Augmented Geometric Distillation for Data-Free Incremental Person ReID

Yichen Lu  Mei Wang  Weihong Deng*
Beijing University of Posts and Telecommunications
{yichen.lu, wangmei1, whdeng}@bupt.edu.cn

Abstract

Incremental learning (IL) remains an open issue for Person Re-identification (ReID), where a ReID system is expected to preserve preceding knowledge while learning incrementally. However, due to the strict privacy licenses and the open-set retrieval setting, it is intractable to adapt existing class IL methods to ReID. In this work, we propose an Augmented Geometric Distillation (AGD) framework to tackle these issues. First, a general data-free incremental framework with dreaming memory is constructed to avoid privacy disclosure. On this basis, we reveal a “noisy distillation” problem stemming from the noise in dreaming memory, and further propose to augment distillation in a pairwise and cross-wise pattern over different views of memory to mitigate it. Second, for the open-set retrieval property, we propose to maintain feature space structure during evolving via a novel geometric way and preserve relationships between exemplars when representations drift. Extensive experiments demonstrate the superiority of our AGD to baseline with a margin of 6.0% mAP / 7.9% R@1 and it could be generalized to class IL. Code is available here†.

1. Introduction

Person re-identification (ReID) aims at identifying all images of the same person as the query from a gallery set of large scale. Training on a certain dataset empirically empowers a ReID system to expert in the corresponding domain. However, it inhibits the ReID system from adapting to the ever-changing environment, especially when dealing with the streamed data or a sequence of ReID tasks from incremental domains. We expect the system can widen its generalization in incremental domain and retain its capability in base domain simultaneously, which is, briefly, to accumulate new knowledge while avoiding Catastrophic Forgetting [12,29]. To overcome such similar limitation, Class Incremental Learning (CIL) [5,10,18,24,33,44] is proposed in classification task and efforts have been devoted to figuring out how to learn incrementally.

Despite the great success in CIL, it still faces challenges when directly adopted to a ReID system due to the strict privacy issues and the open-set retrieval setting. First, in CIL, reminding the networks of previous knowledge via replaying pre-stored exemplars is well-recognized [5,19,33] to alleviate catastrophic forgetting. However, replaying memory of real data faces risks of violating privacy licenses in ReID. Second, on one side, ReID is substantially an open-set retrieval task, which puts more attention on constructing a robust feature space when compared with the close-set classification, since not only representations but also their neighborhoods play key roles in retrieval ranking. On the other side, feeding new knowledge sequentially will inevitably cause semantic drift [49] and distort the preceding feature space, resulting in forgetting. Hence, there exists a critical yet ignored contradiction between stabilizing the feature space for preceding domains and adapting feature space for the incremental domain.

Considering the limitations aforementioned, we conduct further research on Incremental ReID (IL-ReID) [32] and propose a novel Augmented Geometric Distillation (AGD) framework which consists of Augmented Distillation (AD) and Geometric Distillation (GD). First, to tackle the privacy issue, we first construct a general data-free incre-
mental framework for IL-ReID (overview in Fig. 1), in which dreaming memory, generated by DeepInversion [47], drives replaying procedure without access to preceding real data. Unfortunately, due to the poor quality, directly replaying these dreaming exemplars will induce a phenomenon termed “noisy distillation”, during which, noisy knowledge will be transferred into the evolving model and aggravate forgetting. To alleviate this problem, we further propose to augment distillation itself. Enlightened by contrastive learning, we produce different views of memory and distill in a pair-wise and cross-wise pattern to strengthen the robustness and reduce the perturbation.

Second, to handle the contradiction caused by open-set retrieval property, we propose the geometric distillation (GD) tailored-made for retrieval task that our intuition is to maintain the structure of the preceding feature space while drifting instead of to stabilize the whole space and to penalize drift. The structure of preceding space is formulated with exemplars in dreaming memory. To prevent exemplars from drifting in their own manners arbitrarily and “rolling” the space structure, we encourages exemplars to drift in a consistent manner, so that the structure could be maintained via similarity criterion in a novel geometric way. This allows to adapt the feature space for new knowledge while preserving rich preceding information for retrieval, offering a compromise between learning and memorizing.

To conclude, our contributions could be summarized as:

i) We construct a data-free incremental framework for ReID with dreaming memory. It serves without privacy issues;

ii) We propose Augmented Distillation (AD), where distillation is conducted in a pair-wise and cross-wise pattern to address the “noisy distillation” phenomenon in dreaming memory;

iii) We propose Geometric Distillation (GD) to adapt new and preceding knowledge for retrieval tasks while maintaining space structure geometrically when drifting;

iv) We adapt mainstream solutions in CIL to ReID. Extensive experiments indicate that our AGD is superior to baseline with a margin of 6.0% mAP / 7.9% R@1 and it is promising to be generalized to CIL.

2. Related Work

2.1. Incremental Learning

Incremental learning [41] studies the problem of accumulating knowledge sequentially without catastrophic forgetting [12, 34]. To achieve this, methods based on parameters regularization [2, 23] attempted to penalize updating parameters for preceding tasks. Parameter-isolation based methods [1, 22, 28] dedicated extra parameters for new tasks. The recent mainstreams are on insights of replaying memory and distilling knowledge. LwF [24] first introduced distillation into IL. Dhar et al. [9] further proposed to constrain attention. iCaRL [33] and its improved variants [5, 18] introduced replay mechanism, where a memorizer is maintained to store limited samples for replaying. Following up on this, Wu et al. [44] and Hou et al. [19] corrected the bias in classifier. PODNet [10] distilled the pooled intermediate feature maps and GeoDL [36] constrained the geodesic flow in lower dimensions. TOPIC [40] and TPCIL [39] put their emphasis on topology of exemplars. Despite the remarkable insights, a compact memorizer is indispensable to all these replaying-based methods. As data-free frameworks, ARM [21] and ABD [37] replayed generated memory instead, but “noisy distillation” is ignored. SDC [49] measured the semantic drift without memory, but it was oriented to classification not retrieval.

2.2. Data-free Knowledge Transfer

As the seminal work by Hinton et al. [17], a basic solution was proposed to compress knowledge to student networks, on which to base, a line of works [25, 45, 46, 50] has reported more effective solutions. However, most methods above are data-driven. To address this flaw, some works managed to generate images. Lopes et al. [26] synthesized images via meta-data of networks. Bhardwaj et al. [3] synthesized samples via pre-recorded the centroids of classes. Some works [4, 14, 47] discovered constraining the generated images to match BatchNorm [20] statistics in teacher networks could close the gap between real image distribution. Similarly, Yoo et al. [48] and Chen et al. [6] trained a decoder to output class-conditional images. Moreover, some papers [8, 11, 30] transferred knowledge in an adversarial strategy. Despite notable progress above, how to retain knowledge in the pre-trained base model and incrementally learn from new tasks remains under-explored.

3. Background and Data-free Framework

In this section, we define the IL-ReID (Sec. 3.1) and clarify a data-free incremental framework for ReID (Sec. 3.2).

3.1. Problem Definition

In IL-ReID, to provide basic knowledge, $T_1$ leads the task sequence. Following the setting in LUCIR [19], the first task $T_1$ contains samples with a wide variety to achieve a strong base model $f^b_1$. After that, similar to the task-incremental setting [31] in CIL, data from a sequence of ReID tasks $T_2, T_3, T_4$... will be continually presented for incremental learning. In incremental training stage of $T_n$, the base model $f^{n-1}_b$ evolves into $f^n_b$. During this, we can only access to base model $f^{n-1}_b$ and dataset of $T_n$. Especially, no real data of base tasks $T_{n-1} = \{T_j\}^{n-1}_{j=1}$ is available. The dataset of $T_n$ is denoted as $D_{T_n} = \{(x^n_i, y^n_i)\}^{N_n}_{i=1}$, where $(x^n_i, y^n_i)$ is the i-th image and its ID. $N_n$ is the number of images in $D_{T_n}$. 

7330
and replaying to formulate the basic framework objective:

\[ \mathcal{L}_{base}(x, \hat{x}) = \mathcal{L}_{rep}(x, \hat{x}) + \lambda \mathcal{L}_e(\hat{x}), \]

where \( \mathcal{L}_e(\cdot) \) is the knowledge distillation term in replaying.

4. Proposed Method

Based on the design above, we propose a novel Augmented Geometric Distillation (AGD) framework to i): tackle the problem of “noisy distillation” in dreaming memory; ii): learn to adapt knowledge flexibly yet retentively via a geometric way for IL-ReID. In Sec. 4.1, we elaborate why noise is made, visualize how it impacts the distillation and how to alleviate it via our augmented distillation. In Sec. 4.2, we pave a brand-new path to retain knowledge when representations drift via geometrically maintaining structure in euclidean feature space.

4.1. Augmented Distillation

As discussed above, in order to circumvent privacy disclosure, we adopt dreaming data to acts as the memory. We expect data generated by DeepInversion serves as effectively as real data. However, a drawback of it is that mimicking the real image distribution imperfectly causes domain gap. For instance, an evident gap in visual level between dreaming exemplars and raw images exists due to the poor quality. And such domain gap weakens the robustness to data augmentation (e.g. crop, flip and REA [52]) as visualized in Fig. 4 (Left). The unexpected perturbation widens divergence of dreaming exemplars in feature space, which introduces overfitting of noise in typical pair-wise knowledge distillation as Fig. 4 (Right) and hence, aggravates forgetting. Even worse, such perturbation could bring more adverse impacts to our geometric distilling (detailed in Sec. 4.2) due to unstable relationships between dreaming exemplars.

Under such circumstance, the guidance from teacher is noisy, but data augmentation is necessary to promote sample diversity. For the best of both worlds, we propose to augment distillation itself. Specifically, to mitigate perturbation in each iteration, we first follow contrastive learning [7, 13, 15] to build two views of data \( \hat{x}' \) and \( \hat{x}'' \) with independent data augmentation. These views are from the same sample \( \hat{x} \) and should have robust features extracted by the teacher \( f_{\theta}^{n-1} \). However, due to the issue above, teacher outputs features of two views \( f_{\theta}^{n-1}(\hat{x}') \) and \( f_{\theta}^{n-1}(\hat{x}'') \) with divergence. To get more stable distilling effects, we average the gradients from pair-wise distillation of two views. Besides, \( \hat{x}' \) and \( \hat{x}'' \) are sampled from the same distribution and form a congruent pair \( (\hat{x}', \hat{x}'') \). For better consistency between views, we consider distilling across views symmetrically as illustrated in Fig. 3 (Left). The crisscross mechanism provides guidance from at least four views of the same.
Middle: Illustration of Geometric Distillation in Euclidean space that polyhedrons of classes built by feature points are encouraged to keep their similarity when evolving with scale and translation transforms to approximate the similarity of their subspaces. The process is driven by dreaming memory \( M \), which is a set of images generated by features in \( Z^{n-1} \). Right: Geometry interpretation of fundamental AAA similarity criterion.

4.2. Geometric Distillation

During incremental learning, feature space drift is inevitable and the arbitrary drift could roil the space structure (Fig. 5). Despite the success of penalizing the drift (e.g. Equ. 5), there exists a contradiction between preserving preceding knowledge, which detests the drift, and learning new knowledge, which leads to necessary drift. This issue bothers ReID particularly due to open-set retrieval property. To reach a compromise, we propose a brand-new solution, where space drift is not penalized explicitly. Our intuition is to maintain geometry structure of subspace of each class when drifting and keep the most discriminative representations for ranking in retrieval task. Then our method has the flexibility to fit new data and meanwhile preserves rich information in relationships with a geometric approach.

Definition 1 Given a Euclidean space \( Z \), if a bijection \( g(x) = rAx + t \) maps any two points \( x_1 \) and \( x_2 \) in \( Z \) into a Euclidean space \( Z' \) and \( d(g(x_1), g(x_2)) = r \cdot d(x_1, x_2) \), where \( d(\cdot, \cdot) \) is the Euclidean distance, we call \( Z' \) a similarity space to \( Z \) and \( r \) is the scale coefficient.

\(^*1\)A is an orthogonal matrix and \( t \) is a translation vector.
merating all positive pairs and convergence, triangles are
tors to be parallel with orientations. And three paralleled
f
where
i.e.
their own manners and damage their space structures.
Accumulating
New
Knowledge
adaptively and independently in each subspace.
rally scale into the polyhedrons in
A
nc structure simultaneously, given the bijection
shown in Fig. 3 (Middle).
ilarity via preserving geometric structure of polyhedrons, as
Z
ure points in
However, with the dreaming dataset
shown in Fig. 3 (Middle).
starting from the basic form
where
f
is the scale of
, to achieve this, we model
the loss as:
L
r
Z
n
, Z
j
n
f
x
n
i
n
j
n
= (x
i
, x
j
) ∈ P
Z
n
= {z
i
n
, z
j
n
= f
x
n
(ˆ
x
i
)}, Z
n
= {xn
−1 − z
i
n
−1 = f
x
n
(ˆ
x
i
)}. In this constraint, L
r
encourages two vectors to be parallel with orientations. And three paralleled
sides in Equ. 6 fulfill the AAA criterion of triangle similarity in its plane as demonstrated in Fig. 3 (Right). After enumerating all positive pairs and convergence, triangles are chained with shared sides and polyhedrons in
n
gradually scale into the polyhedrons in
n
. And since the scale coefficient
r
is not defined in Equ. 6 explicitly, \( r \) is learnt adaptively and independently in each subspace.
Based on discussion above, we now consider the translation vector \( \mathbf{t} \), which can be viewed as the drift of feature distribution. To allow the drift and maintain the geometric structure simultaneously, given the bijection \( f\hat{\theta}(\hat{x}) = rf\hat{\theta}^{-1}(\hat{x}) + \mathbf{t} \) and two samples \( \hat{x}_i, \hat{x}_j \),
\[
\Delta z_{ij} = z_i^n - z_j^n = f\hat{\theta}(\hat{x}_i) - f\hat{\theta}(\hat{x}_j) = r f\hat{\theta}^{-1}(\hat{x}_i) - r f\hat{\theta}^{-1}(\hat{x}_j) = r \Delta z_{ij},
\]
where \( \Delta z_{ij} \) is the scale of \( \Delta z_{ij}^{-1} \), which is a similar prob-
lem to constraining \( f\theta^n(\hat{x}) = r f\theta^{-1}(\hat{x}) \). Based on Equ. 6, we formulate the constrain with feature drift as:
\[
\mathcal{L}_G^r(Z^{n-1}, Z^n) = \mathcal{L}_G(\Delta Z^{n-1}, \Delta Z^n),
\]
where \( \Delta Z^n = \{\Delta z_{ij}^n | i \neq j, (\hat{x}_i, \hat{x}_j) \in \mathcal{P} \} \) and \( \Delta Z^{n-1} = \{\Delta z_{ij}^{n-1} | i \neq j, (\hat{x}_i, \hat{x}_j) \in \mathcal{P} \} \). Unlike \( \mathcal{L}_r \) only constraining orientations of individual features, in our \( \mathcal{L}_G^r \), features scale and drift in a consistent manner, which is of great importance to maintain the relationship between intra-exemplars. Meanwhile, the scale coefficient \( r \) and translation vector \( \mathbf{t} \) endow impressive plasticity when compared with much more critical criterion as \( \mathcal{L}_r \) or \( \mathcal{L}_s \) (MSE loss), where subspaces are enforced to be congruent with the preceding ones, \( i.e., f\hat{\theta}^n(\hat{x}) = f\hat{\theta}^{-1}(\hat{x}) \). Note that in the case of Equ. 8, the orthogonal matrix \( \mathbf{A} \) is an identity matrix that no rotation and reflection are adopted for simplicity and practicality (more details in Supp.).

4.3. Overall Objective and Algorithm
The overview of our AGD framework to conduct incremental learning is illustrated as Fig. 3. Due to the universal property of our AD mechanism, it is easy to integrate \( \mathcal{L}_G^r \) into it. The overall objective is formulated as:
\[
\mathcal{L}_{AGD}(\mathbf{x}, \hat{\mathbf{x}}) = \mathcal{L}_{rep}(||\mathbf{x} - \hat{\mathbf{x}}||) + \lambda \mathcal{L}_{r} G(\mathbf{x}; \mathcal{L}_G^r),
\]
and optimization procedures are summarized below.

Algorithm 1 Augmented Geometric Distillation (n-th task)
Input: Incremental dataset \( D_{T_n} \) and fixed base model \( f\theta^{-1} \).
Output: Converged evolving model \( f\theta^* \).
1: Generate dreaming memory \( \mathcal{M}_{T_{n-1}} \) with \( f\theta^{-1} \).
2: Initialize the evolving model \( f\theta^* \) with \( f\theta^{-1} \).
3: while not converged do
4: Sample and augment \( \mathbf{a} \subset D_{T_n} \rightarrow \mathbf{x} \).
5: Sample and augment twice \( \mathbf{\hat{a}} \subset \mathcal{M}_{T_{n-1}} \rightarrow \mathbf{\hat{x}}, \mathbf{\hat{a}}' \).
6: Calculate \( \mathcal{L}_{rep} \) (Equ. 1) with \( f\theta^*(\mathbf{x}), f\theta^*(\mathbf{\hat{x}}) \) and \( f\theta^*(\mathbf{\hat{a}}') \).
7: Calculate \( \mathcal{L}_{AD} \) (Equ. 3 and Equ. 8) between \( f\theta^*(\mathbf{x}), f\theta^*(\mathbf{\hat{a}}'), f\theta^*(\mathbf{\hat{x}}) \).
8: Calculate \( \mathcal{L}_{ADG} \) (Equ. 9) and backward.
9: Update \( \theta \) in \( f\theta^* \).
10: end while
11: Fix the evolving model \( f\theta^* \) for the next step as base model.

5. Experiments
5.1. Datasets and Evaluation Protocol
Market-1501 [51] contains 32,668 annotated images of 1,501 identities collected from 6 cameras totally. 12,936 images of 751 identities and 19,732 gallery images are used for training and test respectively.
PersonX [38] is a dataset generated by Unity under controllable cameras and environment. It has 9,840 images / 410 IDs for training and 35,952 images / 856 IDs for test.
Table 1. Comparison with mainstream families of methods in CIL. iCaRL [33] and LUCIR [19]: baseline solutions with Eqn. 2 $\kappa = kl$ (Eq. 4) and $\kappa = cos$ (Eq. 5) respectively. Oracle: training with supervision on according dataset(s). Note that all results are obtained on joint gallery (detailed in Sec. 5.1). For fair comparison, based on the basic representation loss in ReID, we only reproduced distillation parts of Distillation-based methods and tuned hyper-parameters for best performance. Bold and underline: best and second-best results.

5.3. Comparison with Other Methods

After replacing the memory built by real preceding data with dreaming data, we adapt typical methods in CIL to ReID for comparison as summarized in Tab. 1. We will analyze results mainly on MSMT17 $\rightarrow$ Market (M-to-M) and take AVG as overall performance.

Oracle: All results in “Oracle” family are achieved under supervised training protocol. As expected, both base and incremental tasks achieve satisfactory results after “Joint” training.

Finetune: Similar to results in CIL and AKA [32], fine-tuning directly induces catastrophic forgetting in base task. A vanilla approach to alleviate it is decreasing the finetuning directly induces catastrophic forgetting in base task. A vanilla approach to alleviate it is decreasing the finetuning learning rate. Despite the sacrifice of performance on incremental dataset, “1/10 lr” still yields better overall performance (1.5% mAP / 6.3% R@1) over “origin lr”.

Regularization: EWC [23] and MAS [2] constrain updating important parameters explicitly to balance knowledge learning and retaining. The results demonstrate its effectiveness in mitigating forgetting. On the other side, the explicit penalization of parameter updating disturbs the fitting on incremental dataset dramatically (10.4+% mAP / 5.9+% @1) compared with “Finetune”.

Distillation and Ours: Distillation-based methods intend to transfer knowledge from the base model to the evolving model to combat forgetting. This family of methods acts as the mainstream and leads the SOTAs in CIL. LwF [24] and iCaRL [33] focus on the distribution on preceding classes. With dreaming data as memory, iCaRL outperforms LwF greatly (15.5% mAP / 17.3% R@1). AKA [32] leverages a graph to manage knowledge. How-

MSMT17 [43] consists of 126,441 bounding boxes of 4,101 identities, of which 32,621 images of 1,041 identities form training set and the remaining set for testing.

Evaluation Protocol. After learning incrementally, we denote all test sets of seen tasks as $T = \{(Q_i, G_i)\}$, where $(Q_i, G_i)$ is the query set and gallery set of the $i$-th task. To evaluate the performance of model in all domains, we define the joint gallery as the intersection of all individual gallery sets, i.e., $G = \bigcup_{(Q_i, G_i) \in T} G_i$, and evaluate each query set in $G$. Finally, we take average performance as the overall results, i.e., $AVG = \frac{1}{|T|} \sum_{(Q_i, G_i) \in T} eval(Q_i, G_i)$, where $eval(\cdot, \cdot)$ outputs mean Average Precision (mAP) and Cumulated Matching Characteristics (CMC) curve as metrics.

5.2. Implementation Details

Following the baseline BoT [27] in ReID, we employ ResNet50 [16], initialized with parameters pre-trained on ImageNet [35], as our backbone. REA [52] ($sh=0.4$), BNNek [27] are adopted in all training process. Note that stride trick [27] is abandoned for fast training and inference. During inference stage, features after BNNek will be extracted for final ranking. SGD with learning rate of 0.01 is leveraged to update the parameters. We train the first base model $f_{\theta}^1(\cdot)$ for 90 epochs with warmup and decay learning rate at epoch 61. For incremental tasks, optimization lasts 80 epochs and decay occurs at epoch 41. We generate the dreaming memory until all classes have 40 exemplars or $|M|$ reaches 40960. When learning incrementally, batch-size is 128, 64 (16 identities $\times$ 4 samples) from $D_{T_n}$ and $M_{T_{1:n-1}}$ respectively. Settings of hyper-parameters are detailed in Sec. 5.4.
Table 2. Ablation studies. iCaRL (Eq. 2 \( \kappa = kl \)) and LUCIR (Eq. 2 \( \kappa = \cos \)) serve as baselines. “bs x2”; with larger batch size, i.e., 256 (=128+128). “ep x2”: train longer, i.e., 160 epochs. “re /2”: erasing less area of dreaming data (weak augmentation). Comparisons are marked in colors (blue: comparisons with iCaRL, green: comparisons with LUCIR).

<table>
<thead>
<tr>
<th>Method</th>
<th>MSMT17 ( \rightarrow ) Market (M-to-M)</th>
<th>Market ( \rightarrow ) MSMT17 (M-to-M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mAP R@1</td>
<td>mAP R@1</td>
</tr>
<tr>
<td>iCaRL [33]</td>
<td>27.6</td>
<td>30.7</td>
</tr>
<tr>
<td>LUCIR [19]</td>
<td>37.4</td>
<td>62.4</td>
</tr>
<tr>
<td>w/ AD</td>
<td>30.4</td>
<td>53.8</td>
</tr>
<tr>
<td>w/ AD</td>
<td>39.1</td>
<td>64.6</td>
</tr>
<tr>
<td>w/ AD</td>
<td>39.0</td>
<td>64.3</td>
</tr>
<tr>
<td>w/ AD</td>
<td>39.8</td>
<td>65.3</td>
</tr>
<tr>
<td>w/o M</td>
<td>23.5</td>
<td>46.0</td>
</tr>
<tr>
<td>w/ ( L^G_{\text{cos}} ) (bs x2)</td>
<td>41.2</td>
<td>66.8</td>
</tr>
<tr>
<td>w/ ( L^G_{\text{cos}} ) (ep x2)</td>
<td>38.5</td>
<td>64.9</td>
</tr>
<tr>
<td>w/ ( L^G_{\text{cos}} ) (re /2)</td>
<td>40.5</td>
<td>66.2</td>
</tr>
<tr>
<td>AGD</td>
<td>41.9</td>
<td>67.5</td>
</tr>
</tbody>
</table>

5.4. Ablation Studies and Parameter Analysis

In this section, we perform ablation studies and parameter analysis to investigate the contribution of each component in AGD to the final performance gain and evaluations on different settings. Results are shown in Tab. 2.

Effectiveness of Augmented Distillation. Augmented distillation aims at mitigating “noisy distillation”, particularly when driven by dreaming exemplars. Based on both baselines iCaRL and LUCIR, which transfer knowledge from two different perspectives (detailed in Eq. 4 and Eq. 5 in Sec. 4.1), our proposal brings gain of 1.8% mAP / 1.7% R@1 and 0.8% mAP / 1.0% R@1 respectively. When incorporated into our \( L^G_{\text{cos}} \), “AGD” outperforms “w/ \( L^G_{\text{cos}} \)” with a margin of 1.1% mAP and 1.2% R@1. The consistent improvements demonstrate its generalization on different distillation terms. To further investigate the rationale of AD mechanism, we train the networks with larger batch size, longer period and weak augmentation for dreaming data, which are the empirical approaches to stabilize training. However, “w/ \( L^G_{\text{cos}} \) (bs x2)” and “w/ \( L^G_{\text{cos}} \) (ep x2)” both fail to surpass “w/ \( L^G_{\text{cos}} \)”, “w/ \( L^G_{\text{cos}} \)” achieves marginal gain but is not capable to defeat “AGD”. The fact indicates that different from increasing batch size or training longer directly, AD mechanism digs more information in noisy exemplars effectively without hurting knowledge learning in incremental domain. And this is of great advances for such data-limited scenario as incremental learning.

Effectiveness of Geometric Distillation. Casting the distillation term on features is verified to be crucial as aforementioned. To further investigate the necessity of geometric distillation, we adopt \( L_{\text{cos}} \), \( L_{\text{R}} \) and \( L_{\text{F}} \) (MSE loss) to conduct extensive experiments. \( L_{\text{cos}} \) requires input pair-wise features to have the same orientations, while \( L_{\text{R}} \) and \( L_{\text{F}} \) enforce the features to remain unchanged, i.e., the preceding feature space is congruent after evolving (the bijection function is \( f_{\theta}^r(x) = f_{\theta}^{r-1}(x) \)). After tuning weighting parameter \( \lambda \), \( L_{\text{cos}} \), \( L_{\text{R}} \) and \( L_{\text{F}} \) all yield the satisfactory performance on M-to-M task (Tab. 1). But when compared with \( L_{\text{F}} \) (Tab. 2), 0.5% mAP / 1.1% R@1 decreases are shown. When we allow more necessary drift, \( L^G_{\text{cos}} \) achieves another 0.7% mAP / 0.2% R@1 improvements and surpasses LUCIR with advances of 1.2% mAP / 1.3% R@1. The gain justifies the superiority of geometric distillation, which makes our framework flexible yet retentive.

Effectiveness of Dreaming Memory. In our framework, the distillation term is completely driven by dreaming memory \( M \). In addition to the privacy issue, it plays a central role in building the similarity feature subspaces. In Tab. 1, replay-based family of methods surpasses others with a huge margin, which validates the necessity of dreaming memory. To further measure its contribution, we remove \( M \) and cast the distillation term \( L_{\text{cos}} \) directly on incremental dataset \( D_{\text{T}} \). An serious degradation of 26.9% mAP / 20.7% R@1 is observed in Tab. 2, especially in incremental domain, which completely ruins the results. This shows that \( M \) decouples the objectives of learning and reviewing, avoiding the potential interference.

Evaluation on \( \alpha \). \( \alpha \) determines the weight of cross part in AD (Eq. 3). According to the curve in Fig. 6, “\( \alpha = 0.9 \)” performs best, which demonstrates necessity of multi-view guidance for robust feature distillation.

Evaluation on \( \lambda \). \( \lambda \) is the weight factor of overall distillation term. Larger weight leads to less forget and less flexibility. Relatively, our framework is not sensitive to \( \lambda \) and “\( \lambda = 3 \)” yields the best results.

Evaluation on peers. peers, i.e., number of views in each distillation iteration, is fixed to 2 by default. A larger peers will provides guidance from more views and stronger...
regularization. However, as shown in Fig. 6 (Bottom Left), no extra gain is observed and we think it is because that average of guidances from too many views weakens the diversity in each view and over-regularizes the distillation.

**Evaluation on exemplars.** In general, more exemplars report better performance due to the more diversity of memory. In our framework, “≈ 40 exemplar” (40960 in total / 1041 IDs in MSMT17) outperforms other settings. But it is noteworthy that much less exemplars only result in about 0.5% degradation, which confirms the effectiveness of our method from the other side.

**5.5. Further Discussion**

**Learning More Tasks.** When learning incrementally with more tasks, “Finetune” performs similarly that encounters catastrophic forgetting. Dreaming memory alleviates such forgetting to a great extent, which brings iCaRL [33] and LUCIR [19] the huge gain. Furthermore, our AGD makes more advantage of $M$ and yields compelling improvements of 25.0+% mAP / 29.3+% R@1 over “Finetune” on both tasks settings.

### Table 3. Extensive experiments under more tasks settings.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSMT17 → Market → PersonX (M-to-M-to-P)</th>
<th>Market → PersonX (M-to-P-to-M)</th>
<th>PersonX → Market (P-to-M-to-P)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AVG mAP R@1</td>
<td>AVG mAP R@1</td>
<td>AVG mAP R@1</td>
</tr>
<tr>
<td>Oracle-Joint</td>
<td>48.0 73.1</td>
<td>82.3 93.2</td>
<td>83.9 93.4</td>
</tr>
<tr>
<td>Finetune</td>
<td>1.5 4.9</td>
<td>15.0 36.8</td>
<td>85.0 93.8</td>
</tr>
<tr>
<td>iCaRL [33]</td>
<td>16.4 35.2</td>
<td>68.5 85.8</td>
<td>82.6 93.0</td>
</tr>
<tr>
<td>LUCIR [19]</td>
<td>28.3 52.5</td>
<td>68.5 85.8</td>
<td>82.6 93.0</td>
</tr>
<tr>
<td>AGD</td>
<td>36.5 62.4</td>
<td>71.9 87.3</td>
<td>83.6 93.3</td>
</tr>
</tbody>
</table>

**6. Conclusion**

In this work, we have developed the AGD framework, which is an incremental framework tailored-made for ReID. It replays preceding knowledge via dreaming memory without privacy issue, and augments the “noisy distillation” in a novel crisscross pattern, uncovering the potential information in dreaming memory from noise. Moreover, we have stricken a better balance between learning and memorizing in a geometric way, where semantic drift is allowed to adapt new knowledge and preceding knowledge is preserved via maintaining space structure when drifting. Finally, superiority to typical solutions in CIL validates its promising potential when adopted in ReID, open-set incremental tasks and even more conventional CIL.

### Acknowledgement

This work was partially supported by the National Natural Science Foundation of China under Grants No. 61871052 and 62192784 and Alibaba Group through Alibaba Innovation Research Program.
References


