Generalising without Forgetting for Lifelong Person Re-Identification

Guile Wu, Shaogang Gong
Queen Mary University of London
guile.wu@qmul.ac.uk, s.gong@qmul.ac.uk

Abstract

Existing person re-identification (Re-ID) methods mostly prepare all training data in advance, while real-world Re-ID data are inherently captured over time or from different locations, which requires a model to be incrementally generalised from sequential learning of piecemeal new data without forgetting what is already learned. In this work, we call this lifelong person Re-ID, characterised by solving a problem of unseen class identification subject to continuous new domain generalisation and adaptation with class imbalanced learning. We formulate a new Generalising without Forgetting method (GwFReID) for lifelong Re-ID and design a comprehensive learning objective that accounts for classification coherence, distribution coherence and representation coherence in a unified framework. This design helps to simultaneously learn new information, distil old knowledge and solve class imbalance, which enables GwFReID to incrementally improve model generalisation without catastrophic forgetting of what is already learned. Extensive experiments on eight Re-ID benchmarks, CIFAR-100 and ImageNet show the superiority of GwFReID over the state-of-the-art methods.

Introduction

Person re-identification (Re-ID) aims at matching people across non-overlapping camera views. The development of deep learning and the availability of increasingly large-scale Re-ID datasets have significantly advanced person Re-ID in the past decade (Cheng et al. 2020; Song et al. 2019; Wu, Zhu, and Gong 2020). Existing Re-ID methods (Xiao et al. 2016; Yu, Wu, and Zheng 2017; Song et al. 2019) mostly assume that all training data can be prepared in advance for model learning. However, real-world Re-ID data (person images) are inherently captured over time or from different locations/domains, which requires a Re-ID model to be incrementally optimised from sequential learning of piecemeal new data. Potential solutions for solving this problem include fine-tuning a pre-trained model with sequentially inputted new data or assembling all the data (old and new) into a large data pool for joint-training from scratch. Although these solutions are easy to implement, the former leads to forgetting most previously learned knowledge, whilst the latter imposes a huge burden not only on data storage, but also on the need for learning from scratch over combined old and new data many times. This is both wasteful and non-scalable. In essence, a Re-ID model needs to continuously incorporate piecemeal new information while preserving old knowledge without assembling old data again in a lifelong learning process. We call this *lifelong person Re-ID*.

Intuitively, lifelong person Re-ID is related to incremental learning (Li and Hoiem 2018; Zhao et al. 2020) (also known as lifelong learning (Hou et al. 2018)), which aims to incrementally increase a model's capability by a flow of data rather than training a model with a fixed dataset all at once. However, in traditional incremental learning, all test classes are usually seen during training and/or share a common class space as the training data, so the main challenge is leveraging new and old information to optimise a model for a common set of known (seen) classes. In contrast, lifelong person Re-ID has significantly different and harder challenges: (1) Lifelong Re-ID is inherently a zero-shot learning problem (w/o semantic information) where training and test classes (person identities) are non-overlapping, so test classes are unseen in training; (2) Sequential data streams are from different domains with entirely different and new classes (person IDs), which increases the difficulty in balancing information from new and old domains; (3) Class imbalance is more challenging in lifelong Re-ID as compared to a shared common class space in conventional incremental learning, as each Re-ID dataset contains different numbers of person identities (non-overlapping) with different numbers of samples. Image classification datasets usually have balanced class sampling distributions, *e.g.* 100 classes in CIFAR-100 (Krizhevsky and Hinton 2009) all having 600 samples per class.

In this work, we propose a novel Generalising without Forgetting method (GwFReID) for lifelong person Re-ID. In GwFReID, we continuously optimise a Re-ID model to extract more generalisable feature representations for Re-ID at different domains without catastrophic forgetting. To implement this generalising without forgetting principle, our learning paradigm resembles the spirit of the human vision perceptual process. In the human vision system, we normally store in memory a few exemplars of classes observed so far. When we encounter new classes (unseen before), we improve our understanding of these new classes by leveraging both new and old information. In the same princi-

ple, GwFReID exploits both new data and memory exemplars (Rebuffi et al. 2017) of old data to simultaneously learn new knowledge, distil old knowledge and solve class imbalance with a comprehensive learning objective in an end-to-end trainable framework for lifelong Re-ID.

Our **contributions** are: **(1)** We introduce lifelong person Re-ID characterised by solving the problem of unseen class identification subject to continuous new domain generalisation and adaptation with class imbalanced learning, and formulate a novel Generalising without Forgetting (GwFReID) framework for lifelong Re-ID. **(2)** To simultaneously learn new information, distil old knowledge and solve class imbalance in lifelong Re-ID, we incorporate classification coherence, distribution coherence and representation coherence into a comprehensive learning objective for model generalisation learning. **(3)** Extensive experiments on eight person Re-ID benchmarks, CIFAR-100 (Krizhevsky and Hinton 2009) and ImageNet (Russakovsky et al. 2015) show the superiority of GwFReID against the state-of-the-art alternative methods.

Related Work

Lifelong Learning. Lifelong (incremental) learning is a learning strategy capable of continually upgrading a system with a flow of new data (stream), rather than learning once from a fixed set of data (batch) (Rebuffi et al. 2017; Zhao et al. 2020). A key challenge in lifelong learning is to minimise catastrophic forgetting, *i.e.* how to leverage new information for model updating while preserving old knowledge learned in the past. A popular solution is to distil existing capabilities about old knowledge from a frozen model using a modified cross-entropy loss (Li and Hoiem 2018; Rebuffi et al. 2017). Some recent works focus on exploiting cosine normalisation (Hou et al. 2019) for learning a unified classifier or utilising bias correction for estimating the bias in the last fully connected (FC) layer (Wu et al. 2019; Zhao et al. 2020). However, existing methods cannot be readily applied to lifelong Re-ID, because they mainly focus on incremental classifier learning which assumes that training and testing data cover the same class space (seen classes) largely from the same domain, while lifelong Re-ID requires to address unseen class recognition and classifiers are removed during testing (only the feature embedding model is used). In this work, we propose a new Generalising without Forgetting method for lifelong Re-ID. We formulate a unified framework with a comprehensive learning objective to incrementally optimise the feature embedding space for Re-ID matching without catastrophic forgetting of what is already learned.

Person Re-Identification. In the past decade, the development of deep learning and the emergence of large-scale datasets have significantly advanced person Re-ID (Song et al. 2019; Wei et al. 2018; Wu, Zhu, and Gong 2019). Existing Re-ID methods mostly assume that all training data (labelled or unlabelled) are prepared in advance for model learning. For example, in (Song et al. 2019), Song et al. assemble multiple Re-ID benchmarks to optimise a domain invariant mapping network for deployment. In (Xiao et al. 2016), Xiao et al. pretrain a Re-ID model using all available datasets and fine-tune this model on a target domain for deployment. However, real-world Re-ID data are inherently captured over time or from different locations, which poses new challenges to the conventional Re-ID learning. This requires a Re-ID model to be incrementally generalised without forgetting knowledge already learned. In (Sugianto et al. 2019), Sugianto et al. apply the learning without forgetting method (LwF) (Li and Hoiem 2018) for continuous learning in Re-ID, but their method is a straightforward application of LwF, failing to address the inherent challenge of domain incremental generalisation in Re-ID. In this work, we characterise the lifelong Re-ID problem by unseen class recognition, domain generalisation and class imbalanced learning. To address these problems, we propose a novel Generalising without Forgetting method and formulate a comprehensive learning objective to continuously optimise a generalised Re-ID model with sequential input data without forgetting knowledge already learned.

Knowledge Distillation. Knowledge distillation is an effective solution to transfer knowledge between models with different capabilities. In (Hinton, Vinyals, and Dean 2015), Hinton, Vinyals, and Dean compress the information in a large teacher model into a small student model using a distillation loss. In (Romero et al. 2015), Romero et al. use intermediate representations to compress a wider and shallower teacher model into a deeper and thinner student model. In (Li and Hoiem 2018), Li and Hoiem further demonstrate that the modified cross-entropy loss in knowledge distillation can be used for learning without forgetting in multi-task incremental learning. In our work, to regularise the distribution coherence between the new and old models, we also employ a modified cross-entropy loss. But different from existing methods, the proposed GwFReID aims at continually optimising a generalised Re-ID embedding model without forgetting knowledge already learned, rather than learning an incremental classifier or model compression. We formulate a new comprehensive learning objective through an end-to-end collaborative training procedure for lifelong Re-ID.

Methodology

Approach Overview

Fig. 1 shows an overview of GwFReID. We formulate lifelong Re-ID learning as a multi-class classification problem, *i.e.* each person identity is considered as a unique class. GwFReID incrementally learns discriminative information of new classes whilst preserves learned knowledge of old classes. Suppose we have a model M^o which is previously trained on training datasets X^o (with C^o classes). Based on herding selection (Welling 2009; Rebuffi et al. 2017), we construct an exemplar memory E^m by selecting representative exemplars from X^o using M^o, When new datasets X^n (with C^n classes) are available, we use $X = X^n \cup E^m$ as the input for model incremental training. We initialise the new model M^n=M^o and add C^n new output neurons (C^n new classes) to the last classification layer of M^n. Here, M^o plays a role of an expert dedicated to old knowledge, which is frozen during the learning process. As shown in Fig. 1, we

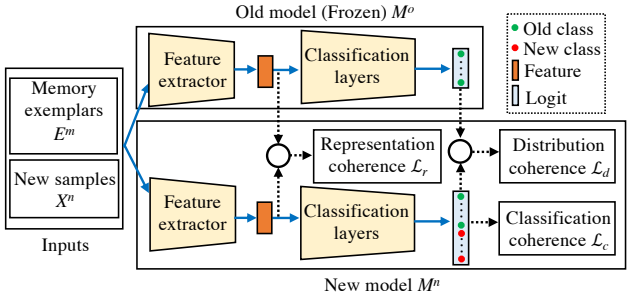


Figure 1: An overview of the proposed Generalising without Forgetting method for lifelong person Re-ID (GwFReID).

compute the features $\{v^o, v^n\}$ and logits $\{z^o, z^n\}$ of each sample x in X using both M^o and M^n. Note that, $\{v^o, z^o\}$ and $\{v^n, z^n\}$ denote outputs of $\{M^o, M^n\}$ rather than outputs of old and new samples. To simultaneously learn new information, distil old knowledge and solve class imbalance, the learning objective consists of three components: (1) The logit outputs z^n from M^n are employed to optimise a classification coherence loss $\mathcal{L}_c$; (2) The logit outputs $\{z^o, z^n\}$ from both M^o and M^n are utilised to learn a distribution coherence loss $\mathcal{L}_d$; (3) The feature representations $\{v^o, v^n\}$ from both M^o and M^n are used to optimise a representation coherence loss $\mathcal{L}_r$. Thus, the overall training objective $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_d + \mathcal{L}_r. \tag{1}$$

In Re-ID deployment, we use the feature extractor to get the feature representation of each person image and employ a generic distance metric $d(\cdot)$ (*e.g.* $L2$ distance) for Re-ID matching. In image classification, we use the latest model with a Nearest-Mean-of-Exemplars (NME) classifier (Rebuffi et al. 2017) to predict the label of each image based on the distance between an image to its nearest class mean.

Classification Coherence

Exemplar Memory. The exemplar memory is analogous to human memory in the vision perceptual process, which helps to improve understanding of new classes via leveraging both new and old information. In our work, we only reserve a tiny number of samples as the memory exemplars to minimise the memory consumption. Based on herding selection (Welling 2009; Rebuffi et al. 2017), we compute mean feature prototypes of each class and generate a sorted list of X^n based on the distance of each sample to the prototype in each class. Then, we select the top-K samples per class in each list as representative exemplars to update E^m.

Classification Coherence. To learn new information in a new model with new samples and reserved exemplars, cross-entropy loss $\mathcal{L}_{ce}$ is usually used for classification:

$$p_i = \frac{exp(z_i^n)}{\sum_{j=1}^{C^n+C^o} exp(z_j^n)}, \tag{2}$$

$$\mathcal{L}_{ce} = -\sum_{i=1}^{C^n+C^o} y_i log(p_i), \tag{3}$$

where p_i is the probability over class i, z_i^n is the logit outputs over class i from the new model, y_i is the ground-truth label. However, in lifelong Re-ID, there are different numbers of classes (person identities) with different numbers of samples in each dataset, which leads to inherent class imbalance, especially when a tiny number of memory exemplars from different datasets are continuously incorporated into X. This problem can be aggravated when simultaneously learning new information and distilling old knowledge. Thus, we replace $\mathcal{L}_{ce}$ (Eq. (3)) with a classification coherence loss $\mathcal{L}_c$ in M^n:

$$\mathcal{L}_c = -\sum_{i=1}^{C^n+C^o} y_i(1-p_i)^{\gamma(1-\omega(e))} log(p_i), \tag{4}$$

where γ is the focusing parameter (Lin et al. 2017), e is the training epoch, $\omega(e)$ is a coordinating factor defined as:

$$\omega(e) = \begin{cases} 1, & e \leq \theta; \\ 0, & e > \theta; \end{cases} \tag{5}$$

where θ is the epoch threshold for balancing new information learning, old knowledge distillation and rebalanced learning. When $\omega(e) = 1$, Eq. (4) becomes Eq. (3).

Remarks. This design results in an end-to-end two-stage learning procedure: The first stage is performed in the first θ epochs to incrementally generalise a model via learning new information without forgetting old knowledge (*continual generalisation stage*), while the second stage is performed in the remaining epochs to further improve model generalisation via rebalanced learning (*rebalanced learning stage*). Intuitively, we can set $\omega(e) \equiv 0$ to use $(1-p_i)^\gamma$ throughout the whole training process. However, this will impair the continual generalisation of M^n, because large modulating weights will be assigned to hard samples (mainly new samples) and small weights will be assigned to easy samples (including memory exemplars), resulting in overfitting to hard new samples and forgetting of old knowledge. Thus, we use a coordinating factor $\omega(e)$ for balancing new information learning, old knowledge distilling and rebalanced learning.

Distribution Coherence

When adapting a Re-ID model to a new dataset, a straightforward solution is fine-tuning the model with new samples. However, this fine-tuned model usually performs poorly on old classes because of catastrophic forgetting of what is already learned. To preserve already learned knowledge, the new model should mimic the behaviours of the old model, *i.e.* the output predictions about old classes should be consistent between M^o and M^n. In our work, we use a distillation loss (Hinton, Vinyals, and Dean 2015; Li and Hoiem 2018; Rebuffi et al. 2017) and the Kullback Leibler (KL) divergence to generate soft probability distributions for the old classes in M^o and M^n, so that we regularise the distribution coherence $\mathcal{L}_d$ between M^o and M^n as:

$$\mathcal{Q}_i^o = \frac{exp(z_i^o/T)}{\sum_{j=1}^{C^o} exp(z_j^o/T)}, \mathcal{P}_i^n = \frac{exp(z_i^n/T)}{\sum_{j=1}^{C^o} exp(z_j^n/T)}, \tag{6}$$

$$\mathcal{L}_d = \omega(e)\beta \sum_{i=1}^{C^o} \mathcal{Q}_i^o \cdot log\frac{\mathcal{Q}_i^o}{\mathcal{P}_i^n}, \tag{7}$$

where T is a temperature (Hinton, Vinyals, and Dean 2015), β is usually set to T^2 as a compensation factor, $\mathcal{Q}^o$ and $\mathcal{P}^n$ are soft probability distributions over old classes in the M^o and M^n. $\omega(e)$ is the coordinating factor as Eq. (5) for balancing continual generalisation and rebalanced learning.

Representation Coherence

Encouraging the distribution coherence between M^n and M^o helps to learn old knowledge, but as more classes are incrementally incorporated into the model, the probability distribution becomes softer, resulting in the ambiguity of decision boundary and performance degradation. To solve this problem, we further use a representation coherence loss $\mathcal{L}_r$ to align the feature embedding space of M^n and M^o. Specifically, we normalise the features of samples $\{v^n, v^o\}$ extracted by M^n and M^o, and compute $d^2(v^n, v^o)$ to measure the distance of features from new and old models. Since M^o is frozen, the new model might overfit to the old feature space, we therefore further regularise the new feature space by pulling the feature representations of samples extracted by M^n close to their hard positive counterparts and pushing away their hard negatives. Thus, $\mathcal{L}_r$ is formulated as:

$$\begin{aligned}\mathcal{L}_r =& \omega(e) max\big(\phi(x) d^2(v^n, v^o), \alpha + \phi(x) d^2(v^n, v^o) \\ &+ d(v^n, v^{n\text{-}p}) - d(v^n, v^{n\text{-}n})\big),\end{aligned} \tag{8}$$

where α is a margin, $v^{n\text{-}p}$ and $v^{n\text{-}n}$ are hard positive and negative counterparts of v^n respectively, $\phi(x)$ is used to control representation coherence learning with reserved exemplars and new samples. For lifelong Re-ID, to avoid overfitting to the old feature embedding space with domain discrepancies, we set $\phi(x) = \lambda$ if $x \in E^m$, else $\phi(x) = 0$ (where λ is a weight parameter).

Summary. The proposed GwFReID is end-to-end trainable for lifelong person Re-ID. When training with the first dataset, we use a classification coherence loss $\mathcal{L}_c$ to optimise a new model (no old knowledge); Thereafter, when sequential new datasets are available, we optimise the model with a comprehensive learning objective $\mathcal{L}$. We summarise the training process in Algorithm 1.

Algorithm 1 GwFReID for Lifelong Person Re-ID.

Input: Sequential input datasets X^n

if The first dataset **then** /*Without old knowledge*/
  Randomly initialise M^n
  $X = X^n$
  for $e = 1 \rightarrow e_{max}$ **do**
    Get classification coherence loss $\mathcal{L}_c$ (Eq.(4))
    Backward to update M^n with $\mathcal{L}_c$
  end for
else /*Generalising without forgetting*/
  $X = X^n \cup E^m$
  Initialise M^n=M^o and modify the last layer in M^n
  for $e = 1 \rightarrow e_{max}$ **do**
    Get classification coherence loss $\mathcal{L}_c$ (Eq.(4))
    Get distribution coherence loss $\mathcal{L}_d$ (Eq.(7))
    Get representation coherence loss $\mathcal{L}_r$ (Eq.(8))
    Backward to update M^n with Eq.(1)
  end for
end if
$M^o \leftarrow M^n$ and update exemplar memory E^m
return: An up-to-date model M^n

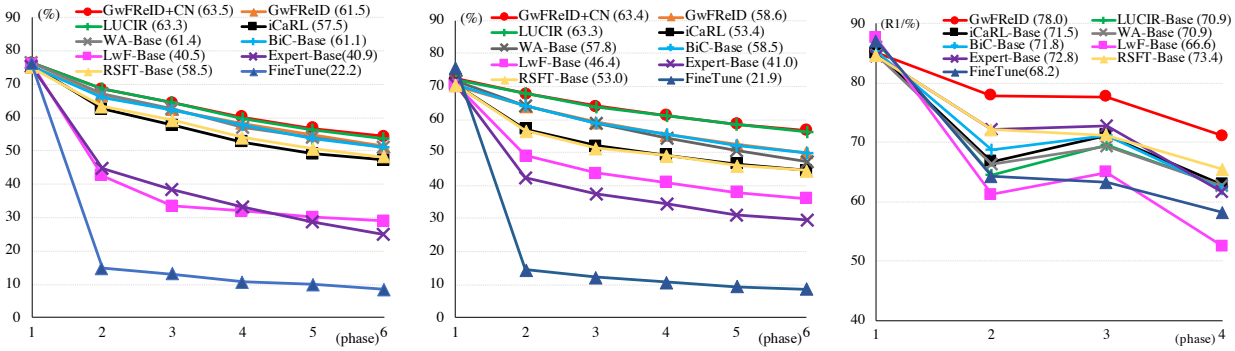
Types	Benchmarks	Total IDs	Total Images	Train IDs	Test IDs
Stream Based Input	Market-1501	1501	36036	751	750
	DukeMTMC	1404	36411	702	702
	CUHK-SYSU	8432	23435	5532	2900
	MSMT17	4101	124068	1041	3060
Unseen New Test	CUHK03	1467	14097	-	100
	iLIDS	119	476	-	60
	VIPeR	632	1264	-	316
	3DPeS	193	1012	-	96

Table 1: Person Re-ID evaluation setting statistics.

Experiments

Datasets. We conducted extensive experiments on eight person Re-ID benchmarks and two image classification datasets. **(1)** Although our method is designed for lifelong person Re-ID, it would be interesting to know the performance of our methods for non Re-ID tasks. Thus, we employed CIFAR-100 (Krizhevsky and Hinton 2009) and ImageNet (Russakovsky et al. 2015) to evaluate the incremental learning performance for image classification. CIFAR-100 consists of 60000 images in 100 classes, with 500 training images and 100 testing images per class. ImageNet with 1000 classes from ILSVRC 2012 (Russakovsky et al. 2015) contains 1.2 million training images and 50000 validation images. On each dataset, we used half of classes as the first dataset for initialisation in the first phase and evenly divided the remaining classes into 5 splits to mimic the lifelong learning process (6 phases in total). Following (Hou et al. 2019), an identical random seed (1993) by NumPy was used for class splitting. **(2)** We used four large-scale Re-ID benchmarks (Market-1501 (Zheng et al. 2015), DukeMTMC-ReID (Zheng, Zheng, and Yang 2017), CUHK-SYSU person search (Xiao et al. 2017) and MSMT17 (Wei et al. 2018)) as sequential input datasets to mimic the lifelong learning process (4 phases). On CUHK-SYSU, we modified the original dataset by using the ground-truth person bounding box annotation rather than using the original images which are used for person search evaluation. For testing on CUHK-SYSU, we fixed both query and gallery sets (w/o distractors) rather than used variable gallery sets. We used 2900 query persons and each person contains at least one image in the gallery. **(3)** We further tested the model (after training with all 4 phases) on four new Re-ID benchmarks (CUHK03 (Li et al. 2014), iLIDS (Zheng, Gong, and Xiang 2009), VIPeR (Gray and Tao 2008) and 3DPeS (Baltieri, Vezzani, and Cucchiara 2011)) to evaluate its lifelong generalised Re-ID performance. On CUHK03, we used the traditional training/testing splits for 20 trials, while on the other benchmarks, we employed the random half training/testing splits for 10 trials. The Re-ID evaluation statistics are summarised in Table 1.

Evaluation Metrics. On image classification evaluation, we computed the classification accuracy, while on person Re-ID evaluation, we computed the Rank-1 accuracy (R1) and



(a) CIFAR-100 (b) ImageNet (c) Re-ID benchmarks

Figure 2: Incremental learning performance evaluation on CIFAR-100, ImageNet, and Re-ID benchmarks.

mean Average Precision (mAP). To evaluate the incremental learning performance (on both image classification and Re-ID), after each training phase, we tested the model on all datasets/classes observed so far and plotted the accuracies in a line graph. The average incremental accuracy (Rebuffi et al. 2017) was shown in the legend of each line graph. To evaluate the lifelong generalised Re-ID performance, at the end of the last phase, we tested the model on all Re-ID benchmarks. Besides, at the end of the last phase, we computed the difference between the accuracies ($A_{R1/mAP}$) of the last (M_{last}) and the first models (M_{first}) on the initialisation dataset, and measured the forgetting ratio (FR) for lifelong Re-ID as $(1 - \frac{A_{R1}(M_{last})+A_{mAP}(M_{last})}{A_{R1}(M_{first})+A_{mAP}(M_{first})}) \times 100\%$.

Implementation Details. We implemented the proposed method using Python 3.6 and PyTorch 0.4, and trained it on NVIDIA TESLA GPUs. On Re-ID, we used ResNet-50 (He et al. 2016) (pretrained on ImageNet) as the backbone model. In each lifelong learning phase, we trained the model with 60 epochs for continual learning and 30 epochs for rebalanced learning (*i.e.* set $\theta = 60$ in Eq. (5) and $e_{max} = 90$). We used SGD as the optimiser with momentum 0.9 and weight decay $5e$-4. We set the initial learning rates to 0.01 for the feature extractor and 0.1 for the classification layers, which decayed by 0.1 after $\{40, 75\}$ epochs. We set batch size to 32, K=2 to construct the exemplar memory, λ=10 and α=0.5 in Eq. (8) to balance representation learning, $\gamma = 2$ in Eq. (4), $T = 2$ in Eqs. (6) and (7) to generate soft distribution. On image classification, we used ResNet-32 and ResNet-18 for CIFAR-100 and ImageNet, respectively. We set batch size to 128, K=20, $\beta = 0.1T^2$, and applied $\phi(x) \equiv \lambda$. For CIFAR-100, we trained the model with θ=160 epochs for continual learning and set e_{max}=200. We set the initial learning rate to 0.1, which decayed by 0.1 after $\{80, 120, 180\}$ epochs. For ImageNet, we trained the model with θ=90 epochs for continual learning and set e_{max}=112. We set the initial learning rate to 0.1, which decayed by 0.1 after $\{30, 60, 100, 110\}$ epochs.

Compared Methods. *iCaRL-Base*: A baseline model (*e.g.* ResNet-50 for Re-ID) with iCaRL (Rebuffi et al. 2017); *WA-Base*: A baseline with iCaRL plus weight aligning (Zhao et al. 2020); *LUCIR-Base*: A baseline with LUCIR (Hou et al. 2019); *BiC-Base*: A baseline with iCaRL plus a bias correction layer (Wu et al. 2019); *LwF-Base*: A baseline with LwF (Li and Hoiem 2018) (multi-class implementation with herding exemplars for the NME classifier); *RSFT-Base*: A baseline with iCaRL plus rebalanced resampling for fine-tuning (Castro et al. 2018); *Expert-Base*: A baseline with iCaRL using an expert distillation loss to replace classification loss (Hou et al. 2018); *FineTune*: A baseline incrementally fine-tuned with new datasets. *Joint-Train-All*: A baseline assembles all datasets/data in advance for joint training once. Here, *BiC-Base* for lifelong Re-ID used resampling to get the balanced validation data for bias correction, because the 9:1 splitting ratio is not applicable when using a tiny number of memory exemplars (*e.g.* K=2).

Incremental Learning Performance Evaluation

Evaluation on CIFAR-100 and ImageNet. To adapt our method for non ReID tasks (*i.e.* image classification in this work), we employed *GwFReID* (used traditional FC layers for classification) and *GwFReID+CN* (used a Cosine Normalisation (CN) classifier for classification and computed distillation loss with the scores before softmax w/o the rebalanced learning stage (Eq.(4)) as (Hou et al. 2019)). We used a NME classifier for testing prediction in all competitors (except FineTune and Joint-Train-All). As shown in Figs. 2(a) and 2(b), GwFReID+CN (red line) and GwFReID (orange line) achieve competitive performance compared with the state-of-the-art methods. Specifically, on CIFAR-100 (Fig. 2(a)), GwFReID+CN achieves the best average incremental accuracy (63.5%), while LUCIR and GwFReID rank the second and the third, respectively. Besides, during lifelong learning, the performance of FineTune drops dramatically, which indicates that FineTune suffers from catastrophic forgetting; By contrast, GwFReID achieves significantly better performance in each phase after the initialisation, which shows the effectiveness of GwFReID for simultaneously learning new information and old knowledge. Here, Joint-Train-All achieves 70.1% classification accuracy. On ImageNet (Fig. 2(b)), GwFReID+CN, LUCIR and GwFReID still perform better than the other competitors, where GwFReID+CN achieves the compelling average incremental accuracy 63.4%. Here, Joint-Train-All achieves

Methods	Train: Market→Duke→Cuhk-Sysu→MSMT17										
	Market		Duke		Cuhk-Sysu		MSMT17		Average		FR
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	
Joint-Train-All	88.2	72.9	76.4	62.3	87.5	85.5	60.7	34.7	78.2	63.9	-
LwF-Base	56.0	30.6	42.8	26.6	79.4	76.2	32.1	14.0	52.6	36.9	45.2
iCaRL-Base	77.0	55.9	56.1	37.4	84.4	81.5	34.4	14.5	63.0	47.3	12.7
LUCIR-Base	76.6	52.8	49.4	31.4	84.2	81.5	40.7	18.1	62.7	46.0	18.4
WA-Base	73.2	52.1	54.6	36.0	83.7	81.1	39.6	17.2	62.8	46.6	17.0
BiC-Base	75.8	53.4	55.4	37.7	84.2	81.4	33.0	13.1	62.1	46.4	15.3
Expert-Base	65.3	39.5	50.9	30.0	79.4	76.5	50.9	25.5	61.6	42.9	30.8
RSFT-Base	75.6	53.6	58.4	40.0	**84.7**	**82.0**	43.2	19.5	65.5	48.8	14.2
FineTune	52.6	26.5	46.0	29.2	75.1	71.3	**59.1**	**31.1**	58.2	39.5	47.4
GwFReID	**81.6**	**60.9**	**66.5**	**46.7**	83.9	81.4	52.4	25.9	**71.1**	**53.7**	**6.7**

Table 2: Lifelong Re-ID generalised performance evaluation on stream-based source domains. The results are reported at the end of the last training phase.

69.1% classification accuracy.

Evaluation on Re-ID Benchmarks. To mimic lifelong person Re-ID, we used four large-scale Re-ID benchmarks as the sequential inputs: Market→Duke→Cuhk-Sysu→MSMT17. As shown in Fig. 2(c), GwFReID achieves compelling performance compared with the state-of-the-art alternative methods. Specifically, after the initialisation (phase:1), GwFReID (red line) achieves the best results in each phase and gets the best average incremental accuracy (78.0%), which significantly outperforms the alternative methods. This demonstrates that GwFReID is effective for learning new information without catastrophic forgetting of old knowledge in lifelong Re-ID. Here, randomly selecting K samples per class at each phase for lifelong Re-ID can only achieve 33.9% average incremental accuracy.

Lifelong Generalised Performance Evaluation

Evaluation on Source Domains. As shown in Table 2, after training with four phases, on the first two source datasets, GwFReID achieves the best performance (81.6%/60.9% in R1/mAP on Market and 66.5%/46.7% in R1/mAP on Duke), which are significantly better than other incremental methods. This shows that GwFReID can preserve old knowledge without catastrophic forgetting. On the third dataset (CUHK-SYSU), most methods (*e.g.* RSFT-Base, iCaRL-Base, LUCIR-Base and GwFReID) achieve close performance because this dataset is less challenging. RSFT-Base ranks the first but it requires to resample datasets for fine-tuning, while GwFReID achieves 83.9%/81.4% in R1/mAP which are close to the performance of RSFT-Base. On the last dataset (MSMT17), FineTune performs significantly better than the other methods because it mainly focuses on learning information on the last dataset without distilling old knowledge, while GwFReID achieves the second-best R1/mAP (52.4%/25.9%) which are close to FineTune. This shows that GwFReID can simultaneously learn new information and distil old knowledge. On average, GwFReID achieves the best R1 (71.1%) and mAP (53.7%), which are significantly better than other incremental methods and are close to the Joint-Train-All. Besides, GwFReID achieves the best FR (6.7%), which shows the effectiveness of GwFReID for solving the catastrophic forgetting problem.

Evaluation on Unseen New Domains. As shown in Table 3, on four new unseen datasets, GwFReID achieves significantly better performance compared with other incremental methods. Specifically, GwFReID achieves the best R1 accuracies on CUHK03 (40.2%), iLIDS (69.5%), VIPeR (43.2%) and 3DPeS (64.9%), which are close to the Joint-Train-All. These results show that GwFReID is capable of generalising a model in lifelong Re-ID with good potential for real-world domain transfer deployment.

Methods	Train: Market→Duke→Cuhk-Sysu→MSMT17			
	CUHK03	iLIDS	VIPeR	3DPeS
Joint-Train-All	45.9	70.3	46.3	65.0
LwF-Base	32.7	64.0	33.6	51.4
iCaRL-Base	33.1	62.3	36.9	58.0
LUCIR-Base	37.2	66.7	39.3	56.7
WA-Base	30.5	65.8	39.8	58.9
BiC-Base	31.2	62.8	38.0	58.8
Expert-Base	33.3	63.0	37.8	56.0
RSFT-Base	33.2	67.0	38.8	58.6
FineTune	31.8	62.7	24.7	51.7
GwFReID	**40.2**	**69.5**	**43.2**	**64.9**

Table 3: Lifelong Re-ID generalised performance evaluation on new unseen domains. The results (R1) are reported at the end of the last training phase. Here, we did not use any training data on new test domains.

Components	mAP	R1
GwFReID (full model)	**53.7**	**71.1**
GwFReID w/o $\{\mathcal{L}_c, \mathcal{L}_d, \mathcal{L}_r\}^\star$	44.3	63.7
GwFReID w/o $\{\mathcal{L}_d, \mathcal{L}_r\}$	45.4	65.1
GwFReID w/o $\mathcal{L}_d$	52.9	70.8
GwFReID w/o $\mathcal{L}_r$	51.0	69.2
GwFReID w/o $\mathcal{L}_c{}^\star$	49.8	64.7

Table 4: Evaluating comprehensive learning objective on four sequential input Re-ID datasets. The average lifelong generalised performance at the end of the last training phase is reported. *: Use $\mathcal{L}_{ce}$ to replace $\mathcal{L}_c$ and set to 0 after $e > \theta$.

Further Analysis and Discussion

Comprehensive Learning Objective. Table 4 shows the evaluation on the comprehensive learning objective of GwFReID. We can see that GwFReID with all the optimisation component achieves the best performance (53.7% in mAP and 71.1% in R1), while GwFReID w/o $\{\mathcal{L}_c, \mathcal{L}_d, \mathcal{L}_r\}$ performs the worst. These results show the importance of distribution coherence, representation coherence and classi-

Components	mAP	R1
GwFReID (full model)	**53.7**	**71.1**
$\omega(e) \equiv 0$ in Eq.(4)	52.7	69.7
$\omega(e) \equiv 1$ in Eq.(4)	53.2	70.3
$\omega(e) \equiv 0$ in Eqs.(7) & (8)	45.4	65.1
$\omega(e) \equiv 1$ in Eqs.(7) & (8)	50.4	65.4

Table 5: Evaluating the coordinating factor on four sequential input Re-ID benchmarks. The average lifelong generalised performance is reported at the end of the last training phase.

Metric	λ=0	λ=10	λ=20	λ=50	$\phi(x{\in}E^m) = \lambda$	$\phi(x) \equiv \lambda$
FR	9.9	6.7	6.5	6.3	6.7	8.6
A_{inc}	77.0	78.0	77.5	77.2	78.0	74.9

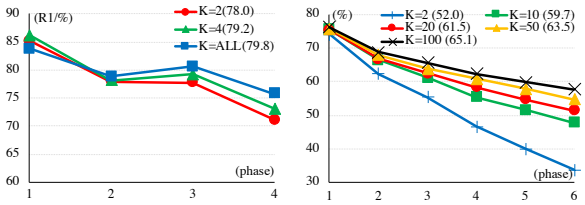
Table 6: Evaluating representation coherence parameters on four sequential Re-ID benchmarks. Forgetting Ratio (FR) and Average incremental accuracy (A_{inc}) are reported.

fication coherence in a unified framework for GwFReID.

Coordinating Factor. From Table 5, we can see that: **(1)** "$\omega(e) \equiv 0$ in Eq.(4)" means using the rebalanced factor throughout the whole training process, which leads to performance degradation due to overfitting to hard samples; **(2)** "$\omega(e) \equiv 1$ in Eq.(4)" means using the standard cross-entropy loss in the continual generalisation stage and then further using it for the rebalanced learning stage, which performs closely to GwFReID (full model), indicating the importance of the rebalanced stage for lifelong Re-ID; **(3)** "$\omega(e) \equiv 0$ in Eqs.(7) and (8)" means without using distribution and representation coherence losses, which results in the worst performance, while "$\omega(e) \equiv 1$ in Eqs.(7) and (8)" means using distribution and representation coherence losses throughout the whole training process, which results in poor performance due to overfitting to old knowledge.

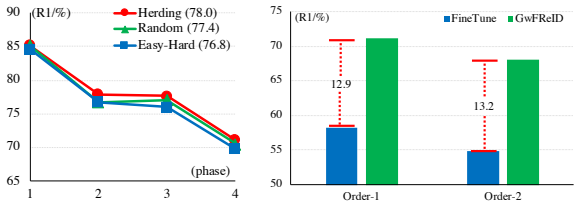
Representation Coherence Parameters. In Eq. (8), $\phi(x)$ and λ control representation coherence among samples. From Table 6, we can see that: **(1)** With the increase of λ, FR gradually decreases resulting in less catastrophic forgetting; **(2)** Setting λ=0 brings inferior FR and A_{inc}, but using a large λ also decreases A_{inc}; **(3)** When using GwFReID w/ $\phi(x) \equiv \lambda$ (align feature representations for all samples), the new feature embedding space overfits to the old one, resulting in inferior A_{inc}, but its FR is still better than GwFReID w/o using feature distillation (λ=0).

Exemplar Memory. Fig. 3 shows the impact of the number of memory exemplars on lifelong Re-ID and incremental image classification. On Re-ID (Fig. 3(a)), GwFReID with K=ALL performs significantly better than GwFReID with a few exemplars (K=2 and K=4). On CIFAR-100 (Fig. 3(b)), the performance of the GwFReID gradually improves when more exemplars are used. Here, when using a fixed memory size (2000 exemplars in total) on CIFAR-100, GwFReID still achieves 62.5% average incremental accuracy. Although using more memory exemplars brings better performance, it requires more storage and computational cost, so we set K=2 for lifelong Re-ID and K=20 for image classification.



(a) Re-ID (b) CIFAR-100

Figure 3: Evaluating exemplar memory on person Re-ID and image classification (Incremental accuracies).



(a) Exemplar selection (b) Sequential input order

Figure 4: Evaluating (a) exemplar selection on Re-ID (Incremental accuracies) and (b) sequential input orders of Re-ID datasets (Average lifelong generalised performance). In (b), Order-1: Market→Duke→Cuhk-Sysu→MSMT17. Order-2: Cuhk-Sysu→MSMT17→Market→Duke.

Exemplar Selection. Fig. 4(a) compares herding selection, random selection and easy-hard selection (select the first and the last samples in the sorted list) for exemplar selection in GwFReID. We can see that GwFReID with herding selection performs slightly better than the other variants.

Sequential Input Order. In Fig. 4(b), we tested with two different sequential input orders. As shown in Fig. 4(b), GwFReID outperforms FineTune by approximately 13% in terms of average lifelong generalised R1 accuracy in both orders, which indicates that GwFReID is applicable to lifelong Re-ID not specific to the order of input datasets.

Conclusion

In this work, we addressed the problem of lifelong person Re-ID, an incrementally generalisable learning approach to more realistic deployment requirements. We characterised lifelong Re-ID by incrementally learning on new domains for unseen class recognition without forgetting old knowledge whilst subject to class imbalanced data. We formulated a novel Generalising without Forgetting method (GwFReID) for lifelong Re-ID, which resembles the spirit of the human vision perceptual process by exploiting new samples and memory exemplars for simultaneously learning new information, distilling old knowledge and solving class imbalance. This is accomplished using a comprehensive learning objective that accounts for classification coherence, distribution coherence and representation coherence in a unified end-to-end trainable framework. Extensive experiments on eight Re-ID benchmarks, CIFAR-100 and ImageNet show the advantages of GwFReID over the state-of-the-arts.

Acknowledgements. This work is supported by Vision Semantics Limited, Alan Turing Institute Turing Fellowship, and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149), Queen Mary University of London Principal's Scholarship.

References

Baltieri, D.; Vezzani, R.; and Cucchiara, R. 2011. 3DPeS: 3D people dataset for surveillance and forensics. In *Proceedings of the joint ACM Workshop on Human Gesture and Behavior Understanding*, 59–64.

Castro, F. M.; Marín-Jiménez, M. J.; Guil, N.; Schmid, C.; and Alahari, K. 2018. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision*, 233–248.

Cheng, Z.; Dong, Q.; Gong, S.; and Zhu, X. 2020. Inter-task association critic for cross-resolution person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2605–2615.

Gray, D.; and Tao, H. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision*, 262–275.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* .

Hou, S.; Pan, X.; Change Loy, C.; Wang, Z.; and Lin, D. 2018. Lifelong learning via progressive distillation and retrospection. In *Proceedings of the European Conference on Computer Vision*, 437–452.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 831–839.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 152–159.

Li, Z.; and Hoiem, D. 2018. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(12): 2935–2947.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001–2010.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3): 211–252.

Song, J.; Yang, Y.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2019. Generalizable Person Re-identification by Domain-Invariant Mapping Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 719–728.

Sugianto, N.; Tjondronegoro, D.; Sorwar, G.; Chakraborty, P.; and Yuwono, E. I. 2019. Continuous Learning without Forgetting for Person Re-Identification. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 1–8.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 79–88.

Welling, M. 2009. Herding dynamical weights to learn. In *Proceedings of the International Conference on Machine Learning*, 1121–1128.

Wu, G.; Zhu, X.; and Gong, S. 2019. Spatio-temporal associative representation for video person re-identification. In *British Machine Vision Conference*.

Wu, G.; Zhu, X.; and Gong, S. 2020. Tracklet self-supervised learning for unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12362–12369.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large scale incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 374–382.

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1249–1258.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3415–3424.

Yu, H.-X.; Wu, A.; and Zheng, W.-S. 2017. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 994–1002.

Zhao, B.; Xiao, X.; Gan, G.; Zhang, B.; and Xia, S.-T. 2020. Maintaining Discrimination and Fairness in Class Incremental Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13208–13217.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, 1116–1124.

Zheng, W.-S.; Gong, S.; and Xiang, T. 2009. Associating groups of people. In *British Machine Vision Conference*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 3754–3762.