Exploring Sampler Strategies in Unsupervised Person Re-identification Training: Insights and Performance Analysis

Ge Cao
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea, Republic of
caoge@islab.ulsan.ac.kr

Qing Tang
Data Science Group
INTERX
Ulsan, Korea, Republic of
tangqing@interxlab.com

Tran Tien Dat
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea, Republic of
tdat@islab.ulsan.ac.kr

Ashraf Uddin Russo
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea, Republic of
russo@islab.ulsan.ac.kr

Adri Priadana
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
priadana3202@mail.ulsan.ac.kr

Kanghyun Jo*
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea, Republic of
acejo@ulsan.ac.kr

***Abstract*—Person re-identification holds significant research value within supervised systems characterized by non-overlapping multiple cameras. In recent years, unsupervised learning has made notable strides and has gradually approached the training efficacy of supervised learning. This paper focuses on exploring the influence and analysis of various sampling strategies on overall unsupervised training. We initially delineate a proxy-level memory bank scheme based on camera labels and employ a hard sample mining strategy for selecting negative pairs in a contrastive learning loss. Various sampling strategies, Random sampling, triplet sampling with dissimilar labels, and group sampling yield markedly distinct outcomes across three large-scale datasets, i.e. Market-1501, DukeMTMC-reID, and MSMT17. Detailed analysis and discussion of these results are provided in this study.**

***Index Terms*—Person Re-identification, unsupervised learning, sampling strategy.**

I. Introduction

In recent years, numerous computer vision tasks based on deep learning have achieved remarkable results [1, 2]. This paper focuses on person re-identification (ReID) tasks with an unsupervised training scheme. With the widely applied convolutional neural network (CNN), unsupervised person ReID achieves increasingly better performance and is gradually approaching the performance of supervised training [3, 4]. Thanks to its characteristic of not requiring manual annotation, unsupervised training mode has gained extensive research and exploration [5, 6, 7, 8]. Among them, unsupervised domain adaptive methods [7, 8] trained the backbone network on the source dataset with the ground truth and then transferred it to the target unlabeled dataset. While fully unsupervised methods [5, 6] directly trained the model on the target unlabeled dataset without leveraging any annotations.

As researchers delve deeper into unsupervised person ReID, a pipeline based on clustering algorithm [9, 10] and contrastive learning [5, 6] has gradually formed and demonstrated excellent performance. Under that pipeline, a backbone network [11] is applied to extract the features. And using clustering algorithms to generate pseudo labels for each training sample. Then the contrastive pipeline constructs a memory bank [12] to store the representations of each cluster. It should be emphasized that this paper follows the camera-aware contrastive pipeline [6], so the clusters would be divided into multiple proxies with the camera ID. And the memory bank also stores the representation of every proxy instead of the cluster.

After obtaining the clustering results, the DBSCAN would divide the whole training set into inliers and outliers. A sampling strategy [6, 13, 14]is applied to choose real training samples for each mini-batch from the inliers. This paper considers four kinds of sampling strategies with the camera-aware contrastive pipeline, i.e. random sampling, triplet sampling with cluster label [5], triplet sampling with camera-ware proxy label [6], and group sampling [14].

In this paper, we analyze the advantages and disadvantages of multiple sampling strategies and show the extensive experiment results on three large-scale person reID datasets, i.e. Market-1501 [15], DukeMTMC-reID [16], and MSMT17 [17].

The remaining content is organized as follows. Section II introduces the related work for the unsupervised person reID methods and sampling strategy. The main methodology is

shown in section III. Extensive experiments are demonstrated in section IV, and the section V concludes this paper.

II. Related Works

A. Unsupervised Person Re-ID

Due to its significant advantage of not requiring manual annotation, unsupervised learning has attracted widespread attention and research. In recent years, the unsupervised person re-ID was divided into many subtasks due to the different training settings, e.g. the unsupervised domain adaptation (UDA) case would train the model on single or multiple source datasets and then transfer to target unlabeled dataset to get great performance [7, 8, 18, 19], the fully unsupervised person re-ID would only focus on the target dataset without leveraging any other data with annotations [5, 6].

The research training with the contrastive learning algorithm achieved great success and gradually became mainstream. In these works, although the sampling plays a crucial role, it has not received focused research attention.

B. Sampling strategy

Given the training set, sampling is a necessary operation to assign samples into successive mini-batch. A good sampling strategy would help reduce bias and accelerate convergence during the training process. Random sampling strategy was applied in many classic papers [11] but was not suitable to the contrastive learning pipeline for unsupervised person re-ID. The triplet sampling [20] is often applied in person re-ID task. In each mini-batch, a certain number of samples from the same clusters are randomly selected [7, 5]. For some works that utilized the camera ID, the triplet sampling would take a certain number of samples from a proxy instead of a cluster randomly [6]. In the paper on group sampling [14], they proposed a new group sampling strategy that takes all the training samples inside the training process and improves the performance of the contrastive pipeline.

III. Methodology

A. Camera-aware contrastive pipeline

Given the target unlabeled dataset $X = \{x_i\}_{i=1}^{N}$, a backbone network is utilized to extract feature embedding $F = \{f_i\}_{i=1}^{N}$, where N denotes the number of samples in the training set. To generate the pseudo labels $Y = \{y_i\}_{i=1}^{N}$ for the samples, a clustering algorithm DBSCAN [9] is utilized. Thanks to the inherent characteristics of DBSCAN, the training set is effectively partitioned into inliers and outliers, minimizing the impact of outliers on the overall data analysis. So the samples would be remarked as outliers when $y_i = -1$. Following [6], the clusters obtained by DBSCAN would be divided into multiple proxies based on the corresponding camera ID. Each inlier sample would obtain a proxy pseudo label $\tilde{y_i}$. The memory bank M is constructed by storing the average embedding of the samples belonging to the same proxy as the representation of each proxy, e.g. $M[\tilde{y_i}]$ denotes the representation of the proxy which contains sample x_i. Successively, the inlier samples would be sampled for each mini-batch and the loss function. Different from the unified contrastive loss function [7], this paper chooses the limited number of negative proxies instead of all the proxies which not belong to the corresponding cluster, which is computed as,

$$L = -\sum_{i=1}^{B}\left(\frac{1}{|P^+|}\sum_{u\in P^+} log\frac{S(u,x_i)}{\sum_{p\in P^+} S(q,x_i)+\sum_{q\in P^-} S(q,x_i)}\right), \tag{1}$$

where $S(u, x_i) = exp(M[u]^T f_i/\tau)$, P^+ and P^- denote the positive proxies set and negative proxies set, simultaneously. $|\cdot|$ denotes the cardinality operation. Note that the positive proxies set contain the proxies which belong to the same clusters. The negative proxies set chooses the K-nearest proxies that do not belong to the corresponding cluster, called hard negative mining [21]. Finally, the memory bank is updated by a moving average scheme [7].

B. Random Sampling

Given the inliers $X' = \{x'_i\}_{i=1}^{N'}$ generated by the clustering algorithm, the sampling method is applied for sampling a fixed number of samples into a mini-batch, where the N' denote the number of inliers.

Random sampling is the simplest sampling strategy but is widely applied in computer vision tasks. When obtaining the training dataset X', random sampling is just randomly shuffling all the samples' order and getting the sampling list. As shown in Fig. 1, the left part of the figure shows a part of the samples in X', and the example sampling results are shown in group (a).

C. Triplet Sampling

Based on its name, triplet sampling [22] samples a certain number K of samples from the same class to construct a complete mini-batch. If the number of samples in a class is more than K, then randomly selected only K samples from the class. Otherwise, some samples would be sampled repeatedly to reach the K samples from one class. Following [7], the papers in the unsupervised person re-ID task often adopt the $P \times K$ mode to construct a mini-batch, where K denotes the number of samples taken from one class as mentioned above, and P denotes the number of classes in a mini-batch, i.e. $B = P \times K$. Additionally, in unsupervised re-ID work, the training set is divided into clusters, so we sample K samples from the P clusters. In this paper, we called it **triplet sampling with cluster labels**.

For the paper which utilized camera ID [6] as mentioned in the Section III. A, **triplet sampling with proxy labels** would sample K samples from P proxies instead of a cluster.

D. Group Sampling

Different from triplet sampling, group sampling considers all the samples in X' into the sampling list. When a proxy contains relatively more samples, triplet sampling would discard the extra samples and only sample the K samples into a mini-batch. Obviously, we cannot make sure the K samples

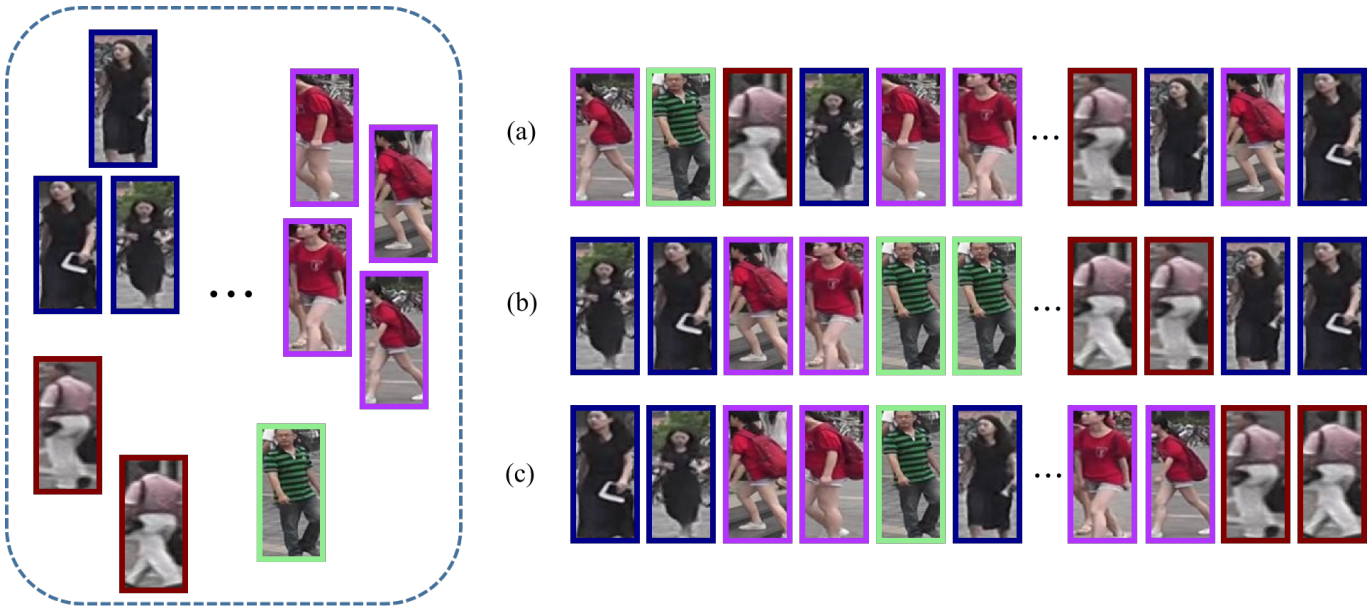

Fig. 1: Illustration of three kinds of sampling strategies, where group (a) denotes the random sampling, group (b) denotes the triplet sampling with $K = 2$ for example, group (c) describes the group sampling with $N = 2$ for instance.

could represent visual information from all the samples in that class and the discarded samples' feature would not be updated iteratively.

Different from the triplet sampling, group sampling would take all the inliers into the sample list. Group sampling will first group inliers together according to clustering labels, with every N inliers combined, where N is referred to as the group number. If the number of samples in a certain cluster cannot be divided evenly by N, the remaining samples will still be treated as one group, and then the order of all groups will be shuffled. Note that in this context, the group sampling used does not mix outliers into the sample list, which differs from [14].

IV. Experiments

A. Datasets and Evaluation Metrics

Dataset. Three large-scale person re-ID datasets are applied in this paper, i.e. Market-1501 [15], DukeMTMC-reID [16], and MSMT17 [17] datasets. Among them, the Market-1501 dataset captured pedestrian samples from 6 cameras and has a relatively smaller inter-camera domain gap. The DukeMTMC-reID dataset used 8 cameras to catch samples and contains relatively more occlusions. The samples in the MSMT17 dataset are derived from 15 cameras and possess a relatively bigger inter-camera domain gap.

Evaluation metrics. The mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) are adopted to evaluate the performance.

B. Implementation Details

Model hyper-parameters. The ResNet-50 [11] is utilized as the backbone network. The final fully connected layer is replaced by a global average pooling, a batch normalization, and an $L2$-normalization layer. The other structure of the backbone network is pre-trained on the ImageNet [24].

Other settings. The maximum distance between the closest samples is set to 0.5. The temperature factor is set to 0.09. The batch size is 32. The experiments for different numbers of K in triplet sampling with cluster labels and proxy labels are set to 1, 2, 4, 8, 12, 16, and 32. The experiments for different group sizes of N in group sampling are set to 1, 2, 4, 8, 12, 16, and 32 (same with triplet sampling to fairly compare). Other settings are following [6].

C. Comparisons with other researches

In this paper, the results of four different sampling strategies are shown in Table. I. The experiment results are compared with the state-of-the-art, i.e. MMCL [19], SpCL [7], GCL [23], and CAP [6]. The results for different sampling strategies are performed by different detailed settings and the best performances are shown in the Table. I. The comparisons of different detailed settings for the various sampling strategies except random sampling are shown in Table. II, Table. III, and Table. IV.

D. Discussions

This paper provides extensive experiments on the different key parameters in triplet sampling with cluster labels, triplet sampling with proxy labels, and group sampling.

In Table. II and Table. III, the experiments for triplet sampling with different labels are shown, where the key parameter is the different number of K. The larger value of K takes more samples from the same cluster/proxy, and vice versa. The range of K is taken from $\{1, 2, 4, 8, 12, 16, 32\}$. Although the

TABLE I: Comparison between the proposed method and state-of-the-art algorithms. The results on three target person Re-ID datasets, Market-1501 [15], DukeMTMC-Re-ID [16], and MSMT17 [17].

Methods	Market1501				DukeMTMC				MSMT17			
	mAP/%	R1/%	R5/%	R10/%	mAP/%	R1/%	R5/%	R10/%	mAP/%	R1/%	R5/%	R10/%
MMCL[19]	45.5	80.3	89.4	92.3	40.2	65.2	75.9	80.0	11.2	35.4	44.8	49.8
SpCL[7]	73.1	88.1	95.1	97.0	-	-	-	-	19.1	42.3	55.6	61.2
GCL[23]	66.8	87.3	93.5	95.5	62.8	82.9	87.1	88.5	21.3	45.7	58.6	64.5
CAP[6]	79.2	91.4	96.3	97.7	67.3	81.1	89.3	91.8	36.9	67.4	78.0	81.4
This paper-Random Sampling	11.5	24.8	33.4	37.6	0.8	2.5	4.0	5.1	9.9	24.3	32.1	36.3
This paper-Triplet Sampling[1]	82.7	93.3	97.1	98.3	69.4	83.4	90.6	92.8	38.6	70.6	80.5	83.9
This paper-Triplet Sampling[2]	83.7	93.3	97.7	98.6	66.2	80.7	88.9	91.2	37.1	68.6	78.8	82.2
This paper-Group Sampling	80.2	92.2	97.2	98.3	63.0	80.2	87.9	90.3	40.2	71.5	80.9	84.0

[1] Triplet sampling strategy with cluster labels.
[2] Triplet sampling strategy with camera-aware proxy labels.

TABLE II: The results obtained from training with triplet sampling using cluster labels of varying K on three person Re-identification (Re-ID) datasets—Market-1501 [15], DukeMTMC-Re-ID [16], and MSMT17 [17]—are presented.

K	Market1501				DukeMTMC				MSMT17			
	mAP/%	R1/%	R5/%	R10/%	mAP/%	R1/%	R5/%	R10/%	mAP/%	R1/%	R5/%	R10/%
1	36.2	61.9	70.5	73.5	23.8	41.8	50.4	53.8	12.0	29.4	40.0	44.7
2	48.8	72.6	79.6	82.2	43.8	63.8	71.9	74.8	22.0	48.3	59.3	64.3
4	78.8	91.2	96.2	97.6	64.5	79.5	86.8	89.2	35.2	64.9	75.4	79.1
8	**82.7**	**93.3**	**97.1**	**98.3**	**69.4**	**83.4**	90.6	92.8	**38.6**	**70.6**	**80.5**	**83.9**
12	80.5	92.3	97.0	98.0	68.1	82.8	**90.8**	**93.0**	32.5	66.5	76.9	80.6
16	45.2	73.8	87.4	91.4	41.9	65.5	78.1	82.3	17.6	47.1	59.1	64.1
32	0.8	2.0	4.5	7.3	0.4	1.0	2.4	4.0	0.1	0.4	1.3	2.1

TABLE III: The results obtained from training with triplet sampling using proxy labels of varying K on three person Re-identification (Re-ID) datasets—Market-1501 [15], DukeMTMC-Re-ID [16], and MSMT17 [17]—are presented.

K	Market1501				DukeMTMC				MSMT17			
	mAP/%	R1/%	R5/%	R10/%	mAP/%	R1/%	R5/%	R10/%	mAP/%	R1/%	R5/%	R10/%
1	65.1	84.2	91.4	93.7	50.8	70.9	79.3	82.8	20.4	46.1	58.8	64.4
2	78.0	89.5	95.3	96.7	58.8	76.0	84.0	85.7	30.7	59.9	71.5	75.6
4	81.6	92.0	96.7	97.7	64.5	79.5	86.8	89.2	**37.1**	**68.6**	**78.8**	**82.2**
8	**83.7**	**93.3**	**97.7**	**98.6**	66.2	80.7	88.9	91.2	36.6	67.9	77.9	81.5
12	80.4	91.5	96.7	97.9	**67.9**	**81.8**	**90.2**	**92.7**	31.9	63.7	74.8	78.8
16	56.3	80.6	90.5	93.7	48.1	68.9	80.7	84.6	2.7	7.7	13.4	16.9
32	0.6	1.2	3.7	5.8	0.3	0.8	2.3	3.5	0.1	0.1	0.6	1.0

TABLE IV: Comparison between the proposed method and state-of-the-art algorithms. The results on three target person Re-ID datasets, Market-1501 [15], DukeMTMC-Re-ID [16], and MSMT17 [17].

N	Market1501				DukeMTMC				MSMT17			
	mAP/%	R1/%	R5/%	R10/%	mAP/%	R1/%	R5/%	R10/%	mAP/%	R1/%	R5/%	R10/%
1	20.9	41.4	52.6	57.4	2.4	7.0	9.9	11.3	6.4	15.9	21.9	25.1
2	37.2	58.5	69.6	74.0	2.4	7.0	10.2	11.5	12.2	30.7	40.3	44.6
4	41.6	62.7	72.6	75.9	3.2	9.4	12.4	13.5	27.1	54.2	65.1	69.4
8	77.9	90.6	96.1	97.2	19.0	34.3	39.0	40.8	**40.2**	**71.5**	80.9	84.0
12	75.8	90.3	96.1	97.3	**63.8**	80.0	87.0	88.9	39.1	71.1	**81.2**	**84.1**
16	**80.2**	**92.2**	**97.2**	**98.3**	63.0	**80.2**	**87.9**	**90.3**	35.7	68.9	79.0	82.4
32	73.9	89.9	95.9	97.4	63.5	79.2	86.8	88.6	31.4	66.1	76.6	80.4

settings for achieving optimal results differ between the two experimental groups, the observed trends in the data variations are similar. For datasets Market1501 and DukeMTMC-reID, optimal outcomes are typically achieved at K=8 or 12, whereas for dataset MSMT17, optimal performance is observed at K=4 or 8. Although the optimal K values for maximizing peaks are not identical, they are remarkably close. This phenomenon arises from variations in the population size and the number of cameras utilized across different datasets. Database MSMT17 notably encompasses a larger population of training individuals compared to Market1501 and DukeMTMC-reID, along with a greater number of cameras, resulting in a lower average sample count per proxy.

Compared to group sampling, triplet sampling exhibits a notable drawback in its sensitivity to changes in the value of K. This is distinctly evident from Tables II and III, where experimental outcomes markedly deteriorate when K is small, and significantly worsen for larger K values (greater than 16), to the extent that the fitting can be deemed nonexistent. Regarding the substantial degradation in experimental performance when K is large, this stems from the excessive

size of the sample list, resulting in a scarcity of traversed clusters/proxies within a single epoch. For instance, when K=32, for MSMT17, the number of clusters traversed in one epoch significantly exceeds 1000 (even during the initial training stages). Consequently, the length of the sample list surpasses 32,000, while within one epoch, only $BatchSize \times N_{iteration}$, i.e., $32 \times 400 = 12800$ samples are processed, where $N_{iteration}$ denotes the number of the iterations in one epoch which is equal to 400 in the experiments. Thus, in such instances, training not only encounters issues with a plethora of duplicates among the K samples extracted from a single cluster or proxy but also faces the challenge of numerous clusters/proxies failing to compute loss, leading to pronounced training imbalances.

From Table IV, it can be observed that the deterioration in experimental performance is less pronounced when we alter the value of N, especially when N is large, in contrast to triplet sampling. From the experimental results, two issues can be discerned: 1) When N is small, the performance of group sampling relative to triplet sampling is comparatively inferior across all three datasets. This is primarily attributed to the scarcity of samples grouped together within a batch, rendering it challenging to provide efficient and accurate gradient descent directions. 2) Under consistent experimental conditions, the optimal performance of group sampling on the Market1501 and DukeMTMC-reID datasets is inferior to triplet sampling, with only MSMT17 exhibiting superior performance to triplet sampling.

V. Conclusions

The significance of sampling strategies in unsupervised person re-identification tasks is elucidated in this paper. Building upon a baseline contrastive learning framework involving camera-aware memory bank construction and hard sample mining, extensive experimentation is conducted on four sampling strategies. Furthermore, a comprehensive analysis is undertaken to delve into the experimental outcome disparities arising from each sampling strategy.

Acknowledgment

This result 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] Qing Tang and Hail Jung. "Reliable Anomaly Detection and Localization System: Implications on Manufacturing Industry". In: *IEEE Access* (2023).

[2] Qing Tang, YoungSeok Lee, and Hail Jung. "The Industrial Application of Artificial Intelligence-Based Optical Character Recognition in Modern Manufacturing Innovations". In: *Sustainability* 16.5 (2024), p. 2161.

[3] Kaiyang Zhou et al. "Omni-scale feature learning for person re-identification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3702–3712.

[4] Shuting He et al. "Transreid: Transformer-based object re-identification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 15013–15022.

[5] Zuozhuo Dai et al. "Cluster contrast for unsupervised person re-identification". In: *Proceedings of the Asian Conference on Computer Vision*. 2022, pp. 1142–1160.

[6] Menglin Wang et al. "Camera-aware proxies for unsupervised person re-identification". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 4. 2021, pp. 2764–2772.

[7] Yixiao Ge et al. "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 11309–11321.

[8] Zhun Zhong et al. "Invariance matters: Exemplar memory for domain adaptive person re-identification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 598–607.

[9] Anant Ram et al. "A density based algorithm for discovering density varied clusters in large spatial databases". In: *International Journal of Computer Applications* 3.6 (2010), pp. 1–4.

[10] Huapeng Cai et al. "Research on unsupervised people re-identification based on k-means clustering". In: *2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT)*. 2021, pp. 405–408. DOI: 10.1109/ISCIPT53667.2021.00087.

[11] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[12] Zhirong Wu et al. "Unsupervised feature learning via non-parametric instance discrimination". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3733–3742.

[13] Qing Tang, Ge Cao, and Kang-Hyun Jo. "Unsupervised Object Re-identification via Irregular Sampling". In: *2022 International Workshop on Intelligent Systems (IWIS)*. 2022, pp. 1–6. DOI: 10.1109/IWIS56333.2022.9920902.

[14] Xumeng Han et al. "Rethinking sampling strategies for unsupervised person re-identification". In: *IEEE Transactions on Image Processing* 32 (2022), pp. 29–42.

[15] Liang Zheng et al. "Scalable person re-identification: A benchmark". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1116–1124.

[16] Zhedong Zheng, Liang Zheng, and Yi Yang. "Unlabeled samples generated by gan improve the person re-identification baseline in vitro". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3754–3762.

[17] Longhui Wei et al. “Person transfer gan to bridge domain gap for person re-identification”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018, pp. 79–88.

[18] Zhun Zhong et al. “Learning to adapt invariance in memory for person re-identification”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.8 (2020), pp. 2723–2738.

[19] Dongkai Wang and Shiliang Zhang. “Unsupervised person re-identification via multi-label classification”. In: *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 2020, pp. 10981–10990.

[20] Hao Luo et al. “A strong baseline and batch normalization neck for deep person re-identification”. In: *IEEE Transactions on Multimedia* 22.10 (2019), pp. 2597–2609.

[21] Joshua Robinson et al. “Contrastive learning with hard negative samples”. In: *arXiv preprint arXiv:2010.04592* (2020).

[22] Weihua Chen et al. “Beyond triplet loss: a deep quadruplet network for person re-identification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 403–412.

[23] Hao Chen et al. “Joint generative and contrastive learning for unsupervised person re-identification”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2004–2013.

[24] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.