VirFace: Enhancing Face Recognition via Unlabeled Shallow Data

Wenyu Li[1,*], Tianchu Guo[*], Pengyu Li, Binghui Chen, Biao Wang, Wangmeng Zuo[1](✉), Lei Zhang[2]
[1]School of Computer Science and Technology, Harbin Institute of Technology, China
[2]Hong Kong Polytechnic University, Hong Kong, China
liwenyu27@gmail.com, antares_tcguo@163.com, lipengyu007@gmail.com,
chenbinghui@bupt.edu.cn, wangbiao225@foxmail.com, wmzuo@hit.edu.cn, cslzhang@comp.polyu.edu.hk

Abstract

Recently, how to exploit unlabeled data for training face recognition models has been attracting increasing attention. However, few works consider the unlabeled shallow data[1] in real-world scenarios. The existing semi-supervised face recognition methods that focus on generating pseudo labels or minimizing softmax classification probabilities of the unlabeled data do not work very well on the unlabeled shallow data. It is still a challenge on how to effectively utilize the unlabeled shallow face data to improve the performance of face recognition. In this paper, we propose a novel face recognition method, named VirFace, to effectively exploit the unlabeled shallow data for face recognition. VirFace consists of VirClass and VirInstance. Specifically, VirClass enlarges the inter-class distance by injecting the unlabeled data as new identities, while VirInstance produces virtual instances sampled from the learned distribution of each identity to further enlarge the inter-class distance. To the best of our knowledge, we are the first to tackle the problem of unlabeled shallow face data. Extensive experiments have been conducted on both the small- and large-scale datasets, e.g. LFW and IJB-C, etc, demonstrating the superiority of the proposed method.

1. Introduction

Deep face recognition has benefit much from loss function [17, 30, 29, 6] and large-scale labeled datasets [28, 2, 10, 1]. Meanwhile, considering the fact that it is easy to obtain large amount of unlabeled face data while annotating these unlabeled data is time-consuming, several works [35, 34, 33, 36] has been proposed trying to enhance face recognition performance via the unlabeled face data. However, in real-world scenarios, the unlabeled face data prefers containing large amount of identities but only very few images per identity, namely shallow face data [8]. For better quantitative analysis, we define the shallow data as the data with no more than 5 images per identity in this paper. In such situation, we find that the existing semi-supervised learning methods [35, 34, 33, 36] do not work well.

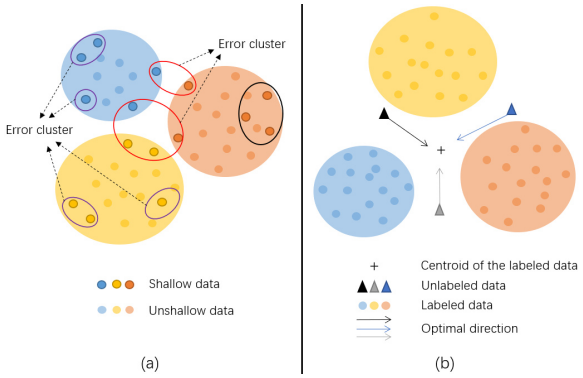


Figure 1. The dilemma of the existing semi-supervised face recognition methods on shallow data. (a) shows the clustering results of shallow data. In this figure, the error clusters are marked with red and purple circles. It is obvious that shallow data is hard to be well clustered. (b) presents the optimal problem of UIR loss. During training, the goal of optimizing these unlabeled data is to move them towards the centroid point of the labeled data. This is hard to optimize and may converge into a trivial solution.

Specifically, the clustering based methods [34, 33, 36] intend to assign the unlabeled data with pseudo-labels, and then combining with the labeled data together to train a new model. However, it is hard to cluster the shallow data as shown in Figure 1(a). In this figure, the red circles denote the error clusters that the samples from different identities are clustered into a same category, while the purple circles represent another kind of the clustering error that the samples from the same identity are clustered into different categories. Another work [35] proposed Unknown Identity Rejection(UIR) loss to make the unlabeled data be rejected by all the identities. However, as shown in Figure 1(b), it is easy to optimize the model to converge into a trivial solu-

[*]Equal contribution.
[1]*Shallow data* means there are only few images per identity [8].

tion, *i.e.* optimizing all the unlabeled data to the centroid of the labeled data. Moreover, Yu Liu *et al.* [19] indicates that UIR loss may learn an identity-irrelevant feature representation.

Additionally, self-learning is also a candidate solution for utilizing the unlabeled shallow data. Methods such as MOCO [12], SimCLR [4], and BYOL [9] have shown their superiority on representation learning in object classification as well as several downstream tasks *e.g.* object detection and segmentation. The core ingredient of above works is data augmentation which is utilized to obtain the positive samples and plays an important role in performance improvement. However, data used in face recognition is always aligned such that some data augmentation methods such as random cropping and rotation which are widely used in self-learning cannot be utilized.

In this paper, we propose VirFace, which consists of VirClass and VirInstance, to improve the supervised face recognition through the unlabeled shallow data. VirClass enlarge the inter-class distance by injecting the unlabeled data as new identities into the labeled space, while VirInstance produces virtual instances sampled from the learned distribution of each identity to further enlarge the inter-class distance. In summary, our proposed VirFace method can effectively utilize the unlabeled shallow data to learn a discriminative feature representation and to improve the performance over the supervised baselines.

The main contribution of this paper can be summarized as follows:

1. We propose a novel face recognition approach named VirFace which is the first to work on the unlabeled shallow data situation.

2. Our proposed VirFace contains VirClass and VirInstance which intend to enlarge inter-class distance and learn a discriminative feature representation.

3. The extensive experiments present significant performance improvement over supervised baselines in unlabeled shallow situation compared to other unlabeled approaches.

2. Related Work

2.1. Semi-Supervised Face Recognition

Most semi-supervised face recognition methods are clustering-based methods [34, 33, 36]. In these works, the unlabeled data is clustered and assigned with pseudo-labels. Then the pseudo-labeled data is combined with the labeled data to re-train the face recognition model. These clustering-based methods have achieved promising performance when the unlabeled data is from unshallow dataset, *e.g.* MS1M [10], VGGFace2 [2], IMDB-SenseTime [28], etc. However, on the shallow data, the existing cluster methods cannot achieve a good clustering performance due to the error clustering issues indicated in Figure 1(a).

Another method that works on the unlabeled face data is UIR [35]. An Unknown Identity Rejection(UIR) loss is proposed to learn a compact feature representation. To facility this, UIR loss minimizes the probabilities of all identities on the unlabeled data which does not belong to any labeled identity. However, this may lead the model converge to the centroid of the labeled dataset which is a single point. Furthermore, as discussed in [22], UIR loss may learn an identity-irrelevant feature representation.

2.2. Self-Supervised Learning and Metric Learning

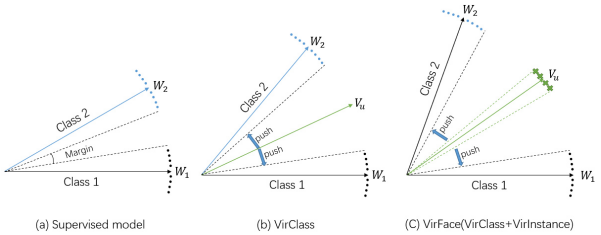
Recently, self-supervised learning has achieved great improvement in object classification and several downstream tasks *e.g.* object detection and segmentation [12, 4, 5, 9]. SimCLR [4, 5] studies the influence of projection head and different data augmentation methods. It also forms a standard data augmentation protocol consisting of random cropping, color distortion and Gaussian blur in order to generate positive samples for self-supervised learning. In order to increase the size of dictionary and keep the dictionary and the encoder in sync, MOCO [12] proposes a first-in-first-out queue dictionary and a momentum update protocol for dictionary model updating. BYOL [9] proposed an implicit contrastive learning method in which only positive pairs are used to simplify the training process. These self-supervised learning methods are all based on the augmentation method proposed in SimCLR [5]. While in face recognition task, since face images are often aligned first, it is impossible to implement the data augmentation which are widely used in self-learning, *e.g.* random cropping and rotation.

Metric learning methods have been implemented in supervised face recognition for a long time [26, 27, 24]. Since only pair-wise labels are needed, metric learning is a possible candidate for applying the unlabeled data. Considering that the number of the pairs used in the metric learning paradigm is restricted by the mini-batch size and the pairs' extraction strategy is tricky, these metric learning methods cannot achieve very high performance. Though most of these metric learning methods only take one negative pair into account, N-pair loss [26] considers multiple negative pairs and has made some progress in face recognition. Thus, we try to utilize the N-pair loss on the unlabeled shallow data along with ArcFace [6], but it doesn't work very well.

3. Proposed Method

In this section, we introduce our proposed VirFace method which contains VirClass and VirInstance. VirClass injects the unlabeled data as new identities into the labeled space to enlarge the inter-class distance, while VirInstance further sparse the inter-class by producing virtual instances sampled from the learned distribution of each identity.

Before we introduce our proposed method in detail, we first summarize the angular-margin based supervised face



(a) Supervised model (b) VirClass (C) VirFace(VirClass+VirInstance)

Figure 2. Geometrical interpretation of VirFace.(a) shows features and centroids from two identities through the supervised pre-trained model. (b) denotes the influence of adding a virtual class. The green arrow represents the virtual class. (c) represents the effect of VirFace which is the combination of VirClass and VirInstance. The crosses here stand for the virtual instances.

recognition losses as

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{f(w_{y_i}, x_i, m)}{f(w_{y_i}, x_i, m) + \sum_{j\neq y_i}^{n} f(w_j, x_i)} \quad (1)$$

where $f(\cdot)$ describes the cosine exponential term in SphereFace [17], AM-softmax [29], CosFace [30], or ArcFace [6]. m is the angular margin. w denotes the weight of the last FC layer and x indicates the output feature of backbone. N and n represent the batch size and the class number, respectively. In this section, we use this $f(\cdot)$ function to describe the cosine exponential term in our loss function for simplicity.

Since it is simple to deduplicate the overlapping identities by adding a softmax layer after the last FC layer and setting a threshold [35], we assume that the unlabeled data has no overlap identities with the labeled data. Section 3.4 introduces the detail of the deduplication method used in this paper.

In the rest of this section, we first give a brief description of VirClass and VirInstance which are two main components in our proposed VirFace method, and then the training strategy and the deduplication method are discussed.

3.1. VirClass

Since the unlabeled data has no label to indicate the exact identity the data belongs to, inspired by the Virtual-Softmax [3], we propose a concept of virtual class to give the unlabeled data a virtual identity in mini-batch. We treat these virtual classes as negative classes and try to find the centroid of each virtual class as the weight w in the last FC layer does which have been discussed in [31, 18]. Since the unlabeled data is shallow such that it is hard to find samples from the same identity in a mini-batch, each unlabeled feature can be a substitute to represent the centroid of its virtual class. Then we inject these centroids into the labeled space and maximize the angles between the labeled samples and the centroids of virtual classes to enlarge the inter-class distance. In order to reduce the storage cost and the computational consumption, we dynamically update the virtual classes along with the mini-batch.

To facilitate this, we add a virtual class term into the angular-based loss:

$$P_{vc_i} = \frac{f(w_{y_i}, x_i, m)}{f(w_{y_i}, x_i, m) + \sum_{j\neq y_i}^{n} f(w_j, x_i) + \sum_{u=1}^{U} f(v_u, x_i)} \quad (2)$$

$$L_{vc} = -\frac{1}{N}\sum_{i=1}^{N}\log P_{vc_i} \quad (3)$$

where U is the number of unlabeled data in mini-batch while N is the number of the labeled data in mini-batch. $v_u = x_u$ is the centroid of the virtual class u.

By optimizing L_{vc}, the inter-class distance can be enlarged by the additional virtual classes as shown in Figure 2(b). We name this virtual classes injecting method as VirClass.

Our further study finds that it is less likely to have samples of the same identity in a mini-batch when the identity number in the unlabeled data is much larger than the batch size. Meanwhile, since the unlabeled data is only worked as negative virtual classes, the samples with same identity in a mini-batch is equivalent to weight the corresponding virtual class. Thus, based on these analyses, our VirClass method can also work in the deep data situation.

3.2. VirInstance

For the purpose of exploiting more potential of the unlabeled data, we propose a further enhancement component VirInstance to get better use of the unlabeled data. VirInstance tends to generate feature distribution of each virtual class and then maximize the distance between the labeled centroids and the feature distributions of virtual classes to enlarge the inter-class distance as shown in Figure 2(c).

According to the central limit theorem [21], regardless the original distribution, the sampling distribution is always close to the normal distribution if the sampling number is large enough. If we treat all face features as a full face feature set, the features of each identity can be regarded as a sampling subset of the full set following a similar distribution form. Thus, we can learn this distribution form from labeled features and then predict the feature distributions of the unlabeled identities, *i.e.* the virtual classes, via the learned distribution form.

In order to formulate the distribution of each virtual class more conveniently, we randomly sample multiple virtual instances from the distribution of each virtual class. The virtual instances from the same virtual class represent the corresponding feature distribution. Thus, maximizing the distances between the labeled centroids and the feature distributions of virtual classes is the same to maximize the

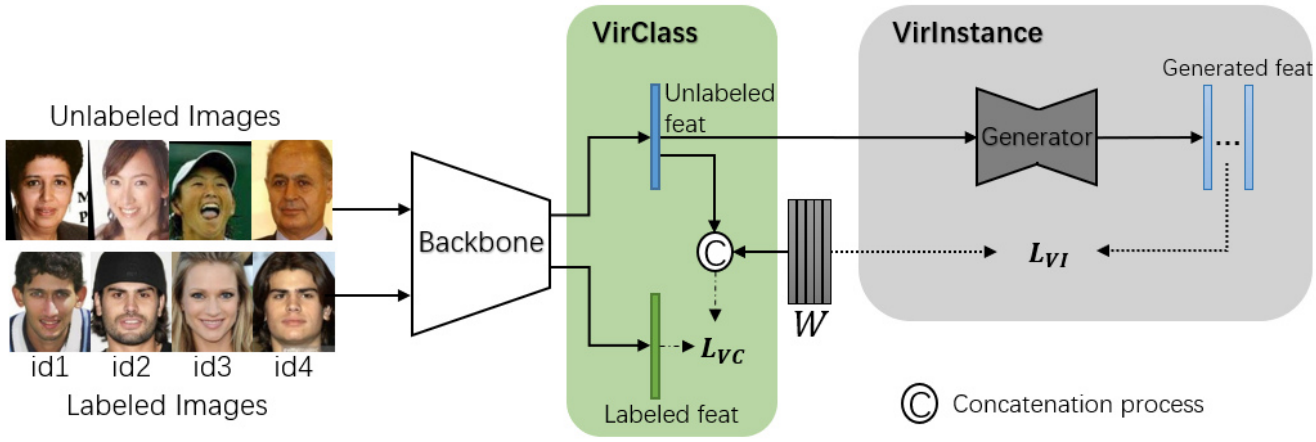


Figure 3. The framework of VirFace. VirClass and VirInstance are marked separately. W indicates the labeled centroids, *i.e.* the weights of the last FC layer in the pre-trained model. C denotes the concatenation operation.

distances between the labeled centroids and the virtual instances of each virtual class.

The loss function can be formulated as follows:

$$P_{vi_{i,s}} = \frac{f(v_i, x_{s_i}, m)}{f(v_i, x_{s_i}, m) + \sum_{u \neq i}^{U} f(v_u, x_{s_i}) + \sum_{j=1}^{n} f(w_j, x_{s_i})} \tag{4}$$

$$L_{vi} = -\frac{1}{U} \sum_{i=1}^{U} \sum_{s=1}^{S} \log P_{vi_{i,s}} \tag{5}$$

where v_i is the centroid of virtual class i, and x_{s_i} denotes the sth generated virtual instance of virtual class i. w_l is the centroid of label l in the labeled data. U and S denote the number of virtual classes and the number of virtual instances sampled from one virtual class.

We implemented a VAE network [16, 23] to predict the feature distribution and generate instances sampled from the feature distribution. For the encoder, our aim is to fit the distribution of each identity instead of the distribution of the whole dataset. Thus, we change the reconstruction loss of the VAE network to make the sampled features closer to the corresponding identity centroid as follows:

$$L_G = \frac{1}{N} \sum_{i=1}^{N} \|w_i - G(F_i)\|^2 + L_{KL} \tag{6}$$

where w_i denotes the centroid of identity i and F_i denotes the input feature. $G(\cdot)$ presents the VAE network. L_{KL} denotes the KL divergence loss.

We train this VAE network with the labeled features. Then, we use this trained VAE network to predict the feature distribution of the unlabeled data and generate virtual instances of the corresponding virtual class. The architecture of the VAE network and the detail of KL divergence loss are shown in supplementary.

3.3. Training strategies

Our VirFace is the combination of VirClass and VirInstance. The loss function is shown below:

$$L_{VirFace} = L_{vc} + L_{vi} \tag{7}$$

Since the $f(\cdot)$ function in L_{vc} and L_{vi} shown in Eq. 2 and Eq. 4 can refer to any of the angular-margin based loss functions, our proposed VirFace can work with any angular-margin based supervised face recognition method.

The framework of our proposed VirFace is shown in Figure 3. We divide the whole training process into two stages: pre-train stage and refining stage. Since the weight w of the last FC layer can denote the centroid of each identity in the pre-trained model [31, 18], we train the backbone network with the labeled data first. We use supervised face recognition loss function in the training process of the backbone network. After this, the VAE network is trained with the centroids of the identities and the labeled features outputing from the pre-trained backbone network. In the refining stage, we use the iterative training strategy to refine the backbone network and update the VAE network in order to keep sync. Algorithm 1 presents a clear pipeline of our training strategy.

3.4. Deduplication Process

Since it is common to have overlapping identities between the labeled data and the unlabeled data, we utilize a simple method to solve this issue. In the pre-train stage, we add a softmax layer following the last FC layer of the backbone network. Since the backbone network is trained with the labeled data, the max activation value of the labeled identities always hold a higher value than that of the unlabeled ones. Thus, we set a threshold to deduplicate the overlapping samples: the samples with a higher value over the threshold will be treated as an overlapping sample and

Algorithm 1: The Pipeline of VirFace

Data: Labeled data D_L, Unlabeled data D_U

/* Pre-train stage */

$Backbone \overset{train}{\longleftarrow} D_L$;

$Deduplication$ if necessary;

$Generator \overset{train}{\longleftarrow} D_L, fixed\, Backbone$;

/* Refining Stage */

Loop

  $Backbone \overset{train}{\longleftarrow} D_L, D_U, fixed\, Generator$;

  $Generator \overset{train}{\longleftarrow} D_L, fixed\, Backbone$;

end

Output: $Backbone$

will be dropped, otherwise the samples will be used as unlabeled data. In our experiments, the threshold is set to 0.8.

4. Experiments

In this section, we first describe the datasets and our implementation details in Section 4.1. Section 4.2 shows evaluation results and analysis of our proposed VirFace and the conventional semi-supervised face recognition methods when training on the unlabeled shallow face data. Then, the ablation study on VirFace is shown in Section 4.3. Finally, section 4.4 presents the performance of VirFace on large scale data and large scale networks.

4.1. Experimental Settings

Table 1 shows details of the training set and the testing set employed in our experiments.

Training Datasets. We employ the public dataset MS1M [10] as our training data. This data has been cleaned via the protocol mentioned in [6] so that the training data has no overlapping identities with the testing data. To construct shallow data, we divide the whole dataset into the labeled data and the unlabeled shallow data. In the real-world situation, the identities and samples of the unlabeled data are much more than those of the labeled data and the labeled data has limited identities but enough samples per identity. In order to fit this situation, we retain all samples which hold the labeled identities to build up the labeled data, and then randomly select 5 samples for each unlabeled identity to construct the unlabeled data. As a result, our label data contains 4,214 IDs and 77,935 samples denoting as MS1M-label, while the unlabeled shallow data consists 80,068 IDs and 400,222 samples denoting as MS1M-unlabel. These two datasets have no overlapping identities for simplicity.

For the purpose of verifying VirFace in large-scale dataset, we also employ Glint360k [1] as the unlabeled shallow data by randomly selecting 5 samples per identity. This unlabeled large-scale dataset consists 358,019 identities and 1,699,393 samples and is denoted as Glint360k-unlabel. In order to evaluate the ability of our proposed method in real-world scenarios, we also collect 4.8M images without label from website denoting as web-collected. We use the whole MS1M dataset as the labeled dataset to evaluate in the large-scale dataset.

Datasets	#Identity	#Images/Video
MS1M-label	4,214	77,935
MS1M-unlabel	80,068	400,222
MS1M [10]	84,282	5,757,574
Glint360k-unlabel	358,019	1,699,393
web collected	—	4,851,311
LFW [14]	5,749	13,233
CFP-FP [25]	500	7,000
CPLFW [37]	5,749	11,652
CALFW [38]	5,749	12,174
SLLFW [7]	5,749	13,233
IJB-B [32]	1,845	76.8K
IJB-C [20]	3,531	148.8K

Table 1. Detail of training and testing datasets.

Testing Datasets. During training, We evaluate our method on face verification datasets (*e.g.* LFW [14], CPLFW [37], CALFW [38], CFP-FP [25]) to check the performance of different settings. Besides, we also compare with other semi-supervised methods on SLLFW [7], IJB-B/C [32, 20], and MegaFace [15].

Implementation Details. For the backbone network architecture, we employ ResNet-50 and ResNet-101 [13, 11] to get 512-*D* embedding feature. We use the state-of-the-art ArcFace loss [6] as our basic loss. In the VAE network, we use one FC layer combining with two symmetric FC layers as an encoder to generate the feature distribution of each identity, while a re-sample layer following by two FC layers are used as a decoder to randomly sample and reconstruct virtual instances. More details of the VAE network architecture are in the supplementary.

In the pre-train process, the backbone network is trained on the labeled dataset for 20 epochs with 0.1 learning rate. The VAE network is trained on the labeled data for 4 epochs with 0.1 learning rate. In the refining process, the backbone network is fine-tuned on both the labeled and the unlabeled dataset with 10 epochs, while the VAE network updates for 4 epochs every 2 backbone training epochs. The learning rate is still 0.1 and decays on the [3rd, 5th, 7th, 9th] epoch. Batch size is setting to 144, and in a mini-batch, the number of the labeled and the unlabeled data are equal.

4.2. Evaluation Results

Table 2 shows the evaluation results of our proposed method and the conventional semi-supervised methods. We

Methods	LFW	CALFW	CPLFW	SLLFW	CFP-FP	IJB-B		IJB-C	
						Ver@1e-4	Id@Rank1	Ver@1e-4	Id@Rank1
ResNet50 Baseline	96.68	82.27	65.75	88.80	83.77	57.84	72.14	61.06	71.80
Cluster all	89.20	66.27	56.80	72.23	73.23	25.79	46.30	29.71	45.47
Cluster (#sample>1)	94.20	75.72	61.08	80.72	78.39	43.42	59.85	46.58	58.58
UIR [35]	96.68	83.25	66.67	89.95	80.60	63.58	74.20	66.74	74.53
N-pair [26]	97.00	83.15	65.68	89.60	84.09	58.67	72.27	62.09	72.27
VirClass	**97.33**	**85.28**	**68.01**	**91.00**	**85.74**	**62.37**	**75.93**	**66.26**	**76.25**
VirFace	**97.40**	**86.40**	**69.03**	**92.56**	**85.74**	**64.34**	**76.23**	**67.67**	**76.31**

Table 2. Evaluation results in RseNet50 and MS1M subsets. Our methods are shown in bold, and the best results are shown in red.

	#ID	#Sample
Ground Truth	80,068	400,222
Clustering Result	267,814	400,222
Clustering Result #sample>1	31,491	163,899
Clustering Result #sample= 1	236,323	236,323

Table 3. Clustering result of MS1M-unlabeled subset.

use MS1M-label as the labeled data and MS1M-unlabel as the unlabeled data. The architecture of backbone used here is ResNet-50. In Table 2, "ResNet50 Baseline" denotes the result of training only on the labeled data. "Cluster" represents the cluster-based methods. In "Cluster all", we combine all the pseudo labeled clustered data with the labeled data to form the training set, while in "Cluster(#sample>1)", we only use the clustered data whose pseudo label holds more than one sample. We also compare with UIR [35] and the modified N-pair loss [26]. "VirClass" and "VirFace" denote our proposed method without/with the VirInstance part. In Table 2, our VirClass model significantly outperforms other methods on all testing datasets. The VirFace which combines VirInstance with VirClass improves the performance of VirClass.

Analysis of Cluster-based Methods. The main reason that the cluster-based methods is inferior to the baseline is the poor quality of the pseudo labels. Table 3 shows the cluster result of Face-GCN [34] in our unlabeled subset which has 5 images per identity. Our unlabeled subset is clustered into 267,814 categories, which is over three times of the ground truth identity number. Moreover, for samples in the same cluster, we assume that the samples are clustered correctly when their ground truth identity holds the most samples in the cluster, while the samples with other ground truth identities are noise samples. Thus, the ratio of the correctly clustered samples over all samples is 35.76% which means almost 2/3 of the samples are noise. Then we go deeper into the clusters and find that 99.24% identities are clustered into several categories, which obviously increase the inter-class noise. Meanwhile, among the categories which have more than one sample, 85.07% holds samples from different identities meaning that the intra-class noise is also at a high level. Therefore, it is normal for the cluster-based methods to fail in the unlabeled shallow situation.

Analysis of UIR [35] and N-pair [26]. As described in Section 2.1, the UIR loss is hard so optimize and may lead to extracting identity-irrelevant features. Thus, it is hard to train and shows limited improvement over the baseline. N-pair is a representative of metric learning method which can work in both the labeled and the unlabeled situation. We use ArcFace [6] combined with N-pair loss to train. The result presents that N-pair loss brings limited improvement.

4.3. Ablation Study

We use MS1M-label and MS1M-unlabel subsets as the labeled and the unlabeled dataset, respectively. The backbone architecture is ResNet-50 in this section. The face verification datasets are LFW [14], CPLFW [37], CALFW [38], CFP-FP [25]. More results are shown in supplementary.

4.3.1 VirClass

In this section, we define several factors of the unlabeled data *i.e.* shallow rate, identity number and the scale of the unlabeled training set, to analyze the influence of these factors to the performance. Shallow rate means the number of samples per identity. Identity number denotes the number of the identities in the unlabeled dataset, and the scale of the unlabeled training set represents the number of samples in the unlabeled dataset. As our experiments, the scale of the unlabeled training set is the greatest factor to impact the face recognition performance.

Influence of Shallow Rate and Identity Number. In this part, we fix the scale of the unlabeled training set to 80,068, and evaluate different combinations of the shallow rate and the identity number. Table 4 shows the evaluation results on various testing datasets. From the results, the performances of different combinations have changed very little which means our VirClass method is not such sensitive to the shallow rate and the identity number when the scale of the unlabeled training set is fixed. Also, when the shallow rate equals to 5 and the identity number is 16,014, the

Methods	LFW	CALFW	CPLFW	CFP-FP
ResNet50 Baseline	96.68	82.27	65.75	83.77
shallow rate = 1 80,068 ids	97.05	84.53	67.10	84.25
shallow rate = 2 40,034 ids	97.15	84.48	67.18	84.11
shallow rate = 5 16,014 ids	97.01	84.38	67.03	83.88

Table 4. Result of different combination of shallow rate and identity number when fixing the scale of the unlabeled training set.

Methods	LFW	CALFW	CPLFW	CFP-FP
ResNet50 Baseline	96.68	82.27	65.75	83.77
80,068 samples shallow rate = 1	97.05	84.53	67.10	84.25
160,136 samples shallow rate = 2	97.21	84.93	67.88	84.64
400,222 samples shallow rate = 5	97.33	85.28	68.01	85.74

Table 5. Results of different scales of the unlabeled training set.

performance is the worst though the reduction is not much. Since this experiment holds the smallest identity number, the reduction may due to the lack of diversity in the training set.

Influence of the Scale of the unlabeled Training Set. In this part, we fix the identity number to 80,068 in order to guarantee the diversity, and evaluate the influence of different scales of the unlabeled training set. The results are shown in Table 5. The performance improves obviously as the scale of the unlabeled training set increases. Though the shallow rate is also changed with the increment of the scale of the unlabeled training set, it is obvious that the shallow rate has little contribution to the improvement when compare the 160,136 samples result in Table 5 with the shallow rate $= 2$ term in Table 4. Thus, the scale of the unlabeled training set is the most important factor on the VirClass improvement.

4.3.2 VirInstance

Improvement of the VirInstance. In this part, we demonstrate the performance of the VirInstance method. We evaluate it on different scales of the unlabeled training set and fix the sampling number of the VAE network to 5. The results are shown in Figure 4. The bars from left to right denote the performance of the ResNet50 baseline, VirClass and VirFace, respectively. The x-axix is the different scales of the unlabeled training set, and the y-axix is the performance. It shows that VirInstance can improve the performance of VirClass.

Sampling Number. We study the effect of different sampling number of the VAE network on the basis of the 1-shallow rate VirClass model. Sampling number means the number of the sampled virtual instances for each virtual

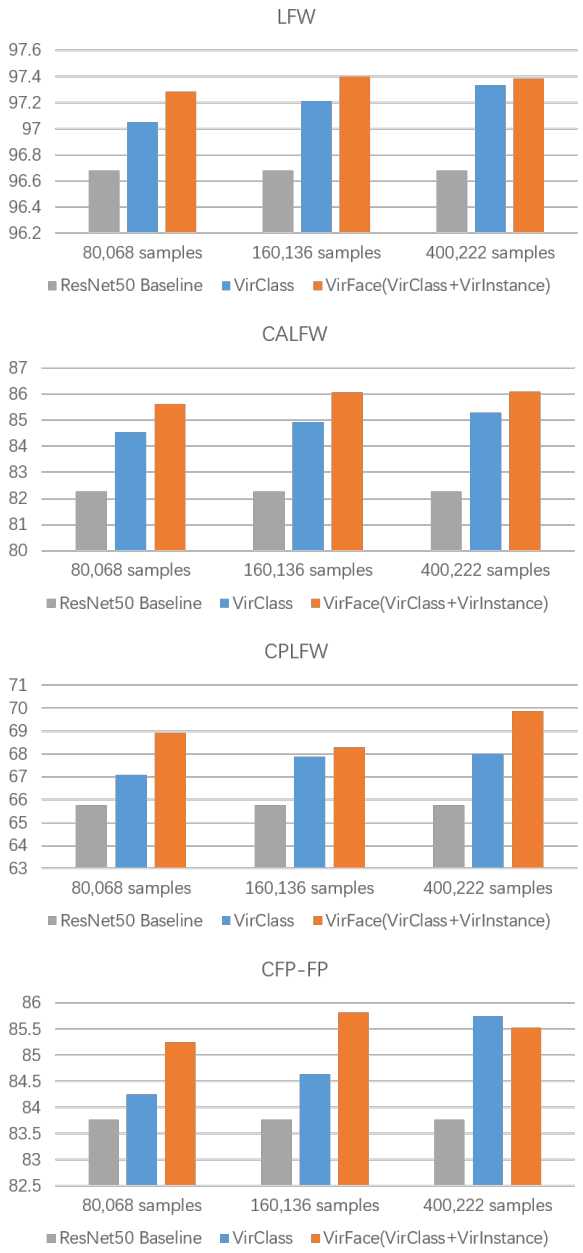


Figure 4. The improvement of VirInstance over VirClass.

class. The results are shown in Table. 7. The performance gets better as the sampling number increases. When sampling number is 5, our VirFace method reaches the best performance. The overall performance tends to be stable when sampling number exceeds 5.

Comparison with Data Augmentation. In this part, we compare the VirInstance method with the traditional data augmentation method. Different from our VirInstance, the data augmentation method generates images instead of features. We implement the random combination of blur and color jitter as data augmentation. They are friendly to

Methods	labeled dataset	unlabeled dataset	LFW	CALFW	CPLFW	SLLFW	CFP-FP	IJB-B		IJB-C	
								Ver@1e-4	Id@Rank1	Ver@1e-4	Id@Rank1
ResNet50 Baseline	MS1M	—	99.55	94.71	83.11	98.88	96.80	85.07	89.63	87.43	90.82
ResNet50 VirFace	MS1M	Glint360k unlabel	***99.61***	***94.96***	*83.35*	***98.96***	***96.78***	*87.54*	*90.23*	*89.60*	*91.34*
ResNet50 VirFace	MS1M	web-collected dataset	*99.58*	*94.80*	***83.85***	*98.90*	***96.78***	***87.94***	***90.54***	***90.18***	***91.65***
ResNet101 Baseline	MS1M	—	99.55	94.88	86.10	99.03	97.30	86.11	90.96	87.86	92.02
ResNet101 VirFace	MS1M	Glint360k unlabel	***99.58***	*95.11*	***86.51***	***99.13***	***97.51***	*88.45*	*91.47*	*90.19*	*92.33*
ResNet101 VirFace	MS1M	web-collected dataset	*99.56*	***95.15***	*86.25*	***99.13***	*97.15*	***88.90***	***91.66***	***90.54***	***92.68***

Table 6. Result on large-scale dataset. The performance improvements are shown in italic, and the best results are in italic & bold.

Methods	LFW	CALFW	CPLFW	CFP-FP
VirClass Baseline	97.05	84.53	67.10	84.25
VirFace (Sampling number = 2)	97.20	84.35	68.18	85.05
VirFace (Sampling number = 5)	97.40	86.40	69.03	85.74
VirFace (Sampling number = 10)	97.26	85.56	68.05	84.85

Table 7. Results of different sampling number.

Methods	LFW	CALFW	CPLFW	CFP-FP
VirClass Baseline	97.05	84.53	67.10	84.25
VirClass + DataAug (Generation number = 5)	97.20	85.06	65.88	83.82
VirFace (Sampling number = 5)	97.40	86.40	69.03	85.74

Table 8. Comparison with data augmentation method.

face recognition task as human face image is highly aligned. We generate 5 instances for both methods. The results are shown in Table. 8. There is little improvement when use the data augmentation method, denoted as "VirClass + DataAug" and VirInstance method outperforms on all the testing datasets. The reason is that the images generated by the data augmentation method are similar, resulting in lack of variation in the feature space which means these features are almost the same as the original one bringing little help.

To clearly demonstrate our analysis, we randomly sample 5,000 unlabeled images and generate 5 feature instances per image via the data augmentation method and our VirInstance method, respectively. Then, we calculate the cosine distance between the generated feature instances and the original features. The mean cosine distance of the data augmentation method is 0.83, while that of the VirInstance is 0.74. Since the data augmentation method holds a larger cosine distance, it is obvious that the feature instances generated by the data augmentation method is more similar to the original ones. Therefore, the VirInstance can increase the variation of the unlabeled data, which can enlarge the inter-class distance and significantly improve the performance.

4.4. Performance on the Large-Scale Training

In this part, we demonstrate that our method can also work in the large-scale labeled and unlabeled dataset. We test on both ResNet-50 and ResNet-101 architectures. The MS1M dataset [10] which has over 5M images is used as the labeled dataset. Glint360k-unlabel dataset is used as the unlabeled dataset. In order to test on the real-world scenario, we also collect 4.8M face images from website and construct an unlabeled dataset. The results are shown in Table 6. From the results, it is obvious that our proposed VirFace method can also work well on the large-scale datasets and the deep network architectures. Moreover, since the web-collected dataset is collected in real-world scenario and our VirFace method shows an improvement on it, it is clear that our proposed VirFace method can also work well in the real-world scenario.

5. Conclusion

In this paper, we proposed a semi-supervised face recognition framework dubbed VirFace to enhance the supervised face recognition performance via exploiting the unlabeled shallow data. We conducted comprehensive analysis and quantitative comparison on each part of our proposed VirFace method. The results validated the effectiveness of our proposed method. Moreover, we compared our proposed VirFace with the conventional semi-supervised face recognition methods and the experiments on various face recognition benchmarks showed the superiority of our proposed method.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under grant No.U19A2073. We hereby give specical thanks to Alibaba Group for their contribution to this paper.

References

[1] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Fu Ying. Partial fc: Training 10 million identities on a single machine. In *Arxiv 2010.05222*, 2020.

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

[3] Binghui Chen, Weihong Deng, and Haifeng Shen. Virtual class enhanced discriminative embedding learning. In *Advances in Neural Information Processing Systems*, pages 1942–1952, 2018.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[7] Weihong Deng, Jiani Hu, Nanhai Zhang, Binghui Chen, and Jun Guo. Fine-grained face verification: Fglfw database, baselines, and human-dcmn partnership. *Pattern Recognition*, 66:63–73, 2017.

[8] Hang Du, Hailin Shi, Yuchi Liu, Jun Wang, Zhen Lei, Dan Zeng, and Tao Mei. Semi-siamese training for shallow face learning. *ECCV*, 2020.

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[10] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[11] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5927–5935, 2017.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[15] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[18] Yu Liu, Guanglu Song, Jing Shao, Xiao Jin, and Xiaogang Wang. Transductive centroid projection for semi-supervised large-scale recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–86, 2018.

[19] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2080–2089, 2018.

[20] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018.

[21] Douglas C Montgomery and George C Runger. *Applied statistics and probability for engineers*. John Wiley and Sons, 2014.

[22] Haoyu Qin. Asymmetric rejection loss for fairer face recognition. *arXiv preprint arXiv:2002.03276*, 2020.

[23] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[25] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[26] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016.

[27] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[28] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face

recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018.

[29] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[30] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[31] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[32] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 90–98, 2017.

[33] Lei Yang, Dapeng Chen, Xiaohang Zhan, Rui Zhao, Chen Change Loy, and Dahua Lin. Learning to cluster faces via confidence and connectivity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13369–13378, 2020.

[34] Lei Yang, Xiaohang Zhan, Dapeng Chen, Junjie Yan, Chen Change Loy, and Dahua Lin. Learning to cluster faces on an affinity graph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2298–2306, 2019.

[35] Haiming Yu, Yin Fan, Keyu Chen, He Yan, Xiangju Lu, Junhui Liu, and Danming Xie. Unknown identity rejection loss: Utilizing unlabeled data for face recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[36] Xiaohang Zhan, Ziwei Liu, Junjie Yan, Dahua Lin, and Chen Change Loy. Consensus-driven propagation in massive unlabeled data for face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 568–583, 2018.

[37] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5, 2018.

[38] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.