VirFace: Enhancing Face Recognition via Unlabeled Shallow Data

Wenyu Li\textsuperscript{1,}\textsuperscript{*}, Tianchu Guo\textsuperscript{*}, Pengyu Li, Binghui Chen, Biao Wang, Wangmeng Zuo\textsuperscript{1}(✉), Lei Zhang\textsuperscript{2}
\textsuperscript{1}School of Computer Science and Technology, Harbin Institute of Technology, China
\textsuperscript{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\textsuperscript{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 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 \cite{17, 30, 29, 6} and large-scale labeled datasets \cite{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 \cite{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 \cite{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 \cite{35, 34, 33, 36} do not work well.

Specifically, the clustering based methods \cite{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 \cite{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 solution.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{Figure1.png}
\caption{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.}
\end{figure}

\textsuperscript{*}Equal contribution.

\textsuperscript{1}Shallow data means there are only few images per identity \cite{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
In this section, we use this function and the centroids of virtual classes to enlarge the inter-class 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:

$$L_{vc} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{f(w_{yi}, x_i, m)}{f(w_{yi}, x_i, m) + \sum_{j \neq yi} f(w_j, x_i) + \sum_{u=1}^{U} f(u, x_i)}$$

where $U$ is the number of unlabeled data in mini-batch while $N$ is the number of the labeled data in mini-batch. $u_u = x_u$ 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 Vir-Class.

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:

\[ P_{vi,s} = \frac{f(v_i, x_{si}, m)}{f(v_i, x_{si}, m) + \sum_{u \neq i}^{U} f(v_u, x_{si}) + \sum_{j=1}^{n} f(w_j, x_{si})} \]  

(4)

\[ L_{vi} = -\frac{1}{U} \sum_{i=1}^{U} \sum_{s=1}^{S} \log P_{vi,s} \]  

(5)

where \( v_i \) is the centroid of virtual class \( i \), and \( x_{si} \) denotes the \( s \)th 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:

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

(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:

\[ L_{VirFace} = L_{vc} + L_{vi} \]  

(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 outputting 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
Table 1 shows details of the training and testing datasets.

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

Table 1. Detail of training and testing datasets.

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.

**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
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 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

<table>
<thead>
<tr>
<th>Methods</th>
<th>LFW</th>
<th>CALFW</th>
<th>CPLFW</th>
<th>SLLFW</th>
<th>CFP-FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50 Baseline</td>
<td>96.68</td>
<td>82.27</td>
<td>65.75</td>
<td>88.80</td>
<td>83.77</td>
</tr>
<tr>
<td>Cluster all</td>
<td>89.20</td>
<td>66.27</td>
<td>56.80</td>
<td>72.23</td>
<td>73.23</td>
</tr>
<tr>
<td>Cluster (#sample&gt;1)</td>
<td>94.20</td>
<td>75.72</td>
<td>61.08</td>
<td>80.72</td>
<td>78.39</td>
</tr>
<tr>
<td>UIR [35]</td>
<td>96.68</td>
<td>83.25</td>
<td>66.67</td>
<td>89.95</td>
<td>80.60</td>
</tr>
<tr>
<td>N-pair [26]</td>
<td>97.00</td>
<td>83.15</td>
<td>65.68</td>
<td>89.60</td>
<td>84.09</td>
</tr>
<tr>
<td>VirClass</td>
<td>97.33</td>
<td>85.28</td>
<td>68.01</td>
<td>91.00</td>
<td>85.74</td>
</tr>
<tr>
<td>VirFace</td>
<td>97.40</td>
<td>86.40</td>
<td>69.03</td>
<td>92.56</td>
<td>85.74</td>
</tr>
</tbody>
</table>

Table 2. Evaluation results in ResNet50 and MS1M subsets. Our methods are shown in bold, and the best results are shown in red.

<table>
<thead>
<tr>
<th>Methods</th>
<th>LFW</th>
<th>CALFW</th>
<th>CPLFW</th>
<th>SLLFW</th>
<th>CFP-FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50 Baseline</td>
<td>57.84</td>
<td>72.14</td>
<td>61.06</td>
<td>71.80</td>
<td></td>
</tr>
<tr>
<td>Cluster all</td>
<td>52.79</td>
<td>46.30</td>
<td>29.71</td>
<td>45.47</td>
<td></td>
</tr>
<tr>
<td>Cluster (#sample&gt;1)</td>
<td>43.42</td>
<td>59.85</td>
<td>46.58</td>
<td>58.58</td>
<td></td>
</tr>
<tr>
<td>UIR [35]</td>
<td>63.58</td>
<td>74.20</td>
<td>66.74</td>
<td>74.53</td>
<td></td>
</tr>
<tr>
<td>N-pair [26]</td>
<td>58.67</td>
<td>63.58</td>
<td>73.23</td>
<td>80.60</td>
<td></td>
</tr>
<tr>
<td>VirClass</td>
<td>62.37</td>
<td>75.93</td>
<td>66.26</td>
<td>76.25</td>
<td></td>
</tr>
<tr>
<td>VirFace</td>
<td>64.34</td>
<td>76.23</td>
<td>67.67</td>
<td>76.31</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Clustering result of MS1M-unlabeled subset.
### Methods

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

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

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

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

- **Methods**
  - LFW
  - CALFW
  - CPLFW
  - CFP-FP

- **Results**
  - 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

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 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
In this part, we demonstrate that our 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.

### 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.

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>labeled dataset</th>
<th>unlabeled dataset</th>
<th>LFW</th>
<th>CALFW</th>
<th>CPLFW</th>
<th>SLLFW</th>
<th>CFP-FP</th>
<th>IJB-B Ver@1e-4</th>
<th>IJB-C Id@Rank1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50</td>
<td>MS1M</td>
<td>—</td>
<td>99.55</td>
<td>94.71</td>
<td>83.11</td>
<td>98.88</td>
<td>96.80</td>
<td>85.07</td>
<td>90.63</td>
</tr>
<tr>
<td>VirFace</td>
<td>MS1M</td>
<td>Glint360k unlabel</td>
<td><strong>99.61</strong></td>
<td><strong>94.96</strong></td>
<td><strong>83.35</strong></td>
<td><strong>98.96</strong></td>
<td><strong>96.78</strong></td>
<td><strong>87.54</strong></td>
<td><strong>90.23</strong></td>
</tr>
<tr>
<td>ResNet50</td>
<td>MS1M</td>
<td>web-collected</td>
<td>99.58</td>
<td>94.80</td>
<td>83.85</td>
<td>98.90</td>
<td>96.78</td>
<td>87.94</td>
<td>90.54</td>
</tr>
<tr>
<td>ResNet101</td>
<td>MS1M</td>
<td>—</td>
<td>99.55</td>
<td>94.88</td>
<td>86.10</td>
<td>99.03</td>
<td>97.30</td>
<td>86.11</td>
<td>90.96</td>
</tr>
<tr>
<td>VirFace</td>
<td>MS1M</td>
<td>Glint360k unlabel</td>
<td><strong>99.58</strong></td>
<td><strong>95.11</strong></td>
<td><strong>86.51</strong></td>
<td><strong>99.13</strong></td>
<td><strong>97.51</strong></td>
<td><strong>88.45</strong></td>
<td><strong>91.47</strong></td>
</tr>
<tr>
<td>ResNet101</td>
<td>MS1M</td>
<td>web-collected</td>
<td>99.56</td>
<td><strong>95.15</strong></td>
<td><strong>86.25</strong></td>
<td><strong>99.13</strong></td>
<td><strong>97.15</strong></td>
<td><strong>88.90</strong></td>
<td><strong>91.66</strong></td>
</tr>
</tbody>
</table>

**Table 7. Results of different sampling number.**

<table>
<thead>
<tr>
<th>Methods</th>
<th>LFW</th>
<th>CALFW</th>
<th>CPLFW</th>
<th>CFP-FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>VirClass Baseline</td>
<td>97.05</td>
<td>84.53</td>
<td>67.10</td>
<td>84.25</td>
</tr>
<tr>
<td>VirFace</td>
<td>97.20</td>
<td>84.35</td>
<td>68.18</td>
<td>85.05</td>
</tr>
<tr>
<td>(Sampling number = 2)</td>
<td>97.40</td>
<td>86.40</td>
<td>69.03</td>
<td>85.74</td>
</tr>
<tr>
<td>VirFace</td>
<td>97.26</td>
<td>85.56</td>
<td>68.05</td>
<td>84.85</td>
</tr>
<tr>
<td>(Sampling number = 5)</td>
<td>97.20</td>
<td>85.06</td>
<td>65.88</td>
<td>83.82</td>
</tr>
<tr>
<td>VirFace</td>
<td>97.40</td>
<td>86.40</td>
<td>69.03</td>
<td>85.74</td>
</tr>
<tr>
<td>(Sampling number = 10)</td>
<td>97.20</td>
<td>85.06</td>
<td>65.88</td>
<td>83.82</td>
</tr>
</tbody>
</table>

**Table 8. Comparison with data augmentation method.**

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 special thanks to Alibaba Group for their contribution to this paper.
References


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


