Continual Representation Learning for Biometric Identification

Bo Zhao*,[1,2], Shixiang Tang*,[2,3], Dapeng Chen[2], Hakan Bilen[1], Rui Zhao[2]
[1]The University of Edinburgh, [2]SenseTime Group Limited, [3]The 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.

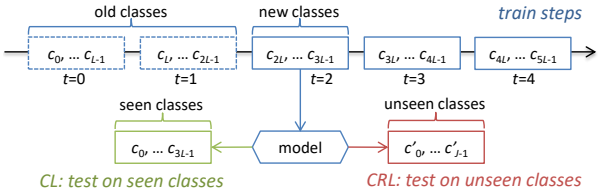

Figure 1: The proposed continual representation learning (CRL) v.s. the traditional continual learning (CL). The model is trained online on newly obtained tasks without access to old classes. Our CRL aims to learn better representation that is generalized to unseen classes, while traditional CL aims to learn and remember more old classes.

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 classification 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 [6, 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 [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-

	Task	Scale	Concept	Classes			Images		
				Train	Test	Total	Train	Test	Total
MNIST [31]	Cls	Small	number	10	10	10	60,000	10,000	70,000
CORe50 [38]			coarse-grained objects	50	50	50	120,000	45,000	165,000
CIFAR-100 [30]			coarse-grained objects	100	100	100	50,000	10,000	60,000
CUB [63]			fine-grained birds	200	200	200	5,994	5,794	11,788
ImageNet [51]		Large	coarse-grained objects	1,000	1,000	1,000	1,281,167	50,000	1,331,167
CRL-face	Rpt		fine-grained face	85,738	5,829	91,567	5,783,772	4,000	5,787,772
CRL-person			fine-grained person	2,494	4,512	7,006	59,706	30,927	90,633

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

marks are small-scale in terms of class ($\leq$ 200) and image ($\leq$ 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 $c_{2L}, \dots c_{3L-1}$. Without accessing old classes $c_0, \dots c_{2L-1}$, the model will gradually forget the knowledge obtained from previous learning steps. In each step, CRL tests the model on previously unseen testing classes $c'_0, \dots c'_{J-1}$ 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 ($\leq$ 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.

CRL-face Benchmark Continual face recognition requires large-scale training data for each learning step. Ms1M dataset [15] is a suitable option for constructing CRL-face benchmark, because there are 85,738 identities and 5,783,772 images in the dataset. We divide identifies 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

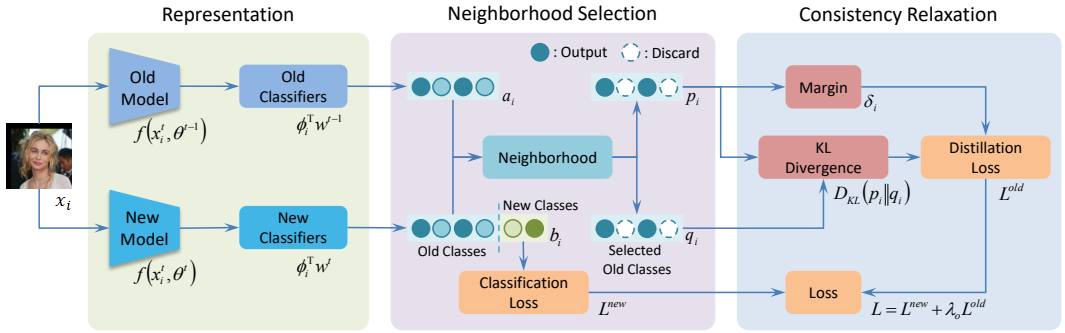

Figure 2: Illustration of the proposed method. Our method consists of three modules, namely representation, neighborhood selection and consistency relaxation. Representation: The image x_i^t is fed into both old and new models, $f(x_i^t, \theta^{t-1})$ and $f(x_i^t, \theta^t)$, followed by corresponding old and new classifiers $\mathbf{w}^{t-1}$ and $\mathbf{w}^t$. The activation of old and new models, $\mathbf{a}_i$ and $\mathbf{b}_i$, are produced. We use the activation of new classes produced by the new model to calculate the classification loss $\mathcal{L}^{new}$. Neighborhood selection: We determine the neighborhood of the given datum based on activation of old model $\mathbf{a}_i$ and choose top ones of old and new models in the neighborhood to calculate KL divergence $\mathcal{D}_{KL}(\mathbf{p}_i||\mathbf{q}_i)$. Consistency Relaxation: The margin δ_i is introduced to KL divergence for consistency relaxation, and the relaxed KL divergence is produced as the distillation loss $\mathcal{L}^{old}$. The overall loss $\mathcal{L}$ is the weighted combination of the classification loss $\mathcal{L}^{new}$ and distillation loss $\mathcal{L}^{old}$.

of 4,512 identities. Mean average precision (mAP) and top1 accuracy are reported in each learning step.

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

$$\begin{aligned} &\mathcal{L}^{new}(\theta^t, \mathbf{w}^t; \mathcal{O}^t) = \frac{1}{n}\sum_{i=1}^{n} l(\phi_i, y_i^t, \theta^t, \mathbf{w}^t), \\ &\phi_i = f(x_i^t, \theta^t), \\ &l(\phi_i, y_i^t, \theta^t, \mathbf{w}^t) = -\log \frac{\exp(\phi_i^\top \mathbf{w}_{y_i^t}^t)}{\sum_{j=1} \exp(\phi_i^\top \mathbf{w}_j^t)}. \end{aligned} \tag{1}$$

w_j^t indicates the classifier for the jth class. Obviously, minimizing the loss in Eq. 1 will result in overfitting to the instances in $\mathcal{O}^t$. 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 $\mathbf{w}^{t-1}$ 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.

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 $\mathbf{w}^{t-1}$ is about L classes, the output probability of y_i^t generated by the old model $f(x_i^t, \theta^{t-1})$ given $\mathbf{w}^{t-1}$ is: $\mathbf{p}_i = \{p_{i,1}, p_{i,2}, ..., p_{i,L}\}$. The cross-entropy loss is utilized to regularize the new probability $\mathbf{q}_i = \{q_{i,1}, q_{i,2}, ..., q_{i,L}\}$ generated by the new model $f(x_i^t, \theta^t)$:

$$\mathcal{L}^{old}(\theta^t, \mathbf{w}^p; \mathcal{O}^t) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L} p_{i,l} \log q_{i,l}, \tag{2}$$

where l is the index of the class.

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 Flexiable Knowledge Distillation(FKD), where we perform neighbour selection and consistency relaxation over the loss related to the old model $\mathcal{L}^{old}$. The proposed method is illustrated in Figure 2.

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:
$\mathcal{O}^t = \{(x_i^t, y_i^t)\}_{i=1}^n$: training data in current learning step;
θ^{t-1} : old model; w^{t-1} : old classifiers;
Ensure:
θ^t : new model; w^t : new classifiers;
Initialize θ^t, w^t by θ^{t-1}, w^{t-1};
for a batch in $\mathcal{O}^t$ **do**
  Compute classification loss $\mathcal{L}^{new}$ using Eq. 1;
  for (x_i^t, y_i^t) in the batch **do**
    Determinate the neighborhood $\mathcal{S}_i$ based on activations $\mathbf{a}_i$ (old) and $\mathbf{b}_i$ (new);
    Compute relaxed KL Divergence $\mathcal{D}'_{KL}$ using Eq. 6;
  Compute distillation loss $\mathcal{L}^{old}$ using Eq. 7 and the final loss $\mathcal{L}$ using Eq. 8;
  Update θ^t and w^t by back-propagation with $\mathcal{L}$.

direct output of the classifiers) by a softmax layer. When the number of classifies 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^t, the activation generated by the old model is denoted by $\mathbf{a}_i = \{a_{i,1}, a_{i,2}, ..., a_{i,L}\}$ and the activation of the new model is denoted by $\mathbf{b}_i = \{b_{i,1}, b_{i,2}, ..., b_{i,L}\}$. We rank the activation units of the old model ($\mathbf{a}_i$) with descending order, select the top K ones, and put their indices into the set $\mathcal{S}_i$, *i.e.*, the neighborhood of the ground-truth class y_i^t. The probabilities generated by the old and new model are $\mathbf{p}_i$ and $\mathbf{q}_i$, which are calculated based on the selected label set $\mathcal{S}_i$:

$$p_{i,l} = \frac{\exp(a_{i,l}/T)}{\sum_{j\in\mathcal{S}_i}\exp(a_{i,j}/T)}, \quad q_{i,l} = \frac{\exp(b_{i,l}/T)}{\sum_{j\in\mathcal{S}_i}\exp(b_{i,j}/T)}, \tag{3}$$

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 $\mathbf{p}_i$ and $\mathbf{q}_i$:

$$\mathcal{D}_{KL}(\mathbf{p}_i||\mathbf{q}_i) = \sum_{l\in\mathcal{S}_i}(p_{i,l}\log p_{i,l} - p_{i,l}\log q_{i,l}). \tag{4}$$

As $\sum_{l\in\mathcal{S}_i} p_{i,l}\log p_{i,l}$ is a constant in the optimization, the KL-divergence is equivalent to the cross-entropy in Eq. 2, and $\mathcal{D}_{KL}(\mathbf{p}_i||\mathbf{q}_i)$ will be 0 if $\mathbf{p}_i$ and $\mathbf{q}_i$ 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 δ_i to relax the consistency constraint. The margin for $\mathbf{x}_i$ is set to be:

$$\delta_i = -\beta\sum_{l\in\mathcal{S}_i} p_{i,l}\log p_{i,l}, \tag{5}$$

where β is the coefficient that controls the magnitude of margin and the term $-\sum_{l\in\mathcal{S}_i} p_{i,l}\log p_{i,l}$ is the minimal value of cross-entropy $-\sum_{l\in\mathcal{S}_i} p_{i,l}\log q_{i,l}$. With the margin, the KL-divergence is relaxed by:

$$\mathcal{D}'_{KL}(\mathbf{p}_i||\mathbf{q}_i) = [\mathcal{D}_{KL}(\mathbf{p}_i||\mathbf{q}_i) - \delta_i]_+, \tag{6}$$

where $[\cdot]_+$ indicates the hinge loss. Minimizing the relaxed KL-divergence $\mathcal{D}'_{KL}(\mathbf{p}_i||\mathbf{q}_i)$ indicates the cross-entropy should be as small as $-\sum_{l\in\mathcal{S}_i} p_{i,l}\log p_{i,l}$ until it is smaller than $-(1+\beta)\sum_{l\in\mathcal{S}_i} p_{i,l}\log p_{i,l}$. With the selection and relaxation, the loss term $\mathcal{L}^{old}(\theta^t, \mathbf{w}^p; \mathcal{O}^t)$ can be reformulated by:

$$\mathcal{L}^{old}(\theta^t, \mathbf{w}^t; \mathcal{O}^t) = -\frac{1}{n}\sum_{i=1}^{n}\mathcal{D}'_{KL}(\mathbf{p}_i||\mathbf{q}_i) \tag{7}$$

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 λ_0, which is used to optimize $\theta^t, \mathbf{w}^t, \mathbf{w}^p$:

$$\mathcal{L} = \mathcal{L}^{new}(\theta^t, \mathbf{w}^t; \mathcal{O}^t) + \lambda_o\mathcal{L}^{old}(\theta^t, \mathbf{w}^t; \mathcal{O}^t). \tag{8}$$

Algorithm 1 shows the main steps for training the new model and classifiers. First, we initialize the new model θ^t and classifiers w^t by copying weights from the old model θ^{t-1} and classifiers w^{t-1}. As the number of new classifiers increases, we randomly initialize the added weights. For a batch of taining data, we compute the classification loss $\mathcal{L}^{new}$ using Eq. 1. Then, for each datum (x_i^t, y_i^t), we do the flexible knowledge distillation. The activation $\mathbf{a}_i = \{a_{i,1}, a_{i,2}, .., a_{i,L}\}$ and $\mathbf{b}_i = \{b_{i,1}, b_{i,2}, .., b_{i,L}\}$ are produced by the old and new models. The valid units in the neighborhood $\mathcal{S}_i$ are selected as the top K ones in $\mathbf{a}_i$, and the rest units are ignored. We also select the corresponding units for $\mathbf{b}_i$ based on $\mathcal{S}_i$. With the selected units, we compute the probabilities $\mathbf{p}_i$ and $\mathbf{q}_i$ using Eq. 3. Then, KL divergence $\mathcal{D}_{KL}$ and margin δ_i are calculated using Eq. 4 and 5 for obtaining the relaxed KL divergence $\mathcal{D}'_{KL}$ using Eq. 6. The distillation loss $\mathcal{L}^{old}$ for the batch is calculated by Eq. 7. The final loss $\mathcal{L}$ is calculated as the weighted combination of $\mathcal{L}^{new}$ and $\mathcal{L}^{old}$ using Eq. 8. With loss $\mathcal{L}$, we update θ^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

λ_o 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 β), we first select the best K without CR module, then we choose the best β 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 112×112. The feature dimension is 256, and the batch size is 384. SGD optimizer with initial learning rate 10^{-2} 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 256×128. 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 3.5×10^{-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.

	Step0	Step1	Step2	Step3	Step4
Finetune	93.38	95.05	94.20	94.08	93.82

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.

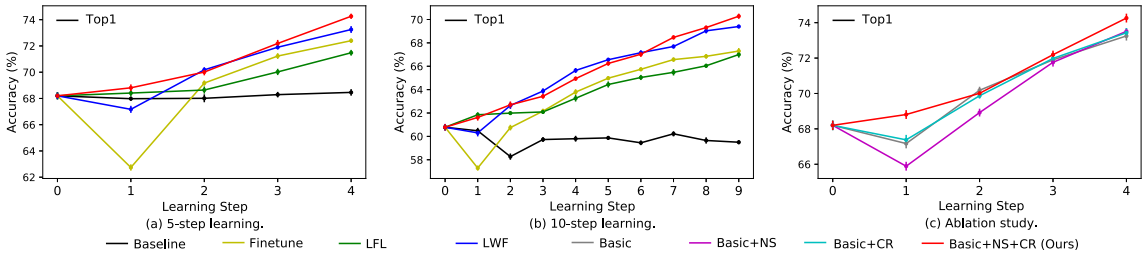

Figure 3: Experimental results on continual face recognition (tesed on Megaface). Top 1 accuracy (%) is reported. Figure (a) and (b) are 5-step and 10-step continual learning. We compare our method to Baseline, Finetune, LFL and LWF. Figure (c) is the ablation study. We compare the variants of our method with/without Neighborhood Selection (NS) and Consistency Relaxation (CR) modules.

		Baseline	Finetune	LFL	LWF	Ours	Upper-bound
5-step	LFW	98.85 ± 0.02	99.00 ± 0.01	98.97 ± 0.02	98.95 ± 0.01	$\mathbf{99.10 \pm 0.01}$	99.42 ± 0.01
	Megaface	68.20 ± 0.21	72.40 ± 0.12	71.48 ± 0.23	73.25 ± 0.22	$\mathbf{74.26 \pm 0.23}$	82.93 ± 0.12
10-step	LFW	98.52 ± 0.02	98.72 ± 0.04	98.65 ± 0.02	98.82 ± 0.02	$\mathbf{99.05 \pm 0.01}$	99.42 ± 0.01
	Megaface	60.78 ± 0.22	66.85 ± 0.21	67.00 ± 0.12	69.03 ± 0.24	$\mathbf{70.28 \pm 0.13}$	82.93 ± 0.12

Table 3: Comparison to SOTA: face recognition. Top 1 accuracy (%) in the final learning step. Upper-bound means joint training.

Continual 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 continual 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 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 continual person re-id. By adding NS module, Basic+NS overwhelms Basic by 0.7% of Top1 and 0.7% of mAP. Meantime, 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 outperforms Basic. Basic+NS+CR has the best performance compared to other variants. Figure 4 (c) illustrates how the two modules influence continual person re-id performance. We find that NS module of Basic+NS stably improves the performance compared to Basic. In addition, CR module also significantly and stably promotes the performance. Although ours (Basic+NS+CR) is slightly weaker ($\leq 0.2\%$ in the final step) than Basic+CR in continual person re-id, ours is more suitable for large-scale continual learning because of its better scalability and efficiency. This priority will be further discussed in Sec. 6.

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

Neighborhood Size. First, we change K in Basic+NS

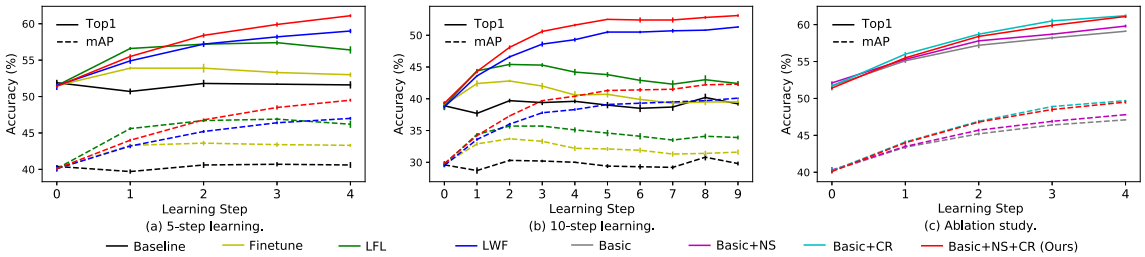

Figure 4: Experimental results on continual person re-id. Top 1 accuracy and mAP (%) are reported. Figure (a) and (b) are 5-step and 10-step continual learning. We compare our method to Baseline, Finetune, LFL and LWF. Figure (c) is the ablation study. We compare the variants of our method with/without Neighborhood Selection (NS) and Consistency Relaxation (CR) modules.

		Baseline	Finetune	LFL	LWF	Ours	Upper-bound
5-step	Top1	51.6 ± 0.5	53.0 ± 0.3	56.4 ± 0.5	59.0 ± 0.3	$\mathbf{61.1 \pm 0.2}$	75.5 ± 0.1
	mAP	40.6 ± 0.4	43.3 ± 0.2	46.2 ± 0.5	47.0 ± 0.2	$\mathbf{49.5 \pm 0.2}$	64.6 ± 0.1
10-step	Top1	39.2 ± 0.3	39.5 ± 0.6	42.4 ± 0.4	51.3 ± 0.1	$\mathbf{53.1 \pm 0.2}$	75.5 ± 0.1
	mAP	29.8 ± 0.3	31.6 ± 0.4	33.9 ± 0.3	40.1 ± 0.1	$\mathbf{42.3 \pm 0.2}$	64.6 ± 0.1

Table 4: Comparison to SOTA: person re-id. Top 1 and mAP accuracy (%) in the final learning step. Upper-bound means joint training.

	Step0	Step1	Step2	Step3	Step4
GSS [4]	68.20	**71.00**	74.52	75.42	77.94
GDumb [47]	68.20	69.80	73.53	75.23	76.16
Ours	68.20	70.20	**74.66**	**76.54**	**78.53**

Table 5: Comparison to SOTA with memory. Top 1 accuracy (%) on Megaface at each learning step. Std deviations are around ± 0.22 %. Ours outperforms GSS and GDumb by 0.59% and 2.37% after 5-step learning.

	Basic	Basic+NS	Basic+CR	Basic+NS+CR (Ours)
LFW	98.95 ± 0.02	99.05 ± 0.02	98.97 ± 0.01	99.10 ± 0.01
Megaface	73.25 ± 0.31	73.51 ± 0.22	73.43 ± 0.24	74.26 ± 0.22

Table 6: Ablation study: face recognition. The Top 1 accuracy (%) of different variants of our method, *i.e.*, with/without NS and CR modules.

	Basic	Basic+NS	Basic+CR	Basic+NS+CR (Ours)
Top1	59.1 ± 0.1	59.8 ± 0.2	61.2 ± 0.2	61.1 ± 0.2
mAP	47.1 ± 0.1	47.8 ± 0.1	49.7 ± 0.1	49.5 ± 0.2

Table 7: Ablation study: person re-id. The Top 1 accuracy (%) and mAP of different variants of our method, *i.e.*, with/without NS and CR modules.

which only includes the neighborhood selection module. The range of K is $\{0, 20, 200, 500, 1000\}$. If the number of old classes in current step is less than K, all old classes will be used. 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 varying β. The range of β 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 β is smooth.

6. Discussion and 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.

K	0	20	200	500	1000
Top1	59.1 ± 0.1	59.1 ± 0.3	$\mathbf{59.8 \pm 0.2}$	59.1 ± 0.3	58.9 ± 0.4
mAP	47.1 ± 0.1	47.4 ± 0.1	$\mathbf{47.8 \pm 0.1}$	47.4 ± 0.1	47.0 ± 0.2
$\beta \times 10^{-3}$	0	2	5	10	50
Top1	59.8 ± 0.2	59.7 ± 0.3	60.7 ± 0.4	$\mathbf{61.1 \pm 0.2}$	59.4 ± 0.2
mAP	47.8 ± 0.1	47.9 ± 0.2	49.0 ± 0.3	$\mathbf{49.5 \pm 0.2}$	48.4 ± 0.2

Table 8: The sensitive analysis of performance (%) w.r.t hyper-parameters K and β. The upper results are based on Basic+NS with varying K. The lower results are based on Basic+NS+CR with fixed $K = 200$ and varying β ($\times 10^{-3}$). The results are of 5-step continual person re-id.

If we use constant K as the neighborhood size, the memory and time cost for back-propagation of popular methods (*e.g.* LWF [36], iCaRL [49] and End2End [6]) which use all old classes for knowledge distillation is $\mathcal{O}(t)$ times of ours, and it increases along with step t. If we use a constant ratio $\frac{1}{r}$ of old classes, where $r > 1$, their memory and time cost is $\mathcal{O}(r)$ times of ours. Although we need do feed-forward on all old classes for neighborhood selection, it is not time-consuming compared to back-propagation. Besides, feed-forward can be implemented on CPU with RAM which is $\times 10$ to $\times 1000$ larger than GPU memory.

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

[1] David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In *International Conference on Machine Learning*, pages 10–19, 2018.

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

[3] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018.

[4] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*, pages 11816–11825, 2019.

[5] Rahaf Aljundi, Marcus Rohrbach, and Tinne Tuytelaars. Selfless sequential learning. *ICLR*, 2019.

[6] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 233–248, 2018.

[7] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2017.

[8] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2016.

[9] Edwin D de Jong. Incremental sequence learning. *ICLR*, 2017.

[10] 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.

[11] Rao Dushyant, Visin Francesco, A. Rusu Andrei, Whye Teh Yee, Pascanu Razvan, and Hadsell Raia. Continual unsupervised representation learning. In *Proceedings of the Conference on Neural Information Processing Systems*, 2019.

[12] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, pages 128–135, 1999.

[13] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

[14] Linting Guan, Yan Wu, Junqiao Zhao, and Chen Ye. Learn to detect objects incrementally. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 403–408. IEEE, 2018.

[15] 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.

[16] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu. Incremental learning in online scenario. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13926–13935, 2020.

[17] 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.

[18] Xu He, Jakub Sygnowski, Alexandre Galashov, Andrei A Rusu, Yee Whye Teh, and Razvan Pascanu. Task agnostic continual learning via meta learning. *arXiv preprint arXiv:1906.05201*, 2019.

[19] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[20] Wenpeng Hu, Zhou Lin, Bing Liu, Chongyang Tao, Zhengwei Tao, Jinwen Ma, Dongyan Zhao, and Rui Yan. Overcoming catastrophic forgetting for continual learning via model adaptation. 2018.

[21] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[22] Anil Jain, Lin Hong, and Sharath Pankanti. Biometric identification. *Communications of the ACM*, 2000.

[23] Khurram Javed and Martha White. Meta-learning representations for continual learning. In *Advances in Neural Information Processing Systems*, pages 1818–1828, 2019.

[24] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. *arXiv preprint arXiv:1607.00122*, 2016.

[25] Woo-Young Kang and BT Zhang. Continual learning with generative replay via discriminative variational autoencoder, 2018.

[26] Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Continual reinforcement learning with complex synapses. *ICML*, 2018.

[27] 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.

[28] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[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

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1931–1939, 2015.

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[31] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

[32] Janghyeon Lee, Donggyu Joo, Hyeong Gwon Hong, and Junmo Kim. Residual continual learning. *AAAI*, 2020.

[33] Timothée Lesort, Hugo Caselles-Dupré, Michael Garcia-Ortiz, Andrei Stoian, and David Filliat. Generative models from the perspective of continual learning. *arXiv preprint arXiv:1812.09111*, 2018.

[34] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[35] Wei-Hong Li, Zhuowei Zhong, and Wei-Shi Zheng. One-pass person re-identification by sketch online discriminant analysis. *Pattern Recognition*, 93:237–250, 2019.

[36] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2018.

[37] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[38] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. *arXiv preprint arXiv:1705.03550*, 2017.

[39] David Lopez-Paz et al. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476, 2017.

[40] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[41] Niki Martinel, Abir Das, Christian Micheloni, and Amit K Roy-Chowdhury. Temporal model adaptation for person re-identification. In *European Conference on Computer Vision*, pages 858–877. Springer, 2016.

[42] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

[43] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[44] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2554–2563. JMLR. org, 2017.

[45] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019.

[46] Anastasia Pentina and Christoph Lampert. A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pages 991–999, 2014.

[47] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, 2020.

[48] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

[49] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[52] Raul Sanchez-Reillo, Carmen Sanchez-Avila, and Ana Gonzalez-Marcos. Biometric identification through hand geometry measurements. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1168–1171, 2000.

[53] 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.

[54] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, pages 2990–2999, 2017.

[55] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[56] Daniel L Silver, Qiang Yang, and Lianghao Li. Lifelong machine learning systems: Beyond learning algorithms. In *2013 AAAI spring symposium series*, 2013.

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[58] Shagun Sodhani, Sarath Chandar, and Yoshua Bengio. On training recurrent neural networks for lifelong learning. *arXiv preprint arXiv:1811.07017*, 2018.

[59] Yuke Sun, Hong Liu, and Qianru Sun. Online learning on incremental distance metric for person re-identification. In *2014 IEEE International Conference on Robotics and Biomimetics*, pages 1421–1426. IEEE, 2014.

[60] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[61] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision*, pages 791–808. Springer, 2016.

[62] Risto Vuorio, Dong-Yeon Cho, Daejoong Kim, and Jiwon Kim. Meta continual learning. *arXiv preprint arXiv:1806.06928*, 2018.

[63] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[64] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. Human-in-the-loop person re-identification. In *European conference on computer vision*, pages 405–422. Springer, 2016.

[65] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[66] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013.

[67] Kun Wei, Cheng Deng, and Xu Yang. Lifelong zero-shot learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 551–557, 2020.

[68] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[69] 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.

[70] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. In *Advances In Neural Information Processing Systems*, pages 5962–5972, 2018.

[71] Ye Xiang, Ying Fu, Pan Ji, and Hua Huang. Incremental learning using conditional adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6619–6628, 2019.

[72] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015.

[73] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2018.

[74] Ju Xu and Zhanxing Zhu. Reinforced continual learning. In *Advances in Neural Information Processing Systems*, pages 899–908, 2018.

[75] Bangpeng Yao, Gary Bradski, and Li Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3466–3473. IEEE, 2012.

[76] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *ICLR*, 2018.

[77] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3987–3995. JMLR. org, 2017.

[78] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1131–1140, 2020.

[79] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[80] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[81] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.