electronics

MDPI

Article

FCL: Pedestrian Re-Identification Algorithm Based on Feature Fusion Contrastive Learning

Yuangang Li [1], Yuhan Zhang [2], Yunlong Gao [2], Bo Xu [3] and Xinyue Liu [2,*]

[1] Faculty of Business Information, Shanghai Business School, Shanghai 200235, China; yuangang.li@sbs.edu.cn
[2] School of Software, Dalian University of Technology, Dalian 116024, China; octopuszyh119@163.com (Y.Z.); yunlong@mail.dlut.edu.cn (Y.G.)
[3] School of Computer Science, Dalian University of Technology, Dalian 116024, China; xubo@dlut.edu.cn
* Correspondence: xyliu@dlut.edu.cn

Abstract: Pedestrian re-identification leverages computer vision technology to achieve cross-camera matching of pedestrians; it has recently led to significant progress and presents numerous practical applications. However, current algorithms face the following challenges: (1) most of the methods are supervised, heavily relying on specific datasets, and lacking robust generalization capabilities; (2) it is hard to extract features because the elongated and narrow shape of pedestrian images introduces uneven feature distributions; (3) the substantial imbalance between positive and negative samples. To address these challenges, we introduce a novel pedestrian re-identification unsupervised algorithm called Feature Fusion Contrastive Learning (FCL) to extract more effective features. Specifically, we employ circular pooling to merge network features across different levels for pedestrian re-identification to improve robust generalization capability. Furthermore, we propose a feature fusion pooling method, which facilitates a more efficient distribution of feature representations across pedestrian images. Finally, we introduce FocalLoss to compute the clustering-level loss, mitigating the imbalance between positive and negative samples. Through extensive experiments conducted on three prominent datasets, our proposed method demonstrates promising performance, with an average 3.8% improvement in FCL's mAP indicators compared to baseline results.

Keywords: re-identification algorithm; feature fusion; contrastive learning

check for updates

Citation: Li, Y.; Zhang, Y.; Gao, Y.; Xu, B.; Liu, X. FCL: Pedestrian Re-Identification Algorithm Based on Feature Fusion Contrastive Learning. *Electronics* **2024**, *13*, 2368. https://doi.org/10.3390/electronics13122368

Academic Editor: Silvia Liberata Ullo

Received: 27 May 2024
Revised: 11 June 2024
Accepted: 14 June 2024
Published: 17 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Pedestrian re-identification is defined as the task of retrieving the images matching a person across non-overlapping cameras based on a specific query sample, which is significant for some application scenarios including surveillance and security. Guided by manual annotations, the supervised pedestrian re-identification methods [1–3] have made remarkable achievements thanks to the boom in deep learning. However, the pedestrian re-identification task still faces challenges. First, existing supervised algorithms predominantly rely on mapping data to associated labels within a dataset to learn image features. Consequently, their performance remains confined to specific datasets. Furthermore, with the limited quantity and scale of current datasets, acquiring massive manual annotations is costly, which is not feasible in real application scenarios. Therefore, to attain robust performance in real-world scenarios, unsupervised pedestrian re-identification has attracted more and more attention in recent years.

Existing unsupervised pedestrian re-identification approaches can be classified into two categories: (1) unsupervised domain adaptation (UDA) methods and (2) fully unsupervised learning (FUL) methods. UDA methods [4,5] take advantage of the unsupervised domain adaptation, which transfers the label information from the labeled source dataset to the unlabeled target dataset. This approach enhances the network's generalization capability and addresses the challenge of dataset transfer in pedestrian re-identification. However, the performance of these UDA pedestrian re-identification methods depends on

the quality of the source dataset, which means that these methods are not data-friendly. FUL methods [3,6] solely rely on unlabeled datasets, aligning more closely with practical deployment requirements, which are more scalable and flexible because of the lack of manual involvement. Most of these methods utilize pseudo-labels to train the network, typically employing clustering algorithms to generate pseudo-labels for training purposes.

Recently, contrastive learning has shown impressive progress in unsupervised visual representation learning [7,8], which capitalizes on the inherent information within the data for network training. These methods revolve around the automatic construction of similar and dissimilar instances to acquire a feature model that brings similar instances closer together in the projection space while simultaneously pushing dissimilar instances farther apart. Some existing works [9–12] have introduced contrastive learning to the field of unsupervised pedestrian re-identification. In [9], the clustering and the contrastive learning are performed separately within and between cameras to mine the intra-camera and inter-camera similarities. In [12], different transformed versions of the original features are drawn to the corresponding cluster centroid. However, these contrastive learning methods usually focus only on mining the intra-class similarity, neglecting the negative effects of clustering noise, as demonstrated in the case of Xiao et al. [13], and the construction of effective positive samples assumes paramount importance in contrastive learning techniques.

We analyzed the influence of negative samples in two aspects. On the one hand, it is difficult to extract features because, in the context of pedestrian re-recognition, the images in the dataset are usually surrounded by bounding boxes obtained from the object detection process. As depicted in Figure 1, the pedestrian image possesses narrow rectangular shapes with a small aspect ratio (width/height). This characteristic often engenders an uneven collection of features during the extraction process, wherein one direction exhibits a dense feature distribution while the other direction displays a comparatively sparse distribution. We simulated the spatial distribution of features within the image. The extracted features exhibit a pronounced horizontal clustering, which accentuates the presence of narrow edges. However, existing local feature-based methods for pedestrian re-identification primarily rely on image segmentation guided by the vertical arrangement of different body parts in pedestrians. This horizontal concentration of features impedes the effective extraction of pedestrian characteristics and undermines the accuracy of detection algorithms. To tackle this challenge, we propose a feature fusion pooling method, which facilitates a more efficient distribution of feature representations across pedestrian images compared to conventional simplistic image enhancement techniques.

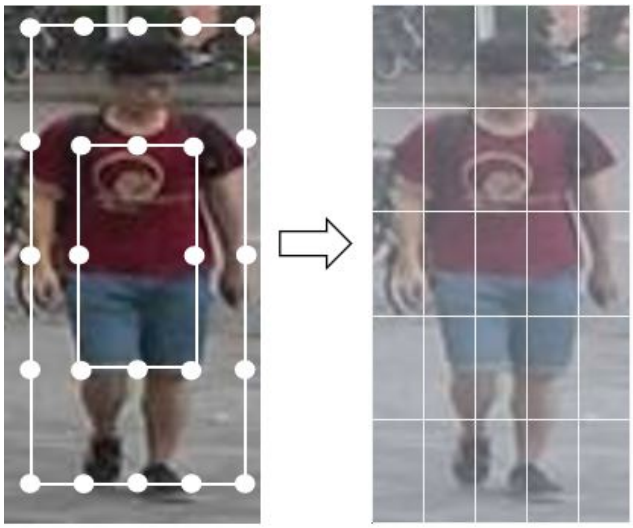


Figure 1. Feature distribution of a pedestrian. The feature distribution in the horizontal direction is very dense while the vertical direction is relatively sparse.

On the other hand, the substantial imbalance between positive and negative samples presents challenges in enhancing detection accuracy. In the realm of pedestrian re-identification, positive samples that pertain to pedestrians are notably scarce in comparison to the profusion of irrelevant samples and interfering factors. This disparity results in an overwhelming surplus of negative samples relative to positive ones. Consequently, in scenarios where the proportion of negative samples significantly surpasses that of positive samples, and most negative samples are straightforward and markedly dissimilar to positive samples, the learning process of the model becomes dominated by these simpler samples. Despite the relatively minor loss value associated with simple negative samples, their sheer abundance profoundly impacts the convergence of the model. Therefore, instead of conventional cross-entropy functions, the decision was made to employ the FocalLoss function. This function dynamically balances the proportion of positive and negative samples while adjusting their respective contributions to the loss.

The principal contributions of this study can be summarized as follows:

- We propose a pedestrian re-identification algorithm, namely, Feature Fusion Contrastive Learning (FCL). With a proposed feature fusion pooling method, FCL facilitates a more efficient distribution of feature representations across pedestrian images.
- To address the imbalance in the proportion of positive and negative samples, we introduce FocalLoss, which computes the comparative loss between sample features and their corresponding cluster centers.
- To validate the effectiveness of our network algorithm, we conducted experiments on three benchmark datasets: Market1501, DukeMTMC-reID, and MSMT17. Through comprehensive comparisons with a series of algorithms, we demonstrate the superiority of FCL.

2. Related Work

2.1. Pedestrian Re-Identification

With the impressive potential of deep neural networks and the guidance of manual annotation, many practical applications [14–16] have sprung up. Among them, pedestrian re-identification has received much attention because it is concerned with surveillance and security in some application scenarios. At present, supervised pedestrian re-identification methods [1–3,17] have achieved promising performance. Specifically, Si et al. [18] proposed a novel Spatial-Driven Network (SDN) to capture discriminative features with abundant semantic information. Wang et al. [19] designed a joint similarity metric to mine the visual semantic information and spatial-temporal information simultaneously. To be more suitable for practical applications, unsupervised pedestrian re-recognition has been proposed and widely studied.

Unsupervised pedestrian re-identification is more scalable and flexible in real-world application scenarios due to its low requirement for manual annotation. These unsupervised methods can be broadly classified into two categories: (1) The unsupervised domain adaptation (UDA) methods necessitate labeled source datasets and unlabeled target datasets. Ge et al. [20] proposed a combined loss function to co-train with samples from the source and target domains and the merging memory bank. Ganin et al. [21] proposed to maximize the inter-domain classification loss and minimize the intra-domain classification loss to learn domain robust features. However, UDA methods are limited by the requirement of the target dataset having a distribution close to that of the source dataset. (2) The fully unsupervised learning (FUL) method solely relies on unlabeled datasets, aligning more closely with the algorithm's implementation requirements. Most FUL methods utilize pseudo-labels to train the network, and various clustering methods such as DBScan [22] and KNN [23] are autonomously employed. Ji et al. [24] exploited a Graph Convolutional Network (GCN) to estimate the pseudo-labels of image pairs. Zeng et al. [5] proposed to learn robust pseudo-labels by integrating the hierarchical clustering and the hard-batch triplet loss into a unified framework. Lin et al. [25] proposed a classification model trained with softened labels to estimate the pairwise similarity. These methods do not fully pay

attention to the influence of negative samples, including the elongated and narrow shape of pedestrian images and the serious imbalance between positive and negative samples. We propose a feature fusion pooling method to facilitate more efficient distribution of feature representations and introduce FocalLoss to compute the clustering-level loss, mitigating the imbalance between positive and negative samples.

2.2. Contrastive Learning

In recent years, with the development and application of the Siamese network, contrastive learning has begun to emerge in the field of unsupervised learning. Different from transfer learning [26], contrastive learning aims at learning a good image representation. Contrastive learning treats the input image and its augmented representations as positive samples, while all other diverse samples in the dataset are regarded as negative. Initially proposed in the paper by Ye et al. [27], the fundamental premise of contrastive learning is that similar images can be grouped together based on the similarity of their features, rather than predefined labels. The objective is to train a network to encode similar data in a manner that maximizes the differences between the encodings of different types of data, facilitating the discrimination of each image.

Currently, research on contrastive learning addresses several key aspects: the construction of similar positive and negative samples, the design of models guided by these principles, the development of effective loss functions, and strategies to prevent model collapse. Regarding methods to prevent model collapse, existing approaches can be broadly categorized into four types: (1) negative-examples-based contrastive learning methods, exemplified by MOCO [28] and SimClR [29]; (2) contrastive-clustering-based methods, such as SWAC [30]; (3) asymmetric network structures-based methods, such as BYOL [31]; and (4) redundancy elimination loss functions-based methods. Moreover, certain methods focus on handling difficult negative samples, as demonstrated by works like [32], and leverage data augmentation techniques [13]. Our work is a contrastive-learning-based framework, operating on the principle of automatically constructing similar and dissimilar instances to train a model, which improves the performance of pedestrian re-identification algorithms and mitigates issues related to dataset limitations and inadequate generalization capabilities. It should be noted that our work is more for pedestrian re-recognition tasks, including the proposed feature fusion pooling method to help extract features, and we designed FocalLoss to mitigate the imbalance between positive and negative samples.

3. Feature Fusion Contrastive Learning

3.1. Overall

FCL is a contrastive-learning-based framework, learning on unlabeled datasets by forming pseudo-labels through clustering and updating them through momentum updating to guide network training, as shown in Figure 2. Specifically, for the input images, FCL first uses two ResNet50 pre-trained on ImageNet to extract features by the data augmentation and feature fusion branches, respectively. Secondly, FCL generates pseudo-labels by data augmentation and fuses features from different levels of the network by the feature fusion branch. Third, FCL optimizes the cluster-level and instance-level loss functions to explore deeper clustering information, To alleviate the problem of an unequal proportion of positive and negative samples, we use the FocalLoss loss function as a cluster-level loss function. The feature extractor ResNet50 and data augmentation methods are widely used. The following mainly introduces the pseudo-label generation method, the feature fusion pooling operation, and the designed loss function.

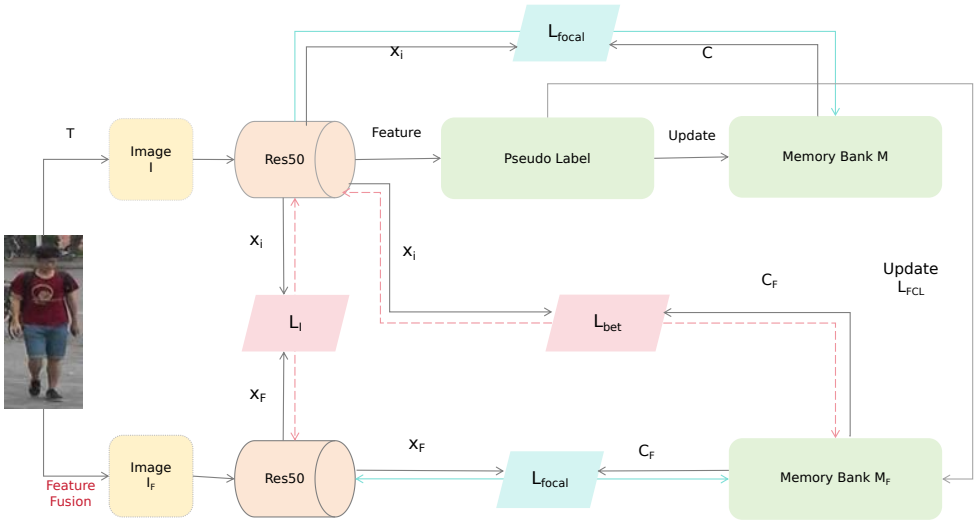


Figure 2. Architecture of Feature Fusion Contrastive Learning framework.

3.2. Pseudo-Label Generation

FCL is an unsupervised method that requires pseudo-labels to be generated on samples to train feature encoders. Following CACL [10], FCL uses data augmentation with clustering to generate pseudo-labels. We represent an image dataset containing N samples without manually labeling as $I = \{I_i\}_{i=1}^{N}$. We generate two samples $T(I)$ and $F(I)$ from the image $I_i \in I$ through different data augmentation strategies and send them into two branches as inputs, where $T(I)$ represents conventional data augmentation methods, including random inversion, random cropping, random erasure, etc. $F(I)$ represents the feature fusion data augmentation method. We represent the output features of the data augmentation branch and the feature fusion branch as X_i and X_F.

To generate pseudo-labels, we use the output feature $X = \{x_1, x_2 \dots x_N\}$ from the data augmentation branch for clustering. Based on the results, we obtain pseudo-labels $Y = \{y_1, y_2, \dots, y_N\}$, which are used to integrate clustering information into contrastive learning and guide the training. In addition, following contrastive learning's setting, we use instance memory banks $M = \{c_i\}_{i=1}^{N}$ and $M_F = \{c_{Fi}\}_{i=1}^{N}$ to store the output features of each branch, respectively. They are initialized by outputs from network branches, and the momentum updates the pseudo-label information in the memory banks using Equations (1) and (2), where γ is a hyperparameter and defaults to 0.2.

$$c_i^t = \gamma c_i^{t-1} + (1-\gamma)x_i \tag{1}$$

$$c_{Fi}^t = \gamma c_{Fi}^{t-1} + (1-\gamma)x_{Fi} \tag{2}$$

3.3. Feature Fusion

As analyzed in the introduction, the data in pedestrian re-identification images are narrow and long, and the extracted features are unevenly distributed, emphasizing horizontal features. However, in pedestrian re-identification, vertical features are more helpful for algorithm calculation. Therefore, we use a special pooling method to alleviate the problem of uneven features.

Unlike traditional pooling methods based on rows and columns within the feature domain, we perform circular pooling as shown in Figure 3. This pooling method arranges pooling features along different square circular spaces. The basic idea is to pool the feature samples of the outer circle from the low level, while the feature samples of the inner circle come from the high level. This multi-level pooling method helps to more fully utilize the low-level and high-level features of the network. This is because in the feature area,

outer ring features that are further away from the center are usually more sensitive to the geometric transformations of the object. Therefore, at these locations, we tend to use low-level features to maximize their advantages in contour features, which are incorporated into the feature map through pooling of the outermost circle. On the contrary, the features in the central part have a larger receptive field and focus more on high-level features for object-level classification, which are more helpful in distinguishing images semantically.

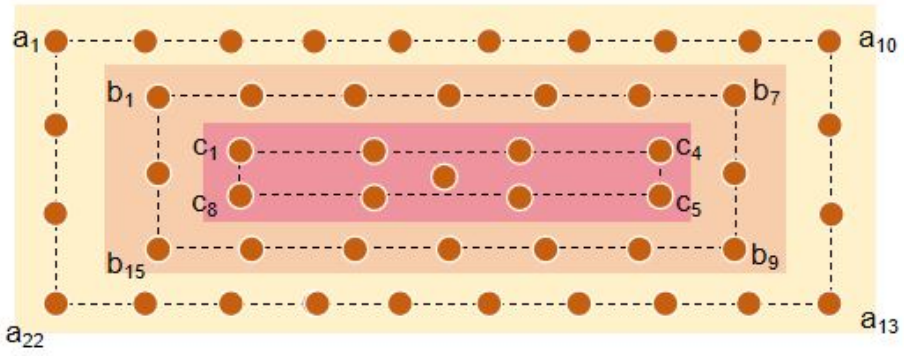


Figure 3. Architecture of square-shaped circles pooling.

Specifically, we perform pooling operations along the shape of a square circle. Unlike traditional max pooling and average pooling methods, we use bilinear interpolation methods. As shown in the brown feature points in Figure 3, we sample a set of feature points uniformly spaced along each square circle. This is because bilinear interpolation can avoid the harmful misaligned rough quantization that may exist between the image and the extracted features. This process ensures the complete output of fixed-size feature values. Taking Figure 3 as an example, three sets of values, $a_1 - a_{24}$, $b_1 - b_{16}$, and $c_1 - c_8$, are sampled from different circles. The space between sampling points along each circle depends on the circumference of the circle, and the distribution of each group of sampling points varies according to their aspect ratio. It is precisely this kind of sampling within the circle that promotes the performance of shape adaptation. To ensure that, the three circles are evenly distributed within the proposal.

In addition, we fuse features from multiple network levels through pooling to obtain rich learned features. This is because the features learned by the shallow and deep layers of convolutional neural networks are different. Deep features are not sensitive to intra-class variations, such as geometric distortions and small displacements. Therefore, this non-interference feature is very beneficial for improving task performance when dealing with classification-matching tasks, we use a multi-layer feature fusion method as one of the data augmentation methods in both network branches to fuse the different levels of features of the image when it passes through the neural network. As shown in Figure 4, directly inputting raw feature maps of different levels from deep convolutional networks can lead to performance degradation. As mentioned in the paper [33], the reason for this situation may be due to the large semantic differences between different levels. Therefore, we "smooth" these differences by performing 2 $\times$ upsampling on higher levels and 3 $\times$ 3 convolution between higher and lower levels.

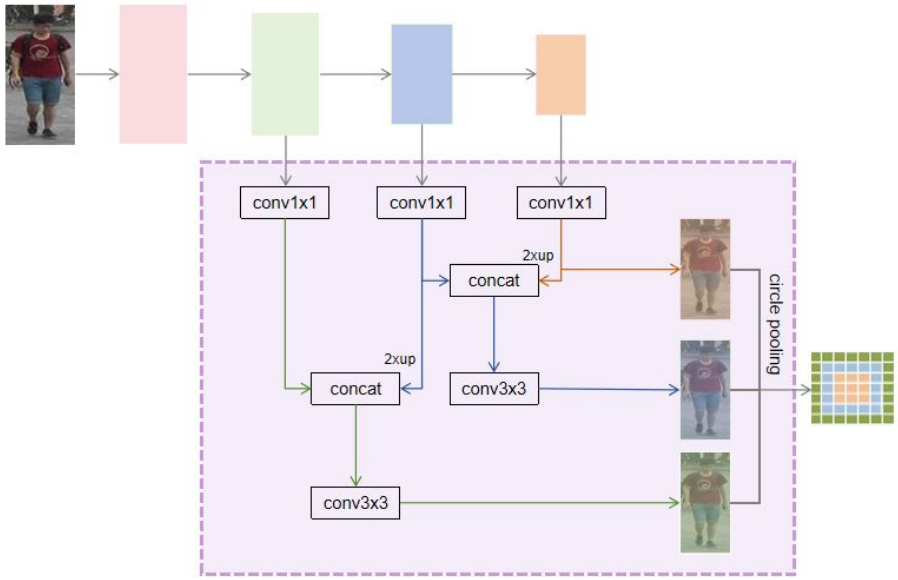


Figure 4. Architecture of multiple network levels of feature fusion.

3.4. Loss Function

Through the data augmentation and feature fusion branches, we obtain two outputs X_i and X_F, and the pseudo-labels are obtained using the clustering method. To better utilize the invariance between two enhanced views, we calculate the instance-level loss and cluster-level comparison loss separately on the loss function.

3.4.1. Instance-Level Contrastive Loss

The instance-level contrastive loss calculates the difference between the output features of the two branch networks. To minimize this loss, it helps to pull in the distance of features obtained from the same image through different data enhancement methods, so that the model can learn to extract the class-related features better. The formula is as follows:

$$L_I = -\frac{x_i^T}{\|x_i\|_2}\frac{x_F}{\|x_F\|_2}, \tag{3}$$

where x_i is the feature output in the data augmentation branch of the network, and x_F is the feature output in the feature fusion branch of the network.

3.4.2. Cluster-Level Contrastive Loss

Cluster-level contrastive learning is aimed at learning the relationship between features and their corresponding cluster centers. Following CACL [10], we divide the cluster-level loss into intra-view loss L_{in} and inter-view loss L_{bet}.

- Inter-view loss L_{in} refers to the loss between features and cluster centers after the same data augmentation. Due to the substantial imbalance between positive and negative samples, positive samples that pertain to pedestrians are notably scarce in comparison to the profusion of irrelevant samples and interfering factors. This disparity results in an overwhelming surplus of negative samples relative to positive ones, making it difficult for the model to learn the characteristics of the positive samples. Therefore, we chose to use FocalLoss to adaptively balance the proportion of positive and negative samples and adjust the contribution of samples to the loss. The FocalLoss formula is as follows:

$$\begin{aligned} FocalLoss &= -(1-p_t)^{\gamma}\log p_t \\ &= -y(1-p_t)^{\gamma}\log p_t - (1-y)p_t^{\gamma}\log(1-p_t), \end{aligned} \tag{4}$$

where p_t is calculated by the following,

$$p_t = \begin{cases} -p & if \;\; y = 1 \\ 1-p & otherwise \end{cases}. \tag{5}$$

This loss introduces $(1-p_t)$ as a modulation coefficient to reduce the weight of easily distinguishable negative samples, making the model more focused on difficult-to-classify samples during training. Specifically, when p_t tends to 1, it indicates that the sample is easily distinguishable and has a small contribution to the loss. At this point, $1-p_t^{gamma}$ tends to 0, which reduces the proportion for easily distinguishable samples. When p_t is very small, the sample is misclassified into other categories, and the modulation factor $1-p_t^{gamma}$ tends to 1, without affecting the calculation of the loss. Compared to the cross-entropy loss, FocalLoss does not change the loss for hard negative samples but decreases the loss for easily distinguishable samples. With p_t, FocalLoss can adjust the weight of positive and negative samples, as well as control the weight of difficult-to-classify samples.
Since the FocalLoss function can reduce the impact of substantial imbalance between positive and negative samples, we calculate inter-view loss L_{in} as follows:

$$L_{in} = -(1-p_i)^4\log p_i - (1-p_{Fi})^4\log p_{Fi}, \tag{6}$$

where p_i and p_{Fi} are defined as:

$$p_i = \frac{exp\left(c_{(I_i)}^T x_i/\tau\right)}{\sum_{l=1}^{m} exp\left(c_l^T x_i/\tau\right)}, \tag{7}$$

$$p_{Fi} = \frac{exp\left(c_{(F_i)}^T x_{Fi}/\tau\right)}{\sum_{l=1}^{m'} exp\left(c_{(F_i)}^T x_{Fl}/\tau\right)}, \tag{8}$$

where c_l and c_{F_l} are the center vectors of the lth cluster of the data augmentation and feature fusion branches of the network:

$$c_l = \frac{1}{|C^l|}\sum_{I_i \in C^l} v_i, \tag{9}$$

$$c_{Fi} = \frac{1}{|C^l|}\sum_{I_i \in C^l} v_{Fi}, \tag{10}$$

where v_i and v_{Fi} refer to the features belonging to the same cluster stored in M and M_F, while C^l refers to the total number of clusters. We calculate the cluster center features by summing and averaging all features.

- Intra-view loss L_{bet} is to pull the distance between the image representation and the enhanced representation in response to the image. Considering the calculation cost, we calculate the picture representation and the corresponding clustering center as follows:

$$L_{bet} = -\frac{x_i^T}{\|x_i\|_2}\frac{c_{Fi}}{\|c_{Fi}\|_2} \tag{11}$$

where x_i is the output feature in the data augmentation branch of the network, and c_{Fi} is the cluster center stored in M_F.

Finally, to summarize the loss functions, we use our FCL network loss function L_{FCL}, denoted as:

$$L_{FCL} = L_{bet} + L_{in} + L_I. \quad (12)$$

4. Experiments

4.1. Dataset

We conducted experiments on the mainstream datasets Market-1051, DukeMTMC-reID, and MSMT17 in pedestrian re-identification. The basic information is shown in the Table 1.

Table 1. Mainstream dataset in pedestrian re-identification.

Dataset	Training Set		Test Set		Cam
	IDs	Images	IDs	Images	
Market-1051	751	12,936	750	19,732	6
DukeMTMC-reID	702	16,522	702	19,889	8
MSMT17	4101	32,621	3060	93,820	15

Market-1501: This dataset contains a total of 751 IDs, with 12,936 images used for training and 750 images used for testing, totaling 19,732 images. All images were taken by five high-resolution cameras and one low-resolution camera.

DukeMTMC-reID: The dataset was collected from eight cameras, and each identity was guaranteed to be observed by at least two cameras. This dataset contains a total of 1404 IDs where 16,522 images with 702 identities were selected as the training set, and 19,889 images with the remaining 702 identities were selected as the test set.

MSMT17: The dataset contains 126,441 images of 4101 identities captured by 15 cameras and currently is the largest ReID dataset. In practice, 32,621 images with 1041 identities were used for training, and 93,820 images with 3060 identities were tested.

4.2. Baslines

We compare our proposed FCL to the state-of-the-art unsupervised domain adaptation methods and fully unsupervised methods for pedestrian re-identification. The unsupervised domain adaptation methods for pedestrian re-identification include: SpCL [20], NRMT [34], MMT [35], and MetaCam [36]. The fully unsupervised methods for pedestrian re-identification include: ITCS [9], CAP [37], PPLR [38], RLCC [39], ICE [40], CIFL [41], GRACL [11], CACL [10], PLRIS [42], and LRMGFS [43].

4.3. Experimental Settings

FCL was implemented using pytorch 1.12 and used ResNet50, which was pre-trained on ImageNet, as the backbone network for the two branches. We initialized two memory banks with pre-trained outputs from network branches on ImageNet and updated them with momentum updates in Equation (12). Following CACL [10], we chose the Adam optimizer to optimize the network and set the weight decay to 0.0005. We trained the network for a total of 80 epochs and set the batch size to 256. The initial learning rate was set to 0.00035 and decreased to one-tenth of the original every 20 epochs. The temperature coefficient τ in Equation (7) was set to 0.05, and the update factor γ in Equations (1) and (2) was set to 0.2. All of our experiments were conducted on a Ubuntu 22.04.1 system with four 24 G Nvidia TITAN RTX GPUs and Intel(R) Xeon(R) Gold 6248R CPU @ 3.00 GHz (five epochs per hour). The results of the experiment are the average of the five seeds.

4.4. Overall Performance

Results on Market-1501 and DukeMTMC-reID datasets: To demonstrate the effectiveness of our proposed method, we first compared the mAP, Rank-1, and Rank-5 accuracy of FCL with several pedestrian re-identification methods on Market-1501 and DukeMTMC-

reID datasets. Table 2 summarizes the results of comparing different methods on these two datasets, the experimental performance of FCL is presented in the last row of the table. It is easy to observe that FCL performed well in all four benchmark tests, which proves the effectiveness of our method.

Table 2. Comparing with other methods on the MARKET-1501 and DUKEMTMC-REID datasets. Bold indicates the highest value.

Method	Market-1501			DukeMTMC-reID		
	mAP	**Rank-1**	**Rank-5**	**mAP**	**Rank-1**	**Rank-5**
SpCL [20] (NIPS20)	73.1%	88.1%	96.3%	65.3%	81.2%	90.3%
NRMT [34] (ECCV20)	71.7%	87.8%	94.6%	62.2%	77.8%	86.9%
MMT [35] (ICLR20)	73.8%	89.5%	96.0%	62.3%	76.3%	87.7%
ITCS [9] (CVPR21)	72.9%	89.5%	95.3%	64.4%	80.0%	89.0%
CAP [37] (AAAI21)	79.2%	91.4%	96.3%	67.3%	81.1%	89.3%
RLCC [39] (CVPR21)	77.7%	90.8%	96.3%	69.2%	83.2%	91.6%
MetaCam [36] (CVPR21)	61.7%	83.9%	92.3%	53.8%	73.8%	84.2%
ICE [40] (ICCV21)	82.3%	93.8%	97.6%	69.6%	83.3%	91.5%
CIFL [41] (TMM22)	82.3%	**93.8%**	97.6%	69.6%	83.3%	91.5%
CACL [10] (TIP22)	80.9%	92.7%	97.4%	69.6%	82.6%	91.2%
GRACL [11] (TCSVT22)	83.7%	93.2%	-	-	-	-
PPLR [38] (CVPR22)	81.5%	92.8%	97.1%	-	-	-
PLRIS [42] (ICIP23)	83.2%	93.1%	-	-	-	-
FCL	**83.7%**	**93.8%**	**97.9%**	**70.3%**	**83.3%**	**92.1%**

Result on the MSMT17 dataset: The MSMT17 dataset is currently the largest dataset, with more IDs and more images than the previous dataset. Table 3 summarizes the latest baseline and FCL results (the mAP and Rank-1 accuracy). It is easy to find that the performance of FCL exceeds that of other methods.

Analysis: Through the experiments of the above three datasets, we can see the strong results of FCL. This is easy to understand because FCL considers the effect of negative samples in unsupervised methods. We paid attention to the particularity of pedestrian images and extracted different scale features of pedestrians by using our feature fusion module. Considering the imbalance of positive and negative samples, we used the Focal-Loss function to balance the influence of positive and negative samples, so that the model could better learn how to distinguish positive samples.

Table 3. Comparing with other methods on the MSMT dataset. Bold indicates the highest value.

Method	MSMT17	
	mAP	**Rank-1**
SpCL [20] (NIPS20)	26.8%	53.7%
MMT [35] (ICLR20)	24.0%	50.1%
RLCC [39] (CVPR21)	27.9%	56.5%
LRMGFS [43] (JEMI23)	27.4%	28.4%
FCL	**30.8%**	**58.1%**

4.5. Ablation Study

In order to explore the effectiveness of each component of FCL, we conducted comprehensive ablation studies on the performance of FCL in both loss function and data augmentation methods.

The impact of FocalLoss: Firstly, we compared the performance improvement of the algorithm with FocalLoss. As shown in Table 4, it is easy to see that the overall performance of the model deteriorates significantly without FocalLoss. To be specific, with FocalLoss,

FCL's rank-1 increased by 1.1%, and the average accuracy mean mAP increased by 1.4% on the Market-1501 dataset. And on the DukeMTMC-ReID dataset, rank-1 improved by 0.4% and the average accuracy mean mAP improved by 0.5%. On the MSMT17 dataset, rank-1 improved by 3.2% and the average accuracy mean mAP improved by 4.3%. This is illustrated by the effect of the positive and negative sample imbalance on performance and also proves the effectiveness of FocalLoss.

Table 4. Ablation study of the FocalLoss. Bold indicates the highest value.

Method	Market-1501			DukeMTMC-reID			MSMT17		
	mAP	**Rank-1**	**Rank-5**	**mAP**	**Rank-1**	**Rank-5**	**mAP**	**Rank-1**	**Rank-5**
FCL	**83.7%**	**94.1%**	**97.9%**	**70.3%**	**83.3%**	**92.1%**	**30.8%**	**58.1%**	**72.1%**
-FocalLoss	82.4%	93.0%	97.5%	69.8%	82.9%	91.6%	26.5%	54.9%	67.9%

The impact of feature fusion: We conducted ablation experiments on our feature fusion as shown in Table 5. It can be seen that the performance was significantly improved by using the feature fusion method, proving its effectiveness. Specifically, compared with FCL without our feature fusion, FCL improved the mAP by 1.7% on average and rank-1 by 0.8% on average on the three datasets.

Table 5. Ablation study of feature fusion. Bold indicates the highest value.

Method	Market-1501			DukeMTMC-reID			MSMT17		
	mAP	**Rank-1**	**Rank-5**	**mAP**	**Rank-1**	**Rank-5**	**mAP**	**Rank-1**	**Rank-5**
FCL	**83.7%**	**94.1%**	**97.9%**	**70.3%**	**83.3%**	**92.1%**	**30.8%**	**58.1%**	**72.1%**
-Feature Fusion	81.6%	92.8%	97.6%	69.7%	82.8%	91.7%	29.4%	57.5%	69.3%

The impact of the feature fusion's various components: To further explore the function of our feature fusion, we conducted a series of ablation experiments on Market-1501, as shown in Table 6. Firstly, we compared our method with another multi-scale fusion method as shown in the first and second rows of Table 6, where "channel concat" represents the direct connection of multi-level features after pooling. In this case, the pooling features at lower and higher levels completely overlap at each position, increasing the total number of feature channels. "opp" represents the opposite arrangement to our method, where internal feature samples come from lower levels, while external samples are pooled from higher levels. It can be seen that compared to FCL, its performance decreased, further proving the effectiveness of our method.

As shown in the third and fifth rows of Table 6, "typ" represents maximum pooling operation, and "cir" represents the square-shaped-circle pooling we used. It can be seen that when eliminating other interference and only using pooling for a single level, the performance of the model improved to a certain extent, resulting in higher rank-1 and mAP values. In the fourth and fifth rows of Table 6, when fusing features from multiple network levels, the performance of the model also improved. Therefore, both parts of the feature fusion we proposed are effective, as shown in rows 3–6 of the Table 6. This is because the square-shaped-circle pooling operation ensures the integrity of the overall contour of pedestrians and integrates the features of pedestrians under different visual fields, and the multi-scale fusion integrates the fine and coarse-grained features.

Table 6. Ablation study of multi-scale and pooling fusion method. Bold indicates the highest value.

Method	Market-1501		
	mAP	**Rank-1**	**Rank-5**
FCL and channel concat	83.4%	93.8%	97.6%
FCL and opp	82.5%	92.6%	97.3%
Single and typ	79.7%	90.9%	96.5%
Multi and typ	82.3%	92.9%	97.5%
Single and cir	82.1%	92.3%	97.1%
Multi and cir(FCL)	**83.7%**	**94.1%**	**97.9%**

5. Conclusions

We proposed an unsupervised pedestrian re-identification algorithm based on Feature Fusion Contrastive Learning (FCL). FCL used a twin-branch ResNet50 network and employed clustering operations to obtain pseudo-labels. We used feature fusion as a data augmentation method. We used a circle pooling method and fused features extracted from different levels of the network to distribute feature representations on pedestrian images. For the loss function, we chose multiple loss functions to obtain more comprehensive and deeper clustering information to guide contrastive learning, including an instance-level loss and cluster-level loss. In order to overcome the imbalance in the proportion of positive and negative samples, we used the FocalLoss function to calculate the comparative loss between sample features and their corresponding cluster centers.

Although the continuous improvement of unsupervised methods has to some extent alleviated the problem of high integration costs for manually annotated data, there is still a lack of datasets in this field. The current datasets still face problems such as having a limited quantity, small-scale, and single-scene images, which are far from practical applications in real life and to some extent restrict the development and implementation of pedestrian re-identification. We believe that with the efforts of many scholars, pedestrian re-identification will eventually be applied in real life, just like facial recognition, providing convenience to our lives.

Author Contributions: Investigation, Y.L. and Y.Z.; methodology, Y.Z. and Y.G.; validation, Y.G. and B.X.; writing—original draft preparation, Y.Z.; writing—review and editing, X.L. and Y.L.; supervision, Y.L. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Liaoning Provincial Social Science Planning Fund (no. L22CTQ002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this paper are publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; Wu, F. Diverse Part Discovery: Occluded Person Re-identification with Part-Aware Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
2. Chen, X.; Fu, C.; Zhao, Y.; Zheng, F.; Song, J.; Ji, R.; Yang, Y. Salience-Guided Cascaded Suppression Network for Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
3. Ming, Z.; Zhu, M.; Wang, X.; Zhu, J.; Cheng, J.; Gao, C.; Yang, Y.; Wei, X. Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image Vis. Comput.* **2022**, *119*, 104394. [CrossRef]
4. Chen, Y.; Zhu, X.; Gong, S. Instance-Guided Context Rendering for Cross-Domain Person Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.

5. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
6. S, S.R.; Prasad, M.V.; Balakrishnan, R. Spatio-Temporal association rule based deep annotation-free clustering (STAR-DAC) for unsupervised person re-identification. *Pattern Recognit.* **2022**, *122*, 108287. [CrossRef]
7. Xie, J.; Zhan, X.; Liu, Z.; Ong, Y.S.; Loy, C.C. Delving into Inter-Image Invariance for Unsupervised Visual Representations. *Int. J. Comput. Vis.* **2022**, *130*, 2994–3013. [CrossRef]
8. Chen, X.; He, K. Exploring Simple Siamese Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
9. Xuan, S.; Zhang, S. Intra-inter camera similarity for unsupervised person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11926–11935.
10. Li, M.; Li, C.G.; Guo, J. Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. *IEEE Trans. Image Process.* **2022**, *31*, 3606–3617. [CrossRef]
11. Zhang, H.; Zhang, G.; Chen, Y.; Zheng, Y. Global relation-aware contrast learning for unsupervised person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8599–8610. [CrossRef]
12. Yu, H.X.; Wu, A.; Zheng, W.S. Unsupervised Person Re-Identification by Deep Asymmetric Metric Embedding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 956–973. [CrossRef] [PubMed]
13. Xiao, T.; Wang, X.; Efros, A.A.; Darrell, T. What should not be contrastive in contrastive learning. *arXiv* **2020**, arXiv:2008.05659.
14. Jawaharlalnehru, A.; Sambandham, T.; Sekar, V.; Ravikumar, D.; Loganathan, V.; Kannadasan, R.; Khan, A.A.; Wechtaisong, C.; Haq, M.A.; Alhussen, A.; et al. Target Object Detection from Unmanned Aerial Vehicle (UAV) Images Based on Improved YOLO Algorithm. *Electronics* **2022**, *11*, 2343. [CrossRef]
15. Khobdeh, S.B.; Yamaghani, M.R.; Sareshkeh, S.K. Basketball action recognition based on the combination of YOLO and a deep fuzzy LSTM network. *J. Supercomput.* **2023**, *80*, 3528–3553. [CrossRef]
16. Sharma, N.; Haq, M.A.; Dahiya, P.K.; Marwah, B.R.; Lalit, R.; Mittal, N.; Keshta, I. Deep Learning and SVM-Based Approach for Indian Licence Plate Character Recognition. *Comput. Mater. Contin.* **2023**, *74*, 881–895. [CrossRef]
17. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. *arXiv* **2017**, arXiv:1703.07737.
18. Si, T.; He, F.; Wu, H.; Duan, Y. Spatial-driven features based on image dependencies for person re-identification. *Pattern Recognit.* **2022**, *124*, 108462. [CrossRef]
19. Wang, G.; Lai, J.; Huang, P.; Xie, X. Spatial-Temporal Person Re-identification. *Proc. Aaai Conf. Artif. Intell.* **2019**, *33*, 8933–8940. [CrossRef]
20. Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; Li, H. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11309–11321.
21. Ganin, Y.; Lempitsky, V. *Unsupervised Domain Adaptation by Backpropagation*; PMLR: New York, NY, USA, 2019.
22. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; Volume 96, pp. 226–231.
23. Coates, A.; Ng, A.Y. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 561–580.
24. Ji, H.; Wang, L.; Zhou, S.; Tang, W.; Zheng, N.; Hua, G. Meta Pairwise Relationship Distillation for Unsupervised Person Re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.
25. Lin, Y.; Xie, L.; Wu, Y.; Yan, C.; Tian, Q. *Unsupervised Person Re-Identification via Softened Similarity Learning*; Cornell University: New York, NY, USA, 2020.
26. Morabbi, S.; Soltanizadeh, H.; Mozaffari, S.; Fadaeieslam, M.J. Improving generalization in deep neural network using knowledge transformation based on fisher criterion. *J. Supercomput.* **2023**, *79*, 20899–20922. [CrossRef]
27. Ye, M.; Zhang, X.; Yuen, P.C.; Chang, S.F. Unsupervised embedding learning via invariant and spreading instance feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6210–6219.
28. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
29. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
30. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9912–9924.
31. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
32. Robinson, J.; Chuang, C.Y.; Sra, S.; Jegelka, S. Contrastive learning with hard negative samples. *arXiv* **2020**, arXiv:2010.04592 .
33. Li, L.; Zhou, Z.; Wang, B.; Miao, L.; Zong, H. A novel CNN-based method for accurate ship detection in HR optical remote sensing images via rotated bounding box. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 686–699. [CrossRef]

34. Zhao, F.; Liao, S.; Xie, G.S.; Zhao, J.; Zhang, K.; Shao, L. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; Proceedings, Part XI 16, pp. 526–544.
35. Ge, Y.; Chen, D.; Li, H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv* **2020**, arXiv:2001.01526.
36. Yang, F.; Zhong, Z.; Luo, Z.; Cai, Y.; Lin, Y.; Li, S.; Sebe, N. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4855–4864.
37. Wang, M.; Lai, B.; Huang, J.; Gong, X.; Hua, X.S. Camera-aware proxies for unsupervised person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 2764–2772.
38. Cho, Y.; Kim, W.J.; Hong, S.; Yoon, S.E. Part-based pseudo label refinement for unsupervised person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7308–7318.
39. Zhang, X.; Ge, Y.; Qiao, Y.; Li, H. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3436–3445.
40. Chen, H.; Lagadec, B.; Bremond, F. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 14960–14969.
41. Pang, Z.; Zhao, L.; Liu, Q.; Wang, C. Camera invariant feature learning for unsupervised person re-identification. *IEEE Trans. Multimed.* **2022**, *25*, 6171–6182 [CrossRef]
42. Li, P.; Wu, K.; Zhou, S.; Huang, Q.; Wang, J. Pseudo Labels Refinement with Intra-Camera Similarity for Unsupervised Person Re-Identification. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 8–11 October 2023; pp. 366–370.
43. Cheng, S.; Chen, Y. Camera sensing unsupervised pedestrian re-recognition method guided by pseudo-label refinement. *JEMI* **2023**, *50*, 230239.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.