



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

**Doctor of Philosophy**

**Novel and Practical Object Re-identification and  
Search in Intelligent Surveillance Systems**

**The Graduate School**

**of the University of Ulsan**

**Department of Electrical, Electronic and Computer Engineering**

**Qing Tang**

Novel and Practical Object Re-identification and Search in Intelligent  
Surveillance Systems

Supervisor: Kang-Hyun Jo

A Dissertation

Submitted to  
the Graduate School of the University of Ulsan  
in Partial Fulfillment of the Requirements  
for the Degree of

Doctor of Philosophy

by

Qing Tang

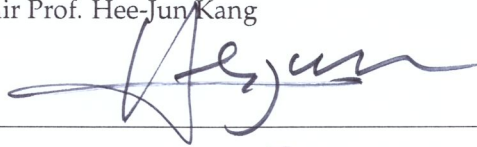
Department of Electrical, Electronic and Computer Engineering  
University of Ulsan

May 2022

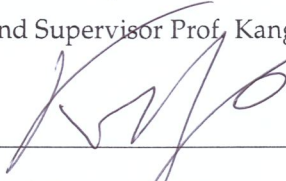
**Novel and Practical Object Re-identification and Search in Intelligent  
Surveillance Systems**

This certifies that the dissertation of Qing Tang is approved:

Committee Chair Prof. Hee-Jun Kang



Committee Member and Supervisor Prof. Kang-Hyun Jo



Committee Member Prof. Young-Soo Suh



Committee Member Prof. Hyun-Deok Kang



Committee Member Prof. Jang-Sik Park



Department of Electrical/Electronic and Computer Engineering

University of Ulsan, South Korea

May, 2022



*"To have life henceforth, the poem of new joys."*

*- Dead Poets Society*

UNIVERSITY OF ULSAN

# ABSTRACT

Graduate School of Electrical Engineering

Department of Electrical, Electronic and Computer Engineering

Doctor of Philosophy

## **Novel and Practical Object Re-identification and Search in Intelligent Surveillance Systems**

by Qing Tang

Some of the most populated cities in all over the world are under a heavy amount of public surveillance systems for monitoring the population surrounding. In order to save labor costs, intelligent surveillance systems have rapidly developed in recent years by supplying and assisting security workers in detecting, analyzing, and predicting undesirable incidents.

Object re-identification (re-ID) and search are the foundation of a wide range of applications in intelligent surveillance systems. The targets of Object re-ID and search systems indicate person and vehicle. It can be used for a cross-camera person or vehicle tracking and search. The work on this manuscript focus on the camera-based object re-ID and search system in public datasets and real-world scenario.

To design a more realistic and practical object re-ID and search system for intelligent surveillance systems, we focus on three aspects. Firstly, this manuscript focus on investigating the unsupervised object re-ID. Secondly, we argue that training a system which able to identify a specific object from full scene images is closer to the real-world applications, therefore we investigate object search systems. Third, this manuscript focus on combining the supervised detection methods and unsupervised object re-ID methods.

We improve the performance of the re-ID systems by designing a more robust sampling strategy, refining pseudo labels, and designing loss functions. Moreover, we improve the performance of the search by designing learning strategies for unlabeled data and designing loss function. Extensive experimental results demonstrate

the effectiveness of the proposed methods and their practicality in real-world unsupervised person re-ID applications. The experimental results of object re-ID have been evaluated on three public person re-ID datasets and one public vehicle re-ID dataset. The experiments are performed in two public person search datasets. Moreover, several outdoor real-world videos are used to validate the performance of the proposed methods in real-world applications.

## *Acknowledgements*

I would like to express my gratitude for my supervisor, Professor Kang-Hyun Jo, who gave me an opportunity to study under his guidance, nurture, encourage, and give abundant support during my life as a graduate student at the University of Ulsan. I would also thank the members of the thesis committee Prof. Hee Jun Kang, Prof. Young Soo Suh, Prof. Hyun-Deok Kang, and Prof. Jangsik Park for their invaluable comments and suggestion to improve the quality of this thesis.

I am grateful to the member of the Intelligent Systems Laboratory for the meaningful discussion and lesson that enhance my knowledge and broaden my understanding in this research field, computer vision and machine learning. Many thanks for the Brain Korea Scholarship program that allow me to attend various conferences around the world and expand my horizon as a student and researcher. Special thanks to my parent who always stays beside me during my study in Korea.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation and Background . . . . .	1
1.2 Disposition . . . . .	4
<b>2 Literature Review</b>	<b>6</b>
2.1 Supervised and Unsupervised Object Re-identification System . . . . .	6
2.2 Unsupervised Object Re-identification System . . . . .	8
2.2.1 Hand-Crafted Feature-Based Methods . . . . .	8
2.2.2 UDA-Based Methods . . . . .	8
2.2.3 Fully Unsupervised Learning Methods . . . . .	9
2.2.4 Pseudo Label Prediction . . . . .	10
2.2.5 Similarity Exploration . . . . .	11
<b>3 Proposed Fully Unsupervised Object Re-identification</b>	<b>13</b>
3.1 Experimental Datasets and Evaluation Metrics . . . . .	13
3.1.1 Person re-ID datasets . . . . .	14
3.1.2 Vehicle re-ID datasets . . . . .	14
3.1.3 Evaluation Metrics . . . . .	14

3.2	Improved Unsupervised Object Re-identification via Irregular Sampling	15
3.2.1	Related works of Sampling Strategy . . . . .	15
3.2.2	Proposed Method . . . . .	17
	Centroids-towards learning . . . . .	17
	Proposed Irregular Sampling Strategy . . . . .	18
3.2.3	Experiments . . . . .	20
	Datasets and Evaluation Metrics . . . . .	20
	Implementation Details . . . . .	21
	Comparisons with The State-of-the-Arts . . . . .	22
3.2.4	Ablation Studies . . . . .	22
	Effectiveness of Irregular Sampling . . . . .	22
3.2.5	CONCLUSIONS . . . . .	23
3.3	Unsupervised Person Re-identification via Mining Label Homogeneity	24
3.3.1	Introduction . . . . .	24
3.3.2	Proposed Method . . . . .	25
	The Baseline Network . . . . .	26
	Auxiliary Module . . . . .	27
	Overall Loss . . . . .	29
3.3.3	Experiments Setting . . . . .	30
	Datasets and Evaluation Metrics . . . . .	30
	Implementation Details . . . . .	30
3.3.4	Experiments Results . . . . .	31
	Comparison with Other Methods . . . . .	31
	Ablation Studies . . . . .	31
3.3.5	Conclusion . . . . .	33

3.4 Fully Unsupervised Person Re-Identification via Centroids and Neighborhoods Joint Learning . . . . .	33
3.4.1 Introduction . . . . .	33
3.4.2 Proposed Method . . . . .	36
Framework Overview . . . . .	36
Joint Label Prediction (Joint-LP) . . . . .	36
Rectified Binary Cross Entropy (ReBCE) Loss . . . . .	38
Overall Loss . . . . .	39
3.4.3 Experiment Setting . . . . .	40
Ablation Study . . . . .	40
Comparision with Other FUL Methods . . . . .	43
3.4.4 Conclusion . . . . .	43
3.5 Unsupervised Person Re-Identification via Multiple Pseudo Labels Joint Training . . . . .	44
3.5.1 Proposed Method . . . . .	46
Framework Overview . . . . .	46
Multiple Pseudo Labels Prediction . . . . .	47
3.5.2 The joint Loss function . . . . .	51
3.5.3 Experiment Settings . . . . .	52
Datasets and Evaluation Metrics . . . . .	52
Implementation Details . . . . .	52
3.5.4 Ablation study of the proposed components . . . . .	52
Performance of the baseline method . . . . .	54
The effectiveness of the clustering-based pseudo label . . . . .	54
The analysis of Adaptive Similarity Measurement-based pseudo label Prediction (ASMP) . . . . .	55

	The analysis of channel-based self-similarity (CSS)	56
	The analysis of joint training strategy	59
3.5.5	Performance	60
	Performance comparison in public datasets	60
	Performance in real-world application	62
3.5.6	Conclusion	62
3.6	Unsupervised Object Re-identification via Instances Correlation Loss	63
3.6.1	Proposed Method	64
	Momentum Contrast Learning	64
	The Proposed Instances Correlation Loss	65
3.6.2	Experiments	68
	Datasets and Evaluation Metrics	68
	Implementation Details	68
	Comparisons with the State-of-the-Arts	68
	Effectiveness of the Instances Correlation Loss	69
3.6.3	Conclusion	69
3.7	Unsupervised Object Re-identification via Relative Hard Samples Learning	70
3.7.1	Related Work in Hard Sample Mining Strategy	70
3.7.2	Proposed Method	71
	Centroids-towards learning	71
	Relative Hard Samples (RHS) Learning	72
3.7.3	Experiments	73
	Datasets and Evaluation Metrics	73
	Implementation Details	75
	Comparisons with The State-of-the-Arts in Three Datasets	75



3.7.4	Ablation Studies . . . . .	76
	Comparison with different $\tau_h$ . . . . .	76
	Comparison with other hard sample learning methods . . . . .	76
3.7.5	Conclusions . . . . .	77
<b>4</b>	<b>Supervised Object Search System</b>	<b>78</b>
4.1	Literature Review . . . . .	80
4.1.1	The Online Instance Matching (OIM) . . . . .	80
4.1.2	Sequential End-to-end Network (SeqNet) . . . . .	81
4.2	Proposed Person Search via Background and Foreground Contrastive Learning . . . . .	82
4.2.1	The Proposed BFCL loss . . . . .	82
	Architecture Overview . . . . .	82
4.2.2	Experiments . . . . .	84
	Datasets . . . . .	84
	Ablation Study . . . . .	85
	Comparison with the state-of-the-art Methods . . . . .	87
4.2.3	Conclusion . . . . .	87
<b>5</b>	<b>Weakly Supervised Object Search System</b>	<b>89</b>
5.1	Weakly Supervised Object Search with Region Siamese Networks . . . . .	90
5.1.1	Instance-Level Consistency Learning . . . . .	91
5.1.2	Cluster-Level Contrastive Learning . . . . .	92
<b>6</b>	<b>Conclusion</b>	<b>93</b>
6.1	Future Works . . . . .	94
<b>A</b>	<b>Publications</b>	<b>95</b>

A.1 Journal . . . . .	95
A.2 Conference . . . . .	95
<b>Bibliography</b>	<b>98</b>

# List of Figures

1.1	The examples of object re-ID images. Green boxes denote the matching identity between query and gallery. . . . .	2
1.2	The examples of person search systems on outdoor real-world videos. The query image (re-ID target) is shown as a sub-figure at the bottom of each frame. The green bounding boxes are the search results, and the classification scores are written on the boxes. . . . .	3
1.3	Labeled information of supervised, unsupervised setting of object re-ID and search tasks. . . . .	3
2.1	Illustration of (a) supervised object re-ID, (b) UDA-based object re-ID, and (c) FUL-based object re-ID. <b>ID:</b> identity. . . . .	7
3.1	The illustration of the fully unsupervised object re-ID framework with sampling. Before every training epoch, cluster algorithm DBSCAN (Ester et al., 1996) is used to roughly cluster every sample in the whole dataset into $N_c$ classes. $C = \{c_1, \dots, c_{N_c}\}$ represents centroids of $N_c$ classes. In every training iteration, mini-batches are generated from a dataset using sampling strategy based on clustering results. The encoder is fine-tuned according to $C$ via centroids-towards learning, and the momentum encoder is updated by the encoder by momentum update in (He et al., 2019). . . . .	16
3.2	The illustrations of sampling strategies. (a) Random Sampling (b) Triplet Sampling (c) Our proposed Irregular Sampling. The same color represents the samples sharing the same identity class. . . . .	16

3.3	The model performance of different sampling strategies in different $P$ . The X-axis represents different numbers of positive samples $P$ , and Y-axis represents the model performance. Graphs in the first-row report performance in mAP (%), and the second-row report performance in Rank-1 (%). Graphs in different columns report performance on different datasets. . . . .	20
3.4	The framework of our proposed SNNet. The black line indicates the baseline network. Apart from the predicted main pseudo label $y_i$ , we proposed an auxiliary module to seek auxiliary labels $y_i^{sym}$ and $y_i^{nh}$ as additional supervisions to optimize the network. . . . .	25
3.5	The illustration of pseudo label memory (PLM). $n = 5$ is assumed in this figure. Ideally, $y_i = y_i^{sym}$ because of symmetric homogeneity constrain, and $y_i = y_i^{nh}$ because of neighbor homogeneity constrain. . .	26
3.6	The illustrations of (a) C-LP (class centroids-towards learning) (b) SM-LP (neighborhoods-towards learning) (b) proposed Joint-LP (both). . .	35
3.7	General framework for FUL person re-ID methods. . . . .	35
3.8	The illustration of our proposed Jointly Label Prediction Module (Joint-LP). . . . .	36
3.9	The t-SNE (Maaten and Hinton, 2008) visualization on features representation of 10 identities. The different color points are denoted identities. . . . .	41
3.10	The framework of FUL person re-ID method. (a) Baseline method, MLReID (Wang and Zhang, 2020). (b) The proposed MLJT. . . . .	45
3.11	The illustration of different pseudo label prediction methods. . . . .	46
3.12	The illustrations of self-similarities exploration strategies in unsupervised re-ID. (a) Part-based self-similarities used in SSG (Fu et al., 2019b) and SSL (Lin et al., 2020). (b) Our proposed Channel-based self-similarities. (c) The examples of human location variance in Market-1501. <b>GAP:</b> Global Average Pooling. $P = 2$ and $G = 2$ are assumed in this figure. . .	47

3.13	The histogram of similarity distribution. The horizontal axis is similarity score. The right graph is zoomed in from the blue rectangular area of the left graph. . . . .	49
3.14	The t-SNE (Maaten and Hinton, 2008) plot of 10 identities. Different colors denotes different identities. . . . .	53
3.15	The visualization results of feature maps with different similarity exploration methods. . . . .	58
3.16	Examples of the top-10 person re-ID results. The first and second rows are generated by the baseline method (Wang and Zhang, 2020) and the proposed MLJT, respectively. The green boxes denote the true positive results, and the red boxes denote the false positive results. . . . .	60
3.17	The comparison re-ID results of the baseline method (Wang and Zhang, 2020) and our proposed method MLJT on two outdoor real-world videos. Each column refers to the different frames of videos. The query image (re-ID target) is shown as a sub-figure at the bottom of each frame. The human regions are detected by YOLOv5 (al, Apr. 2021). The green bounding boxes are the re-ID results, and the classification scores are written on the boxes. Frames with a red border refer to false re-ID results. . . . .	61
3.18	The illustration of the proposed fully unsupervised object re-ID framework. . . . .	63
3.19	The example of $M$ : instances correlation matrix, and $T$ : target matrix. . . . .	64
3.20	The illustration of the proposed fully unsupervised object re-ID framework. Before every training epoch, cluster algorithm DBSCAN (Ester et al., 1996) is used to roughly cluster every sample in the whole dataset into $N_c$ classes. $C = \{c_1, \dots, c_{N_c}\}$ represents centroids of $N_c$ classes. In every training iteration, the encoder is fine-tuned according to $C$ via centroids-towards learning and RHS learning, and the momentum encoder is updated by the encoder as Equation 3.33 by momentum update (He et al., 2019). . . . .	71

3.21	The illustration of the proposed RHS selection. The same color represents the samples containing the same identity. . . . .	72
4.1	Illustration of the object re-ID and search system. . . . .	79
4.2	Examples of backgrounds (red boxes) and foregrounds (green boxes) RoIs in (a)-(c) three different input images. The number at the top left corner represents the identity $i$ . $i = 0$ indicates background, and $i > 0$ indicates foreground. Different $i$ means different identity. . . . .	82
4.3	The architecture of the proposed person searching framework. The component in yellow is newly proposed by us. Our proposed Backgrounds and Foregrounds Contrasting Loss $L_{bfc}^3$ aims to push an RoI far away from other RoIs with different $i$ and backgrounds. . . . .	83
5.1	Comparisons between two person search settings. (a) Supervised setting. The images are annotated with both bounding boxes and person identities. Note that some identity annotations have lacked in original person search datasets. (b) The proposed weakly supervised setting. The images only have bounding box annotations. . . . .	90
5.2	Illustration of our R-SiamNet (Han et al., 2021a) . . . . .	91

# List of Tables

3.1	DATASET STATISTICS. <b>IDs</b> : IDENTITIES. <b>CAMS</b> : THE NUMBERS OF CAMERAS. . . . .	13
3.2	Experimental results of our proposed method and other fully unsupervised re-ID methods on two person re-ID datasets. The top result is highlighted in bold. . . . .	21
3.3	Experimental results of our proposed method and other fully unsupervised re-ID methods on vehicle re-ID datasets VeRi-776. . . . .	22
3.4	Ablation study on different sampling methods. † denotes results are obtained by our experiments using publicly available source code. . . . .	23
3.5	Comparison with other fully unsupervised person re-ID methods on Market-1501 and DukeMTMC-ReID Dataset. “*”: Baseline method, reproduced by us based on the authors’ code. “↑”: Results that outperforms baseline. Results that surpass all methods are <b>bold</b> . . . . .	29
3.6	Methods comparison when tested on Market-1501 and DukeMTMC-reID. <b>Baseline</b> : Baseline model trained with main pseudo labels. <b>S</b> : Auxiliary symmetric labels $y_i^{sym}$ . <b>N</b> : Auxiliary neighbor label $y_i^{nh}$ . . . . .	30
3.7	Evaluation with different values of $\lambda^{sym}$ . . . . .	31
3.8	Evaluation with different values of $\lambda^{nh}$ . . . . .	32
3.9	Ablation study on outliers. “X”: Training without outliers. “✓”: Training each outlier as an individual class as Eq(2). . . . .	40
3.10	Comparison with different label prediction methods . . . . .	40
3.11	Comparison with different loss function . . . . .	41

3.12 Performance comparison with other FUL person re-ID methods on Market-1501 and DukeMTMC-ReID. “LPM”:Label prediction methods. “*”: The MetaCam algorithm in DSCE(Yang et al., 2021) is not considered in this table, because MetaCam requires camera IDs. This paper performs comparison experiments in unknown camera IDs environment. . . . .	42
3.13 Ablation study on individual proposed modules in the proposed MLJT.	54
3.14 Analysis of hyper-parameters $k$ on Market-1501. . . . .	55
3.15 Comparison with different pseudo label prediction methods. . . . .	56
3.16 Comparison of different splitting methods and number of splitting groups in CSS on Market-1501. . . . .	57
3.17 Comparison with different similarity exploration methods. . . . .	57
3.18 Performance comparison with state-of-the-art FUL-based methods. The first and second best results are marked in <b>bold</b> and <b>blue</b> , respectively. . . . .	59
3.19 Experimental results of our proposed method and state-of-the-art fully unsupervised re-ID methods on Market-1501 and DukeMTMC-reID. ICL: Instances Correlation Loss. The top result is highlighted in bold and the second best result is shown in blue. . . . .	66
3.20 Experimental results of our proposed method and state-of-the-art fully unsupervised re-ID methods on MSMT17. . . . .	67
3.21 Experimental results of our proposed method on vehicle re-ID datasets VeRi-776. . . . .	67
3.22 Ablation study on using different loss functions for instance-to-instance learning . . . . .	67
3.23 Experimental results of our proposed method and other fully unsupervised re-ID methods on vehicle re-ID datasets VeRi-776. . . . .	74
3.24 Experimental results of our proposed method and other fully unsupervised re-ID methods on two person re-ID datasets. The top result is highlighted in bold. . . . .	74



3.25	Experimental results of our proposed method with different values of $\tau_h$ .	76
3.26	Ablation study on using different loss functions for hard sample learning on Market-1501.	76
4.1	Ablation experiments on <b>CQ</b> : Circular Queue (Xiao et al., 2017) and <b>BFCL</b> : our proposed Background and Foreground Contrastive Loss.	86
4.2	Performance of our framework with different values of $\tau_c$ in Equation 4.5.	86
4.3	Comparison with state-of-the-art methods on two person searching datasets. The top result is highlighted in bold.	87

## Chapter 1

# Introduction

### 1.1 Motivation and Background

Surveillance systems are used to monitor specific areas, such as homes, buildings, and borders. A standard surveillance system consists of a large number of cameras and security workers must continuously watch the real-time videos to check if there are undesirable incidents. In order to reduce labor costs, intelligent surveillance systems have rapidly developed in recent years to supply and assist security workers in detecting, analyzing, and predicting undesirable incidents.

Intelligent security and surveillance systems are useful techniques that are utilized in various sectors in public and privacy places, such as specific person and vehicle searching, a specific person and vehicle searching tracking, detecting and identifying abnormal actions or situations, criminal identification, advertising, and many more (Shahbaz and Jo, 2021; Yang et al., 2017; Feizi, 2017; Li et al., 2019a).

The increasing demand and excellent performance of intelligent surveillance systems make it become an active and fast-growing research area in recent years. The main reasons are (1) the widespread camera network in public places, (2) the increasing demand for public safety, (3) the rapid development of deep learning, and (4) the expensive human labor.

Some of the most populated city in all over the world are under a heavy amount of public surveillance systems (Comparitech, 2021). Moreover, according to a report (Liza Lin, 2019) in the end of 2019, hundreds of millions more surveillance cameras will be set. The surveillance systems serve as a primary tool to monitor the population surrounding and to fight crime and terrorism.

The rapid development of deep learning methods also conducted a new wave into intelligent surveillance systems. Deep learning methods achieve breakthrough performance in gathering data and making predictions by utilizing computer vision, pattern recognition, and artificial intelligence technologies.

Object re-identification (re-ID), including person re-ID and vehicle re-ID, can be used for a cross-camera person or vehicle tracking and search. As an important application in intelligent security and surveillance systems, object re-ID aims to re-identify an object across multiple non-overlapping cameras. In other words, the re-ID system aims to retrieve images containing the same identity. The illustration of person and vehicle re-ID tasks is shown in Figure 1.1.



FIGURE 1.1: The examples of object re-ID images. Green boxes denote the matching identity between query and gallery.

As mentioned above, the object re-id system is trained to match an investigated object (query) with the gallery of well-cropped images which is far from real-world applications. Because object detectors in intelligent systems might produce wrong-cropped images in practical applications, which leads to a bad re-id performance.

Identifying a specific object from a whole scene of images is closer to the real-world applications, therefore recent researches (Xu et al., 2014; Xiao et al., 2017;

Zhang, Li, and Zhang, 2021; Yan et al., 2021; Li and Miao, 2021; Chen et al., 2020) tend to solve person detection and re-id jointly, namely person search. The simplified view of person search is illustrated in Figure 1.2.

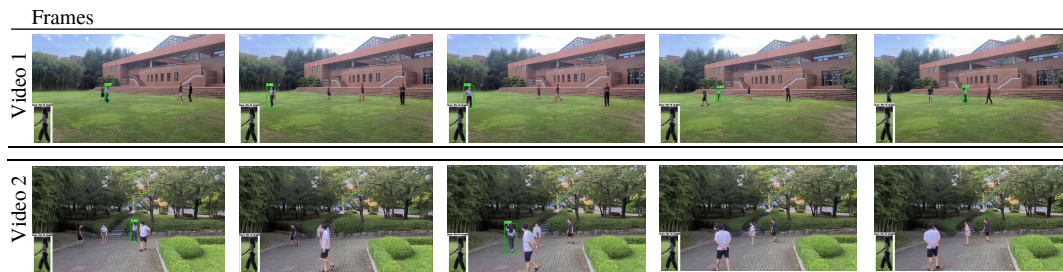


FIGURE 1.2: The examples of person search systems on outdoor real-world videos. The query image (re-ID target) is shown as a sub-figure at the bottom of each frame. The green bounding boxes are the search results, and the classification scores are written on the boxes.

The object search system aims to detect the specific object regions from realistic and uncropped images. Then, based on detected object regions, the system retrieves the specific object regions that contained the same identity as a query image by matching detected regions with query images. The object search can be considered as an integrated task of object detection and object re-identification.



FIGURE 1.3: Labeled information of supervised, unsupervised setting of object re-ID and search tasks.

In the past decade, object search works mostly focused on supervised learning

and person-based search, which achieved significant progress (Li and Miao, 2021; Chen et al., 2020; Xiao et al., 2017). The supervised object search requires substantial labeled bounding boxes and identities for achieving satisfying performance. The labeled information of supervised and unsupervised setting systems is shown in Figure 1.3. In the supervised setting manner, both coordination of bounding boxes and identities are provided to train the system.

However, it is time-consuming and difficult to annotate every object and its identity across multiple cameras, especially for identities. Therefore, some recent works (**weakly**; Han, Ko, and Sim, 2021a) focus on combining the supervised detection methods and unsupervised object re-ID methods. As shown in Figure 1.3, only positions of bounding boxes are required in the unsupervised setting manner. In other words, the identity information is unknowable in an unsupervised setting manner. Therefore, the unsupervised setting does not require annotating the identity for each image. It is relatively easier to acquire a large amount of unlabeled data by public surveillance systems in the real world.

The following section of this manuscript focus on discussing various methods related to the unsupervised object re-ID, supervised object search, and self-supervised object search.

## 1.2 Disposition

This part explains the organization of this manuscript. The following section is discussing various methods related to the utilization of deep learning models in unsupervised object re-id, supervised object search, and weakly supervised object search.

Section 2 discusses previous publications in object re-ID that influence the recent and our proposed research works.

Section 3 explains the proposed approaches in developing more robust unsupervised object re-ID models. Our proposed approaches focus on the sampling strategy design, generated pseudo labels refinement, learning strategy optimization, loss function design.

Section 4 discusses the existing supervised object search works, which integrate the object detection and re-ID as one complete pipeline. Moreover, our proposed

object search methods are introduced.

Section 5 introduces the weakly supervised person search works, which only require the coordination of bounding boxes. In other words, only positions of bounding boxes are provided and identities information is unknowable during the whole training process.

Section 6 concludes the manuscript. The discussion and the directions of our future works of object re-id and search systems in intelligent security and surveillance systems are further presented.

## Chapter 2

# Literature Review

### 2.1 Supervised and Unsupervised Object Re-identification System

The object re-ID system aims to retrieve matched people from different cameras or different occasions. The idea of object re-ID is to matching features of images, therefore extracting discriminative features is critical and challenging for the object re-ID.

Based on training strategy, current object re-ID methods can be summarized in three categories: supervised object re-ID (Huang et al., 2020; Zhou et al., 2020; Fu et al., 2019a), Unsupervised Domain Adaptive (UDA) based object re-ID (Tahir, 2019; Zhong et al., 2019; Zhong et al., 2018; Fan, Zheng, and Yang, 2017; Ge, Chen, and Li, 2020), and the Fully Unsupervised Learning (FUL) based object re-ID (Yu, Wu, and Zheng, 2017; Yu, Wu, and Zheng, 2020; Lin et al., 2019; Wang and Zhang, 2020; Tang and Jo, 2021; Yang et al., 2021; Ji et al., 2020; Fu et al., 2019b; Lin et al., 2020; Ding, Khan, and Tang, 2019). The illustrations of these methods are shown in Figure 2.1.

In past decades, object re-ID works mostly focused on supervised object re-ID. The supervised object re-ID methods (Huang et al., 2020; Zhou et al., 2020; Fu et al., 2019a) train the network on a labeled dataset. Annotating object identities for every image is required in supervised methods, as shown in Figure 2.1(a).

Some recent works focus on unsupervised object re-ID, which do not require labeled information in a target dataset. It is expensive to manually annotate identity across multiple cameras. Because of the lack of labeled information, the unsupervised methods can not achieve satisfying performance as supervised methods. Several unsupervised methods (Tahir, 2019; Zhong et al., 2019; Zhong et al., 2018; Fan,

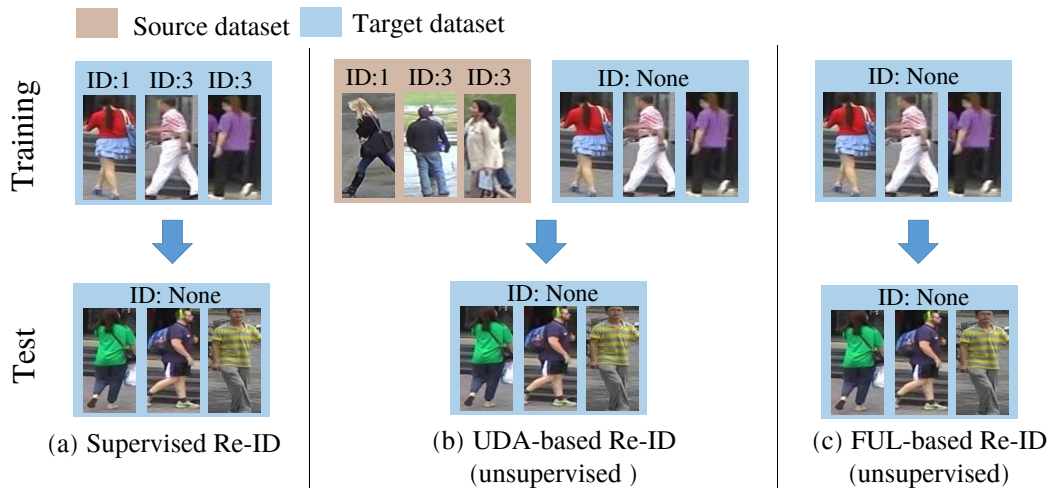


FIGURE 2.1: Illustration of (a) supervised object re-ID, (b) UDA-based object re-ID, and (c) FUL-based object re-ID. **ID:** identity.

Zheng, and Yang, 2017; Ge, Chen, and Li, 2020) utilized the Unsupervised Domain Adaption (UDA) to improve the model performance. The idea of UDA-based methods is to transfer the knowledge from a labeled source dataset to the unlabeled target dataset to improve the model performance on the target dataset. A common operation is to train the network on the labeled source and the unlabeled target dataset simultaneously (Zhong et al., 2018; Ge, Chen, and Li, 2020), as shown in Figure 2.1(b). However, the model performance will significantly decline if the domain gap between the source and target datasets is large.

The above two factors make supervised object re-ID and UDA-based unsupervised object re-ID are difficult to meet the requirement of practical industry application. Conversely, the FUL-based object re-ID methods (Yu, Wu, and Zheng, 2017; Yu, Wu, and Zheng, 2020; Lin et al., 2019; Wang and Zhang, 2020; Tang and Jo, 2021; Yang et al., 2021; Ji et al., 2020; Fu et al., 2019b; Lin et al., 2020; Ding, Khan, and Tang, 2019) enjoy two merits, which make them more suitable for real-world application. (1) Training a FUL-based re-ID system does not require any labeled information, as shown in Figure 2.1(c). It is relatively easier to acquire a large number of unlabeled images and videos by public surveillance systems in the real world. (2) FUL-based re-ID methods do not need to consider the domain gap between the source and target datasets, because they can be trained directly in any unlabeled target dataset. Because of the above two merits, people tend to pay more attention to FUL-based object re-ID.



## 2.2 Unsupervised Object Re-identification System

### 2.2.1 Hand-Crafted Feature-Based Methods

In the past, researchers used traditional manual features (Liao et al., 2015; Zheng et al., 2015a; Peng et al., 2016) to conduct feature extraction. The hand-crafted feature-based methods have demonstrated the effectiveness on small datasets, but can not achieve satisfying performance on large datasets. Zheng et al. (Zheng et al., 2015a) proposed an unsupervised Bag-of-Words (BoW) descriptor which extracted image features using the Color Names (CN) descriptor. Liao et al. (Liao et al., 2015) proposed an effective feature representation called Local Maximal Occurrence (LOMO) which is robust to illumination and viewpoint changes. Peng et al. (Peng et al., 2016) considered training the model only with labeled image pairs limits model's scalability in real-world applications and therefore developed a cross-dataset transfer learning re-ID approach which able to learn a dataset-shared but target-data-biased feature representation. The performance of the above hand-crafted feature-based methods are poor, because these hand-craft features are not robust enough for learning discriminative features, especially without human-annotated labels. Feature extraction is critical and challenging for unsupervised person re-ID.

### 2.2.2 UDA-Based Methods

To learn to extract discriminative features, one of the most popular solutions is unsupervised domain adaptation (UDA). Recent researches (Wei et al., 2018a; Zhong et al., 2018; Zhang et al., 2019; Qi et al., 2019; Tahir, 2019; Jiang et al., 2020; Li et al., 2019b; Fu et al., 2019b; Yu et al., 2019; Zhong et al., 2019; Wang and Zhang, 2020) on UDA-based methods adopt the Convolutional Neural Network (CNN) to extract features. UDA methods transfer knowledge from a labeled source dataset to a fully unlabeled target datasets. The key of UDA is reducing the domain gap between the source dataset and target dataset. Jiang et al. (Wei et al., 2018a) bridge the domain gap using a Generative Adversarial Network (GAN). Zhong et al. (Zhong et al., 2018) proposed a Hetero-Homogeneous Learning (HHL) method which enforces camera invariance and domain connectedness on target dataset to improve

the generalization ability of re-ID model. Zhang et al. (Zhang et al., 2019) introduced a self-training method with progressive augmentation framework (PAST) to promote the model performance progressively on the target dataset. Li et al. (Li et al., 2019b) proposed a Pose Disentanglement and Adaptation Network (PDA-Net) which learns deep image representation with pose and domain information properly disentangled. Zhong et al. (Zhong et al., 2019) discovered that the intra-domain variations in the target dataset also influenced the person re-ID performance; thus, Zhong et al. designed three optimization functions to constrain network learn more knowledge in target domain. The unsupervised re-ID model is very sensitive to challenging scenarios like complex human poses, occlusion, and complex backgrounds, because of lack pre-annotated labels. To more accurately extract features from the target identity, Fu et al. (Fu et al., 2019b) proposed a Self-similarity Grouping (SSG) model, which self-generated local information from global information in spatial domain by splitting global feature maps into upper human body feature maps and lower human body feature maps. Moreover, the performance of re-ID model also can be enhanced by improving the quality of generated pseudo labels. Jiang et al. (Yu et al., 2019) learn the soft multilabels from a labeled source dataset then use it to label each image in target dataset. (Yu et al., 2019) used the soft multilabel mine the potential label information between two datasets to reduce domain gaps. Qi et al. (Qi et al., 2019) developed a camera-aware domain adaptation method to better integrate source and target domains, and proposed an unsupervised online in-batch triplet generation method to explore the underlying discriminative information in unlabeled target domain.

### 2.2.3 Fully Unsupervised Learning Methods

Although UDA-based methods have achieved superior performance, but it still need an additionally labeled dataset. It is not the fully unsupervised person Re-ID system. Several fully unsupervised methods (Yu, Wu, and Zheng, 2017; Yu, Wu, and Zheng, 2020; Ding, Khan, and Tang, 2019; Lin et al., 2019; Wang and Zhang, 2020) are proposed, which train the model without any manually annotated labels. Yu et al. (Yu, Wu, and Zheng, 2017) learn an unsupervised asymmetric metric based on asymmetric clustering on cross-view person images to deal with the view-specific interference. Yu et al. (Yu, Wu, and Zheng, 2020) followed the similar idea of (Yu,

Wu, and Zheng, 2017) but embed the asymmetric metric into a deep neural network to further improve the performance. Ding et al. (Ding, Khan, and Tang, 2019) proposed an effective clustering approach for re-ID by exploring the dispersion in statistics. Lin et al. (Lin et al., 2019) proposed a Bottom-Up Clustering (BUC) approach to merge a fixed number of clusters and fine-tune the model until convergence. In order to ease the impact of noisy clusters, Ji et al. (Ji et al., 2020) proposed a Two-stage Clustering Method to trained the network only using reliable samples. Without using an additional clustering algorithm, Wang et al. (Wang and Zhang, 2020) formulated re-ID as a multi-label classification task by generating pseudo labels via similarity measurement. Based on the MLCReID, Tang et al. (Tang and Jo, 2021) leveraged the eligible neighbors as additional reference information to further boost the model performance in ranking accuracy. Yang et al. (Yang et al., 2021) designed a noise-tolerant loss function, which enables the model robust against noisy pseudo labels. Generally, existing FUL-based re-ID researches mainly aim to design an accurate pseudo label prediction method or a noise-robust loss function.

#### 2.2.4 Pseudo Label Prediction

Based on the pseudo label prediction methods, unsupervised person re-ID can be generally divided into two categories: clustering-based label prediction and similarity measurement-based label prediction.

The core idea of clustering-based label prediction (Yang et al., 2021; Fan, Zheng, and Yang, 2017; Ge, Chen, and Li, 2020) is that they perform a clustering algorithm on Convolutional Neural Network (CNN) features to generate pseudo labels for training. Clustering-based methods pull samples close to their class centroids. Fan et al. (Fan, Zheng, and Yang, 2017) proposed a progressive unsupervised learning (PUL) method, which can be seen as an original clustering-based re-ID work. They iterated clustering and fine-tune CNN step by step until convergence. Because the clustering results may noisy, subsequent researches (Yang et al., 2021; Ge, Chen, and Li, 2020; Ji et al., 2020) mainly focus on refining noisy labels.

The core idea of similarity measurement-based label prediction (Zhong et al., 2019; Lin et al., 2020; Wang and Zhang, 2020) is that they estimate similarities among

samples to select positive samples for training. Similarity measurement-based methods pull samples close to their near neighbors. The existing methods selected positive samples based on some fixed rules. ECN (Zhong et al., 2019) is the original similarity measurement-based re-ID method. ECN (Zhong et al., 2019) and SSL (Lin et al., 2020) selected  $k$ -Nearest Neighbors ( $k$ -NN) of each image as its positive samples. Subsequently, MLCReID (Wang and Zhang, 2020) proposed to select positive samples using a pre-defined and fixed similarity threshold  $t$ .

The existing researches only focus on one of the pseudo label prediction methods. As mentioned above, clustering-based and similarity measurement-based methods lead the model to learn in different directions, and thus they can be complementary to each other. This paper predicts clustering-based and similarity measurement-based pseudo labels for each image simultaneously to train the model jointly to achieve better performance.

### 2.2.5 Similarity Exploration

Exploring similarity correctly is critical to the re-ID performance. Supervised object re-ID methods can use additional human pose labels (Zhou et al., 2020) or human body part segmentation regions (Huang et al., 2020) to help model learn human part similarity. However, these pre-annotated auxiliary labels are unknowable in the unsupervised learning task.

In order to boost model learning discriminative features without human part labels, several methods, including supervised method in (Fu et al., 2019a), UDA-based unsupervised method Self-Similarity Grouping (SSG) (Fu et al., 2019b), and FUL-based unsupervised method Softened Similarity Learning (SSL) (Lin et al., 2020) explored the self-defined part-based self-similarities. The part-based methods (Fu et al., 2019a; Fu et al., 2019b; Lin et al., 2020) split  $w_f \times h_f \times d$  global feature maps into  $P$  horizontal stripes to compute  $P$  numbers of part-based self-similarities. In SSG (Fu et al., 2019b),  $P = 2$  are used for exploring the potential similarities of upper human body and lower human body of unlabeled samples. They (Fu et al., 2019a; Fu et al., 2019b; Lin et al., 2020) ignore a factor that the human region does not always locate in the center of the image, as shown in Fig. 4(c). The human location variance makes the inconsistency in each partition. Moreover, data augmentation strategies, such as

random crop, random erasing, and random rotation, further increases human location variance.

## Chapter 3

# Proposed Fully Unsupervised Object Re-identification

In this chapter, we first introduce the experimental datasets and evaluation Metrics in Section 3.1. Then, the proposed methods are introduced. We focus on designing three aspects of an object search system. First, a more effective and robust sampling strategy, called Irregular Sampling (IS) is designed to avoid information leakage and training imbalance. Second, multiple pseudo labels refinement strategies are proposed to mitigate the effects of label noise. Third, two learning strategies are investigated and integrated to enforce instances centroid-towards learning and neighborhoods-towards learning simultaneously. Fourth, instance-to-instance loss functions and relative hard samples learning are designed.

### 3.1 Experimental Datasets and Evaluation Metrics

We evaluate the proposed method on the three large-scale and mainstream person re-ID datasets and one vehicle re-ID dataset.

TABLE 3.1: DATASET STATISTICS. **IDS:** IDENTITIES. **CAMS:** THE NUMBERS OF CAMERAS.

Dataset	IDs			Images			Cams
	Training	Query	Gallery	Training	Query	Gallery	
Market-1501	751	750	751	12,936	3,368	16,364	6
DukeMTMC-reID	702	702	1,110	16,522	2,228	17,661	8
MSMT17	1,041	3,060	3,060	32,621	11,659	82,161	15
VeRi-776	576	200	200	37,778	1,678	11579	20

### 3.1.1 Person re-ID datasets

Every person re-ID datasets consist of three subsets, including a training set, a query set, and a gallery set. The number of identities (IDs) and images in the two datasets are reported in Table 3.1. The training set is used for training, and both the query and gallery sets are used for testing. The identities of the training set are different from the query and gallery sets.

**Market-1501** (Zheng et al., 2015b) (Market) has 6 cameras and 32,668 person images of 1,501 identities in total. 751 identities with 12,936 images are used for training, and 750 identities with 19,732 images are used for testing.

**DukeMTMC-reID** (Zheng, Zheng, and Yang, 2017; Ristani et al., 2016) (Duke) has 8 cameras and 36,411 person images of 1,404 identities in total. 702 identities with 16,522 images are used for training, and 702 identities with 19,889 images are used for testing.

**MSMT17** (Wei et al., 2018b) (MSMT) is a person re-ID dataset, which has 15 cameras and 126,441 person images of 4,101 identities in total. 1,041 identities with 32,621 images are used for training, and 3,060 identities with 93,820 images are used for testing.

### 3.1.2 Vehicle re-ID datasets

**VeRi-776** (Liu et al., 2016) (VeRi) is a vehicle re-ID dataset, which has 20 cameras and 51,003 vehicle images of 775 identities in total. 576 identities with 37,778 images are used for training, and 200 identities with 13,257 images are used for testing.

### 3.1.3 Evaluation Metrics

Two evaluation metrics are used to measure model performance. The first one is Mean Average Precision (mAP) (%), which is the average value of Average Precision (AP) over all query images. AP is the area under the precision-recall curve. Another evaluation metric 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.

## 3.2 Improved Unsupervised Object Re-identification via Irregular Sampling

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 sampling strategy. Here we introduce a more effective and robust sampling strategy, called Irregular Sampling (IS).

The purpose of the sampling strategy is to split a whole dataset into mini-batches for training. Instance discrimination tasks (He et al., 2019; Wu et al., 2018; Silva, 2020) used random sampling, which treats each instance as a single class and samples  $P = 1$  numbers of instance in each mini-batch. Random sampling results in a bad performance in re-ID task because of lack of intra-/inter-class learning (Ge et al., 2020; Han et al., 2021b). State-of-the-art re-ID methods (Ge et al., 2020; Chen, Lagadec, and Bremond, 2021) performed triplet sampling to perform intra-/inter-class learning and achieved impressive performance. Triplet sampling (Schroff, Kalenichenko, and Philbin, 2015; Han et al., 2021b) samples a fixed number  $P > 1$  of same identity instances in each mini-batch. Therefore, small clusters need to be sampled repeatedly to ensure  $P$  instances in each mini-batch. Here we found out and demonstrated that triplet sampling harms the network performance because 1) same patterns in a mini-batch leads model over-fitting and 2) selecting repeat samples introduces an imbalanced problem between small and large clusters. Therefore, we introduce a more effective and robust sampling strategy - Irregular Sampling (IS).

Extensive experiments are performed on one vehicle re-ID dataset and two mainstream person re-ID datasets.

### 3.2.1 Related works of Sampling Strategy

Instead of designing learning frameworks or loss functions, recent works (Han et al., 2021b; Wu et al., 2017) showed that sampling strategies also play an important role in model performance. Random sampling is a widely used, simple and suitable sampling strategy for self-supervised instance discrimination task (Wu et al., 2018; He et al., 2019; Silva, 2020). Random sampling randomly selects samples from the



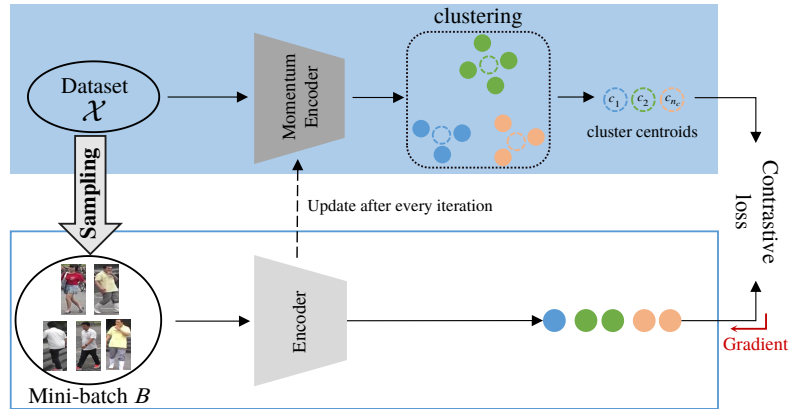


FIGURE 3.1: The illustration of the fully unsupervised object re-ID framework with sampling. Before every training epoch, cluster algorithm DBSCAN (Ester et al., 1996) is used to roughly cluster every sample in the whole dataset into  $N_c$  classes.  $C = \{c_1, \dots, c_{N_c}\}$  represents centroids of  $N_c$  classes. In every training iteration, mini-batches are generated from a dataset using sampling strategy based on clustering results. The encoder is fine-tuned according to  $C$  via centroids-towards learning, and the momentum encoder is updated by the encoder by momentum update in (He et al., 2019).

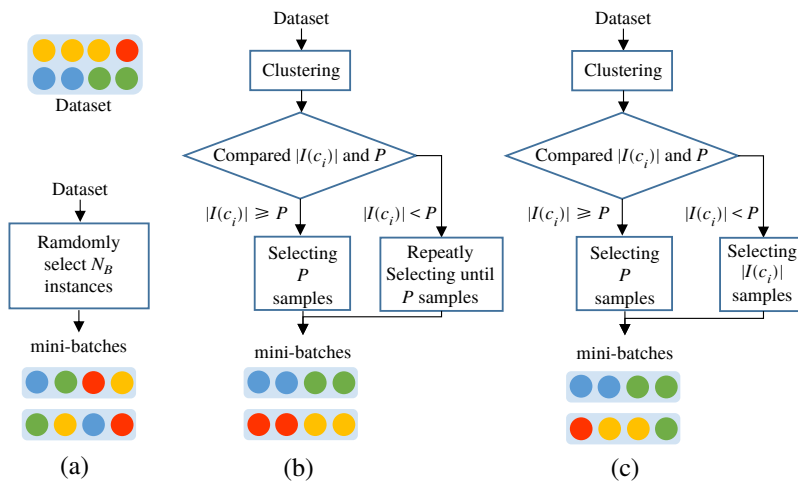


FIGURE 3.2: The illustrations of sampling strategies. (a) Random Sampling (b) Triplet Sampling (c) Our proposed Irregular Sampling. The same color represents the samples sharing the same identity class.

whole dataset to form each mini-batch, because the instance discrimination task considers each image as a distinct class for enforcing inter-class contrasting learning, as illustrated in Figure 3.2.

Random sampling is also adopted in early re-ID works (Wang and Zhang, 2020; Tang and Jo, 2021; Zhong et al., 2019), however, it can not achieve satisfying performance. Random sampling can not ensure that every mini-batch contains inter-class samples. Random sampling leads to deteriorated over-fitting and harms the object re-ID performance because of neglecting intra-class learning (Ge et al., 2020; Han et al., 2021b).

Therefore, recent researches (Ge et al., 2020; Ge, Chen, and Li, 2020; Wang et al., 2021; Xuan and Zhang, 2021a; Chen, Lagadec, and Bremond, 2021; Hu, Zhu, and He, 2021) adopted triplet sampling in re-ID task. They first roughly classify all samples into clustered inliers or unclustered outliers by clustering algorithm DBSCAN (Ester et al., 1996) or k-means. Then,  $P$  numbers of same class instances are selected to form mini-batch, intra-class and inter-class learning are performed simultaneously in every training iteration. Subsequently, Han et al. (Han et al., 2021b) proposed a Group Sampling strategy by addressing the deteriorated over-fitting problem in triplet sampling.

### 3.2.2 Proposed Method

In this section, we first introduce the fully unsupervised object re-ID architecture with the centroids-towards learning. Then, we will describe the proposed sampling strategy - Irregular Sampling (IS).

#### Centroids-towards 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 (He et al., 2019) with InfoNCE loss (Oord, Li, and Vinyals, 2018) is used as the baseline to enforce centroids-towards learning. The framework of the proposed method is illustrated in Figure 3.1.

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,

$$\theta_{me} = m\theta_{me} + (1 - m)\theta_e \quad (3.1)$$

Before each training epoch starts, given an unlabeled training dataset  $X = \{x_1, \dots, x_N\}$ , all images representations  $F_{me} = \{f_{me,1}, \dots, f_{me,N}\}$  are extracted by the momentum encoder. Then, unsupervised dense-based clustering algorithm DBSCAN (Ester et al., 1996) clusters  $F_{me}$  into  $N_C$  numbers of clusters. After that, cluster centroids  $C = \{c_0, \dots, c_{N_C}\}$  are computed as the mean vector of all instances in the cluster. This clustering results are also used to split  $X$  into mini-batches by irregular sampling.

In each training iteration, given an irregular sampled mini-batch  $B$ ,  $F_e = \{f_{e,1}, \dots, f_{e,N_B}\}$  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 centroids-towards learning  $L_C$  of an instance is designed based on InfoNCE loss (Oord, Li, and Vinyals, 2018) as follows,

$$L_C = -\log \frac{\exp(f_{e,i} \cdot c^+) / \tau}{\exp(f_{e,i} \cdot C) / \tau} \quad (3.2)$$

, 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.

### Proposed Irregular Sampling Strategy

The MoCo-based self-supervised contrastive learning framework (He et al., 2019) is adopted as the baseline framework in this work, as shown in Figure 3.1. Before every training epoch, unsupervised cluster algorithm DBSCAN (Ester et al., 1996) is used to roughly cluster every samples in the whole dataset into  $C = \{c_1, \dots, c_{N_c}\}$  classes.  $N_c$  denotes the numbers of clusters. Then, generated  $c_i$  fine-tunes the encoder model iteration by iteration until convergence.

The purpose of the sampling strategy is to split a whole dataset  $X$  into mini-batches  $\{B_1, \dots, B_{N_B}\}$  for contrastive training batch by batch.  $N_B$  denotes the numbers of batches,  $N_B = \frac{|X|}{|B_i|}$ . Previous works (Ge et al., 2020; Chen, Lagadec, and Bremond, 2021; Xuan and Zhang, 2021a; Yang et al., 2021; Wang et al., 2021) utilized triplet sampling to enforce intra-class and inter-class learning simultaneously in every training iteration by sampling  $P$  numbers of same class samples in every mini-batch, as illustrated in Figure 3.2 (b). Different colors represent samples having different  $c_i$  classes. We denote the samples within the same cluster/class of  $c_i$  as  $I(c_i)$ , and the numbers of samples in  $I(c_i)$  is represented as  $|I(c_i)|$ . If there is a large cluster ( $|I(c_i)| \geq P$ ), only  $P$  instances are sampled. If there is a small cluster ( $|I(c_i)| < P$ ), instances will be repeatedly sampled.  $P$  is a hyper-parameter, which plays an important role in model performance. (Han et al., 2021b) demonstrated that larger  $P$  brings benefits in Memory-based re-ID framework by strengthening statistical stability of each class in a mini-batch. However, we found out that larger  $P$  harms model performance in MoCo-based re-ID framework, especially in  $P = 16$ .

To investigate the impact of  $P$  in the MoCo-based unsupervised re-ID framework, we perform experiments and report the results in Figure 3.3. With the increase of  $P$ , triplet sampling harms the model performance consistently in three datasets, especially in  $P = 16$ . The dramatically declines are caused by the resampling in triplet sampling in two ways: 1) information leakage within a batch, and 2) training imbalance between small and large clusters.

**Information leakage within a batch.** Repeatedly sampling in triplet sampling brings same and repeat patterns in a mini-batch, hence the model may learn to exploit such simple and repeat mini-batch information instead of learning correct representations 3.3. A perturbation factor  $\sigma_p \sim N(0, 0.5)$  is added  $P$  to disturb the patterns in a mini-batch. More specifically,  $P + \sigma_p$  same class samples are sampled in every mini-batch. Figure 3.3 shows that the triplet sampling with perturbation factor makes good re-ID results in  $P = 16$ .

**Training imbalance between small and large clusters.** As mentioned in above, triplet sampling resamples samples from small cluster ( $|I(c_i)| < P$ ) to form every mini-batch. Therefore, samples from smaller clusters have a higher proportion of updated than samples from large clusters in every training iteration. To investigate

the impact of this imbalance problem, we further perform experiments “triplet sampling + w/o resampling” in Figure 3.3. Without resampling from clusters, the steady, continued and consistent performance rise is observed with the increase of  $P$ .

The experimental results demonstrate the negative effect of resampling operation in triplet sampling. It is interesting to observe that without resampling indirectly breaks the same pattern in a mini-batch, as shown in Figure 3.2 (c). Therefore, here we introduce a simple but robust sampling strategy, called Irregular Sampling (IS). Different from triplet sampling, IS selects  $|I(c_i)|$  numbers of instances for small clusters to avoid repeated sampling.

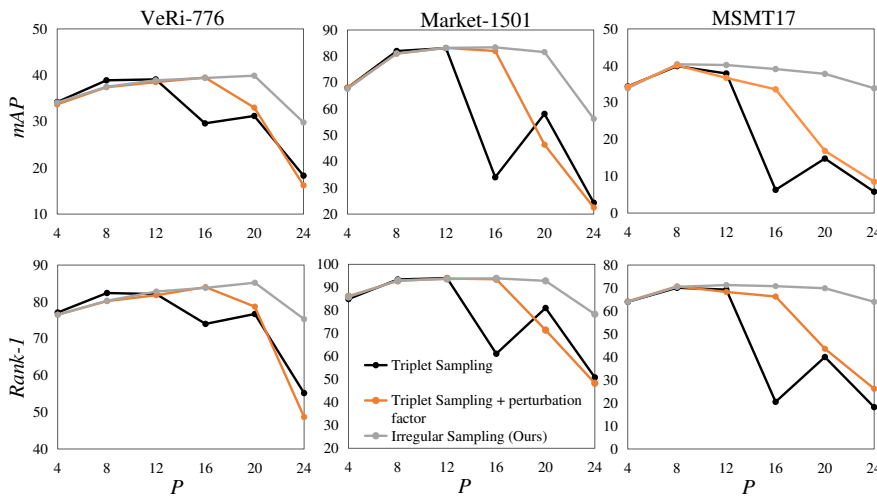


FIGURE 3.3: The model performance of different sampling strategies in different  $P$ . The X-axis represents different numbers of positive samples  $P$ , and Y-axis represents the model performance. Graphs in the first-row report performance in mAP (%), and the second-row report performance in Rank-1 (%). Graphs in different columns report performance on different datasets.

investigate and analyze the performances of the current sampling strategy in different numbers of positive samples in a mini-batch under the same learning framework and loss function, then we proposed a more effective and robust sampling strategy - Irregular Sampling (IS).

### 3.2.3 Experiments

#### Datasets and Evaluation Metrics

We evaluate the proposed method on three large-scale and mainstream datasets, i.e., one vehicle re-ID datasets, and two person re-ID datasets. The details are mentioned

Method	Market-1501		MSMT17	
	mAP	Rank-1	mAP	Rrank-1
BUC (Lin et al., 2019)	29.6	61.9	-	-
HCT (Zeng et al., 2020)	56.4	80.0	-	-
MMCL (Wang and Zhang, 2020)	45.5	80.3	11.2	35.4
DSCE (Yang et al., 2021)	61.7	83.9	15.5	35.2
SpCL (Ge et al., 2020)	79.1	88.1	19.1	42.3
CAP (Wang et al., 2021)	79.2	91.4	36.9	67.4
GS (Han et al., 2021b)	79.2	92.3	24.6	56.2
ICE (Chen, Lagadec, and Bremond, 2021)	82.3	93.8	38.9	70.2
CCL (Dai et al., 2021)	82.1	92.3	27.6	56.0
HHCL (Hu, Zhu, and He, 2021)	84.2	93.4	-	-
Ours (w/ IS)	83.4	<b>93.9</b>	<b>40.2</b>	<b>71.3</b>

TABLE 3.2: Experimental results of our proposed method and other fully unsupervised re-ID methods on two person re-ID datasets. The top result is highlighted in bold.

in Section 3.1.1 and Section 3.1.2. Two evaluation metrics are mentioned in Section 3.1.3

### 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  $m = 0.999$ ,  $\tau = 0.05$ , are used in all experiments for fair comparisons. Other hyper-parameters are selected for each datasets for achieving the best performance. In VeRi-776,  $P = 20$  and  $\tau_c = 0.15$ . In Market-1501,  $P = 16$  and  $\tau = 0.1$ . In MSMT17,  $P = 8$  and  $\tau = 0.1$ . The model performance with different hyperparameters are reported in Figure 3.3. 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 6 hours on MSMT17 and VeRi-776.

Method	VeRi-776	
	mAP	R-1
SSML (Yu and Oh, 2021)	26.7	74.5
SpCL (Ge et al., 2020)	36.9	79.9
Ours (w/ IS)	39.9	<b>85.2</b>

TABLE 3.3: Experimental results of our proposed method and other fully unsupervised re-ID methods on vehicle re-ID datasets VeRi-776.

### Comparisons with The State-of-the-Arts

The comparisons with the state-of-the-arts fully unsupervised methods on one vehicle re-ID dataset VeRi-776 in Table 3.3. We obtain  $mAP=39.9\%$  and  $Rank-1=85.2\%$ , which considerably outperforms SpCL (Ge et al., 2020). The superior performance indicates the effectiveness of our proposed Irregular sampling.

Comparisons are also performed in two person re-ID datasets, i.e., Market-1501 and MSMT17, which are reported in Table 3.3. On Market-1501, our method achieves the best performance with  $mAP=83.4\%$  and  $Rank-1=93.9\%$ . Compared to the best MoCo-based re-ID method IC, our method achieves good and competitive results. Specifically, our method outperforms ICE by 1.3% in mAP and 2.5% in Rank-1 in the largest and most difficult person re-ID datasets MSMT17.

### 3.2.4 Ablation Studies

#### Effectiveness of Irregular Sampling

We illustrate the model performance using triplet sampling and our proposed irregular sampling with different numbers of instances  $P$  in Figure 3.3. It can be observed that  $P$  plays an important role in model performance. Small or large  $P$  indicates less or more instances belonging to the same class are selected in mini-batches, respectively. With the increase of  $P$  in triplet sampling, the model performance is increased first and then decreased rapidly. The performance increase is because selecting more positive instances helps the model learn more intra-class information and brings more statistical stability (Han et al., 2021b). However, selecting more positive instances also causes a more serious imbalanced situation between small clusters and large clusters because more small clusters are re-sampled.

Method	mAP	Rank-1
Random Sampling	55.3	76.5
Group Sampling † (Han et al., 2021b)	76.0	91.0
Irregular Sampling	<b>83.4</b>	<b>93.9</b>

TABLE 3.4: Ablation study on different sampling methods. † denotes results are obtained by our experiments using publicly available source code.

The above situation is not be observed after we remove the re-sampling operation. The performances do not decrease rapidly with the increase of  $P$  in irregular sampling. The experiments show that irregular sampling is a more effective and robust sampling strategy than triplet sampling.

We further compare our proposed irregular sampling with random sampling (He et al., 2019) and group sampling (Han et al., 2021b) in Table 3.4. It can be clearly seen that random sampling has poor performance because it does select multiple positive instances in mini-batches, leading the model to neglect intra-class information. Group sampling can not achieve satisfying performance because it is designed based on Memory Bank architecture (Ge et al., 2020; Wu et al., 2018). Table 3.4 shows that group sampling is not suitable in MoCo-based architecture. Finally, irregular sampling achieves the best performance in MoCo-based architecture.

### 3.2.5 CONCLUSIONS

In this work, we proposed a fully unsupervised object re-ID method, which can be trained without using any labeled information. The existing sampling strategies are investigated and compared. Based on the drawbacks of existing methods, we propose an effective and robust sampling strategy - irregular sampling. Experimental results on one vehicle and two person re-ID datasets show the effectiveness of the proposed methods.



### 3.3 Unsupervised Person Re-identification via Mining Label Homogeneity

#### 3.3.1 Introduction

To make unsupervised training possible, the model needs to self-generate pseudo labels by clustering algorithm (Lin et al., 2019; Fu et al., 2019b; Zhang et al., 2019) or exemplar memory-based algorithm (Wang and Zhang, 2020; Zhong et al., 2019). Unlike human-annotated labels used in supervised learning, the self-generated pseudo labels contain noisy labels that substantially hinder the model’s capability because the network learns to extract discriminative features based on these pseudo labels. Therefore, unsupervised person re-ID performance still significantly falls behind the supervised person re-ID (Zhou et al., 2019; Zhou et al., 2020; Huang et al., 2020).

Consequently, the key to improving the unsupervised person re-ID model performance is to generate high-quality pseudo labels which can represent the unlabeled data domain distribution. Several studies (SML; Wang and Zhang, 2020; Fu et al., 2019b; Zhang et al., 2019; Zhong et al., 2019; Wei et al., 2018a; Zhong et al., 2018; Qi et al., 2019; Jiang et al., 2020; Li et al., 2019b) transferred knowledge from a labeled source dataset to an unlabeled target dataset by transfer learning, thereby learning better image representation. Subsequently, a fully unsupervised method MMCL (Wang and Zhang, 2020) formulated unsupervised person Re-ID as a multi-label classification task and achieved good re-ID performance. MMCL did not require any manually labeled dataset and learn re-identification information only from unlabeled images, thus it is easily deployed to a new scenario.

In fact, after the networks can roughly capture the training data distribution and predict pseudo labels, the predicted pseudo labels also can be served as auxiliary labels. In this study, we mine the relations among generated pseudo labels and use them as additional supervisions for an unlabeled image to further refine noisy labels. We construct an Pseudo Label Memory (PLM) for storing the up-to-date generated pseudo labels  $\hat{Y} = \{y_1, y_2, \dots, y_n\}$  of all images in  $X = \{x_1, x_2, \dots, x_n\}$ . With the help of PLM, we inquire two auxiliary labels based on two discovered underlying relations, i.e, Symmetric Homogeneity (S) and Neighbor Homogeneity (N), thereby refining the main label. Figure 3.4 shows our re-ID framework SNNet which optimizes the

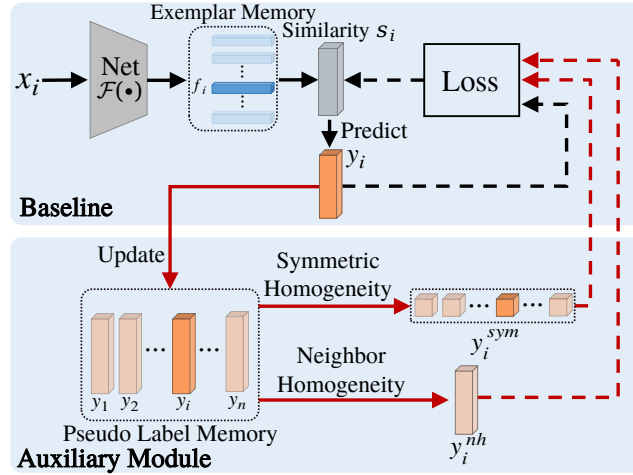


FIGURE 3.4: The framework of our proposed SNNNet. The black line indicates the baseline network. Apart from the predicted main pseudo label  $y_i$ , we proposed an auxiliary module to seek auxiliary labels  $y_i^{sym}$  and  $y_i^{nh}$  as additional supervisions to optimize the network.

network  $\mathcal{F}(\cdot)$  under the joint supervisions of main pseudo label  $y_i$  and two auxiliary pseudo labels  $y_i^{sym}$  and  $y_i^{nh}$ . Different from training several student networks collaboratively and mutually in deep mutual learning (Zhang et al., 2018), our proposed SNNNet framework trains one single network via labels of different samples, thereby training different samples mutually.

Based on the above aspects, we propose a fully unsupervised person re-ID training strategy. The contributions could be summarized as three-fold. (1) We discovered two underlying label homogeneity constraints in the exemplar memory-based person re-ID method. (2) To make the network learn two label homogeneities possible, we design an unsupervised mutual label learning framework, which optimizes the network under the joint supervisions of main pseudo labels and auxiliary labels. (3) The proposed fully unsupervised person re-ID framework SNNNet shows exceptionally strong performances.

### 3.3.2 Proposed Method

Our proposed re-ID framework is shown in Figure 3.4. The model is mainly divided into two parts: the baseline network and the proposed auxiliary module. The baseline network is the general exemplar memory-based re-ID network (Wang and Zhang, 2020). Given an unlabeled dataset  $X = \{x_1, x_2, \dots, x_n\}$ , the baseline network computes the similarity score  $s_i$  to seek main pseudo label  $y_i$  for each image  $x_i \in X$ .

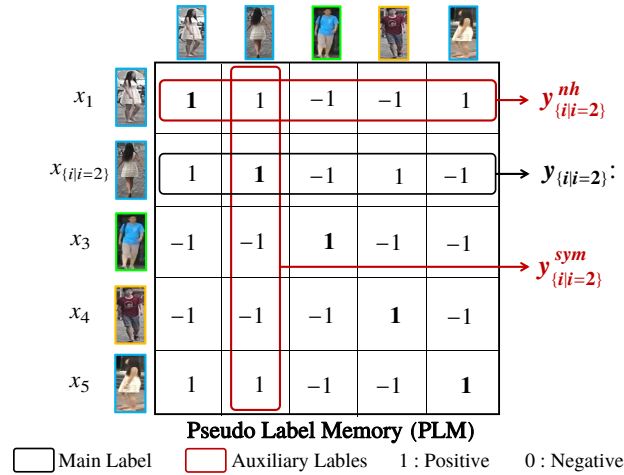


FIGURE 3.5: The illustration of pseudo label memory (PLM).  $n = 5$  is assumed in this figure. Ideally,  $y_i = y_i^{sym}$  because of symmetric homogeneity constrain, and  $y_i = y_i^{nh}$  because of neighbor homogeneity constrain.

Unlike the human-annotated label, generated main pseudo label  $y_i$  contain the noisy. To mitigate the effects of noisy pseudo labels, we propose an auxiliary module that seeks two auxiliary pseudo labels  $y_i^{sym}$  and  $y_i^{nh}$  to optimize the neural network with  $y_i$  jointly.

Subsequently, we first introduce the fully unsupervised object re-ID baseline architecture with the centroids-towards learning. Then, we will describe the proposed auxiliary module.

### The Baseline Network

Given a set of unlabeled person images  $\{x_1, x_2, \dots, x_n\} \in X$ , the  $d$ -dimensional feature  $f_i$  of  $x_i$  are extracted by backbone network  $\mathcal{F}(\cdot)$  to form the exemplar memory  $\mathcal{M}$ .  $n$  is the number of images in  $X$ . The size of  $\mathcal{M}$  is  $n \times d$ . Using  $\mathcal{M}$ , the cosine similarity between  $x_i$  and the other image  $x_j \in X$  can be computed as,

$$s_i[j] = f_i \times f_j^\top, \quad j = 1, \dots, n. \quad (3.3)$$

where  $s_i$  is an  $n$ -dimensional vector, range in  $[-1, 1]$ . Based on  $s_i$ , the Memory-based Positive Label Prediction (MPLP) (Wang and Zhang, 2020) is used to predict the main pseudo multi-class label  $y_i$  for  $x_i$ .

In each training iteration,  $y_i$  is predicted and used to fine-tune the model until

convergence. The Memory-based Multi-label Classification Loss (MMCL) (Wang and Zhang, 2020) is used in baseline network to directly regress similarity score  $s_i$  to  $y_i$  as follows:

$$\mathcal{L}_{baseline} = \|s_i - y_i\|^2 \quad (3.4)$$

### Auxiliary Module

The re-ID model trained only with  $y_i$  is usually sensitive to the noise in  $y_i$ , which hinders the feature learning in  $\mathcal{F}(\cdot)$ . To mitigate the effects of noisy pseudo labels, we proposed the auxiliary module, which seeks two auxiliary labels by investigating two underlying constraints in PLM, i.e., symmetric homogeneity and neighbor homogeneity. Ideally, the generated main label and two auxiliary labels should be the same because of two homogeneity constraints. Based on them, we design a mutual label learning training strategy to enforce one single network mutual training with different labels.

**Pseudo Label Memory (PLM):** To mine the generated labels relation on the whole dataset, the Pseudo Label Memory (PLM) is first constructed in this paper for storing the up-to-date main pseudo labels  $\hat{Y} = \{y_1, y_2, \dots, y_n\}$  of all images in  $X$ . The PLM is notated as  $\mathcal{P}$ , contains  $n$  slots, in which each slot storing a  $n$ -dimensional pseudo label  $y_i$ . Therefore, the size of PLM is  $n \times n$ . It is noteworthy that our proposed PLM is different with exemplar memory  $\mathcal{M}$  in (Zhong et al., 2019; Wang and Zhang, 2020),  $\mathcal{M}$  is used to store the up-to-date features of all images in  $X$ . An illustration of PLM is shown in Figure 3.5 which assumes  $n = 5$ .

To store and inquire up-to-date labels for all images, PLM is updated using generated  $y_i$  in every training iteration through,

$$\mathcal{P}[i] \leftarrow y_i \quad (3.5)$$

**Symmetric Homogeneity:** Ideally, two images  $x_i$  and  $x_j$  should be mutual positive samples or mutual negative samples for each other. However, because the network is updated in each iteration based on the input batch size, the  $x_i$  and  $x_j$  might not be predicted as mutual positive or negative samples if they are fed-forward into

**Algorithm 1:** SNNet Algorithm

---

```

1  $x_i$ : Input image in unlabeled dataset  $X$  Initialize weighting factors  $\lambda^{sym}$  and  $\lambda^{nh}$ 
2 Initialize  $\mathcal{P}$  by identity matrix with size  $n \times n$ 
3 for  $epoch$  in  $[1, num\_epochs]$  do
4   for  $iter$  in  $[1, num\_iter]$  do
5     1. Predict  $y_i$  for input image  $x_i$ 
6     2. Inquire  $y_i^{sym}$  from  $\mathcal{P}$  based on symmetric homogeneity
7     3. Inquire  $y_i^{nh}$  from  $\mathcal{P}$  based on neighbor homogeneity
8     4. Update  $\mathcal{P}$  using predicted pseudo labels  $y_i$ 
9     5. Joint update  $\mathcal{F}(\cdot)$  by the loss function Equation 3.10

```

---

the network in a different iteration. To effectively avoid the asymmetric error amplification, we seek auxiliary symmetric label  $y_i^{sym}$  of image  $x_i$  to enforce symmetric constraint into the network via our proposed mutual label learning strategy.

An illustration of the symmetric homogeneity of labels is shown in Figure 3.5. The main pseudo label  $y_{\{i|i=2\}}$  of  $x_{\{i|i=2\}}$  is save in PLM. The auxiliary symmetric label of  $y_i$  is inquired from PLM as follows,

$$y_i^{sym} = \mathcal{P}[:, i] \quad (3.6)$$

If there is no noisy pseudo labels,  $\mathcal{P}$  is a symmetric matrices; in other words,  $y^i = y_i^{sym}$ , ideally. We utilize the symmetric constraint of generated pseudo label  $y_i$  in  $\mathcal{P}$  to mitigate the effects of noisy pseudo labels by also regressing similarity score  $s_i$  to  $y_i^{sym}$  as follows:

$$\mathcal{L}_{sym} = \left\| s_i - y_i^{sym \top} \right\|^2 \quad (3.7)$$

**Neighbor Homogeneity:** For an image  $x_i \in X$ , there may have another image  $x_e$  in  $X$  share same multi-class label with  $x_i$ . We aim to seek the  $x_e$  and exploit its main pseudo label  $y_e$  as an auxiliary label of  $x_i$  to further mitigate the effects of noise in  $y_i$ . To achieve this objective, we find the nearest neighbors of  $x_i$  according to the similarity between  $x_i$  and other images in  $X$ . The image share the highest similarity with  $x_i$  is the nearest neighbor of  $x_i$ . Then, we define the indexes of the nearest neighbor as  $e$ . Ideally, the predicted labels of  $x_i$  and its nearest neighbor  $x_e$  are same. Based on this constraint, we utilize the main label of  $x_e$  as the auxiliary neighbor label  $y_i^{nh}$  of image  $x_i$  to enforce neighbor homogeneity constraint into the network via our proposed mutual label learning strategy. The  $y_i^{nh}$  is inquired from PLM as

TABLE 3.5: Comparison with other fully unsupervised person re-ID methods on Market-1501 and DukeMTMC-ReID Dataset. “\*”: Baseline method, reproduced by us based on the authors’ code. “↑”: Results that outperforms baseline. Results that surpass all methods are **bold**.

Method	Reference	Market		Duke	
		R-1	mAP	R-1	mAP
CAMEL (Yu, Wu, and Zheng, 2017)	ICCV17	54.5	26.3	-	-
DECAMEL (Yu, Wu, and Zheng, 2020)	TPAMI18	60.2	32.4	-	-
BUC (Lin et al., 2019)	AAAI19	66.2	38.3	47.4	27.5
DBC (Ding, Khan, and Tang, 2019)	BMVC19	69.2	41.3	51.5	30.3
SSL (Lin et al., 2020)	CVPR20	71.7	37.8	52.5	28.6
MMCL* (Resnet-50)	CVPR20	79.1	44.1	63.6	39.0
MMCL* (OSNet)	CVPR20	80.3	45.0	66.3	41.9
SNNNet (Resnet-50)	<b>Ours</b>	80.2↑	48.3↑	66.1↑	41.5↑
SNNNet (OSNet)	<b>Ours</b>	<b>85.1↑</b>	<b>58.3↑</b>	<b>69.2↑</b>	<b>46.7↑</b>

follows,

$$y_i^{nh} = \mathcal{P}[e] \quad (3.8)$$

An illustration of auxiliary neighbor label  $y_i^{nh}$  is shown in Figure 3.5. Apart from the  $y_i$  and  $y_i^{sym}$ , We further mitigate the effects of noisy pseudo labels by regressing similarity score  $s_i$  to  $y_i^{nh}$  as follows:

$$\mathcal{L}_{nh} = \left\| s_i - y_i^{nh} \right\|^2 \quad (3.9)$$

### Overall Loss

Different from training several student networks collaboratively and mutually in deep mutual learning (Zhang et al., 2018), our proposed mutual label learning trains one single network  $\mathcal{F}(\cdot)$  via the main pseudo label  $y_i$  and two auxiliary labels in a collaborative training manner.

The overall loss is the summation of  $\mathcal{L}_{baseline}$ ,  $\mathcal{L}_{sym}$ , and  $\mathcal{L}_{nh}$ , which combines Equation 3.4, Equation 3.7, and Equation 3.9 and is formulated as,

$$\mathcal{L} = \frac{1}{\eta} (\lambda^{base} \mathcal{L}_{baseline} + \lambda^{sym} \mathcal{L}_{sym} + \lambda^{nh} \mathcal{L}_{nh}) \quad (3.10)$$

where  $\lambda^{base}$ ,  $\lambda^{sym}$  and  $\lambda^{nh}$  are the weighting parameters of  $\mathcal{L}_{baseline}$ ,  $\mathcal{L}_{sym}$  and  $\mathcal{L}_{nh}$ .  $\eta$  is the normalized parameter,  $\eta = (\lambda^{base} + \lambda^{nh} + \lambda^{sym})$  to keep the scale of the gradient of loss = 1.0. Our model update steps are summarized in Algorithm 1.

TABLE 3.6: Methods comparison when tested on Market-1501 and DukeMTMC-reID. **Baseline**: Baseline model trained with main pseudo labels. **S**: Auxiliary symmetric labels  $y_i^{sym}$ . **N**: Auxiliary neighbor label  $y_i^{nh}$ .

Method	Market-1501				DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Baseline	80.3	89.9	93.0	45.0	66.3	77.7	81.1	41.9
Baseline + S	83.8	91.3	93.8	49.0	67.5	78.5	81.8	44.1
Baseline + S + N	<b>85.1</b>	<b>92.6</b>	<b>94.3</b>	<b>58.3</b>	<b>69.2</b>	<b>79.6</b>	<b>83.2</b>	<b>46.7</b>

### 3.3.3 Experiments Setting

#### Datasets and Evaluation Metrics

We evaluate the proposed method on two large-scale and mainstream datasets, i.e., Market-1501 (Zheng et al., 2015b) (Market) and DukeMTMC-reID (Zheng, Zheng, and Yang, 2017; Ristani et al., 2016) (Duke). The details are mentioned in Section 3.1.1 and Section 3.1.2. Two evaluation metrics are mentioned in Section 3.1.3.

#### Implementation Details

The experiments are performed using one NVIDIA GeForce Titan 1080Ti GPU with 11 GB of memory. The experiments are implemented on PyTorch. The Resnet-50 (He et al., 2016) and OSNet (Zhou et al., 2019) is adopted as the backbone network. Following the previous works (Zhong et al., 2019; Wang and Zhang, 2020), we remove the subsequent layers after the pooling-5 layer and add a batch normalization layer. The backbone network is pre-trained on ImageNet (Deng et al., 2009). During training, the initial learning rate is 0.1. The learning rate is divided by ten after 30 epochs. The training batch size is 128 with ResNet-50, and 64 with OSNet backbones because of the limited memory of the GPU. The network is trained in an end-to-end fashion by the Stochastic Gradient Descent (SGD). The weighting parameters  $\lambda^{sym} = 1.0$ ,  $\lambda^{sym} = 0.8$  and  $\lambda^{nh} = 0.6$  are set for achieving the best performance.

TABLE 3.7: Evaluation with different values of  $\lambda^{sym}$ .

$\lambda^{sym}$	Market-1501				DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
0.0 (baseline)	80.3	89.9	93.0	45.0	66.3	77.0	81.5	41.9
0.2	81.9	90.9	93.4	47.6	66.9	77.8	82.1	42.6
0.4	82.2	91.1	93.3	49.6	67.1	77.5	80.9	43.3
0.6	82.9	90.6	93.2	49.3	<b>67.5</b>	78.2	<b>82.2</b>	43.5
0.8	<b>83.8</b>	<b>91.3</b>	<b>93.8</b>	49.0	<b>67.5</b>	<b>78.5</b>	81.8	<b>44.1</b>
1.0	83.5	91.2	93.2	<b>50.2</b>	66.1	77.5	81.8	42.8

### 3.3.4 Experiments Results

#### Comparison with Other Methods

As shown in Table 3.5, we compare our proposed SNNNet with other fully unsupervised methods with ResNet-50 (He et al., 2016) and OSNet (Zhou et al., 2019) backbones. We do not compare our method with the UDA-based methods, because UDA-based methods still required a labeled source dataset which is not the fully unsupervised method. The weighting parameters  $\lambda^{sym} = 1.0$ ,  $\lambda^{sym} = 0.8$  and  $\lambda^{nh} = 0.6$  are set in Table 3.5 for achieving the best performance.

MMCL\* (Wang and Zhang, 2020) is the baseline approach in here which is reproduced by us based on the authors' code. Compared to the baseline, the proposed SNNNet improves the model performance with ResNet-50 and OSNet on two datasets consistently. More specifically, on DukeMTMC-reID, 2.7% Rank-1 and 2.9% mAP improvements with ResNet-50 backbone, and 3.1% Rank-1 and 5.2% mAP improvements with OSNet backbone are observed. The results indicate the importance of our proposed two label homogeneity constraints. Moreover, the proposed SNNNet achieves the best performance among the compared methods with ResNet-50 and OSNet on Market-1501 and DukeMTMC-reID. The superior performance with different backbones indicates the robustness and effectiveness of our designed mutual label learning structure.

#### Ablation Studies

In this section, we evaluate each components in SNNNet by conducting ablation studies on Market-1501 and DukeMTMC-reID with OSNet (Zhou et al., 2019) backbones.



TABLE 3.8: Evaluation with different values of  $\lambda^{nh}$ .

$\lambda^{nh}$	Market-1501				DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
0.0 (baseline)	80.3	89.9	93.0	45.0	66.3	77.0	81.5	41.9
0.2	<b>84.8</b>	<b>91.7</b>	<b>94.1</b>	<b>55.7</b>	<b>69.2</b>	<b>79.8</b>	83.0	46.1
0.4	82.8	90.6	92.8	55.2	68.6	78.5	82.2	45.7
0.6	80.1	88.7	91.9	51.7	69.1	79.7	81.8	46.2
0.8	80.3	89.3	91.6	51.4	68.8	79.5	82.5	46.0
1.0	78.5	88.2	91.1	50.5	69.1	<b>79.8</b>	<b>83.3</b>	<b>46.7</b>

**Effectiveness of Mutual Label Learning:** we conduct ablation studies to investigate the effectiveness of the proposed mutual label learning in Table 3.6. The weighting parameters  $\lambda^{sym} = 1.0$ ,  $\lambda^{sym} = 0.8$  and  $\lambda^{nh} = 0.6$  are set in Table 3.6 for achieving the best performance. First, we report the result of the baseline network as “Baseline” in Table 3.6, which not incorporates auxiliary labels into the training. Our proposed “Baseline + S” and “Baseline + S + N” consistently improve the results over “baseline”, e.g., from baseline 80.3% to 85.1% in rank-1 and 45.0% to 58.3% on Market-1501, and 66.3% to 69.2% in rank-1 and 41.9% to 46.7% on DukeMTMC-reID. The results demonstrate the proposed mutual label learning with auxiliary labels is an effective way to improve the re-ID model performance, and the necessity of high-quality pseudo labels in the unsupervised person re-ID task.

**Effectiveness of Symmetric Homogeneity:** In Table 3.7, we compare different values of  $\lambda^{sym}$  by keeping  $\lambda^{nh} = 0.0$  in Equation 3.10 for analyzing the effect of symmetric homogeneity to baseline network. When  $\lambda^{sym} = 0.0$ , the re-ID model is only trained with main pseudo labels  $y_i$ . Comparing  $\lambda^{sym} = 1.0$  to  $\lambda^{sym} = 0.0$ , we observe 3.2% Rank-1 and 5.2% mAP improves on Market-1501. Comparing  $\lambda^{sym} = 0.8$  to  $\lambda^{sym} = 0.0$ , we observe 1.2% Rank-1 and 2.2% mAP improves on DukeMTMC-reID. The improvements demonstrate the effectiveness of our proposed symmetric homogeneity. We observe that higher  $\lambda^{sym}$  value achieves better result on Market-1501. Different to Market-1501, we achieve best results on DukeMTMC-reID when  $\lambda^{sym} = 0.8$  rather than  $\lambda^{sym} = 1.0$  because of over-fitting.

**Effectiveness of Neighbor Homogeneity:** Table 3.8, investigates the effect of neighbor homogeneity in our method by varying  $\lambda^{nh}$  from 0.0 to 1.0 on both Market-1501 and DukeMTMC-reID datasets. We keep  $\lambda^{sym} = 0.0$  in Equation 3.10 for analyzing the effect of neighbor homogeneity to baseline network. Using auxiliary neighbor

labels significantly boosts the performance on Market-1501 and DukeMTMC-reID, e.g., from baseline 80.3% to 84.8% in rank-1 and 45.0% to 55.7% on Market-1501. With the increasing of  $\lambda^{nh}$ , we observe the model is very easy to over-fitting. However, Comparing  $\lambda^{nh} = 1.0$  to  $\lambda^{nh} = 0.0$ , we still observe improvements on both Market-1501 and DukeMTMC-reID. The significant improvements demonstrate the necessity of our proposed neighbor homogeneity constraint.

### 3.3.5 Conclusion

This paper introduces an exemplar memory-based fully unsupervised method for person re-ID task via mining two underlying label homogeneities, symmetric homogeneity and neighbor homogeneity. We design a mutual label learning framework to enforce two label homogeneities constraints into the network training by optimizing the network under the joint supervision of the main pseudo label and two auxiliary labels in every training iteration. The experiment results on Market-1501 and DukeMTMC-reID demonstrate the effectiveness of our approach.

## 3.4 Fully Unsupervised Person Re-Identification via Centroids and Neighborhoods Joint Learning

### 3.4.1 Introduction

Based on the pseudo label prediction methods, unsupervised person re-ID can be generally divided into Clustering-based Label Prediction (C-LP) (Yang et al., 2021; Ge et al., 2020; Fan, Zheng, and Yang, 2017; Xuan and Zhang, 2021b; Ge, Chen, and Li, 2020) and Similarity Measurements-based Label Prediction (SM-LP) (Zhong et al., 2019; Lin et al., 2020; Wang and Zhang, 2020; Tang and Jo, 2021), where the C-LP methods maintain state-of-the-art performance to date by introducing an additional unsupervised clustering algorithm.

The core idea of C-LP is performing a clustering algorithm on Convolutional Neural Network (CNN) features to generate pseudo labels for training. Fan et al. (Fan, Zheng, and Yang, 2017) can be seen as an original work studying C-LP methods. They proposed a Progressive Unsupervised Learning (PUL) method to iterate

clustering and fine-tune CNN step by step until convergence. Because the clustering results may be noisy, subsequent researches (Yang et al., 2021; Ge, Chen, and Li, 2020; Ge et al., 2020; Xuan and Zhang, 2021b) mainly focus on refining noisy labels. Ge et al. (Ge, Chen, and Li, 2020) first grouped the features into  $M_t$  classes by clustering algorithm k-means (Agarwal and Mustafa, 2004), then they trained the model using the hard and soft pseudo-classes jointly to mitigate the effects of noisy labels. Yang et al. (Yang et al., 2021) generated pseudo-classes by clustering algorithm DBSCAN (Ester et al., 1996), then it further proposed a Dynamic and Symmetric Cross-Entropy loss (DSCE) to deal with noisy samples. In here, DBSCAN (Ester et al., 1996) is adopted because of its strong robustness against noisy samples.

There are two weaknesses in C-LP that are valuable to discuss but were ignored in the existing methods. (1) The intervals of the pseudo label prediction and model optimization are out of sync. More specifically, the model parameters are updated in every training iteration but labels are predicted before every training epoch. This asynchronism hinders the model's performance because the model can not be updated based on sync updated labels. (2) The intra-class inliers can not perform intra-class differential learning because intra-class inliers share the same labels. As illustrated in Figure 3.6 (a), intra-class inliers are enforced class centroid-towards learning without considering neighborhood information. Moreover, how to deal with the un-clustered outliers is still an open question.

To tackle the weaknesses of C-LP, we propose a Joint Label Prediction (Joint-LP) to bound C-LP and SM-LP together to utilize the merits of SM-LP. Although SM-LP (Zhong et al., 2019; Lin et al., 2020; Wang and Zhang, 2020) achieve poorer performance than C-LP, SM-LP still enjoys three merits that are complementary to the C-LP. (1) The intervals of the label prediction and model optimization are synchronous. (2) the pseudo label of each sample is different to enforce samples learning towards their own nearest neighbor. (3) SM-LP assigns the label for every sample including the outliers. As illustrated in Figure 3.6 (b), every sample is enforced learning towards their own nearest neighbor.

Moreover, we discover that the traditional Binary Cross Entropy (BCE) loss achieved satisfying performance in supervised learning methods because of correct human-annotated labels, but BCE loss achieves poor performance in unsupervised learning methods because of extensive noisy pseudo labels. Therefore, to remedy this issue,

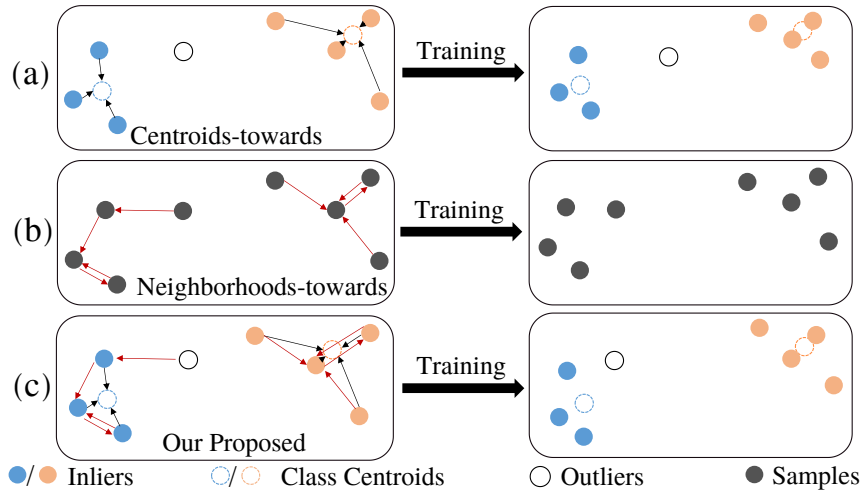


FIGURE 3.6: The illustrations of (a) C-LP (class centroids-towards learning) (b) SM-LP (neighborhoods-towards learning) (b) proposed Joint-LP (both).

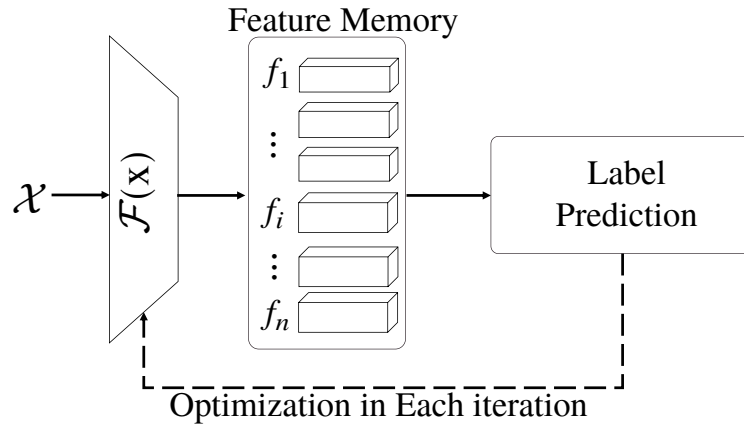


FIGURE 3.7: General framework for FUL person re-ID methods.

we further propose a Rectified BCE (ReBCE) loss to make unsupervised training with BCE loss possible by alleviating model excessive attention on noise.

Our contributions are summarized as three-fold. (1) We propose a Joint-LP method to predict high-quality pseudo labels by utilizing complementarities between C-LP and SM-LP. (2) We propose a ReBCE loss to avoid the model pay more attention to noisy labels. (3) The proposed unsupervised person re-ID method achieves superior person Re-ID performance under the FUL setting on two large-scale datasets.

To the best of our knowledge, this study is an original work studying and utilizing the complementarities between C-LP and SM-LP.

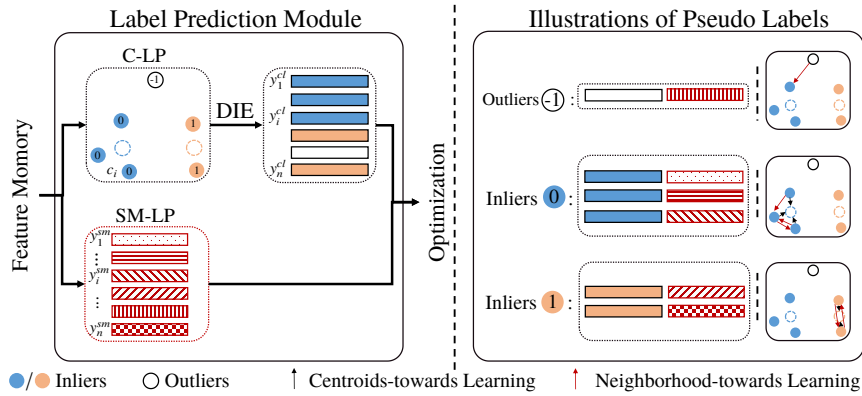


FIGURE 3.8: The illustration of our proposed Jointly Label Prediction Module (Joint-LP).

### 3.4.2 Proposed Method

#### Framework Overview

The general framework for FUL person re-ID is shown in Figure 3.7. Given an unlabeled person image  $x_{\{i|i=1,2,\dots,n\}} \in \mathcal{X}$ ,  $d$ -dimensional feature  $f_i$  are extracted by backbone network  $\mathcal{F}(\cdot)$  to form the feature memory  $\mathcal{M}$ ,  $f_{\{i|i=1,2,\dots,n\}} \in \mathcal{M}$ .  $i$  means the index of the image, which is fixed throughout training process, and  $n$  is the total numbers of images in  $\mathcal{X}$ .  $\mathcal{M}$  store features for all images in  $\mathcal{X}$ , the size of  $\mathcal{M}$  is  $n \times d$ .

Using  $\mathcal{M}$ , the proposed label prediction method Joint-LP predicts the pseudo label for every image in  $\mathcal{X}$ . Finally,  $\mathcal{F}(\cdot)$  is optimized progressively with the proposed ReBCE loss based on pseudo labels step by step. Consequently, the key to improving model performance is to generate high-quality pseudo labels which can represent the unlabeled data domain distribution.

#### Joint Label Prediction (Joint-LP)

To generate high-quality pseudo labels, we propose a Joint-LP in here. The structure of Joint-LP is shown in Figure 3.8. The Joint-LP consists of three components: C-LP, the proposed Dimension Increment pseudo-class Encoding method (DIE), and SM-LP, which will be introduced one by one.

**C-LP:** The unsupervised clustering algorithm k-means (Agarwal and Mustafa, 2004) and DBSCAN (Ester et al., 1996) are widely used to generate pseudo labels in recent studies. Following the previous works (Ge, Chen, and Li, 2020; Ge et al., 2020;

Yang et al., 2021), DBSCAN is used in here because it has stronger robustness against noisy samples. As illustrated in Figure 3.8, given the feature memory  $\mathcal{M}$ , DBSCAN assigns pseudo-class  $c_{\{i|i=1,2,\dots,n\}}$  for every image  $\in \mathcal{X}$  before every training epoch. DBSCAN assigns pseudo-class  $c_i \geq 0$  to clustered inliers, and remains  $c_i = -1$  to un-clustered outliers.

**DIE Pseudo-class Encoding:** The DBSCAN-based methods still face one challenge that the numbers of pseudo-classes keep changing during the whole training process. The centroid-based clustering algorithm K-means (Agarwal and Mustafa, 2004) generates certain cluster centroids, therefore a Fully Connected layer (FC-layer) can easily be adopted to output a probability vector for computing the cross-entropy classification loss or triplet loss (Schroff, Kalenichenko, and Philbin, 2015) as (Fan, Zheng, and Yang, 2017; Ge, Chen, and Li, 2020). However, the number of pseudo-class predicted by DBSCAN (Ester et al., 1996) are constantly changing because DBSCAN only considers high-confident samples as clustered inliers. To address this issue, we proposed a Dimension Increment pseudo-class Encoding method (DIE) to equivalently encode 1-dimensional pseudo-class  $c_i$  to  $n$ -dimensional clustering-based pseudo label  $y_i^{cl}$ . Then,  $n$  independent binary classifiers can be adopted to compute loss functions easily.

The intuition of the proposed DIE is that, there may exist a number of inliers sharing the same pseudo-class, DIE directly sets these samples as mutual positive samples. For the inliers ( $c_i \geq 0$ ), the  $c_i$  is encoded to  $y_i^{cl}$  using DIE as follows,

$$\text{Inliers: } y_i^{cl}[j] = \begin{cases} 1 & c_j = c_i \\ -1 & c_j \neq c_i; \end{cases} \quad i, j = 1, \dots, n. \quad (3.11)$$

If a sample  $x_j$  has same pseudo-class with  $x_i$  ( $c_j = c_i$ ), the  $x_j$  is a positive sample of  $x_i$ , therefore  $y_i^{cl}[j]$  set to 1; Otherwise,  $y_i^{cl}[j] = -1$ . For the outliers ( $c_i = -1$ ), DIE encodes  $c_i$  as follows,

$$\text{Outliers: } y_i^{cl}[j] = \begin{cases} 1 & j = i \\ -1 & j \neq i; \end{cases} \quad i, j = 1, \dots, n. \quad (3.12)$$

where each outlier can be trained as an individual class. This operation is repeated until all samples in  $\mathcal{X}$  are enumerated.

Finally, we obtain  $n$  numbers of  $n$ -dimensional clustering-based pseudo label  $y_i^{cl}$  as illustrated in Figure 3.8. Our proposed DIE ensures equivalency between  $c_i$  and  $y_i^{cl}$ , and the value in  $y_i^{cl}$  points to the index of the samples that have the same pseudo-class with  $x_i$  in the meantime.

**SM-LP:** SM-LP methods (Zhong et al., 2019; Lin et al., 2020; Wang and Zhang, 2020) predicted positive labels by measuring the similarity among samples. Given the feature memory  $\mathcal{M}$ , the similarity of image  $x_i$ , notated as  $s_i$ , can be computed as:

$$s_i = \mathcal{M}[i] \times \mathcal{M}^\top \quad (3.13)$$

where  $s_i$  is an  $n$ -dimensional vector.  $s_i[j]$  represents the similarity scores between  $x_i$  and the image  $x_{\{j|j=1,\dots,n\}} \in \mathcal{X}$ .

Existing positive sample selection strategies (Zhong et al., 2019; Lin et al., 2020; Wang and Zhang, 2020) selected positive samples for  $x_i$  based on its similarity  $s_i$  using some fixed rules. We use the latest and best positive sample selection methods MPLP (Wang and Zhang, 2020). The MPLP (Wang and Zhang, 2020) used a pre-defined similarity threshold  $t = 0.6$  and the cycle consistency to select positive neighbors for  $x_i$ . Finally, the similarity measurement-based pseudo label  $y_i^{sm}$  for  $x_i$  can be generated as:

$$y_i^{sm}[j] = \begin{cases} 1 & \text{if } x_j \text{ is a positive neighbor} \\ -1 & \text{Otherwise.} \end{cases} \quad (3.14)$$

where  $y_i^{sm}$  is an  $n$ -dimensional vector. SM-LP assigns distinct pseudo labels  $y_i^{sm}$  to every sample.

### Rectified Binary Cross Entropy (ReBCE) Loss

To simplify the expression, we use asterisk symbol “\*” to represent clustering-based information and similarity measurement-based information. For example,  $L^*$  can represent the loss of  $y^{cl}$  or  $y^{sm}$ , and  $y^*$  can represent  $y^{cl}$  or  $y^{sm}$ .

In supervised learning, the Binary Cross Entropy (BCE) loss with ground-truth labels has been well studied in previous researches (Zhang and Zhou, 2014; Durand, Mehrasa, and Mori, 2019). Inspired by (Feng et al., 2020), we discover that BCE loss

poses a great challenge in unsupervised learning because of extensive noisy pseudo-labels. The BCE loss of classifying image  $x_i$  to its positive sample  $x_j$  can be computed as Equation 3.15. The gradient of  $L_{bce}^*$  are represented as Equation 3.16,

$$L_{bce}^* = -y_i^*[j] \times \log(s_i[j]) \quad (3.15)$$

$$\frac{\partial L_{bce}^*}{\partial \theta} = -y_i^*[j] \times \frac{1}{s_i[j]} \times \partial_{\theta} s_i[j] \quad (3.16)$$

where  $\theta$  means current network parameters. From Equation 3.15 and Equation 3.16, we can see a factor in BCE loss that samples with smaller similarity  $s_i[j] \rightarrow 0$  are weighted more than higher similarity for gradient update. In supervised learning, this factor helps the model paying more attention to difficult samples. However, in unsupervised learning, small similarity samples may contain many false-positive noisy pseudo labels. Therefore, using BCE loss might cause the model pay more attention to noises, thereby leading the model to fail.

We hence propose a Rectified Binary Cross Entropy (ReBCE) loss to address the above issue. The ReBCE loss is formulated as,

$$L_{Rebce}^* = \begin{cases} -y^*[j] \times \log(\alpha) & \text{if } s_j[j] < \alpha \\ -y^*[j] \times \log(s_i[j]) & \text{if } s_j[j] \geq \alpha. \end{cases} \quad (3.17)$$

where  $\alpha \in [0, 1]$  is a pre-defined rectified parameter to control the small similarity score amplify gradient excessively by rectifying very small  $s_i[j]$  to  $\alpha$ .

### Overall Loss

The effectiveness of Memory-based Multi-label Classification Loss (MMCL) in unsupervised multi-label person re-ID task is demonstrated in previous research (Wang and Zhang, 2020). Therefore, the network is simultaneously optimized with respect to the MMCL  $L_{mmcl}^*$  and the proposed ReBCE loss  $L_{Rebce}^*$  to achieve optimal model performances. The overall loss  $L$  can be computed by combining Equation ?? and Equation ?? as follows,

$$L_{mmcl}^* = \|s_i[j] - y_i^{sm}[j]\|^2 \quad (3.18)$$

$$L = \frac{1}{\eta} (L_{mmcl}^{cl} + L_{Rebce}^{cl} + L_{mmcl}^{sm} + L_{Rebce}^{sm}) \quad (3.19)$$



TABLE 3.9: Ablation study on outliers. “ $\times$ ”: Training without outliers.  
“ $\checkmark$ ”: Training each outlier as an individual class as Eq(2).

Outliers	Methods	Market-1501				DukeMTMC-reID			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
$\times$	C-LP	63.9	79.7	84.8	35.4	57.1	69.8	73.5	31.9
	<b>Joint-LP (ours)</b>	<b>75.7</b>	<b>87.9</b>	<b>90.9</b>	<b>48.1</b>	<b>64.5</b>	<b>75.5</b>	<b>79.6</b>	<b>40.3</b>
$\checkmark$	C-LP	72.4	86.0	89.9	44.7	63.2	75.0	79.0	40.6
	<b>Joint-LP (ours)</b>	<b>78.4</b>	<b>88.7</b>	<b>91.7</b>	<b>51.3</b>	<b>66.2</b>	<b>77.6</b>	<b>81.1</b>	<b>42.8</b>

TABLE 3.10: Comparison with different label prediction methods

Methods	Market-1501				DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
C-LP	72.4	86.0	89.9	44.7	63.2	75.0	79.0	40.6
SM-LP	77.4	87.3	90.3	41.8	63.6	75.0	79.4	39.0
<b>Joint-LP (ours)</b>	<b>78.4</b>	<b>88.7</b>	<b>91.7</b>	<b>51.3</b>	<b>66.2</b>	<b>77.6</b>	<b>81.1</b>	<b>42.8</b>

where  $\eta = 4$  is a normalized coefficient to normalize the scale of the overall loss.

### 3.4.3 Experiment Setting

We perform experiments on the two person re-ID datasets, Market-1501 (Market) and DukeMTMC-reID (Duke). The details are mentioned in Section 3.1.1. Two evaluation metrics are mentioned in Section 3.1.3.

The experiments are performed using one NVIDIA 1080Ti GPU with 11 GB of memory. The ResNet-50 (He et al., 2016) are adopted as the backbone network, which is pre-trained on ImageNet. The setting of backbone network follows the same setting in (Zhong et al., 2019; Lin et al., 2020; Wang and Zhang, 2020). The input images are resized to  $256 \times 128$ . The training batch size is 64. The total training epoch is 40. The initial learning rate is 0.03, and it is divided by 10 after 30 epochs. We set the rectified parameter  $\alpha = 0.2$  in ReBCE loss to achieve the best performance.

#### Ablation Study

**Importance of Outliers:** As shown in Table 3.9, directly discarding outliers from training data cannot achieve satisfying results on both datasets. There are two reasons. (1) discarding outliers leads to a poor initial model because there are many outliers during the whole training process, especially in early epochs. (2) discarding outliers inhibits the model learning on difficult samples. Therefore, we train

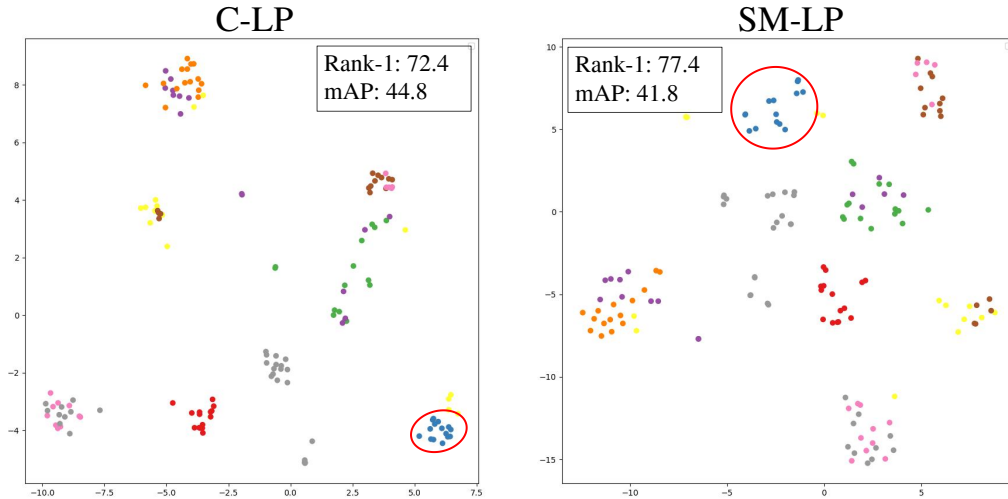


FIGURE 3.9: The t-SNE (Maaten and Hinton, 2008) visualization on features representation of 10 identities. The different color points are denoted identities.

TABLE 3.11: Comparison with different loss function

Methods (Loss function)	Market-1501				DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
MMCL	78.4	88.7	91.7	51.3	66.2	77.6	81.1	42.8
BCE	1.0	3.9	6.0	0.5	0.0	0.3	1.1	0.1
BCE + MMCL	0.9	2.7	4.7	0.3	0.6	1.6	2.4	0.3
<b>ReBCE (ours)</b>	78.8	89.5	92.8	50.6	66.1	77.8	81.5	42.3
<b>ReBCE + MMCL (ours)</b>	<b>80.3</b>	<b>90.8</b>	<b>93.2</b>	<b>55.1</b>	<b>67.8</b>	<b>78.4</b>	<b>81.6</b>	<b>44.0</b>

each outlier as an individual class, as mentioned in Equation (2). The results verify the effectiveness of our proposed DIE, which treats each un-clustered outlier as an individual class.

**Effectiveness of Joint-LP:** In order to verify the complementarities between C-LP and SM-LP, and the effectiveness of our proposed Joint-LP, we report comparison results of different label prediction methods in Table 3.10 and illustrate the t-SNE (Maaten and Hinton, 2008) visualization results in Figure 3.9.

From the comparison between C-LP and SM-LP, three observations are obtained. 1) In table 3.10, C-LP achieves better performance in mAP on two datasets. 2) SM-LP achieves better performance in Rank- $k$  accuracy on two datasets. 2) In Figure 3.9, C-LP generates closer and more compacter intra-class features than SM-LP. The reasons are that C-LP enforces centroid learning by assigning the same pseudo labels to the samples in the same cluster, therefore C-LP obtains higher clustering accuracy

TABLE 3.12: Performance comparison with other FUL person re-ID methods on Market-1501 and DukeMTMC-ReID. “LPM”: Label prediction methods. “\*”: The MetaCam algorithm in DSCE(Yang et al., 2021) is not considered in this table, because MetaCam requires camera IDs. This paper performs comparison experiments in unknown camera IDs environment.

LPM	Method	Reference	Market				Duke			
			R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
C-LP	BUC	AAAI19	66.2	79.6	84.5	38.3	47.4	62.6	68.4	27.5
	DBC	BMVC19	69.2	83.0	87.8	41.3	51.5	64.6	70.1	30.3
	DSCE*	CVPR21	74.9	-	-	53.9	62.8	-	-	43.4
SM-LP	SSL	CVPR20	71.7	83.8	87.4	37.8	52.5	63.5	68.9	28.6
	MMCL	CVPR20	80.3	89.4	92.3	45.5	65.2	75.9	80.0	40.2
	NNCT	ICIP21	<b>82.0</b>	90.0	92.9	48.4	64.8	75.7	79.2	40.7
Combined	<b>Joint-LP + ReBCE</b>		80.3	<b>90.8</b>	<b>93.2</b>	<b>55.1</b>	<b>67.8</b>	<b>78.4</b>	<b>81.6</b>	<b>44.0</b>

(in mAP) than SM-LP. Conversely, SM-LP enforces neighborhood learning by mining reliable positive samples around the sample, therefore SM-LP achieves higher ranking accuracy (in  $R-k$ ) than C-LP. These results demonstrate that C-LP and SM-LP lead model to learn in different directions, and thus they can be complementary to each other in  $R-k$  accuracy and in mAP to achieve better performance. It is an important discovery for the current and future object re-ID research.

Based on the above discovery, we propose the Joint-LP in here. The proposed Joint-LP achieves the best performance by enforcing centroid-towards and neighborhood-towards learning collaboratively. It is also interesting to observe that, with the help of SM-LP, the upper bounds of mAP are also increased from 44.7% to 51.3% on Market-1501, and from 40.6% to 42.8% on DukeMTMC-reID. Same improvements are also observed that, with the help of C-LP, the upper bounds of ranking accuracy 77.4% and 63.6% are also increased on two datasets, respectively. The improvements further demonstrate the complementarity between C-LP and SM-LP, and the proposed Joint-LP can overcome their demerits and utilize their merits at the same time.

*Effectiveness of ReBCE Loss:* We bring out that the traditional BCE loss cannot be directly adopted in the unsupervised multi-label classification task because of extensive noisy pseudo-labels. To demonstrate the above conjecture, we report the experimental results of different loss functions in Table 3.11. Table 3.11 shows that using BCE loss (w/ or w/o MMCL) leads the experiments to fail on two datasets. The main reason is that BCE forces the model to pay more attention to noisy labels

which leads to serious overfitting on noisy labels. Table 3.11 shows that the proposed ReBCE can solve this problem well, and ReBCE with or without MMCL achieves satisfying performance. These results demonstrate the effectiveness of the proposed ReBCE in the unsupervised multi-label classification task.

### Comparison with Other FUL Methods

As shown in Table 3.12, we compare our method with other FUL person re-ID methods. Three C-LP based methods, BUC (Lin et al., 2019), DBC(Ding, Khan, and Tang, 2019), and DSCE (Yang et al., 2021) are reported. Three SM-LP based methods, SSL (Lin et al., 2020), MMCL (Wang and Zhang, 2020), and NNCT (Tang and Jo, 2021) are reported.

Compared to the SM-LP based method MMCL (Wang and Zhang, 2020), our method significantly outperforms it in mAP by 9.6% on Market-1501 and by 3.8% on DukeMTMC-reID because of the adding centroids-towards learning. The best SM-LP based method NNCT (Tang and Jo, 2021) achieved the best R-1 accuracy 82.0% in Market-1501. NNCT cannot achieve satisfying performance in mAP because it lacks the clustering information to enforce centroids-towards learning.

Compared to the best C-LP based method DSCE (Yang et al., 2021), our method significantly outperforms DSCE in R-1 accuracy by 5.4% on Market-1501 and by 5.0% on DukeMTMC-reID because our method measures similarities among samples. It is noteworthy that, these specific and consistent improvements again demonstrate the importance of combining C-LP and SM-LP for the current and future object re-ID research. Finally, our method achieves the best performance with Rank-1 = 80.3%, mAP = 55.1% on Market-1501, and Rank-1 = 67.8%, mAP = 44.0% on DukeMTMC-reID.

### 3.4.4 Conclusion

In here, we have presented a superior fully unsupervised person re-ID method. To the best of our knowledge, this letter is an original work that (1) investigates the relation and difference between different label prediction methods, C-LP and SM-LP, (2) demonstrates the failure reason of BCE loss in unsupervised learning is because

BCE loss leads the noisy labels are weighted more for gradient update. Finally, comparisons with recent FUL methods demonstrate the superiority of our method.

### 3.5 Unsupervised Person Re-Identification via Multiple Pseudo Labels Joint Training

To mitigate the effects of noisy labels, Therefore, we propose a Multiple pseudo Labels Joint Training (MLJT) method based on MLCReID (Wang and Zhang, 2020). MLCReID (Wang and Zhang, 2020) formulated unsupervised person Re-ID as a Multi-Label Classification task and achieve state-of-the-art performance. The framework of MLCReID is shown in Figure 3.10 (a). The MLCReID (Wang and Zhang, 2020) can be mainly divided into three stages: feature extraction by backbone network, pseudo labels prediction, and fine-tuning backbone network with computed loss. Unlike human-annotated true labels, these self-generated pseudo labels contain noises that hinder the model’s capability on extracting discriminative features.

The proposed person re-ID architecture is shown in Figure 3.10 (b). The MLJT can be mainly divided into four stages: feature extraction by backbone network, channel-wise matching for exploring channel-based self-similarity, multiple pseudo labels prediction, fine-tuning backbone network with computed joint losses.

Unlike predicting a single pseudo label for an input image in MLCReID, the proposed MLJT predicts multiple pseudo labels for the input image by mining potential similarities in different ways. The combination of multiple pseudo labels is more robust than a single pseudo label. According to invariance constraints among these predicted multiple pseudo labels, the MLJT optimizes the network via multiple pseudo labels in a joint manner. In general, there are three types of pseudo labels in MLJT, as shown in Figure 3.10 (b). The first one is the clustering-based pseudo label  $y_i^{cl}$ . The second one is adaptive similarity measurement-based pseudo label  $y_i^{sm}$ . The third one is pseudo sub-labels  $y_i^{\{g|g=0,\dots,G\}}$ . We will introduce these three pseudo labels one by one in detail. As a summary, the contributions is three-fold:

1. Unlike predicting a single pseudo label in the baseline method, this work proposes to predict multiple pseudo labels for an input image by mining multiple potential similarities among samples.

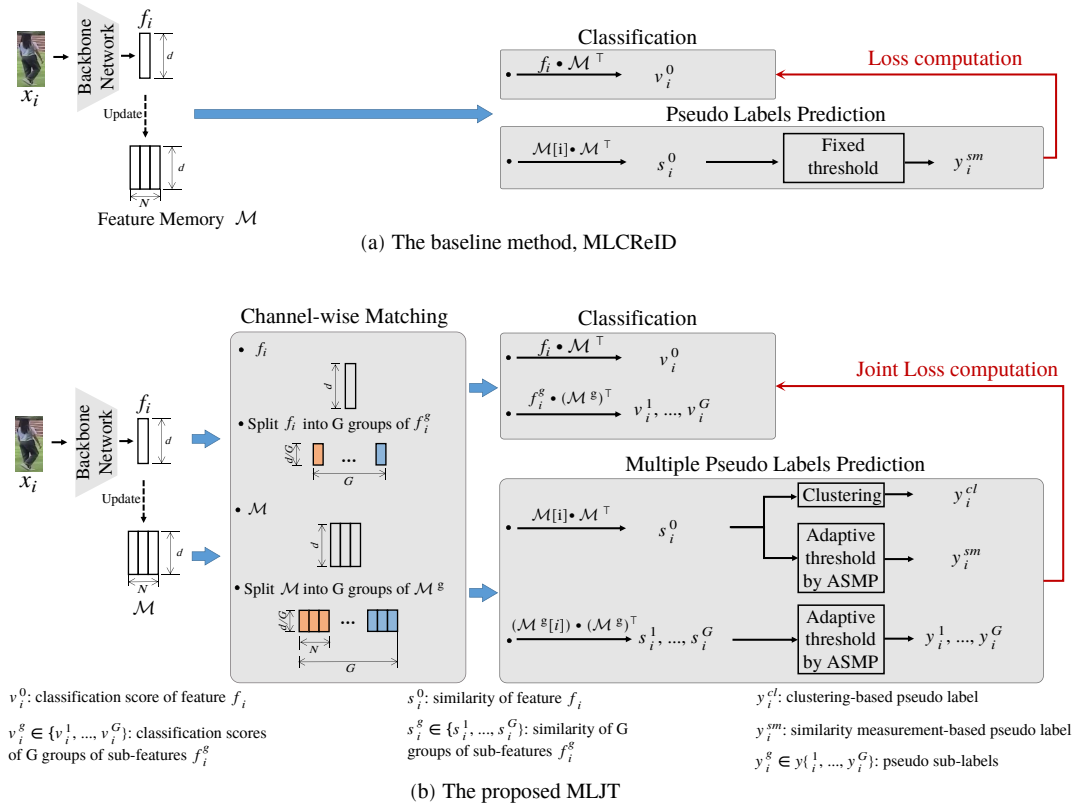


FIGURE 3.10: The framework of FUL person re-ID method. (a) Baseline method, MLCReID (Wang and Zhang, 2020). (b) The proposed MLJT.

2. Based on the invariance constraints among multiple pseudo labels, this work proposes to train the backbone network jointly to refine pseudo labels effectively.
3. In order to avoid neglecting positive labels of difficult samples, this paper proposes an Adaptive Similarity Measurement-based pseudo label Prediction (ASMP) method to adaptively select positive labels based on the degree of difficulty of samples.

This work improves the performance of baseline method (Wang and Zhang, 2020) by considerable margins on two mainstream and public datasets, Market-1501 (Zheng et al., 2015b) (Market) and DukeMTMC-reID (Zheng, Zheng, and Yang, 2017; Ristani et al., 2016) (Duke). Testing experiments are also performed in outdoor real-world videos to show the practicality of this work in real-world applications.

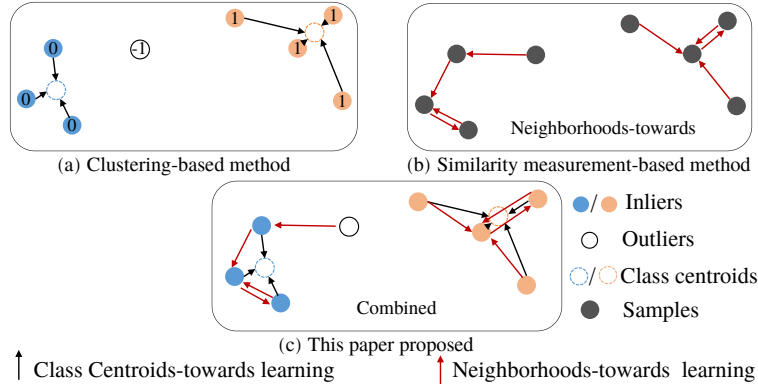


FIGURE 3.11: The illustration of different pseudo label prediction methods.

### 3.5.1 Proposed Method

#### Framework Overview

The framework of the proposed method MLJT is shown in Figure 3.10 (b), which can be mainly divided into five stages: feature extraction by backbone network, feature splitting using channel-wise matching, multiple pseudo labels prediction, classification, network optimization using joint losses.

Given an unlabeled person images  $x_{\{i=1,2,\dots,N\}} \in \mathcal{X}$ ,  $d$ -dimensional feature  $f_i$  is extracted by the backbone network.  $i$  is the index of the image in an unlabeled object re-ID dataset  $\mathcal{X}$ ,  $N$  indicates the number of images in  $\mathcal{X}$ . A feature memory  $\mathcal{M}$  is used to stores up-to-date features for all images in  $\mathcal{X}$  to make compute similarities among all images possible. The size of  $\mathcal{M}$  is  $N \times d$ .  $\mathcal{M}$  is updated by  $f_i$  after every training iteration as,

$$\mathcal{M}^t[i] = \alpha \mathcal{M}^{t-1}[i] + (1 - \alpha) f_i. \quad (3.20)$$

The superscript  $t$  in Equation 3.20 indicates  $t$ -th training iteration.  $\alpha$  is the updating rate,  $\alpha \in [0.0, 1.0]$ . Updating with the moving averaged weights makes  $\mathcal{M}$  more stable to predict pseudo labels (Zhong et al., 2019; Wang and Zhang, 2020). After updating, the feature of  $x_i$  in  $\mathcal{M}$  are notated as  $f_i^M$ , where  $\mathcal{M}[i] = f_i^M$ .

Then, multiple pseudo labels  $y_i = \{y_i^{cl}, y_i^{sm}, y_i^s\}$  of  $x_i$  can be predicted.  $y_i^{cl}$  and  $y_i^{sm}$  are the clustering-based pseudo label and the adaptive similarity measurement-based pseudo label predicted using  $\mathcal{M}$ , respectively.  $y_i^s$  are sub-labels predicted by mining channel-based self-similarities using sub-feature memories  $\mathcal{M}^s$ , which are obtained by channel-wise matching, as shown in Figure 3.10 (b).

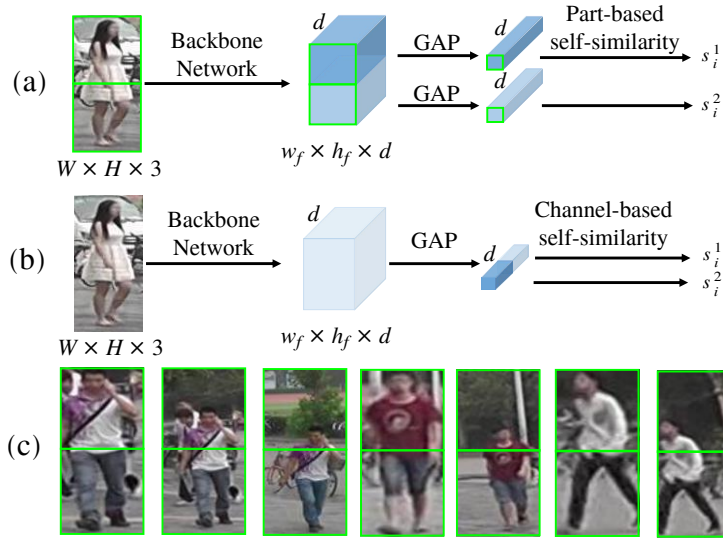


FIGURE 3.12: The illustrations of self-similarities exploration strategies in unsupervised re-ID. (a) Part-based self-similarities used in SSG (Fu et al., 2019b) and SSL (Lin et al., 2020). (b) Our proposed Channel-based self-similarities. (c) The examples of human location variance in Market-1501. **GAP**: Global Average Pooling.  $P = 2$  and  $G = 2$  are assumed in this figure.

The backbone network is optimized progressively by the loss between multiple pseudo labels  $y_i = \{y_i^{cl}, y_i^{sm}, y_i^s\}$  and their corresponding classification score  $v_i = \{v_i^0, v_i^s\}$  in a joint training manner. The classification scores  $v_i^0$  and  $v_i^s$  are computed using the whole feature  $f_i$  and the sub-features  $f_i^s$ , respectively.

### Multiple Pseudo Labels Prediction

1. **Clustering-based Pseudo Label  $y_i^{cl}$** : Following the previous works (Yang et al., 2021; Ester et al., 1996), the unsupervised clustering algorithm DBSCAN (Ester et al., 1996) is used in this paper. The clustering results is illustrated in Figure 3.11. Before each training epoch, DBSCAN assigns pseudo-classes  $c_i$  to all samples in  $\mathcal{X}$  by computing distances among their features  $f_{\{i|1, \dots, N\}}^M \in \mathcal{M}$ . DBSCAN is a density-based clustering algorithm, it treats high-confident samples as clustered inliers ( $c_i \geq 0$ ), and treats low-confident samples as unclustered outliers ( $c_i = -1$ ).

There is one challenge in the DBSCAN-based clustering method: the numbers of clustered inliers keep changing during the whole training process. Uncertain numbers of clusters make loss function is difficult to design. To address this issue, we



encode 1-dimensional pseudo-class  $c_i$  to  $N$ -dimensional clustering-based pseudo label  $y_i^{cl}$  as follows,

$$\text{Inliers: } y_i^{cl}[j] = \begin{cases} 1 & c_j = c_i \\ -1 & c_j \neq c_i; \end{cases} \quad j = 1, \dots, N. \quad (3.21)$$

After encoding, the multi-label classification loss function (Wang and Zhang, 2020) can be easily adopted to compute the loss between  $v_i^0$  and  $y_i^{cl}$  against the changing of clusters.

**Adaptive Similarity Measurement-based Pseudo Label** The cosine similarity of  $x_i$  and all images  $x_{\{j|j=1,\dots,N\}} \in \mathcal{X}$  is computed using their features in  $\mathcal{M}$  as follows,

$$s_i^0 = \mathcal{M}[i] \times \mathcal{M}^\top. \quad (3.22)$$

According to the  $s_i^0$ , the adaptive similarity measurement-based pseudo label  $y_i^{sm}$  is predicted by our proposed Adaptive Similarity Measurement-based pseudo label Prediction (ASMP).

As mentioned above, the previous pseudo label prediction methods selected positive labels (Zhong et al., 2019; Lin et al., 2020; Wang and Zhang, 2020) using fixed rules, i.e., a fixed  $k$  number of positive labels in ECN (Zhong et al., 2019) and SSL (Lin et al., 2020), or fixed label selection threshold  $t$  in MLCReID (Wang and Zhang, 2020). Using fixed rules (Zhong et al., 2019; Lin et al., 2020; Wang and Zhang, 2020) to select positive labels is unsuitable, because the similarity distribution of every image is different and keeps changing during the whole training process. More specifically, difficult samples always share relatively low similarity with its  $k$ -Nearest Neighbors ( $k$ -NN) because of challenging situations, such as complex or uncommon human poses, occlusion, and complex backgrounds, etc. The examples of similarity distributions of image A (simple case) and image B (difficult case) are illustrated in Figure 3.13. The similarities among image B and its  $k$ -NN are much lower than the similarities among image A and its  $k$ -NN, as shown in the right graph of Figure 3.13. If positive samples are selected using the fixed threshold  $t = 0.6$  for all images as (Wang and Zhang, 2020), the potential positive labels of the difficult sample with similarity less than 0.6 might be neglected continuously.

To avoid neglect the positive label of the difficult sample, we proposed the ASMP,

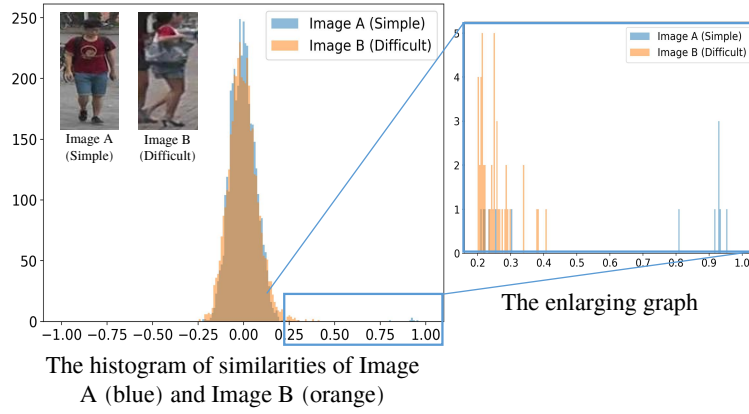


FIGURE 3.13: The histogram of similarity distribution. The horizontal axis is similarity score. The right graph is zoomed in from the blue rectangular area of the left graph.

which first distinguish the difficult sample based on the degree of difficulty  $\theta_i$  of the sample, then assigns a lower and adaptive threshold to the difficult sample for selecting more positive labels. Our idea is that, when the model can roughly capture the data distribution and predict similarity, we can therefore utilize the obtained similarity distribution to estimate the degree of difficulty of the image.

ASMP first computes the similarities among  $x_i$  and its  $k$ -NN as follows,

$$\begin{cases} R_i = \text{argsort}(s_i^0) \\ K_i = R_i[1 : k] \end{cases} \quad (3.23)$$

where  $R_i$  is the rank list by sorting the similarity  $s_i^0$  by descending order.  $K_i$  is the collection of  $k$ -NN of  $x_i$ .  $k$  is a hyper-parameter, controls the numbers of samples in  $K_i$ . The analysis of  $k$  are reported in the Sec. V.

The degree of difficulty  $\theta_i$  is estimated according to statistical characteristics of  $K_i$ , i.e., the mean  $m_i$  and standard deviation  $\sigma_i$ , as follows,

$$\begin{cases} m_i = \frac{\sum_{s_i^0[j] \in K_i} (s_i^0[j])}{k} \\ \sigma_i = \sqrt{\frac{\sum_{s_i^0[j] \in K_i} (s_i^0[j] - m_i)^2}{k}} \end{cases} \quad (3.24)$$

$$\theta_i = m_i + \sigma_i \quad (3.25)$$

If  $\theta_i$  is low,  $x_i$  is considered a difficult sample. The computed  $\theta_i$  will be directly used to select more positive labels for  $x_i$  as follows,

$$y_i^{sm}[j] = \begin{cases} 1, & \text{if } s_i^o[j] \geq \min(0.6, \theta_i) \\ -1, & \text{otherwise} \end{cases}, \quad j = 1, \dots, n. \quad (3.26)$$

In this case, more difficult samples can be assigned a lower  $\theta_i$  to mitigate the neglect to potential positive labels. Finally, adaptive similarity measurement-based label  $y_i^{sm}$  are generated. It is noteworthy that  $\theta_i$  for every sample is different according to statistical characteristics of the sample, therefore ASMP is better than fixed rules in (Zhong et al., 2019; Lin et al., 2020; Wang and Zhang, 2020).

**Pseudo Sub-labels by Channel-based Self-Similarity Exploration (CSS)** The third pseudo labels in this paper is pseudo sub-labels  $y_i^s$ , which is predicted by mining channel-based self-similarities.

The baseline method MLCReID (Wang and Zhang, 2020) only compared the similarity between two images by global features. To mine more similarity information existing in the unlabeled dataset, several methods (Fu et al., 2019b) split images into horizontal parts to represent human partial regions, as illustrated in Figure 3.12 (a). Then, unlabeled images are additionally compared using these partial features. The precondition of part-based self-similarity computation is that the human region should always locate in the center of the image. Uncertain human location may cause inconsistency in similarity mining and matching.

The critical idea of self-similarity computation is to compute a more robust similarity score by additionally comparing partial features. In order to avoid the impact of human location variance and mine the potential similarity as well, we propose to explore channel-based self-similarity in this paper, as shown in Figure 3.12 (b).

To formulate the proposed channel-based self-similarity computation, a channel-wise matching module is proposed in this paper. The channel-wise matching module is attached after feature  $f_i \in \mathbb{R}^d$  and feature memory  $\mathcal{M} \in \mathbb{R}^{N \times d}$  to split them into  $G$  groups of sub-features  $f^g$  and sub-feature memories  $\mathcal{M}^g$ , respectively. As illustrated in Figure 3.10 (b), each sub-feature  $f_i^{\{g|g=1, \dots, G\}} \in \mathbb{R}^{\frac{d}{G}}$ , and each sub-feature memory  $\mathcal{M}^{\{g|g=1, \dots, G\}} \in \mathbb{R}^{N \times \frac{d}{G}}$ . A channel shuffle operation is used before splitting. The  $f_i$  and  $\mathcal{M}$  share the sample shuffle index in every training iteration to maintain

the comparison consistency.

The classification scores  $v_i^{\{g|g=1,\dots,G\}}$  are computed using their corresponding sub-features  $f_i^{\{g|g=1,\dots,G\}}$  as follows,

$$v_i^g = f_i^g \times (\mathcal{M}^g)^\top, \quad g = 1, \dots, G. \quad (3.27)$$

The  $G$  groups of self-similarities of  $x_i$  are computed using their corresponding sub-feature memories  $\mathcal{M}^g$  as follows,

$$s_i^g = \mathcal{M}^g[i] \times (\mathcal{M}^g)^\top, \quad g = 1, \dots, G. \quad (3.28)$$

Based on  $s_i^g$ , the corresponding pseudo sub-labels  $y_i^g$  can be predicted using the proposed ASMP as Equation 3.23 - Equation 3.26 as follows,

$$y_i^g = \text{ASMP}(s_i^g, k), \quad g = 1, \dots, G. \quad (3.29)$$

### 3.5.2 The joint Loss function

The Memory-based Multi-label Classification Loss (MMCL) (Wang and Zhang, 2020)  $\mathcal{L}^*$  is used to regress classification scores  $v_i$  to predicted pseudo label  $y_i$  as follows:

$$\mathcal{L}^*(v_i, y_i) = \frac{\delta}{|P_i|} \sum_{v_i[j] \in P_i} (v_i[j] - y_i[j]) + \frac{1}{|N_i|} \sum_{v_i[k] \in N_i} (v_i[k] - y_i[k]) \quad (3.30)$$

$P_i$  indicates the positive samples of  $x_i$ , of where  $y_i[j] = 1$ .  $N_i$  indicates the hard negative samples of  $x_i$ , of where  $y_i[k] = -1$ . To solve the sparsity of  $y_i$ , only top 1% negative samples are chosen as hard negative samples to compute  $\mathcal{L}^*$ .  $\delta$  is the balance parameter between positive hard negative sample loss. Following same setting in MLCReID (Wang and Zhang, 2020),  $\delta = 5$  in our paper.

As a summary, for an unlabeled image  $x_i$ , three types of pseudo labels  $y_i = \{y_i^{cl}, y_i^{sm}, y_i^g\}$  can be predicted by Equation 3.21, Equation 3.26, and Equation 3.29. The MLJT is trained using the losses between multiple pseudo labels  $y_i = \{y_i^{cl}, y_i^{sm}, y_i^g\}$  and their corresponding classification scores  $v_i = \{v_i^0, v_i^g\}$  in an end-to-end and joint

manner. The overall joint loss function is formulated as follows,

$$\mathcal{L} = \mathcal{L}^*(v_i^0, y_i^{cl}) + \mathcal{L}^*(v_i^0, y_i^{sm}) + \frac{\sum_{g=1}^G \mathcal{L}^*(v_i^g, y_i^g)}{G}. \quad (3.31)$$

### 3.5.3 Experiment Settings

#### Datasets and Evaluation Metrics

We evaluate the proposed method on Market-1501 (Zheng et al., 2015b) (Market) and DukeMTMC-reID (Zheng, Zheng, and Yang, 2017; Ristani et al., 2016) (Duke). The details are mentioned in Section 3.1.1. Two evaluation metrics are mentioned in Section 3.1.3

#### Implementation Details

Following the previous researches (Zhong et al., 2019; Wang and Zhang, 2020; Tang and Jo, 2021; Ji et al., 2020; Lin et al., 2020; Yang et al., 2021), we use an ImageNet (Deng et al., 2009) pre-trained ResNet-50 (He et al., 2016), provided by Pytorch official, as the backbone network to conduct fair comparisons. A  $1 \times 1$  CNN layer and a batch normalization layer are added after the last global pooling layer of ResNet-50 to generate 2048-dimensional L2-normalized features. The input images are resized to  $256 \times 128 \times 3$ . The training batch size is 64. The network is trained by the Stochastic Gradient Descent (SGD) with a learning rate of 0.03, 40 epochs in total. The CamStyle (Tahir, 2019) is used as a data augmentation strategy. To achieve the best performance, hyper-parameter  $k = 80$ . The experiments are performed using an Intel Core i5-6600 3.30-GHz CPU and one NVIDIA GeForce Titan 1080Ti GPU with 11 GB of memory. The total training time is around 4 hours on Market-1501 and DukeMTMC-reID. The test architecture is the same as the baseline method (Wang and Zhang, 2020), therefore our method does not increase any computation during testing.

### 3.5.4 Ablation study of the proposed components

To demonstrate the effectiveness of the proposed multiple pseudo labels and joint training strategy, extensive ablation studies are reported in this section. Ablation

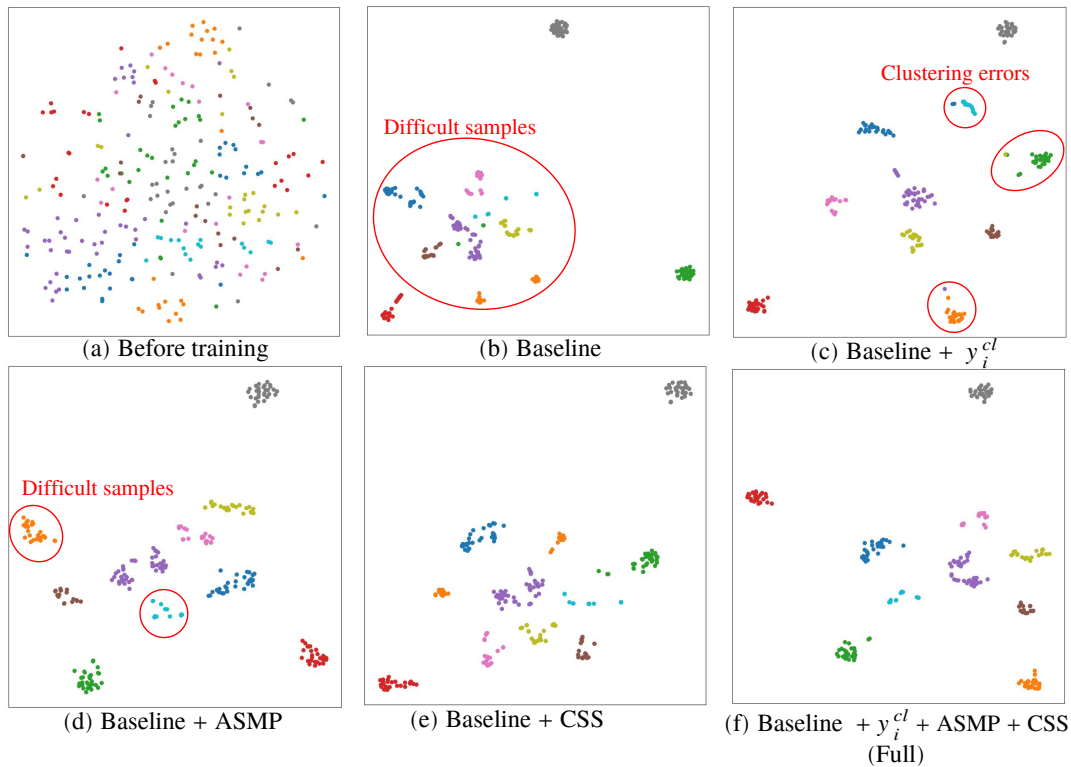


FIGURE 3.14: The t-SNE (Maaten and Hinton, 2008) plot of 10 identities. Different colors denotes different identities.

studies are presented in four aspects: 1) The effectiveness of the clustering-based pseudo label  $y_i^{cl}$ , 2) The analysis of ASMP, 3) the analysis of CSS, and 4) the analysis of multiple pseudo labels joint training strategy.

We summarize the performance of each proposed component in Table 3.13, and visualize the t-SNE (Maaten and Hinton, 2008) features in Figure 3.14. Features in Figure 3.14 (a) only contain weak clustering information before training on a specific unlabeled person re-ID dataset. This is because the ImageNet (Deng et al., 2009) pre-trained model was trained to distinguish humans and other objects. Therefore, the model still can not distinguish human identities with their appearances without training on specific person re-ID dataset. As shown in Figure 3.14 (b)-(f), features of the same identity are progressively gathered after training with predicted pseudo labels. It indicates the predicted pseudo labels can train the model on the unlabeled dataset to distinguish human identities. More accurate pseudo labels help the model learn to extract more discriminative features to re-identify persons.

TABLE 3.13: Ablation study on individual proposed modules in the proposed MLJT.

No.	Proposed Modules			Market-1501				DukeMTMC-reID			
	Clustering	ASMP	CSS	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
1				43.0	78.9	79.0	91.7	37.4	63.5	67.5	71.8
2	✓			50.9	78.1	88.5	91.8	40.3	64.5	75.5	79.6
3		✓		49.1	81.6	90.0	92.3	40.8	65.8	76.9	80.6
4			✓	46.4	80.4	89.6	92.7	40.2	65.4	76.2	80.3
5	✓	✓		51.4	79.7	89.3	92.4	41.3	64.7	76.1	79.3
6	✓		✓	52.3	81.2	<b>91.3</b>	<b>94.2</b>	41.7	65.4	76.6	80.5
7		✓	✓	50.1	80.8	89.6	92.8	41.4	66.3	76.3	80.0
8 (Full)	✓	✓	✓	<b>55.3</b>	<b>81.6</b>	90.4	93.5	<b>42.9</b>	<b>66.3</b>	<b>77.4</b>	<b>80.7</b>

### Performance of the baseline method

The *No. 1* in Table 3.13 and Figure 3.14 (b) show the results of the baseline method, which does not use clustering-based pseudo labels  $y_i^{cl}$ , adaptive threshold  $\theta_i = 0.6$  in Equation 3.26, and sub-labels  $y_i^s$  by channel-based self-similarity exploration. In Table 3.13, the baseline method *No. 1* produces unsatisfactory performance on two datasets. For instance, it achieves 43.0% in mAP and 78.9% in R-1 on Market-1501, and 37.4% in mAP and 63.5% in R-1 on DukeMTMC-reID. It shows that using a single pseudo label is not enough to provide robust information. As shown in Figure 3.14 (b), the model still can not produce discriminative features to distinguish difficult samples.

To improve the performance of the baseline method, this paper proposes three types of pseudo labels. We further perform experiments to investigate the effectiveness of them in detail by adding them into the baseline model. The results are reported in Table 3.13 and Figure 3.14.

### The effectiveness of the clustering-based pseudo label

The effectiveness of the clustering-based pseudo label  $y_i^{cl}$  is presented in here. From the comparison of *No. 1* and *No. 2* in Table 3.13, it is clear that the model performance is improved with the clustering-based pseudo label  $y_i^{cl}$ . Specifically, the performance improves in mAP from 43.0% to 50.9% and from 37.4% to 40.3% on Market-1501 and DukeMTMC-reID, respectively. The significant improvement is because samples are additionally enforced learning towards their corresponding class centroids under the supervision of clustering labels  $y_i^{cl}$ , as illustrated in Figure 3.11 (c). Similar

TABLE 3.14: Analysis of hyper-parameters  $k$  on Market-1501.

$k$	20	40	60	80	100	120	140	160
mAP	44.8	48.1	48.5	49.1	<b>49.3</b>	48.3	46.9	46.8
R-1	80.0	<b>82.0</b>	81.3	81.6	80.7	80.1	79.6	78.6

observation also can be found from the comparison of Figure 3.14 (b) and (c). Features of the same identity are more compact and independent from other clusters by using  $y_i^{cl}$ . The improvement demonstrates the effectiveness of combining similarity measurement-based and clustering-based label prediction in this paper.

### The analysis of Adaptive Similarity Measurement-based pseudo label Prediction (ASMP)

The ASMP is analyzed in three aspects. First, the effectiveness of ASMP is evaluated in Table 3.13 and Figure 3.14. Second, the robustness of hyper-parameter  $k$  in Equation 3.23 is analyzed in Table 3.14. Third, we compare our proposed ASMP with other label prediction methods in Table 3.15.

*The effectiveness of ASMP:* ASMP is proposed to adaptively select positive labels for a sample according to the similarity distribution between the sample and its neighbors. If a sample shares low similarities with neighbors, ASMP will consider the sample as the difficult sample and compute a low threshold to choose more positive labels for it. Comparison results between the model trained with or without ASMP are illustrated in Table 3.13 and Figure 3.14. In Table 2, No. 3 surpasses the No. 1 by 6.1% in mAP and 2.7% in R-1 accuracy on Market-1501, and by 3.4% in mAP and 2.3% in R-1 accuracy on DukeMTMC-reID. Similarly, when we compare Figure 3.14 (b) and (d), we observe that the model with ASMP can gather features of the same identity and enlarges the distances among different identities. It is because that using a fixed and high threshold in the baseline method makes dispersed samples (e.g. blue and orange samples in Figure 3.14 (b) lack positive supervisory signals for training, thereby leading them to be neglected. The results demonstrate the necessity of setting a lower threshold to choose more positive labels for difficult samples in unsupervised person re-ID tasks and the effectiveness of our proposed ASMP.



TABLE 3.15: Comparison with different pseudo label prediction methods.

Method	Market		Duke	
	mAP	R-1	mAP	R-1
Fixed numbers $k$ (Zhong et al., 2019; Lin et al., 2020)	36.7	72.4	36.2	62.2
Fixed threshold $t$ (Wang and Zhang, 2020)	43.0	78.9	37.4	63.5
Adaptive threshold $\theta_i$ (our ASMP)	<b>49.1</b>	<b>81.6</b>	<b>40.8</b>	<b>65.8</b>

*Comparison of different  $k$  in ASMP:* The hyper-parameter  $k$  controls the numbers of near neighbors are chosen to build  $K_i$  to estimate the degree of difficulty of the sample in Equation 3.23. We validate the influence of different  $k$  in Table 3.14 by vary  $k$  from 20 to 160. Compared with baseline results in Table 3.13, it is clear that using any  $k$  enhances model performance consistently, which demonstrates the necessity of adaptive thresholds in ASMP. When  $k = 80$ , the optimal performance can be obtained. Too small  $k$  decreases the performance because the limited number of near neighbors is not enough to represent the statistical characteristics of similarity distribution of the sample for estimating its degree of difficulty  $\theta_i$ . Also, too large  $k$  causes similar statistical characteristics of simple and difficult images, thereby distinguishing them difficultly.

*Comparison with different label prediction methods:* We further compare the proposed ASMP with other label prediction methods, i.e., a fixed  $k$  number of positive labels in ECN (Zhong et al., 2019) and SSL (Lin et al., 2020), and a fixed label selection threshold  $t$  in MLCReID (Wang and Zhang, 2020) in Table 3.15. Table shows that using the fixed threshold  $t$  ( $t=0.6$ , which achieves the best performance) outperforms using the fixed numbers  $k$  ( $k=10$ , which achieves the best performance) by large margins. Our proposed adaptive threshold  $\theta_i$  in ASMP achieves the best performance. The superior performance demonstrates that, compared with fix rules in existing works, our proposed adaptive threshold is a more reasonable and effective method for pseudo label prediction.

### The analysis of channel-based self-similarity (CSS)

The CSS is proposed to predict pseudo sub-labels  $y_i^s$  by exploring a more robust and precise similarity relation among images. The CSS is analyzed in three aspects here. First, the effectiveness of CSS is evaluated in Table 3.13 and Figure ???. Second, different splitting methods and different numbers of group  $G$  in channel-wise matching

TABLE 3.16: Comparison of different splitting methods and number of splitting groups in CSS on Market-1501.

Splitting Method	Groups	mAP	Rank-1	Rank-5	Rank-10
In order	G=2	41.3	75.3	87.4	90.8
	G=4	44.8	78.1	88.9	92.2
	G=8	46.0	80.2	89.5	92.3
Shuffle	G=2	40.8	75.5	87.1	90.7
	G=4	45.7	78.6	89.3	92.4
	G=8	<b>46.4</b>	<b>80.4</b>	<b>89.6</b>	<b>92.7</b>

TABLE 3.17: Comparison with different similarity exploration methods.

Methods	Market		DukeMTMC	
	mAP	R-1	mAP	R-1
w/o self-similarity	43.0	78.9	37.4	63.5
w/ Part-based self-similarity	41.9	74.6	35.1	61.3
w/ Channel-based self-similarity (ours)	<b>46.4</b>	<b>80.4</b>	<b>40.2</b>	<b>65.4</b>

are analyzed in Table 3.16. Third, we compare our proposed channel-based self-similarity with part-based self-similarity in existing works (Fu et al., 2019b; Lin et al., 2020) in Table 3.17.

*The effectiveness of CSS:* As reported in Table 3.13, No. 4 surpasses the baseline No. 1 by 3.4% in mAP and 1.5% in R-1 accuracy on Market-1501, and by 2.8% in mAP and 1.9% in R-1 accuracy on DukeMTMC-reID. As shown in Figure 3.14 (e), the model can distinguish different identities better than the baseline by mining similarities from global feature to channel-wise partial feature. The improvements and visualization results demonstrate that the proposed CSS is a simple and effective method for mining self-similarity in an unsupervised manner.

*Comparison of different splitting methods and different G:* In channel-wise matching, the feature  $f_i$  and the feature memory  $\mathcal{M}$  can be split into  $G$  groups sub-features  $f_i^g$  and sub-feature memories  $\mathcal{M}^g$  in order or shuffle, respectively. Table 3.16 reports the performance comparisons of these two splitting methods with different splitting groups  $G$ . The results show that no significant difference in performance between splitting in order and shuffle. It is clear that a large  $G$  consistently performs better than a small  $G$ . When we set  $G = 2$ , we observe that small  $G$  slows down the model convergence speed. Thus, the model is difficult to converge to achieve satisfactory performance even if we set the training epoch very high. Finally, we split the feature memory in a shuffling manner and use  $G = 8$ .

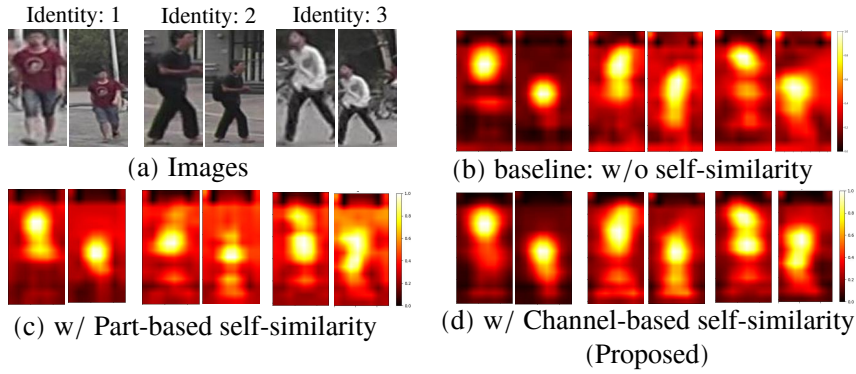


FIGURE 3.15: The visualization results of feature maps with different similarity exploration methods.

**Comparison between part-based and channel-based self-similarity:** We compare the proposed channel-based self-similarity with part-based self-similarity in (Fu et al., 2019b; Lin et al., 2020) in Table 3.17. For a straightforward comparison, we further visualize the hot maps of feature maps before the last GAP layer of the backbone network in Figure 3.15. Figure 3.15 (a) illustrates six different input images with three different identities. The original height and width of the feature maps are  $w_f \times h_f \times d : 8 \times 16 \times 2048$ , as mentioned in Figure 3.12. We resize them to the same size as input images  $W \times H : 128 \times 256$  to more straightforwardly comparisons. Brighter color means the model extracts more features from the regions.

Table 3.13 reports that using part-based self-similarity drops the model performance from the baseline 43.0% to 41.9% in mAP and 78.9% to 74.6% in R-1 accuracy on Market-1501. Similar, the performance declines are observed on DukeMTMC-reID. It shows that part-based self-similarity method is not robust. The same situations are observed by visualization examples in Figure 3.15 (c). Mining part-based self-similarity makes the model can not accurately capture features from the human region, especially if the identity does not locate in the center of the image.

The proposed channel-based self-similarity can help the model learn to extract more discriminative features, as compared in Figure 3.15. The superior performance of the proposed CSS are mainly reflected in four aspects. Firstly, Table 3.13 reports that using CSS enhance the model performance consistently on Market-1501 and DukeMTMC-reID. Secondly, compared with Figure 3.15 (b), features of the foreground (human region) are more accurate and brighter in Figure 3.15 (d). Thirdly, compared with Figure 3.15 (c), the background area is darker (lower importance) in Figure 3.15 (d). Fourthly, feature extraction capability of the model is not affected by

TABLE 3.18: Performance comparison with state-of-the-art FUL-based methods. The first and second best results are marked in **bold** and **blue**, respectively.

Method	Market-1501				DukeMTMC-reID			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
BoW (Zheng et al., 2015a)	14.8	35.8	52.4	60.3	8.3	17.1	28.8	34.9
UDML (Peng et al., 2016)	12.4	34.5	52.6	59.6	7.2	18.5	31.4	37.6
LOMO (Liao et al., 2015)	18.0	27.2	41.6	49.1	4.8	12.3	21.3	26.6
CAMEL (Yu, Wu, and Zheng, 2017)	26.3	54.5	-	-	-	-	-	-
DECAMEL (Yu, Wu, and Zheng, 2020)	32.4	60.2	76.0	81.1	-	-	-	-
SSL (Lin et al., 2020)	37.8	71.7	83.8	87.8	28.6	52.5	63.5	68.9
BUC (Lin et al., 2019)	38.3	66.2	79.6	84.5	27.5	47.4	64.6	68.4
ADTC (Ji et al., 2020)	38.8	59.5	71.6	76.9	37.9	59.4	70.0	74.1
DBC (Ding, Khan, and Tang, 2019)	41.3	69.2	83.0	87.8	30.3	51.5	64.6	70.1
MLCReID (Wang and Zhang, 2020)	45.5	80.3	89.4	92.3	40.2	65.2	75.9	80.0
NNCT (Tang and Jo, 2021)	48.4	<b>82.0</b>	<b>90.0</b>	<b>92.9</b>	40.7	<b>64.8</b>	<b>75.7</b>	<b>79.2</b>
DSCe (Yang et al., 2021)	<b>53.9</b>	74.8	-	-	<b>43.4</b>	62.8	-	-
MLJT (Ours)	<b>55.3</b>	<b>81.6</b>	<b>90.4</b>	<b>93.5</b>	<b>42.9</b>	<b>66.3</b>	<b>77.4</b>	<b>80.7</b>

human location. The visualization results demonstrate that the proposed channel-based self-similarity helps the model extract discriminative features and avoid the impact of human location variance as well.

### The analysis of joint training strategy

The *No. 5-No. 8* in Table 3.13 shows that combinations of each individual proposed module bring greater improvements. Finally, the full version of the proposed MLJT achieves the best performance 55.3% in mAP and 81.6% in R-1 accuracy on Market-1501, and 42.9% in mAP and 66.3% in R-1 accuracy on DukeMTMC-reID. The t-SNE features generated by the full version of the proposed MLJT are shown in Figure 3.14 (f). It shows that joint training strategy can overcome demerits and utilize merits of each individual module. Specifically, compared with Figure 3.14 (c), the clustering errors are eased in Figure 3.14 (f). Compared with Figure 3.14 (d) and (e), Figure 3.14 (f) shows that the model with joint training strategy produces more compact feature clusters. These verify that our proposed multiple pseudo labels joint training strategy is able to fully utilize each individual module for learning better and discriminative features.

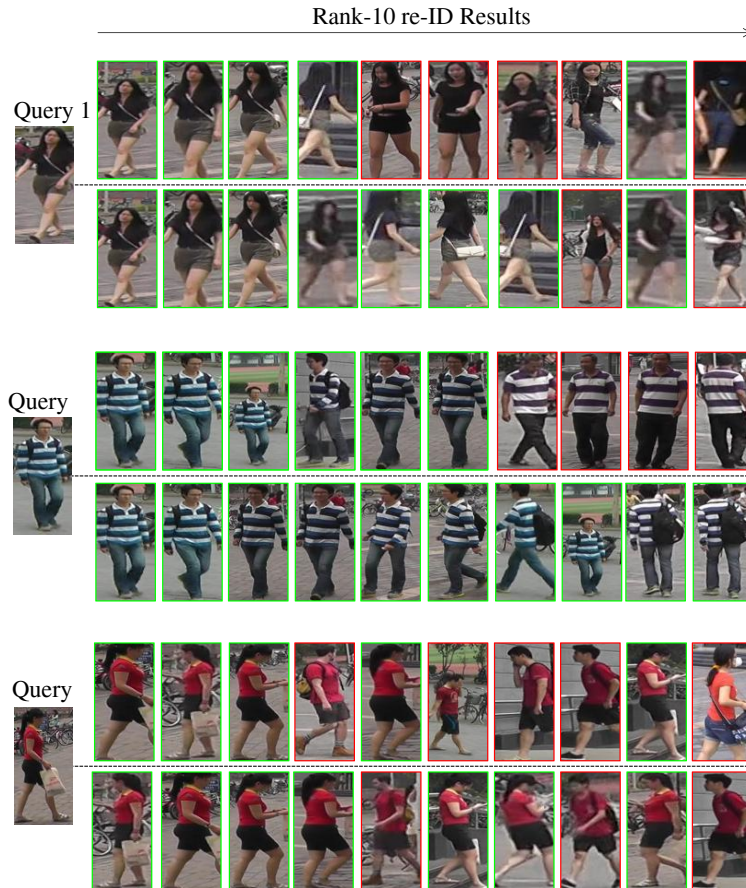


FIGURE 3.16: Examples of the top-10 person re-ID results. The first and second rows are generated by the baseline method (Wang and Zhang, 2020) and the proposed MLJT, respectively. The green boxes denote the true positive results, and the red boxes denote the false positive results.

### 3.5.5 Performance

#### Performance comparison in public datasets

We compare our proposed MLJT against state-of-the-art unsupervised person re-ID models in Market-1501 and DukeMTMC-reID in Table 3.18. We first compare our method MLJT with the hand-crafted feature-based methods, including BoW (Zheng et al., 2015a), UDML (Peng et al., 2016), and LOMO (Liao et al., 2015), which do not require CNN and any labeled dataset to extract features. The performances of the hand-crafted feature-based methods are not satisfactory enough, because it is difficult to manually design discriminative features with good generalization and robustness, especially on large-scale datasets.

We also compare the MLJT with CNN-based fully unsupervised learning-based





FIGURE 3.17: The comparison re-ID results of the baseline method (Wang and Zhang, 2020) and our proposed method MLJT on two outdoor real-world videos. Each column refers to the different frames of videos. The query image (re-ID target) is shown as a sub-figure at the bottom of each frame. The human regions are detected by YOLOv5 (al, Apr. 2021). The green bounding boxes are the re-ID results, and the classification scores are written on the boxes. Frames with a red border refer to false re-ID results.

methods, including CAMEL (Yu, Wu, and Zheng, 2017), DECAMEL (Yu, Wu, and Zheng, 2020), SSL (Lin et al., 2020), BUC (Lin et al., 2019), ADTC (Ji et al., 2020), DBC (Ding, Khan, and Tang, 2019), the baseline method MLCReID (Wang and Zhang, 2020), the best similarity measurement-based method NNCT (Tang and Jo, 2021), and the best clustering-based method DSCE (Yang et al., 2021). We summarize three observations. 1) The clustering-based methods BUC, DBC, ADTC and DSCE can achieve better clustering accuracy (in mAP) because they assigned the same labels to the samples in the same cluster. During training, intra-class samples are enforced learning towards their class centroids to force these samples to get more compact. 2) The similarity measurement-based methods SSL, MLCReID, and NNCT can achieve higher ranking accuracy (in R-k) because they assign labels by mining reliable positive neighbors around the sample. The similarity measurement-based methods enforce learning towards reliable neighbors. 3) our method MLJT outperforms other FUL-based person re-ID methods in mAP and R-k accuracy. The superior performance demonstrate the effectiveness of our proposed multiple pseudo labels joint training strategy for unsupervised person re-ID.

Moreover, we compare the results of the baseline method (Wang and Zhang, 2020) and the proposed method MLJT by showing their top-10 retrieved images of three query images in Figure 3.16. The green and red boundaries denote correct

and false re-ID results, respectively. As illustrated, for the same query images, the MLJT can retrieve correct images more accurately than the baseline. For example, the baseline is easily confused by the cloth with white and purple stripes. Overall, the comprehensive comparison results indicate the effectiveness and superiority of the MLJT.

### Performance in real-world application

Finally, we test the proposed re-ID method on two outdoor real-world videos to evaluate the model performance in a real-world application. The full testing demo can be found at [https://drive.google.com/drive/folders/1RvNaEiy6tF18\\_RcgTncjE7jJ6eGy8sZL?usp=sharing](https://drive.google.com/drive/folders/1RvNaEiy6tF18_RcgTncjE7jJ6eGy8sZL?usp=sharing).

In a real-world application, YOLOv5 (al, Apr. 2021) is adopted to detect human regions because person regions are a prerequisite for person re-ID. Then, the re-ID network retrieves the same person by matching a pre-defined query image with every detected person region in the frame.

The visualization of re-ID results of the baseline method (Wang and Zhang, 2020) and our proposed MLJT on two outdoor real-world videos are compared in Figure 3.17. The query image (re-ID target) is shown at the left bottom of each frame. From the comparison, we notice that MLJT retrieves correct persons more accurately than the baseline, and MLJT outputs more confident (higher) classification scores than the baseline in two videos.

The runtime of the person detection and re-identification system is shown at the left top of each frame, notated as Frame Per Second (FPS). The system can work in real-time with a processing speed of about 128 FPS with YOLOv5 (al, Apr. 2021) detector.

### 3.5.6 Conclusion

In this work, we proposed an end-to-end fully unsupervised person re-ID method, which can be trained without using any labeled information. The proposed method achieves superior performance benefit from three aspects. 1) Selecting positive labels adaptively according to similarity distribution of samples. 2) Estimating similarity

precisely by the channel-based self-similarities exploration strategies. 3) Optimizing network jointly using multiple pseudo labels to mitigate the impact of noises in a single pseudo label. Extensive experiments and comprehensive analysis demonstrate the effectiveness of the proposed method MLJT.

### 3.6 Unsupervised Object Re-identification via Instances Correlation Loss

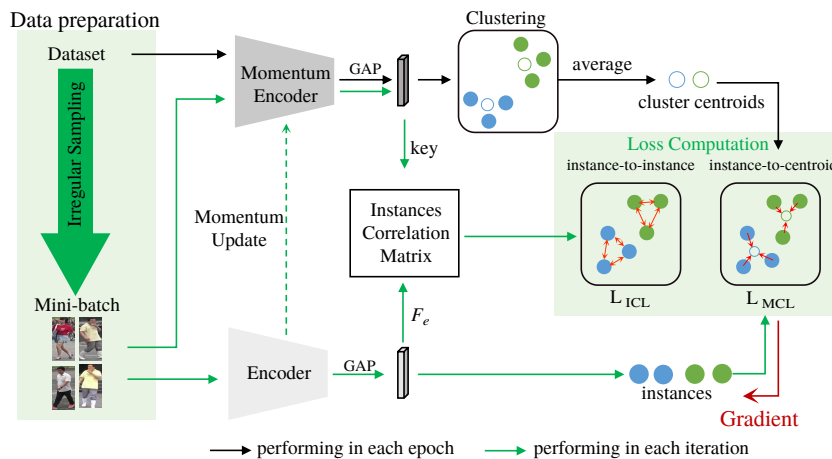


FIGURE 3.18: The illustration of the proposed fully unsupervised object re-ID framework.

The state-of-the-art unsupervised re-ID method (Ge et al., 2020; Chen, Lagadec, and Bremond, 2021; Wang et al., 2021) achieved significant success by utilizing strong self-supervised contrastive learning mechanisms, i.e., the Memory Bank approaches (Wu et al., 2018) and Momentum Contrast (MoCo) (He et al., 2019) 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 (Ge et al., 2020; Chen, Lagadec, and Bremond, 2021; Wang et al., 2021).

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



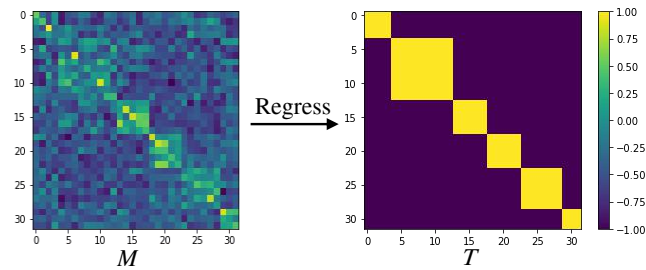


FIGURE 3.19: The example of  $M$ : instances correlation matrix, and  $T$ : target matrix.

are neglected in current framework. MoCo (He et al., 2019) demonstrated the importance of maintaining consistent representation for unsupervised learning, however, the representation of instances (query) and centroids are less consistent, as illustrated in Figure 3.18. 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 propose a Instance Correlation Loss (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.

### 3.6.1 Proposed Method

The MoCo-based re-ID contrastive learning framework in ICE (Chen, Lagadec, and Bremond, 2021) is adopted in this work as the baseline framework as shown in Figure 3.18.

#### 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 (He et al., 2019) with InfoNCE loss (Oord, Li, and Vinyals, 2018) is used as the baseline to enforce instance-to-centroid learning. The framework of the proposed method is illustrated in Figure 3.18.

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,

$$\theta_{me} = m\theta_{me} + (1 - m)\theta_e \quad (3.32)$$

Before each training epoch starts, given an unlabeled training dataset  $X = \{x_1, \dots, x_N\}$ , all images representations  $F_{me} = \{f_{me,1}, \dots, f_{me,N}\}$  are extracted by the momentum encoder. Then, unsupervised dense-based clustering algorithm DBSCAN (Ester et al., 1996) clusters  $F_{me}$  into  $N_C$  numbers of clusters. After that, cluster centroids  $C = \{c_0, \dots, c_{N_C}\}$  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  $B$ ,  $F_e = \{f_{e,1}, \dots, f_{e,N_B}\}$  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 (Oord, Li, and Vinyals, 2018) as follows,

$$L_{MCL} = -\log \frac{\exp(f_{e,i} \cdot c^+) / \tau}{\exp(f_{e,i} \cdot C) / \tau} \quad (3.33)$$

, 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.

### The Proposed Instances Correlation Loss

Training re-ID model only using momentum contrast learning with Equation 3.33 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

Method	Market-1501				DukeMTMC-reID			
	mAP	R-1	R-5	R-10	mAP	R-1	R5	R-10
BUC (Lin et al., 2019)	29.6	61.9	73.5	78.2	22.1	40.4	52.5	58.2
HCT (Zeng et al., 2020)	56.4	80.0	91.6	95.2	50.7	69.6	83.4	87.4
MMCL (Wang and Zhang, 2020)	45.5	80.3	89.4	92.3	40.2	65.2	75.9	80.0
DSCE (Yang et al., 2021)	61.7	83.9	92.3	-	53.8	73.8	84.2	-
SpCL (Ge et al., 2020)	79.1	88.1	95.1	97.0	65.3	81.2	90.3	92.2
CAP (Wang et al., 2021)	79.2	91.4	96.3	97.7	67.3	81.1	89.3	91.8
Group Sampling (Han et al., 2021b)	79.2	92.3	96.6	97.8	69.1	82.7	91.1	93.5
ICE (Chen, Lagadec, and Bremond, 2021)	82.3	93.8	97.6	98.4	69.9	83.3	91.5	94.1
Ours	<b>84.5</b>	<b>94.5</b>	<b>98.2</b>	<b>98.8</b>	69.3	82.7	90.1	92.3

TABLE 3.19: Experimental results of our proposed method and state-of-the-art fully unsupervised re-ID methods on Market-1501 and DukeMTMC-reID. ICL: Instances Correlation Loss. The top result is highlighted in bold and the second best result is shown in blue.

learning is ignored. Mining similarity relationship among instances is also beneficial to re-ID model performance (Wang and Zhang, 2020; Tang and Jo, 2021). 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 methods (Chen, Lagadec, and Bremond, 2021) designed an instance contrastive loss to enforce instance-to-instance learning, in which only one hardest positive and multiple negative samples are involved. It neglects the relationship among positive samples. Instead of using contrastive loss, our proposed instances of correlation loss involves multiple positive samples to increase intra-class compactness.

For the mini-batch  $B$ , L2 normalized key  $K = \{k_1, \dots, k_{N_B}\}$  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,

$$L_{ICL} = \|M - T\|^2 \quad (3.34)$$

, where  $T$  consists by  $-1$  and  $1$ , has same size with  $M$  as illustrated in Figure 3.19. 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  $-1$ . The overall loss of our proposed framework is the summation of Equation 3.33 and Equation 3.34.

Method	MSMT17			
	mAP	R-1	R-5	R-10
MMCL (Wang and Zhang, 2020)	11.2	35.4	44.8	49.8
DSCE (Yang et al., 2021)	15.5	35.2	48.3	-
SpCL (Ge et al., 2020)	19.1	42.3	55.6	61.2
CAP (Wang et al., 2021)	36.9	67.4	78.0	81.4
Group Sampling (Han et al., 2021b)	24.6	56.2	67.2	71.5
ICE (Chen, Lagadec, and Bremond, 2021)	38.9	70.2	80.5	84.4
Ours	<b>42.4</b>	<b>72.7</b>	<b>82.0</b>	<b>85.2</b>

TABLE 3.20: Experimental results of our proposed method and state-of-the-art fully unsupervised re-ID methods on MSMT17.

Method	VeRi-776			
	mAP	R-1	R-5	R-10
SpCL (Ge et al., 2020)	36.9	79.9	86.8	89.9
Ours (w/o $L_{ICL}$ )	38.6	81.9	86.5	88.9
Ours	<b>39.5</b>	<b>83.7</b>	<b>88.4</b>	<b>90.7</b>

TABLE 3.21: Experimental results of our proposed method on vehicle re-ID datasets VeRi-776.

Loss function	Market1501		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)	<b>84.5</b>	<b>94.5</b>	<b>69.3</b>	<b>82.7</b>

TABLE 3.22: Ablation study on using different loss functions for instance-to-instance learning

## 3.6.2 Experiments

### Datasets and Evaluation Metrics

We evaluate the proposed method on three large-scale and mainstream person re-id datasets, i.e., Market-1501 (Zheng et al., 2015b) (Market), and DukeMTMC-reID (Zheng, Zheng, and Yang, 2017; Ristani et al., 2016) (Duke), and MSMT17 (Wei et al., 2018b). The details are mentioned in Section 3.1.1. Moreover, vehicle re-id datasets VeRi (Liu et al., 2016) is used, and the details are mentioned in Section 3.1.2. Two evaluation metrics are mentioned in Section 3.1.3.

### 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 and Duke, and 6 hours on MSMT and Veri.

### 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 3.19 and Table 3.20. 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 3.21. We obtain  $Rank-1 = 83.7\%$  and  $mAP = 39.5\%$ , which considerably outperforms SpCL. The superior performance indicates that the

effectiveness of our proposed Irregular sampling and instance-to-instance learning loss  $L_{ICL}$ .

### 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 in (Chen, Lagadec, and Bremond, 2021). The results are reported in Table 3.22. 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 = 83.6% 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 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 (Chen, Lagadec, and Bremond, 2021) 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.

### 3.6.3 Conclusion

In this work, we proposed a fully unsupervised object re-ID method, which can be trained without using any labeled information. The current sampling strategies are analyzed. 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 person and vehicle re-ID datasets show the effectiveness of the proposed methods.

## 3.7 Unsupervised Object Re-identification via Relative Hard Samples Learning

The discrepancy in learning strategy causes the advantage of self-supervised contrastive learning mechanism to be not fully utilized in object re-ID tasks. One aspect is hard sample learning. Previous works (Ge, Chen, and Li, 2020; Ge et al., 2020; Hu, Zhu, and He, 2021; Chen, Lagadec, and Bremond, 2021) achieved impressive performance by integrating unsupervised clustering algorithms (Ester et al., 1996) and self-supervised contrastive learning (He et al., 2019; Wu et al., 2018), however they did not fully exploit information of hard samples. To utilize hard samples and mine more discriminative information, recent works (Chen, Lagadec, and Bremond, 2021; Hu, Zhu, and He, 2021) proposed hard instance contrastive loss. Chen et al. (Chen, Lagadec, and Bremond, 2021) performed contrastive learning between the hardest positive sample and all negative samples in every training mini-batch. Hu et al. (Hu, Zhu, and He, 2021) consider that mining hard negative samples within the mini-batch is not enough, therefore they proposed a Hard-sample guided Hybrid Contrast Learning (HHCL) to mine more negative samples from all negative clusters. The above methods (Chen, Lagadec, and Bremond, 2021; Hu, Zhu, and He, 2021) achieved state-of-the-art performance but they did not consider situations of 1) multiple hardest positive samples and 2) the relative information between positive and negative samples. Based on these two situations, we design an adaptive and more stable hard sample selection method, called as Relative Hard Samples (RHS) Selection, to enforce RHS learning in this paper. The MoCo-based self-supervised contrastive learning framework (He et al., 2019) is adopted as the baseline framework in this work. The illustration of our proposed framework is shown in Figure 3.20.

### 3.7.1 Related Work in Hard Sample Mining Strategy

The hard sampling mining strategy is widely used in many deep learning algorithms. Recently, some methods (Chen, Lagadec, and Bremond, 2021; Hu, Zhu, and He, 2021) incorporated the conception of hard sample mining into unsupervised person re-ID task to further distinguish easily confused samples by pulling hard positive samples closer and pushing hard negative samples away. Chen et al.

(Chen, Lagadec, and Bremond, 2021) mined the hardest positive instance and all negative instances in each mini-batch. To mine more negative samples, (Hu, Zhu, and He, 2021) proposed to select global hard samples online from all negative clusters rather than from a mini-batch. Our method also focuses on the hard sample mining strategy. Our proposed mining strategy, RHS selection, is more flexible than previous strategies by considering the statistical characteristics of each cluster.

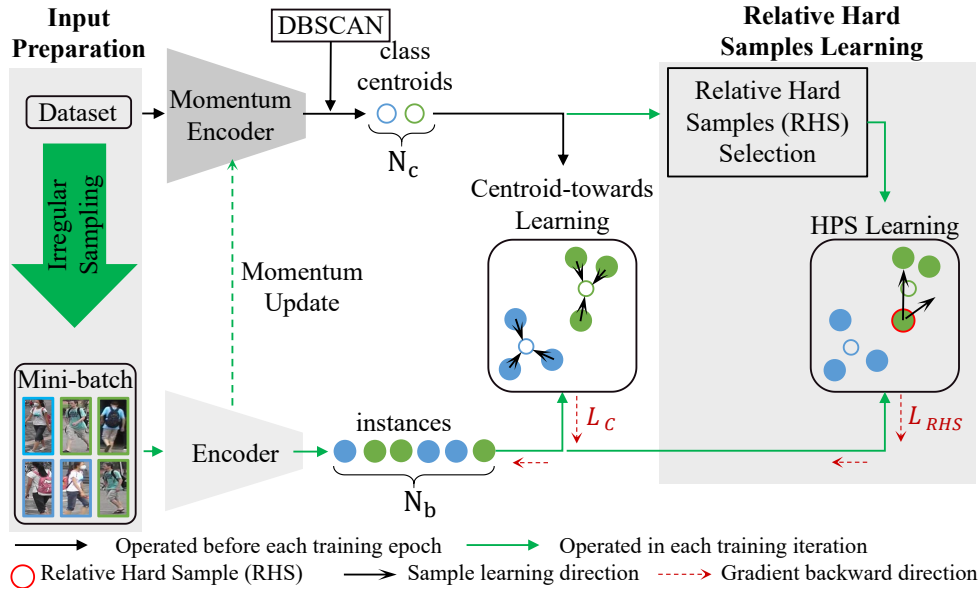


FIGURE 3.20: The illustration of the proposed fully unsupervised object re-ID framework. Before every training epoch, cluster algorithm DBSCAN (Ester et al., 1996) is used to roughly cluster every sample in the whole dataset into  $N_c$  classes.  $C = \{c_1, \dots, c_{N_c}\}$  represents centroids of  $N_c$  classes. In every training iteration, the encoder is fine-tuned according to  $C$  via centroids-towards learning and RHS learning, and the momentum encoder is updated by the encoder as Equation 3.33 by momentum update (He et al., 2019).

## 3.7.2 Proposed Method

### Centroids-towards learning

Same as Section 3.6.1, the MoCo-based self-supervised contrastive learning framework (He et al., 2019) is adopted as the baseline framework to perform momentum contrast learning with InfoNCE loss (Oord, Li, and Vinyals, 2018) for centroids-towards learning.



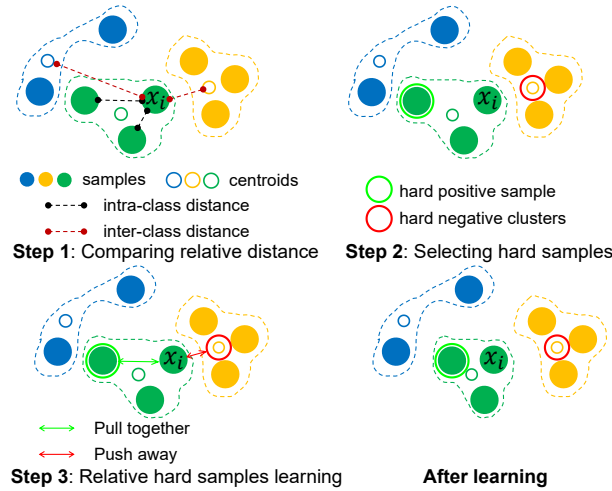


FIGURE 3.21: The illustration of the proposed RHS selection. The same color represents the samples containing the same identity.

### Relative Hard Samples (RHS) Learning

Training re-ID model only using momentum contrast learning with Equation still has three 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 (He et al., 2019). 2) The instance-to-instance learning is ignored. Mining similarity relationship among instances is also beneficial to re-ID model performance (Wang and Zhang, 2020; Tang and Jo, 2021; Tang, Cao, and Jo, 2021). 3) Hard samples are not exploited. Thus, we proposed a Relative Hard Samples (RHS) Learning to solve the above problems. RHS learning is performed to enforce instance-to-instance learning in each training iteration. We first introduce RHS selection, then RHS loss  $L_{RHS}$ .

**RHS Selection:** A reliable sample should have compactness with its intra-class samples, and it also should have independence with its inter-class samples. Therefore, our intuition is to select easily confused samples, i.e., hard positive samples and hard negative samples, by measuring the intra-class distances and inter-class distances. The illustrations of steps of the proposed RHS selection are shown in Figure 3.21.

**Step 1:** Given a feature  $f_{e,1}$  of samples  $x_i$ , we measure the intra-class distances between  $x_i$  and its intra-class samples by computing their cosine similarity,  $D_i^{intra}$  denotes intra-class distances of  $x_i$ . The inter-class distances of  $x_i$  are measured by

computing cosine similarity between  $x$  and its inter-class centroids,  $D_i^{inter}$  denotes inter-class distances of  $x_i$ .

**Step 2:** The hard positive samples and hard negative centroids are selected by comparing all distances in  $D_i^{intra}$  and  $D_i^{inter}$ . If there is an intra-class distance in  $D_i^{intra}$  larger than the inter-class distance in  $D_i^{inter}$ , the intra-class sample is assigned as a hard positive sample of  $x_i$ , and vice versa. Specifically if there is a inter-class distance in  $D_i^{inter}$  larger than intra-class distance in  $D_i^{intra}$ , the intra-class centroid is assigned as a hard negative sample of  $x_i$ . We use  $H_i^{hps}$  represents the selected hard positive sample of  $x_i$ , and  $H_i^{hns}$  represents the hard negative samples of  $x_i$ .

**Step 3:** The RHS learning are performed by minimizing the RHS loss as follows,

$$L_{RHS} = \mathbb{E} \left[ \sum_{pos \in H_i^{hps}} -\log \left( \frac{\langle pos, x_i \rangle / \tau_h}{\langle pos, x_i \rangle / \tau_h + \langle H_i^{hns}, x_i \rangle / \tau_h} \right) \right] \quad (3.35)$$

where  $\langle, \rangle$  denotes cosine similarity of features of two samples, and  $\tau_h$  is a temperature hyper-parameter.

**RHS Learning:** By minimizing the Equation ??, the proposed  $L_{RHS}$  can pull  $x_i$  and its multiple hard positive samples closer. Meanwhile,  $L_{RHS}$  pushes  $x_i$  and its hard negative clusters away by pushing cluster centroids of the hard negative clusters. With our proposed RHS selection, if a sample belongs to lower compactness and lower independence cluster, the sample could have a higher probability to choose more hard positive and hard negative samples. In other words, RHS learning increases intra-class compactness and inter-class separability based on the characteristics of samples.

The overall loss of our proposed framework is the summation of Equation 3.33 and Equation ??.

### 3.7.3 Experiments

#### Datasets and Evaluation Metrics

We evaluate the proposed method on three large-scale and mainstream person re-identification datasets, i.e., Market-1501 (Zheng et al., 2015b) (Market), and DukeMTMC-reID

Method	VeRi-776	
	mAP	R-1
SSML (Yu and Oh, 2021)	26.7	74.5
SpCL (Ge et al., 2020)	36.9	79.9
Ours (w/ IS)	39.9	<b>85.2</b>
Ours (w/ IS & RHS Learning)	<b>40.4</b>	85.1

TABLE 3.23: Experimental results of our proposed method and other fully unsupervised re-ID methods on vehicle re-ID datasets VeRi-776.

Method	Market-1501		MSMT17	
	mAP	Rank-1	mAP	R-rank-1
BUC (Lin et al., 2019)	29.6	61.9	-	-
HCT (Zeng et al., 2020)	56.4	80.0	-	-
MMCL (Wang and Zhang, 2020)	45.5	80.3	11.2	35.4
DSCE (DSCE)	61.7	83.9	15.5	35.2
SpCL (Ge et al., 2020)	79.1	88.1	19.1	42.3
CAP (Wang et al., 2021)	79.2	91.4	36.9	67.4
GS (Han et al., 2021b)	79.2	92.3	24.6	56.2
ICE (Chen, Lagadec, and Bremond, 2021)	82.3	93.8	38.9	70.2
CCL (Dai et al., 2021)	82.1	92.3	27.6	56.0
HHCL (Hu, Zhu, and He, 2021)	84.2	93.4	-	-
Ours (w/ IS)	83.4	93.9	40.2	71.3
Ours (w/ IS & RHS Learning)	<b>84.5</b>	<b>94.3</b>	<b>42.4</b>	<b>72.7</b>

TABLE 3.24: Experimental results of our proposed method and other fully unsupervised re-ID methods on two person re-ID datasets. The top result is highlighted in bold.

(Zheng, Zheng, and Yang, 2017; Ristani et al., 2016) (Duke), and MSMT17 (Wei et al., 2018b). The details are mentioned in Section 3.1.1. Moreover, vehicle re-id datasets VeRi (Liu et al., 2016) is used, and the details are mentioned in Section 3.1.2. Two evaluation metrics are mentioned in Section 3.1.3

### 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  $m = 0.999$ ,  $\tau = 0.05$ , are used in all experiments for fair comparisons.

Other hyper-parameters are selected for each datasets for achieving the best performance. In VeRi-776,  $P = 20$  and  $\tau = 0.15$ . In Market-1501,  $P = 16$  and  $\tau = 0.1$ . In MSMT17,  $P = 8$  and  $\tau_c = 0.1$ . The model performance with different  $P$  and  $\tau_h$  are reported in Figure 4 and Table IV.

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 6 hours on MSMT17 and VeRi-776.

### Comparisons with The State-of-the-Arts in Three Datasets

The comparisons with the State-of-the-Arts fully unsupervised methods on one vehicle re-ID dataset VeRi-776 in Table 3.23. We obtain mAP= 40.4% and Rank-1 = 85.1%, which considerably outperforms SpCL. The superior performance indicates the effectiveness of our proposed Irregular sampling and RHS selection and Learning by  $L_{RHS}$ .

Comparisons are also performed in two person re-ID datasets, i.e., Market-1501 and MSMT17, which are reported in Table 3.24. On Market-1501, our method achieves the best performance with mAP= 84.5% and Rank-1 = 94.3%. Compared to the best MoCo-based re-ID method IC, our method achieves good and competitive results.

$\tau_h$	Market-1501		VeRi-776	
	mAP	Rank-1	mAP	Rrank-1
0.05	83.3	93.2	39.4	83.8
0.10	<b>84.5</b>	<b>94.3</b>	40.1	85.0
0.15	84.0	93.7	<b>40.4</b>	85.1
0.20	83.7	93.8	39.9	<b>85.5</b>
0.25	83.8	93.6	39.5	85.0

TABLE 3.25: Experimental results of our proposed method with different values of  $\tau_h$ .

Method	mAP	Rank-1
Baseline	83.4	93.9
Hard Instance Contrastive Loss (Chen, Lagadec, and Bremond, 2021)	83.9	93.8
RHS Loss (Ours)	84.5	94.5

TABLE 3.26: Ablation study on using different loss functions for hard sample learning on Market-1501.

Specifically, 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.

### 3.7.4 Ablation Studies

#### Comparison with different $\tau_h$

The  $\tau_h$  is a temperature hyper-parameter in Equation 3.35, which controls the strength of penalties on hard negative samples (**temp**). More specifically, small  $\tau_h$  tends to concentrate more on hard negative samples, and large  $\tau_h$  tends to concentrate more on hard positive samples. We finally set  $\tau_h = 0.15$  for VeRi-776, and  $\tau_h = 0.10$  for Market-1501.

#### Comparison with other hard sample learning methods

To test the validity of the proposed hard sample learning methods RHS learning, we compare it against the baseline method and Hard Instance Contrastive loss in (Chen, Lagadec, and Bremond, 2021). The results are reported in Table 3.26, in where our proposed irregular sampling is adopted. The baseline method only uses  $L_C$  to perform centroids-towards learning, which outputs performance in mAP= 83.4% and in Rank-1 = 93.9% on Market-1501.

Two hard samples learning losses, including hard instance contrastive loss (Chen, Lagadec, and Bremond, 2021) and our proposed RHS Loss, boost the model performance from the baseline. The consistent improvements demonstrated the importance of mining information among instances. The idea of loss in (Chen, Lagadec, and Bremond, 2021) 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 RHS loss in Equation ?? can involve multiple positive samples and negative samples. Moreover, the proposed RHS selection automatically selects hard positive samples and hard negative samples for each sample according to the statistical characteristics of the sample, i.e., intra-class distances and inter-class distances. Intuitively, a sample in a low compactness and independence cluster could have a higher proportion of RHS learning by choosing more hard positive and hard negative samples. The performance of RHS loss remarkably surpasses the hard instance contrastive loss (Chen, Lagadec, and Bremond, 2021). which demonstrates the effectiveness of the proposed RHS loss.

### 3.7.5 Conclusions

In this work, we proposed a fully unsupervised object re-ID method, which can be trained without using any labeled information. Relative hard sample learning is proposed to adaptively increase intra-class compactness and inter-class separability based on the characteristics of samples. Experimental results on one vehicle and two person re-ID datasets show the effectiveness of the proposed methods.

## Chapter 4

# Supervised Object Search System

The specific object search is the foundation of a wide range of applications in intelligent security and surveillance systems. The object re-id aims to retrieve images containing the same identity. Although extensive object re-id methods (Zhou et al., 2019; Tang and Jo, 2021; Tang, Cao, and Jo, 2021; Wang and Zhang, 2020; Zhong et al., 2019) have been proposed, recent researchers (Xu et al., 2014; Xiao et al., 2017; Zhang, Li, and Zhang, 2021) found out that there is still a big gap between the object re-id system setting and real-world application. The object re-id systems are trained using well-cropped images, as shown in Figure 4.1 (a). However, object detectors might produce wrong-cropped images in practical applications. To close the gap, recent researches (Xu et al., 2014; Xiao et al., 2017; Zhang, Li, and Zhang, 2021; Yan et al., 2021; Li and Miao, 2021; Chen et al., 2020) tend to solve object detection and re-id jointly, namely object search.

An comparison illustration of the object search system is shown in Figure 4.1 (b). The object search system aims to detect the specific object regions from realistic and uncropped images. Then, based on detected object regions, the system retrieves the specific object regions that contained the same identity as a query image by matching detected regions with query images.

The person search can be considered as an integrated task of person detection and person re-identification. The existing person search framework can be summarized into two categories: two-step framework and end-to-end framework. Two-step methods (Chen et al., 2018; Han et al., 2019) tackled the detection and re-id with two separate models. End-to-end methods (Xiao et al., 2017; Zhang, Li, and Zhang, 2021; Yan et al., 2021; Li and Miao, 2021; Chen et al., 2020) unified detection and re-id tasks in one model by attaching the original detection and re-id branches parallelly.

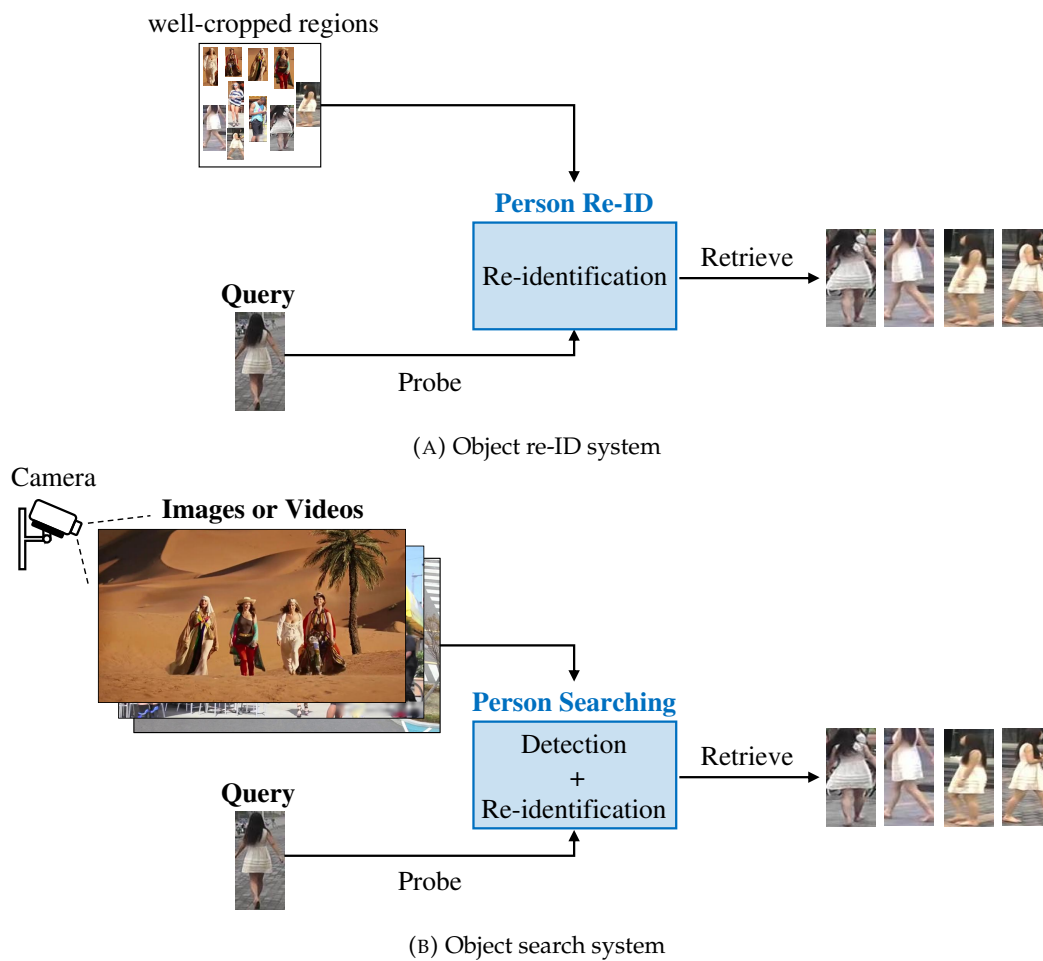


FIGURE 4.1: Illustration of the object re-ID and search system.

In general, two-step methods yield better performance but they are time-consuming and heavy, and end-to-end methods are faster and simpler but they can not obtain satisfactory re-id results. It is because the inconsistent objectives (Chen et al., 2018; Liang et al., 2020; Han, Ko, and Sim, 2021b; Zhang et al., 2021) between detection and re-id. More specifically, detection tends to produce similar features for person regions to distinguish them from backgrounds, but re-id tends to produce different features for person regions to further subdivide them into identities.

Chen et al. (Chen et al., 2018) first revealed the above goal conflict between the detection and re-ID. They argued that sharing features between the detection and re-ID tasks is not appropriate and therefore two-step methods yield better performance than end-to-end methods. Chen et al. presented a Mask-Guided Two-Stream (MGTS) method to eliminate the conflict. Wang et al. (Wang et al., 2020) considered the consistency between detection and re-ID stages and introduced a Task-Consist



Two-Stage (TCTS) framework. Recent end-to-end works also start to tackle the goal conflict between detection and re-id and further improve the person search performance. Chen et al. (Chen et al., 2020) proposed a Norm-Aware Embedding to decompose embedding into norm and angle for detection and re-id respectively. Li et al. (Li and Miao, 2021) proposed a Sequential End-to-end Network (SeqNet), which employed an extra detection head to provide high-quality Region of Interests (RoIs) and embedding for re-id. Han et al. (Han, Ko, and Sim, 2021b) proposed an Adaptive Gradient Weighting Function (AGWF) to control the weight of the back-propagated gradients according to the quality of detection results.

## 4.1 Literature Review

### 4.1.1 The Online Instance Matching (OIM)

The Online Instance Matching (OIM) loss (Xiao et al., 2017) is widely used in traditional person search methods. For a dataset with  $N$  numbers of identity classes, a  $N \times 256$  sized look-up table  $\mathcal{M}$  is built and maintained to store features for  $N$  classes.  $\mathcal{M}$  is updated in every training iteration by  $\theta$  as follows,

$$\mathcal{M}^t[i] = \alpha \mathcal{M}^{t-1}[i] + (1 - \alpha)\theta \quad (4.1)$$

where the superscript  $t$  denotes the  $t$ -th training iteration.  $\alpha \in [0, 1]$ , is the updating rate.  $i$  indicates the identity class of  $\theta$ . During the whole training process. Then, the similarity  $s$  of the current input image and  $N$  identity classes can be computed using  $\mathcal{M}$ . OIM loss aims to pull  $\theta$  close to its identity class  $i$  and push  $\theta$  far away from other identity classes. In other words, when the similarity between  $\theta$  and  $\mathcal{M}[i]$  is high and the similarities between  $\theta$  and other features in  $\mathcal{M}$  is low, the OIM loss is small. OIM loss is defined as follows,

$$L_{oim}^3 = -\log \frac{\exp(\langle \theta, \mathcal{M}^+ \rangle) / \tau}{\sum \exp(\langle \theta, \mathcal{M} \rangle) / \tau} \quad (4.2)$$

where  $\langle, \rangle$  denotes cosine similarity of features, restricted between  $[1, 1]$ .  $\mathcal{M}^+ = \mathcal{M}[i]$  is the  $i$ -th class that  $\theta$  belongs to.  $\tau$  is a temperature hyper-parameter.  $\langle \theta, \mathcal{M} \rangle$  is a  $N$ -dimensional vector, indicates the similarity between  $\theta$  and  $N$  identity classes.

### 4.1.2 Sequential End-to-end Network (SeqNet)

SeqNet extracts the  $2048d$  features using Faster RCNN (Ren et al., 2015), which contains a backbone network ResNet50 (He et al., 2016), a Region Proposal Network (RPN).

During training step, there are four loss in SeqNet, i.e.,  $L_{reg}^1, L_{cls}^1, L_{reg}^2, L_{cls}^2$ . Superscripts <sup>1</sup> and <sup>2</sup> indicate the first and second head of SeqNet, and subscripts <sub>reg</sub> and <sub>cls</sub> indicate the regression and classification loss, respectively.

For an input image  $x$ , 128 numbers of proposals are selected then aligned into  $1024 \times 14 \times 14$  RoIs by RoIAlign. The res5 in ResNet50 extracted these RoIs into  $2048d$  features to calculate the box regression loss. Following the previous work NAE (Chen et al., 2020),  $256d$  features  $f$  is extracted from  $2048d$  by fully connection to perform classification and re-id loss. To overcome the goal conflict between the classification and re-id, NAE (Chen et al., 2020) decomposing  $f$  into norm  $r$  and angle  $\theta$  in the polar coordinate system as follows:

$$f = r \cdot \theta \quad (4.3)$$

where norm  $r$  is  $1d$  value and angle  $\theta$  is a  $256d$  unit vector. To represent classification confidence using  $r \in [0, +\infty)$ , NAE normalize it to  $|r| \in [0, 1]$ . Four losses, i.e., regression loss  $L_{reg}^3$ , classification loss  $L_{cls}^3$ , online instance matching  $L_{oim}^3$  for re-id, and the proposed BFCL  $L_{bfcl}^3$  are used in the third head. The total learning objective function is then formulated as,

$$L = \lambda_1 L_{reg}^1 + \lambda_2 L_{cls}^1 + \lambda_3 L_{reg}^2 + \lambda_4 L_{cls}^2 + \lambda_5 L_{reg}^3 + \lambda_6 L_{cls}^3 + \lambda_7 L_{oim}^3 + \lambda_8 L_{bfcl}^3 \quad (4.4)$$

Following the SeqNet,  $\lambda_1 = 10$ , and the others are set to 1.

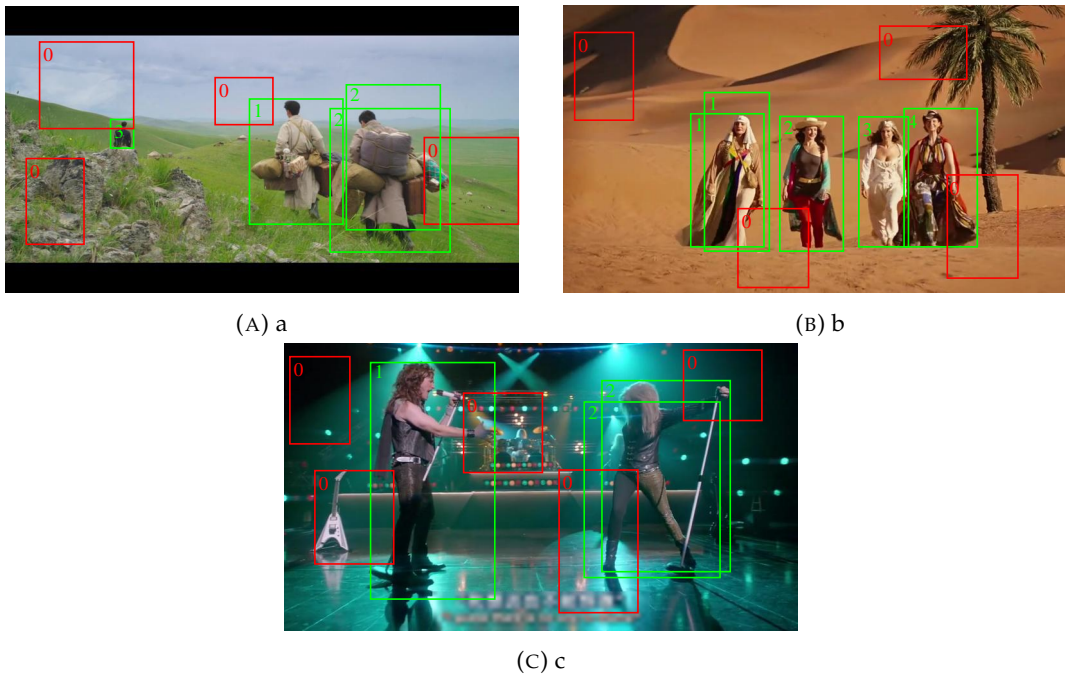


FIGURE 4.2: Examples of backgrounds (red boxes) and foregrounds (green boxes) RoIs in (a)-(c) three different input images. The number at the top left corner represents the identity  $i$ .  $i = 0$  indicates background, and  $i > 0$  indicates foreground. Different  $i$  means different identity.

## 4.2 Proposed Person Search via Background and Foreground Contrastive Learning

### 4.2.1 The Proposed BFCL loss

In this section, we revisit the end-to-end person search network SeqNet in Section 4.1.2. The SeqNet is the baseline framework of our work. Then, we introduce the overview of person search architecture and the proposed Background and Foreground Contrastive Loss (BFCL) in detail.

#### Architecture Overview

SeqNet (Li and Miao, 2021) with OIM (Xiao et al., 2017) loss is the baseline framework of our work. The architecture of the SeqNet with our proposed BFCL is illustrated in Figure 4.3.

Although previous methods achieved good results, the relationship among RoIs

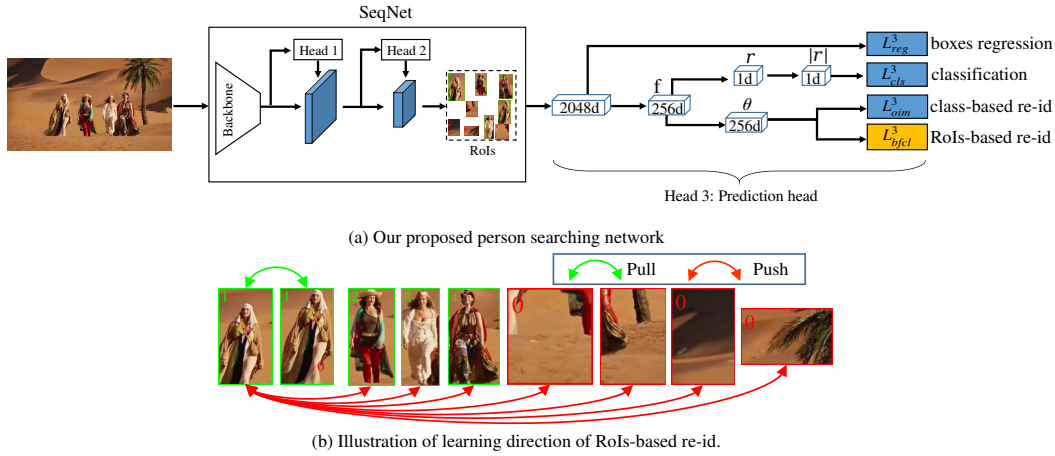


FIGURE 4.3: The architecture of the proposed person searching framework. The component in yellow is newly proposed by us. Our proposed Backgrounds and Foregrounds Contrasting Loss  $L_{bfcl}^3$  aims to push an ROI far away from other RoIs with different  $i$  and backgrounds.

of an image has not been explored. OIM loss can be considered as class-based learning by pulling a feature closer to its corresponding class feature but OIM loss did not consider the relationship among RoIs of an image. Intuitively, an input image has its own characteristic patterns and therefore RoIs from the image have high probabilities of containing similar patterns. Three examples are illustrated in Figure 4.2. The RoIs in Figure 4.2 (a) mainly contain sky, grassland, or stone. RoIs in Figure 4.2 (b) contain desert. RoIs in Figure 4.2 (c) contain a stage with green light. Moreover, person regions (foregrounds) from the same image are more difficult to classify than person regions from different images because of similar patterns. The similar patterns lead the RoIs from the same image to be more difficult to classify. Therefore, we explore the relationship among RoIs additionally to help the model learn discriminative identity features from the foreground. To address this issue, we propose Background and Foreground Contrastive Loss (BFCL) to further boost re-id performance by leveraging inter-RoIs pairwise similarity. With the help of BFCL, the person search model is able to differentiate similar RoIs for re-id. The learning directions of the proposed BFCL loss are illustrated in Figure 4.3 (b).

For one ROI feature  $\theta$  in an input image, the BFCL is computed as follows,

$$L_{bfcl}^3 = \frac{1}{|\theta^+|} \sum_1^{|\theta^+|} -\log \frac{\exp(\langle \theta, \theta^+ \rangle) / \tau_c}{\sum \exp(\langle \theta, \mathcal{U} \rangle) / \tau_c} \quad (4.5)$$

where  $\mathcal{U} = \{\theta_1, \dots, \theta_{128}\}$  indicates the collection of all IoU features in one input image.  $\theta^+$  indicates the positive RoIs in  $\mathcal{U}$  that have the same class with  $\theta$ . Intuitively, the above  $L_{bfc}^3$  encourages the  $\theta$  to approach its positive RoIs, and leave its negative RoIs.

## 4.2.2 Experiments

The experiments are performed on two widely used datasets, CUHK-SYSU (Xiao et al., 2017) and PRW (Zheng et al., 2017). (Chen et al., 2018).

### Datasets

**CUHK-SYSU:** CUHK-SYSU (Xiao et al., 2017) contains two data sources to diversity the scenes. The first one contains street snaps in an urban city, is shot by hand-held cameras. The second data source is collected from movie snapshots, which contain person images with abundant variations of viewpoints, lighting, and background conditions. StreetSnap images and MovieTV screenshots. The datasets contains 18,184 uncropped images, 96,143 person bounding boxes with 8,432 labeled identities in total. The training set has 11,206 images, 55,272 persons with 5,532 different identities. The test set has 6,978 images, 40,871 persons with 2,900 different identities.

**PRW:** PRW (Zheng et al., 2017) are collected in Tsinghua university for about 10 hours with 6 cameras. The PRW aims to simulate real-world situations where pedestrians appear or disappear in different cameras. The datasets contains 11,816 uncropped images, 43,110 person bounding boxes with 484 labeled identities in total. Both CUHK-SYSU and PRW contain unlabeled identities. For example. In our paper, unlabeled identities are used in regression and classification losses but not used in re-id and our proposed BFCL.

**Evaluation Metrics:** Same with re-id task, 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 CMC (%) of Top-1 is reported, which represents the probability of top-1 ranked gallery samples containing the query identity. Following the previous works, the detection

evaluation metrics are not used to evaluate the performance of the person search model.

**Implementation Details:** Faster R-CNN (Ren et al., 2015) is adopted as the backbone network, in which ResNet-50 (He et al., 2016) pretrained on ImageNet is used. The SeqNet (Li and Miao, 2021) is the baseline framework of our method. The backbone network contains the res1, res2, res3, and res4 blocks of ResNet-50, and the output features of res4 are used for the first prediction head. The output features of res5 are used for the second prediction head.

The input images are resized to  $900 \times 1500$ . The batch size is 5. The network is trained by the Stochastic Gradient Descent (SGD) with a learning rate of 0.003 which is warmed up during the first epoch and decreased by 10 at the 16-th epoch. The model is trained for 20 epochs in CUHK-SYSU and 18 epochs in PRW. The circular queue size of OIM is not used here because the circular queue did not enhance model performance consistently in two datasets, the experimental results are reported in Table I. The updating rate  $\alpha$  in Equation 4.1 and  $\tau$  in Equation 4.2 are set to 0.5 and  $1/3$ , respectively.

The experiments are performed on one NVIDIA Tesla V100 GPU with 32 GB of memory. The total training time is around 24 hours on CUHK-SYSU, and 17 hours on PRW.

### Ablation Study

Ablation studies are performed to demonstrate the effectiveness of the proposed BFCL and analyze the effectiveness of different temperature values  $\tau_c$ .

**Effectiveness of Circular Queue (CQ):** We implement the analysis of Circular Queue (CQ) (Xiao et al., 2017) on SeqNet-base model (Li and Miao, 2021). The results are reported in Table 4.1. Our re-implementation of the SeqNet model without CQ is notated as “SeqNet-base” in Table 4.1. Xiao et al. (Xiao et al., 2017) proposed CQ to store the features of unlabeled identities situation. They demonstrate the effective use of CQ in their framework. However, we found out CQ did not enhance SeqNet performance consistently in two datasets, as reported in “SeqNet-base” and “SeqNet-base + CQ” in Table 4.1. Adding CQ to SeqNet-base yields a gain of +0.3

Methods	CUHK-SYSU		PRW	
	mAP	Top-1	mAP	Top-1
SeqNet-base	93.6	94.1	47.2	83.6
SeqNet-base + CQ	93.9	94.5	46.7	83.4
SeqNet-base + BFCL	<b>94.0</b>	<b>94.6</b>	<b>48.7</b>	<b>84.4</b>

TABLE 4.1: Ablation experiments on **CQ**: Circular Queue (Xiao et al., 2017) and **BFCL**: our proposed Background and Foreground Contrastive Loss.

$\tau_c$ in Equation 4.5	CUHK-SYSU		PRW	
	mAP	Top-1	mAP	Top-1
0.03	<b>94.0</b>	<b>94.6</b>	48.5	84.1
0.05	93.6	94.2	<b>48.7</b>	<b>84.4</b>
0.10	92.7	93.5	47.9	84.1

TABLE 4.2: Performance of our framework with different values of  $\tau_c$  in Equation 4.5.

in mAP and +0.4 in Top-1 in CUHK-SYSU but decreases  $-0.5$  in mAP and  $-0.2$  in Top-1 in PRW. Therefore, The CQ is not used in our paper.

*Effectiveness of Proposed BFCL:* We implement the analysis of our proposed BFCL on SeqNet-base model (Li and Miao, 2021). The results are reported in Table 4.1. It is clear that adding BFCL to the SeqNet-base yields consistent gain in two datasets. Specifically, Adding BFCL to SeqNet-base yields a gain of +0.4 in mAP and +0.5 in Top-1 in CUHK-SYSU but decreases +1.5 in mAP and +0.8 in Top-1 in PRW.

*Comparison with Different Temperature  $\tau_c$  in Equation 4.5:* Almost all contrastive learning-based methods (Chen, Lagadec, and Bremond, 2021; He et al., 2019; Li and Miao, 2021) used the temperature value and have similar effects. (Wang and Liu, 2021) demonstrated the contrastive loss is a hardness-aware loss function, and the temperature value  $\tau_c$  controls the strength of penalties on hard negative samples. Small  $\tau_c$  tends to pay more attention to the hard negative samples. Large  $\tau_c$  tends to pay less attention to the hard negative samples, in other words, less sensitive to the hard negative samples. Our person search framework performance in two datasets with different temperature  $\tau_c$  are reported in Table 4.2. For CUHK-SYSU, when  $\tau_c = 0.03$ , our framework achieves the best results 94.0% in mAP and 94.6% in Top-1. For PRW, when  $\tau_c = 0.05$ , our framework achieves the best results 48.7% in mAP and 84.4% in Top-1. The different optimal value of  $\tau_c$  in CUHK-SYSU and PRW demonstrates that paying more attention to the hard negative samples helps



Method	Reference	CUHK-SYSU		PRW	
		mAP	Top-1	mAP	Top-1
OIM (Xiao et al., 2017)	CVPR17	75.5	78.7	21.3	49.9
NAE (Chen et al., 2020)	CVPR20	91.5	92.4	43.3	80.9
AGWF (Han, Ko, and Sim, 2021b)	ICCV21	93.3	94.2	53.3	87.7
AlignPS (Yan et al., 2021)	CVPR21	93.1	93.4	45.9	81.9
SeqNet (Li and Miao, 2021)	AAAI21	93.8	<b>94.6</b>	46.7	83.4
BFCL	Ours	<b>94.0</b>	<b>94.6</b>	<b>48.7</b>	<b>84.4</b>

TABLE 4.3: Comparison with state-of-the-art methods on two person searching datasets. The top result is highlighted in bold.

model performance in CUHK-SYSU. On the other hand, paying less attention to the hard negative samples helps model performance in PRW.

### Comparison with the state-of-the-art Methods

We compare our method against state-of-the-art person search models in CUHK-SYSU and PRW in Table 4.3. As the baseline of recent person search works (Chen et al., 2020; Yan et al., 2021; Li and Miao, 2021), OIM (Xiao et al., 2017) is the first paper that proposed Online Instance Matching (OIM) loss function to end-to-end train the re-id with detection jointly. NAE (Chen et al., 2020) notice the end-to-end training strategy enhance the goal conflict between detection and re-id, therefore NAE (Chen et al., 2020) decompose feature  $f$  into norm  $r$  and angle  $\theta$ . Based on two head method NAE, SeqNet (Li and Miao, 2021) added one prediction head to improve the detection accuracy for providing high-quality RoIs. Based on SeqNet, our proposed BFCL further boosts the person search performance in both CUHK-SYSU and PRW datasets, especially in PRW. The consistent improvements demonstrate the necessity of mining relationships among RoIs of an image and the effectiveness of our proposed BFCL.

### 4.2.3 Conclusion

We introduced an end-to-end person search model in this paper. To strengthen the re-id capability of the model, we propose a Background and Foreground Contrastive Loss (BFCL) which can leverage similarity relationships among RoIs to learn to distinguish similar backgrounds and foregrounds. Moreover, we demonstrate that the widely used CQ can not enhance the performance of our model consistently in two



datasets. In the future, we wish to integrate the proposed algorithm in high-level video surveillance tasks.

## Chapter 5

# Weakly Supervised Object Search System

In the past decade, object search works mostly focused on supervised learning, which achieved significant progress (Li and Miao, 2021; Chen et al., 2020; Xiao et al., 2017). The supervised object search requires substantial labeled training data for satisfying performance. The training process of supervised object search requires strong supervision in terms of bounding boxes and identity information. However, it is expensive and time-consuming to annotate bounding boxes, especially in labeling identities across multiple cameras. More specifically, large-scale labeled training data is often difficult to collect, especially for identities. A lot of existing researches have been dedicated to train the model with incomplete labeled (Zhong et al., 2019) or fully unlabeled dataset (He et al., 2019; Chen, Lagadec, and Bremond, 2021) in the object re-ID field. However, relevant researches are missing in the field of person search. To fill the gap, some recent works focus on using the weakly supervised object search method (Han et al., 2021a; Han, Ko, and Sim, 2021a). Comparisons between fully supervised setting and weakly supervised setting are illustrated in Figure 5.1.

As shown in Figure 5.1 (a), some identity annotations have lacked in original person search datasets CUHK-SYSU (Xiao et al., 2017) with the supervised setting. As shown in Figure 5.1 (b), only the positions of bounding boxes are provided. In other words, the weakly supervised setting alleviates the burden of obtaining manually labeled identities. Without human-annotated correct identity labels, the person search system is difficult to learn robust and discriminative features for re-ID.



FIGURE 5.1: Comparisons between two person search settings. (a) Supervised setting. The images are annotated with both bounding boxes and person identities. Note that some identity annotations have lacked in original person search datasets. (b) The proposed weakly supervised setting. The images only have bounding box annotations.

## 5.1 Weakly Supervised Object Search with Region Siamese Networks

(Han et al., 2021a) set up a strong and effective weakly supervised baseline termed Region Siamese Networks (R-SiamNets), which learns useful representations for person re-id in the absence of identity labels. They supervise the R-SiamNet with instance-level consistency loss and cluster-level contrastive loss. The architecture of R-SiamNet is illustrated in Figure 5.2.

There are two branches in R-SiamNet. The upper branch is fed with the whole scene image and then extracts the RoI features of the person region. The below branch is fed with the cropped person region and then extracts its features. Based on these two branches, there are two consistency learning strategies. The first consistency learning strategy is instance-level consistency learning. The R-SiamNet is constrained to extract consistent features from the whole scene image and the cropped person region. The second consistency learning strategy is inter-instance similarity

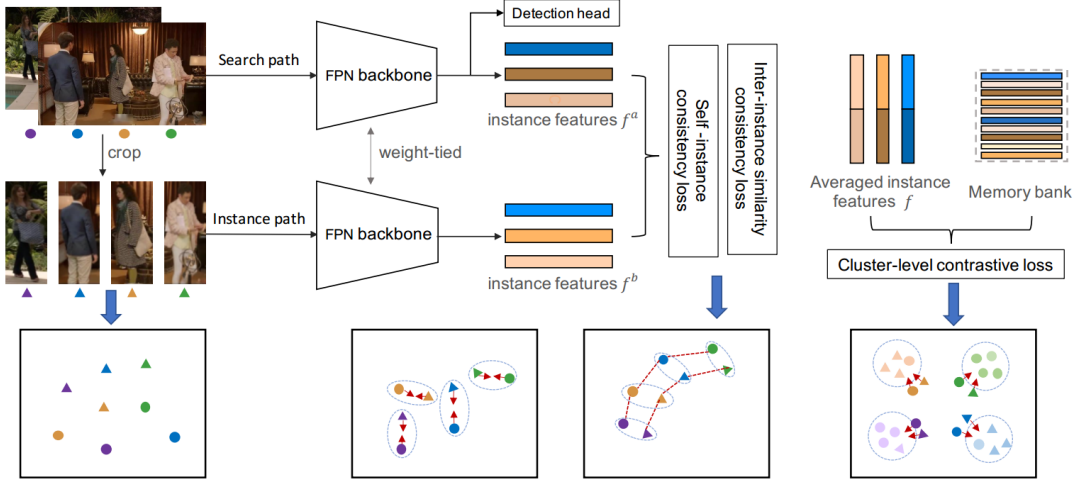


FIGURE 5.2: Illustration of our R-SiamNet (Han et al., 2021a)

consistency learning. Moreover, cluster-level contrastive learning is introduced for clustering and separating instances in every training iteration.

### 5.1.1 Instance-Level Consistency Learning

The upper branch, called as search path, is fed with the whole scene image and then extracts the RoIs features of person regions. The output features of search path are denoted as  $F_s = \{f_1^s, \dots, f_B^s\}$ . The below branch, called as instance path, is fed with the well-cropped person regions. The cropped person regions are cropped by the bounding box labels. The L2-normalized output features of instance path are denoted as  $F_i = \{f_1^i, \dots, f_B^i\}$ .  $B$  is the number of cropped regions in a mini-batch in each training iteration. Two consistency learning strategies are designed using  $F_s$  and  $F_i$ : self-instance consistency loss and inter-instance similarity consistency loss.

**Self-instance consistency loss:** Self-instance consistency loss is proposed to maximize the cosine similarity of each set of corresponding terms in  $F_s$  and  $F_i$  as follows,

$$L_{ins} = \frac{1}{B} \sum_{i=1}^B (1 - \langle f_i^s, f_i^i \rangle) \quad (5.1)$$

, where  $\langle, \rangle$  indicates the cosine similarity of two features.

**Inter-instance similarity consistency loss:** Inter-instance similarity consistency loss is proposed to learn the relationship among different features. The similarity matrix  $S^s \in \mathbb{R}^{B \times B}$  is computed by  $S^s = F_s \times F_s^\top$ . In the same way, The similarity

matrix  $S^i \in \mathbb{R}^{B \times B}$  is computed by  $S^i = F_i \times F_i^\top$ . The aim of Inter-instance similarity consistency loss is to keep the consistency of  $S^s$  and  $S^i$ . The Kullback–Leibler divergence (KL divergence) are used as follows,

$$L_{int} = D_{KL}(S^s || S^i) + D_{KL}(S^i || S^s) \quad (5.2)$$

### 5.1.2 Cluster-Level Contrastive Learning

To striking a balance between clustering and separation, pseudo labels are assigned to each instance by find their positive and negative samples. A positive sample  $x_j^p$  of an instance  $x_j$  should satisfy two conditions simultaneously.

1.  $x_j^p$  and  $x_j$  should come from different whole scene image.
2.  $x_j^p$  is the nearest neighbor of  $x_j$  or  $x_j$  is the nearest neighbor of  $x_j^p$ .

Following previous works (**SimSLR**; Wang and Zhang, 2020; Zhong et al., 2019; Lin et al., 2020), a memory bank  $M$  is adopted to store the embeddings of all instances in dataset, where  $M \in \mathbb{R}^{N \times d}$ .  $d$  denotes the feature dimensions. Memory bank is updated after each training iteration to ensure the up-to-date information using  $M_t \leftarrow \lambda M_t + (1 - \lambda)F^s$ . Based on the memory bank, the cluster-level contrastive loss are use to enforce clustering and separating instances as follows,

$$L_{cls} = \log[1 + \sum_K^{i=1} \sum_J^{j=1} \exp((s_n^j - s_p^i)/\tau)] \quad (5.3)$$

, where  $\tau$  is the scale factor.  $s_p^i$  is similarity between current instance and its positive samples.  $s_n^j$  is similarity between current instance and its negative samples. The Equation 5.3 aims to make the  $s_p^i$  greater than  $s_n^j$ .

In summary, the overall training objective is the summation of Equation 5.1, Equation 5.2, and Equation 5.3 and the detection loss.

## Chapter 6

# Conclusion

The works in this thesis focus on researching object re-id and search systems in intelligent surveillance systems. Object re-id and search systems, concentrated in the field of the person and vehicle re-id and search, which have been broadly used in academics and quite a few industry implementations, such as the person and vehicle tracking and search, persons and vehicles behavior analysis, detecting and identifying abnormal actions or situations, crime and terrorism identification, etc.

Object re-ID, including person and vehicle re-ID, aims to re-identify the specific object across multiple non-overlapping cameras. The object re-id system is trained to match a query image with the human manually cropped and well-cropped images and thus it is far from real-world applications. However, object detectors in intelligent systems might produce wrong-cropped images in practical applications, which leads to a bad re-id performance.

Therefore, training a system which able to identify a specific object from full scene images is closer to the real-world applications, therefore we further investigate object search systems in this manuscript. Object search integrates detection and re-ID tasks into one model and therefore it is more realistic and applicable in real-world practical applications.

In order to further increase the practicability of the system, we also focus on investigating the unsupervised learning strategy. We focus on combining the supervised detection methods and unsupervised object re-ID methods. In other words, only coordination of bounding boxes is provided and identity information is unknowable. Unsupervised learning does not require annotating the identity for each person or vehicle image. Moreover, it is relatively easier to acquire a large amount

of unlabeled data by public surveillance systems in the real world. Therefore, it is closer to real-world applications in intelligent surveillance systems.

In a summary, we improve the performance of the unsupervised object re-ID systems in three ways: (1) sampling strategy, (2) generated pseudo labels refinement, and (3) loss function design. We improve the performance of the object search system in two ways: (1) usage of unlabeled targets and (2) loss function design. Several aspects including speed improvement and quality enhancement are left as future works as described in the following section.

## 6.1 Future Works

In the future, we mainly focus on below three aspects to enforce the practicability of object re-id and search in intelligent surveillance systems.

1. Integrating the object search system into low-cost devices. Currently, the existing object search system is trained and run on GPU, which is expensive and heavy for most surveillance systems. Therefore, performing object search experiments with low-cost devices are left as future work.
2. Searching for a more autonomous and adaptive matching (retrieval) strategy in evaluation. The existing matching strategy is not practically applicable in real-world applications. One strategy in the existing methods is to use the nearest neighbor but it will fail when the target does not exist. Another strategy is to use a fixed similarity threshold but it might predict multiple results. Therefore, designing a more practical retrieval strategy for real-world surveillance systems is left as future work.
3. Applying weakly supervised object searching in real-world surveillance systems. Compared with it is relatively easier to acquire a large amount of unlabeled and suitable data by public surveillance systems in the real world.

## Appendix A

# Publications

### A.1 Journal

1. Q. Tang, G. Cao and K. -H. Jo, "Fully Unsupervised Person Re-Identification via Multiple Pseudo Labels Joint Training," in *IEEE Access*, vol. 9, pp. 165120-165131, 2021. (IF: 3.369)
2. Q. Tang, G. Cao and K. -H. Jo, "Integrated Feature Pyramid Network With Feature Aggregation for Traffic Sign Detection," in *IEEE Access*, vol. 9, pp. 117784-117794, 2021. (IF: 3.369)

### A.2 Conference

1. Q. Tang and K. -H. Jo, "Unsupervised Person Re-Identification Via Nearest Neighbor Collaborative Training Strategy," 2021 IEEE International Conference on Image Processing (ICIP), 2021, pp. 1139-1143. (h-index: 96)
2. Q. Tang and K. Jo, "Person Search via Background and Foreground Contrastive Learning," 15th IEEE International Conference on Human System Interaction (HSI), 2022. (accepted).
3. Q. Tang and K. Jo, "Unsupervised Person Re-identification via Mining Label Homogeneity," 23rd IEEE International Conference on Industrial Technology (ICIT), 2022. (accepted)
4. Q. Tang and K. Jo, "Fully Unsupervised Person Re-Identification via Centroids and Neighborhoods Joint Learning," 2022 IEEE International Symposium on Industrial Electronics (ISIE), 2022. (accepted)



5. Tang Qing, Youlkyeong Lee, Kang-Hyun Jo, Occlusion Assistant Unit for Convolution Neural Network in Occluded Object Classification, IWIS 2020, Ulsan, Korea, Dec 13, 2020.
6. Ge Cao, Qing Tang, and Kang-hyun Jo. 2020. Aggregated Deep Saliency Prediction by Self-attention Network. In *Intelligent Computing Methodologies: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part III*. Springer-Verlag, Berlin, Heidelberg, 87–97.
7. Tang, Q., Cao, G., Jo, K. Accurate and Efficient Traffic Sign Detection with a Guided Region Enlarging Algorithm. In: Huang, DS., Bevilacqua, V., Hussain, A. (eds) *Intelligent Computing Theories and Application. ICIC 2020. Lecture Notes in Computer Science*, vol 12463. (2020) Springer, Cham.
8. Tang, Q., Lee, Y., Jo, K. (2019). Occluded Object Classification with Assistant Unit. In: Huang, DS., Huang, ZK., Hussain, A. (eds) *Intelligent Computing Methodologies. ICIC 2019. Lecture Notes in Computer Science*, vol 11645. Springer, Cham.
9. Tang Qing, Laksono Kurnianggoro, Kang-Hyun Jo, Traffic Sign Classification with Dataset Augmentation and Convolutional Neural Network, ICGIP 2017, Qingdao, China, Oct 14, 2017.
10. Qing Tang and Kang-Hyun Jo, "Analysis of Various Traffic Sign Detectors Based on Deep Convolution Network," 2019 IEEE/SICE International Symposium on System Integration (SII), 2019, pp. 507-511.
11. Qing Tang, Laksono Kurnianggoro, and Kang-Hyun Jo "Traffic sign classification with dataset augmentation and convolutional neural network", Proc. SPIE 10615, Ninth International Conference on Graphic and Image Processing (ICGIP 2017), 106152W (10 April 2018)
12. Tang Qing, Kang-Hyun Jo, Analysis of Traffic Sign Classification using Multiple Image Preprocessing Method, CICIRO2017, Changwon, Korea, Nov 30, 2017.
13. Tang Qing, Laksono Kurnianggoro, Nguyen Quang Huy, Kang-Hyun Jo, Vehicle Detection Using LiDAR in Real Road Based on Artificial Neural Network, ICT-ROBOT 2016, Busan, Korea, Sep 7, 2016.

14. Laksono Kurnianggoro, Nguyen Quang Huy, Tang Qing, and Kang-Hyun Jo, Comparative Study of Various Machine Learning Algorithms for Object Recognition on Single Scan 2D LIDAR, ICT-ROBOT 2016 ICT-Robot 2016, Busan, Korea, Sep 7, 2016.
15. Q. Tang, L. Kurnianggoro and K. Jo, "Statistical and geometrical features for LiDAR-based vehicle detection," 2016 IEEE/SICE International Symposium on System Integration (SII), 2016, pp. 192-197.
16. Tang Qing, Kang-Hyun Jo, Analysis of Traffic Sign Classification using Multiple Image Preprocessing Method, CICIRO2017, Changwon, Korea, Nov 30, 2017.

# Bibliography

- Agarwal, Pankaj K. and Nabil H. Mustafa (2004). “k-means projective clustering”. In: *PODS '04*.
- al, Glenn Jocher et (Apr. 2021). “ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models”. In: *ArXiv*.
- Chen, Di et al. (2018). “Person Search via A Mask-Guided Two-Stream CNN Model”. In: *ECCV*.
- Chen, Di et al. (2020). “Norm-Aware Embedding for Efficient Person Search”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12612–12621.
- Chen, Hao, Benoit Lagadec, and François Bremond (2021). “ICE: Inter-Instance Contrastive Encoding for Unsupervised Person Re-Identification”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14960–14969.
- Comparitech (2021). *The world's most-surveilled cities*. URL: <https://www.comparitech.com/blog/vpn-privacy/us-surveillance-camera-statistics/> (visited on 05/01/2022).
- Dai, Zuozhuo et al. (2021). “Cluster Contrast for Unsupervised Person Re-Identification”. In: *ArXiv abs/2103.11568*.
- Deng, Jia et al. (2009). “ImageNet: A large-scale hierarchical image database”. In: *CVPR*.
- Ding, Guodong, Salman Hameed Khan, and Zhen min Tang (2019). “Dispersion based Clustering for Unsupervised Person Re-identification”. In: *BMVC*.
- Durand, Thibaut, Nazanin Mehrasa, and Greg Mori (2019). “Learning a Deep ConvNet for Multi-Label Classification With Partial Labels”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 647–657.
- Ester, Martin et al. (1996). “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *KDD*.

- Fan, Hehe, Liang Zheng, and Yi Yang (2017). "Unsupervised Person Re-identification: Clustering and Fine-tuning". In: *arXiv: Computer Vision and Pattern Recognition*.
- Feizi, Asghar (2017). "High-Level Feature Extraction for Classification and Person Re-Identification". In: *IEEE Sensors Journal* 17, pp. 7064–7073.
- Feng, Lei et al. (2020). "Can Cross Entropy Loss Be Robust to Label Noise?" In: *IJCAI*.
- Fu, Yang et al. (2019a). "Horizontal Pyramid Matching for Person Re-identification". In: *ArXiv abs/1804.05275*.
- Fu, Yang et al. (2019b). "Self-Similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-Identification". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6111–6120.
- Ge, Yixiao, Dapeng Chen, and Hongsheng Li (2020). "Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification". In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=rJln0hVYPS>.
- Ge, Yixiao et al. (2020). "Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID". In: *Advances in Neural Information Processing Systems*.
- Han, Byeong-Ju, Kuhyeun Ko, and Jae-Young Sim (2021a). "Context-Aware Unsupervised Clustering for Person Search". In: *ArXiv abs/2110.01341*.
- (2021b). "End-to-End Trainable Trident Person Search Network Using Adaptive Gradient Propagation". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 905–913.
- Han, Chuchu et al. (2019). "Re-ID Driven Localization Refinement for Person Search". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9813–9822.
- Han, Chuchu et al. (2021a). "Weakly Supervised Person Search with Region Siamese Networks". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11986–11995.
- Han, Xumeng et al. (2021b). "Rethinking Sampling Strategies for Unsupervised Person Re-identification". In.
- He, Kaiming et al. (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.
- He, Kaiming et al. (2019). "Momentum Contrast for Unsupervised Visual Representation Learning". In: *arXiv preprint arXiv:1911.05722*.

- Hu, Zheng, Chuang Zhu, and Gang He (2021). "Hard-sample Guided Hybrid Contrast Learning for Unsupervised Person Re-Identification". In: *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, pp. 91–95.
- Huang, Houjing et al. (2020). "Improve Person Re-Identification With Part Awareness Learning". In: *IEEE Transactions on Image Processing* 29, pp. 7468–7481.
- Ji, Zilong et al. (2020). "An Attention-Driven Two-Stage Clustering Method for Unsupervised Person Re-identification". In: *ECCV*.
- Jiang, Kongzhu et al. (2020). "Self-Supervised Agent Learning for Unsupervised Cross-Domain Person Re-Identification". In: *IEEE Transactions on Image Processing* 29, pp. 8549–8560.
- Li, Dangwei et al. (2019a). "A Richly Annotated Pedestrian Dataset for Person Retrieval in Real Surveillance Scenarios". In: *IEEE Transactions on Image Processing* 28, pp. 1575–1590.
- Li, Yu-Jhe et al. (2019b). "Cross-Dataset Person Re-Identification via Unsupervised Pose Disentanglement and Adaptation". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7918–7928.
- Li, Zhengjia and Duoqian Miao (2021). "Sequential End-to-end Network for Efficient Person Search". In: *AAAI*.
- Liang, Chao et al. (2020). "Rethinking the competition between detection and ReID in Multi-Object Tracking". In: *ArXiv abs/2010.12138*.
- Liao, Shengcai et al. (2015). "Person re-identification by Local Maximal Occurrence representation and metric learning". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2197–2206.
- Lin, Yutian et al. (2019). "A bottom-up clustering approach to unsupervised person re-identification". In: *AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 2, pp. 1–8.
- Lin, Yutian et al. (2020). "Unsupervised Person Re-Identification via Softened Similarity Learning". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3387–3396.
- Liu, Xinchun et al. (2016). "A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance". In: *ECCV*.
- Liza Lin, Newley Purnell (2019). *A World With a Billion Cameras Watching You Is Just Around the Corner*. URL: <https://www.wsj.com/articles/a-billion->

- surveillance - cameras - forecast - to - be - watching - within - two - years - 11575565402 (visited on 05/01/2022).
- Maaten, Laurens van der and Geoffrey E. Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9, pp. 2579–2605.
- Oord, Aäron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation Learning with Contrastive Predictive Coding". In: *ArXiv abs/1807.03748*.
- Peng, Peixi et al. (2016). "Unsupervised Cross-Dataset Transfer Learning for Person Re-identification". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1306–1315.
- Qi, Lei et al. (2019). "A Novel Unsupervised Camera-Aware Domain Adaptation Framework for Person Re-Identification". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8079–8088.
- Ren, Shaoqing et al. (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, pp. 1137–1149.
- Ristani, Ergys et al. (2016). "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking". In: *ArXiv abs/1609.01775*.
- Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015). "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823.
- Shahbaz, Ajmal and Kang-Hyun Jo (2021). "Enhanced Unsupervised Change Detector for Industrial Surveillance Systems". In: *IEEE Transactions on Industrial Electronics* 68, pp. 8973–8981.
- Silva, Thalles Santos (2020). "Exploring SimCLR: A Simple Framework for Contrastive Learning of Visual Representations". In: <https://sthalles.github.io>.
- Tahir, Rabia (2019). "Multi-domain Cross-dataset and Camera Style Transfer for Person Re-Identification". In: *International Journal of Advanced Trends in Computer Science and Engineering*.
- Tang, Qing, Ge Cao, and Kang-Hyun Jo (2021). "Fully Unsupervised Person Re-Identification via Multiple Pseudo Labels Joint Training". In: *IEEE Access* 9, pp. 165120–165131. DOI: [10.1109/ACCESS.2021.3134181](https://doi.org/10.1109/ACCESS.2021.3134181).
- Tang, Qing and Kang-Hyun Jo (2021). "Unsupervised Person Re-Identification Via Nearest Neighbor Collaborative Training Strategy". In: *2021 IEEE International*

- Conference on Image Processing (ICIP)*, pp. 1139–1143. DOI: [10.1109/ICIP42928.2021.9506109](https://doi.org/10.1109/ICIP42928.2021.9506109).
- Wang, Cheng et al. (2020). “TCTS: A Task-Consistent Two-Stage Framework for Person Search”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wang, Dongkai and Shiliang Zhang (2020). “Unsupervised Person Re-Identification via Multi-Label Classification”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10978–10987.
- Wang, Feng and Huaping Liu (2021). “Understanding the Behaviour of Contrastive Loss”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2495–2504.
- Wang, Menglin et al. (2021). “Camera-aware Proxies for Unsupervised Person Re-Identification”. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.
- Wei, Longhui et al. (2018a). “Person Transfer GAN to Bridge Domain Gap for Person Re-identification”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 79–88.
- (2018b). “Person Transfer GAN to Bridge Domain Gap for Person Re-identification”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 79–88.
- Wu, Chao-Yuan et al. (2017). “Sampling Matters in Deep Embedding Learning”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Wu, Zhirong et al. (2018). “Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination”. In: *CoRR* abs/1805.01978. arXiv: [1805.01978](https://arxiv.org/abs/1805.01978). URL: <http://arxiv.org/abs/1805.01978>.
- Xiao, Tong et al. (2017). “Joint Detection and Identification Feature Learning for Person Search”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3376–3385.
- Xu, Yuanlu et al. (2014). “Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness”. In: *Proceedings of the 22nd ACM international conference on Multimedia*.
- Xuan, Shiyu and Shiliang Zhang (2021a). “Intra-Inter Camera Similarity for Unsupervised Person Re-Identification”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11921–11930.

- Xuan, Shiyu and Shiliang Zhang (2021b). "Intra-Inter Camera Similarity for Unsupervised Person Re-Identification". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11921–11930.
- Yan, Yichao et al. (2021). "Efficient Person Search: An Anchor-Free Approach". In: *ArXiv abs/2109.00211*.
- Yang, Fengxiang et al. (2021). "Joint Noise-Tolerant Learning and Meta Camera Shift Adaptation for Unsupervised Person Re-Identification". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4853–4862.
- Yang, Jiachen et al. (2017). "A Fast Image Retrieval Method Designed for Network Big Data". In: *IEEE Transactions on Industrial Informatics* 13, pp. 2350–2359.
- Yu, Hong-Xing, Ancong Wu, and Weishi Zheng (2017). "Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 994–1002.
- (2020). "Unsupervised Person Re-Identification by Deep Asymmetric Metric Embedding". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, pp. 956–973.
- Yu, Hong-Xing et al. (2019). "Unsupervised Person Re-Identification by Soft Multilabel Learning". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2143–2152.
- Yu, Jongmin and Hyeontaek Oh (2021). "Unsupervised Vehicle Re-Identification via Self-supervised Metric Learning using Feature Dictionary". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3806–3813.
- Zeng, Kaiwei et al. (2020). "Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13654–13662.
- Zhang, Min-Ling and Zhi-Hua Zhou (2014). "A Review on Multi-Label Learning Algorithms". In: *IEEE Transactions on Knowledge and Data Engineering* 26, pp. 1819–1837.
- Zhang, Xinyu et al. (2019). "Self-Training With Progressive Augmentation for Unsupervised Cross-Domain Person Re-Identification". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8221–8230.
- Zhang, Xinyu et al. (2021). "Diverse Knowledge Distillation for End-to-End Person Search". In: *AAAI*.



- Zhang, Yaqing, Xi Li, and Zhongfei Zhang (2021). “Efficient Person Search via Expert-Guided Knowledge Distillation”. In: *IEEE Transactions on Cybernetics* 51, pp. 5093–5104.
- Zhang, Ying et al. (2018). “Deep Mutual Learning”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328.
- Zheng, Liang et al. (2015a). “Scalable Person Re-identification: A Benchmark”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124.
- (2015b). “Scalable Person Re-identification: A Benchmark”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124.
- Zheng, Liang et al. (2017). “Person Re-identification in the Wild”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3346–3355.
- Zheng, Zhedong, Liang Zheng, and Yi Yang (2017). “Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3774–3782.
- Zhong, Zhun et al. (2018). “Generalizing a Person Retrieval Model Hetero- and Homogeneously”. In: *ECCV*.
- Zhong, Zhun et al. (2019). “Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-Identification”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 598–607.
- Zhou, Kaiyang et al. (2019). “Omni-Scale Feature Learning for Person Re-Identification”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3701–3711.
- Zhou, Qinqin et al. (2020). “Fine-Grained Spatial Alignment Model for Person Re-Identification With Focal Triplet Loss”. In: *IEEE Transactions on Image Processing* 29, pp. 7578–7589.