Continual Representation Learning for Biometric Identification

Bo Zhao¹,², Shixiang Tang¹,²,³, Dapeng Chen², Hakan Bilen¹, Rui Zhao²
¹The University of Edinburgh, ²SenseTime Group Limited, ³The University of Sydney
*Equal Contribution
{bo.zhao, hbilen}@ed.ac.uk, {tangshixiang, chendapeng, zhaorui}@sensetime.com

Abstract

With the explosion of digital data in recent years, continuously learning new tasks from a stream of data without forgetting previously acquired knowledge has become increasingly important. In this paper, we propose a new continual learning (CL) setting, namely "continual representation learning", which focuses on learning better representation in a continuous way. We also provide two large-scale multi-step benchmarks for biometric identification, where the visual appearance of different classes are highly relevant. In contrast to requiring the model to recognize more learned classes, we aim to learn feature representation that can be better generalized to not only previously unseen images but also unseen classes/identities. For the new setting, we propose a novel approach that performs the knowledge distillation over a large number of identities by applying the neighbourhood selection and consistency relaxation strategies to improve scalability and flexibility of the continual learning model. We demonstrate that existing CL methods can improve the representation in the new setting, and our method achieves better results than the competitors.

1. Introduction

Biometric identification [22, 52], including face recognition [10, 37, 69] and person re-identification (re-id) [34, 68, 73], has achieved significant progress in the recent years due to the advances in modern learnable representations [7, 8, 10, 19, 34, 53, 61, 69] and emerging large datasets [15, 21, 27, 29, 68, 79, 81]. In particular, deep neural networks (DNN) [17, 50, 57, 60] are shown to learn features that encode complex and mosaic biometrics traits and achieve better feature generalization ability, when trained on large-scale datasets. However, the paradigm of training DNNs offline becomes impractical and inefficient with the increase in stream data such as surveillance videos and online images/texts. For example, the intelligent security system [66, 80] in a city or an airport captures millions of new images every day. In this scenario, training a model with all the images in one step can never be realized. To continuously improve our model with limited computational and storage resource, we expect the model to be trained online only with the newly obtained data.

Motivated by this, we propose a new but realistic setting named “continual representation learning” (CRL) for this real-world biometric identification problem. The new setting aims to learn from continuous stream data, meanwhile continuously improving model’s generalization ability on unseen classes/identities.

In the standard offline learning paradigm, the model cannot preserve the previous knowledge well when being continuously trained on new tasks without access to old tasks, which is known as the Catastrophic Forgetting [12, 13, 42, 43, 48]. Continual learning (CL) [6, 39, 45, 49, 54, 55] becomes an important research topic to alleviate such problem. For image recognition, continual learning is typically formulated as the class-incremental classification task. The training process includes a sequence of training steps, and each step involves training with the images from new classes. Once the model is trained on the data of new classes, its performance is measured on a set of images from both old and new classes. The classes in the testing set are all previously seen (appeared) in the training set, thus the main goal is to recognize as many classes as possible without forgetting old classes.

Yet, this setting is not ideal for the biometric identification problems for various reasons. First, biometric identification typically consists of train and test sets which are dis-
joint in terms of classes (or identities). The performance on learned old classes in CRL is easily kept high while learning new classes, which means CRL suffers little from the traditional forgetting problem, i.e., forgetting old classes. Hence, it is not suitable to measure model’s generalization ability on seen (old and new) classes in the new setting. We give the experimental evidence in Sec. 5.2. Thus, the main goal of our CRL is to generalize to previously unseen classes which is in contrast to the typical CL setting. The comparison of two settings is illustrated in Figure 1. Second, biometric identification focuses on a more challenging setting which is similar to fine-grained classification [2, 65, 72, 75]. The intra-class appearance variations are significantly subtler than the standard object classes in the commonly used CIFAR-100 [30] and ImageNet [51] datasets. Hence, it is particularly challenging for continual representation learning, as the model has to learn better representation during many learning steps and improve the ability to discriminate unseen classes/identities.

Most existing CL benchmarks, illustrated in Table 1, are for either small-scale (e.g., MNIST [31], CIFAR-100 [30], CUB [63]) or coarse-grained (e.g., CORE50 [38] and ImageNet [51]) object recognition. To the best of our knowledge, there is no large-scale continual learning benchmark for biometric identification. To simulate the continuous stream setting for biometric identification, we propose two large-scale benchmarks for face recognition and person re-id, which contain around 92K and 7K identities respectively. As shown in Table 1, the proposed two benchmarks are larger than all existing CL benchmarks when both class and image numbers are considered.

The traditional continual learning methods usually learn classifiers for small-scale seen classes, and they are hardly scalable to a large number of identities in the real scenario. For example, the popular LWF method [36] regularizes the consistency of outputs of all old classifiers in old and new models for knowledge distillation, in addition to minimizing the classification loss for learning new classes. A large number of identities will prohibit the usage of previous methods, because the limited memory and computation resources of GPU cannot handle the huge fully connected connection layer. To solve this problem, we also propose a method that implements knowledge distillation regarding the outputs of selected classes instead of all classes. In particular, the knowledge distillation is based on KL divergence instead of cross-entropy to regularize the difference between the outputs of old and new models. We then relax the regularization by an adaptive margin to give the model more flexibility to learn new knowledge.

In summary, our contributions are two-fold: (1) We propose a new continual learning setting for learning better representation in biometric identification. Such setting requires a large-scale multi-step training set and a third-party testing set with identities that have never appeared in the training set. For this reason, we introduce two large-scale benchmarks for continual face recognition and continual person re-id. (2) To address the new setting, we propose a novel method with neighbourhood selection (NS) and consistency relaxation (CR) for knowledge distillation, which significantly improves the scalability and learning flexibility. Extensive experiments show that the representation can actually be improved in the continual representation learning setting by existing knowledge distillation strategies, and the proposed method achieves better results.

2. Related Work

Biometric Identification. Much progress has been achieved in biometric identification including face recognition [10, 37, 69] and person re-id [34, 68, 73] by learning better representation with different losses, e.g., softmax-based losses [10, 34, 69], triplet-based losses [8, 19, 53] and other kinds of losses [7, 61], on large-scale image datasets [15, 21, 27, 29, 68, 79, 81]. Different from object recognition, biometric identification focuses on learning better representation for large-number of fine-grained identities.

However, few works concern how to learn better representation from biometric data stream. The existing related works are different from our setting in terms of goal, training/testing protocol and dataset scale. Some methods [41, 59, 64] were proposed for online person re-id. Unfortunately, all observed training data need to be stored. In contrast, data of old classes are not accessible in our CRL setting. [35] proposed an online-learning method for one-pass person re-id. They used a fixed feature extractor, while we aim to continually learn better representation.

Continual Learning. Continual learning is also named lifelong learning [46, 56, 58, 67], incremental learning [5, 49, 55] and sequential learning [5, 9] in previous works. Existing continual learning works focus on general object recognition [6, 28, 49], object detection [14, 55], image generation [33, 70], reinforcement learning tasks [1, 26, 74] and unsupervised learning tasks [1, 11]. The popular continual learning setting is to continuously learn new data/classes and test on all seen (both old and new) classes, and it suffers from the catastrophic forgetting problem.

A number of methods are proposed to avoid the catastrophic forgetting of deep models. Generally speaking, they can be divided into two kinds. The one is based on rehearsal [6, 39, 49] or pseudo-rehearsal [25, 54, 70], which requires an extra memory or generative model to remember old task data. The other one is based on the regularization on weights [3, 28, 77], features [24], and outputs [6, 36].

The popular benchmarks for evaluating these CL methods are (original or permuted) MNIST [31], CORE50 [38], CIFAR-100 [30], CUB [63] and ImageNet [51]. Except ImageNet (with 1K classes and 1.3M images), all other bench-
marks are small-scale in terms of class (≤ 200) and image (≤ 170K) numbers. In addition, except CUB, all benchmarks are about coarse-grained objects. Hence, they are not suitable for evaluating the representation ability of deep models. Different from the popular CL setting and benchmarks, the proposed CRL aims to continuously learn more generalized representation model for identifying many unseen classes/identities. The proposed two benchmarks are the first large-scale benchmarks for CRL, and they are much larger than existing CL benchmarks in terms of class (92K and 7K) and image (5.8M and 91K) numbers.

3. CRL Setting and Benchmarks

CRL Setting. As illustrated in Figure 1, the model will be trained for in total $T = 5$ steps starting from step 0. Each training step includes $L$ classes, and training classes of different steps are disjoint. The model can only access the training data of current learning step $t$. For example, assuming current step $t = 2$, the model is trained only on training data of new classes $c_{2L}, \ldots, c_{2L-1}$. Without accessing old classes $c_0, \ldots, c_{2L-1}$, the model will gradually forget the knowledge obtained from previous learning steps. In each step, CRL tests the model on previously unseen testing classes $c'_{0}, \ldots, c'_{j-1}$ for evaluating model’s generalization ability, which is frequently used as the performance metric in biometric identification tasks.

We present continual representation learning benchmarks for two popular biometric identification tasks, namely, face recognition and person re-id. The statistics of the two benchmarks are shown in Table 1. The presented benchmarks are different from existing continual learning benchmarks in three main aspects.

- The proposed CRL benchmarks are the first ones designed for biometric (face and person) identification in continual learning.
- The number of classes in our benchmarks (92K and 7K) is much larger than existing benchmarks (≤ 1K).
- We test the model on novel identities that have never appeared in the training set, while existing benchmarks test on new images of learned (seen) classes.

<table>
<thead>
<tr>
<th>Task</th>
<th>Scale</th>
<th>Concept</th>
<th>Classes</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST/1K</td>
<td>Cls</td>
<td>Small</td>
<td>18,092</td>
<td>13,088</td>
</tr>
<tr>
<td>CUBAR Face</td>
<td>Rpt</td>
<td>Large</td>
<td>8,573,772</td>
<td>5,783,772</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>50,000</td>
<td>45,000</td>
</tr>
<tr>
<td>CUB</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,331,167</td>
<td>1,281,167</td>
</tr>
<tr>
<td>ImageNet</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CUB AR</td>
<td>Rpt</td>
<td>Coarse-grained</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>CRL-person</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>2,494</td>
<td>2,494</td>
</tr>
<tr>
<td>CRL-face</td>
<td>Rpt</td>
<td>Fine-grained</td>
<td>85,738</td>
<td>85,738</td>
</tr>
</tbody>
</table>

Table 1: Statistics of popular CL benchmarks and the proposed CRL-face and CRL-person. Cls: Classification. Rpt: Representation.

face benchmark, because there are 85,738 identities and 5,783,772 images in the dataset. We divide identities in Ms1M into 5 and 10 subsets randomly and equally. Each split subset has around 17,148 identities for 5-step setting and 8,573 identities for 10-step setting, respectively. The number of images in each subset varies because the number of images associated with each identity is not equal. Each subset serves as the training set in each learning step.

Two testing datasets, namely LFW [21] and Megaface [27], are used for evaluating the representation ability of models. The LFW dataset is the most widely used as the testing benchmark that contains 6,000 testing pairs from 5,749 identities. We follow the unrestricted with labelled outside data protocol, where features are trained with additional data and the verification accuracy is estimated by a 10-fold cross validation scheme. 9 folders are combined as the validation set to determine the threshold , and the 10th folder is used for testing. The Megaface benchmark is another challenge for face recognition. It contains 1M images of 690K different individuals as the gallery set and 100K photos of 530 unique individuals from FaceScrub as the probe set. For testing, the target set has 4000 images of 80 identities, and the distractor set has over 1M images of different identities. The Top 1 accuracy is reported.

CRL-person Benchmark To obtain enough identities for implementing continual learning, we combine three popular person re-id datasets, namely, Market1501 [79], DukeMTMC-reID [81] and MSMT17_V2 [68]. The mixed dataset (CRL-person) contains 2,494 training identities. Specifically, the three datasets contribute 751, 702 and 1,041 training identities respectively. In total, 59,706 training images of the 2,494 identities are employed as the training set. The training set is split into 5 subsets and 10 subsets for 5-step and 10-step learning respectively.

For testing, we combine the testing sets of the three datasets. However, evaluating the model on all testing data is computational expensive, as there are 17,255 query images and 119,554 gallery images in total. Thus, we apply two strategies to reduce the image number: (1) We keep all query identities and remove those identities only in the gallery set. (2) We randomly select at most one image for each identity under each camera in query and gallery set respectively. After applying the two strategies, the final testing set has 11,351 query images and 19,576 gallery images.
4. Flexible Knowledge Distillation for CRL

According to our protocol, we need to continuously train our identification model for multi-steps. The tth step provides new data \( O^t = \{ (x_i^t, y_i^t) \}_{i=1}^n \), where each instance \( (x_i^t, y_i^t) \) is composed by an image \( x_i^t \in \mathcal{X}^t \) and a label \( y_i^t \in \mathcal{Y}^t \). \( n = |O^t| \) is the number of all new data. The goal of CRL is to construct an embedding function \( f \), which can compute a feature representation \( \phi_i \) to better associate with \( y_i^t \). To accomplish this, we consider the \( f(x_i^t, \theta^t) \) parameterized by \( \theta^t \), and define a classification loss based on the tth step data \( O^t \):

\[
\begin{align*}
L^{new}(\theta^t, w^t; O^t) &= \frac{1}{n} \sum_{i=1}^{n} l(\phi_i, y_i^t, \theta^t, w^t), \\
\phi_i &= f(x_i^t, \theta^t), \\
l(\phi_i, y_i^t, \theta^t, w^t) &= -\log \frac{\exp(\phi_i^T w_{y_i^t})}{\sum_{j=1}^{l} \exp(\phi_i^T w_j)}.
\end{align*}
\]

where \( l = |\mathcal{L}| \) and \( w_j \) indicates the classifier for the jth class. Obviously, minimizing the loss in Eq. 1 will result in overfitting to the instances in \( O^t \). As an alternative, we could additionally maintain a memory data set to keep the predictions at the past steps invariant, which will lead to the problem on how to select the most useful samples from the past data. This paper focuses on the scenarios where there are no memory data. We only have the model \( f(x_i^t, \theta^{t-1}) \) and classifiers \( w^{t-1} \) in the last step. It is suitable to employ knowledge distillation(KD) to optimize a loss function based on the old model and the current data.

\subsection{4.1. Knowledge Distillation}

The idea of knowledge distillation was found by Hinton et al., which works well for encouraging the output of one network to approximate that of another network. Suppose \( w^{t-1} \) is about \( L \) classes, the output probability of \( y_i^t \) generated by the old model \( f(x_i^t, \theta^{t-1}) \) given \( w^{t-1} \) is: \( p_i = \{ p_{i,1}, p_{i,2}, ..., p_{i,L} \} \). The cross-entropy loss is utilized to regularize the new probability \( q_i = \{ q_{i,1}, q_{i,2}, ..., q_{i,L} \} \) generated by the new model \( f(x_i^t, \theta^t) \):

\[
L^{old}(\theta^t, w^t; O^t) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{L} p_{i,l} \log q_{i,l},
\]

where \( l \) is the index of the class.

\textbf{Discussion}. Compared with the existing CL scenarios, a crucial difference for the proposed setting is that it handles a large number of classes/identities. For example, existing CL methods are evaluated on 10 classes of MNIST, 100 classes of CIFAR and 1,000 classes of ImageNet, while the model for biometric identification usually needs to be trained over thousands to millions classes. The scalability and efficiency of training methods become important due to the limited memory and computation resources. Furthermore, the current knowledge distillation require the strict consistency between the outputs of new and old models, and it largely restricts the ability to learn new knowledge. Based on the above concerns, we propose Flexible Knowledge Distillation(FKD), where we perform neighbour selection and consistency relaxation over the loss related to the old model \( L^{old} \). The proposed method is illustrated in Figure 2.

\subsection{4.2. Neighbourhood Selection}

In the standard knowledge distillation(Eq.2), the probability distribution is calculated based on the activation (the
Algorithm 1 Flexible Knowledge Distillation

Require:
\[ O' = \{(x'_i, y'_i)\}_{i=1}^{n} : \text{training data in current learning step}; \]
\[ \theta^{-1} : \text{old model}; \ w^{-1} : \text{old classifiers}; \]

Ensure:
\[ \theta^t : \text{new model}; \ w^t : \text{new classifiers}; \]
1: Initialize \( \theta^t, w^t \) by \( \theta^{-1}, w^{-1} \);
2: for a batch in \( O' \) do
3: Compute classification loss \( L_{\text{new}} \) using Eq. 1;
4: for \( (x'_i, y'_i) \) in the batch do
5: Determine the neighborhood \( S_i \) based on activations \( a_i \) (old) and \( b_i \) (new);
6: Compute relaxed KL Divergence \( D'_{KL} \) using Eq. 6;
7: Compute distillation loss \( L_{dist} \) using Eq. 7 and the final loss \( L \) using Eq. 8;
8: Update \( \theta^t \) and \( w^t \) by back-propagation with \( L \).

direct output of the classifiers) by a softmax layer. When the number of classes increases to thousands even millions, its not scalable for maintaining such a large fully-connected classifier layer. At the same time, the computation of softmax probability is not effective for such a large number of classes, as the probability values are weakened by many unrelated classes. Hence, we select a few "similar" classes from all old classes to implement selected knowledge distillation.

Given a sample \( x'_i \), the activation generated by the old model is denoted by \( a_i = \{a_{i,1}, a_{i,2}, ..., a_{i,L}\} \) and the activation of the new model is denoted by \( b_i = \{b_{i,1}, b_{i,2}, ..., b_{i,L}\} \). We rank the activation units of the old model \( a_i \) with descending order, select the top \( K \) ones, and put their indices into the set \( S_i \), i.e., the neighborhood of the ground-truth class \( y'_i \). The probabilities generated by the old and new model are \( p_i \) and \( q_i \), which are calculated based on the selected label set \( S_i \):
\[
p_{i,l} = \frac{\exp(a_{i,l}/T)}{\sum_{j \in S_i} \exp(a_{i,j}/T)}, \quad q_{i,l} = \frac{\exp(b_{i,l}/T)}{\sum_{j \in S_i} \exp(b_{i,j}/T)},
\]
where \( T \) is the hyper-parameter of knowledge distillation. Instead of using cross-entropy loss, we utilize the Kullback Leibler(KL) divergence, i.e. KLD, to measure the difference between \( p_i \) and \( q_i \):
\[
D_{KL}(p_i || q_i) = \sum_{l \in S_i} p_{i,l} \log p_{i,l} - p_{i,l} \log q_{i,l}.
\]
As \( \sum_{l \in S_i} p_{i,l} \log p_{i,l} \) is a constant in the optimization, the KL-divergence is equivalent to the cross-entropy in Eq. 2, and \( D_{KL}(p_i || q_i) \) will be 0 if \( p_i \) and \( q_i \) are the same.

4.3. Consistency Relaxation

The new model needs to learn knowledge from both old and new classes. The best parameters of new model should not be exactly the same as the old model. Hence, we introduce an adaptive margin \( \delta_i \) to relax the consistency constraint. The margin for \( x_i \) is set to be:
\[
\delta_i = -\beta \sum_{l \in S_i} p_{i,l} \log p_{i,l},
\]
where \( \beta \) is the coefficient that controls the magnitude of margin and the term \( -\sum_{l \in S_i} p_{i,l} \log p_{i,l} \) is the minimal value of cross-entropy. \( -\sum_{l \in S_i} p_{i,l} \log q_{i,l} \). With the margin, the KL-divergence is relaxed by:
\[
D'_{KL}(p_i || q_i) = [D_{KL}(p_i || q_i) - \delta_i],
\]
where \([\cdot]_+\) indicates the hinge loss. Minimizing the relaxed KL-divergence \( D'_{KL}(p_i || q_i) \) indicates the cross-entropy should be as small as \(-\sum_{l \in S_i} p_{i,l} \log p_{i,l} \) if it is smaller than \(-(1 + \beta) \sum_{l \in S_i} p_{i,l} \log q_{i,l} \). With the selection and relaxation, the loss term \( L_{dist}(\theta^t, w^t; O'i) \) can be reformulated by:
\[
L_{dist}(\theta^t, w^t; O'i) = -\frac{1}{n} \sum_{i=1}^{n} D'_{KL}(p_i || q_i).
\]

4.4. Learning Algorithm

The overall objective function combines the classification loss (Eq. 1) and flexible knowledge distillation loss (Eq. 7) by a balance weight \( \lambda_0 \), which is used to optimize \( \theta^t, w^t, w^t; \)
\[
L = L_{\text{new}}(\theta^t, w^t; O'i) + \lambda_0 L_{dist}(\theta^t, w^t; O'i).
\]

Algorithm 1 shows the main steps for training the new model and classifiers. First, we initialize the new model \( \theta^t \) and classifiers \( w^t \) by copying weights from the old model \( \theta^{-1} \) and classifiers \( w^{-1} \). As the number of new classifiers increases, we randomly initialize the added weights. For a batch of training data, we compute the classification loss \( L_{\text{new}} \) using Eq. 1. Then, for each datum \( (x'_i, y'_i) \), we do the flexible knowledge distillation. The activation \( a_i = \{a_{i,1}, a_{i,2}, ..., a_{i,L}\} \) and \( b_i = \{b_{i,1}, b_{i,2}, ..., b_{i,L}\} \) are produced by the old and new models. The valid units in the neighborhood \( S_i \) are selected as the top \( K \) ones in \( a_i \), and the rest units are ignored. We also select the corresponding units for \( b_i \) based on \( S_i \). With the selected units, we compute the probabilities \( p_i \) and \( q_i \) using Eq. 3. Then, KL divergence \( D_{KL} \) and margin \( \delta_i \) are calculated using Eq. 4 and 5 for obtaining the relaxed KL divergence \( D'_{KL} \) using Eq. 6. The distillation loss \( L_{dist} \) for the batch is calculated by Eq. 7. The final loss \( L \) is calculated as the weighted combination of \( L_{\text{new}} \) and \( L_{dist} \) using Eq. 8. With loss \( L \), we update \( \theta^t \) and \( w^t \) by back-propagation.

5. Experiments

5.1. Implementation Details

We use ResNet-50 as the backbone for all experiments. The temperature \( T \) is set to be 2 for all knowledge distillation based methods in experiments. The balance weight
\(\lambda_s\) is chosen from \(10^{(-2,-1,0,1)}\) for different methods individually. We use hold-off validation data to determine the two hyper-parameters (K and \(\beta\)), we first select the best \(K\) without CR module, then we choose the best \(\beta\) based on the selected \(K\). The details will be given in Sec. 5.4. The classification/retrieval of testing data is based on the similarity (Euclidean distance) of feature embeddings. For all experiments, we repeat five times and report the mean value and standard deviation. We run experiments on four NVIDIA GTX1080-Ti GPUs.

**Continual Face Recognition.** All images are aligned and then resized to \(112 \times 112\). The feature dimension is 256, and the batch size is 384. SGD optimizer with initial learning rate \(10^{-4}\) is used in face experiments. We follow the face testing protocol in [10,27]. For 5-step setting, we train the model for 20000 iterations in each learning step, and the learning rate is reduced by \(\times 0.1\) at 8000th and 16000th iteration. For 10-step setting, the model is trained for 10000 iterations and the learning rate is reduced by \(\times 0.1\) at 4000th and 8000th iteration.

**Continual Person Re-id.** All images are resized to \(256 \times 128\). The feature dimension is 2,048, and the batch size is 256. We use the popular person re-id testing protocol [40]. In each training batch, we randomly select 64 identities and sample 4 images for each identity. Adam optimizer with learning rate \(3.5 \times 10^{-4}\) is used. We train the model for 50 epochs, and we decrease the learning rate by \(\times 0.1\) at the 25th and 35th epoch.

### 5.2. Preliminary Experiment

We first give a simple preliminary experiment to illustrate that Catastrophic Forgetting of old classes is not the main problem of CRL and the performance on old classes is not suitable to evaluate CRL methods. In this experiment, the model is continually finetuned without any regularization, on 5 subsets of CRL-face, and evaluated on the hold-off testing data of the first subset (Step0). In other words, the finetuned model is always evaluated on classes of Step0. According to Table 2, the performance on classes of the first subset does not show obvious decrease after finetuning on other subsets. The performance even increased on Step1 due to learning new classes. It means that CRL models suffer little from Catastrophic Forgetting of old classes. Hence, it is more suitable to evaluate model’s generalization ability on unseen classes in the CRL setting.

<table>
<thead>
<tr>
<th></th>
<th>Step0</th>
<th>Step1</th>
<th>Step2</th>
<th>Step3</th>
<th>Step4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finetune</td>
<td>93.38</td>
<td>95.05</td>
<td>94.20</td>
<td>94.08</td>
<td>93.82</td>
</tr>
</tbody>
</table>

**Table 2:** The performance (%) on old classes on CRL-face dataset. The training set of CRL-face is split into 5 subsets with 17,148 classes per subset. The performances are evaluated on classes of the first subset (Step0).

### 5.3. Comparison to the State-of-the-art

Our goal is not improving SOTA face recognition or person reid performance, instead we aims to extend continual learning to biometric identification and propose a scalable CL method. Unless otherwise stated, no extra memory is provided to store old data. In this main setting, we compare to four popular CL methods with the same backbone and classification loss, namely, Baseline, Finetune, LFL [24] and LWF [36]. As our method is compatible with any memory rehearsal mechanism, we also compare to two memory based methods GSS [4] and GDumb [47] by integrating our method with the memory rehearsal mechanism in [47].

We choose LFL and LWF as competitors because they are efficient enough for large-scale training on 5.8M images and 86K classes, while those generative model [54, 70, 71], meta-learning [18, 20, 23, 62] and dynamic-network [5, 44, 76] based methods are not salable or efficient to train on such large benchmarks. As shown in the most recent works [16, 32, 78], LWF is still a competitive method under the same memory setting. Baseline means that the model is trained from scratch (without old model) in every learning step. In this method, the model totally forgets knowledge learned from old classes. Finetune is a naive continual learning method in which the model is updated by finetuning the old model on new classes. LFL aims to restrict the difference of features produced by old and new models. In this way, the new model can produce similar features like the old model. LWF is based on knowledge distillation, which minimizes the cross-entropy between the outputs of old and new models. We also provide the upper-bound of each experimental setting. The upper-bound is calculated by jointly training on all data (of all steps). The performance in the final learning step is given in Table 3 and 4, while the detailed results of every learning step are illustrated in Figure 3 and 4. Comparison to memory based methods is shown in Table 5.

**Continual Face Recognition.** Table 3 shows the final results of our method on LFW and Megaface, compared to the state-of-the-art. Clearly, on the same dataset, performance of 10-step learning is worse than that of 5-step learning, because of fewer training data per step. Generally speaking, our method outperforms all other methods in all settings. Especially, when tested on Megaface, ours overwhelms the runner-up (LWF) by 1.01% and 1.25% on 5-step and 10-step settings respectively. As 5-step learning on LFW dataset is easy, Finetune also achieves good performance. However, on harder setting (10-step learning) and dataset (Megaface), the gap between Finetune and others widens. Figure 3 (a) and (b) illustrate the performance evaluated on Megaface in every learning step. We find that, except Baseline, performance of all methods increases after learning more classes. Our method shows obvious advantages compared to others.
Continental Person Re-id. Table 4 shows Top1 and mAP performance of 5-step and 10-step learning settings on the proposed dataset. The gap between different methods is obvious. Our method outperforms the runner-up (LWF) by around 2% on all settings. Compared to Baseline, our method improves the performance by 13.9% (Top1) and 12.5% (mAP) on 10-step learning, which means our method effectively leverages knowledge from old classes. However, our results are still obviously lower than the upper-bound. The gap indicates the challenging of continental person re-id on the proposed benchmark. Figure 4 (a) and (b) illustrate the performance (Top1 and mAP) of different methods in every learning step. Our method shows obvious advantage compared to other methods, especially on 10-step learning.

Comparison to SOTA with memory. Our method is also compatible with memory based rehearsal, and the performance on unseen testing set can be improved. We compare to the recent works GSS [4] and GDumb [47] with a fixed size of memory (25000 images) under the setting of 5-step face recognition evaluated on Megaface. We use the memory strategy presented in [47]. We simplify GSS and compute the similarity of two images based on gradients of the compute the similarity of two images based on gradients of the final layer instead of the whole network, because of its huge computational cost on large-scale datasets. Table 5 shows that our method outperforms GSS [4] and GDumb [47] by 0.59% and 2.37% after 5-step learning.

5.4. Ablation Study

Effectiveness of Proposed Modules. We do ablation study on two modules of the proposed method, namely, Neighborhood Selection (NS) and Consistency Relaxation (CR). To verify the effectiveness of two modules, we compare the four variants of our method: (1) Basic: Plain knowledge distillation without NS or CR; (2) Basic+NS: Basic with Neighborhood Selection; (3) Basic+CR: Basic with Consistency Relaxation; (4) Basic+NS+CR (Ours): Basic with both Neighborhood Selection and Consistency Relaxation.

We do ablation study on 5-step continual face recognition and person re-id. Table 6 and 7 show the results on the two benchmarks. Clearly, both NS and CR modules benefit the final performance. The improvement is obvious in continental person re-id. By adding NS module, Basic+NS overwhelms Basic by 0.7% of Top1 and 0.7% of mAP. Meanwhile, Basic+CR outperforms Basic by 2.1% of Top1 and 2.6% of mAP. In continual face recognition, Basic+NS+CR outperforms Basic+NS by 0.13% and 0.83% on LFW and Megaface respectively.

Figure 3 (c) shows the results of four variants in continual face recognition. NS module may hinder knowledge transfer in first three steps because of fewer distillation bases (selected old classes). Finally, Basic+NS+CR outperforms Basic. Basic+NS+CR has the best performance compared to other variants. Figure 4 (c) illustrates how the two modules influence continental person re-id performance. We find that NS module of Basic+NS stably improves the performance compared to Basic. In addition, CR module also significa-ably and stably promotes the performance. Although ours (Basic+NS+CR) is slightly weaker (≤ 0.2% in the final step) than Basic+CR in continual person re-id, ours is more suitable for large-scale continental learning because of its better scalability and efficiency. This priority will be further discussed in Sec. 6.

Sensitivity of Hyper-parameters. We further analyze the sensitivity of performance w.r.t. the two key hyper-parameters, namely, $K$: neighborhood size and $\beta$: margin magnitude. These experiments are based on 5-step continual person re-id.

Neighborhood Size. First, we change $K$ in Basic+NS
the continual person re-id is less than 10%. As shown in Table 8, when \( K = 200 \), Basic+NS achieves the best performance. The best \( K \) is about 10% of the number of all old classes in the final step.

### Margin Magnitude
For simplicity, we analyze the performance of Basic+NS+CR with fixed \( K \) and \( \beta \) varying. The range of \( \beta \) is \( \{0.2, 5, 10, 50\} \times 10^{-3} \). According to Table 8, \( \beta = 10^{-2} \) is the best parameter when \( K = 200 \). Overall, the changing of performance w.r.t. \( K \) and \( \beta \) is smooth.

### 6. Discussion & Conclusion

#### Scalability & Efficiency
In continual learning, the old classes accumulate quickly along with more learning steps. Especially, in face recognition and person re-id, thousands even millions of identities are involved in real applications.

#### Conclusion & Future Work
In this paper, we propose the continual representation learning for biometric identification with two large-scale benchmarks. Flexible knowledge distillation with Neighborhood Selection and Consistency Relaxation modules are proposed for better scalability and flexibility in large-scale continual learning. Extensive experiments show that our method outperforms the state-of-the-art on two benchmarks. Effectiveness of the two modules is verified by ablation study. In the future, more effort should be devoted to improving the generalization ability and scalability of continual learning models in large-scale real-world applications.

#### Acknowledgement
This work is partially supported by China Scholarships Council (Student No. 201806010331).
References


[29] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In...


