Towards Discriminative Representation Learning for Unsupervised Person Re-identification
Supplementary Material

Takashi Isobe[1,2], Dong Li[1], Lu Tian[1],
Weihua Chen[3], Yi Shan[1], Shengjin Wang[2*]
[1]Xilinx Inc., Beijing, China.
[2]Department of Electronic Engineering, Tsinghua University
[3]Machine Intelligence Technology Lab, Alibaba Group
{dongl, lutian, yishan}@xilinx.com jbj18@mails.tsinghua.edu.cn
wgsg@tsinghua.edu.cn kugang.cwh@alibaba-inc.com

1. Overview

In this supplementary material, we present additional experimental results and analysis.

- We report the results on two additional unsupervised person re-ID benchmarks.
- We report the results of different network backbones on multiple benchmarks.
- We present detailed ablation study on the effect of different training losses.
- We discuss the relationships between our cluster-wise contrastive learning (CCL) algorithm and other unsupervised feature learning methods.
- We show more curves of our progressive domain adaptation (PDA) strategy in terms of clustering and re-ID performance.

2. Additional Benchmarks

We conduct experimental evaluations on two additional unsupervised person re-ID benchmarks to supplement this paper, i.e., MSMT-to-Duke and MSMT-to-Market. Table 1 presents the results of our approach and previous state-of-the-art methods. For fair comparisons, we re-implement MMT [5] using the same DBSCAN clustering algorithm and train the same 80 epochs with our method. Equipped with the same ResNet-50 backbone, our approach outperforms prior state-of-the-art methods by a large margin, e.g., surpassing MMT [5] and SpCL [6] by 7.9% and 5.6% on MSMT-to-Market, respectively.

3. Different Network Backbones

We report results with IBN-ResNet-50 as backbone on multiple benchmarks. The IBN-ResNet-50 is an improved network by instance normalization and batch normalization modules. Table 4 shows that using the stronger IBN-ResNet-50 backbone can achieve better results than ResNet-50. Compared to previous state-of-the-art methods with the same IBN-ResNet-50, our method achieves consistent performance improvement on all the eight benchmarks, e,g. +8.1% mAP over MMT and +6.3% mAP over SpCL on MSMT-to-Market.

4. More Ablation Study

Table 2 compares the results with different training losses. The experiment of (v) means the baseline method with the common cross entropy and triplet losses. Combined with our CCL loss, we can improve the baseline by 5.9% and 7.5% mAP on Market-to-Duke and Duke-to-Market, respectively.

5. Different Unsupervised Methods

Compared to instance-wise unsupervised methods. Recently, the self-supervised contrastive learning framework has been explored in these unsupervised feature learning methods [9, 7, 4]. They typically rely on *instance-wise* pairs to maximize agreement between differently augmented views of the same image. Our method adopts the similar N-pair loss formulation with these contrastive learning based methods. However, we cluster different images with similar semantics into a group. We hypothesize that directly applying instance-wise pairs may hinder the learning of discriminative features for distinguishing semantic classes due to the different optimization targets. We add the improved

[*] Corresponding author

MoCoV2 baseline [4] in Table 3 and find it also fails on our unsupervised re-ID task, which validates our intuition.

Compared to cluster-based unsupervised methods. We discuss the relationships between our CCL algorithm with other cluster-based unsupervised feature learning methods [1, 3, 2]. First, SwAV [3] clusters the differently augmented views of the *same* image to obtain prototypes vectors and constructs a swapped prediction problem without contrastive learning. Differently, we cluster *different* images to multiple pseudo classes and use a momentum network for contrastive learning iteratively. Second, Both DeepCluster [2] and SeLa [1] alternate between feature learning and clustering (self-labeling). In the feature learning step, they only use the common cross entropy loss but we integrate the proposed CCL. Besides, they handle the degenerate solution of clustering by heuristics such as sampling training data based on a uniform distribution over the pseudo classes [2], or maximizing the information between data indices and labels [1]. The degenerate solutions of clustering need be considered in their unsupervised feature learning setting where the network is trained from scratch. However, we focus on a unsupervised domain adaptation problem where the network can be trained with the supervised source data. With this meaningful initialization, we do not need particular design on the clustering algorithm during the CCL process.

6. Effect of Progressive Training

Figure 1 compares our progressive domain adaptation strategy and the two-stage training baseline on multiple benchmarks. According to the clustering performance, the results imply that our method can gradually reduce label noise and yield cleaner clusters than baseline. According to the re-ID performance, the results imply that our method can learn better features gradually and achieve higher recognition performance.

Table 1. Performance comparisons on two additional benchmarks for unsupervised person re-ID. † means our re-implementation of [5] with the DBSCAN clustering algorithm for fair comparisons. The results of "Ours*" are obtained by combining the proposed method with soft cross-entropy loss, soft triplet loss and mutual learning strategy introduced by [5]

Methods	MSMT-to-Duke		MSMT-to-Market	
	mAP	rank-1	mAP	rank-1
MMT † [5]	60.6	75.0	75.2	90.4
SpCL [6]	-	-	77.5	89.7
Ours	**67.5**	**80.6**	**83.1**	**94.1**
Ours*	**68.4**	**81.2**	**84.1**	**94.5**

Table 2. Ablation studies of training losses. (v) means the "Baseline" and (viii) means the "Baseline+CCL" methods in the paper.

Methods	$\mathcal{L}_{\text{ce}}$	$\mathcal{L}_{\text{tri}}$	$\mathcal{L}_{\text{ccl}}$	Markt-to-Duke		Duke-to-Market	
				mAP	rank-1	mAP	rank-1
(i)				31.1	48.8	33.7	62.3
(ii)	✓			50.6	67.8	61.7	83.5
(iii)		✓		23.9	47.0	27.3	45.4
(iv)			✓	56.8	71.9	67.5	84.2
(v)	✓	✓		53.7	69.9	63.6	82.5
(vi)	✓		✓	58.8	74.1	70.4	86.5
(vii)		✓	✓	57.9	72.6	68.7	85.1
(viii)	✓	✓	✓	**59.6**	**75.0**	**71.1**	**87.8**

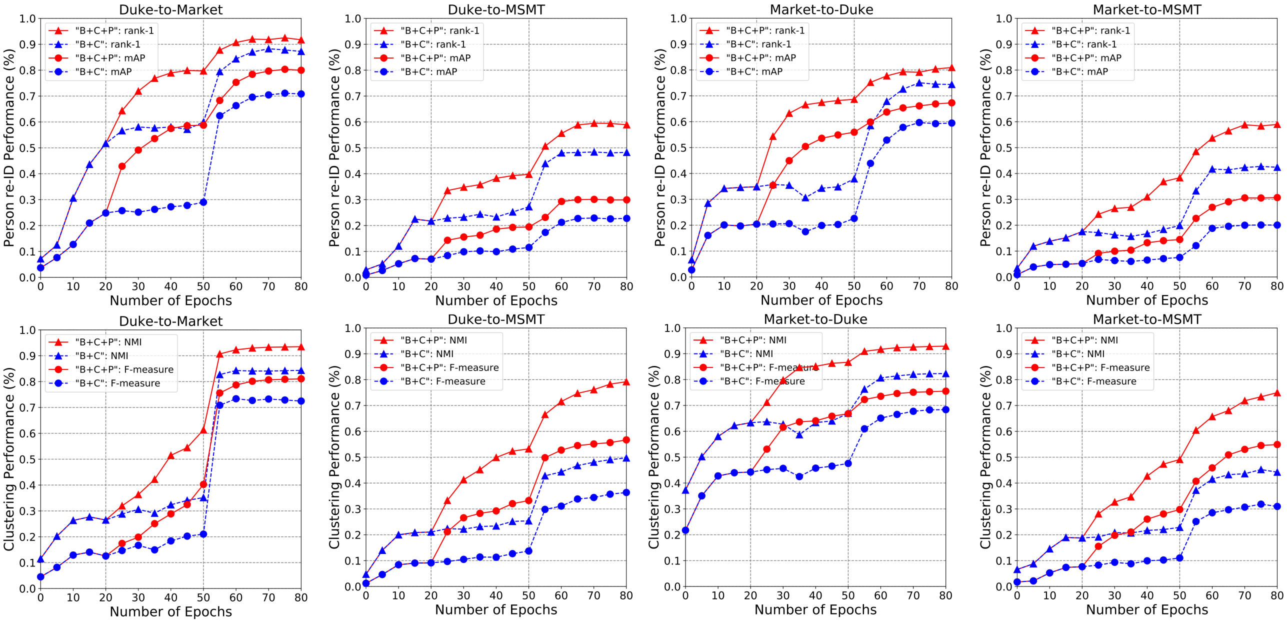
Table 3. Performance comparisons with other contrastive learning methods and our method. "†" means our implementation based on the official code. The cross-entropy and triplet losses are not used for all the experiments here.

Method	Market-to-Duke		Duke-to-Market	
	mAP	rank-1	mAP	rank-1
SupCon† [8]	66.0	79 .4	75.4	88.1
InstDisc† [9]	1.9	4.1	2.4	5.9
MoCo† [7]	10.3	17.7	11.1	26.2
MoCoV2† [4]	9.5	17.4	10.8	25.2
CCL (Ours)	**56.8**	**71.9**	**67.5**	**84.2**

Table 4. Performance comparisons with different network backbones. We re-implementation [5] with the DBSCAN clustering algorithm for fair comparisons.

Backbone	Methods	Market-to-Duke		Duke-to-Market		Market-to-MSMT		Duke-to-MSMT	
		mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
ResNet-50	MMT [5]	62.7	76.8	73.5	89.7	24.4	50.7	25.2	53.2
	Ours	**70.8**	**83.5**	**83.4**	**94.2**	**35.8**	**65.8**	**36.3**	**66.6**
IBN-ResNet-50	MMT [5]	65.2	79.6	76.9	92.0	30.4	56.0	30.6	60.1
	Ours	**72.9**	**85.8**	**86.3**	**95.8**	**40.1**	**70.4**	**40.1**	**70.7**

Backbone	Methods	MSMT-to-Duke		MSMT-to-Market		PersonX-to-Market		PersonX-to-MSMT	
		mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
ResNet-50	MMT [5]	60.6	75.0	75.2	90.4	70.7	86.2	18.2	39.5
	SpCL [6]	-	-	77.5	89.7	73.8	88.0	22.7	47.7
	Ours	**68.4**	**81.2**	**84.1**	**94.5**	**79.6**	**92.5**	**28.9**	**53.2**
IBN-ResNet-50	MMT [5]	63.3	78.1	78.1	91.9	73.8	90.1	21.5	44.8
	SpCL [6]	-	-	79.9	92.0	77.9	90.5	25.4	50.6
	Ours	**70.9**	**83.2**	**86.2**	**96.3**	**81.5**	**94.2**	**31.3**	**55.8**



(a) Duke-to-Market

(b) Duke-to-MSMT

(c) Market-to-Duke

(d) Market-to-MSMT

Figure 1. Illustration of the proposed progressive domain adaptation (PDA) in term of the re-ID and clustering performance. The first line shows the re-ID performance (mAP and rank-1 accuracy) on the test set. The second line shows the clustering performance (NMI and F-measure) on the training set. "B", "C" and "P" represent the baseline, the proposed CCL, and PDA methods, respectively. Both "B+C" and "B+C+P" methods use the same ResNet-50 backbone, the same DBSCAN clustering algorithm and the same 80 training epochs in total. Without PDA, we train the baseline model by 50 epochs on source and then fine-tune it on target until 80 epochs.

References

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020.

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[5] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020.

[6] Yixiao Ge, Dapeng Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[8] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

[9] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.