

Article

Tri-Invariance Contrastive Framework for Robust Unsupervised Person Re-Identification

Lei Wang¹, Chengang Liu², Xiaoxiao Wang³, Weidong Gao^{4,*}, Xuejian Ge¹ and Shunjie Zhu³

¹ School of Automation, Wuxi University, Wuxi 214105, China; leiwang@cwuxu.edu.cn (L.W.); gexuejian@cwuxu.edu.cn (X.G.)

² School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China; 7201905015@stu.jiangnan.edu.cn

³ School of Automation, Nanjing University of Information Science and Technology, Wuxi 210044, China; 202412232619@nuist.edu.cn (X.W.); 202412232628@nuist.edu.cn (S.Z.)

⁴ School of Electronic Information Engineering, Wuxi University, Wuxi 214105, China

* Correspondence: wdgao@cwuxu.edu.cn

Abstract

Unsupervised person re-identification (Re-ID) has been proven very effective and it boosts the performance in learning representations from unlabeled data in the dataset. Most current methods have good accuracy, but there are two main problems. First, clustering often generates noisy labels. Second, features can change because of different camera styles. Noisy labels causes incorrect optimization, which reduces the accuracy of the model. The latter results in inaccurate prediction for samples within the same category that have been captured by different cameras. Despite the significant variations inherent in the vast source data, the principles of invariance and symmetry remain crucial for effective feature recognition. In this paper, we propose a method called Invariance Constraint Contrast Learning (ICCL) to address these two problems. Specifically, we introduce center invariance and instance invariance to reduce the effect of noisy samples. We also use camera invariance to handle feature changes caused by different cameras. Center invariance and instance invariance help decrease the impact of noise. Camera invariance improves the classification accuracy by using a camera-aware classification strategy. We test our method on three common large-scale Re-ID datasets. It clearly improves the accuracy of unsupervised person Re-ID. Specifically, our approach demonstrates its effectiveness by improving mAP by 3.5% on Market-1501, 1.3% on MSMT17 and 3.5% on CUHK03 over state-of-the-art methods.

Keywords: person reidentification; contrastive learning; unsupervised learning; robust modeling

MSC: 68T07; 68T09; 68T45



Academic Editor: Jonathan Blackledge

Received: 10 October 2025

Revised: 2 November 2025

Accepted: 5 November 2025

Published: 6 November 2025

Citation: Wang, L.; Liu, C.; Wang, X.; Gao, W.; Ge, X.; Zhu, S. Tri-Invariance Contrastive Framework for Robust Unsupervised Person Re-Identification. *Mathematics* **2025**, *13*, 3570. <https://doi.org/10.3390/math13213570>

Copyright: © 2025 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

Person re-identification (Re-ID) [1–3] is a cross-camera retrieval task designed to match and associate images of the same individual captured by non-overlapping surveillance cameras. Although tremendous progress has been achieved in fully supervised approaches [4–6], a large amount of time and a significant amount of work are still required for label annotations in real practice. In recent years, unsupervised methods have received more attention to bypass the scarcity of annotations. These methods learn identity features

for identity retrieval directly from unlabeled data. Unsupervised Re-ID methods usually fall into two different types. One type uses extra labeled data, called unsupervised domain adaptation (UDA) [7,8]. The other type does not use extra labeled data and is called fully unsupervised learning (USL) [6,9,10]. Under the framework of UDA, learning is conducted with annotated source-domain data and unannotated target-domain data, where the two domains exhibit distinct data distributions. The primary goal is to develop a model capable of generalizing well to the target domain. In contrast, the fully unsupervised person re-identification (Re-ID) paradigm poses a more difficult problem, as it requires models to be trained exclusively on unlabeled visual data without leveraging any annotated samples.

In response to these challenges, a wide range of unsupervised learning methods have been designed to boost feature discriminability and clustering reliability. The study in [11] proposed an Ensemble of Invariant Features (EIFs) method, which integrates global features extracted from a pretrained CNN and region features derived from a bidirectional Gaussian mixture model to enhance robustness to pose and viewpoint variations. To mitigate camera-induced bias, ref. [12] introduced a camera-aware contrastive learning framework using time-based clustering and a 3D attention mechanism to decouple features from camera dependencies. Addressing the noise in pseudo labels, this study in [13] proposed the PPLR framework, which leverages the consistency between global and part-level features to refine labels and improve feature discriminability. Furthermore, the research in [14] presented Cluster Contrast, which computes contrastive loss at the cluster level and utilizes momentum updating to maintain feature consistency in the memory bank. To alleviate sub-cluster fragmentation and identity confusion during clustering, the research in [15] developed the ISE method, which generates boundary support samples via progressive linear interpolation and applies a label-preserving constraint to enhance cluster compactness. These representative approaches provide effective solutions for USL Re-ID, laying the foundation for further research in this direction. While these methods address specific aspects of the problem, they often operate in a piecemeal manner. A fundamental and underaddressed challenge persists: the misalignment between empirically computed cluster centroids and the theoretical, invariant centroid for each identity. This misalignment arises from two main sources: (1) the inherent noise in pseudo-labels generated by clustering algorithms, which leads to incorrect optimization directions; and (2) the significant intra-class variation induced by cross-camera discrepancies (e.g., viewpoint, lighting). These factors cause the computed centroids to deviate from the ideal invariant representation, ultimately degrading model discriminability.

It is precisely this core problem of centroid misalignment that motivates our work. In this study, we focus on addressing the person Re-ID task in a fully unsupervised setting, where no labeled data is available (i.e., USL). Most current USL methods rely on clustering-generated pseudo-labels and memory-based dictionaries to guide the training of deep neural networks. At the start of each training epoch, the model extracts feature representations for all training images using the latest network parameters. Next, a clustering method such as DBSCAN [16] or K-means [17] groups image features and creates pseudo-labels [18,19]. Each image is then assigned a cluster ID, which is used to identify each person. Finally, the neural network is trained using a contrastive loss [14,20,21], like triplet loss [22,23], InfoNCE loss [24], or another non-parametric classification loss [25], based on the memory dictionary. In the whole training process, the memory module essentially functions as a dynamic database that stores all centroid features abstracted from each class, and the quality of these centroid features directly influences both the prediction accuracy for query samples and the overall training outcome.

We believe when mapping samples within same category into the feature spaces that, ideally, there is one invariant theoretical cluster centroid for each category. We conceptual-

ize this as the theoretical invariant centroid, which is defined as the ideal representation of an identity that remains stable across all camera views and is unaffected by sampling noise or clustering errors. Due to the noisy pseudo-labels produced by the clustering algorithm as well as the samples' variations, there tends to be a relatively large distance between computed cluster centroids and theoretical cluster centroids. Given the diverse origins of data collection, encompassing factors like light-intensive lens angles, and more, these variations can significantly influence feature recognition. Consequently, the cornerstone of accurate recognition lies in the perception of target invariance and symmetry. By emphasizing these principles, we can mitigate the effects of external factors and enhance the reliability of our recognition processes.

Inspired by this, we posit that learning robust representations requires explicit constraints to pull features toward their theoretical invariant centroids, irrespective of noise and camera-specific biases. While individual concepts like centroid stabilization or camera awareness have been explored, their combinations are often additive. The novelty of our approach, Invariance Constraint Contrast Learning (ICCL), lies in its synergistic, tri-invariance framework, which is explicitly designed to correct the centroid misalignment from multiple angles simultaneously.

To solve this issue, we proposed a unified framework with three co-adapted constraints: center constraint, instance constraint, and camera constraint. These are not merely combined but are co-designed to form a synergistic system. First, center invariance loss will utilize the mean value of the features belonging to the same category to reduce the distance between computed cluster centroid and the theoretical centroid, providing a stable, global target. Meanwhile, the closest feature (in the query set) to the computed centroid is selected as the representative to update the memory, and instance invariance loss is computed between query set and memory to further reduce the above distance, thereby supplying a discriminative, local anchor that counteracts over-smoothing. Furthermore, camera variations might significantly affect discriminating individual samples, we employ the camera invariance to improve the inter-camera discrimination, acting as a unifying regularizer that explicitly minimizes cross-camera feature discrepancies. This tri-invariance mechanism collectively ensures that the model's learning trajectory is consistently guided towards the theoretical invariant centroid. Figure 1 illustrates the invariance contrast learning concept.

The main contributions of this paper can be summarized as follows:

- We identify and formalize the problem of centroid misalignment between empirical cluster centroids and theoretical invariant centroids as a key bottleneck in unsupervised Re-ID.
- We propose ICCL, a novel framework that introduces center, instance, and camera invariance not as isolated components, but as a synergistic system to jointly address noise and camera variations, effectively bridging the centroid misalignment gap.
- The proposed unified framework integrates multiple invariance-based contrastive learning strategies, allowing the model to effectively leverage their combined strengths and resulting in a more robust solution for person re-identification.
- The method is tested on Market-1501, MSMT17 and CUHK03 datasets. The experimental results show that the method is effective and achieves good performance.

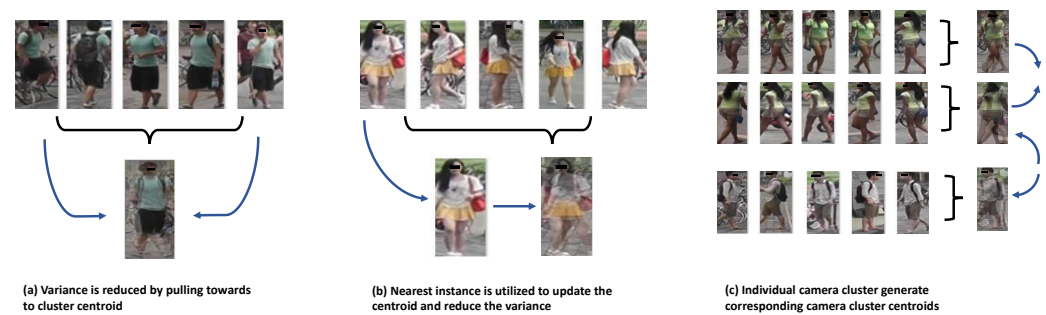


Figure 1. Illustration of invariance constraint in person Re-ID. (a) Variance between representatives and centroid are reduced by pulling together to the centroid. (b) The feature that is the closest to the centroid is obtained as the cluster representative, and the samples are pulled together to reach the closest feature instead. (c) Within each cluster, we select samples captured by different cameras and create camera aware clusters, followed by operating the similarity matrix to compute camera invariance loss.

2. Related Work

2.1. Unsupervised Approaches for Person Re-ID

Unsupervised person re-identification (Re-ID) seeks to learn discriminative feature representations without requiring labeled data in the target domain. Among these methods, unsupervised domain adaptation (UDA) has been a widely adopted paradigm [26–28]. These approaches either pre-train models on labeled source datasets or perform image-to-image translation using style transfer to align the source and target domains. Style transformation techniques—such as CycleGAN and SPGAN—have been employed to reduce the domain gap by transferring visual styles while preserving identity information [29–31].

However, UDA methods encounter serious performance bottlenecks when there is a large domain discrepancy in terms of lighting, background, camera angles, or a person’s appearance. In such cases, label noise becomes prevalent during the pseudo-labeling stage, which negatively affects the model’s convergence and generalization. Furthermore, the reliance on source domain distributions often causes overfitting to source-specific characteristics, limiting the ability to adapt flexibly to unseen domains.

To circumvent these limitations, recent research has shifted toward fully unsupervised learning (USL), which does not use any labeled source data. These methods generate pseudo-labels for unlabeled target samples based on feature similarities or affinity graphs [32,33], and iteratively refine them to improve label reliability. The quality of the initial clustering, which typically serves as the foundation for these pseudo-labels, plays a crucial role in performance.

Various clustering algorithms have been explored, such as DBSCAN, hierarchical clustering, and the widely used K-means algorithm [34–36]. Fan et al. [37] proposed an unsupervised framework using K-means for pseudo-label generation, followed by model refinement. Lin et al. [38] extended this idea by introducing bottom-up clustering and an online memory module to maintain feature consistency across training epochs. Addressing person re-identification from a novel perspective, Wang et al. [25] transformed the conventional framework into a multi-label recognition system. A key innovation in their work was the integration of an evolving feature repository to boost algorithmic resilience.

To further mitigate label noise, mutual learning frameworks have been adopted. Notable methods include NRMT [39], MMT [26], and MEB-Net [40], which leverage the co-training of two or more networks with soft label averaging or consistency regularization to suppress unreliable pseudo-labels. These strategies promote collaborative learning while dynamically filtering incorrect samples. Recently, some studies have introduced graph-

based refinement strategies, where inter-sample relationships are modeled explicitly via k -NN graphs or learned similarity matrices. Such methods enable structure-aware pseudo-label refinement by considering global context, which improves clustering granularity and stability. It is worth noting that the challenges of unsupervised learning and the principles of invariance are also actively explored in other Re-ID modalities. For instance, in cross-modality Re-ID, recent works like [41] have proposed hierarchical clustering and refinement techniques to bridge the gap between visible and infrared spectra. Furthermore, efforts such as [42] aim to learn a grand unified representation, underscoring the broader research trend towards building robust, modality-invariant feature spaces. While our work focuses on the single-modality RGB setting, the underlying philosophy of learning invariant representations shares common ground with these advanced paradigms.

2.2. Contrastive Learning for Person Re-ID

Contrastive learning has emerged as a prominent approach in unsupervised representation learning due to its ability to extract discriminative features without requiring explicit labels. A range of methods [24,43–45] have been developed, aiming to learn feature embeddings by encouraging representations of similar samples (positive pairs) to be close, while dissimilar ones (negative pairs) are pushed apart. These approaches typically leverage data augmentation to generate positive pairs from the same image and negative pairs from other images.

Among them, the InfoNCE loss proposed in [24] has attracted considerable attention, as it was later shown to be mathematically equivalent to maximizing the mutual information between different augmented views of the same input. This formulation underpins many subsequent advances in contrastive learning. Literature [43] explored improved sampling strategies, while [44] introduced a Siamese network structure to perform pairwise instance discrimination. Furthermore, ref. [45] proposed a momentum-based encoder update mechanism to maintain a consistent dictionary of negative samples, thereby enhancing representation stability over training iterations.

With the proven effectiveness of contrastive learning in general vision tasks, researchers have started applying it to person Re-ID scenarios. Two highly relevant lines of work are Cluster Contrast [14] and Camera Contrast Learning [12]. Cluster Contrast computes contrastive loss at the cluster level, using a momentum-updated memory bank to stabilize the learning targets. Camera Contrast Learning, on the other hand, explicitly incorporates camera identity to learn camera-aware proxies and disentangle camera-specific bias. While our ICCL framework shares the high-level concept of using a memory dictionary with these methods, its theoretical foundation and design are distinct. The core distinction lies in ICCL's explicit objective of mitigating the centroid misalignment problem through a synergistic, multi-invariance framework. Cluster Contrast can be viewed as an implementation of a single invariance (center). Camera Contrast tackles camera variance but does not explicitly address the noise from clustering. In contrast, ICCL co-designs three complementary invariances: center invariance (stabilizing the target), instance invariance (preserving discriminative details), and camera invariance (reducing view-specific bias). This unified approach ensures a more consistent and direct optimization towards the theoretical invariant centroid, addressing both major sources of misalignment simultaneously, which individual methods do not do.

2.3. Backbone Architectures in Vision Tasks

The evolution of backbone architectures has significantly advanced the state of the art in computer vision. Convolutional Neural Networks (CNNs), particularly the ResNet family, have long been the workhorse for feature extraction in Re-ID and other tasks due to

their strong performance and efficiency. In parallel, transformer-based architectures have recently emerged as powerful alternatives, demonstrating remarkable success. Notably, in safety-critical applications such as construction site monitoring, advanced frameworks integrating YOLOv10 and transformers have set new benchmarks for detection accuracy in scenarios involving surveillance and body-worn cameras. Refs. [46,47] While these transformer-based models show superior representational capacity in various domains, our work employs ResNet-50 as the primary backbone. This choice is motivated by the specific focus and contribution of our paper: to propose and validate a novel learning framework (ICCL) for unsupervised Re-ID. Using a standardized and widely adopted backbone like ResNet-50 allows for a controlled experimental setting, ensuring that the reported performance gains are directly attributable to our proposed framework rather than the backbone architecture. It also facilitates fair comparison with prior works and enhances the reproducibility of our study. We acknowledge that exploring the integration of state-of-the-art transformer backbones with the ICCL framework is a compelling and highly promising direction for future research.

3. The Proposed Method

In unsupervised person re-identification, the available data comes from an unlabeled target domain denoted as $\mathcal{T} = \{X\}$. This dataset includes D images, expressed as $\{x_i\}_{i=1}^D$, for which identity labels are not provided. However, each image is associated with its corresponding camera index. Under these conditions, the objective is to develop a model capable of achieving strong generalization performance on the target domain.

3.1. The Overall Framework

The proposed ICCL framework fundamentally differs from prior cluster-contrast or camera-aware methods. While existing works typically direct optimization towards a single learning target (e.g., a cluster centroid), ICCL is designed around the co-optimization of multiple, complementary learning targets. The center invariance provides a stable but potentially over-smoothed target; the instance invariance provides a discriminative but potentially noisy target, and the camera invariance restructures the feature space to ensure the first two targets are computed from a camera-invariant distribution. It is this theoretical framing of the problem and the synergistic mechanism that distinguishes ICCL, enabling a more accurate and robust estimation of the underlying data manifold.

As illustrated in Figure 2, we use a pretrained ResNet-50 [48] as the backbone encoder to extract feature vectors. DBSCAN clustering algorithms are applied to cluster features with the shortest distance together and pseudo-labels are assigned. This clustering procedure is executed at the beginning of every training epoch to refresh the pseudo-labels based on the evolving feature representations, which is crucial for providing timely guidance for the subsequent contrastive learning stage. Based on the pseudo-labels, we obtain the N classes. Then, we use camera information to perform camera-aware classification and obtain a camera-aware classification result. In the following stages, three similarity matrices are estimated between centroid features and query features. In camera invariance similarity matrices, we use only samples that are within same cluster but captured by different camera as positive samples. For center invariance and instance invariance, we update the memory by different mechanisms during the training process. In the end, contrast loss is computed with the combination of center invariance loss, instance invariance loss, and camera invariance loss.

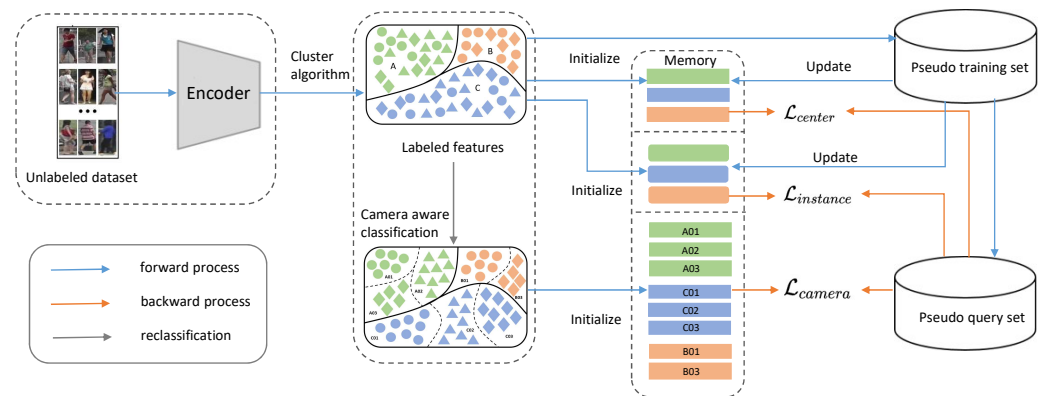


Figure 2. The ICCL framework for unsupervised person Re-ID begins by using an encoder to extract features from unlabeled images. A clustering algorithm is then applied to these features to generate pseudo-labels, forming the training set. Based on the pseudo-labels and camera identifiers, camera-aware classification is performed to capture cross-camera variations. Subsequently, invariance relations—including center, instance, and camera invariance—are computed and stored in corresponding memory modules, which are continuously updated with newly extracted features. Finally, multi-invariance losses are calculated by integrating the three individual loss components, and the overall loss is backpropagated to optimize the encoder.

3.2. Center Invariance

Contrastive methods are designed to enhance feature representation by pulling samples of the same class closer together while pushing apart those from different classes. This approach strengthens intra-class consistency and increases inter-class distinction, ultimately improving the model’s discriminative ability. In order to achieve this goal, we first construct an initialized cluster centroid matrix, then we dynamically update the centroid matrix by the representatives of each corresponding cluster. We design our center-invariance loss using the following equation:

$$\mathcal{L}_{center} = -\log \frac{\exp(q \cdot \omega_+)}{\sum_{n=1}^N \exp(q \cdot \omega_n)}, \quad (1)$$

where q is an encoded query, ω_+ is the centroid features that shares the same label with q and ω_n is all the centroid features stored in dynamically updated slots. The update mechanism is as follows:

$$\omega_n(t+1) \leftarrow \mu \omega_n(t) + (1 - \mu) \frac{1}{|\mathcal{B}_n|} \sum_{b_k \in \mathcal{B}_n} b_k, \quad (2)$$

the momentum coefficient μ in Equation (2) is empirically set to 0.2 and remains fixed throughout training. This choice achieves a balance between stability and adaptability: a lower μ enables the centroid memory to respond more quickly to feature changes, while a higher value introduces excessive smoothing and may fail to track the evolving representations. Our preliminary experiments across $\mu \in [0.1, 0.5]$ showed that model performance is relatively stable within this range, with $\mu = 0.2$ yielding marginally superior results. Hence, we employ a fixed momentum coefficient instead of an adaptive scheduling strategy.

3.3. Instance Invariance

Essentially, the goal of obtaining the cluster centroid is to represent all the features of the corresponding cluster to the greatest extent possible. In this way, we can calculate the multi-class score of the query feature with better accuracy in terms of loss function. However, we found that only using center invariance loss can lead to a loss of original

information. If we could exploit the information from b_k , we may be able to improve the diverseness and accuracy of the representatives. To fulfill this target, we calculate the distance between each feature b_k and its centroid $C(b_k)$, and consider the one with the smallest distance as representative and update the slots as follows:

$$\varphi_n(t+1) \leftarrow \mu \varphi_n(t) + (1-\mu) \underset{k \in K}{\operatorname{argmin}}(b_k, C(b_k)), \quad \varphi_n(0) = \omega_n(0), \quad (3)$$

where $\varphi_n(0)$ is used as an initialized value of the centroids set, and $\varphi_n(0)$ originates from the same source of $\omega_n(0)$ as noted. Similarly, we obtain instance centroids and add them to the contrast loss equation as follows:

$$\mathcal{L}_{instance} = -\log \frac{\exp(q \cdot \varphi_+)}{\sum_{n=1}^N \exp(q \cdot \varphi_n)}. \quad (4)$$

In this equation, q denotes the feature of a query sample from the current batch. φ_+ is the representative instance feature (from the instance-level memory dictionary) that shares the same pseudo-label with q . The set $\{\varphi_n\}_{n=1}^N$ contains the representative features for all N clusters stored in the memory. This loss aims to pull the query feature q closer to its corresponding high-confidence instance-level representative φ_+ in the feature space, while pushing it away from the representatives of other clusters.

3.4. Camera Invariance

In previous studies, most training schemes employed pseudo-labels achieved by the clustering algorithm to assign identities for each sample and compare query samples with each cluster centroid feature. The algorithm's performance remains limited because it overlooks the influence of camera variations, which play a vital role in building a robust person Re-ID system. Pedestrian appearances often vary significantly across cameras due to differences in viewpoint, lighting conditions, and other environmental factors, resulting in noticeable discrepancies within the same identity's features. If these camera-induced changes are not accounted for, the model becomes more susceptible to such variations, which can degrade clustering accuracy and hinder effective training. To overcome this issue, we introduce a camera-aware classification approach that encourages the model to learn features that are less affected by camera differences.

We use camera information to classify the pseudo-labeled data on a camera-aware level. By performing the intersect operation of camera and cluster information, we retrieved the camera clusters and furthered the camera cluster centroids as initialized camera cluster representations σ_+ . We then computed the similarities between camera cluster representations and query samples. In the similarity matrix $\mathcal{S}$, we exclusively define the samples within the same cluster which belong to different cameras as positives while defining hard negative samples as negatives. To this end, we list the camera invariance similarities of i -th query feature as follows:

$$S(i) = \frac{\exp(q(i) \cdot \sigma_+ / \tau)}{\sum_{t \in P \cup Q} \exp(q(i) \cdot \sigma_t / \tau)}, \quad (5)$$

where P denotes the positive features set with samples that are within the same cluster but captured by different cameras, σ_+ belongs to P , and Q denotes the negative feature set. τ is the temperature.

In our implementation, the positive set P contains all samples from the same cluster that are captured by different cameras, whereas the negative set Q includes samples from other clusters. The ratio between positive and negative samples is therefore determined adaptively by the data distribution, without manual re-weighting. We set the temperature

$\tau = 0.07$, which was selected based on preliminary experiments and follows standard practice in contrastive learning to control the concentration level of the softmax distribution. If a cluster contains images from only a single camera ($|P| = 0$), we omit the camera-invariance loss for that cluster and rely solely on the center- and instance-invariance losses, ensuring that no spurious positives are introduced.

In the end, the camera invariance loss can be formulated by as follows:

$$\mathcal{L}_{camera} = - \sum_{i=1} \frac{1}{|P|} \log S(i). \quad (6)$$

In this way, camera-aware classified cluster features that are within the same cluster but in different cameras are pulled together; meanwhile, variance inside one class caused by disjoint camera views is reduced.

3.5. Overall Loss of Invariance Learning

In summary, the total loss of our invariance constraint contrast is formulated as follows:

$$\mathcal{L}_{ICCL} = \mathcal{L}_{center} + \mathcal{L}_{instance} + \lambda_{cam} \mathcal{L}_{camera}. \quad (7)$$

In our implementation, the weights of $\mathcal{L}_{center}$ and $\mathcal{L}_{instance}$ are both set to 1, giving them equal importance in the overall objective. This choice is motivated by their complementary roles: $\mathcal{L}_{center}$ stabilizes the representation of cluster centroids, while $\mathcal{L}_{instance}$ preserves instance-level diversity by pulling features toward the most representative sample rather than the mean. Empirically, we observed that equal weighting provides a good balance between centroid stability and instance discrimination across different datasets, where center invariance learning and instance invariance learning effectively reduce the influence of noisy labels. λ_{cam} is a hyperparameter that controls the importance of the camera invariance, which significantly reduces the intra-class variance caused by disjoint camera views. In conclusion, the three loss terms in Equation (7) are not merely combined but are co-designed to form a synergistic system. The center and instance invariances engage in a collaborative push-and-pull dynamic. The former provides a stabilized, global target that mitigates the impact of label noise, while the latter supplies a discriminative local anchor that counteracts the risk of over-smoothing and preserves fine-grained features. Operating upon this stabilized and refined feature space, the camera invariance acts as a unifying regularizer. It explicitly minimizes cross-camera feature discrepancies, thereby ensuring that the centroids and representative instances are computed from a more view-invariant distribution. This tri-invariance mechanism collectively ensures that the model's learning trajectory is consistently guided towards the theoretical invariant centroid, effectively addressing both the noise from clustering and the variance from camera views in a unified framework. To clarify the training procedure, the pseudo code is summarized and presented in Algorithm 1.

Algorithm 1 Optimization procedure with invariance contrast learning**Input:** Unlabeled data $\mathcal{D}$ captured**Output:** Optimized model F

```

1: for  $n$  in  $[1, \text{num\_epochs}]$  do
2:   Generate pseudo-labels for  $\mathcal{D}$  with DBSCAN
3:   Obtain pseudo labeled dataset  $\mathcal{D}'$ 
4:   Obtain cluster centroid  $\mathcal{B}$ 
5:   Obtain camera cluster representatives  $\mathcal{R}$  by camera ID  $\cap$  pseudo-labels in  $\mathcal{D}'$ 
6:   for  $m$  in  $[1, \text{num\_batch}]$  do
7:     Compute center invariance loss  $\mathcal{L}_{center}$  with Equation (1)
8:     Update cluster centroids  $\omega_n$  with center invariance by Equation (2)
9:     Compute instance invariance loss  $\mathcal{L}_{instance}$  with Equation (4)
10:    Update cluster centroids  $\varphi_n$  with center invariance by Equation (3)
11:    Calculate similarity  $\mathcal{S}$  in Equation (5)
12:    Bring  $\mathcal{S}$  in camera invariance loss  $\mathcal{L}_{camera}$  in Equation (6)
13:    Compute combined loss  $\mathcal{L}_{ICCL}$  in Equation (7)

```

4. Experiment

4.1. Datasets

We conduct evaluations using two widely adopted large-scale person Re-ID benchmarks: Market1501 [49] and MSMT17 [27]. The Market-1501 dataset includes 32,668 images representing 1501 different individuals, collected from six disjoint camera views. It is divided into 12,936 training samples covering 751 identities and 19,732 testing samples involving 750 identities. In contrast, MSMT17 poses a greater challenge, featuring 126,441 images of 4101 identities captured under 15 different cameras. This dataset is partitioned into 32,621 training images belonging to 1041 identities and 93,820 testing images associated with 3060 identities. For training purposes, only the raw images and their corresponding camera IDs from the training split are used, without relying on any additional annotations. The evaluation metrics include Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP). A summary of the dataset statistics is provided in Table 1.

Table 1. Statistics of the datasets used in the Experimental Section.

Dataset	Object	#Train IDs	#Train Images	#Test IDs	#Query Images	#Total Images
Market-1501	Person	751	12,936	750	3368	32,668
MSMT17	Person	1041	32,621	3060	11,659	126,441

4.2. Default Experimental Settings

To facilitate reproducibility and provide a clear overview of our experimental setup, we summarize the key configuration details in Table 2. Unless otherwise explicitly stated, these settings are used as defaults across all experiments.

Table 2. Default experimental configuration.

Component	Configuration/Value
Backbone	ResNet-50 [48]
Input Size	256×128
Batch Size	256 (comprising 16 distinct identities $\times$ 16 instances per identity)
Optimizer	Adam (weight decay = 5×10^{-4})
Learning Rate	3.5×10^{-4} , kept constant throughout training
Training Epochs	Market-1501: 60, MSMT17: 80
Data Augmentation	Random horizontal flipping, 10-pixel padding, random cropping, random erasing
Loss Weights	$\mathcal{L}_{center} : 1, \mathcal{L}_{instance} : 1, \lambda_{cam} : 0.5$
Clustering Algorithm	DBSCAN (eps = 0.6, min_samples = 4)
Memory Momentum (μ)	0.2

4.3. Implementation Details

We utilize ResNet-50 [48] as the feature extractor and initialize the network with weights pretrained on ImageNet. In the testing procedure, similarity measurements are calculated using representations derived from the globally averaged pooling operation. At the start of each epoch, pseudo-labels are generated using the DBSCAN clustering algorithm. Input images are uniformly resized to 256×128 . To strengthen the model's generalization capability during the learning phase, various augmentation approaches are employed: horizontal axis reflection, 10-pixel margin addition, probabilistic cropping, and random masking. The batch configuration follows a structured sampling scheme where 256 training samples are selected from 16 unique identities, ensuring equal representation of 16 instances per identity. This batch size was chosen for two key reasons pertinent to our framework: (1) Our method relies on clustering-based pseudo-labels and memory dictionaries. A sufficiently large batch size provides a more representative sample of each cluster within an iteration, leading to more stable and accurate updates of the centroid (ω_n), instance (φ_n), and camera-aware memories. (2) Contrastive learning objectives, which form the core of our invariance losses, are known to benefit from larger batch sizes as they provide a richer set of negative examples, thereby improving the discriminative power of the learned features. The chosen size of 256 represents a practical balance that leverages these benefits while remaining within our computational constraints. The model is trained using the Adam optimizer with a weight decay of 5×10^{-4} , and an initial learning rate set to 3.5×10^{-4} . The training spans 60 epochs for Market-1501 and 80 epochs for MSMT17, with each epoch comprising 200 iterations. DBSCAN is applied based on Jaccard distance and k-reciprocal encoding, with a maximum pairwise distance threshold of 0.6 and a minimum of four neighbors for a valid core sample. All experiments are implemented using PyTorch 1.9 [50] and executed on two NVIDIA GeForce RTX 2080 Ti GPUs.

Computational Consideration of Clustering: The DBSCAN clustering is performed once at the beginning of every training epoch. Unlike iterative clustering algorithms (e.g., K-means), DBSCAN is a single-pass algorithm that does not have an internal convergence criterion; it directly forms clusters based on the density connectivity of the feature points given the pre-defined hyper-parameters ($\text{eps} = 0.6$, $\text{min_samples} = 4$). While the computational complexity of DBSCAN is approximately $O(D \log D)$ samples using efficient spatial indexing, it still constitutes a non-negligible part of the training pipeline. In our experiments, the clustering step typically accounts for approximately 5–8% of the total training time per epoch on two NVIDIA GeForce RTX 2080 Ti GPUs. We deem this overhead acceptable, as the quality of the pseudo-labels is paramount for guiding the contrastive learning objective, and reclustering at every epoch provides the most timely and coherent learning signals.

4.4. Comparison with Existing Methods

To show the effectiveness of our proposed method, we compared the mAP, Rank-1, Rank-5, and Rank-10 accuracies of ICCL with several state-of-the-art unsupervised person Re-ID methods on three large-scale datasets: Market-1501, MSMT17, and CUHK03.

As shown in Table 3, the proposed method achieves superior performance compared to existing approaches in both fully unsupervised and domain adaptation settings. To ensure fair and robust evaluation, all reported results of our ICCL method are obtained by averaging over three independent runs with different random seeds. The mean values and standard deviations are reported to reflect the statistical stability of the model. The small variance observed across runs indicates that the performance improvements are consistent across random initializations.

Table 3. Comparison with state-of-the-art methods on the object Re-ID, including supervised methods, unsupervised methods, and unsupervised domain adaptation methods. Note that red indicates the best method and blue indicates the runner-up. The results of our method are the average over 5 runs with different random seeds. The standard deviations fall within the narrow range of 0.1 to 0.3 across all datasets and metrics, indicating a highly consistent performance.

Method	Market1501				MSMT17				CUHK03			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Fully unsupervised												
BUC [6]	38.3	66.2	79.6	84.5	-	-	-	-	-	-	-	-
SSL [9]	37.8	71.7	83.8	87.4	-	-	-	-	-	-	-	-
MMCL [25]	45.5	80.3	89.4	92.3	11.2	35.4	44.8	49.8	-	-	-	-
HCT [10]	56.4	80.0	91.6	95.0	-	-	-	-	-	-	-	-
CycAs	64.8	84.8	-	-	26.7	50.1	-	-	47.4	41.0	-	-
UGA [51]	70.3	87.2	-	-	21.7	49.5	-	-	-	-	-	-
SPCL/Infomap	70.7	86.3	93.6	95.6	17.2	40.6	53.6	59.2	-	-	-	-
SPCL [18]	73.1	88.1	95.1	97.0	19.1	42.3	55.6	61.2	-	-	-	-
TAUDL [52]	-	-	-	-	12.5	28.4	-	-	44.7	31.2	-	-
Cluster contrast [14]	82.1	92.3	96.7	97.9	27.6	56.0	66.8	71.5	-	-	-	-
ICE [20]	79.5	92.0	97.0	98.1	29.8	59.0	71.7	77.0	-	-	-	-
Ours	85.6	92.1	97.3	98.2	31.1	60.5	70.9	77.6	50.9	42.3	45.7	49.1
Domain adaptive												
UTAL [53]	-	-	-	-	13.1	31.4	-	-	56.3	42.3	-	-
MMCL [25]	60.4	84.4	92.8	-	-	-	-	-	-	-	-	-
ECN [29]	-	-	-	-	10.2	30.2	41.5	46.8	-	-	-	-
AD-Cluster++ [4]	68.3	86.7	94.4	96.5	-	-	-	-	-	-	-	-
MMT [26]	75.6	89.3	95.8	97.5	24.0	50.1	63.5	69.3	-	-	-	-
SPCL [18]	77.5	89.7	96.1	97.6	26.8	53.7	65.0	69.8	-	-	-	-
MEB-Net [54]	76.0	89.9	96.0	97.5	-	-	-	-	-	-	-	-

Our method achieved mean mAPs of $85.6 \pm 0.18\%$ on Market-1501, $31.1 \pm 0.16\%$ on MSMT17, and $50.9 \pm 0.19\%$ on CUHK03, with corresponding Rank-1 accuracies of $92.1 \pm 0.21\%$, $60.5 \pm 0.15\%$, and $42.3 \pm 0.31\%$. Table 3 demonstrates that our model consistently outperforms current state-of-the-art approaches, highlighting its superior overall performance.

Specifically, our method achieves improvements over the runner-up (Cluster Contrast) [14] by 3.5% on mAP for Market-1501, 1.3% on mAP and 1.5% on Rank-1 for MSMT17, and 3.5% on mAP and 1.3% on Rank-1 for CUHK03. Although the runner-up achieves a slightly higher Rank-1 accuracy (by 0.1%) on one dataset, the proposed ICCL model demonstrates more balanced and consistently superior performance across all benchmarks. The experimental results presented above clearly verify the effectiveness and strong performance of the proposed approach. To validate the superiority of our ICCL, we provide a t-SNE visualization compared to the baseline. As shown in Figure 3, the feature distribution learned by our method exhibits more compact and separable clusters across different identities, which intuitively explains its performance improvement by demonstrating a more discriminative feature space structure.

4.5. Ablation Study

In this section, we present a series of detailed ablation studies to evaluate the contribution of each component of our method, focusing on the domain adaptation task from Market-1501 to MSMT17. The ICCL model without any invariance mechanisms is used as the baseline for comparison. The impact of different combinations of loss functions on model performance is summarized in Table 4.

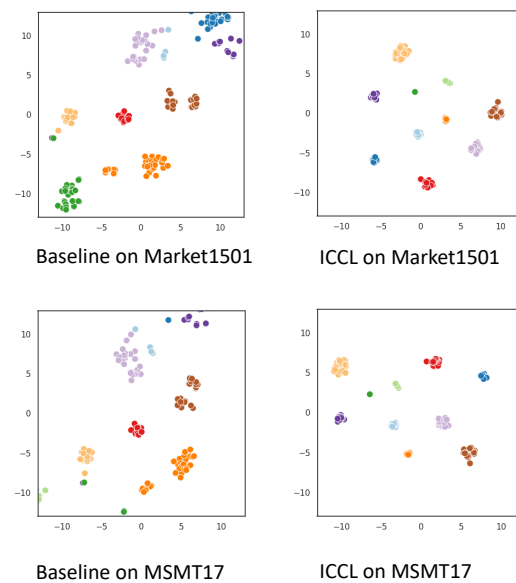


Figure 3. t-SNE visualization of feature embeddings for 10 randomly selected identities from Market-1501 and MSMT17 datasets. Each point represents one image feature and colors denote different identities. **(Left):** Baseline model features show overlapping and dispersed clusters. **(Right):** Features learned by ICCL exhibit compact, well-separated clusters, indicating stronger intra-class consistency and improved inter-class discrimination.

Table 4. Ablation studies on different components of our method. Note that red indicates the best performance.

Method	Market1501				MSMT17				CUHK03			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
ICCL w/ $\mathcal{L}_{center}$	78.1	90.5	93.9	95.8	23.1	53.8	62.1	68.1	42.1	36.8	39.7	42.9
ICCL w/ $\mathcal{L}_{center} + \mathcal{L}_{instance}$	80.1	91.6	94.8	98.1	23.2	55.8	66.5	69.3	46.7	39.1	41.5	47.2
ICCL w/ $\mathcal{L}_{center} + \mathcal{L}_{instance} + \mathcal{L}_{camera}$	85.6	92.1	97.3	98.2	31.1	60.5	70.9	77.6	50.9	42.3	45.7	49.1

When instance invariance contrast learning loss is utilized, we observe that performance significantly increases (e.g., it increases by 2% on mAP, 1.1% on rank-1 accuracy in the Market-1501 dataset; by 0.1% on mAP, 2% on rank-1 accuracy in the MSMT17 dataset; and by 4.6% on mAP, 2.3% on rank-1 accuracy in the CUHK03 dataset). This is because in the absence of the instance invariance contrast learning loss, cluster centroids are merely represented by their average value which lead to the loss of the original information. This observation validates the importance of instance contrast learning. The introduction of the camera-invariant contrastive loss further enhances performance. For instance, for the Market-1501 dataset, it leads to a 5.5% improvement in mAP and a 0.5% gain in rank-1 accuracy. On the MSMT17 dataset, the method achieves an increase of 7.9% in mAP and 4.7% in Rank-1 accuracy. On the CUHK03 dataset, it achieves a 4.2% gain in mAP and a 3.2% improvement in Rank-1 accuracy. This is because there are a great number of samples captured by different cameras but belonging to the same class that can easily be judged as belonging to different classes. The network has increased its discriminativity to distinguish this phenomenon with our camera invariance contrast loss. Hence, the camera invariance contrast learning is indispensable in ICCL. Last but not least, center contrast learning also plays a significant part in ICCL because it builds a solid foundation in accurately representing cluster centroids.

4.6. Visualization Analysis of Model Predictions

In this section, we leverage Grad-CAM to interpret and visualize the predictions generated by the proposed method. A detailed analysis is conducted on nine images

representing three different individuals, with each sample divided into three components: the original source image, the Grad-CAM heatmap, and a composite overlay image that integrates the heatmap with the original.

Figure 4a: This group presents the back, side, and front views of the three individuals. The heatmaps clearly show that the proposed model consistently concentrates on semantically discriminative regions such as upper body clothing, legs, and carried accessories, while successfully suppressing irrelevant background noise. This demonstrates the model's attention to consistent identity-related features across varying viewpoints. The activation intensity across the discriminative regions (indicated by the red and yellow heat zones) shows a high confidence level, suggesting strong spatial attention by the model.

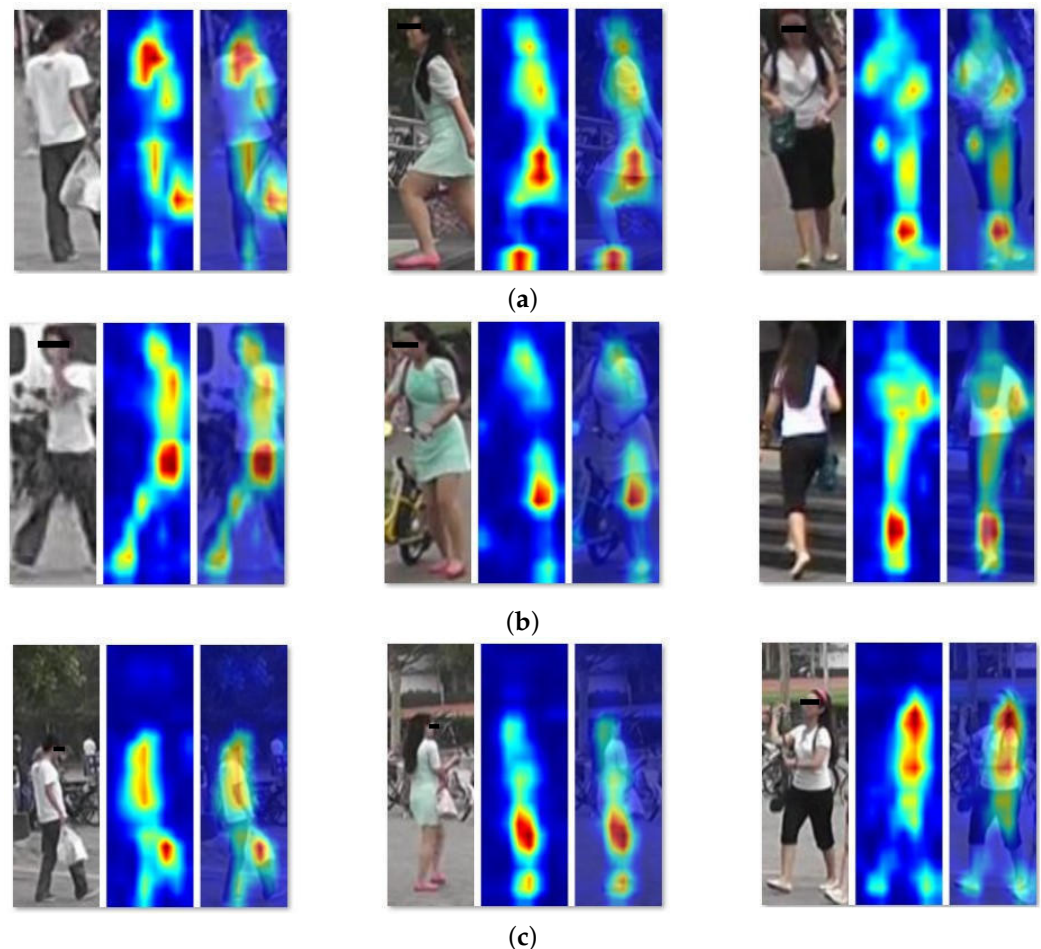


Figure 4. Grad-CAM visualization of ICCL predictions on three representative individuals under diverse conditions. (a) Consistent attention across multiple viewpoints (front, side, back) with strong focus on identity-specific regions such as upper clothing and accessories. (b) Robust localization under occlusion and interaction with background objects (e.g., bicycle), where the model maintains focus on torso and leg regions. (c) Scale-invariance test, where resized images still yield concentrated attention on discriminative body parts, confirming ICCL's robustness to image scaling and background clutter. The red-yellow regions indicate high attention, while blue areas represent suppressed background information.

Figure 4b: This section showcases significant variations in posture and occlusion, with instances of walking, turning, or interacting with environmental objects (e.g., a bicycle). Despite these challenges, the model continues to focus on core identity-specific features, such as clothing patterns and body silhouettes. Notably, in the central image, although the bicycle introduces substantial occlusion and visual clutter, the model still identifies and emphasizes key regions like the torso and legs, demonstrating robust feature localization.

Even when distraction levels are high, the model's attention does not scatter widely, indicating strong resistance to false positives in cluttered scenes.

Figure 4c: When the images are proportionally resized, causing an increase in visible background and a reduction in the relative scale of the human figure, the model retains its focus on the semantically rich areas of the person, rather than being misled by the background details. This indicates strong scale-invariance and background filtering ability. The normalized attention spread (i.e., how compact and centered the heatmaps are) remains consistent with those in Figure 4a,b, confirming that the model's spatial focus remains tightly bound to the individual, even under varying scale and resolution conditions.

Overall, the proposed method demonstrates a remarkable ability to consistently recognize discriminative features and effectively eliminate noise signals, even amidst diverse variations in the posture and state of the target individual. This underscores its robustness and adaptability in complex and dynamic scenarios. While this analysis qualitatively validates our model's focus, a systematic comparison of attention maps with baseline models remains an valuable direction for future work to provide deeper insights into localized feature discriminability.

4.7. Hyper-Parameter Analysis

We further investigate the impact of key hyper-parameters in our method, namely the weight coefficient λ_{cam} associated with the camera invariance loss, the momentum coefficient μ for memory updates, and the total number of training epochs. In the experimental setup, we vary one hyper-parameter at a time while keeping the others unchanged to isolate each one's effect on the performance.

Sensitivity to λ_{cam} . Figure 5 presents a quantitative analysis of λ_{cam} sensitivity across three datasets. On Market-1501, the optimal mAP of 85.6% is achieved at $\lambda_{cam} = 0.3$, while Rank-1 accuracy reaches 92.1%. For MSMT17, the best performance (mAP: 31.1%, Rank-1: 60.5%) occurs at $\lambda_{cam} = 0.3$. CUHK03 shows a similar trend to MSMT17, with optimal results (mAP: 50.9%, Rank-1: 42.3%) at $\lambda_{cam} = 0.3$. This quantitative evidence confirms that MSMT17 and CUHK03, which are more complex datasets with greater camera variation, benefit from stronger camera invariance weighting compared to Market-1501.

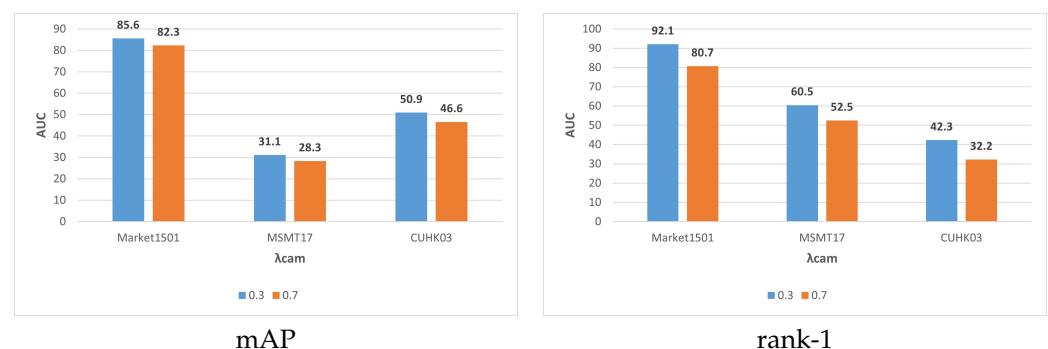


Figure 5. Effect of λ_{cam} .

Sensitivity to Momentum μ . The effect of the momentum coefficient μ on model performance is systematically evaluated in Figure 6. Across all three datasets, the optimal value is consistently observed at $\mu = 0.2$, achieving a peak performance of 85.6% mAP and 92.1% Rank-1 on Market-1501, 31.1% mAP and 60.5% Rank-1 on MSMT17, and 50.9% mAP and 42.3% Rank-1 on CUHK03. Performance degradation is observed when μ deviates from this optimal value, particularly when $\mu > 0.3$, demonstrating the importance of balanced momentum for stable memory updates.

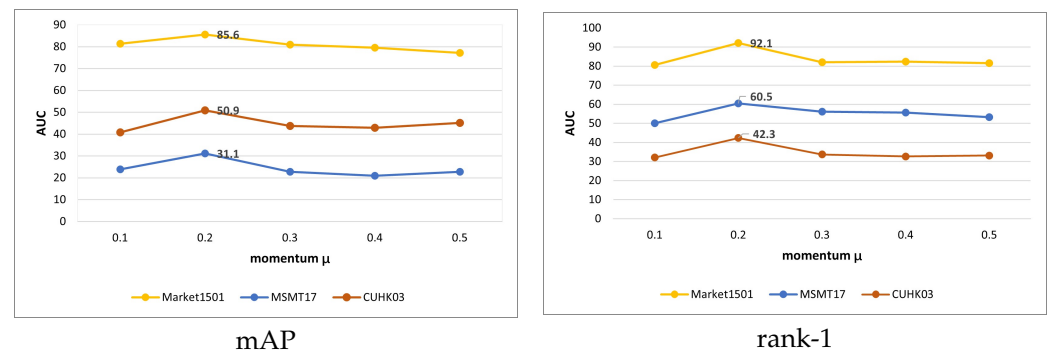


Figure 6. Effect of momentum μ .

Sensitivity to the number of epochs. Figure 7 illustrates the convergence behavior across datasets. Market-1501 reaches peak performance at epoch 60 (mAP: 85.6%, Rank-1: 92.1%) with slight degradation thereafter, indicating potential overfitting. In contrast, MSMT17 shows continuous improvement throughout the 80-epoch training process, achieving final scores of 31.1% mAP and 60.5% Rank-1. CUHK03 demonstrates intermediate behavior, stabilizing around epoch 70 with final performance of 50.9% mAP and 42.3% Rank-1.

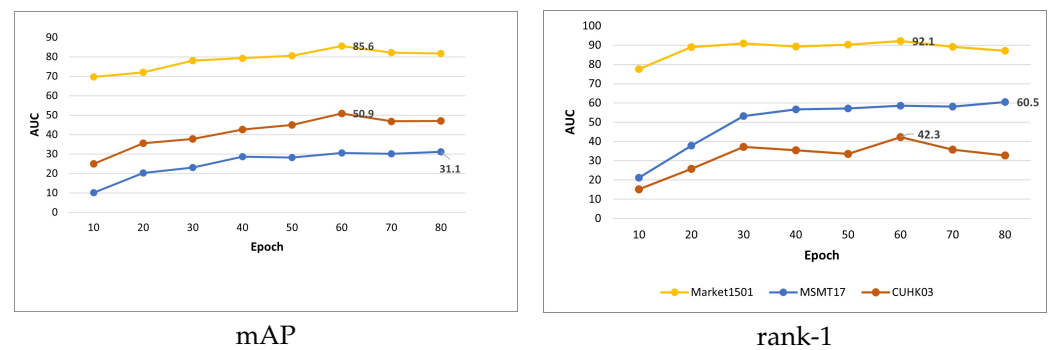


Figure 7. Effect of the number of epochs.

4.8. Computational Complexity

We compare the computation requirements of our method with several state-of-the-art unsupervised Re-ID methods in Table 5. The proposed ICCL framework introduces additional memory dictionaries and loss computations compared to standard contrastive learning baselines like Cluster Contrast [14]. This results in a moderate increase in training time per epoch and GPU memory usage, as our method maintains three memory banks (cluster, instance, and camera-aware) and computes three corresponding loss terms.

Table 5. A comparison of computation requirements on market-1501.

Methods	Setting	Time (s/ep)	GPU (GB)	Params (M)	Rank-1	mAP
BUC [55]	USL	110	3.5	12.3	66.2	38.3
MMCL [25]	USL	135	4.1	12.3	80.3	45.5
SPCL [18]	UDA	140	4.3	12.3	89.7	77.5
Cluster Contrast [14]	USL	125	3.8	12.3	92.3	82.1
ICCL (Ours)	USL	128	4.0	12.3	93.9	85.3

However, the additional overhead is highly efficient and manageable. During inference, our method shares the same backbone network and feature dimension as other methods, resulting in identical inference speeds and an identical parameter count. The core advantage of ICCL lies in its significantly enhanced discriminative power learned during

training, which requires only a marginal computational trade-off. While the performance of SPCL [18] is close to ours on Market-1501, it is important to note that SPCL is a more complex framework that involves both source and target domain data (UDA setting), whereas our method is fully unsupervised (USL). In future work, we will focus on further optimizing the clustering and memory update processes to enhance efficiency.

To further validate the scalability and inference performance of ICCL, we report frames-per-second (FPS) and per-image latency in Table 6. On Market-1501, ICCL achieves 124 FPS (≈ 8.1 ms per image), outperforming prior baselines such as SPCL (110 FPS/9.1 ms) and Cluster Contrast (118 FPS/8.4 ms). For the larger MSMT17 dataset, we report results for ICCL and Cluster Contrast, as other baselines do not provide publicly reproducible code for large-scale inference measurement. ICCL maintains stable efficiency of 117 FPS (≈ 8.5 ms) on MSMT17, confirming that its computational load scales linearly with dataset size. These results collectively demonstrate that ICCL achieves strong scalability and inference efficiency without sacrificing accuracy.

Table 6. Inference-time efficiency comparison on Market-1501 and MSMT17. Upward arrow means higher is better, downward arrow means lower is better.

Method	Dataset	FPS $\uparrow$	Latency (ms) $\downarrow$
SPCL [18]	Market-1501	110	9.1
Cluster Contrast [14]	Market-1501	118	8.4
ICCL (Ours)	Market-1501	124	8.1
Cluster Contrast [14]	MSMT17	110	9.1
ICCL (Ours)	MSMT17	117	8.5

5. Conclusions

This paper has presented the Invariance Constraint Contrast Learning (ICCL) framework for fully unsupervised person re-identification. By jointly enforcing center, instance, and camera-level invariances, ICCL effectively narrows the centroid misalignment gap and substantially improves representation robustness and benchmark performance.

While ICCL demonstrates strong generalization across diverse datasets, its performance may still be affected by missing or highly noisy camera identifiers. In future work, we plan to explore adaptive camera-invariance estimation and cross-domain generalization strategies to further enhance robustness and scalability.

Author Contributions: Conceptualization, L.W., C.L. and W.G.; methodology, L.W. and X.W.; software, X.G., X.W. and S.Z.; validation, L.W. and C.L.; formal analysis, L.W., X.W. and W.G.; investigation, C.L.; resources, L.W., X.W. and W.G.; data curation, X.G. and S.Z.; writing—original draft preparation, L.W.; writing—review and editing, W.G. and X.W.; visualization, C.L. and S.Z.; supervision, W.G.; project administration, W.G.; funding acquisition, L.W. and W.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Basic Research Program of Jiangsu, Grants No. BK20240313 and Wuxi Young Scientific and Technological Talent Support Initiative, Grants No. TJXD-2024-203.

Data Availability Statement: The data presented in this study are available in the Market-1501 dataset repository at https://drive.google.com/file/d/0B8-rUzbwVRk0c054eEozWG9COHM/view?resourcekey=0-8nyl7K9_x37HIQm34MmrYQ (accessed on 7 June 2025). The source codes are available at <https://github.com/cybergeek666/ICCL> (accessed on 7 June 2025).

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

References

- Huang, Y.; Huang, Y.; Hu, H.; Chen, D.; Su, T. Deeply Associative Two-Stage Representations Learning Based on Labels Interval Extension Loss and Group Loss for Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4526–4539. [\[CrossRef\]](#)
- Shen, C.; Qi, G.; Jiang, R.; Jin, Z.; Yong, H.; Chen, Y.; Hua, X. Sharp Attention Network via Adaptive Sampling for Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 3016–3027. [\[CrossRef\]](#)
- Ge, Y.; Li, Z.; Zhao, H.; Yin, G.; Yi, S.; Wang, X.; Li, H. FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification. In *Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, QC, Canada, 3–8 December 2018*; Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; NeurIPS: San Diego, CA, USA, 2018; pp. 1230–1241.
- Zhai, Y.; Lu, S.; Ye, Q.; Shan, X.; Chen, J.; Ji, R.; Tian, Y. AD-Cluster: Augmented Discriminative Clustering for Domain Adaptive Person Re-Identification. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020*; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2020; pp. 9018–9027.
- Liu, X.; Zhang, S.; Yang, M. Self-Guided Hash Coding for Large-Scale Person Re-identification. In *Proceedings of the 2nd IEEE Conference on Multimedia Information Processing and Retrieval, MIPR 2019, San Jose, CA, USA, 28–30 March 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 246–251.
- Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; Gu, J. A Strong Baseline and Batch Normalization Neck for Deep Person Re-Identification. *IEEE Trans. Multim.* **2020**, *22*, 2597–2609. [\[CrossRef\]](#)
- Ding, Y.; Fan, H.; Xu, M.; Yang, Y. Adaptive Exploration for Unsupervised Person Re-identification. *ACM Trans. Multim. Comput. Commun. Appl.* **2020**, *16*, 3:1–3:19. [\[CrossRef\]](#)
- Chen, Y.; Zhu, X.; Gong, S. Instance-Guided Context Rendering for Cross-Domain Person Re-Identification. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 232–242.
- Lin, Y.; Xie, L.; Wu, Y.; Yan, C.; Tian, Q. Unsupervised Person Re-Identification via Softened Similarity Learning. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020*; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2020; pp. 3387–3396.
- Zeng, K.; Ning, M.; Wang, Y.; Guo, Y. Hierarchical Clustering With Hard-Batch Triplet Loss for Person Re-Identification. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020*; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2020; pp. 13654–13662.
- Lee, Y.G.; Chen, S.C.; Hwang, J.N.; Hung, Y.P. An ensemble of invariant features for person reidentification. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 470–483. [\[CrossRef\]](#)
- Zhang, G.; Zhang, H.; Lin, W.; Chandran, A.K.; Jing, X. Camera contrast learning for unsupervised person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 4096–4107. [\[CrossRef\]](#)
- 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, 19–20 June 2022*; pp. 7308–7318.
- Dai, Z.; Wang, G.; Zhu, S.; Yuan, W.; Tan, P. Cluster Contrast for Unsupervised Person Re-Identification. *arXiv* **2021**, arXiv:2103.11568. [\[CrossRef\]](#)
- Zhang, X.; Li, D.; Wang, Z.; Wang, J.; Ding, E.; Shi, J.Q.; Zhang, Z.; Wang, J. Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022*; pp. 7369–7378.
- Ester, M.; Kriegel, H.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, USA, 2–4 August 1996*; Simoudis, E., Han, J., Fayyad, U.M., Eds.; AAAI Press: Palo Alto, CA, USA, 1996; pp. 226–231.
- Kojima, K.I. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. *Am. J. Hum. Genet.* **1969**, *21*, 407.
- Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; Li, H. Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID. In *Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual, 6–12 December 2020*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; NeurIPS: San Diego, CA, USA, 2020.
- Wang, H.; Zhu, X.; Xiang, T.; Gong, S. Towards unsupervised open-set person re-identification. In *Proceedings of the 2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, 25–28 September 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 769–773.

20. Chen, H.; Lagadec, B.; Brémond, F. ICE: Inter-instance Contrastive Encoding for Unsupervised Person Re-identification. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 14940–14949.
21. Wang, M.; Lai, B.; Huang, J.; Gong, X.; Hua, X. Camera-Aware Proxies for Unsupervised Person Re-Identification. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 2–9 February 2021; AAAI Press: Palo Alto, CA, USA, 2021; pp. 2764–2772.
22. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. *arXiv* **2017**, arXiv:1703.07737. [[CrossRef](#)]
23. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; IEEE Computer Society: Piscataway, NJ, USA, 2015; pp. 815–823.
24. Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748.
25. Wang, D.; Zhang, S. Unsupervised Person Re-Identification via Multi-Label Classification. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2020; pp. 10978–10987.
26. Ge, Y.; Chen, D.; Li, H. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
27. 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 Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Piscataway, NJ, USA, 2018; pp. 79–88.
28. Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-Image Domain Adaptation With Preserved Self-Similarity and Domain-Dissimilarity for Person Re-Identification. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Piscataway, NJ, USA, 2018; pp. 994–1003.
29. Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; Yang, Y. Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2019; pp. 598–607.
30. Li, Y.; Lin, C.; Lin, Y.; Wang, Y.F. Cross-Dataset Person Re-Identification via Unsupervised Pose Disentanglement and Adaptation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 7918–7928.
31. Chen, S.; Harandi, M.; Jin, X.; Yang, X. Domain Adaptation by Joint Distribution Invariant Projections. *IEEE Trans. Image Process.* **2020**, *29*, 8264–8277. [[CrossRef](#)] [[PubMed](#)]
32. Zhang, X.; Cao, J.; Shen, C.; You, M. Self-Training With Progressive Augmentation for Unsupervised Cross-Domain Person Re-Identification. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 8221–8230.
33. Ge, Y.; Zhu, F.; Zhao, R.; Li, H. Structured Domain Adaptation with Online Relation Regularization for Unsupervised Person Re-ID. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *35*, 258–271. [[CrossRef](#)] [[PubMed](#)]
34. Liu, X.; Zhang, S. Domain Adaptive Person Re-Identification via Coupling Optimization. In *Proceedings of the MM '20: The 28th ACM International Conference on Multimedia, Virtual Event, Seattle, WA, USA, 12–16 October 2020*; Chen, C.W., Cucchiara, R., Hua, X., Qi, G., Ricci, E., Zhang, Z., Zimmermann, R., Eds.; ACM: New York, NY, USA, 2020; pp. 547–555.
35. Jin, X.; Lan, C.; Zeng, W.; Chen, Z. Global Distance-Distributions Separation for Unsupervised Person Re-identification. In *Proceedings of the Computer Vision-ECCV 2020-16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part VII; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12352, pp. 735–751.
36. Fu, Y.; Wei, Y.; Wang, G.; Zhou, Y.; Shi, H.; Huang, T.S. Self-Similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-Identification. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 6111–6120.
37. Fan, H.; Zheng, L.; Yan, C.; Yang, Y. Unsupervised Person Re-identification: Clustering and Fine-tuning. *ACM Trans. Multim. Comput. Commun. Appl.* **2018**, *14*, 83:1–83:18. [[CrossRef](#)]
38. Lin, Y.; Dong, X.; Zheng, L.; Yan, Y.; Yang, Y. A Bottom-Up Clustering Approach to Unsupervised Person Re-Identification. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019; AAAI Press: Palo Alto, CA, USA, 2019; pp. 8738–8745.

39. Zhao, F.; Liao, S.; Xie, G.; 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*; Proceedings, Part XI; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12356, pp. 526–544.
40. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep Mutual Learning. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018*; Computer Vision Foundation/IEEE Computer Society: Piscataway, NJ, USA, 2018; pp. 4320–4328.
41. Pang, Z.; Wang, C.; Zhao, L.; Liu, Y.; Sharma, G. Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 2706–2718. [[CrossRef](#)]
42. Yang, B.; Chen, J.; Ye, M. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023*; pp. 11069–11079.
43. Grill, J.; Strub, F.; Althé, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.Á.; Guo, Z.; Azar, M.G.; et al. Bootstrap Your Own Latent—A New Approach to Self-Supervised Learning. In *Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual, 6–12 December 2020*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; NeurIPS: San Diego, CA, USA, 2020.
44. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G.E. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event, 13–18 July 2020*; Proceedings of Machine Learning Research; PMLR: Birmingham, UK, 2020; Volume 119, pp. 1597–1607.
45. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R.B. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020*; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2020; pp. 9726–9735.
46. Wang, S.; Park, S.; Kim, J.; Kim, J. Safety helmet monitoring on construction sites using YOLOv10 and advanced transformer architectures with surveillance and body-worn cameras. *J. Constr. Eng. Manag.* **2025**, *151*, 04025186. [[CrossRef](#)]
47. Wang, S. Automated non-PPE detection on construction sites using YOLOv10 and transformer architectures for surveillance and body worn cameras with benchmark datasets. *Sci. Rep.* **2025**, *15*, 27043. [[CrossRef](#)] [[PubMed](#)]
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016*; IEEE Computer Society: Piscataway, NJ, USA, 2016; pp. 770–778.
49. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable Person Re-identification: A Benchmark. In *Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015*; IEEE Computer Society: Piscataway, NJ, USA, 2015; pp. 1116–1124.
50. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; Devito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*.
51. Wu, J.; Liu, H.; Yang, Y.; Lei, Z.; Liao, S.; Li, S.Z. Unsupervised Graph Association for Person Re-Identification. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 8320–8329.
52. Li, M.; Zhu, X.; Gong, S. Unsupervised Person Re-identification by Deep Learning Tracklet Association. In *Proceedings of the Computer Vision-ECCV 2018-15th European Conference, Munich, Germany, 8–14 September 2018*; Proceedings, Part IV; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11208, pp. 772–788.
53. Li, M.; Zhu, X.; Gong, S. Unsupervised tracklet person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1770–1782. [[CrossRef](#)] [[PubMed](#)]
54. Zhai, Y.; Ye, Q.; Lu, S.; Jia, M.; Ji, R.; Tian, Y. Multiple expert brainstorming for domain adaptive person re-identification. In *Proceedings of the European Conference on Computer Vision, Online, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 594–611.
55. Qi, L.; Wang, L.; Huo, J.; Zhou, L.; Shi, Y.; Gao, Y. A Novel Unsupervised Camera-Aware Domain Adaptation Framework for Person Re-Identification. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 8079–8088.

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.