Supplementary Materials: 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*

1. Hungarian Algorithm

In our method, 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:

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

This is a linear assignment problem, which can be solved by the Hungarian algorithm. Please refer to [8, 1, 13] for the details of Hungarian algorithm. In fact, the Hungarian algorithm have been implemented in the package "scipy". The pseudo code of solving the optimal matrix $\boldsymbol{\pi}^*$ is:

Listing 1. The python code of solving optimal matrix $\boldsymbol{\pi}^*$

```
from scipy.optimize import linear_sum_assignment
import numpy as np

def solver(cost_matrix):
    """
    cost_matrix: the dimensions are [m, n], m <= n
    return: the optimal strategy
    """
    m, n = cost_matrix.shape
    optimal_matrix = np.zeros((m, n))
    temp, _ = linear_sum_assignment(cost_matrix)
    for i in range(m):
        optimal_matrix[i] = temp[i]
    return optimal_matrix
```

* Corresponding author

Method	Data size	Market-1501		MSMT17		CUHK03	
		R1	mAP	R1	mAP	R1	mAP
TrackContrast [6]	∼15.4M	72.7	36.2	–	–	–	–
Ours	∼15.4M	**81.4**	**58.6**	**41.3**	**18.3**	**21.9**	**23.2**
LUP† [2]	∼4.2M	3.3	1.0	0.3	0.1	0.1	0.5
Ours	∼4.2M	**61.8**	**33.5**	**24.5**	**8.1**	**9.3**	**10.1**
LUPnl† [3]	∼40.6M	13.8	3.8	0.6	0.2	0.4	0.8
Ours	∼10.6M	**80.3**	**56.1**	**39.9**	**17.2**	**20.3**	**21.7**
CycAs [17]	∼47.8M	80.3	57.5	43.9	20.2	25.8	26.5
Ours	∼47.8M	**85.1**	**65.1**	**45.7**	**21.2**	**26.1**	**27.4**

Table 1. Comparisons with several unsupervised algorithms based on the same-sized training data.

2. Comparison v.s. Data Size

To further show the effectiveness of our method, we compare our method with several unsupervised algorithms based on the same-sized training data, including TrackContrast [6], LUP [2], LUPnl [3] and CycAs [17]. The results are shown in Table 1. It can be seen that our method significantly outperforms these methods. For example, our method outperforms TrackContrast [6] by 8.7% Rank-1 on Market-1501. These results demonstrate the effectiveness of our method when serving as an unsupervised method for the ReID task.

3. Cloth-Changing Setting

To further verify the effectiveness of our method, we conduct experiments on the cloth-changing ReID dataset LTCC [14]. The results are reported in Table 2.

LTCC [14] contains 17,119 images from 152 identities. Each identity contains 112 images on average. It includes different viewpoints, backgrounds, and occlusions. Each identity may wear different clothes, and different people may wear the same clothes.

Compared to supervised methods under the intra-domain evaluation. As far as we know, there are no methods to conduct experiments on cloth-changing datasets under

Methods	Sup.	Training set	Cross-domain Evaluation	LTCC	
				R1	mAP
PCB [16]	✔	LTCC	✘	65.1	30.6
HACNN [10]	✔	LTCC	✘	60.2	26.7
GI-ReID* [7]	✔	LTCC	✘	63.2	29.4
BOT [12]	✔	M+D+C3+MT	✔	60.6	26.9
Ours (R50)	✘	Unsup-videos	✔	**65.7**	**31.6**
Ours (Swin)	✘	Unsup-videos	✔	**68.8**	**34.7**

Table 2. Results on the cloth-changing dataset LTCC [14]. "Sup." indicates the method is supervised or unsupervised. "*" indicates that what we report is the result when GI-ReID uses ResNet50 as the backbone. "Unsup-videos" indicate our large-scale unlabeled video dataset. M: Market-1501 [20], D: DukeMTMC [15], C3: CUHK-3 [9], MT: MSMT17 [19].

the domain generalizable settings. Therefore, we mainly compare our method with three supervised methods, *i.e.*, PCB [16], HACNN [10] and GI-ReID [7], which are tested under the intra-domain evaluation settings. For example, when testing PCB on the LTCC dataset, the model is trained using the training set of LTCC. We can see that our method shows superior performance than PCB, HACNN, and GI-ReID. For example, on LTCC, our method outperforms PCB, HACNN, GI-ReID in terms of the Rank-1 score. Especially, with Swin-Transformer as the backbone, our method outperforms PCB, HACNN and GI-ReID by 3.7%, 8.6%, and 5.6% Rank-1 on LTCC, respectively. Note that our method is trained by large-scale videos in an unsupervised manner. In our training data, the clothing of a pedestrian hardly changes. It is satisfactory that our method can achieve such results when directly tested on cloth-changing datasets. This demonstrates that our method not only learns the clothes information of a pedestrian but also other discriminative information for the pedestrian representation.

Compared to supervised methods under the cross-domain evaluation. We also compared our method with a strong supervised baseline BOT [12] under the cross-domain evaluation. The BOT is trained using the combination of Market-1501[20], DukeMTMC [15], CUHK03 [9], and MSMT17 [19] and directly tested on the cloth-changing dataset LTCC. We can see that, with ResNet50 backbone, our method outperforms BOT by 5.1% Rank-1 on LTCC, respectively. The gap is further widened when adopting the more data-hungry Swin-Transformer [11]. This shows that our method learns a more domain-generalizable person representation. More importantly, the results verify that applying unsupervised contrastive learning on the domain-diverse large-scale unlabeled data is indeed one of the valuable directions for domain-generalizable ReID, which learns a domain-generalizable person representation even better than some supervised methods trained with limited training data.

4. Comparisons with MoCo [4]

Here, we conduct experiments to compare our method with the conventional unsupervised contrastive learning method MoCo [4]. ***Note that for fair comparisons, both our method and MoCo use the same training data,* i.e.*, our collected videos.*** The experiments are conducted under domain-generalizable settings. The backbone is ResNet50 [5].

To illustrate the superiority of our method in more detail, we conduct experiments with two versions of MoCo:

- **The original MoCo:** adopt the framework in [4] that regards two augmented views of an image as a positive pair.
- **The improved MoCo:** adopt our proposed framework that considers inter-frame images belonging to the same identity as positive pairs.

We report related results in Table 3, from which we can make several observations.

- The improved MoCo outperforms the original MoCo by a large margin (*e.g.*, +44.3% Rank-1 in Market-1501 [20]). This demonstrates the effectiveness of our proposed framework for the ReID task. The original MoCo enforces the model to learn a unique representation for each image, learning to *instance discrimination*, which conflicts with the objective of ReID. Our framework requires inter-frame images with the same identity to have similar representations, which aligns with the ReID required *identity discrimination*.
- Our method outperforms the improved MoCo by a significant margin (*e.g.*, +30.3% Rank-1 in Market-1501). The biggest difference between them is that our method utilizes the proposed reliability-guided contrastive loss $\mathcal{L}_{\mathrm{RC}}$. The result demonstrates that our $\mathcal{L}_{\mathrm{RC}}$ can effectively suppress the adverse impact of mined noisy positive pairs and learn more identity-discriminative representations.

5. Ablation Study

5.1. Data Augmentation

As described in the main text, suitable data augmentations benefit unsupervised contrastive learning. Here, we study the impact of data augmentations on our method. We select three commonly used data augmentations in supervised ReID methods to explore, *i.e.*, random horizontal flipping, random color jittering, and random erasing [21]. We report the results in Table 4. From the results, we can make several observations. (1) Comparing *row-1* and *row-2*, we can see that random horizontal flipping slightly improves

Methods	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
MoCo* [4] (Original)	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
MoCo† [4] (Improved)	54.8	32.5	18.5	6.7	10.4	9.9	46.5	59.4	33.0	44.5	39.7	50.4	75.3	81.6
ISR (Ours)	**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**

∗ Original MoCo: regarding two augmented views of an image as a positive pair

† Improved MoCo: adopting our proposed framework, *i.e.*, constructing positive pairs from inter-frame images

Table 3. Comparison with MoCo under DG settings. The original MoCo adopts the framework in [4] that regards two augmented views of an image as a positive pair. The improved MoCo adopts our proposed framework that considers inter-frame images belonging to the same identity as positive pairs. We can make several observations: *(i)* the comparisons between the improved MoCo and the original MoCo demonstrate the effectiveness of our proposed framework; *(ii)* the comparisons between our method and the improved MoCo demonstrate the effectiveness of our proposed reliability-guided contrastive loss $\mathcal{L}_{\mathrm{RC}}$.

#	HF	CJ	RE	Market-1501		MSMT17		CUHK03	
				R1	mAP	R1	mAP	R1	mAP
1	✗	✗	✗	78.2	54.6	38.5	17.5	22.7	23.6
2	✓	✗	✗	79.1	56.1	39.4	18.3	23.8	24.5
3	✗	✓	✗	83.5	62.7	43.9	20.2	25.0	26.9
4	✓	✓	✗	**85.1**	**65.1**	**45.7**	**21.2**	**26.1**	**27.4**
5	✓	✓	✓	81.7	60.9	40.5	18.6	21.0	22.7

Table 4. Ablation study of data augmentation under the domain generalizable settings. The backbone is ResNet50. HF: random horizontal flipping; CJ: random color jittering; RE: random erasing [21].

the performance. (2) Comparing *row-1* and *row-3*, it can be seen that random color jittering significantly improves the performance. For example, on Market-1501, the Rank-1 and mAP are improved by 5.3% and 8.1%, respectively. This is because that random color jittering can effectively enrich the diversity of domain distributions in training data and strengthen the contrastive information between positive pairs, improving the generalization of the learned representation. (3) Combining random horizontal flipping and color jittering, the performance further improved, which is the data augmentation approach used in our paper. (4) Comparing *row-4* and *row-5*, we find that random erasing degrades the performance of our method. This may be because, after random erasing, similarities between images belonging to the same identity are greatly reduced, making the mined positive pairs error-prone. At the same time, the differences between the reliability scores decrease, which degrades the proposed reliability-guided contrastive loss $\mathcal{L}_{\mathrm{RC}}$.

5.2. Hyper-parameter τ

As described in Eq.1 in the main text, the reliability score of a positive pair is formulated as:

$$p(\boldsymbol{x}_i) = \frac{\sum_{j=1}^{n} \boldsymbol{\pi}_{ij}^{*} \exp(\boldsymbol{x}_i \cdot \boldsymbol{y}_j / \tau)}{\sum_{j=1}^{n} \exp(\boldsymbol{x}_i \cdot \boldsymbol{y}_j / \tau)}$$

where τ is the temperature hyper-parameter. τ is a key factor to the reliability score. If τ is too large, the reliability score of a positive pair will be small even if its corresponding samples are very similar. If τ is too small, the reliability score of the positive pair will be sharpened to be close to 0 or 1, resulting in that $\mathcal{L}_{\mathrm{RC}}$ approximates to 0. Very small τ loses a lot of information useful for training and compromises the effectiveness of contrastive learning. The selection of τ should be related to the number of instances in $\boldsymbol{Y}$. Following [18], we adopt a batch-related τ:

$$\tau_{\mathrm{br}} = \frac{\epsilon}{\log(n+1)} \tag{2}$$

where n is the number of instances in $\boldsymbol{Y}$ and $\epsilon = 0.4$. We compare τ_{br} and fixed τ and report the results in Table 5. We can make several observations. (1) When τ gradually decreases from 0.091 to 0.040, the performance decreases accordingly. This shows too small τ degrades the contrastive learning. (2) When τ increases from 0.091 to 0.200, the model is crashed. This shows too large τ is fatal for our method. (3) Finally, batch-related τ_{br} achieves the best performance. Note that such a τ_{br} is also adaptive to different batch sizes. When using the Swin-Transformer as the backbone, it also works.

5.3. Hyper-parameter k

As described in Sec.3.3 in the main text, for $\boldsymbol{x}_i$, we select k hard negative samples from the representation queue to perform the contrastive loss $\mathcal{L}_{\mathrm{Q}}$. Here, we vary the size

	Market-1501		MSMT17		CUHK03	
	R1	mAP	R1	mAP	R1	mAP
$\tau = 0.200$	–	–	–	–	–	–
$\tau = 0.091$	84.6	64.4	44.8	21.0	25.9	27.2
$\tau = 0.067$	81.9	60.8	43.0	19.9	22.9	23.6
$\tau = 0.050$	72.1	48.9	33.5	13.8	15.1	15.8
$\tau = 0.040$	62.3	39.4	26.1	10.1	9.8	11.0
τ_{br}	**85.1**	**65.1**	**45.7**	**21.2**	**26.1**	**27.4**

Table 5. Ablation study of temperature hyper-parameter τ. The backbone is ResNet50. When $\tau = 0.200$, the model crashes with NAN parameters.

	Market-1501		MSMT17		CUHK03	
	R1	mAP	R1	mAP	R1	mAP
$k = 3$	**85.8**	**65.3**	**46.2**	**21.7**	25.4	26.8
$k = 5$	85.1	65.1	45.7	21.2	26.1	27.4
$k = 10$	85.1	63.8	45.5	21.3	**27.3**	**28.1**
$k = 15$	84.3	63.8	45.2	21.1	26.8	27.9
$k = 20$	84.3	63.2	44.5	20.4	26.6	27.4

Table 6. Ablation study of temperature hyper-parameter k. The backbone is ResNet50.

of k to investigate how the performance changes. We conduct experiments under the domain generalizable settings and adopt ResNet50 as the backbone. We report the results in Table 6. From the results, we can see that our method is robust to the hyper-parameter k. When k varies from 3 to 20, the changes in the results are limited, especially on CUHK03 [9]. One observation is that when k increases, the performance on Market-1501 [20] and MSMT17 [19] drops. This may be because when k increases, the role of extremely hard negative samples is gradually weakened. Finally, to balance the performance of these datasets, we set $k = 5$ in our experiments.

6. Performance *vs*. Model complexity

Here, we investigate how the complexity of the model affects performance. We select four models with increasing complexity as the backbones, *i.e*., ResNet18 [5], ResNet34 [5], ResNet50 [5] and Swin-Transformer [11]. Experiments are conducted under domain-generalizable settings. We report the results in Table. 7. We can make several observations from the results: (1) ResNet18, ResNet34, and ResNet50 are from the same family. Comparisons between them can better illustrate the impact of the model complexity on performance. We can see that when the model complexity increases, the performance accordingly increases. (2) Swin-Transformer is a more complex and advanced architecture. The superiority of the Swin-transformer is significantly obvious. For example, our method with Swin-transformer backbone outperforms our method with ResNet50 backbone by 1.9%, 10.7% and 10.5% Rank-1 on Market-1501, MSMT17 and CUHK03, respectively. This shows that data-hungry architectures are more suitable for our method, which can make better use of large-scale training data.

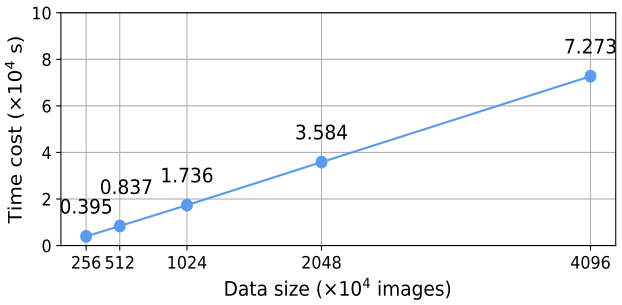

Figure 1. Time cost *vs*. data scale on ResNet50 backbone. We use 4×NVIDIA 3090 GPU with 2× AMD EPYC 7742 CPU. The training cost of method scales approximately linearly with the data size, making it feasible to utilize large-scale data for training.

7. Time cost *vs*. Data size

In this part, we conduct experiments to show that the training cost of our method is roughly linear with the data size. We adopt ResNet50 as the backbone and adopt distributed training on 4×NVIDIA 3090 GPU with 2× AMD EPYC 7742 CPU. We vary the training data size from 256×10^4 to 4096×10^4 to see how the time cost changes. The results are reported in Figure 1. From the results, we can see that the training cost of our method scales approximately linearly with the size of the data. This makes it possible to drive large-scale data for training.

Backbone	Params	Flops	Market-1501		MSMT17		CUHK03	
			R1	mAP	R1	mAP	R1	mAP
ResNet18	1.99×10^9	10.9M	80.8	56.5	39.7	16.7	20.9	22.3
ResNet34	3.65×10^9	20.5M	83.6	62.9	43.6	19.7	25.8	26.9
ResNet50	4.06×10^9	23.4M	85.1	65.1	45.7	21.2	26.1	27.4
Swin-Transformer	1.50×10^{10}	83.2M	**87.0**	**70.5**	**56.4**	**30.3**	**36.6**	**37.8**

Table 7. Performance *vs*. Model complexity. The more complex and advanced the model, the higher the performance.

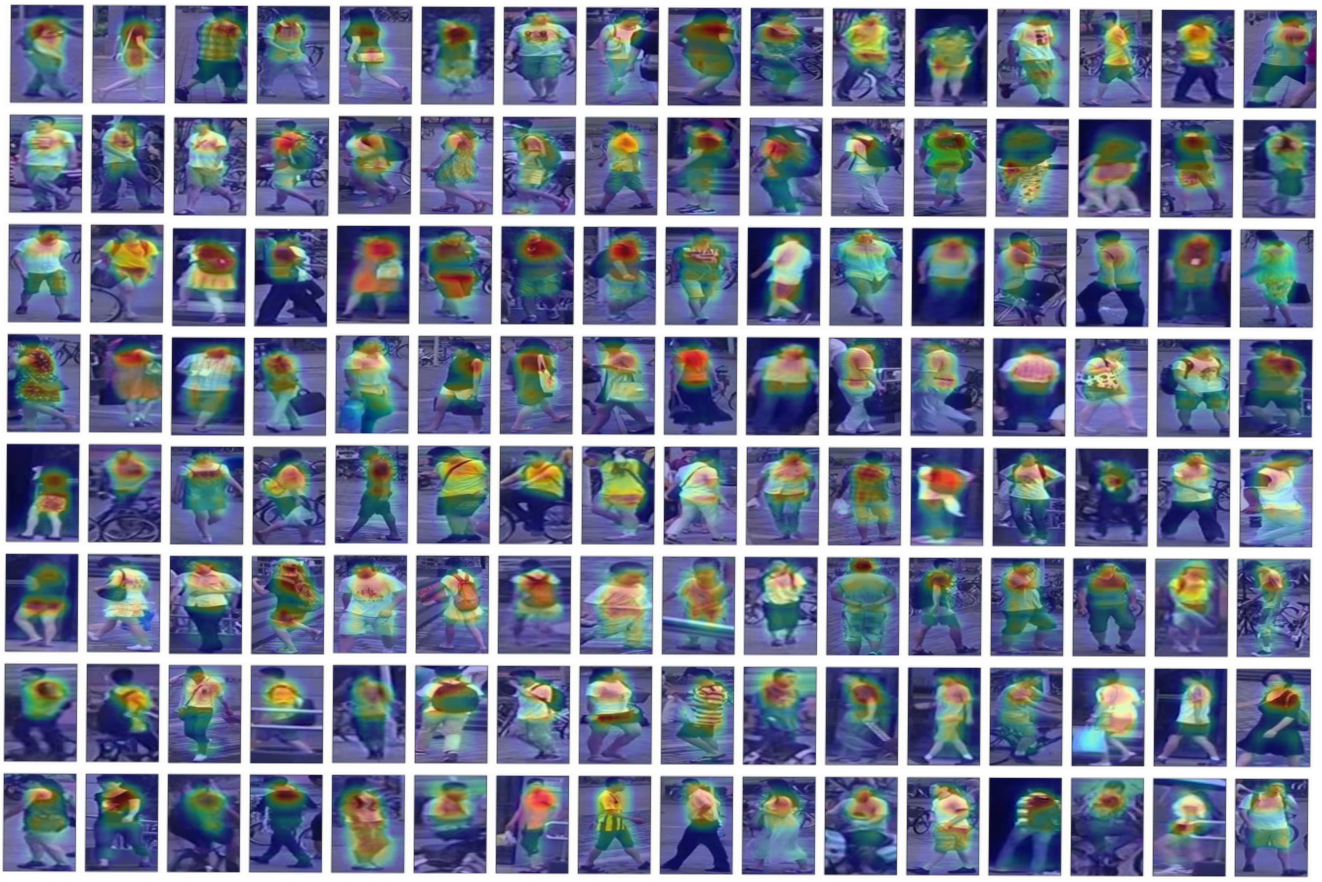

Figure 2. Activation maps when adopting ResNet50 as the backbone. Images are randomly selected from the query set of Market-1501. It can be seen that our method focuses on the foreground region of pedestrians, effectively ignoring the negative effect of the background.

8. Activation Maps

A generalizable and robust person ReID model should focus on the pedestrian regions in images and effectively ignore the background. To further verify that the model learned by our method has this property, we investigate the activation maps of the images. Specifically, give an image $\boldsymbol{I}$, we input it into the model (ResNet50) to obtain the feature maps $\boldsymbol{F} \in \mathbb{R}^{d \times h \times w}$, where d, h and w are the channels, height and width of the feature maps. We regard the norm of the vector at each spatial position as the corresponding activation value. For example, for a spatial position (x, y) in the feature maps, its activation value is $\boldsymbol{A}(x, y) = \|\boldsymbol{F}(x, y)\|_2$, where $\boldsymbol{A} \in \mathbb{R}^{h \times w}$ is the activation map. Then the elements in the activation map are mapped to $[0, 1]$:

$$\hat{\boldsymbol{A}} = \frac{\boldsymbol{A} - a_{\min}}{a_{\max} - a_{\min}} \tag{3}$$

where $a_{\max}$ and $a_{\min}$ are the maximum and minimum values in $\boldsymbol{A}$, respectively. We randomly select many images from the query set of Market-1501 and visualize their $\hat{\boldsymbol{A}}$ in Figure 2. For almost all selected images, we can see that the region where the model is activated is the pedestrian region. This shows that our learned model focuses on the foreground region of pedestrians, effectively ignoring the negative effect of the background. Note that our model is

Figure 3. Retrieval examples of our method on Market-1501 with **ResNet50** as the backbone. Red/green boxes indicate the true/false matches in the gallery. **Note that the false matches are hard samples.**

Figure 4. Retrieval examples of our method on Market-1501 with **Swin-Transformer** as the backbone. Red/green boxes indicate the true/false matches in the gallery. **Note that the false matches are hard samples.**

trained without any identity and foreground labels. We consider such results to be very gratifying and satisfying.

9. Retrieval Examples

Here, we show some retrieval examples on Market-1501 [20] under the domain generalizable settings. We adopt ResNet50 and Swin-Transformer as the backbone and report the results in Figure 3 and Figure 4, respectively. We can see that our method can effectively retrieve positive images in the gallery. This shows that our method learns the fundamental concept of pedestrian representation. Moreover, the retrieved false matches are very similar to the query, and some are even difficult to distinguish manually. Therefore, we consider the retrieval results to be satisfactory.

References

[1] Ivanda Zevi Amalia, Ahmad Saikhu, and Rully Soelaiman. A fast dynamic assignment algorithm for solving resource allocation problems. *Jurnal Online Informatika*, 2021. 1

[2] 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. 1

[3] 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. 1

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

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 4

[6] 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. 1

[7] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *CVPR*, pages 14278–14287, 2022. 2

[8] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, pages 83–97, 1955. 1

[9] 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. 2, 4

[10] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 2

[11] 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, 4

[12] 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. 2

[13] G Ayorkor Mills-Tettey, Anthony Stentz, and M Bernardine Dias. The dynamic hungarian algorithm for the assignment problem with changing costs. *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27*, 2007. 1

[14] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *ACCV*, 2020. 1, 2

[15] 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. 2

[16] 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. 2

[17] 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. 1

[18] 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

[19] 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, 4

[20] 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. 2, 4, 6

[21] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020. 2, 3