# Unsupervised Object Re-identification via Instances Correlation Loss

Qing Tang, Kang-Hyun Jo
*Department of Electrical, Electronic and Computer Engineering*
*University of Ulsan*
Ulsan, Korea
tangqing@islab.ulsan.ac.kr; acejo@ulsan.ac.kr

***Abstract*—This paper studies the fully unsupervised object re-identification (re-ID) problem which can learn re-ID without any human-annotated labeled data. Recent works show that self-supervised momentum contrastive learning is an effective method for unsupervised object re-ID, but they neglect to optimize one important component - the similarity relationships among instances. Previous works focus on enforcing instance-to-centroid learning, which does not fully utilize the inter-instances information. Thus, we propose an Instances Correlation Loss (ICL) to enforce instance-to-instance learning in each training iteration. Experimental results show that the proposed ICL effectively boost the performance, which demonstrates that learning strategy is also a central importance to unsupervised re-ID task. Extensive experiments are performed on three mainstream person re-ID datasets and one vehicle re-ID dataset.**

***Index Terms*—Person re-identification, fully unsupervised learning, vehicle re-identification**

## I. Introduction

Object re-identification (re-ID), is a fundamental task in intelligent surveillance systems, aims to retrieve a particular object instance across different camera views or scenes. Common object re-ID problems include person re-ID and vehicle re-ID, as illustrated in Fig. 1. Given a query image, the goal of an object re-ID system is to find images from the gallery which contain the same identity as the query image. Green boxes in Fig. 1 denote the matching identity between query and gallery.

In the past decade, supervised object re-ID achieved significant progress, however, supervised methods required substantial human-annotated labeled data for satisfying performance. Therefore, recent works are beginning to focus on unsupervised object re-ID methods.

The state-of-the-art unsupervised re-ID method [5]–[7] achieved significant success by utilizing strong self-supervised contrastive learning mechanisms, i.e., the Memory Bank approaches [8] and Momentum Contrast (MoCo) [9] approach. The Memory Bank and MoCo are designed for the unsupervised instance discrimination task, which learn the discriminative features of an image by matching its random augmented views. Different from the instance discrimination task, contrastive learning-based object re-ID tasks first roughly classify all images into clusters then conduct instance-to-centroids learning in feature space [5]–[7].

The discrepancy in learning strategy causes the advantage of self-supervised contrastive learning mechanism to be not fully utilized in object re-ID tasks. The problems is that similarity relationships among instances in each training iteration are neglected in current framework. MoCo [9] demonstrated the importance of maintaining consistent representation for unsupervised learning, however, the representation of instances (query) and centroids are less consistent, as illustrated in Fig. 2. The representation of instances is extracted by encoder in every training iteration, but the representation of cluster centroids are generated by momentum encoder before every training epoch. To mine similarity relationships from consistent representation, we further propose a Instance Correlation Loss


(a) Person re-ID from Market-1501

(b) Person re-ID from DukeMTMC-reID

(c) Person re-ID from MSMT17

(d) Vehicle re-ID from VeRi-776

Fig. 1. The examples of object re-ID images from three person re-ID datasets and one VeRi-776 dataset. Green boxes denote the matching identity between query and gallery. (a) Market-1501 [1], (b) DukeMTMC-reID [2], (c) MSMT17 [3], and (d) VeRi-776 datasets [4].

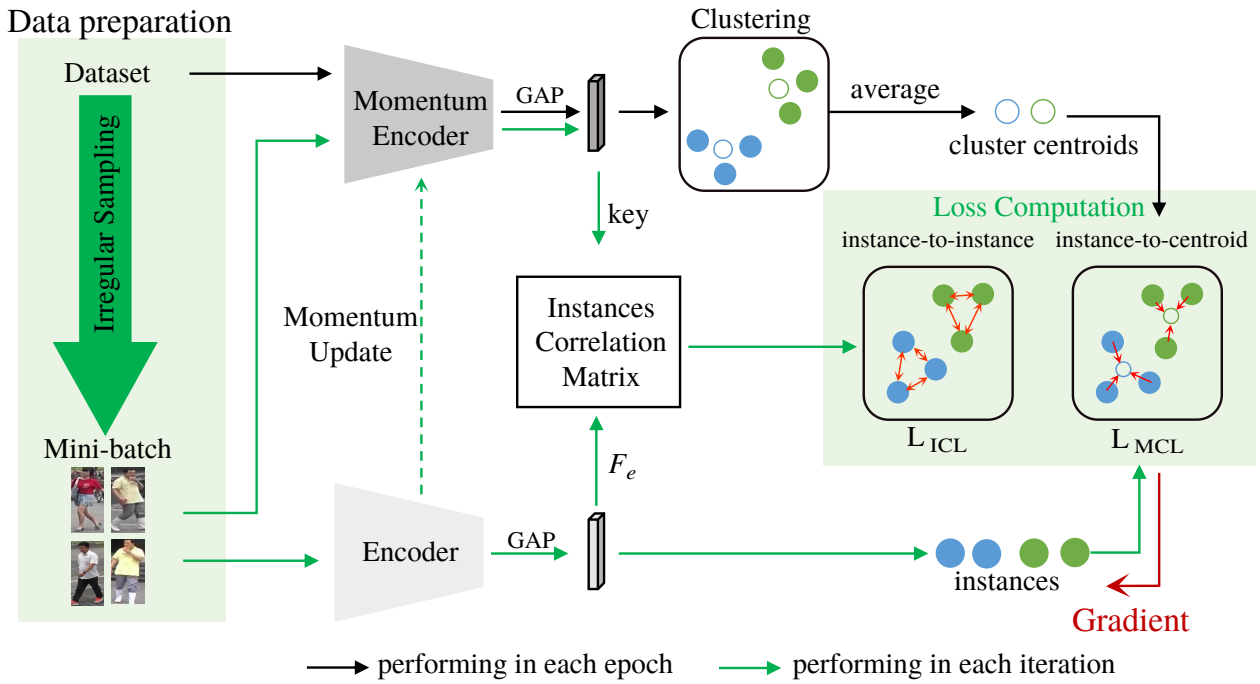


Fig. 2. The illustration of the proposed fully unsupervised object re-ID framework.

(ICL) $L_{ICL}$ to increase compactness of intra-class instances. Here we implement state-of-the-art self-supervised contrastive learning mechanism MoCo as baseline framework to perform experiments.

In summary, the contributions of this work are two-fold.

- We proposed an Instance Correlation Loss (ICL) to solve the inconsistency problem by enforcing instance-to-instance learning in each training iteration.
- Extensive experiments demonstrate that the proposed object re-ID framework shows exceptionally strong performances in four object re-ID datasets.

The remainder of this paper is organized as follows. Section II summarizes the related work. Section III describes the re-ID method. In section IV, the implementation details and extensive experimental results are reported and analyzed. Finally, Section V concludes this paper.

## II. Related Works

### A. Unsupervised object re-ID

Common unsupervised object re-ID problems include person re-ID [5], [6], [10] and vehicle re-ID [5], [11]. Based on the training strategy, current unsupervised re-ID methods can be summarized in two categories: unsupervised domain adaptation (UDA) and fully unsupervised methods.

UDA methods [11] learn to extract discriminative features for re-ID by transferring knowledge from a labeled source dataset to the unlabeled target dataset. The key is to reduce the gap between source and target domain. A common operation is to train the network on the labeled source and the unlabeled target dataset simultaneously [12]. UDA methods had taken the lead in performance but they still need labeled information.

Fully unsupervised re-ID has recently gained attention because it does not require any labeled information. Lin et al. [13] first proposed a fully unsupervised method for re-ID, called Bottom-Up Clustering (BUC), which merges a fixed number of clusters to fine-tune the model step by step. Wang et al. [14] formulated re-ID as a multi-label classification task, optimized the network under the supervision of self-predicted and online pseudo multi-class labels. Based on the [14], Tang et al. [15] leveraged the eligible neighbors as additional reference information to further boost the model performance in ranking accuracy. Recently, contrastive learning methods have pushed the fully unsupervised re-ID performance to a new height. To improved the quality pseudo labels, Tang et al. [16] proposed a Multiple pseudo Labels Joint Training (MLJT) strategy to predict multiple pseudo labels for each image by similarity measurement and clustering algorithm.

### B. Contrastive Learning in Re-ID

Contrastive learning [8], [9] has long been studied in the area of unsupervised visual representation learning. The main idea is to form positive pairs by augmenting the same sample, and to form negative pairs by treating each sample as a single class. Then, the model is trained by contrastive loss to pull positive samples together and push negative samples away.

Recent unsupervised re-ID works adopt contrastive learning method, i.e., Memory Bank [8] or MoCo [9] architecture, as the general pipeline. SpCL [5] considered the instance-level contrastive loss does not perform well on object re-ID, therefore they proposed instance-to-centroid contrastive learning. SpCL first clustered all samples into clustered inliers or unclustered outliers by clustering algorithm DBSCAN [2].

SpCL proposed a self-paced contrastive learning framework, which involved cluster-level and un-clustered instance-level features into a hybrid memory for contrastive learning. CAP [7] and IICS [10] noticed that features distribution discrepancy among cameras harms model learning discriminative features among object identity. To address this issue, they introduced camera information to build camera-aware proxies to perform instance-to-proxy contrastive learning. [6] considered above methods only focus on cluster or proxy centroids-towards learning but neglect inter-instance relationship. Therefore, [6] proposed a hard instance contrastive loss to pull a hard positive sample closer and push negative samples away in a mini-batch.

## III. Proposed Method

The MoCo-based re-ID contrastive learning framework in ICE [6] is adopted in this work as the baseline framework as shown in Fig. 2.

### A. Momentum Contrast Learning

The objective of our work is to obtain a superior re-ID network, which can produce similar features for the same identity and produce distinct features for different identities. To achieve this goal, momentum contrast learning architecture MoCo [9] with InfoNCE loss [17] is used as the baseline to enforce instance-to-centroid learning. The framework of the proposed method is illustrated in Fig. 2.

The encoder and the momentum encoder are used to generate representations of instances and cluster centroids, respectively. We denote parameters of the Encoder as $\theta_e$, and parameters of the Momentum Encoder as $\theta_{me}$. $\theta_e$ is updated in each training iteration by gradient back-propagation. The momentum encoder, served as a robust encoder, updated by $\theta_e$ with a momentum coefficient $m$ after every iteration as follows,

$$\theta_{me} = m\theta_{me} + (1-m)\theta_e \tag{1}$$

Before each training epoch starts, given an unlabeled training dataset $X = \{x_1, ..., x_N\}$, all images representations $F_{me} = \{f_{me,1}, ..., f_{me,N}\}$ are extracted by the momentum encoder. Then, unsupervised dense-based clustering algorithm DBSCAN [2] clusters $F_{me}$ into $N_C$ numbers of clusters. After that, cluster centroids $C = \{c_0, ..., c_{N_C}\}$ are computed as the mean vector of all instances in the cluster. This clustering results are used to split $X$ into mini-batches.

In each training iteration, given an sampled mini-batch $B$, $F_e = \{f_{e,1}, ..., f_{e,N_B}\}$ are extracted by the encoder as representations of instances.

To pull intra-class instances close to their corresponding centroids and push other centroids away, the loss of momentum contrast learning $L_{MCL}$ of an instance is designed based on InfoNCE loss [17] as follows,

$$L_{MCL} = -log\frac{exp(f_{e,i} \cdot c^+)/\tau}{exp(f_{e,i} \cdot C)/\tau} \tag{2}$$

, where $f_{e,i} \cdot c^+$ computes the distance between the instance $x_i$ and its corresponding cluster centroid $c^+$, where $c^+ \in C$. $f_{e,i} \cdot C$ represents distances among $x_i$ and all cluster centroids. $\tau$ is the temperature hyper-parameter.

### B. Instances Correlation Loss

Training re-ID model only using momentum contrast learning with Eq. (2) still has two open problems.

1) The representations $f_{e,i}$ and $C$ are less consistent in updating states. More specifically, $f_{e,i}$ was extracted by the encoder in each training iteration, but the $C$ are generated by momentum encoders all over the past iteration. Although the momentum update is performed after every iterations, the momentum encoder is not fully made use of in the re-ID task.
2) The instance-to-instance learning is ignored. Mining similarity relationship among instances is also beneficial to re-ID model performance [14], [15].

Thus, we proposed an Instance Correlation Loss $L_{ICL}$ to solve the inconsistency problem by enforcing instance-to-instance learning in each training iteration.

Previous method ICE designed an instance contrastive loss to enforce instance-to-instance learning, in which only one hardest positive and multiple negative samples are involved [6]. Instance contrastive loss neglects the relationship among positive samples. In other words, the instance contrastive loss lacks intra-class compactness because it did not enforce positive samples learning close to each others. Instead of using instance contrastive loss in [6], our proposed instances correlation loss involves multiple positive samples to increase intra-class compactness.

For the mini-batch $B$, L2 normalized key $K = \{k_1, ..., k_{N_B}\}$ are additionally extracted by momentum encoder. We then compute cosine similarities to build correlation matrix $M = F_e \cdot K^\top$, size as $N_B \times N_B$. $M$ is bounded by $[-1, 1]$. The $L_{ICL}$ is computed by directly regress the M to target matrix $T$ as follows,

$$L_{ICL} = ||M - T||^2 \tag{3}$$

. The rules of initialize $T$ is simple. If two instances in $B$ have same class $c_i$, the corresponding value in $T$ equals to 1; Otherwise, the value equals to $-1$. Therefore, the target matrix $T$ is a binary matrix, which consists by $-1$ and 1. $T$ has same size with $M$ as illustrated in Fig. 3. The overall loss of our proposed framework is the summation of Eqn. (2) and Eqn. (3) as follows,

$$L = L_{MCL} + L_{ICL} \tag{4}$$

.

## IV. Experiments

### A. Datasets and Evaluation Metrics

The experiments are performed on four large-scale and mainstream datasets, i.e., three person re-ID datasets, and one vehicle re-ID datasets.

Market-1501 [1] (Market) is a person re-ID dataset, which has 6 cameras and 32,217 person images of 1,501 identities in total.

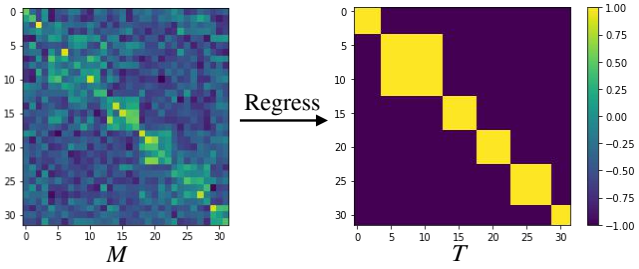


Fig. 3. The example of $M$: instances correlation matrix, and $T$: target matrix.

| Loss function | Market-1501 | | DukeMTMC-reID | |
|---|---|---|---|---|
| | mAP | Rank-1 | mAP | Rank-1 |
| Baseline | 83.3 | 93.6 | 66.9 | 81.5 |
| Contrastive loss | 83.9 | 93.8 | 67.5 | 82.0 |
| ICL (Ours) | **84.5** | **94.5** | **69.3** | **82.7** |

TABLE I. Ablation study on using different loss functions for instance-to-instance learning

| Method | Market-1501 | | | | DukeMTMC-reID | | | | MSMT17 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R5 | R-10 | mAP | R-1 | R-5 | R-10 |
| BUC [13] | 29.6 | 61.9 | 73.5 | 78.2 | 22.1 | 40.4 | 52.5 | 58.2 | - | - | - | - |
| HCT [18] | 56.4 | 80.0 | 91.6 | 95.2 | 50.7 | 69.6 | 83.4 | 87.4 | - | - | - | - |
| MMCL [14] | 45.5 | 80.3 | 89.4 | 92.3 | 40.2 | 65.2 | 75.9 | 80.0 | 11.2 | 35.4 | 44.8 | 49.8 |
| DSCE [19] | 61.7 | 83.9 | 92.3 | - | 53.8 | 73.8 | 84.2 | - | 15.5 | 35.2 | 48.3 | - |
| SpCL [5] | 79.1 | 88.1 | 95.1 | 97.0 | 65.3 | 81.2 | 90.3 | 92.2 | 19.1 | 42.3 | 55.6 | 61.2 |
| CAP [7] | 79.2 | 91.4 | 96.3 | 97.7 | 67.3 | 81.1 | 89.3 | 91.8 | 36.9 | 67.4 | 78.0 | 81.4 |
| Group Sampling [20] | 79.2 | 92.3 | 96.6 | 97.8 | 69.1 | 82.7 | 91.1 | 93.5 | 24.6 | 56.2 | 67.2 | 71.5 |
| ICE [6] | 82.3 | 93.8 | 97.6 | 98.4 | **69.9** | **83.3** | **91.5** | **94.1** | 38.9 | 70.2 | 80.5 | 84.4 |
| Ours | **84.5** | **94.5** | **98.2** | **98.8** | 69.3 | 82.7 | 90.1 | 92.3 | **42.4** | **72.7** | **82.0** | **85.2** |

TABLE II. Experimental results of our proposed method and state-of-the-art fully unsupervised re-ID methods on three person re-ID datasets. IS: Irregular Sampling. ICL: Instances Correlation Loss. The top result is highlighted in bold and the second best result is shown in blue.

| Method | VeRi-776 | | | |
|---|---|---|---|---|
| | mAP | R-1 | R-5 | R-10 |
| SpCL [5] | 36.9 | 79.9 | 86.8 | 89.9 |
| Ours | **39.5** | **83.7** | **88.4** | **90.7** |

TABLE III. Experimental results of our proposed method on vehicle re-ID datasets VeRi-776.

DukeMTMC-reID [21](Duke) is a person re-ID dataset, which has 8 cameras and 36,411 person images of 1,404 identities in total.

MSMT17 [3] (MSMT) is a person re-ID dataset, which has 15 cameras and 126,441 person images of 4,101 identities in total.

VeRi-776 [4] (VeRi) is a vehicle re-ID dataset, which has 20 cameras and 51,003 vehicle images of 775 identities in total.

Two evaluation metrics are used to measure model performance. The first one is Mean Average Precision (mAP) (%). Another one is the Cumulative Matching Characteristic (CMC) curve. The CMCs (%) of Rank-1 (R-1), Rank-5 (R-5), and Rank-10 (R-10) are reported, which represents the probability of top-1, top-5, and top-10 ranked gallery samples containing the query identity, respectively.

### B. Implementation Details

ImageNet pre-trained ResNet-50 is used as the encoder and the momentum encoder. A batch normalization layer and an $L_2$-normalization layer are added after the last global pooling layer of ResNet-50 to generate 2048-dimensional features. The input images are resized to $256 \times 128 \times 3$. The size of training mini-batch $N_b$ is 32. The network is trained by the Stochastic Gradient Descent (SGD) with a learning rate of 0.00055, 50 epochs in total. Hyper-parameters $m = 0.999$, $\tau = 0.05$, $P = 12$ are used in all experiments for fair comparisons, except in hyper-parameter analysis experiments. The experiments are performed on one NVIDIA Titan 1080Ti GPU with 11 GB of memory. The total training time is around 3 hours on Market-1501 and DukeMTMC-reID, and 6 hours on MSMT17 and VeRi-776.

### C. Effectiveness of the Instances Correlation Loss

To test the validity of the proposed instances correlation loss, we compare it against the baseline method and contrastive loss [6]. The results are reported in Table I. The baseline method only using $L_{MCL}$ to perform instance-to-centroids learning, which outputs unsatisfactory performance in mAP$= 83.3\%$ and in Rank-1 $= 93.6\%$ on Market-1501, and in mAP$= 66.9\%$ and Rank-1 $= 81.5\%$ on DukeMTMC-reID.

Two instance-to-instance learning loss, including contrastive loss and our proposed ICL, boost the model performance from the baseline. The consistent improvements demonstrated the importance of mining information among instances.

The idea of contrastive loss in [6] is to pull the hardest neighbor closer and push all negative samples in the same mini-batch away. Limited by the function of contrastive loss, only one positive sample can be involved. Our proposed ICL involves all positive samples and negative samples by directly regressing the correlation matrix of each mini-batch to its target matrix. The performance of ICL remarkably surpasses the contrastive loss. The improvements demonstrate the necessity of involving more positive samples and the effectiveness of our proposed instances correlation matrix and ICL loss.

### *D. Comparisons with the State-of-the-Arts*

The comparisons with the State-of-the-Arts fully unsupervised methods on Market-1501, DukeMTMC-reID, and MSMT17 are reported in Table II. On Market-1501, our method achieves the best performance with mAP$= 84.5\%$ and Rank-1 $= 94.5\%$. Compared to the best fully unsupervised method ICE, our method achieve good and competitive results on DukeMTMC-reID. Moreover, our method outperforms ICE by $3.5\%$ in mAP and $2.5\%$ in Rank-1 in the largest and most difficult person re-ID datasets MSMT17. Comparisons are also performed in vehicle re-ID dataset VeRi-776 in Table III. We obtain mAP$= 39.5\%$ and Rank-1 $= 83.7\%$, which considerably outperforms SpCL. The superior performance indicates that the effectiveness of our proposed instance-to-instance learning loss $L_{ICL}$.

## V. Conclusion

In this work, we proposed a fully unsupervised object re-ID method, which can be trained without using any labeled information. Based on the drawbacks of existing methods, we propose an instances correlation loss is proposed to enforce instance-to-instance learning with consistent features. Experimental results on three person re-ID datasets and one vehicle re-ID dataset demonstrate the effectiveness of the proposed method.

## Acknowledgment

This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003).

## References

[1] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015.

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD*, 1996.

[3] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 79–88, 2018.

[4] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *ECCV*, 2016.

[5] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," in *Advances in Neural Information Processing Systems*, 2020.

[6] H. Chen, B. Lagadec, and F. Bremond, "Ice: Inter-instance contrastive encoding for unsupervised person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 14 960–14 969.

[7] M. Wang, B. Lai, J. Huang, X. Gong, and X.-S. Hua, "Camera-aware proxies for unsupervised person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[8] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance-level discrimination," *CoRR*, vol. abs/1805.01978, 2018. [Online]. Available: http://arxiv.org/abs/1805.01978

[9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.

[10] S. Xuan and S. Zhang, "Intra-inter camera similarity for unsupervised person re-identification," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 921–11 930, 2021.

[11] J. Peng, Y. Wang, H. Wang, Z. Zhang, X. Fu, and M. Wang, "Unsupervised vehicle re-identification with progressive adaptation," *ArXiv*, vol. abs/2006.11486, 2020.

[12] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 654–13 662, 2020.

[13] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 2, 2019, pp. 1–8.

[14] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 978–10 987, 2020.

[15] Q. Tang and K.-H. Jo, "Unsupervised person re-identification via nearest neighbor collaborative training strategy," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 1139–1143.

[16] Q. Tang, G. Cao, and K.-H. Jo, "Fully unsupervised person re-identification via multiple pseudo labels joint training," *IEEE Access*, vol. 9, pp. 165 120–165 131, 2021.

[17] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *ArXiv*, vol. abs/1807.03748, 2018.

[18] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 654–13 662, 2020.

[19] F. Yang, Z. Zhong, Z. Luo, Y. Cai, Y. Lin, S. Li, and N. Sebe, "Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4853–4862, 2021.

[20] X. Han, X. Yu, G. Li, J. Zhao, G. Pan, Q. Ye, J. Jiao, and Z. Han, "Rethinking sampling strategies for unsupervised person re-identification," 2021.

[21] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *ArXiv*, vol. abs/1609.01775, 2016.