicted GCN scores for the proposals to partition an unlabeled dataset into a set of clusters. Please see [42] for further details.

3.3 Joint Data Re-training with Clustering Uncertainty

We seek to incorporate the uncertainty of whether a pseudo-labeled (*i.e.* clustered) sample was correctly labeled into the face recognition model re-training. Let a face drawn from the unlabeled dataset be $\mathbf{x}_i \in \mathcal{X}^{\mathcal{UD}}$. The feature representation for that face using the baseline supervised model is denoted as $\Phi(\mathbf{x}_i)$. Let cluster assignments obtained on $\mathcal{X}^{\mathcal{UD}}$ be $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K\}$, for K clusters. We train a logistic regression classifier to estimate $P(\mathcal{C}_k \mid \Phi(\mathbf{x}_i))$, for $k = 1, 2, ...K$,

$$P(\mathcal{C}_k \mid \Phi(\mathbf{x}_i)) = \frac{\exp(\omega_k^\top \Phi(\mathbf{x}_i))}{\sum_j \exp(\omega_j^\top \Phi(\mathbf{x}_i))} \tag{2}$$

where ω_k are the classifier weights for the k-th cluster. Intuitively, we wish to determine how well a simple linear classifier on top of discriminative face descriptors can fit the cluster assignments. We compare the following uncertainty metrics: **(1)** *Entropy* of the posteriors across the K clusters, *i.e.* $\sum_k P(\mathcal{C}_k \mid \Phi(\mathbf{x}_i)) \log P(\mathcal{C}_k \mid \Phi(\mathbf{x}_i))$; **(2)** *Max-logit:* the largest logit value over the K clusters, **(3)** *Classification margin:* difference between the max and the second-max logit, indicating how easily a sample can flip between two clusters.

We consider two kinds of incorrect clustering corresponding to notions of precision and recall: **(1) Outliers**, samples whose identity does not belong to the identity of the cluster; **(2) Split-ID**, where samples from the same identity are spread over several clusters (Fig. 4(a)). In a controlled setting with known ground-truth identities, we validate our hypothesis that the uncertainty measures can distinguish between correct and incorrect cluster assignments (Fig. 4(b)). Note that Split-ID makes up the bulk of incorrectly-clustered samples, while outliers are about 10%. Fig. 4 (c) shows the distribution of class-margin on pseudo-labeled data on one split of the MS-1M dataset. Intuitively, samples that do not have a large classification margin are likely to be incorrect pseudo-labels, resulting in a bi-modal distribution – noisily labeled samples in one mode, and correctly labeled samples in the other. Notice that similar to overlapping v.s. disjoint identity, this is another distribution separation problem. A Weibull is fit to the lower portion of the distribution (orange curve), with an initial mode-separating threshold obtained from Otsu's method (black vertical line). The probability of sample $\mathbf{x}_i$ being incorrectly clustered is estimated by:

$$p^-(\mathbf{x}_i) = P(g(\mathbf{x}_i) \mid \theta^-_{Wb}), \tag{3}$$

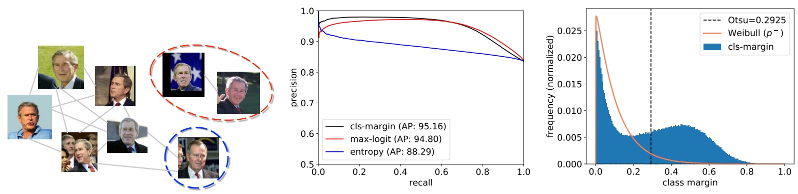
where θ^-_{Wb} are the parameters of the learned Weibull model, $g(.)$ denotes the measure of uncertainty, *e.g.*class-margin. Note, ground-truth labels are not required for this estimation. We propose to associate the above uncertainty with the pseudo-labeled samples and set up a probabilistic face recognition loss.

The large margin cosine loss [41] is used for training:

$$\mathcal{L}(\mathbf{x}_i) = -\log \frac{\exp(\alpha(\mathbf{w}_j^\top \mathbf{f}_i - m))}{\exp(\alpha(\mathbf{w}_j^\top \mathbf{f}_i - m)) + \sum_{k\neq j} \exp(\alpha \mathbf{w}_k^\top \mathbf{f}_i)} \tag{4}$$

where $\mathbf{f}_i$ is the deep feature representation of the i-th training sample $\mathbf{x}_i$, $\mathbf{w}_j$ is the learned classifier weight for the j-th class, $m \in [0, 1]$ is an additive margin and α is a scaling factor; $\|\mathbf{f}_i\|$ and $\|\mathbf{w}_j\|$ are set to 1. For $\mathbf{x}_i \in \mathcal{X}^{\mathcal{U}}$, we modulate the training loss with the clustering uncertainty $p^-(\mathbf{x}_i)$, where γ controls the weighting curve shape:

$$\mathcal{L}^p(\mathbf{x}_i) = (1 - p^-(\mathbf{x}_i))^\gamma \mathcal{L}(\mathbf{x}_i), \tag{5}$$



(a) (b) (c)

Fig. 4: **Clustering uncertainty. (a)** Illustration of incorrect pseudo-labels – an image of *George Bush Sr.* is included in a cluster of *George W Bush* images (outlier circled in blue); some *George W Bush* images are spread across multiple clusters ("split ID" circled in red). **(b)** Precision-recall curves showing Average Precision (AP) of predicting if a cluster assignment is correct using class-margin, max-logit and entropy. **(c)** Estimating clustering error $p^-(\mathbf{x}_i)$ from the distribution of class-margin (orange curve).

4 Experiments

We augment supervised models trained on labeled data with additional pseudo-labeled data under various scenarios. We summarize the main findings first — ***(i)*** the baseline supervised model benefits from additional pseudo-labeled training data; ***(ii)*** re-training on clustering without handling overlapping IDs can hurt performance, and our approach of separating overlaps is shown to be effective empirically; ***(iii)*** increasing diversity of training data by using unlabeled data from outside the distribution of the labeled set helps more than comparable amounts of within-domain unlabeled data; ***(iv)*** scaling up to using the entire MS-Celeb-1M [7] dataset (or MS1M for short) as labeled training set, as typically done by most deep face models, we see significant gains in performance *only* when the volume of unlabeled samples is comparable to the size of MS1M itself.

Experimental setup. Table 1 summarizes the training data sources. The cleaned version of MS1M dataset contains 84,247 identities and 4,758,734 samples in total. Partitioning on the identities, the full MS1M dataset is split into 10 parts with approximately 8.4k identities and 470k samples per split. We create the following settings:

- ***Controlled disjoint*** *(Sec. 4.1):* Both labeled and unlabeled data are drawn from splits of MS1M (Table 1 *MS1M* splits 1 and 2, respectively). Thus, they have the same distribution and have no overlapping identities by construction, similar to the setting in [42]. We compare baseline clustering methods and the effect of clustering uncertainty on re-training the face recognition model.
- ***Controlled overlap*** *(Sec. 4.2):* we introduce simulated identity overlap between the two datasets (Table 1 *MS-Celeb-1M* splits *1-O* and *2-O*), showing the detrimental effect of naïvely clustering and re-training in this case, and the efficacy of our proposed approach.
- ***Semi-controlled*** *(Sec. 4.3):* we have limited labeled data (split-1 of MS1M) with unlabeled data from another dataset, i.e., VGGFace2 [3], containing 8.6k identities and 3.1 million images. This is closer to the realistic scenario, with potential identity overlaps and distribution shift between data sources.

Table 1: Statistics for training datasets.

Dataset	#IDs	Images
MS-Celeb-full	84k	4.7M
MS-Celeb-split-1	8.4k	505k
MS-Celeb-split-2	8.4k	467k
MS-Celeb-split-1-O	16.8k	729k
MS-Celeb-split-2-O	16.8k	705k
VGGFace2 [3]	8.6k	3.1M
CASIA-WebFace [43]	10.5k	455k
IMDB-SenseTime [40]	51k	1M
GlintAsian [5]	94k	2.8M

Table 2: **Controlled: Face clustering baselines.** Comparing the GCN-based method with standard clustering algorithms. The GCN is trained on MS-Celeb-1M split 1, tested on split 2.

Method	Prec	Rec	F1	#Clstr
K-means	55.77	87.56	68.14	5k
FastHAC	99.32	64.66	78.32	117k
DBSCAN	99.62	46.83	63.71	352k
GCN	95.87	79.43	86.88	45k
GCN-iter2	97.94	87.28	92.30	32k

– ***Uncontrolled*** *(Sec. 4.5):* close to the real-world setting, we use *all* the labeled data at our disposal (entire MS-Celeb-1M) and try to improve performance further by including unlabeled data from other datasets – VGGFace2 [3], IMDB-SenseTime [40], CASIA [43] & GlintAsian [5], by completely ignoring their ground truth labels. Note, this setting is not addressed in prior art on pseudo-labeling faces [42, 44].

Evaluation. We report results on the following: verification accuracy on *Labeled Faces in the Wild* (LFW) [11, 17] and *Celebrity Frontal to Profile* (CFP) [35]; identification at rank-1 and rank-5, and True Accept Rate (TAR) at False Accept Rates (FAR) of 1e-3 and 1e-4 on the challenging *IARPA Janus* benchmark (IJB-A) [16]. For clustering metrics we adopt the protocol used in [42].

Training details. We use the high-performing CosFace model [41] for face recognition, with a 118-layer ResNet backbone, trained for 30 epochs on labeled data. Re-training is done from scratch with identical settings. Face-GCN uses the publicly available code of GCN-D [42]. For further details please refer to the supplementary materials.

4.1 Controlled Disjoint: MS-Celeb-1M splits

In controlled setting, Split-1 of MS-Celeb-1M is used as the labeled dataset to train the face recognition model in a fully supervised fashion. The face clustering module is also trained in a supervised way on the labeled Split-1 data. The unlabeled data is from Split-2 of MS-Celeb-1M: ground truth labels are ignored, features are extracted on all the samples and the trained GCN model provides the cluster assignments.

Clustering. The performance of various clustering methods are summarized in Table 2, i.e., ***K-means*** [34], ***FastHAC*** [23] and ***DBSCAN*** [6, 33], with optimal hyper-parameter settings [4]. The GCN is clearly better than the baseline clustering approaches. GCN typically provides an over-clustering of the actual number of identities – the precision is comparably higher than the recall (95.87%

[4] K-means: K=5k, FastHAC: dist=0.85, DBSCAN: minsize=2, eps=0.8

Table 3: **Controlled disjoint**: Re-training CosFace on the union of labeled and pseudo-labeled data (*+GCN*), pseudo-label on second iteration (*+GCN-iter-2*), with an upper bound from ground truth (*GT-2*). ↑ indicates improvement from baseline.

Model	**LFW**	↑	**CFP-fp**	↑	**IJBA-idt.** Rank-1, 5	↑	**IJBA-vrf.** FAR@1e-3,-4	↑
Baseline GT-1	99.20	-	92.37	-	92.66, 96.42	-	80.23, 69.64	-
+K-means	99.47	0.27	94.11	1.74	93.80, 96.79	1.14, 0.37	87.03, 78.00	6.80, 8.36
+FastHAC	99.42	0.22	93.56	0.90	93.84, **96.81**	1.18, **0.39**	84.78, 75.21	4.55, 5.57
+GCN	99.48	0.28	**95.51**	**3.14**	94.11, 96.55	1.45, 0.13	87.60, 77.67	7.37, 7.93
+GCN-*iter-2*	**99.57**	**0.37**	94.14	1.77	**94.46**, 96.40	**1.80**, -0.02	**88.00**, **78.78**	**7.77**, **9.14**
+GT-2 (bound)	99.58	0.38	95.56	3.19	95.24, 97.24	2.58, 0.82	89.45, 81.02	9.22, 11.38

Table 4: **Controlled overlaps:** Re-training with overlapping identity unlabeled data.

Model	**LFW**	↑	**CFP-fp**	↑	**IJBA-idt.** Rank-1, 5	↑	**IJBA-vrf.** FAR@1e-3,-4	↑
Baseline	99.45	-	95.17	-	94.52, 96.60	-	87.36, 75.06	-
+GCN(naive)	99.37	-0.08	93.17	-2.0	93.72, 96.65	-0.80, 0.05	87.02, 79.39	-0.34, 4.33
+GCN(disjoint)	99.57	0.12	95.01	-0.16	**94.83**, 96.98	**0.31**, 0.38	89.29, 82.64	1.93, 7.58
+GCN(overlap)	**99.58**	**0.13**	94.30	-0.87	94.47, 96.64	-0.05, 0.04	86.93, 78.42	-0.43, 3.36
+GCN(both)	**99.58**	**0.13**	**95.36**	**0.19**	94.81, **97.05**	0.29, **0.45**	**89.43**, **82.86**	**2.07**, **7.80**

versus 79.43%), indicating high purity per cluster, but samples from the same identity end up being spread out across multiple clusters ("split ID").

Re-training. The results are summarized in Table 3. Re-training CosFace on labeled Split-1 and pseudo-labeled Split-2 data (*+GCN*) improves over training on just the labeled Split-1 (*Baseline GT-1*) across the benchmarks. The performance is upper-bounded when perfect labels are available on Split-2 (*+GT-2*). Note that re-training on cluster assignments from simpler methods like K-Means and HAC also improve over the baseline.

Re-train w/ iterative clustering. We perform a second iteration of clustering, using the re-trained CosFace model as feature extractor. The re-trained CosFace model has more discriminative features, resulting in better clustering (Table 2 *GCN-iter2* versus *GCN*). However, another round of re-training CosFace on these cluster-assignments yields smaller gains (Table 3 *+GCN-iter2* v.s. *+GCN*).

Insights. With limited labeled data, training on clustered faces significantly improves recognition performance. Simpler clustering methods like K-means are also shown to improve recognition performance – if training Face-GCN is not practical, off-the-shelf clustering algorithms can also provide pseudo-labels. A second iteration gives small gains, indicating diminishing returns.

4.2 Controlled Overlap: Overlapping Identities

We simulate the real-world overlapping-identity scenario mentioned in Sec. 3.1 to empirically observe its impact on the "pseudo-labeling by clustering" pipeline.

Table 5: **Disjoint/overlap clustering.** Results of clustering the entire unlabeled $\mathcal{X}^{\mathcal{U}}$ ("Split-2-O") and clustering the estimated ID-disjoint portion $\mathcal{X}^{\mathcal{UD}}$.

Data	Prec.	Rec.	F1	#IDs	#Clstr	#Img
$\mathcal{X}^{\mathcal{U}}$	98.7	84.8	91.2	16.8k	60k	693k
$\mathcal{X}^{\mathcal{UD}}$	98.8	85.2	91.5	11.7k	39k	464k

Table 6: **Semi-controlled: clustering.** Comparing performance on "within-domain" splits of MS-Celeb-1M vs. VGGFace2 data.

Train	**Test**	**Prec.**	**Rec.**	**F1**	**#clstr**
split-1	split-2	95.87	79.43	86.88	45k
split-1	VGG2	97.65	59.62	74.04	614k
full	VGG2	98.88	72.76	83.83	224k

We create two subsets of MS1M with around 16k identities each, having about 8.5k overlapping identities (suffix "O" for *overlaps* in Table 1). The labeled subset $\mathcal{X}^L$ contains around 720k samples (Split-1-O). The unlabeled subset, Split-2-O, contains approximately 467k *disjoint-identity* ($\mathcal{X}^{UD}$) and 224k *overlapping-identity* ($\mathcal{X}^{UO}$) samples.

Disjoint/Overlap. Modeling the disjoint/overlapping identity separation as an out-of-distribution problem is an effective approach, especially on choosing the max-logit score as the feature for OOD. A simple Otsu's threshold provides acceptably low error rates, i.e., **6.2%** false positive rate and **0.69%** false negative rate, while using 95% confidence intervals from Weibulls, we achieve much lower error rates of **2.3%** FPR and **0.50%** FNR.

Clustering. Table 5 shows the results from clustering all the unlabeled data (*Naive*) versus separating out the identity disjoint portion of the unlabeled data and then clustering (*Disjoint*). On both sets of unlabeled samples, the GCN clustering achieves high precision and fairly high recall, indicating that the clusters we use in re-training the face recognition engine are of good quality.

Re-training. The results are shown in Table 4. Naively re-training on the additional pseudo-labels clearly hurts performance (*Baseline* v.s. *GCN(naive)*). Adding pseudo-labels from the *disjoint* data improves over the baseline across the benchmarks. Merging the *overlapping* samples with their estimated identities in the labeled data is done based on the softmax outputs of the baseline model, causing improvements in some cases (*e.g.* LFW and IJBA verif.) but degrading performance in others (*e.g.* IJBA ident. and YTF). Merging overlapping identities as well as clustering disjoint identities also shows improvements over the baseline across several benchmarks.

Insights. Overlapping identities with the labeled training set clearly has a detrimental effect when retraining and must be accounted for when merging unlabeled data sources – the choice of modeling max-logit scores for this separation is shown to be simple and effective. Overall, discarding overlapping samples from re-training, and clustering *only* the disjoint samples, appears to be a better strategy. Adding pseudo-labeled data for classes that exist in the labeled set seems to have limited benefits, versus adding more identities.

4.3 Semi-controlled: Limited Labeled, Large-scale Unlabeled Data

MS-Celeb-1M Split 1 forms the labeled data, while the unlabeled data is from VGGFace2 (Table 1). We simply discard VGGFace2 samples estimated to have

Table 7: **Semi-controlled: MS-Celeb-1M split 1 and VGGFace2**. Note that similar volume of pseudo-labeled data from MS-Celeb-1M split 2 (*+MS1M-GCN-2*) gives lower benefits compared to data from VGGFace2 (*+VGG-GCN*) in challenging settings like IJB-A verification at FAR=1e-4, IJB-A identification Rank-1.

Model	**LFW**	↑	**CFP-fp**	↑	**IJBA-idt.** Rank-1, 5	↑	**IJBA-vrf.** FAR@1e-3,-4	↑
MS1M-GT-1	99.20	-	92.37	-	92.66, 96.42	-	80.23, 69.64	-
+MS1M-GCN-2	99.48	0.28	**95.51**	**3.14**	94.11, 96.55	1.45, 0.13	87.60, 77.67	7.37, 12.03
+VGG-GCN (ours)	**99.55**	**0.35**	94.60	2.23	**94.72**, **96.97**	**2.06**, **0.55**	**88.12**, **82.48**	**7.89**, **12.84**
+VGG-GT (bound)	99.70	0.50	97.81	5.44	96.93, 98.25	4.27, 1.83	93.20, 84.67	12.97, 15.03

overlapping identities with MS-Celeb-1M Split-1. Out of the total 3.1M samples, about 2.9M were estimated to be identity-disjoint with MS-Celeb-1M Split-1.

Clustering. The same GCN model trained on Split-1 of MS-Celeb-1M in Sec. 4.1 is used to obtain cluster assignments on VGGFace2. Table 6 compares the clustering on MS-Celeb-1M Split2 (controlled) v.s. the current setting. The F-score on VGGFace2 is reasonable – 74.04%, but lower than the F-score on Split-2 MS-Celeb-1M (86.88%) – we are no longer dealing with within-dataset unlabeled data.

Re-training. To keep similar volumes of labeled and pseudo-labeled data we randomly select 50 images per cluster from the largest 8.5k clusters of VGGFace2. Re-training results are in Table 7. We generally see benefits from VGGFace2 data over both *baseline* and *MS1M-split-2*: YTF: 93.82% → 94.64% → **95.14%**, IJBA idnt. rank-1: 92.66% → 94.11% → **94.72%**, IJBA verif. at FAR 1e-4: 69.635% → 77.665% → **82.484%**. When the full VGGFace2 labeled dataset is used to augment MS1M-split-1, *VGG-GT(full)*, we get the upper bound performance.

Insights. Ensuring the *diversity* of unlabeled data is important, in addition to other concerns like clustering accuracy and data volume: pseudo-labels from VGGFace2 benefit more than using more data from within MS1M.

4.4 Soft Labels for Clustering Uncertainty

Table 8 shows results of re-training the face recognition model with our proposed cluster-uncertainty weighted loss (Sec. 3.3) on the pseudo-labeled samples (*GCN-soft*). We set $\gamma = 1$ (ablation in supplemental). We empirically find that incorporating this cluster uncertainty into the training loss improves results in both controlled and large-scale settings (3 out of 4 evaluation protocols). In the controlled setting, the soft pseudo-labels, MS1M-GCN-*soft*, improves over MS1M-GCN (hard cluster assignments) on challenging IJB-A protocols (77.67% → **78.78**% @FAR 1e-4) and is slightly better on LFW. In the large-scale setting, comparing VGG-GCN and VGG-GCN-*soft*, we again see significant improvements on IJB-A (81.85% → **90.16**% @FAR 1e-4) and gains on the LFW benchmark as well. Qualitative analyses of clustering errors and uncertainty estimates $p^{-}(\mathbf{x}_i)$ are included in the supplemental.

Table 8: **Effect of Cluster Uncertainty**: Re-training CosFace with the proposed clustering uncertainty (*GCN-soft*) shows improvements in both controlled (*MS1M-GT-split1*) and large-scale settings, *MS1M-GT-full* (CosFace [41]).

Model	LFW	↑	CFP-fp	↑	IJBA-idt. Rank-1, 5	↑	IJBA-vrf. FAR@1e-3,-4	↑
MS1M-GT-***split1***	99.20	-	92.37	-	92.66, 96.42	-	80.23, 69.64	-
+MS1M-GCN (ours)	99.48	0.28	**95.51**	**3.14**	94.11, 96.55	1.45, 0.13	87.60, 77.67	7.37, 12.03
+MS1M-GCN-*soft* (ours)	**99.50**	**0.30**	94.71	2.34	**94.76**, **97.10**	**2.10**, **0.68**	**87.97**, **79.43**	**7.74**, **9.79**
MS1M-GT-***full*** (CosFace)	99.70	-	**98.10**	-	95.47, 97.04	-	92.82, 80.68	-
+VGG-GCN (ours)	99.73	0.03	97.63	-0.47	95.87, 97.45	0.40, 0.41	93.88, 81.85	1.06, 1.17
+VGG-GCN-*soft* (ours)	**99.75**	**0.05**	97.57	-0.53	**96.37**, **97.70**	**0.90**, **0.66**	**93.94**, **90.16**	**1.12**, **9.48**

Table 9: **Uncontrolled: pseudo-labels.** Showing the clusters and samples in the uncontrolled setting with full-MS1M and unlabeled data of increasingly larger volume – (1) VGG2 [3]; (2) merging CASIA [43] & IMDB-SenseTime [40] with VGG2; (3) merging GlintAsian [5] with all the above.

Dataset:	VGG2	+(CASIA, IMDB)	+Glint
True classes	8631	57,271	149,824
Clusters	224,466	452,598	719,722
Samples	1,257,667	2,133,286	3,673,517
Prec.	98.88	91.35	88.16
Rec.	72.76	77.53	66.93
F-score	83.83	83.88	76.09

4.5 Uncontrolled: Large-scale Labeled and Unlabeled Data

The earlier cases either had limited labeled data, unlabeled data from an identical distribution as the labeled data by construction, or both aspects together. Now, the *entire* MS-Celeb-1M is used as labeled training data for training the baseline CosFace model as well as the GCN. We gradually add several well-known face recognition datasets (ignoring their labels) to MS-Celeb-1M labeled samples during re-training (Table 9) [5]. Along with more data, these datasets bring in more *varied* or *diverse* samples (analysis in supplemental).

Re-training. The re-training results are shown in Table 10. As expected, we get limited benefits from adding moderate amounts of unlabeled data when the baseline model is trained on a large labeled dataset like MS-Celeb-1M. When incorporating data from only VGGFace2, there are improvements on LFW (99.7% → 99.73%), and on IJBA, ident. (95.47% → 95.87%) and verif. (80.68% → 81.85%). There are however some instances of decreased performance on the smaller scale dataset CFP-fp. When the volume of unlabeled data is of comparable magnitude (4.7M labeled versus 3.6M unlabeled) by merging all the other datasets (VGGFace2, CASIA, IMDB-SenseTime and GlintAsian), we get

[5] In particular, we estimated a 40% overlap in identities between MS-Celeb and VGG2.

Table 10: **Uncontrolled: re-training.** Merging unlabeled training samples with the entire MS-Celeb-1M labeled data consistently surpasses the fully-supervised MS1M-GT-*full* (CosFace [41]) trained on the entire labeled MS-Celeb-1M dataset.

Model	**LFW**	↑	**CFP-fp**	↑	**IJBA-idt.** Rank-1, 5	↑	**IJBA-vrf.** FAR@1e-3,-4	↑
MS1M-GT-***full*** (CosFace)	99.70	-	98.10	-	95.47, 97.04	-	92.82, 80.68	-
+VGG-GCN (ours)	**99.73**	**0.03**	97.63	-0.47	95.87, 97.45	0.40, 0.41	93.88, 81.85	1.06, 1.17
+CASIA-IMDB (ours)	**99.73**	**0.03**	97.81	-0.29	96.66, 97.89	1.19, 0.85	93.79, 89.58	0.97, 8.90
+GlintAsian (ours final)	**99.73**	**0.03**	**98.24**	**0.14**	**96.94**, **98.21**	**1.47**, **1.17**	**94.89**, **92.29**	**2.07**, **11.61**

a clear advantage on the challenging IJBA benchmarks (rank-1 identification: 95.47% → **96.94%**, verification TAR at FAR 1e-4: 80.68% → **92.29%**).

Insights. The crucial factors in improving face recognition when we have access to all available labeled data from MS1M appear to be *both* diversity and volume – it is only when we merged unlabeled data from all the other data sources, reaching comparable number of samples to MS1M, that we could improve over the performance attained from training on just the ground-truth labels of MS1M, suggesting that current high-performing face recognition models can benefit from even larger training datasets. While acquiring datasets of such scale purely through manual annotation is prohibitively expensive and labor-intensive, using pseudo-labels is shown to be a feasible alternative.

5 Conclusion

The pseudo-labeling approach described in this paper provides a recipe for improving fully supervised face recognition, i.e., CosFace, leveraging large unlabeled sources of data to augment an existing labeled dataset. The experimental results show consistent performance gains across various scenarios and provide insights into the practice of large-scale face recognition with unlabeled data – **(1)** we require comparable volumes of labeled and unlabeled data to see significant performance gains, especially when several million labeled samples are available; **(2)** overlapped identities between labeled and unlabeled data is a major concern and needs to be handled in real-world scenarios; **(3)** along with large amounts of unlabeled data, greater gains are observed if the new data shows certain domain gap w.r.t. the labeled set; **(4)** incorporating scalable measures of clustering uncertainty on the pseudo-labels is helpful in dealing with label noise. Overall, learning from unlabeled faces is shown to be an effective approach to further improve face recognition performance.

Acknowledgments. This research was partly sponsored by the AFRL and DARPA under agreement FA8750-18-2-0126. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. 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 expressed or implied, of the AFRL and DARPA or the U.S. Government.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Unsupervised label noise modeling and loss correction. In: International Conference on Machine Learning (ICML) (June 2019) 4
2. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1563–1572 (2016) 4
3. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: IEEE FG (2018) 8, 9, 13
4. De Haan, L., Ferreira, A.: Extreme value theory: an introduction. Springer Science & Business Media (2007) 6
5. DeepGlint: Glint asian. http://trillionpairs.deepglint.com/overview, accessed: 2019-11-11 9, 13
6. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp. 226–231 (1996) 9
7. 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) 3, 5, 8
8. He, K., Gkioxari, G., , P.D., Girshick, R.: Mask r-cnn. In: ICCV (2017) 6
9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017) 3, 4, 5
10. Hendrycks, D., Mazeika, M., Wilson, D., Gimpel, K.: Using trusted data to train deep networks on labels corrupted by severe noise. In: Advances in neural information processing systems. pp. 10456–10465 (2018) 4
11. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., University of Massachusetts, Amherst (2007) 9
12. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern recognition letters **31**(8), 651–666 (2010) 3
13. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. arXiv preprint arXiv:1712.05055 (2017) 4
14. Jin, S., Su, H., Stauffer, C., Learned-Miller, E.: End-to-end face detection and cast grouping in movies using erdos-renyi clustering. In: ICCV (2017) 3
15. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017) 6
16. 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: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1931–1939 (2015) 9
17. Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G.: Labeled faces in the wild: A survey. In: Advances in face detection and facial image analysis, pp. 189–248. Springer (2016) 9
18. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in Neural Information Processing Systems. pp. 7167–7177 (2018) 3, 4
19. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.J.: Learning from noisy labels with distillation. In: ICCV. pp. 1928–1936 (2017) 4

20. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: ICLR (2017) 3, 4
21. Lin, W.A., Chen, J.C., Castillo, C.D., Chellappa, R.: Deep density clustering of unconstrained faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8128–8137 (2018) 3
22. Lin, W.A., Chen, J.C., Chellappa, R.: A proximity-aware hierarchical clustering of faces. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). pp. 294–301. IEEE (2017) 3
23. Müllner, D., et al.: fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. Journal of Statistical Software **53**(9), 1–18 (2013) 9
24. Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A.: Learning with noisy labels. In: Advances in neural information processing systems. pp. 1196–1204 (2013) 4
25. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979) 6
26. Otto, C., Wang, D., Jain, A.K.: Clustering millions of faces by identity. IEEE transactions on pattern analysis and machine intelligence **40**(2), 289–303 (2017) 3
27. Patrini, G., Rozza, A., Menon, A.K., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR). pp. 2233–2241 (2017) 4
28. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. arXiv preprint arXiv:1803.09050 (2018) 4
29. Rudd, E.M., Jain, L.P., Scheirer, W.J., Boult, T.E.: The extreme value machine. IEEE transactions on pattern analysis and machine intelligence **40**(3), 762–768 (2017) 4
30. Scheirer, W., Rocha, A., Micheals, R., Boult, T.: Robust fusion: Extreme value theory for recognition score normalization. In: European Conference on Computer Vision. pp. 481–495. Springer (2010) 4
31. Scheirer, W.J., Rocha, A., Micheals, R.J., Boult, T.E.: Meta-recognition: The theory and practice of recognition score analysis. IEEE transactions on pattern analysis and machine intelligence **33**(8), 1689–1695 (2011) 4
32. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015) 1
33. Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X.: Dbscan revisited, revisited: why and how you should (still) use dbscan. ACM Transactions on Database Systems (TODS) **42**(3), 19 (2017) 9
34. Sculley, D.: Web-scale k-means clustering. In: Proceedings of the 19th international conference on World wide web. pp. 1177–1178. ACM (2010) 9
35. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2016) 9
36. Shi, Y., Otto, C., Jain, A.K.: Face clustering: representation and pairwise constraints. IEEE Transactions on Information Forensics and Security **13**(7), 1626–1640 (2018) 3
37. Sohn, K., Shang, W., Yu, X., Chandraker, M.: Unsupervised domain adaptation for distance metric learning. In: ICLR (2019) 1, 5
38. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR (2014) 1

39. Toneva, M., Sordoni, A., Combes, R.T.d., Trischler, A., Bengio, Y., Gordon, G.J.: An empirical study of example forgetting during deep neural network learning. In: ICLR (2019) 4
40. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Change Loy, C.: The devil of face recognition is in the noise. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 765–780 (2018) 3, 4, 9, 13
41. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018) 3, 4, 7, 9, 13, 14
42. Yang, L., Zhan, X., Chen, D., Yan, J., Loy, C.C., Lin, D.: Learning to cluster faces on an affinity graph. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 2, 3, 5, 6, 8, 9
43. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014) 9, 13
44. Zhan, X., Liu, Z., Yan, J., Lin, D., Change Loy, C.: Consensus-driven propagation in massive unlabeled data for face recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 568–583 (2018) 1, 2, 3, 5, 9
45. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 4