Identity-Seeking Self-Supervised Representation Learning for Generalizable Person Re-identification

Zhaopeng Dou[1,2], Zhongdao Wang[1,2], Yali Li[1,2], and Shengjin Wang[1,2*]

[1] Department of Electronic Engineering, Tsinghua University, China

[2] Beijing National Research Center for Information Science and Technology (BNRist), China

dcp19@mails.tsinhua.edu.cn   zhongdwang@gmail.com

liyali13@tsinghua.edu.cn   wgsgj@tsinghua.edu.cn*

Abstract

*This paper aims to learn a domain-generalizable (DG) person re-identification (ReID) representation from large-scale videos **without any annotation**. Prior DG ReID methods employ limited labeled data for training due to the high cost of annotation, which restricts further advances. To overcome the barriers of data and annotation, we propose to utilize large-scale unsupervised data for training. The key issue lies in how to mine identity information. To this end, we propose an Identity-seeking Self-supervised Representation learning (ISR) method. ISR constructs positive pairs from inter-frame images by modeling the instance association as a maximum-weight bipartite matching problem. A reliability-guided contrastive loss is further presented to suppress the adverse impact of noisy positive pairs, ensuring that reliable positive pairs dominate the learning process. The training cost of ISR scales approximately linearly with the data size, making it feasible to utilize large-scale data for training. The learned representation exhibits superior generalization ability. **Without human annotation and fine-tuning, ISR achieves 87.0% Rank-1 on Market-1501 and 56.4% Rank-1 on MSMT17**, outperforming the best supervised domain-generalizable method by 5.0% and 19.5%, respectively. In the pre-training→fine-tuning scenario, ISR achieves state-of-the-art performance, with 88.4% Rank-1 on MSMT17. The code is at* https://github.com/dcp15/ISR_ICCV2023_Oral.

1. Introduction

Person re-identification (ReID) aims to retrieve a person across non-overlapping camera views [60, 6, 86, 39, 14]. Although current full-supervised ReID methods [28, 31, 34, 42, 65, 22] have shown encouraging results on public
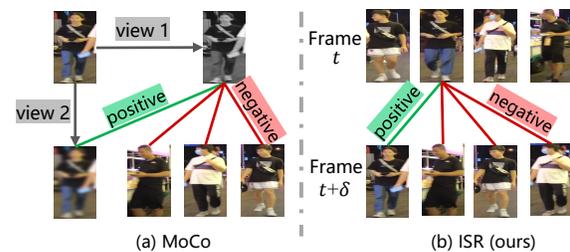
---

* Corresponding author

Figure 1. The core idea. (a) MoCo [25] learns instance discrimination, which aims to learn a unique representation for each image. (b) ISR learns identity discrimination, which aims to learn similar representations for inter-frame images with the same identity.

benchmarks, their performance significantly declines when applied to unseen domains. To enhance the generalization ability, unsupervised domain adaptation (UDA) [15, 84, 19, 1, 53] and domain generalizable (DG) ReID [57, 33, 12] techniques are extensively studied. However, both methods have limitations that prevent their scaling in real-world applications. UDA ReID necessitates well-organized data from the target domain for adaptation. DG ReID, on the other hand, typically employs small-scale labeled training data due to the high cost of annotation, which hinders further progress. To this end, we propose to break through the constraints of human annotation and target domain adaptation on the generalization of person ReID.

This paper aims to learn domain-generalizable representations ***without any annotation***. The representation can be applied directly in arbitrary domains without fine-tuning. Unlike prior DG ReID methods [57, 33, 12] that require labeled training data, we employ vast amounts of unlabeled internet videos for training. Specifically, our training set comprises 47.8 million unlabeled person images extracted from 74,000 video clips. The feasibility comes from two key factors: the low cost of acquiring videos from the internet and the diverse domains present in large-scale videos. We learn a domain-generalizable ReID representation from

the large-scale unsupervised training data. The representation is robust to unseen domains, exhibiting high application potential and value for open-world scenarios.

The core issue is learning *identity discrimination*. Inspired by the great success of contrastive learning [25, 21, 8, 10] in large-scale unsupervised data, we propose to learn identity discrimination through contrastive information between samples. It is not feasible to directly apply conventional unsupervised contrastive learning methods to ReID, because the objective of their pretext task (*i.e.*, *instance discrimination*) conflicts with ReID required identity discrimination. For example, as in Figure 1(a), MoCo [25] regards two augmented views of an image as a positive pair, which prompts the model to learn a unique representation for each image. However, ReID aims to discriminate between identities instead of instances, which expects multiple images from the same identity to have similar representations.

To achieve the objective of identity discrimination, we propose an Identity-seeking Self-supervised Representation learning (ISR) method. ISR aims to learn similar representations for inter-frame images belonging to the same identity, as illustrated in Figure 1(b). We first formulate the instance association between two frames as a maximum-weight bipartite matching problem, where positive pairs are mined by solving the optimal matching strategy. However, these positive pairs inevitably contain noise due to the unsupervised construction, which considerably undermines representation learning. Next, we introduce a reliability-guided contrastive loss to mitigate the adverse impact of noisy positive pairs, ensuring that reliable positive pairs dominate the learning process. Moreover, the training cost of ISR scales approximately linearly with the size of the data, making it feasible to train on large-scale data.

Extensive experiments demonstrate the effectiveness of ISR. Under the domain-generalizable settings, ISR with ResNet50 [26] backbone achieves 45.7% Rank-1 and 21.2% mAP on the most challenging dataset MSMT17 [74], outperforming the best-supervised domain-generalizable method trained with multiple labeled datasets by +8.8% and +6.5%, respectively. When using a more data-hungry backbone, *i.e.*, Swin-Transformer [44], the performance is even further improved, achieving 56.4% Rank-1 and 30.3% mAP on MSMT17. We also evaluate ISR under other practical settings, such as pre-training for supervised fine-tuning or few-shot learning. ISR shows consistent improvements upon existing methods. For example, when serving as a pre-trained model for supervised fine-tuning, ISR achieves 88.4% Rank-1 on MSMT17, setting a new state-of-the-art.

The main contributions of this paper are:

- We propose an Identity-seeking Self-supervised Representation learning (ISR) method that can learn a domain-generalizable ReID representation from large-scale video data *without any annotation*.

- We propose a novel reliability-guided contrastive loss that effectively mitigates the adverse impact of noisy positive pairs, making reliable positive pairs dominate the representation learning process.

- Extensive experiments verify the effectiveness of ISR. Notably, ISR achieves 87.0% Rank-1 on Market-1501 and 56.4% Rank-1 on MSMT17 without human annotation and fine-tuning, outperforming the best supervised DG method by 5.0% and 19.5%, respectively.

2. Related Work

Person Re-identification. Person re-identification (ReID) [60, 86, 14] aims to retrieve a person in a large database from disjoint cameras. Deep learning [28, 31, 34, 65] has dominated the ReID community due to its powerful representation capabilities. The deep learning methods can be roughly divided into but not limited to, deep metric methods [9, 29, 59], part-based methods [60, 63, 14, 23], and attention-based methods [2, 4, 67]. Although these methods have achieved promising results on public benchmarks, they show poor generalization ability on unseen domains. As a result, unsupervised domain adaption (UDA) ReID [53, 1, 84] is proposed. UDA ReID aims at fine-tuning the source-trained model to the target domain using unlabeled target domain data. However, it is still not powerful enough for practical applications because it is sometimes difficult to collect well-organized target domain data for fine-tuning.

Domain Generalizable (DG) ReID. Domain generalizable ReID [57, 36, 11, 81, 12, 49, 35, 79, 77] aims to learn a robust model on the source domain and test it directly on unseen target domains without fine-tuning. It has attracted extensive attention due to its great potential in practical applications. DIMN [57] designs a Domain-Invariant Mapping Network to learn domain-invariant representation under a meta-learning pipeline. MetaBIN [11] and SNR [36] study the normalization layer or module to improve the model generalization. RaMoE [12] leverages the relevance between the target domain and multiple source domains to improve the model's generalization. MDA [49] aligns source and target feature distributions to a prior distribution. These methods are trained with small-scale domain-starved labeled data. Unlike them, we aim to learn a DG ReID model from large-scale domain-diverse unlabeled data.

Synthetic Data for ReID. The performance of ReID models is limited by the high cost of collecting annotated data from the real world. To address this challenge, several methods have turned to using synthetic data. Notably, PersonX [58] contains 1,266 identities with 273,456 images captured from various viewpoints, enabling exploration of viewpoint impacts on ReID systems. RandPerson [70] offers 8,000 identities with 228,655 images from 19 cameras, while UnrealPerson [78] provides 3,000 identities with

120,000 images from 34 cameras, and ClonedPerson [69] includes 5,621 identities with 887,766 images from 24 cameras. These synthetic datasets have proven valuable for supervised learning, as they enhance the generalization ability of ReID models. DomainMix [66] further establishes that combining labeled synthetic data with unlabeled real-world data during training is a promising direction for DG ReID. However, there remains a significant domain gap between synthetic and real-world data, impeding the seamless application of models trained on synthetic data to authentic real-world scenarios. To bridge this disparity, we propose to use vast amounts of unlabeled real-world data for training.

Unsupervised Representation Learning. In recent years, unsupervised representation learning methods [25, 8, 21, 24, 10] have shown promising results in learning pre-training representations from large-scale unlabeled data. MoCo [25] introduces a memory queue to generate negative samples. SimCLR [8] leverages a projection head and rich data augmentations to achieve robust pre-training representation. BYOL [21] learns the representation without using negative pairs. However, if these methods are applied directly to ReID, only a pre-training model can be learned, which shows extremely low accuracy when tested directly. The core reason is that they regard two views of an image as a positive pair or reconstruct masked pixels in an image, resulting in instance discrimination. This contradicts the ReID objective of identity discrimination. Unlike them, we regard inter-frame images with the same identity as positive pairs to align with the identity discrimination objective. A closely related work is CycAs [72] and its improved version [71]. CycAs enforces cycle consistency of instance association between video frame pairs to learn ReID representation. However, CycAs exhibits weaknesses in scenarios where certain instances lack corresponding matches or where association errors occur. Our method significantly differs from CycAs as we mine positive pairs and effectively suppress noise within them, offering a robust solution to learning a generalizable ReID representation.

3. Identity-Seeking Represenatation Learning

We present an Identity-seeking Self-supervised Representation (ISR) learning method to learn a domain-generalizable ReID representation from extensive videos without any annotation. ISR constructs positive pairs from inter-frame images to align with the ReID objective of *identity discrimination* (Sec. 3.1). To mitigate the adverse effects of noisy positive pairs, we incorporate a reliability-guided contrastive loss (Sec. 3.2). The training cost of ISR scales approximately linear with the amount of training data, making it feasible to train with large-scale data.
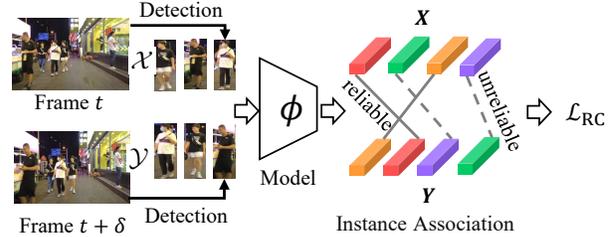

Figure 2. The overview of ISR. We first crop two image sets $\mathcal{X}$ and $\mathcal{Y}$ from a video frame pair. They are fed into the model ϕ to obtain the features $\boldsymbol{X}$ and $\boldsymbol{Y}$. We mine the positive pairs by modeling the association between $\boldsymbol{X}$ and $\boldsymbol{Y}$ as a maximum-weight bipartite matching problem. We propose a reliability-guided contrastive loss $\mathcal{L}_{\mathrm{RC}}$ to suppress the effect of noisy pseudo labels.

3.1. Overview

The core of our method is to consider inter-frame images with the same identity as positive pairs for contrastive learning. As shown in Figure 2, we crop the person images from two frames of a video with the help of an off-the-shelf pedestrian detection model [73] to form two image sets $\mathcal{X}$ and $\mathcal{Y}$. The time interval δ between two frames requires a trade-off. If δ is too large, $\mathcal{X}$ and $\mathcal{Y}$ might not have identity overlap. Conversely, if δ is too small, $\mathcal{X}$ and $\mathcal{Y}$ are too similar to provide sufficient contrastive information. In practice, we define a maximum time interval $\delta_{\max}$. We randomly select two frames as long as their interval does not exceed $\delta_{\max}$. $\mathcal{X}$ and $\mathcal{Y}$ are fed into a model ϕ to extract the features $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m] \in \mathbb{R}^{d \times m}$ and $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n] \in \mathbb{R}^{d \times n}$, where d is the feature dimension, m and n are the number of images in $\mathcal{X}$ and $\mathcal{Y}$, respectively. We assume $m \leq n$; otherwise, we can swap $\boldsymbol{X}$ and $\boldsymbol{Y}$ to ensure this. All features are l_2-normalized.

The positive pairs are mined by formulating the instance association between $\boldsymbol{X}$ and $\boldsymbol{Y}$ as a maximum-weight bipartite matching problem. Let $\boldsymbol{\pi} \in \mathbb{R}^{m \times n}$ be a boolean association matrix, where $\boldsymbol{\pi}_{ij} = 1$ if $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are matched; otherwise, $\boldsymbol{\pi}_{ij} = 0$. The cost of matching $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ is defined as the cosine distance between them, *i.e.*, $c_{ij} = 1 - \boldsymbol{x}_i \cdot \boldsymbol{y}_j$. One of our insights is that if the association matrix is optimal, the total matching cost should be minimum, which is formulated as:

$$
\begin{aligned}
\min_{\boldsymbol{\pi}} \quad & \sum_{ij} \boldsymbol{c}_{ij} \boldsymbol{\pi}_{ij} \\
\text{s.t.} \quad & \sum_j \boldsymbol{\pi}_{ij} = 1, \qquad i \in \{1, 2, \ldots, m\} \\
& \sum_i \boldsymbol{\pi}_{ij} = 0 \text{ or } 1, \qquad j \in \{1, 2, \ldots, n\}
\end{aligned} \quad (1)
$$

The optimal association matrix $\boldsymbol{\pi}^*$ can be solved in polynomial time by the Hungarian algorithm [37] (please see supplementary materials for details). In general, there are finite pedestrians in a frame, so solving the optimal $\boldsymbol{\pi}^*$ brings limited computational overhead. This allows positive pairs

to be mined synchronously during training, benefiting from each other with the learning of ReID model.

After solving the optimal association matrix $\boldsymbol{\pi}^*$, we define two matched samples as a positive pair. An intuitive way is to apply the loss function of MoCo [25] to enforce the positive pairs to be similar and negative pairs to be dissimilar. However, such a way yields unsatisfactory results in our experiments (Table 5). We find such failure is caused by the inevitable noisy positive pairs. To address this issue, we propose a reliability-guided contrastive loss (Sec. 3.2) to suppress the impact of noisy positive pairs during training.

Identity discrimination *vs*. instance discrimination. We analyze the feasibility of directly applying conventional unsupervised contrastive learning methods [25, 8, 21, 10] to ReID. Without loss of generality, we take the widely used MoCo [25] as an example. MoCo considers two augmented views of an image to be a positive pair, which enforces the model to learn a unique representation for each image, leading to *instance discrimination*. However, ReID necessitates *identity discrimination*. Our method regards inter-frame images with the same identity as positive pairs and requires them to have similar representations, which aligns with the objective of ReID. As shown in Table 1, under domain-generalizable settings, our method outperforms MoCo by a large margin, exhibiting great superiority in the ReID task.

3.2. Noisy Positive Pair Suppression

There is inevitable noise in the mined positive pairs. Noisy positive pairs mainly come from two aspects: (1) the feature extractor ϕ is imperfect, which makes the matching cost inaccurate, especially in the early stages of training; (2) some pedestrians appearing in $\boldsymbol{X}$ do not appear in $\boldsymbol{Y}$, which is fatal because bipartite matching strategy constrains that for every sample in $\boldsymbol{X}$, there must be a sample in $\boldsymbol{Y}$ associated to it. Noisy positive pairs will seriously disrupt the learning of embedding space. To remedy this, we propose to measure the reliability of positive pairs and further suppress the adverse impact of noisy positive pairs during training.

Our intuition is that if two images belong to the same identity, their similarity should be large and vice versa. The similarity of two samples in the positive pair can reflect the reliability. If the similarity between two samples within a mined positive pair is very low, the positive pair is likely to be noisy. Besides, since the positive pairs are mined by solving the bipartite matching problem, the reliability should also consider the similarities between the target sample and other samples. Combing these two factors, for $\boldsymbol{x}_i$, the reliability score of the corresponding positive pair is,

$$p(\boldsymbol{x}_i) = \frac{\sum_j \boldsymbol{\pi}^*_{ij} \exp(\boldsymbol{x}_i \cdot \boldsymbol{y}_j / \tau)}{\sum_j \exp(\boldsymbol{x}_i \cdot \boldsymbol{y}_j / \tau)} \quad (2)$$

where $\boldsymbol{\pi}^*$ is the optimal association matrix and τ is the temperature hyper-parameter. The reliability is the softmax-
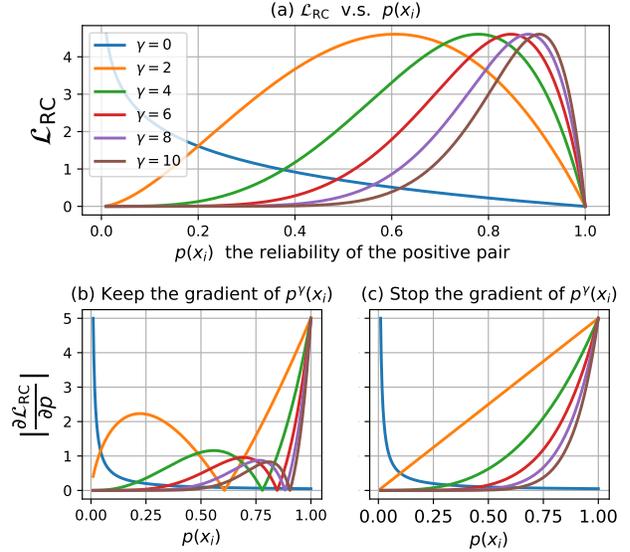

Figure 3. (a) The contrastive loss versus the reliability of the positive pair, *i.e.*, $\mathcal{L}_{\mathrm{RC}}(\boldsymbol{x}_i)$ *vs*. $p(\boldsymbol{x}_i)$. Setting $\gamma > 0$ suppresses the adverse impacts of positive pairs with low reliability scores. (b) The absolute value of $\partial \mathcal{L}_{\mathrm{RC}}(\boldsymbol{x}_i) / \partial p(\boldsymbol{x}_i)$ when stopping the gradient of $p^\gamma(\boldsymbol{x}_i)$. (c) The absolute value of $\partial \mathcal{L}_{\mathrm{RC}}(\boldsymbol{x}_i) / \partial p(\boldsymbol{x}_i)$ when keeping the gradient of $p^\gamma(\boldsymbol{x}_i)$. For a better observation, we scale up the curves under different γ so that their maxima are equal.

based probability that tries to align $\boldsymbol{x}_i$ with its positive one in the set $\boldsymbol{Y}$. The calculation of reliability is extremely low overhead, making it suitable for large-scale training.

We propose the reliability score $p(\boldsymbol{x}_i)$ to be a modulation factor to down-weight the noisy positive pairs and thus focus training on reliable positive pairs. At the same time, the modulation factor should not disrupt the gradient of the loss, so we break its gradient in back-propagation to ensure the stability of the training process. Formally, the reliability-guided contrastive loss is defined as,

$$\mathcal{L}_{\mathrm{RC}}(\boldsymbol{x}_i) = -p^\gamma_{\leftarrow}(\boldsymbol{x}_i) \log(p(\boldsymbol{x}_i)) \quad (3)$$

where $p^\gamma_{\leftarrow}(\boldsymbol{x}_i)$ indicates that the gradient of $p^\gamma(\boldsymbol{x}_i)$ is cut-off in the back-propagation. $\gamma \geq 0$ is the hyper-parameter controlling the strictness for reliability. The larger the γ, the more the loss function focuses on reliable positive pairs. In practice, the loss function $\mathcal{L}_{\mathrm{RC}}(\boldsymbol{x}_i)$ will decrease as γ increases. To maintain the size of the loss during training, we scale up the average $\mathcal{L}_{\mathrm{RC}}$ in the mini-batch,

$$\mathcal{L}_{\mathrm{RC}} = \frac{\alpha}{m} \sum_{i=0}^m \mathcal{L}_{\mathrm{RC}}(\boldsymbol{x}_i) \quad (4)$$

where $\alpha = \left(\sum_j \mathcal{L}_{\mathrm{RC}, \gamma=0}(\boldsymbol{x}_j) \right)_{\leftarrow} / \left(\sum_j \mathcal{L}_{\mathrm{RC}}(\boldsymbol{x}_j) \right)_{\leftarrow}$. This dynamically calibrates the size of the loss, weakening the influence of the learning rate in the optimal γ analysis.

Properties of $\mathcal{L}_{\text{RC}}(\boldsymbol{x}_i)$. We visualize $\mathcal{L}_{\text{RC}}(\boldsymbol{x}_i)$ and the absolute gradient $|\partial\mathcal{L}_{\text{RC}}(\boldsymbol{x}_i)/\partial p(\boldsymbol{x}_i)|$ for $\gamma \in [0,10]$ in Figure 3. We note several key properties of $\mathcal{L}_{\text{RC}}(\boldsymbol{x}_i)$:

Firstly, as in Figure 3a, when $\gamma = 0$, $\mathcal{L}_{\text{RC}}$ degenerates to a version without the reliability guidance. Noisy positive pairs comprise the majority of the loss, which seriously degrades representation learning. When $\gamma > 0$, as shown in Figure 3b, reliable positive pairs dominate the gradient back-propagation process, effectively mitigating the adverse impact of noisy positive pairs during learning.

Secondly, it is necessary to stop the gradient of $p^\gamma(\boldsymbol{x}_i)$. If the gradient of $p^\gamma(\boldsymbol{x}_i)$ is maintained, a slump point will occur for curves with $\gamma > 0$, as in Figure 3c. Consequently, positive pairs whose reliability lies in the trap lose the ability to optimize the model. Some noisy positive pairs exhibit large back-propagation gradients, severely compromising the representation learning. In our experiments, maintaining the gradient of $p^\gamma(\boldsymbol{x}_i)$ will result in NAN issues.

Lastly, $\mathcal{L}_{\text{RC}}(\boldsymbol{x}_i)$ significantly differs from the prior focal loss [43]. Focal loss is designed for training data with *high-quality labels* by assigning large training weights to hard samples. $\mathcal{L}_{\text{RC}}(\boldsymbol{x}_i)$ is designed for training data with *noisy pseudo-labels*, *i.e.*, positive pairs with noise, by assigning large training weights to reliable positive pairs. The modulating factor of focal loss is $(1 - p)^\gamma$, while the modulating factor of $\mathcal{L}_{\text{RC}}(\boldsymbol{x}_i)$ is $p_{\leftarrow}^\gamma$. The gradient cutoff of p^γ is specific and crucial for $\mathcal{L}_{\text{RC}}(\boldsymbol{x}_i)$. Besides, results in Table 6 validate that $\mathcal{L}_{\text{RC}}$ is superior to focal loss in our settings.

3.3. Training and inference

Training. Inspired by [68, 25], we adopt a memory queue to provide hard negative samples for contrastive learning. The queue uses a first-in, first-out update strategy. We regard two samples as a negative pair if they come from different videos. For $\boldsymbol{x}_i$, we select the most dissimilar k negative samples from the queue, denoted as $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_k\}$, to strengthen contrastive information,

$$\mathcal{L}_{\text{Q}}(\boldsymbol{x}_i) = \frac{1}{k}\sum_{j=1}^{k}\log(1 + \exp(\boldsymbol{x}_i \cdot \boldsymbol{f}_j)) \qquad (5)$$

Here, we introduce a "super" frame sampling strategy to improve the efficiency of training. For video b, two cropped image sets are denoted as $\mathcal{X}_b$ and $\mathcal{Y}_b$. The "super" frame pair is formed by merging several sets from different videos, *i.e.*, $\mathcal{X} = \bigcup_b \mathcal{X}_b$ and $\mathcal{Y} = \bigcup_b \mathcal{Y}_b$. Note that the positive pairs are mined between two frames of the same video. The "super" frame sampling strategy guarantees the domain diversity at each iteration. Finally, the total loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{RC}} + \lambda\mathcal{L}_{\text{Q}} \qquad (6)$$

where λ is a hyper-parameter to balance $\mathcal{L}_{\text{RC}}$ and $\mathcal{L}_{\text{Q}}$.

Training cost analysis: The number of pedestrians in a frame is limited, so the computational complexity of solving the bipartite-graph matching problem can be approximately regarded as $\mathcal{O}(1)$. Assuming that there are N frames in our training data, the training complexity of our method is $\mathcal{O}(N)$, which is roughly linear with the data size. This makes it possible to drive large-scale data for training. Please refer to the supplementary for related experiments.

Inference. If not specified, the learned representation is directly tested on target domains without fine-tuning.

4. Experiments

4.1. Implementation

Training dataset. Following LUP [16], LUPnl [17], and CycAs [71], we crawl the videos from youtube as our training data. The videos are searched by queries like "cityname+streeview". The "cityname" includes 100 big cities around the world, ensuring the diversity of the domains. Following [71], to build a high-quality training dataset, the processing includes: (1) use the PySceneDetect[1] toolkit to cut the raw video into multiple slices, eliminating the effects of shot changes within a video; (2) use the off-of-shelf pedestrian detection model JDE [73] to crop the person images every seven frames to reduce the temporal redundancy; (3) calculate the average number of pedestrians detected in five uniformly sampled frames in a video, removing videos with less than five pedestrians. Statistically, the training set contains 47.8M person images from 74K video clips.

Test datasets and protocols. We conduct extensive experiments on public ReID datasets. For domain generalizable (DG) ReID, following [12], we select two common protocols. Protocol-1 is the leave-one-out setting for Market-1501[82], DukeMTMC [54], CUHK03 [41], and MSMT17 [74], which selects one domain for testing (only the testing set) and all the other domains for training (including the training and testing sets). Protocol-2 selects all images in Market-1501, DukeMTMC, CUHK02 [40], CUHK03, and CUHK-SYSU [75] (including the training and testing sets) as the training images. Four small datasets, *i.e.*, PRID [30], GRID [47], VIPeR [20], and iLIDs [64] are used for testing. Following [57], we report the average of 10 random splits of gallery and probe sets. ***Since DukeMTMC has been withdrawn, we do not report the evaluation results on DukeMTMC***, which is introduced just to illustrate the training settings of other methods.

Training details. If not specified, we use ResNet50 [26] as our backbone. Swin-Transformer [44] is also adopted for fair comparisons. Similar to [48], the stride of ResNet50's last layer is set to 1. Inspired by [50], we add one IN [61] layer to the end of layer1 and layer2 of ResNet50, respectively, to improve the model generalization. For ResNet50 (Swin-Transformer), input images are resized to 256×128

---

[1] http://scenedetect.com/

Methods	Sup.	Protocol-1						Protocol-2							
		Market-1501		MSMT17		CUHK03		PRID		GRID		VIPeR		iLIDs	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
CrossGrad* [55]	✔	–	–	–	–	–	–	18.8	28.2	9.0	16.0	20.9	30.4	49.7	61.3
MLDG* [38]	✔	–	–	–	–	–	–	24.0	35.4	15.8	23.6	23.5	33.5	53.8	65.2
PPA* [52]	✔	–	–	–	–	–	–	31.9	45.3	26.9	38.0	45.1	54.5	64.5	72.7
DIMN* [57]	✔	–	–	–	–	–	–	39.2	52.0	29.3	41.1	51.2	60.1	70.2	78.4
SNR [36]	✔	–	–	–	–	–	–	52.1	66.5	40.2	47.7	52.9	61.3	84.1	89.9
ACL [77]	✔	–	–	–	–	–	–	63.0	73.4	55.2	65.7	**66.4**	**75.1**	81.8	86.5
MetaBIN [11]	✔	–	–	–	–	–	–	**74.2**	**81.0**	48.4	57.9	59.9	68.6	81.3	87.0
MDA [49]	✔	–	–	–	–	–	–	–	–	**61.2**	**62.9**	63.5	71.7	80.4	84.4
DDAN [5]	✔	–	–	–	–	–	–	62.9	67.5	46.2	50.9	56.5	60.8	78.0	81.2
DTIN [35]	✔	–	–	–	–	–	–	71.0	79.7	51.8	60.6	62.9	70.7	81.8	87.2
META [76]	✔	–	–	–	–	–	–	61.9	71.7	51.0	56.6	53.9	60.4	79.3	83.9
M³L [81]	✔	75.9	50.2	**36.9**	**14.7**	33.1	32.1	–	–	–	–	–	–	–	–
DML [12]	✔	75.4	49.9	24.5	9.9	32.9	32.6	47.3	60.4	39.4	49.0	49.2	58.0	77.3	84.0
RaMoE [12]	✔	**82.0**	**56.5**	34.1	13.5	**36.6**	**35.5**	57.7	67.3	46.8	54.2	56.6	64.6	**85.0**	**90.2**
TrackContrast [32]	✘	72.7	36.2	–	–	–	–	–	–	–	–	–	–	–	–
LUP† [16]	✘	3.3	1.0	0.3	0.1	0.1	0.5	1.5	3.7	1.2	4.0	1.4	5.0	36.7	43.0
MoCo (R50) [25]	✘	10.5	2.6	0.5	0.2	0.3	0.7	6.5	10.9	2.8	6.9	4.0	7.5	38.8	46.4
LUPnl† [17]	✘	13.8	3.8	0.6	0.2	0.4	0.8	8.1	12.2	3.1	7.4	4.6	9.2	43.3	49.8
CycAs (R50) [71]	✘	80.3	57.5	43.9	20.2	25.8	26.5	58.8	67.7	52.5	62.3	57.3	66.0	85.2	90.4
Ours (R50)	✘	**85.1**	**65.1**	**45.7**	**21.2**	**26.1**	**27.4**	59.7	70.8	55.8	65.2	**58.0**	**66.6**	87.6	91.7
CycAs (Swin) [71]	✘	82.2	60.4	49.0	24.1	36.3	37.1	71.5	79.2	55.8	66.4	60.2	68.8	87.4	91.3
Ours (Swin)	✘	**87.0**	**70.5**	**56.4**	**30.3**	**36.6**	**37.8**	**74.5**	**83.0**	**62.7**	**72.0**	**68.4**	**75.5**	87.5	91.5

Table 1. Comparison with state-of-the-art under DG settings. The performances of the methods marked by "*" are from [57]. "Sup." indicates the method is supervised or unsupervised. † indicates the results are tested with the pre-trained models published on their Github.

Method	Sup.	Training set	PersonX₄₅₆		UnrealPerson		ClonedPerson		RandPerson	
			R1	mAP	R1	mAP	R1	mAP	R1	mAP
BOT [48]	✔	M+D+C3+MT	87.7	72.7	56.1	43.5	41.0	7.9	35.8	16.8
MGN [63]	✔	M+D+C3+MT	91.2	77.8	58.1	45.7	49.1	11.9	50.1	27.1
MoCo (R50) [25]	✘	Unsup-videos	16.0	1.4	0.7	0.2	0.5	0.2	0.2	0.1
Ours (R50)	✘	Unsup-videos	**92.5**	**81.6**	**64.6**	**52.8**	**49.8**	**12.1**	**54.2**	**28.1**
Ours (Swin)	✘	Unsup-videos	**95.0**	**88.5**	**67.3**	**58.1**	**57.6**	**16.6**	**68.3**	**40.1**

Table 2. Comparison on the synthetic datasets: PersonX₄₅₆ [58], UnrealPerson [78], ClonedPerson [69] and RandPerson [70]. We use UnrealPerson-v1.3 and the subset of RandPerson for testing, each partitioned into query and gallery sets with a ratio of 1:4. "unsup-videos" indicates our large-scale unlabeled video dataset. M: Market [82], D: Duke [54], C3: CUHK03 [41], MT: MSMT17 [74].

(224×224). During training, images are augmented by random horizontal flipping and color jitter. We find suitable data augmentation is beneficial for unsupervised contrastive learning. Please refer to supplementary materials for related ablation studies. To guarantee the stability of training, the number of images in a "super" frame is 80 (more than 80 will be truncated) and we sample three frames from each video and team each other up to form three frame pairs. We adopt distributed training on $4 \times$ NVIDIA 3090GPU, so the batch size is $960 = 80 \times 3 \times 4$ (for Swin-Transformer, it is $336 = 28 \times 3 \times 4$). In an epoch, we sample every video 16 times. The model is trained for 50 epochs with the AdamW [46]. We use the Cosine Annealing learning rate [45] initialized with $1e^{-4}$. We set the hyper-parameter $\gamma = 6$ and $\lambda = 5$ in Eq. 3 and Eq. 6, respectively.

Evaluation metrics: Cumulative matching characteris-

tic (CMC) curve and mean average precision (mAP).

4.2. Domain Generalization (DG) Setting.

There are few unsupervised methods for domain generalizable (DG) ReID. Therefore we mainly compare our method with supervised DG ReID methods. We also compare five unsupervised methods that utilize large-scale video data similar to ours. ***Notably, our method uses the crawled unsupervised videos for training, and tests directly on these datasets.*** We report the results in Table 1. Results on DukeMTMC are not reported as it is retracted.

Compared to supervised methods: (1) When adopting ResNet50 as the backbone, the state-of-the-art supervised method to be compared is RaMoE [12]. Compared with RaMoE, our method with ResNet50 backbone outperforms it on most datasets (six out of seven). Especially in the

Datasets	Pre-train	Small-scale					Few-shot				
		10%	30%	50%	70%	90%	10%	30%	50%	70%	90%
Market	IN sup.	76.9 / 53.1	90.8 / 75.2	93.5 / 81.5	94.5 / 84.8	95.2 / 86.9	41.8 / 21.1	87.6 / 68.1	92.8 / 80.2	94.0 / 84.2	94.6 / 86.7
	IN unsup.	81.7 / 58.4	91.9 / 76.6	94.1 / 82.0	94.5 / 85.4	95.5 / 87.4	36.1 / 18.6	87.8 / 69.3	90.9 / 78.3	94.1 / 84.4	95.2 / 87.1
	MoCo [25]	81.8 / 58.8	92.3 / 77.7	94.3 / 84.0	95.4 / 87.3	95.9 / 89.2	41.5 / 22.0	87.7 / 69.8	93.9 / 83.1	94.8 / 86.8	96.0 / 89.3
	LUP [16]	85.5 / 64.6	93.7 / 81.9	94.9 / 85.8	95.9 / 88.8	96.4 / 90.5	47.5 / 26.4	92.1 / 78.3	93.9 / 84.2	95.5 / 88.4	96.3 / 90.4
	LUPnl [17]	88.8 / 72.4	94.2 / 85.2	95.5 / 88.3	96.2 / 90.1	96.4 / 91.3	61.6 / 42.0	94.0 / 83.7	95.2 / 88.1	96.3 / 90.5	96.4 / 91.6
	Ours	**90.5 / 75.3**	**94.8 / 86.2**	**96.2 / 89.3**	**96.6 / 90.9**	**96.7 / 91.8**	**75.9 / 54.3**	**94.1 / 84.9**	**96.0 / 89.4**	**96.4 / 91.3**	**96.8 / 92.1**
MSMT17	IN sup.	50.2 / 23.2	70.8 / 41.9	76.9 / 50.3	81.2 / 56.9	84.2 / 61.9	34.1 / 14.7	71.1 / 44.5	79.5 / 56.2	82.8 / 60.9	84.5 / 63.4
	IN unsup.	48.8 / 22.6	68.7 / 40.4	75.0 / 49.0	79.9 / 55.7	83.0 / 60.9	29.2 / 13.2	67.1 / 41.4	77.6 / 53.3	81.5 / 59.1	83.8 / 62.4
	MoCo [25]	46.8 / 21.5	69.4 / 41.7	76.7 / 50.8	81.3 / 58.3	84.7 / 63.9	23.4 / 9.9	65.9 / 40.7	77.4 / 54.2	82.7 / 61.5	85.0 / 65.3
	LUP [16]	51.1 / 25.5	71.4 / 44.6	77.7 / 53.0	81.8 / 59.5	85.0 / 63.7	36.0 / 17.0	73.6 / 49.0	80.5 / 57.4	83.5 / 62.9	85.1 / 65.0
	LUPnl [17]	51.1 / 28.2	71.2 / 47.7	77.2 / 55.5	81.8 / 61.6	84.8 / 66.1	42.7 / 24.5	74.4 / 53.2	81.0 / 62.2	83.8 / 65.8	85.3 / 67.4
	Ours	**64.2 / 36.2**	**78.1 / 52.9**	**82.5 / 60.0**	**85.4 / 65.7**	**87.5 / 69.6**	**59.4 / 33.4**	**80.9 / 59.3**	**85.6 / 66.3**	**87.6 / 69.7**	**88.1 / 70.9**

Table 3. Comparisons under the *small-scale* settings and *few-shot* settings. Results are shown as *Rank-1 / mAP*.

most challenging and largest dataset MSMT17, our method outperforms it by 11.6% Rank-1 and 7.7% mAP. (2) When adopting a more advanced network, *i.e.*, Swin-Transformer, as the backbone, the performance of our method is further improved and outperforms all considered supervised DG methods on all listed datasets without any human annotation, refreshing state-of-the-art. This demonstrates that data-hungry networks can make better use of large-scale training data, which is more suitable for our method. (3) These results show the paradigm of combining unsupervised contrastive learning and domain-diverse large-scale unlabeled training data is indeed one of the feasible directions for DG ReID, which can learn a DG ReID model even better than supervised methods with limited training data.

Compared to unsupervised methods: we also compare our method with five unsupervised methods, *i.e.*, TrackContrast [32], LUP [16], LUPnl [17], MoCo [25] and CycAs [71]. Our method outperforms TrackContrast by a large margin. For example, our method with ResNet50 backbone outperforms it by 28.9% mAP on Market-1501. Models trained by LUP and LUPnl can only serve as pre-training models and perform poorly when used for *direct testing*. The results of LUP[2] and LUPnl[3] are tested by us using their public models. CycAs uses the same data scale as ours. With the ResNet50 backbone, our method outperforms CycAs by 7.6% mAP in Market-1501. With the Swin-transformer backbone, our method suppresses CycAs by 10.1% mAP in Market-1501. The comparisons between our method and MoCo are fair because the MoCo is trained with our crawled video data. Our method outperforms MoCo significantly on all listed datasets. This demonstrates that the instance discrimination learned by MoCo and the identity discrimination required by ReID are misaligned. Our method focuses on learning identity-discriminative representations, which is consistent with the objective of ReID.

---

[2] https://github.com/DengpanFu/LUPerson
[3] https://github.com/DengpanFu/LUPerson-NL

Methods	CUHK03		Market-1501		MSMT17	
	R1	mAP	R1	mAP	R1	mAP
PCB [60]	63.7	57.5	93.8	81.6	–	–
MGN† [63]	71.2	70.5	95.1	87.5	85.1	63.7
BOT [48]	–	–	94.5	85.9	–	–
DSA [80]	78.9	75.2	95.7	87.6	–	–
ABDNet [7]	–	–	95.6	88.3	82.3	60.8
OSNet [85]	72.3	67.8	94.8	84.9	78.7	52.9
MHN [3]	77.2	72.4	95.1	85.0	–	–
BDB [13]	79.4	76.7	95.3	86.7	–	–
GCP [51]	77.9	75.6	95.2	88.9	–	–
ISP [86]	76.5	74.1	95.3	88.6	–	–
GASM [27]	–	–	95.3	84.7	79.5	52.5
ESNET [56]	–	–	95.7	88.6	80.5	57.3
*IN sup.+MGN†	71.2	70.5	95.1	87.5	85.1	63.7
*IN unsup.+MGN†	67.0	67.1	95.3	88.2	84.3	62.7
MoCo [25]+MGN†	73.2	71.8	96.2	90.3	85.2	65.5
LUP [16]+MGN†	75.4	74.7	96.4	91.0	85.5	65.7
CycAs [71]+MGN†	76.9	76.3	96.5	91.2	86.1	65.8
LUPnl [17]+MGN†	80.9	80.4	96.6	91.9	86.0	68.0
Ours+MGN†	**82.7**	**81.7**	**96.9**	**92.3**	**88.4**	**71.5**

Table 4. Comparisons under the pre-training → fine-tuning settings. Results of methods marked by "*" are from [17]. "IN sup."/"IN unsup." indicates model that is pretrained on ImageNet in a supervised/unsupervised manner. MGN† refers to the re-implementation of MGN in fast-reid.

4.3. DG Setting on Synthetic Dataset

We further conduct experiments on the synthetic datasets, including PersonX$_{456}$ [58], UnrealPerson [78], ClonedPerson [69] and RandPerson [70]. Since there is almost no method to conduct DG experiments on these synthetic datasets, we compare our method with two strong supervised baselines BOT [48] and MGN [63], and one unsupervised method MoCo [25]. Both BOT and MGN use the combination of Market-1501, Duke, CUHK03, and MSMT17 for supervised training. Our method and MoCo are trained with the crawled large-scale unlabeled video data and directly tested on these synthetic datasets without

fine-tuning. Results are reported in Table 2. Our method outperforms BOT, MGN and MoCo by a large margin, especially in terms of mAP. This shows that our method generalizes well, even if the target and source domains are essentially different. Our method learns the core concepts of pedestrian representation in an unsupervised manner.

4.4. Small-scale and Few-shot Settings.

Following LUP [16] and LUPnl [17], we also conduct experiments under two small data settings: the *small-scale* setting and the *few-shot* setting. The small-scale setting controls the percentage of usable IDs, and the few-shot setting controls the percentage of images for each ID. These two settings can reflect the adaptability of the model in new scenarios. We vary the usable percentages from 10% to 90% and use the MGN implemented in fast-reid as the baseline. Results are reported in Table 3. Our method comprehensively outperforms all other pre-trained models. This shows our model can adapt to new environments more efficiently and has a more substantial potential for real-world applications. *Note* that under the few-shot setting with 10% images per identity for fine-tuning, on Market-1501, the performance is lower than direct evaluation, *i.e.*, 75.9/54.3 *vs.* 85.1/65.1 (in Table 1). This is because when the usable images for each identity are few, it is insufficient to train the classifier with multiple classes well, which in turn compromises the feature extractor. Therefore, when the data for fine-tuning is noisy or poorly organized, we recommend using our model for direct testing without fine-tuning, which further demonstrates the superiority of our method.

4.5. Pre-training→Fine-tuning Setting.

Following [16] and [17], we evaluate the effectiveness of the model learned by our method as a pre-trained model. We conduct experiments on the strong baseline MGN [63] implemented by fast-reid with six different pre-training models, *i.e.*, "IN sup.", "IN unsup.", LUP [16], LUPnl [17], MoCo [25] and CycAs [71]. "IN sup."/"IN unsup." indicates the model is pre-trained on ImageNet in a supervised/unsupervised manner. MoCo is pre-trained on our collected data. The results are shown in Table 4. Note that post-processing such as IIA [18] and RR [83] are not applied. We can see that our method outperforms all the other methods by clear margins. This demonstrates that our model has promising transferability.

4.6. Ablation Study

The effectiveness of each component. Table 5 reports the ablation study of each component. There are three main components in our method, *i.e.*, CP: cross-frame positive pairs; $\mathcal{L}_{RC}$: reliability-guided contrastive loss in Eq. 3; $\mathcal{L}_Q$: memory-based contrastive loss in Eq. 5. We regard MoCo [25] as a comparable baseline. We can make sev-

Method	CP	$\mathcal{L}_{RC}$	$\mathcal{L}_Q$	Market-1501		MSMT17		CUHK03	
				R1	mAP	R1	mAP	R1	mAP
MoCo [25]	✗	✗	✗	10.5	2.6	0.5	0.2	0.3	0.7
	✓	✗	✗	54.5	31.9	17.8	3.7	9.5	9.8
Ours	✓	✓	✗	81.5	58.7	41.4	18.2	24.9	26.0
	✓	✗	✓	55.7	32.2	19.2	6.9	11.1	11.7
	✓	✓	✓	**85.1**	**65.1**	**45.7**	**21.2**	**26.1**	**27.4**

Table 5. Ablation study of the components under the DG settings. CP: Cross-frame positive Pairs; $\mathcal{L}_{RC}$: reliability-guided contrastive loss; $\mathcal{L}_Q$: memory-based contrastive loss.

Loss function	Market-1501		MSMT17		CUHK03	
	R1	mAP	R1	mAP	R1	mAP
Focal loss [43]	26.7	10.1	4.6	1.6	0.7	1.2
$\mathcal{L}_{RC}$	**85.1**	**65.1**	**45.7**	**21.2**	**26.1**	**27.4**

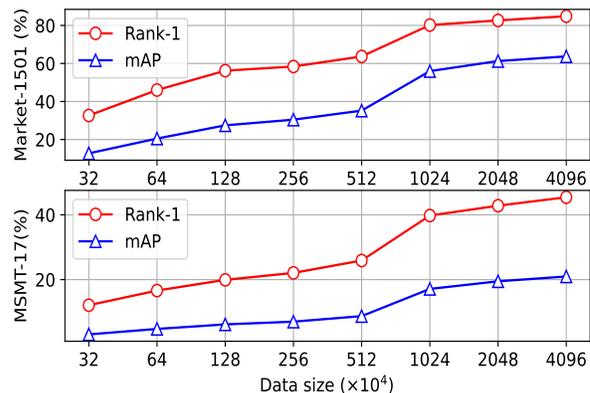Table 6. Comparisons with focal loss [43] under the DG settings.

Figure 4. Performance *vs*. data scale. The performance keeps improving as the training data increases and does not saturate.

eral observations. Firstly, even without using $\mathcal{L}_{RC}$ (set γ=0 in Eq. 3) and $\mathcal{L}_Q$, our framework outperforms MoCo by a large margin (*e.g.*, +44.0% Rank-1 in Market-1501). This confirms the effectiveness of our framework in learning identity discrimination by constructing positive pairs from inter-frame images instead of considering two augmented two augmented views of an image as a positive pair. Secondly, when $\mathcal{L}_{RC}$ is applied, the performance is significantly improved, *e.g.*, from 54.5% to 81.5% (+27.0%) Rank-1 in Market-1501. This demonstrates that our $\mathcal{L}_{RC}$ can effectively suppress the adverse impact of noisy positive pairs, resulting in more identity-discriminative representations. Thirdly, $\mathcal{L}_Q$ helps to improve the discrimination of learned representations. For example, when adding $\mathcal{L}_Q$ to $\mathcal{L}_{RC}$, the Rank-1 on Market-1501 is improved from 81.5% to 85.1%. In conclusion, constructing positive pairs from adjacent frames is the basis of our method. $\mathcal{L}_{RC}$ is much more important than $\mathcal{L}_Q$. Combining all proposed components achieves the best performance.

$\delta_{\max}$	0.5s	1.0s	2.0s	4.0s	8.0s
Market	73.7/48.4	81.4/58.4	83.8/63.2	**85.1/65.1**	84.5/64.4
MSMT17	34.6/13.8	41.2/17.9	43.5/19.0	**45.7/21.2**	44.8/20.1

Table 7. Impact of the maximum time interval $\delta_{\max}$. *Rank-1/mAP*

Comparison with focal loss [43]. We conduct comparative experiments with focal loss to verify the effectiveness of our $\mathcal{L}_{\text{RC}}$. Specifically, we only replace $\mathcal{L}_{\text{RC}}$ with focal loss, and keep other settings unchanged. As shown in Table 6, $\mathcal{L}_{\text{RC}}$ outperforms focal loss by a considerable margin. This supports the notion that it is error-prone to assign large training weight to hard samples in the presence of noisy labels since the label of the hard sample is likely to be incorrect. Instead, it is more appropriate to guide the model to learn from reliable samples.

Performance vs. Data size. Here, we study the effect of the training data size for our method. Under the DG ReID settings, we vary the training data size from 32×10^4 to 2048×10^4 to see how the performance changes. The backbone is ResNet50. We report the results in Figure 4. As we can see, with the increase of the data scale, the performance on Market-1501 and MSMT17 consistently keeps improving. Even at the large scale of 4096×10^4 unlabeled images, the performance does not saturate. There is still a clear upward trend in Rank-1 of MSMT17. This shows that our method has excellent scalability to large-scale unlabeled data, which has the potential to achieve better results as the training data continues to grow.

Time interval between two frame. In this part, we investigate the impact of the maximum time interval $\delta_{\max}$. As in Table 7, the choice of the maximum time interval involves trade-offs. When $\delta_{\max}$ is small, the performance increases as $\delta_{\max}$ increases because richer contrastive information is provided. When $\delta_{\max}$ is too large, the performance degrades because the mined positive pairs are error-prone, misleading the representation learning. Finally, we set $\delta_{\max} = 4.0s$ in our experiments.

Feature distribution. Here, we investigate the feature distribution to validate the capability of our method in learning identity discrimination. We collect samples of 15 identities that are randomly selected from the gallery of the Market-1501 [82], and visualize their features via t-SNE [62] in Figure 5. MoCo [25] is employed as a comparison. It can be seen that the features extracted by our model are closely clustered for all images of a given identity, while the features extracted by MoCo are dispersed. This observation verifies that MoCo can only learn instance-discriminative representations, which is determined by the way positive pairs are constructed. More importantly, this observation provides compelling evidence that our method effectively learns identity-discriminative representations.

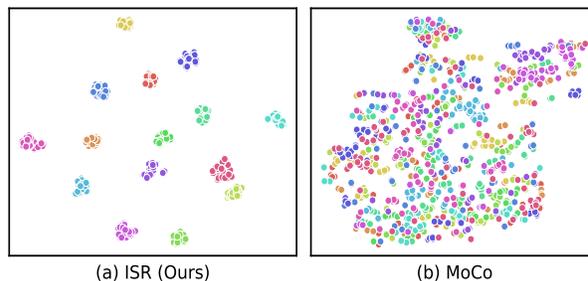
(a) ISR (Ours)   (b) MoCo

Figure 5. The t-SNE [62] visualization of the representation. Our method learns identity discrimination, while MoCo [25] does not.

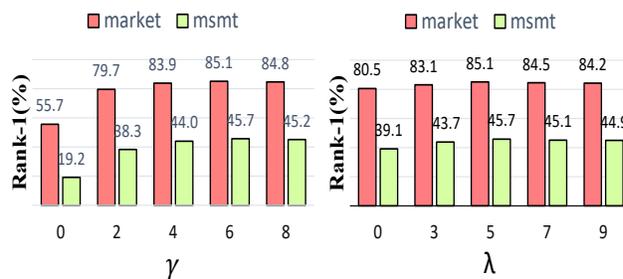Figure 6. Analysis of hyper-parameters γ in Eq. 3 and λ in Eq. 6.

Impact of hyper-parameters. Here, we show how γ in Eq. 3 and λ in Eq. 6 affect the performance. Experiments are conducted on Market-1501 and MSMT17 under the DG ReID settings. The backbone is ResNet-50. We report the results in Figure 6. We can see that $\gamma = 2, 4, 6, 8$ improves $\gamma = 0$ a lot and $\lambda = 3, 5, 7, 9$ outperforms $\lambda = 0$. This again shows the effectiveness of the proposed reliability-guided contrastive loss and memory-based contrastive loss. Our method is robust to λ and γ. Finally, the optimal values of γ and λ are 6 and 5, respectively.

5. Conclusions

This paper proposes an identity-seeking self-supervised representation learning method for learning a generalizable ReID representation from large-scale unlabeled videos. Unlike conventional contrastive learning methods that learn instance discrimination, our method learns identity discrimination by enforcing inter-frame instances with the same identity to have similar representations. The learned representation generalizes well to unseen domains without fine-tuning. In addition, the learned representation exhibits promising adaptability. We hope our method will provide new insights and attract more interest in large-scale unsupervised learning for domain-generalizable ReID.

References

[1] Yan Bai, Ce Wang, Yihang Lou, Jun Liu, and Ling-Yu Duan. Hierarchical connectivity-centered clustering for unsupervised domain adaptation on person re-identification. *IEEE TIP*, pages 6715–6729, 2021. 1, 2

[2] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, pages 371–381, 2019. 2

[3] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, pages 371–381, 2019. 7

[4] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *ICCV*, pages 9637–9646, 2019. 2

[5] Peixian Chen, Pingyang Dai, Jianzhuang Liu, Feng Zheng, Mingliang Xu, Qi Tian, and Rongrong Ji. Dual distribution alignment network for generalizable person re-identification. In *AAAI*, pages 1054–1062, 2021. 6

[6] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, pages 8351–8361, 2019. 1

[7] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, pages 8351–8361, 2019. 7

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 2, 3, 4

[9] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017. 2

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 2, 3, 4

[11] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *CVPR*, pages 3425–3435, 2021. 2, 6

[12] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*, pages 16145–16154, 2021. 1, 2, 5, 6

[13] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *ICCV*, pages 3691–3701, 2019. 7

[14] Zhaopeng Dou, Zhongdao Wang, Weihua Chen, Yali Li, and Shengjin Wang. Reliability-aware prediction via uncertainty learning for person image retrieval. In *ECCV*, pages 588–605, 2022. 1, 2

[15] Hao Feng, Minghao Chen, Jinming Hu, Dong Shen, Haifeng Liu, and Deng Cai. Complementary pseudo labels for unsupervised domain adaptation on person re-identification. *IEEE TIP*, pages 2898–2907, 2021. 1

[16] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *CVPR*, pages 14750–14759, 2021. 5, 6, 7, 8

[17] Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. Large-scale pre-training for person re-identification with noisy labels. In *CVPR*, pages 2476–2486, 2022. 5, 6, 7, 8

[18] Dengpan Fu, Bo Xin, Jingdong Wang, Dongdong Chen, Jianmin Bao, Gang Hua, and Houqiang Li. Improving person re-identification with iterative impression aggregation. *IEEE TIP*, pages 9559–9571, 2020. 8

[19] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*, pages 6112–6121, 2019. 1

[20] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275, 2008. 5

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, pages 21271–21284, 2020. 2, 3, 4

[22] Hongyang Gu, Jianmin Li, Guangyuan Fu, Chifong Wong, Xinghao Chen, and Jun Zhu. Autoloss-gms: Searching generalized margin-based softmax loss function for person re-identification. In *CVPR*, pages 4744–4753, 2022. 1

[23] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, pages 3642–3651, 2019. 2

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 3

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1, 2, 3, 4, 5, 6, 7, 8, 9

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 5

[27] Lingxiao He and Wu Liu. Guided saliency feature learning for person re-identification in crowded scenes. In *ECCV*, pages 357–373, 2020. 7

[28] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021. 1, 2

[29] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[30] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102, 2011. 5

[31] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Feature completion for occluded person re-identification. *IEEE TPAMI*, 2021. 1, 2

[32] Weiquan Huang, Yan Bai, Qiuyu Ren, Xinbo Zhao, Ming Feng, and Yin Wang. Large-scale unsupervised person re-identification with contrastive learning. *arXiv preprint arXiv:2105.07914*, 2021. 6, 7

[33] Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. Frustratingly easy person re-identification: Generalizing person re-id in practice. *arXiv preprint arXiv:1905.03422*, 2019. 1

[34] Mengxi Jia, Xinhua Cheng, Yunpeng Zhai, Shijian Lu, Siwei Ma, Yonghong Tian, and Jian Zhang. Matching on sets: Conquer occluded person re-identification without alignment. In *AAAI*, pages 1673–1681, 2021. 1, 2

[35] Bingliang Jiao, Lingqiao Liu, Liying Gao, Guosheng Lin, Lu Yang, Shizhou Zhang, Peng Wang, and Yanning Zhang. Dynamically transformed instance normalization network for generalizable person re-identification. In *ECCV*, pages 285–301, 2022. 2, 6

[36] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, pages 3143–3152, 2020. 2, 6

[37] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, pages 83–97, 1955. 3

[38] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 6

[39] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined depth space based architecture search for person re-identification. In *CVPR*, pages 6729–6738, 2021. 1

[40] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601, 2013. 5

[41] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 5, 6

[42] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, pages 2898–2907, 2021. 1

[43] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, pages 2980–2988, 2017. 5, 8, 9

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2, 5

[45] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[47] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *IJCV*, pages 106–129, 2010. 5

[48] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE TMM*, pages 2597–2609, 2019. 5, 6, 7

[49] Hao Ni, Jingkuan Song, Xiaopeng Luo, Feng Zheng, Wen Li, and Heng Tao Shen. Meta distribution alignment for generalizable person re-identification. In *CVPR*, pages 2487–2496, 2022. 2, 6

[50] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, pages 464–479, 2018. 5

[51] Hyunjong Park and Bumsub Ham. Relation network for person re-identification. In *AAAI*, pages 11839–11847, 2020. 7

[52] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, pages 7229–7238, 2018. 6

[53] Hamza Rami, Matthieu Ospici, and Stéphane Lathuilière. Online unsupervised domain adaptation for person re-identification. In *CVPR*, pages 3830–3839, 2022. 1, 2

[54] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016. 5, 6

[55] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. 6

[56] Dong Shen, Shuai Zhao, Jinming Hu, Hao Feng, Deng Cai, and Xiaofei He. Es-net: Erasing salient parts to learn more in re-identification. *IEEE TIP*, pages 1676–1686, 2020. 7

[57] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, pages 719–728, 2019. 1, 2, 5, 6

[58] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, pages 608–617, 2019. 2, 6, 7

[59] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020. 2

[60] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 1, 2, 7

[61] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, pages 6924–6932, 2017. 5

[62] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The journal of machine learning research*, pages 3221–3245, 2014. 9

[63] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018. 2, 6, 7, 8

[64] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703, 2014. 5

[65] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *AAAI*, pages 2540–2549, 2022. 1, 2

[66] Wenhao Wang, Shengcai Liao, Fang Zhao, Kangkang Cui, and Ling Shao. Domainmix: Learning generalizable person re-identification without human annotations. In *BMVC*, 2021. 3

[67] Wenhao Wang, Fang Zhao, Shengcai Liao, and Ling Shao. Attentive waveblock: complementarity-enhanced mutual networks for unsupervised domain adaptation in person re-identification and beyond. *IEEE TIP*, pages 1532–1544, 2022. 2

[68] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *CVPR*, pages 6388–6397, 2020. 5

[69] Yanan Wang, Xuezhi Liang, and Shengcai Liao. Cloning outfits from real-world images to 3d characters for generalizable person re-identification. In *CVPR*, pages 4900–4909, 2022. 3, 6, 7

[70] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *ACM MM*, pages 3422–3430, 2020. 2, 6, 7

[71] Zhongdao Wang, Zhaopeng Dou, Jingwei Zhang, Liang Zhen, Yifan Sun, Yali Li, and Shengjin Wang. Generalizable re-identification from videos with cycle association. *arXiv preprint arXiv:2211.03663*, 2022. 3, 5, 6, 7, 8

[72] Zhongdao Wang, Jingwei Zhang, Liang Zheng, Yixuan Liu, Yifan Sun, Yali Li, and Shengjin Wang. Cycas: Self-supervised cycle association for learning re-identifiable descriptions. In *ECCV*, 2020. 3

[73] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *ECCV*, pages 107–122, 2020. 3, 5

[74] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 2, 5, 6

[75] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016. 5

[76] Boqiang Xu, Jian Liang, Lingxiao He, and Zhenan Sun. Mimic embedding via adaptive aggregation: learning generalizable person re-identification. In *ECCV*, pages 372–388, 2022. 6

[77] Pengyi Zhang, Huanzhang Dou, Yunlong Yu, and Xi Li. Adaptive cross-domain learning for generalizable person re-identification. In *ECCV*, pages 215–232, 2022. 2, 6

[78] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Unrealperson: An adaptive pipeline towards costless person re-identification. In *CVPR*, pages 11506–11515, 2021. 2, 6, 7

[79] Yi-Fan Zhang, Zhang Zhang, Da Li, Zhen Jia, Liang Wang, and Tieniu Tan. Learning domain invariant representations for generalizable person re-identification. *IEEE TIP*, 2022. 2

[80] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, pages 667–676, 2019. 7

[81] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, pages 6277–6286, 2021. 2, 6

[82] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 5, 6, 9

[83] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017. 8

[84] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, pages 598–607, 2019. 1, 2

[85] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019. 7

[86] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*, pages 346–363, 2020. 1, 2, 7