Meta reconciliation normalization for lifelong person re-identification

Pu, N.; Liu, Y.; Chen, W.; Bakker, E.M.; Lew, M.S.K.

Meta Reconciliation Normalization for Lifelong Person Re-Identification

Nan Pu*
n.pu@liacs.leidenuniv.nl
LIACS Media Lab, Leiden University
Leiden, The Netherlands

Yu Liu
liuyu8824@dlut.edu.cn
International School of Information Science & Engineering, Dalian University of Technology
Dalian, China

Wei Chen
w.chen@liacs.leidenuniv.nl
LIACS Media Lab, Leiden University
Leiden, The Netherlands

Erwin M. Bakker
E.M.Bakker@liacs.leidenuniv.nl
LIACS Media Lab, Leiden University
Leiden, The Netherlands

Michael S. Lew
m.s.k.lew@liacs.leidenuniv.nl
LIACS Media Lab, Leiden University
Leiden, The Netherlands

ABSTRACT

Lifelong person re-identification (LReID) is a challenging and emerging task, which concerns the ReID capability on both seen and unseen domains after learning across different domains continually. Existing works on LReID are devoted to introducing commonly-used lifelong learning approaches, while neglecting a serious side effect caused by using normalization layers in the context of domain-incremental learning. In this work, we aim to raise awareness of the importance of training proper batch normalization layers by proposing a new meta reconciliation normalization (MRN) method specifically designed for tackling LReID. Our MRN consists of grouped mixture standardization and additive rectified rescaling components, which are able to automatically maintain an optimal balance between domain-dependent and domain-independent statistics, and even adapt MRN for different testing instances. Furthermore, inspired by synaptic plasticity in human brain, we present a MRN-based meta-learning framework for mining the meta-knowledge shared across different domains, even without replaying any previous data, and further improve the model's LReID ability with theoretical analyses. Our method achieves new state-of-the-art performances on both balanced and imbalanced LReID benchmarks.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Lifelong Person Re-Identification, Meta-Learning, Normalization

ACM Reference Format:
Nan Pu, Yu Liu, Wei Chen, Erwin M. Bakker, and Michael S. Lew. 2022. Meta Reconciliation Normalization for Lifelong Person Re-Identification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, Lisbon, Portugal,, 9 pages. https://doi.org/10.1145/3503161.3548234

*Correponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3548234

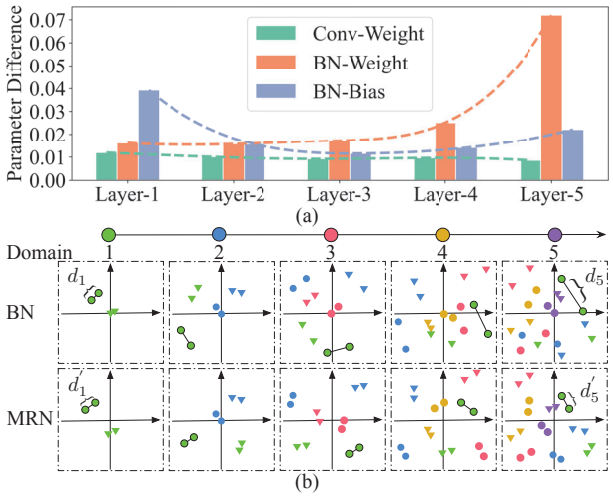

Figure 1: (a) Illustration of feature distribution shifts across five different domains. Different shapes indicate different person identities and different colors represent different domains. Due to continual adaptation, the feature distribution of Domain-1 suffers a significant shift, resulting in $d_1 \leq d_5$. (b) Comparing the changes in three types of network parameters when fine-tuning the model from Domain-1 to Domain-2.

1 INTRODUCTION

Lifelong person re-identification (LReID) is a practical extension of conventional ReID tasks. Different from other ReID variants, LReID aims to continually learn feature representations in a domain-incremental fashion, and reduces backward catastrophic forgetting on seen domains while improving the forward generalization ability on unseen domains. Due to significant and accumulated distribution shifts across multiple domains, the core difficulty for LReID is how to overcome the well-known catastrophic forgetting on seen domains and at the same time retain the generalization ability on

unseen domains. Recent state-of-the-art approaches [26, 31, 35] address this difficulty by leveraging sophisticated lifelong learning techniques, such as knowledge distillation [19] and data replay [27]. Although these methods have achieved promising performances on the recent CNN architectures like ResNet variants and graph network, they neglect the side effect caused by normalization layers, e.g., batch normalization (BN) [11], in the context of lifelong learning [42].

To clarify this problem further, we count the changes in the network parameters in the case of fine-tuning the model from one domain to a new one (Fig. 1(a)). In particular, we analyze the changes in terms of three types of parameters, i.e., convolutional weights, BN weights and BN bias. From Fig. 1(a), we witness several findings: (1) compared to convolutional weights, BN weights display a larger variation after training on a new domain; (2) the BNs in shallow layers are prone to performing translation transformation while the ones in top layers tend to scale transformation that heavily influences the relative positions of features; (3) the BN weights in top layers suffer from more dramatic parameter changes than those in shallow layers, which implies that top layers might include more domain-specific information than shallow layers. Based on these promising findings, we can expect that *the parameter changes in BN layers lead more to catastrophic forgetting for LReID*, since BN layers tend to capture the domain-specific characteristics[28]. To solve the limitation of BN, we thereby propose to learn a new reconciliation normalization (RN) consisting of grouped mixture standardization (GMS) and additive rectified rescaling (ARR). The former allows the model to automatically integrate BN with instance normalization (IN) [30] that is independent on domain-specific statistics. The latter endows RN with the ability to adaptively rectify the rescaling process based on changing statistics, so as to accomplish an instance-aware normalization. In short, our RN is a new replay-free approach toward reconciling domain-dependent and domain-independent statistics in the context of domain-incremental learning, without accessing previous data.

Even though RN is more flexible than BN, there is still an optimization bottleneck caused by stability-plasticity dilemma (SPD) [24], *i.e.*, improving less-forgetting ability always leads to degrading performance on new tasks. However, the synaptic plasticity in human brain has more complex mechanisms so as to protect against interference between old and new knowledge [8]. Motivated by the explorations in computational neuroscience[5], we explore a meta-learning framework to endow our RN with synaptic plasticity through a process of emulating "*reconciliation*" [14] in our brain. Our approach is called meta reconciliation normalization (MRN). Specifically, MRN first experiences an one-step visual update towards the objective of learning new knowledge only. This step is analogous to the hippocampus rapidly learning and acquiring new experiences. Next, we employ a balanced knowledge distillation to examine whether the one-step updated parameters are consistent with the optimized direction of reviewing old knowledge. This examination imitates the summarizing and consolidating process in our brains. Furthermore, based on the theoretical analysis in Sec. 3.3, we find that meta-optimization implicitly introduces an additional regularization term beneficial for coordinating the two optimization objectives in the SPD, compared with the conventional optimization (e.g., weighted sum of loss functions). Eventually, our MRN learns to reconcile the SPD and encourage the model to learn more common meta-knowledge, thereby generating less domain-dependent feature representation. From the comparison in Fig. 1(b), the parameters in our proposed MRN lead to smaller fluctuation than those of BN. This behavior is important to mitigate the catastrophic forgetting on old domains and leverage generalization ability on unseen domains. Our contributions in this work are three-fold:

- We analyse and address the problems in LReID from a new perspective of domain dependence, and design a new reconciliation normalization layer that consists of GMS and ARR components, to improve performance on both seen and unseen domains.
- We propose a meta-learning framework for LReID, which allows RN to imitate the hippocampus in our brain and learn synaptic plasticity that protects against interference between old and new knowledge, and theoretically analyse the mechanism of meta-optimization.
- Extensive experiments show that our method achieves new state-of-the-art performances on both balanced and imbalanced evaluation protocols in LReID tasks.

2 RELATED WORK

2.1 Lifelong Person Re-identification

With the demand of learning in non-stationary scenarios, LReID has become increasingly important in the ReID community. Recently, the work in [26] proposed a new benchmark for evaluating LReID by reorganizing several popular ReID datasets. Compared to the conventional lifelong learning tasks, LReID considers not only the catastrophic forgetting problem but also the ability to generalise on unseen classes and even unseen domains. Pu *et al.* introduced an adaptive knowledge accumulation [26] framework, which enables new-old knowledge graphs to communicate mutually, leading to knowledge accumulation. Wu *et al.* integrated multiple incremental learning techniques, *e.g.*, exemplar replay, balanced finetune and knowledge distilation (KD), to make the lifelong training process coherent [35]. Zhao *et al.* proposed to improve KD with selectively distilling knowledge by neighborhood selection [37].

However, these methods address the LReID task by learning from conventional class-incremental tasks, neglecting the characteristics of LReID. Inspired by extensive explorations of normalization in ReID[3, 12, 23, 41, 42], we analyse the reason of catastrophic forgetting from a perspective of normalization and propose a new reconciliation normalization to approximate domain-independent normalization, thereby overcoming the challenges of person re-identification in the context of lifelong learning.

2.2 Normalization in Person Re-Identification

In the last few years, ReID research has seen great progress and been evolved into various tasks, such as fully-supervised (FS) task, unsupervised domain adaption (UDA), and domain generalization (DG). Meanwhile, normalization, as an important technique in deep neural network, has been exploited to adapt to these different tasks as well. 1) In the FS task, Pan *et al.* presented IBN-Net[23] to enhance learning and generalization capacities by carefully integrating instance normalization (IN) [30] and batch normalization (BN) [11] as building blocks; 2) For the UDA task, Zhuang *et al.* proposed

a new camera-specific BN [42] to narrow the gap across different domains or cameras, and indicated that utilizing unlabeled test data to adapt the statistics in BN achieves considerable improvements. Later, Bai *et al.* designed multiple domain-specific BNs[1] for handling domain characteristics; 3) In the DG task, Jia *et al.* provided a sample yet efficient proposal [12] by selectively replacing BN with IN. The most related work [3] proposed a meta batch-instance normalization (MBIN) for DG, which leveraged meta-learning paradigm to balance the output of BN and IN for better generalization. Whereas this method needs to access multiple domain sources at once, which violates the setting of LReID. Furthermore, when directly employing MBIN in our framework, it still tends to over-fit on current domain without our meta rectified scaling, as experimentally demonstrated in Sec. 4.4.

BN revisiting. The BN[11] layer is designed to reduce the internal covariate shifting. In training, it first standardizes each input feature with the mini-batch statistics and records them for approximating the global statistics; Then, BN restores the representation power of the standardized features by utilizing learnable scaling parameters. During testing, given an input feature, the output of the BN layer is:

$$x_{out} = \gamma_{(B)} \frac{x_{in} - \mu_{(B)}}{\sqrt{\sigma^2_{(B)} + \epsilon}} + \beta_{(B)}, \tag{1}$$

where x_{in} and $x_{out} \in \mathbb{R}^{C\times H\times W}$ are the input and output feature maps with C channels, respectively. H and W denote the height and width of the feature maps. $\gamma_{(B)}$ and $\beta_{(B)}$ are scaling parameters learned during training. Note that $\mu_{(B)}$ and $\sigma_{(B)}$ are the approximated global mean and standard deviation of the current training domain, which mainly causes BN's domain dependence.

BN's limitations for LReID tasks. BN assumes and requires that all testing images are subject to the same training distribution. Considering the evaluative criteria of LReID tasks, however, this assumption is satisfied only when evaluating on current domain, omitting intra-domain distribution discrepancies (*e.g.*, the distribution shift between train-test splits and distribution gaps caused by different cameras[42]). Thus, BN's limitations in LReID tasks are two-fold: 1) in testing, BN's standardization often fails to generalise on unseen domains, since the unseen domains usually subject to entirely different distribution; 2) when evaluating the model's less-forgetting ability, BN is prone to a domain-specific normalization so that the normalized features extracted from previous domains suffer from significant domain shifts. This limitation further aggravates catastrophic forgetting, due to the fixed statistics of BN in testing.

3 PROPOSED METHOD

Given a stream of domains $\mathcal{D} = \{D^t\}_{t=1}^T$, LReID model is required to continually learn T domains. The dataset of the t-th domain is represented as $D^{(t)} = \{D^t_{tr}, D^t_{te}\}$, where $D^t_{tr} = \{(x_i, y_i)\}_{i=1}^{N^t_{tr}}$ contains N^t_{tr} training images and their corresponding label set Y^t_{tr}, and D^t_{te} indicates the testing set with Y^t_{te}. The training and testing classes are disjoint, *i.e.*, $Y^t_{tr} \cap Y^t_{te} =$). Note that, only D^t_{tr} is available at the t-th training step, and the data from previous domains are *not* available any more, as shown in Fig. 2. For evaluation, we test retrieval performance on all encountered domains with their corresponding testing sets. In addition, the generalization ability is evaluated via new and unseen domains D_{un} with unseen identities Y_{un}. Henceforth, we will drop the subscript $\{tr, te\}$ for simplifying notation.

3.1 Balanced Knowledge Distillation

To mitigate catastrophic forgetting, we leverage knowledge distillation (KD) [19] to preserve knowledge learned on previous domains. As person re-identification task is usually a long-tailed representation learning problem [20] with imbalanced data, we propose a balanced knowledge distillation (BKD) loss $\mathcal{L}_{bkd}$ to reinforce the learned knowledge of tail classes which will be severely forgotten during the lifelong learning process [13]. Specifically, given a feature extractor $h\left(\cdot;\theta_f,\phi\right)$ with the normalization parameters ϕ and the convolutional parameters θ_f, and a classifier $g\left(\cdot;\theta_c\right)$. The probabilities generated from the old and new model are defined as $\mathbf{p}^{t-1} = g\left(h\left(\cdot;\theta_f^{t-1},\phi^{t-1}\right);\theta_c^{t-1}\right)$ and $\mathbf{p}^{t} = g\left(h\left(\cdot;\theta_f^{t},\phi^{t}\right);\theta_c^{t}\right)$, where θ_f^{t-1}, ϕ^{t-1} and θ_c^{t-1} are copied from θ_f^t, ϕ^t and θ_c^t before current-step training, respectively. The proposed BKD loss is formulated as:

$$\mathcal{L}_{bkd} = -\sum_{x\in D^t}\sum_{j=1}^{m} \frac{1-\pi}{1-\pi^{n_j}} \mathbf{p}_j^{t-1}(x) \log \mathbf{p}_j^t(x) \tag{2}$$

where n_j is the number of training images in the j-th class and $m = \sum_{i=1}^{t-1} \left|Y^{(i)}\right|$ is the number of the old classes. π is a hyper-parameter to control the re-balance strength. Moreover, we use the commonly-used cross-entropy loss $\mathcal{L}_c$ for learning on current domain. Thus, the total objective of the proposed base-train stage is:

$$\mathcal{L}_{base}(\mathcal{D}^t;\theta,\phi) = \mathcal{L}_c(\mathcal{D}^t;\theta,\phi) + \lambda\mathcal{L}_{bkd}(\mathcal{D}^t;\theta,\phi), \tag{3}$$

where λ is a trade-off factor for the knowledge distillation loss and the cross-entropy loss. For a fair comparison, we set it to 1 in all of our experiments.

3.2 Reconciliation Normalization

Based on the drawbacks expatiated in Sec. 2.2, we propose a sparse mixture standardization and a additive rectified rescaling to mitigate domain-specific dependence of BN.

Grouped Mixture Standardization (GMS). Inspired by DSON [28], instance normalization (IN)[30] has the capability to remove domain-specific characteristics and standardize the representations. Nevertheless, since IN removes domain-specific information while filtering out partial discriminative information, the model with only INs struggles to learn and accumulate the ReID knowledge of the current domain well. To this end, we propose to learn an appropriate coefficient for mixing instance and batch standardization instead of their outputs. Considering the sparsity and the cost of additional parameters, we group the mixed coefficients by reduction rate of C_m, denoted by $\phi_\rho \in \mathbb{R}^{C/C_m}$, instead of channel-wise mixed coefficients. Thus, our SMS is formulated as:

$$\hat{x} = (1-\phi_\rho)\frac{x_{in} - \mu_{(I)}}{\sqrt{\sigma^2_{(I)} + \epsilon}} + \phi_\rho \frac{x_{in} - \mu_{(B)}}{\sqrt{\sigma^2_{(B)} + \epsilon}}, \tag{4}$$

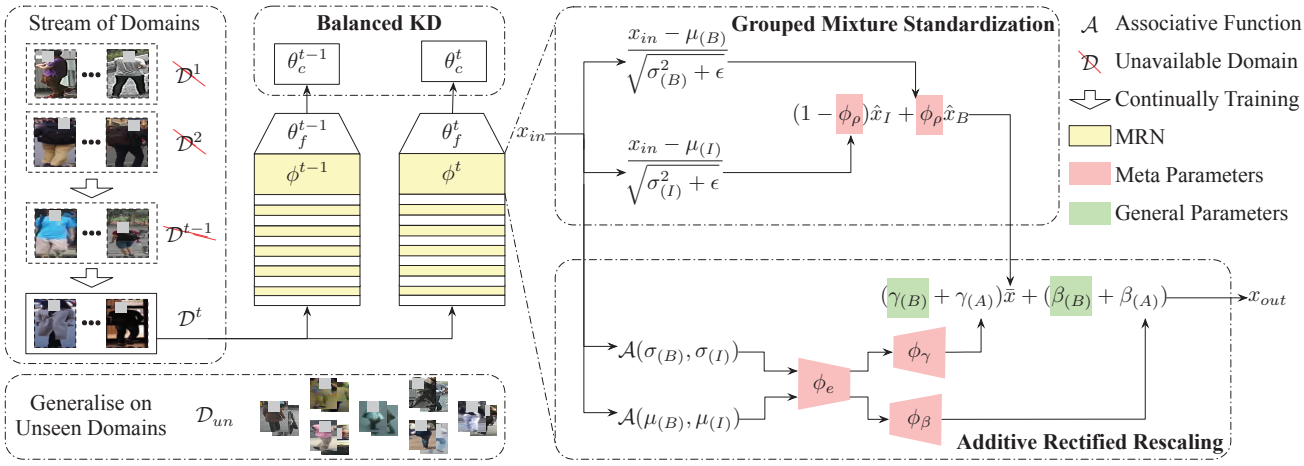

Figure 2: Left: LReID model is required to continually learning on a steam of domains and generalise on unseen domains. For each step, only the current-domain data are available. Right: illustration of our MRN. General and meta parameters are optimized alternatively.

where $\hat{x}$ is the feature standardized by the proposed SMS. $\mu_{(I)}$ and $\sigma_{(I)}$ are the the channel-wise mean and standard deviation of the input feature, respectively. ρ is optimized to balance the domain-dependent factors caused by BNs and the domain-independent factors drawn from instances themselves. Ideally, based on an optimal trade-off, the data from the current, previous and even unseen domains should be effectively standardized. With less bias caused by domain shifts, the model is encouraged to learn more general knowledge to generalise on any domains well. Note that unlike MBIN [3] that optimizes a combination of the outputs of BN and IN, we balance two statistics in standardization to avoid to introduce the IN's rescaling parameters, which still tend to be dominated by domain dependence.

Additive Rectified Rescaling (ARR). Despite the proposed SMS with less domain dependence, the conventional rescaling parameters still have a high risk to over-fit on current training domain, especially for top layers as shown in Fig. 1. Hence, we propose a new ARR to learn to rectify the learned rescaling parameters based on the domain-independent statistics (DIS) and domain-dependent statistics (DDS). Specifically, MR first associates DIS and DDS by an associative encoder parameterized by ϕ_e, then modeling the relationship between standardization and rescaling processes, and finally producing the rectified factors $\gamma_{(A)}$ and $\beta_{(A)}$ by two decoders ϕ_γ and ϕ_β, respectively. The proposed ARR is formulated as:

$$\begin{aligned} x_{out} &= (\gamma_{(B)} + \gamma_{(A)})\hat{x} + (\beta_{(B)} + \beta_{(A)}), \\ \beta_{(A)} &= \phi_\beta(\phi_e(\mathcal{A}(\mu_{(B)}, \mu_{(I)}))), \\ \gamma_{(A)} &= \phi_\gamma(\phi_e(\mathcal{A}(\sigma_{(B)}, \sigma_{(I)})), \end{aligned} \tag{5}$$

where $\mathcal{A}$ is an associative function. Based on experiential exploration in Tab. 3, we choose subtraction function in all the experiments. Similar to the grouped standardization, we first group statistics by a reduction rate of C_s, and then embed them into C/C_e-dimensional subspace. Then, ARR infers the rectified factors from the relationship between DIS and DDS. By summing the conventional scaling parameters and rectified factors, we use the rectified scaling parameters to recover the representation capability. This brings two benefits: 1) when testing on unseen domains, the model can perform an instance-aware normalization with less domain dependence, which significantly improves generalization ability; 2) the learned features are more general and can be shared across multiple previously-training domains, which allows models to change less over training domains, so as to alleviate catastrophic forgetting.

3.3 Meta Reconciliation Normalization

To discover common knowledge in the context of domain-incremental LReID, we propose a meta-learning framework to mine and optimize the meta-knowledge from both old and new domains. Combining this idea with the proposed RN, we further propose a meta RN (MRN) that learns to normalize features based on different instances and better reconcile the dual objectives in the SPD [29].

The main idea of our meta-learning framework is to simulate when learning on the new domain, how model can keep the distributional consistence between old and new knowledge. By doing so, the model can forget old knowledge less while facilitating to learn more common knowledge from new domains. Considering the computational efficiency of meta optimization, we find that fine-tuning only partial parameters in MRN by meta gradients (*i.e.,* second-order gradients) achieves considerable performance gain. Specifically, we denote these parameters as meta parameters $\phi = \{\phi_\rho, \phi_e, \phi_\gamma, \phi_\beta\}$, as illustrated in Fig. 2. As summarized in Algorithm 1, when starting with an incremental training step, the training procedure in an iteration includes base-train, meta-train and meta-test stage.

Base-train: we update all parameters except for the meta parameters ϕ and sample a class-balanced mini-batch D_{base} from current domain D^t according to their identities. The $\mathcal{L}_{base}$ in Eq. (3) is used for updating the general parameters in backbone network and the classifier optimized by SGD with the learning rate α_{base}.

Algorithm 1: Meta Reconciliation Normalization

Input: Model Parameters $\theta = \{\theta_f, \theta_c\}$, Meta Parameters $\phi = \{\phi_e, \phi_\gamma, \phi_\beta, \phi_\rho\}$, Multiple Domain Data $\{\mathcal{D}\}$, learning rate $\alpha = \{\alpha_{base}, \alpha_{mtr}, \alpha_{mte}\}$

. **Output:** θ, ϕ

```
for t = 1 in total_domains do
    if t == 1 then
        for i = 1 in domain_iterations do
            Sample mini-batch D_base from D^t;
            Base-train:;                                  // Eq. (3)
            θ ← θ − α_base Δ_θ L_base(D_base; θ, φ);
            φ ← φ − α_base Δ_φ L_base(D_base; θ, φ);
        end
    else
        for i = 1 in domain_iterations do
            Sample D_base, D_mtr, D_mte from D^t;
            Base-train:;                                  // Eq. (3)
            θ ← θ − α_base Δ_θ L_base(D_base; θ, φ);
            Meta-train:;                                  // Eq. (6)
            φ' = φ − α_mtr Δ_φ L_mtr(D_mtr; θ, φ);
            Meta-test:;                                   // Eq. (9)
            φ ← φ − α_mte Δ_φ L_mte(D_mtr; θ, φ');
        end
    end
end
```

Meta-train: This stage is to simulate the scenario where the meta parameters are optimized toward fitting only new knowledge. We first sample a mini-batch D_{mtr} randomly from current domain D^t, then use the $\mathcal{L}_c$ loss as meta-train loss to guild MRNs to learning on only new knowledge, and finally inner-update the meta parameters from ϕ to ϕ':

$$\mathcal{L}_{mtr}(\mathcal{D}_{mtr}; \theta, \phi) = \mathcal{L}_c(\mathcal{D}_{mtr}; \theta, \phi), \tag{6}$$

$$\phi' = \phi - \alpha_{mtr}\Delta_\phi \mathcal{L}_{mtr}(\mathcal{D}_{mtr}; \theta, \phi). \tag{7}$$

Meta-test: After updating the meta parameters in the inner-level optimization step, we next examine MRNs on a new mini-batch D_{mte} drawn from D^t. In this step, the MRNs are required to ensure the learned knowledge in last step is common and beneficial for both new and old knowledge. Furthermore, motivated by that sparse neural activation is a key mechanism to summarize and consolidate knowledge in human cognitive processes [4, 32], we encourage the model to learn refined knowledge or sparse knowledge so as to effectively generalize on new scenarios. Thus, we employ sparse constraints $\mathcal{L}_s$ and the BKD loss in Eq. (2) with the updated meta parameters ϕ' for meta-test:

$$\mathcal{L}_s = \frac{1}{L}\sum_{l=1}^{L}\frac{\left\|\phi_\rho'^l\right\|_1}{C^l/C_m^l}, \tag{8}$$

$$\mathcal{L}_{mte}(\mathcal{D}_{mte}; \theta, \phi') = \mathcal{L}_{bkd}(\mathcal{D}_{mte}; \theta, \phi') + \lambda_s\mathcal{L}_s, \tag{9}$$

$$\phi \leftarrow \phi - \alpha_{mte}\Delta_\phi \mathcal{L}_{mte}(\mathcal{D}_{mtr}; \theta, \phi'), \tag{10}$$

where L is the number of norm layers and λ_s is a weight factor for the sparse objective. By using Eq. (10), we meta-update the meta parameters to overcome the incomplete optimization in the meta-train simulation. Eventually, our MRN is optimized to be an approximated domain-independent normalization by the significant second-order gradients.

Table 1: The comparison in terms of the number of identities between balanced and imbalanced setting in LReID-Seen datasets.

LReID-Seen	Abbr.	The Number of Training Identities	
		Balanced	Imbalanced
Market-1501[39]	MA	500	751
CUHK-SYSU LReID[36]	SY	500	5532
MSMT17_V2 [33]	MS	500	1041
CUHK03[18]	CU	500	767

Analysis: Apart from the inspiration from computational neuroscience [5], we further analyze the reason why meta-learning can reconcile these two different objective functions. Along a similar vein [15], we derive the first-order Taylor expansion for the final objective of meta-train and meta-test meanwhile omitting θ that are constants in meta-optimization:

$$\begin{aligned}\mathcal{L}_{mte}(\phi - \alpha_{mte}\Delta_\phi \mathcal{L}_{mtr}(\phi)) = \\ \mathcal{L}_{mte}(\phi) - \underbrace{\alpha_{mte}\Delta_\phi \mathcal{L}_{mtr}(\phi)\cdot\Delta_\phi \mathcal{L}_{mte}(\phi)}_{\textbf{Reconciliation Term}},\end{aligned} \tag{11}$$

where the second term in Eq. (11) is a dot product of $\Delta_\phi \mathcal{L}_{mtr}(\phi)$ and $\Delta_\phi \mathcal{L}_{mte}(\phi)$. In this paper, we call it reconciliation term (RT). When minimizing the above objective, the RT is expected to become large. Though $\Delta_\phi \mathcal{L}_{mtr}(\phi)$ and $\Delta_\phi \mathcal{L}_{mte}(\phi)$ are not normalized, the dot product is still larger if these vectors/gradients are in a similar direction. Then, the similar direction means the direction of optimization on both non-forgetting old knowledge and learning new knowledge is similar. Thus, the overall objective can minimize both two objectives meanwhile keeping their gradient descents in a reconciled way. In contrast to the conventional optimization in Eq. (3) that is not constrained by RT, the meta optimizer trends to updates to weights in which the two optimization surfaces agree on the gradient. It reduces over-fitting to ether remembering old knowledge or learning new knowledge by finding a optimization path to minimization in which both objectives approximately are consistent on the direction at all points along the path.

As a consequence, the advantages of MRN can be summarized as two-fold: 1) we experimentally demonstrate that the learned features with less domain dependence realize a dramatic improvement of generalization ability and mitigate forgetting problem by a considerable degree; 2) since the base-train and the meta train-test stage are alternately executed, this endows our MRN to adaptively rectify the way of normalization under variational training distributions caused by not only the domain shift but also the instability in the early stage of incremental training.

4 EXPERIMENTS

4.1 Implementation Details

We remove the classifier of ResNet-50 and use the retained layers as a feature extractor. The feature dimension is 2,048, and the batch

Table 2: Lifelong ReID evaluation on balanced and imbalanced protocols. We test the model after sequentially training on four seen domains. The reported results are reproduced in our setting by using their official codes. "Meta." denotes optimizing by meta objective.

Training Order				MA		SY		MS		CU		Average Seen ↑		Average FR (%)↓		Average Unseen ↑	
Protocol	Method	Replay	Meta.	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	δ(mAP)	δ(R-1)	mAP	R-1
Balanced [26]	SFT			18.9	40.8	59.7	62.8	2.0	6.0	**56.2**	**60.1**	34.2	42.4	47.4	40.9	43.5	40.1
	LwF[19]			45.3	67.7	74.0	77.7	3.8	11.3	28.1	27.6	37.9	46.1	27.1	20.9	44.2	41.4
	iCaRL[27]	√		42.7	65.2	66.7	69.6	3.7	10.4	34.6	36.5	36.9	45.4	34.2	28.2	37.0	33.2
	DER++[2] *with* BN[11]	√		38.9	60.8	68.2	71.9	4.2	11.1	41.0	42.8	38.1	46.7	35.2	29.1	40.6	36.8
	- *with* CN[25]	√		48.4	70.5	70.2	73.1	6.6	18.6	39.8	41.6	41.3	51.0	29.4	23.1	42.6	39.0
	AKA[26] *with* BN[11]			44.4	67.7	73.8	79.8	4.1	11.8	33.9	34.0	39.1	48.3	28.2	22.4	46.5	43.1
	- *with* DN[12]			52.5	78.0	69.9	72.9	9.2	25.7	33.1	33.4	41.2	52.5	15.7	9.6	45.6	41.2
	- *with* BIN[22]			56.2	78.9	**78.9**	81.1	14.2	34.4	37.2	37.9	46.6	58.1	14.6	9.4	54.6	51.0
	- *with* CN[25]			56.9	78.4	78.2	**83.9**	10.4	26.6	32.9	32.1	44.6	55.3	16.0	11.7	49.9	46.0
	- *with* RN(Ours)			55.1	78.5	78.5	80.7	14.2	35.1	37.5	37.9	46.3	58.1	13.8	8.9	54.3	51.0
	AKA[26] *with* MBIN[3]		√	54.7	78.2	78.4	80.1	13.6	32.8	40.3	40.4	46.8	57.9	12.7	10.4	53.5	49.7
	- *with* MRN(Ours)		√	**57.6**	**80.6**	77.5	79.8	**16.5**	**39.9**	42.9	43.7	**48.6**	**61.0**	**11.9**	**6.6**	**56.1**	**52.7**
Imbalanced (Ours)	SFT			25.2	49.1	68.0	71.5	13.5	10.7	**67.6**	**70.6**	41.1	50.5	46.2	38.1	52.6	48.9
	LwF[19]			51.7	72.8	77.9	80.8	4.9	13.6	34.1	34.1	42.2	50.3	25.9	19.9	50.6	47.5
	iCaRL[27]	√		59.8	78.0	88.0	89.4	11.5	25.0	52.2	53.4	52.9	61.5	23.5	18.5	59.2	54.3
	DER++[2] *with* BN[11]	√		60.4	79.2	**88.9**	**90.5**	14.2	30.0	39.5	39.5	50.8	59.8	21.9	16.4	62.6	58.1
	- *with* CN[25]	√		64.0	82.0	88.5	90.2	15.0	33.1	39.5	39.5	51.8	61.2	19.3	13.8	**62.7**	**58.4**
	AKA[26] *with* BN[11]			50.7	71.3	76.8	79.8	4.8	13.5	39.5	39.5	43.0	51.0	28.7	22.1	49.4	48.4
	- *with* DN[12]			64.2	82.8	81.6	83.9	11.5	28.4	38.0	38.4	48.8	58.4	16.7	11.8	53.4	49.0
	- *with* BIN[22]			64.8	84.2	82.7	80.7	15.4	33.8	40.0	40.0	50.7	59.7	13.7	10.8	55.7	52.4
	- *with* CN[25]			62.0	83.6	77.5	79.9	12.2	31.1	37.3	37.4	47.3	58.0	12.2	7.7	54.3	50.6
	- *with* RN(Ours)			63.2	83.4	81.7	84.2	14.2	34.3	39.8	39.2	49.7	60.3	14.3	9.2	57.5	53.6
	AKA[26] *with* MBIN[3]		√	65.5	84.9	83.5	85.8	14.1	36.1	44.4	44.8	51.9	63.0	16.7	9.6	58.0	53.9
	- *with* MRN(Ours)		√	**67.2**	**85.7**	84.7	86.7	**18.3**	**41.8**	42.9	43.7	**53.3**	**64.5**	**10.5**	**6.2**	60.6	56.8

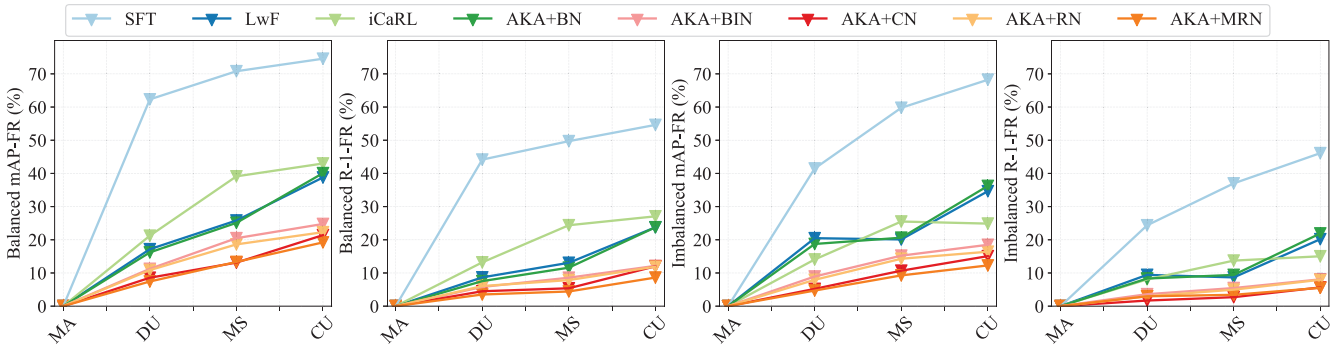

Figure 3: The trend of forgetting rate on different setting. *Left*: visualization of the forgetting rate of mAP and R-1 score on balanced setting. *Right:* visualization of forgetting rate of mAP and R-1 score on imbalanced setting. The models are trained following *Order-1*.

size is 32. We randomly select 8 identities and sample 4 images for each identity to form a class-balanced mini-batch. All images are resized to 256×128. The shared encoder is a fully-connected layer followed by a h-swish [10] activation function. The ϕ_γ and ϕ_β are implemented by a fully-connected layer followed by a hard-sigmoid [10] function and Tanh function, respectively. To balance computational cost and performance, we set the reduction rate, $C_m = 2$, $C_s = 16$ and $C_e = 16$, for mixture coefficients, grouped statistics and associative embeddings, respectively. If the embedding dimension is smaller than 2, then the embedding dimension is set to 2. Following the popular person ReID training strategy, we train the model for 50 epochs, and decrease the learning rate by $\times$ 0.1 at the 25^{th} and 35^{th} epoch. The learning rates α_{base}, α_{mtr} and α_{mte} are set to 1.75×10^{-4}, 1×10^{-2} and 1×10^{-2}, respectively. We follow [26] to set the trade-off factor λ as 1 and fix the intensity of rebalance π and sparsity λ_s to 0.999 and 0.1 in all experiments.

4.2 Evaluation Protocols for LReID

Considering that the scale of each dataset varies largely in the wild, we follow the official codes of the authors of [26] to conduct experiments under both balanced and imbalanced evaluation protocols. Instead of randomly choosing a unified amount of identities in each domain [26], in imbalanced protocol the model is trained on each complete dataset regarded as a long-tail dataset. [20].

Balanced LReID-Seen: note that since the DukeMTMC-reID dataset [40] has been withdrawn by the authors, we do not follow previous lifelong ReID benchmarks [26, 35]. Instead, we select four large-scale datasets: MA [39], SY [36], MS [33], and CU [18], and randomly sample 500 identities from each dataset. This results in the balanced

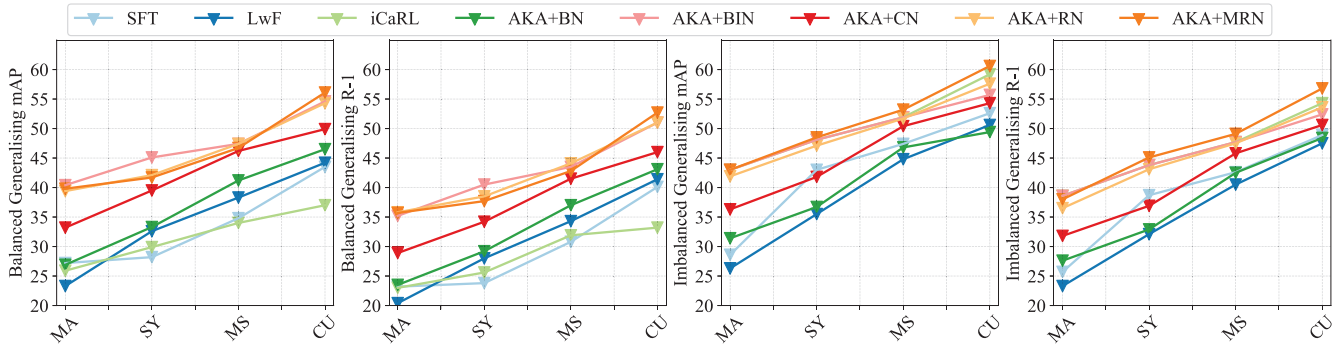

Figure 4: The trend of generalising performance on different settings. *Left*: visualization of the mAP and R-1 score on balanced setting. *Right:* visualization of the mAP and R-1 score on imbalanced setting. The models are trained by following *Order-1*.

LReID training set that includes 2,000 identities. We follow the train-test protocol of SY in [26] and sample the identities that include at least 4 images for training.

Imbalanced LReID-Seen: different from the balanced LReID-Seen set that randomly chooses a unified amount of identities in each domain [26], in imbalanced protocol the model is trained on each complete dataset. Consequently, 65,637 images of the 8,091 identities are employed for the imbalanced LReID training set. Note different from the train-test protocol of SY in [26], we reorganize this dataset but use all available identities for training instead of sampling a subset, called CUHK-SYSU LReID in Tab. 1.

LReID-Unseen: Note that the generalization evaluation are same in both balance and imbalance protocols. Specifically, we follow [26] to test models on the combination of the testing sets of seven ReID datasets, VIPeR [7], PRID [9], GRID [21], i-LIDS [34], CUHK01 [17], CUHK02 [16] and SenseReID [38].

Training Order: we randomly select ***Order-1*** (MA→SY→MS→CU) as the primary training order and show the experimental results in Sec. 4.3. Moreover, we explore more training orders in Appendix C.

Evaluation metrics. In LReID experiments, we use mean average precision (mAP) and rank-1 (R-1) accuracy to evaluate the performance. Moreover, we adapt forgetting rate δ_s^t on the s-th seen domains after the t-th training step, to measure the degree of forgetting. The forgetting rate (FR) is:

$$\delta_s^t(a) = \frac{\max_{k\in\{1,\cdots,t-1\}} a_{k,s} - a_{t,s}}{\max_{k\in\{1,\cdots,t\}} a_{l,s}}, \quad \forall s < t, \tag{12}$$

where a is mAP or R-1. We calculate the average $\bar{\delta}$ over on the seen domain to get a comprehensive evaluative criteria.

4.3 Comparative results of LReID Evaluation

We compare our MRN against several state-of-the-art methods, including three groups: 1) conventional lifelong learning methods, *e.g.*, sequential fine-tune (SFT), LwF [19], iCaRL [27], and Dark Experience Replay++ (DER++) [2] and its variants with continual normalization (CN) [25]; 2) very recent LReID methods, AKA [26] and its variants with different normalization layers, e.g., DN [12] and BIN [22]; 3) a meta-learning based domain generalization method, MBIN [3]. For a reasonable comparison, we employ the norm layers in the MBIN but optimize these layers by our meta objective since the original objective of MBIN impractically requires accessing multi-domain data at once.

Seen-domain Non-forgetting Evaluation: 1) compared with the conventional lifelong methods, MRN ranks first in both balanced and imbalanced settings and even outperforms DER++, which is a strong replay-based method; 2) in AKA-based comparison, MRN outperforms BN by 12.7% in R-1 of balanced setting and 13.5% in R-1 of imbalanced setting; 3) although MIBN optimized by our meta-optimization objective gets competitive results, MRN still outperforms MIBN on FR by a margin of 3.4%, due to our designs specific for preventing forgetting.

Unseen-domain Generalising Evaluation: 1) owing to storing the data of previous domains, the replay-based methods have superior performances on generalising evaluation since the models are trained by large-scale and diverse data during the whole lifelong training process. Even so, MRN ranks the first in balanced setting on generalising test; 2) in AKA-based comparison, MRN outperforms BN by 9.6% in mAP of balanced setting and 11.2% in mAP of imbalanced setting; 3) compared with MBIN, MRN shows better ability to generalize on unseen domains in both settings, due to our instance-aware normalization mechanism.

Balanced & Imbalanced: 1) from the results of SFT and LwF in both settings, we find they are unstable when training on a larger-scale dataset, especially in the imbalanced setting, which indicates our imbalanced setting is more challenging; 2) the replay-based methods benefit from their exemplar sampling strategy, which is favorable for handling the imbalanced setting. 3) since a lengthy training epoch leads learnable knowledge graph to be over-fitting, AKA cannot effectively preserve seen-domain knowledge, so that AKA's performance degrades to the same level as LwF in the imbalanced setting. Surprisingly, by replacing BN with our RN, this drawback is improved to some extent.

Exemplar Replay & Replay-Free: although "DER++ with CN" in Tab. 2 stores old data to replay in order to mitigate catastrophic forgetting, our MRN is superior to it by 2.5% on average mAP and 8.8% on average FR of seen domains, without accessing to any previous data. Note that replay-based methods tend to perform well on SY dataset. This is because most identities in SY have few instances and storing exemplars for each identity is almost equal to store the whole dataset. Moreover, maintaining a memory buffer is expensive and impractical in LReID due to privacy policy.

Table 3: Evaluation of the proposed components. The effectiveness is verified in imbalanced LReID setting. "Meta." denotes optimizing by meta objective.

Evaluation	Average FR (%)		Average Unseen	
Configuration	δ(mAP) ↓	δ(R-1) ↓	mAP↑	R-1↑
AKA *with* KD	28.6	24.5	49.3	48.2
AKA *with* BKD	26.7	22.1	49.4	48.4
w/o Meta. & ARR	13.7	10.8	55.7	52.4
w/o Meta. & GMS	21.1	14.2	52.5	51.3
w/o Meta.	14.3	9.2	57.7	53.6
w/o $\mathcal{L}_s$	11.4	6.9	59.8	56.0
Full (AKA *with* MRN)	**10.5**	**6.2**	**60.6**	**56.8**

Table 4: Selection of hyper-parameters and the comparison of our MRN and MBIN[3] in imbalanced LReID setting. "Position" denotes the index of the layers in ResNet-50. "$[\cdot,\cdot]$" indicates concatenation operation.

Associative Function $\mathcal{A}$			δ(mAP) ↓	mAP↑
$\mu_I - \mu_B, \sigma_I - \sigma_B$			**10.5**	**60.6**
$\mu_I/\mu_B, \sigma_I/\sigma_B$			12.3	58.9
$[\mu_I, \mu_B], [\sigma_I, \sigma_B]$			10.9	60.4
Normalization	Position	#Parameter	δ(mAP) ↓	mAP↑
BN	1,2,3,4,5	25,557.0k	26.7	49.4
MBIN[3]	1,2,3,4,5	+79.7k	16.7	58.0
	5	+33.8k	20.0	56.0
	4,5	+64.5k	18.9	57.2
	1,5	+34.0k	18.3	56.8
MRN	1,2,3,4,5	+48.4k	**10.5**	**60.6**
	5	+24.7k	13.8	58.4
	4,5	+41.5k	11.3	59.9
	1,5	+24.8k	13.7	58.6

4.4 Ablation Study

We conduct three groups of ablation experiments to study the effectiveness of our method: 1) the first group is to verify the improvement of the designed RN and meta-learning optimization. Analysing the performances of "*w/o* ARR." and "*w/o* Meta. & GMS" in Tab. 3, we find both ARR and GMS modules improve the model's capabilities in a complimentary way. Moreover, the performances of "*w/o* Meta." and "Full" demonstrate the meta-learning optimization further enhances the model's ability to prevent forgetting and generalise on the unseen domains, simultaneously; 2) as shown in Tab. 4, the second group is to experimentally select associative function in Sec. 3.2. The results show that the subtraction function is better than other solutions; 3) the last group is to explore the impact of the position in which we replace original batch normalization layers in ResNet-50 by MBINs or MRNs. Specifically, "1,2,3,4,5" denotes that we replace all the 53 BNs included in five convolutional layers of ResNet-50. The results demonstrate our MRN is a lightweight design compared to MBIN and replacing the BN in top and bottom blocks with MRN can obtain considerable performance gains with negligible overhead.

5 DISCUSSIONS AND LIMITATIONS

In this section, we discuss the advantages and limitations of our proposed methods by the following questions:

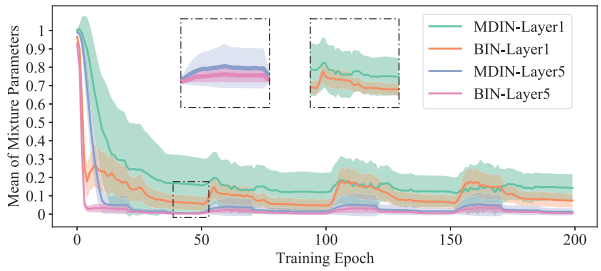

Figure 5: Visualization of mixture coefficients. The coefficients in Layer-1 have higher means than these in Layer-5. The top layers tend to IN's statistics, in order to achieve domain independence.

Q1: Is it feasible to meta-optimise all the network parameters? We assume the knowledge of ReID (*e.g.*, extracting areas of interest or objects, matching patterns and abstracting features) is contained in the weights of convolutional layers, while the normalization layer handles training noise (*e.g.*, distribution shift, gradient vanishing and outliers). In LReID models, it is reasonable to regard the parameters of normalization layers as meta parameters. Moreover, optimising all parameters by meta gradients may be too computationally expensive when training large-scale models, even using first-order approximations [6] in meta-optimization. Thus, we think applying our meta-optimization to whole network is inefficient.
Q2: What can the MRN benefit from meta optimization? We summarize the advantages from two aspects. 1) we visualize the learned mixture parameters in Fig. 5. Compared to BIN without meta optimization, MRN's mixture parameters have a larger variance, which increases the diversity of different feature channels. This enables the model to learn more common knowledge so as to acquire a better ability to generalize on the seen and unseen domains; 2) without the constraint of reconciliation term derived in Eq. (11), the model lacks informative second-order gradients, tending to a worse optimization. Hence, our MRN can mitigate the SPD.

6 CONCLUSION

In this paper, we proposed a new method named meta reconciliation normalization (MRN) and applied it to the LReID task, which is able to continuously learn more common knowledge with less domain dependency. Furthermore, to alleviate the stability-plasticity dilemma, we designed a meta-learning framework, allowing MRN to imitate the hippocampus in human brain and learn synaptic plasticity. It reconciles the knowledge interference between old and new tasks, so as to accomplish an effective knowledge transfer to learn well on new domains while mitigating catastrophic forgetting on old domains. Extensive experiments showed that our method outperforms other competitors on LReID tasks.

ACKNOWLEDGMENTS

This work was supported mainly by the LIACS Media Lab at Leiden University, and partially by the China Scholarship Council and the NSF of China under grant 62102061.

REFERENCES

[1] Zechen Bai, Zhigang Wang, Jian Wang, Di Hu, and Errui Ding. 2021. Unsupervised Multi-Source Domain Adaptation for Person Re-Identification. In *CVPR*. 12914–12923.

[2] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems* 33 (2020), 15920–15930.

[3] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. 2021. Meta Batch-Instance Normalization for Generalizable Person Re-Identification. In *CVPR*.

[4] Rosemary A Cowell, Morgan D Barense, and Patricnow S Sadil. 2019. A roadmap for understanding memory: Decomposing cognitive processes into operations and representations. *Eneuro* 6, 4 (2019).

[5] Florian Fiebig and Anders Lansner. 2014. Memory consolidation from seconds to weeks: a three-stage neural network model with autonomous reinstatement dynamics. *Frontiers in computational neuroscience* 8 (2014), 64.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*. PMLR, 1126–1135.

[7] Douglas Gray and Hai Tao. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 262–275.

[8] Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. 2020. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences* (2020).

[9] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. 2011. Person re-identification by descriptive and discriminative classification. In *scandinavian conference on image analysis*. Springer, 91–102.

[10] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *ICCV*. 1314–1324.

[11] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*. PMLR, 448–456.

[12] Jieru Jia, Qiuqi Ruan, and Timothy M. Hospedales. 2019. Frustratingly Easy Person Re-Identification: Generalizing Person Re-ID in Practice. In *BMVC*. BMVA Press, 117. https://bmvc2019.org/wp-content/uploads/papers/0702-paper.pdf

[13] Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim. 2020. Imbalanced continual learning with partitioning reservoir sampling. In *ECCV*. Springer, 411–428.

[14] Takashi Kitamura, Sachie K Ogawa, Dheeraj S Roy, Teruhiro Okuyama, Mark D Morrissey, Lillian M Smith, Roger L Redondo, and Susumu Tonegawa. 2017. Engrams and circuits crucial for systems consolidation of a memory. *Science* 356, 6333 (2017), 73–78.

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[16] Wei Li and Xiaogang Wang. 2013. Locally aligned feature transforms across views. In *CVPR*. 3594–3601.

[17] Wei Li, Rui Zhao, and Xiaogang Wang. 2012. Human Reidentification with Transferred Metric Learning. In *ACCV*.

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*. 152–159.

[19] Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 12 (2017), 2935–2947.

[20] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. 2020. Deep Representation Learning on Long-tailed Data: A Learnable Embedding Augmentation Perspective. In *CVPR*. 2970–2979.

[21] Chen Change Loy, Tao Xiang, and Shaogang Gong. 2010. Time-delayed correlation analysis for multi-camera activity understanding. *Int. J. Comput. Vis.* 90, 1 (2010), 106–129.

[22] Hyeonseob Nam and Hyo-Eun Kim. 2018. Batch-Instance Normalization for Adaptively Style-Invariant Neural Networks. In *NeurIPS*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/018b59ce1fd616d874afad0f44ba338d-Paper.pdf

[23] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*. 464–479.

[24] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks* 113 (2019), 54–71.

[25] Quang Pham, Chenghao Liu, and HOI Steven. 2021. Continual normalization: Rethinking batch normalization for online continual learning. In *International Conference on Learning Representations*.

[26] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. 2021. Lifelong Person Re-Identification via Adaptive Knowledge Accumulation. In *CVPR*. 7901–7910.

[27] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *CVPR*. 2001–2010.

[28] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. 2020. Learning to optimize domain specific normalization for domain generalization. In *ECCV*. Springer, 68–83.

[29] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. 2020. Few-Shot Class-Incremental Learning. In *CVPR*. 12183–12192.

[30] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*. 6924–6932.

[31] Riccardo Volpi, Diane Larlus, and Gregory Rogez. 2021. Continual Adaptation of Visual Representations via Domain Randomization and Meta-Learning. In *CVPR*. 4443–4453.

[32] Wei-Chun Wang, Nadia M Brashier, Erik A Wing, Elizabeth J Marsh, and Roberto Cabeza. 2018. Knowledge supports memory retrieval through familiarity, not recollection. *Neuropsychologia* 113 (2018), 14–21.

[33] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*. 79–88.

[34] Zheng Wei-Shi, Gong Shaogang, and Xiang Tao. 2009. Associating groups of people. In *BMVC*. 23–1.

[35] Guile Wu and Shaogang Gong. 2021. Generalising without Forgetting for Lifelong Person Re-Identification. In *AAAI*, Vol. 35. 2889–2897.

[36] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2016. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850* 2, 2 (2016).

[37] Bo Zhao, Shixiang Tang, Dapeng Chen, Hakan Bilen, and Rui Zhao. 2021. Continual representation learning for biometric identification. In *WACV*. 1198–1208.

[38] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*. 1077–1085.

[39] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *ICCV*. 1116–1124.

[40] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*. 3754–3762.

[41] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2021. Learning generalisable omni-scale representations for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).

[42] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. 2020. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *ECCV*. Springer, 140–157.