# Supplementary: Unleashing Potential of Unsupervised Pre-Training with Intra-Identity Regularization for Person Re-Identification

Zizheng Yang Xin Jin Kecheng Zheng Feng Zhao*
University of Science and Technology of China
{yzz6000,jinxustc,zkcys001}@mail.ustc.edu.cn fzhao956@ustc.edu.cn

# Appendix

## A.1. Datasets

### A.1.1. Pre-training Dataset

**LUPerson** [5] consists of 4,180,243 person images of over 200K identities extracted from 46,260 YouTube videos. YOLO-v5 trained on MS-COCO is utilized to extract each person instance in the sampled frame. It is worth noting that the LUPerson is large enough to support unsupervised person ReID feature learning.

### A.1.2. Fine-tuning Datasets

**CUHK03** [10] contains 13,164 images of 1,360 pedestrians. Each identity is observed by 2 cameras. Note that CUHK03 offers both hand-labeled and DPM-detected bounding boxes, and the former is adopted in this paper.
**Market1501** [22] contains 32,668 person images of 1,501 identities captured by 6 cameras. The training set consists of 12,936 images of 751 identities, the query set consists of 3,368 images, and the gallery set consists of 19,732 images of 750 identities.
**PersonX** [14] is a large-scale data synthesis engine, which contains 1,266 manually designed identities and editable visual variables. Each identity is captured by 6 cameras.
**MSMT17** [18] contains of 126,441 images of 4,101 identities captured by 15 cameras. The training set consists of 30,248 person images of 1,041 identities, the query set consists of 11,659 images, and the gallery consists of 82,161 images of 3,060 identities.

## A.2. More Details about Data Augmentation

Data augmentation plays a crucial role in self-supervised contrastive learning. We adopt popular augmentation operations including resizing, cropping, random grayscale, Gaussian blurring, horizontal flipping, and RandomErasing. Note that we abandon color jitter since person ReID is extremely dependent on color information [5].

*Corresponding Author.

## A.3. Additional Results

### A.3.1. More Results for Supervised ReID

In Section 4.2 of the main body, we demonstrate that our UP-ReID can benefit the supervised ReID methods and show the results in Table 1. Here, we present the remaining results. Table A1 shows the results of using different pre-trained models in the supervised fine-tuning ReID method PCB [15] on CUHK03, Market1501, and PersonX.

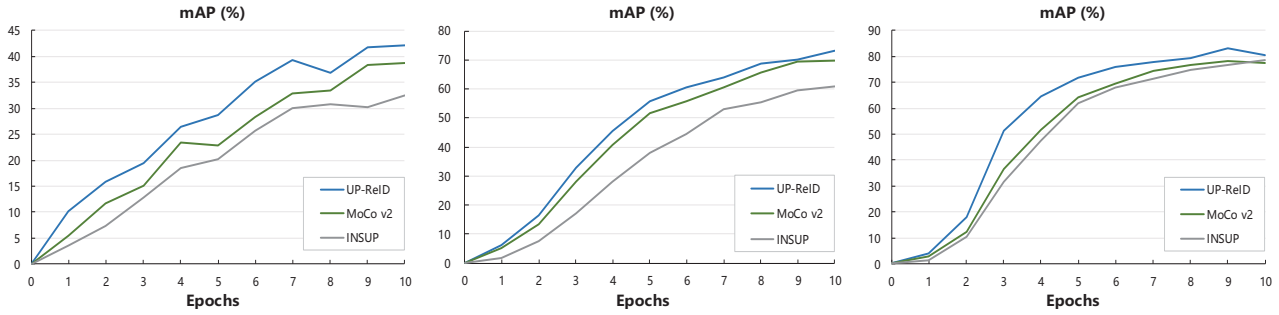
Table A1. Comparison of PCB method using different pre-trained models on three datasets in terms of mAP/Rank1 (%).

| Model | CUHK03 | Market1501 | PersonX |
|---|---|---|---|
| INSUP | 59.5/69.9 | 78.0/92.6 | 80.9/92.7 |
| MoCo v2 | 58.3/72.8 | 79.3/92.9 | 80.7/92.9 |
| UP-ReID | **60.1/74.1** | **80.0/93.1** | **81.7/93.2** |

We also show the comparison of the convergence speed of applying different pre-trained models in method MGN [17] at the early stage of fine-tuning in Figure A1. As can be seen, UP-ReID achieves a faster convergence rapidity compared with MoCo v2 and INSUP on all the three datasets, which further demonstrates that the proposed UP-ReID can better benefit downstream ReID tasks.

### A.3.2. More Comparisons with State-of-the-Arts

In Section 4.4 of the main body, we have shown some comparison results between our UP-ReID and state-of-the-art methods. Here, we extend the results in Table 3 and show the complete results of the comparison between UP-ReID and state-of-the-art methods in Table A2 on three datasets, including CUHK03, Market1501, and MSMT17. As we can see, MGN with our UP-ReID outperforms the other methods by at least **7.9%/6.5%** and **1.6%/1.0%** in terms of mAP/Rank1 on CUHK03 and Market1501, respectively. On the MSMT17 dataset, the TransReID [8] achieves better performance. However, TransReID adopts transformer-based network and utilizes camera information additionally.




(a) mAP learning curve on CUHK03 (b) mAP learning curve on Market1501 (c) mAP learning curve on PersonX

Figure A1. mAP learning curves of different pre-trained models in MGN [17] on three datasets (CUHK03, Market1501, and PersonX) with the same training schedule.

Table A2. Complete performance (%) comparisons with state-of-the-art approaches on CUHK03, Market1501, and MSMT17. The best results are marked as bold and the second ones are masked by underline.

| Method | CUHK03 | | Market1501 | | MSMT17 | |
|---|---|---|---|---|---|---|
| | mAP | cmc1 | mAP | cmc1 | mAP | cmc1 |
| PCB [15] (ECCV'18) | 57.5 | 63.7 | 81.6 | 93.8 | - | - |
| MGN [17] (ACM MM'18) | 70.5 | 71.2 | 86.9 | 95.7 | - | - |
| ABDNet [2] (ICCV'19) | - | - | 88.3 | 95.6 | 60.8 | 82.3 |
| BDB [4] (ICCV'19) | 76.7 | 79.4 | 86.7 | 95.3 | - | - |
| OSNet [24] (ICCV'19) | 67.8 | 72.3 | 84.9 | 94.8 | 52.9 | 78.7 |
| P2Net [6] (ICCV'19) | 73.6 | 78.3 | 85.6 | 95.2 | - | - |
| SCAL [1] (ICCV'19) | 72.3 | 74.8 | 89.3 | 95.8 | - | - |
| DSA [19] (CVPR'19) | 75.2 | 78.9 | 87.6 | 95.7 | - | - |
| DGNet [23] (CVPR'19) | - | - | 86.0 | 94.8 | 52.3 | 77.2 |
| GCP [13] (AAAI'20) | 75.6 | 77.9 | 88.9 | 95.2 | - | - |
| SAN [9] (AAAI'20) | 76.4 | 80.1 | 88.0 | 96.1 | 55.7 | 79.2 |
| ISP [25] (ECCV'20) | 74.1 | 76.5 | 88.6 | 95.3 | - | - |
| GASM [7] (ECCV'20) | - | - | 84.7 | 95.3 | 52.5 | 79.5 |
| RGA-SC [20] (CVPR'20) | 77.4 | 81.1 | 88.4 | 96.1 | - | - |
| HOReID [16] (CVPR'20) | - | - | 84.9 | 94.2 | - | - |
| AMD [3] (ICCV'21) | - | - | 87.1 | 94.8 | - | - |
| PGFL-KD [21] (ICCV'21) | - | - | 87.2 | 95.3 | - | - |
| TransReID [8] (ICCV'21) | - | - | 89.5 | 95.2 | **67.4** | **85.3** |
| PAT [11] (CVPR'21) | - | - | 88.0 | 95.4 | - | - |
| MGN+R50 (UP-ReID) | **85.3** | **87.6** | **91.1** | **97.1** | 63.3 | 84.3 |

## A.4. Discussion about Hard Mining Strategy

In Section 3.4 of the main body, we introduce our hard mining strategy in detail and experimentally prove its effectiveness in Section 4.5. Here we further discuss two points and give more insights about this design. The first one is that we choose hard positive samples and hard negative queues in a fixed way, which is an offline scheme instead of an online scheme. Would an online scheme be better? The second one comes from the positive samples selection. In Section 3.3 of the main body, we emphasize that all $2M$ patch-level instances are partitioned from the input image $x$ actually. So, for each patch feature $q_i \in \mathcal{X}_q$, any of patch feature $k_p^+ \in \mathcal{X}_k$ $(i, p \in \{1, ..., M\})$ could be its positive sample. So, why do we have to choose patches at the same horizontal position instead of other patches as the positive samples?

To answer the aforementioned questions and verify the reasonableness of our selection strategy, we compare it with several other schemes. ***Random Positive Selection***: for patch $i$, we randomly select a patch partitioned from the same pedestrian but located differently as the positive sample. ***Online Positive Selection***: instead of finding a hard positive patch sample for each query patch $i$, we only select the hardest positive pair among all the $M \times M$ posi-

tive pairs. ***Horizontally Symmetric Positive Selection***: the proposed selection strategy wherein two horizontally symmetric patches are selected as a positive pair. Note that all three schemes have the same rule to select negative samples. We show the curves of the patch-wise contrastive loss in the intrinsic contrastive constraint under these three selection strategies in Figure A2. As we can see, the loss value in the scheme of "Random-P" is unstable and cannot reach a convergence. On the other hand, the loss value in the scheme of "Online-P" converges extremely slowly.

We analyze that the scheme of "Random Positive Selection" and "Online Positive Selection" suffer from misalignment and can not guarantee that the selected positive pairs have similar visual information. Take the "Random Positive Selection" as an example, for $q_i \in \mathcal{X}_q$, we randomly select $k_p^+ \in \mathcal{X}_k$ as the corresponding positive sample. However, without any constraint, the visual information contained in $q_i$ and $k_p^+$ may be very different (*e.g.*, $q_i$ represents the head of a person, while $k_p^+$ represents the shoes), which has a negative impact on the pre-training process.

Our hard mining strategy (*i.e.*, Horizontally Positive Selection) is based on the prior knowledge that persons are horizontally symmetric, which assures that the positive pairs are semantically matched. This avoids the negative impact caused by misalignment on the pre-training process.

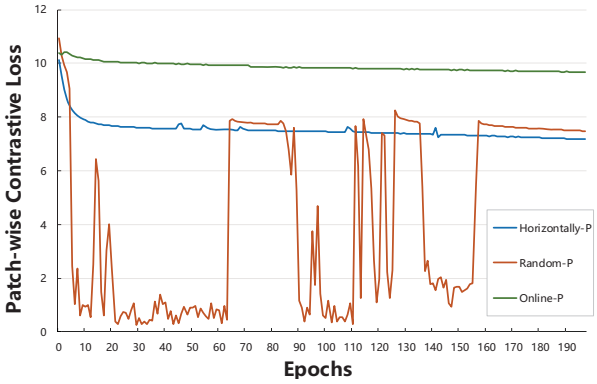



Figure A2. The curves of the patch-wise contrastive loss in different selection strategies. "Horizontally-P", "Random-P", and "Online-P" mean Horizontally Symmetric Positive Selection, Random Positive Selection, and Online Positive Selection, respectively.

## A.5. Feature Visualization

As discussed in the main body, model pre-trained by our UP-ReID has better discriminative feature learning ability than that pre-trained by MoCo v2. We fine-tune these two models in BOT [12] on Market1501 for a few epochs, respectively. Then, we visualize the feature responses of our UP-ReID and MoCo v2 in Figure A3. As we can see, in the downstream tasks, UP-ReID pre-trained model could capture identity-related attributes (*e.g.*, trouser color) and fine-grained features (*e.g.*, shoes color) better than MoCo v2 pre-trained model, which demonstrates the effectiveness of the proposed designs, like the intrinsic contrastive constraint, in our UP-ReID.

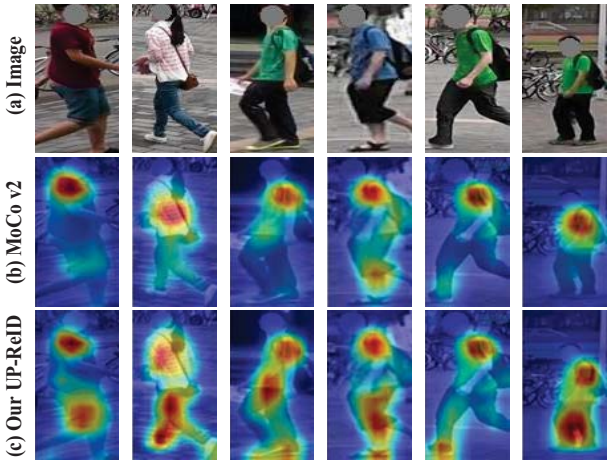



Figure A3. Visualization of the features corresponding to the MoCo v2 and our UP-ReID schemes.

## A.6. Broader Impacts

As for positive impact, we demonstrate that a suitable pre-trained model can benefit downstream person ReID tasks with higher accuracy and faster convergence speed. This will improve efficiency and effectiveness of a series of ReID tasks and save human costs in these areas.

As for negative impact, many public ReID datasets are coming from unauthorized surveillance data, which may cause an invasion of privacy and other security issues. Thus, the collection process should be public and make sure that human subjects in the datasets are aware that they are being recorded. Strict regulation should also be established for ReID datasets to avoid ethical issues.

## References


[1] 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

[2] 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. 2

[3] Xiaodong Chen, Xinchen Liu, Wu Liu, Xiao-Ping Zhang, Yongdong Zhang, and Tao Mei. Explainable person re-identification with attribute-guided metric distillation. In *ICCV*, pages 11813–11822, 2021. 2

[4] 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. 2

[5] 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

[6] 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

[7] Lingxiao He and Wu Liu. Guided saliency feature learning for person re-identification in crowded scenes. In *ECCV*, pages 357–373. Springer, 2020. 2

[8] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*, 2021. 1, 2

[9] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, volume 34, pages 11173–11180, 2020. 2

[10] 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. 1

[11] 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. 2

[12] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, pages 0–0, 2019. 3

[13] Hyunjong Park and Bumsub Ham. Relation network for person re-identification. In *AAAI*, volume 34, pages 11839–11847, 2020. 2

[14] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, pages 608–617, 2019. 1

[15] 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

[16] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*, pages 6449–6458, 2020. 2

[17] 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. 1, 2

[18] 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. 1

[19] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, pages 667–676, 2019. 2

[20] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3186–3195, 2020. 2

[21] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Jiawei Liu, Zhizheng Zhang, and Zheng-Jun Zha. Pose-guided feature learning with knowledge distillation for occluded person re-identification. In *ACM MM*, pages 4537–4545, 2021. 2

[22] 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. 1

[23] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 2

[24] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019. 2

[25] 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. Springer, 2020. 2