Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19868-4

Consistency-aware unsupervised label learning for cross-domain person re-identification

Yanbing Geng[1] · Yongjian Lian[1] · Fangshu Cui[1] · Xiaowei Zhang[2] · Mingliang Zhou[3] · Geao Zhang[4]

Received: 6 May 2023 / Revised: 23 June 2024 / Accepted: 14 July 2024

Abstract
Un-refined pseudo labels always disturb the cross-domain Re-ID performance in unsupervised clustering methods. In this paper, we propose a consistency-aware unsupervised label learning network to refine noisy labels for promising domain adaptation. Specifically, we use a hetero-generated label consistency constraint to mine possible label information, keep a reference consistency of the target pairs on the relative comparative characteristics. We then use a cross-granularity attention consistency constraint to refine the pseudo labels meanwhile learning multi-granularity discriminative representation, keeping a self-similarity consistency of the corresponding parts on the discriminative regions. Our model is fine-tuned by the refined pseudo-labels to reduce the domain gap while coping with the intra-domain variation. Experimental results demonstrate that the proposed method can achieve superior performance on several benchmark datasets.

Keywords Cross-domian person re-identification · Consistency constraint · Hetero-generated label consistency · Cross-granularity attention consistency

1 Introduction

Person Re-Identification (Re-ID) is about associating the same person captured by different cameras. Recently, deep learning techniques have boosted the Re-ID performance for single-domain person in a supervised manner [1–7]. However, the Re-ID model trained on the source domain always has a huge performance drop on the target domain. Data distribution discrepancy between domains is one factor; moreover, the absence of labeled data in the target domain impedes the universality of data-fitting models in the practical application. Self-supervised learning [8–10] tries to reduce domain gaps for unsupervised cross-domain person Re-ID. With labeled source set and unlabeled target sets, Unsupervised Domain Adaption (UDA) method [11–17] usually urges to minimize the domain discrepancy between domains in the deep neural network, such as Maximum Mean Discrepancy (MMD) [18] and Joint MMD [13]. Their assumption that the same class between domains is unreachable in the person Re-ID task, since overlapping classes hardly coexist in different domain datasets.

Yongjian Lian contributed equally to this work.

Extended author information available on the last page of the article

Published online: 25 July 2024

Releasing the dependence of the overlapped identification in different sets, style transfer [19–21] is proposed to pull closer the domain distributions by transferring the labeled image domain to the unlabeled image domain. Part alignment [1, 2, 22] is devoted to improving model scalability by using constrained part alignment. These methods are used to minimize the domain gap but ignore intra-domain image variations in the target representation learning. Intra-domain variation is effectively overcome with pairwise identities as supervision information in supervised Re-ID task, but intra-domain variation becomes hard to handle, because of label information shortage in the unlabeled target domain. In addition, the feature enhancement model with the attention mechanism [23–27] makes efforts to improve the generalization ability of models. For example, Jia et al. [23] proposes a normalization and enhancement (NE) module to suppress domain gaps without any target data. But these methods are highly dependent on labeled data, ignoring the valuable clues on the target data. It is critical work to mine the potential supervision information for unsupervised Re-ID.

Recently, supervision information has been mined by label estimation [28–35] in the unlabeled target domain. Clustering-based label estimation is commonly used. By multiple independent clusters, Fu et al. [36] assigned pseudo identities on unlabeled samples from part to whole body, and this clustering guided semi-supervised training has boosted the process of cross-domain adaption. Fan et al. [37] built progressive unsupervised learning (PUL) with the iteration of person clustering and model fine-tuning to improve the model scalability on the cross-domain application. Based on cross-view clustering for cross-domain specific feature transformation. Yu et al. [38] learned an unsupervised asymmetric distance metric. However, these clustering-based methods are unsatisfactory in discovering the identity discriminative information, since clustering is easily distorted by dramatic intra-domain variation and high inter-domain similarity. To generate robust pseudo labels against intra-domain variation, Yu et al. [35] leveraged the similarity of pair-wise target images with a set of labeled people, and investigates a similarity consistency constraint to mine hard negative samples for pushing the different people away. Analogously, Li et al. [30] generated reliable pseudo labels by measuring the similarities with the sample itself and their neighbors; these methods ignore the hard positive samples, which are of the same importance to refine the noisy pseudo labels by pulling images of the same identities close to each other. With the hypothesis that the same instances have the same pseudo labels for their corresponding local parts, cross-granularity consistency constraint [39] was introduced to refine pseudo label noise. However, for the same instances, the similarity of their corresponding parts is always destroyed by human pose variation, occlusion, and so on.

To address the above problems, we propose a consistency constraint to learn a discriminative embedding for the cross-domain Re-ID task. Firstly, we investigate a hetero-generated label consistency constraint to mine the hard negative and positive samples simultaneously, and feed the mined hard samples into the TriHard loss [40] to rectify the false pseudo labels through the model fine-tuning. Inspired that self-attention [41] can improve the person Re-ID accuracy by introducing it into the convolutional layers. Secondly, a cross-granularity attention consistency constraint is imposed on each part to mine hard samples and learn rich and discriminative feature representations. Using the similarity consistency of the discriminative regions in the corresponding parts, matching pairs should have a consistent similarity, both for both holistic discriminative features and for local discriminative features. Otherwise, they should be as hard samples to further rectify false pseudo labels through the model fine-tuning. Finally, an alternate optimization strategy is performed on the proposed consistency constraint. Specifically, the hetero-generated label consistency constraint is first used to refine the noisy pseudo label during model fine-tuning. Then feature extraction and unsupervised clustering algorithm [42, 43] are successively re-performed on the improved model to obtain multi-granularity features and refined pseudo labels. A cross-granularity attention consistency constraint is subsequently imposed on each part to refine the noisy pseudo labels. Thereby, during the model training, model scalability is improved on cross-domain Re-ID by iteratively performing pseudo label refinement.

The main contributions are as follows.

- A hetero-generated label consistency constraint(HGLCC) is used to refine noisy pseudo labels. To this end, the similarity consistency is extracted from the underlying target pairs on absolute visual characteristics and relative comparative characteristics.
- A cross-granularity attention consistency constraint(CGACC) is used on each part to further improve the robustness against noisy pseudo labels. By the joint spatial attention model, fine-grained similarity consistency is extracted from the discriminative regions in the corresponding parts for matching pairs.
- By performing an alternate optimization strategy, our Re-ID model uses two consistency constraints alternatively to fine-tuning, and our Re-ID model scalability is improved on cross domain unlabeled images.

The remainder of this paper is organized as follows. Related work is reviewed and discussed in Section 2. The proposed method is illustrated in Section 3. Experiments and comparisons are discussed in Section 4. Conclusions are drawn in Section 5.

2 Related work

2.1 Unsupervised person re-identification

Unsupervised person Re-ID refers to directly utilizing unlabeled data for Re-ID under an unsupervised manner. Being domain-independent but sensitive to noise inference, hand-crafted features [44–46] were extracted in an earlier unsupervised study. Thus performances are inferior to those of deep learning based methods. Benefiting from the strong learning ability of deep learning, a deep convolutional network is used to address unsupervised domain adaptation. GAN-based style transfer [19, 47–49] is the main solution to bridge domain gaps in the image level for cross-domain Re-ID. Deng et al. [47] proposed a similarity preserving generative adversarial network (SPGAN) to translate images from the source domain to the target domain; thus, labeled images were generated for supervised training in the target domain. However, these methods ignore the intra-domain variations in the target domain. To overcome the intra-domain variations, Zhong et al. [48] investigated a hetero-homogeneous learning (HHL) method for style transformation within the target domain to deal with the camera variance and domain shift simultaneously. However, the generated labels tend to be different from the real labels in practical application. Unsupervised learning [14, 35–37] is introduced at the feature level of the target data to identify each unlabeled image. Yu et al. [35] identified each unlabeled sample, with a set of relative similarity cores on labeled auxiliary people, but ignores the visual feature consistences mining from both between whole bodies and between parts for unlabeled image pairs; these feature consistences are the critical characteristic against intra-domain image variations. Without any target domain identity label, universal domain invariant information [22, 50] is assisted to optimize Re-ID model to overcome the domain shift. Huang et al. [22]introduced Part Segmentation(PS) constraint to improve the model generation and reach domain adaptation. Zhuang et al. [50] proposed a camera batch normalization to learn camera-invariant features and then achieved a domain adaption. Despite the fact that impressive progresses are achieved by these unsupervised domain adaptation approaches, performances are still far from those of fully supervised methods.

2.2 Pseudo label generation

Pseudo label [15, 36, 38] is used to predict the identity of an unlabeled person and to fine-tune a pre-trained mode. This approach also achieves a promising improvement on the cross-domain Re-ID task. Clustering-based models estimate pseudo labels for unlabeled data by grouping similar features of target images. To tackle the challenging domain adaption in person re-ID, Fu et al.[36]proposed a Self-Similarity Grouping (SSG) approach to assign multiple independent labels for each unlabeled image by exploiting multiple clusters of different part cues. Yu et al. [38] proposed an asymmetric metric clustering to estimate potential label for an unlabeled target image. Based on the iterations between k-means clustering and CNN fine-tuning, Fan et al.[37] constructed a progressive unsupervised learning method (PUL) to improve the scalability of Re-ID model in unseen domain. Song et al. [51] utilized a reranked distance-based clustering for pseudo label generation on unlabeled target images. Zhai et al. [52] proposed an augmented discriminative clustering (AD-Cluster) approach to estimate and augment clusters in target domains, so as to enforce the discrimination ability on unseen target set by pulling inter-person away while pushing intra-person close. However, clustering-based pseudo labels always include noisy labels, since visual features similarity is influenced by intra-domain variation.

In summary, the aforementioned methods do not consider the collaboration of visual feature and relative comparative similarity, so they are deficient in rectifying the noisy pseudo labels. We have quite different emphases with the existing clustering-based unsupervised person Re-ID method as the following aspects:

- For our HGLCC, there is a similar idea with the literature [35] on the hard negative pairs mining, but differences are as follows: (1) Generated by feature clustering with DBSCAN, pseudo labels are more reliable than the metric of pure feature similarity in MAR [35]; (2) Hard triplet loss is calculated to learn a discriminative representation during our model fine-tuning process. In addition to the hard negative pairs, hard positive pairs are also exploited to contribute to hard triplet loss.
- Our CGACC benefits from cross-granularity consistency (CGC) [39], but the differences are as follows: We introduce the spatial attention mechanism into CGC to make the consistency work even under the part occlusion, whereas the assumption that positive pairs always have consistent salient regions for both the whole body and the corresponding part.

3 Proposed method

3.1 Problem definition and overview

For a Re-ID model pre-trained with N_s labeled auxiliary images $D_s = \{x(s_i), y(s_i)\}_{i=1}^{N_s}$, $l(s_i)$ labels each image $x(s_i)$, our goal is to enhance its scalability on unlabeled target set $D_t = \{x(t_i)\}_{i=1}^{N_t}$. Considering the underlying similarity consistency possessed in unlabeled pair, consistent-aware unsupervised learning is proposed to learn discriminative representation for the cross-domain Re-ID task.

Accordingly, this paper proposes a consistent-aware unsupervised learning for a Re-ID model to predict reliable labels. As shown in the red frame of Fig. 1, hetero-generated labels consistency is used to improve the quality of the cluster under the first consistency fact. Referring to the second consistency fact, cross-granularity attention consistency is used to rich feature representation with re-fined pseudo labels, just as shown in the green frame of Fig. 1. Using a novel alternate optimization framework, we combine the benefit of the two consistency constraints to progressively generate more robust pseudo labels in the alternate iteration, just as shown in the red and green arrows.

Our framework is based on a supervised pre-trained ResNet50 model on the source domain. Different from the backbone network of existing label estimation [35, 38, 39]methods, spatial attention model is adopted on high-level feature of Resnet50 to capture discriminative regions. Target images are then input into ResNet50 to extract features. Pseudo label is assigned to each target image by adopting an unsupervised clustering algorithm [42] on the extracted features. Clustering-generated pseudo identities can effectively boost the process of domain adaption, but are easily noisy with intra-domain variation. Noisy pseudo labels influence the model generalization ability. Intrinsically, images of the same person contain two consistent facts: (1) Image pairs with similar visual features should have a consistent similarity to any other reference person; (2) with a global similarity, discriminative regions of instances are prone to have a local similarity in the corresponding part.

3.2 Spatial attention guided feature extraction

In this section, spatial attention model is used to extract global informative features and local discriminative features simultaneously. Theoretically, matching pairs should have a consistent similarity, whether for the holistic or local part, and discriminative regions in the corresponding parts always keep more stable consistency than the others. Based on this observation, spatial attention model is used to learn the discriminative features.

Given the 3-D tensor $f(s) \in R^{h \times w \times c}$, where h, w and c represent the height, width and channels, respectively, in Res 4 block as input, as shown in Fig. 2, global feature mapping of the Whole Body (WB) in the conv4_2 layer is horizontally split into two local feature mappings (one each for the Upper Part (UP) and Lower Part (LP)), and we feed the three feature mappings into the spatial attention model for salient features extraction, respectively.

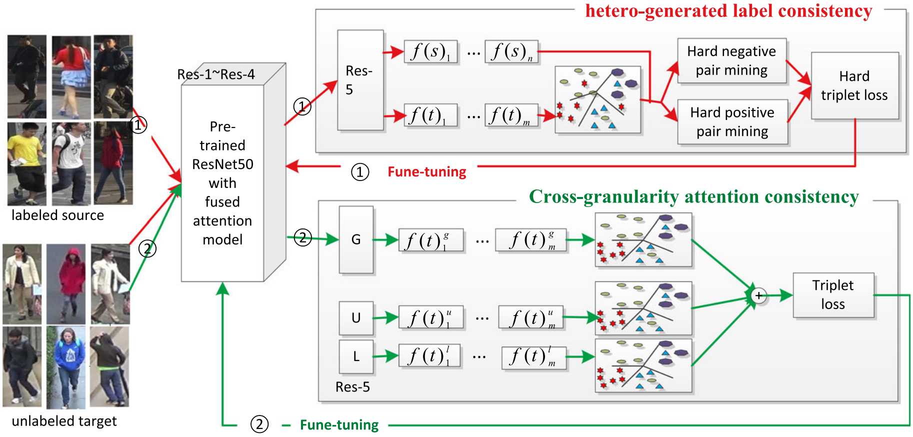

Fig. 1 The overview of our consistency-aware architecture

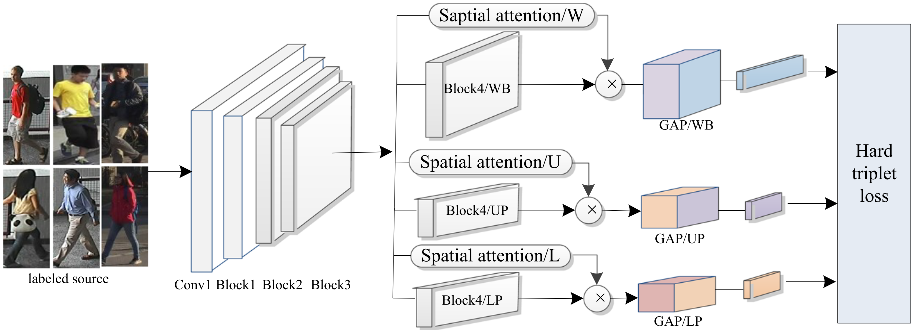

Fig. 2 The pre-trained backbone ResNet50 combined with the spatial attention on the labeled source person

As defined with (1), spatial attention mapping $M(s_i)_v$ is generated as follows: Global Average Pooling(GAP) and Global Maximum Pooling(GMP) are concurrently performed to squeeze features along the channel axes, and then forwarded to a convolutional layer with 3×3 filter and stride 1, followed by a sigmoid activation function

$$M(s_i)_v = softmax(conv_{3\times3}[Avg(f(s_i)_v)) : Max(f(s_i)_v)]), v \in (g, u, l) \tag{1}$$

With (2), three salient feature vectors $\{f(s_i)_g), f(s_i)_u), f(s_i)_l)\}_{i=1}^{N_s}$ are separately obtained by element-wise multiplying $M(s_i)_v$ and $f(s_i)_v$, and hard-batch triplet loss[40] is employed with them to mine the mis-classified samples for learning a discriminative feature representation.

$$sf(s_i)_v) = M(s_i)_v \otimes f(s_i)_v), v \in (g, u, l) \tag{2}$$

3.3 Hetero-generated pseudo label

Hetero-generated pseudo label in our work is used to label an unlabeled instance. It is tagged with a clustering-generated label and a referenced likelihood label. First and foremost, we feed each source images $x(s_i)$ into the pre-trained model to obtain feature vectors $\{f(s_i)\}_{i=1}^{N_s}$. Analogously, feature vectors $\{f(t_j)\}_{j=1}^{N_t}$ for the target image $x(t_j)$.

Clustering-generated label: DBSCAN [42] is used to generate pseudo labels for each target image. Specifically, features $\{f(t_j)\}_{j=1}^{N_t}$ of all targets are clustered with DBSCAN, and each target instance is labeled as $(x(t_j), l_c(t_j))$ according to the clustering result.

Referenced likelihood label: Following the work in the literature [35], referenced likelihood label tags each unlabeled target image by a set of similarities with all labeled source person images. Just as defined as (3), similarities of the i-th target image are firstly calculated with the inner product between $f(t_i)$ and each $f(s_j)_{j=1}^{N_s}$, and then normalized with softmax. Referenced likelihood label $l_r(t_i)$ uses the normalized similarities vector to tag the i-th target image.

$$l_r(t_i) =< softmax(f(t_i) \circ f(s_j)) >_{j=1}^{N_s} \tag{3}$$

3.4 Hetero-generated label consistency constraint

Mis-grouped instance disturbs the Re-ID performance during the model training process. Among them, mis-grouped negative pairs (i.e., hard negative pairs) are always mined to rectify the mislabeled samples during the Re-ID model fine-tuning in most clustering based methods; e.g., SML [35] mined hard negative pairs by utilizing the similarity violation of visual features and referenced likelihood consistency. However, mis-grouped positive pairs (i.e., hard positive pairs) are of the same importance, but always neglected.

By mining the hard positive pairs and negative pairs simultaneously, we apply a hetero-generated label consistency constraint to refine the original clustering-generated pseudo labels during the model training.

According to SML, if the similarity is violated between the absolute visual representation and the relative comparative characteristics for a pair of images, they probably are considered hard samples.On this condition, referenced likelihood label should be inconsistent for the same clustering-generated label. We use the referenced likelihood labels agreement $A(.,.)$ [35] proposed in SML to measure the similarity of the referenced likelihood labels, A(.,.) is defined as follows:

$$\begin{aligned} A(l_r(t_i), l_r(t_j)) &= \sum_m (min(l_r(t_i)^m, l_r(t_j)^m)) \\ &= 1 - \frac{\parallel l_r(t_i) - l_r(t_j) \parallel_1}{2} \end{aligned} \tag{4}$$

in which, $l_r(t_i)^m = softmax(f(t_i) \circ f(s_m))$, and m refers to the m-th source image. Referenced likelihood label agreement means that, target pairs are more similar when theirs referenced likelihood vectors are more consistent than others.

In a mini-batch with k instances of C clusters target sets, there are N_p positive pairs which belong to the same cluster, $N_p = C \times \frac{k(k-1)}{2}$. we mine hard negative pairs H_N as follows:

$$H_N = A(l_r(t_i), l_r(t_j)) < \alpha \tag{5}$$

in which, target image $x(t_i)$ and $x(t_j)$ belong to the same cluster, i.e., α is the threshold of positive pairs, formulated as $\rho N_p - th$ the pair in a descending order. ρ is the mining ratio of hard sample pair.

Analogically, hard positive pairs is determined, through make sure target pairs belong to the different clusters but have a higher reference similarity. In a mini-batch with k instances of C clusters target sets, N_N negative pairs are in the different clusters separately, $N_n = \frac{\sum_{i=1}^{C} \sum_{j=1}^{C} k_i K_j}{2}, i \neq j$. Hard positive pairs are determined as higher similar image pairs of referenced likelihood labels $(l_r(t_m), l_r(t_n))$ between the different clustering. We mine hard positive pairb H_P as follows:

$$H_p = A(l_r(t_m), l_r(t_n)) > \beta \tag{6}$$

in which, target image $x(t_m)$ and $x(t_n)$ belong to the different clusters, i.e., $l_c(t_m) \neq l_c(t_n)$. β is the threshold of the negative pairs, formulated as the $\rho N_n - th$ pair in an ascending order according to A(.,.). rho is the mining ratio of hard sample pair.

We mine hard negative pairs and hard positive pairs for hard triplets, which preform the discriminative representation learning by minimizing a batch-hard triplet loss [40]. Therefore, the HGLCC loss can be formulated as follows:

$$L_{HGLCC} = L_{tri}(\{x(t_i), l_c(t_i)\}_{i=1}^{H_N}) + L_{tri}(\{x(t_i), l_c(t_i)\}_{i=1}^{H_P}) \tag{7}$$

3.5 Cross-granularity attention consistency constraint

Some works [36, 37, 39, 53–56] try to explore the multi-granularity clues to learn robust representation against intra-domain variations for unlabeled instances. On the basis of the assumption that each part should have the same clustering-generated pseudo labels with the whole body, Li et al.[39] proposed a cross-granularity consistency to mitigate the negative effects of noisy pseudo labels. However, this assumption is not tenable, once pedestrian is partially occluded. Occluded part may be assigned a different pseudo label from the others.

We find that discriminative regions of the corresponding part always keep similarity consistency for a matched pair. By introducing spatial attention mechanism into cross-granularity consistency, we propose a cross-granularity attention consistency constraint to refine pseudo labels meanwhile learning multi-granularity discriminative representation.

With our pre-trained Resnet50 model mentioned in Section 3.2, multi-granularity discriminative features are extracted from the target image $x(t_i)$, i.e., $\{f(t_i)_g, f(t_i)_u, f(t_i)_l\}_{i=1}^{N_t}$. For the feature sets on each part, DBSCAN [42] is separately used to cluster pseudo labels, and image is labeled as $(x(t_i), < l_c(t_i)_g, l_c(t_i)_u, l_c(t_i)_l >)$.

To improve the robustness of multi-granularity features against intra-domain variation, we introduce CGACC into triplet loss to pull different instances far away, while pushing the same ones closer. Triplet loss is obtained by joining in the three part consistency losses:

$$L_{CGACC}(tri) = L^g_{CGACC}(tri) + L^u_{CGACC}(tri) + L^l_{CGACC}(tri) \tag{8}$$

where $L^g_{CGACC}(tri)$, $L^u_{CGACC}(tri)$ and $L^l_{CGACC}(tri)$ refer to the triplet loss on the whole body, upper part and lower part, respectively.

Since the original cluster-generated label may be different for different parts, our CGACC attempts to offset this divergence. Therefore, $L^g_{CGACC}(tri)$ focuses on the instance with consistent labels on each part and combines the triplet loss on each part. Specifically, for global representation learning, we first mine triplets according to the pseudo labels of global parts. To utilize local clues, we further mine triplets of global parts by using the pseudo labels of upper parts and lower parts. Defined as:

$$L^v_{CGACC}(tri) = L^g_{tri}(x(t_i)_v, l_c(t_i)_g) + L^u_{tri}(x(t_i)_v, l_c(t_i)_u) + L^l_{tri}(x(t_i)_v, l_c(t_i)_l), \quad v \in (g, u, l) \tag{9}$$

for the i-th training target sample $x(t_i)$, $x(t_i)_g$, $x(t_i)_u$ and $x(t_i)_l$ refer to its global part, upper part and lower part, respectively. Correspondingly, $l_c(t_i)_g$, $l_c(t_i)_u$ and $l_c(t_i)_l$ refer to its cluster-generated label of each part, respectively, N_B is the batch size.

For matched pair, discriminative regions of corresponding part can keep more stable similarity consistency than the others. Different from literature[39], each triplet loss is calculated on the discriminative features, just as follows:

$$L^g_{tri}(x(t_i)_v, l_c(t_i)_g) = \frac{1}{N_B}\sum_{i=1}^{N_B}[\varepsilon + \|sf(t_i)_g - sf(t_{p_g})_g\|_2 - \|sf(t_i)_g - sf(t_{n_g})_g\|_2], \quad p_g = l_c(t_i)_g \neq n_g \tag{10}$$

where p_g and n_g refer to the positive sample and negative sample over the anchor sample label $l_c(t_i)_g$, $sf(t_i)_g$ means its global discriminative feature, L^u_{tri} and L^l_{tri} are measured as a similar formulation with L^g_{tri}, N_B is the batch size.

3.6 Progressive learning with alternative optimization

With our pre-trained model and unlabeled target data, progressive learning is used to optimize our model parameters through an iterative pattern with pseudo label generation and model fine-tuning. During the process, we alternatively minimize losses with the two consistency constraints until model convergence. As described in Algorithm 1.

Algorithm 1 Consistent-aware unsupervised label learning.

```
Input: Pre-trained Resnet50 on labeled source data
       Unlabeled target set:$v \in (a, p, n)$, $trainSet = \{x(s_i)_{i=1}^{N_s}, x(t_j)_{i=1}^{N_t}\}$
Onput:
 while not converged do
    for trainSet=1 to n do
       feature extraction on target set
       Model fine-tuning with HGLCC:
       while not converged do
          pseudo labels with $(x(t_i), l_c(t_i))$
          Referenced likelihood labels with (3).
          for trainSet=1 to n do
             hard negative pairs mining with (4)-(5) .
             hard positive pairs mining with (6).
             parameter optimization with (7).
          end for
       end while
       Model fine-tuning with CGACC:
       while not converged do
          pseudo labels with $(x(t_i), < l_c(t_i)_g), l_c(t_i)_u), l_c(t_i)_l) >)$
          for trainSet=1 to n do
             Triplet loss with (8)-(9).
             parameter optimization with (10).
          end for
       end while
    end for
 end while
```

Based on the assumption that positive pairs should have similar consistency characteristic of visual features and referenced likelihood. Model is fine-tuned with HGLCC to make the model focus on the instances with consistent labels. During the process, we simultaneously mine hard negative pairs and hard positive pairs to minimize the HardTriplet loss. Next, spatial attention is adopted on global and local features of Res 4 block to extract multi-granularity discriminative features, which are used to obtain the clustering-generated multi-granularity pseudo labels. Subsequently, according to the observation that positive pairs should have similar discriminative regions, their global and local parts should belong to a consistent clustering group. Our Re-ID model is fine-tuned with CGACC to learn multi-granularity discriminative feature embedding to be robust within-domain variation. During the process,a triplet loss is constructed to learn finer-grained feature representation while further mitigating the inference of noisy pseudo labels.

4 Experiments

4.1 Datasets setting

Datesets. We conduct our experiments in two widely used Re-ID datasets: Market-1501(Market) [57] and DukeMTMC-reID(Duke) [58]. Market set includes 32,668 images with 1,501 people captured by 6 cameras on campus, and each person is observed by at least 2 cameras. Duke contains 36,411 images with 1,812 identities captured from 8 different view points, this set has more challenging within-domain variation relative to the others. Following the evaluation protocol [58], each set is divided into two equally half, and one half for training, the other half for testing. Experiments are performed with the setting of source to target domain on both Market→ Duke and Duke→ Market, For a fair comparison, we adopt ResNet50 or IBN-ResNet50 [59] pre-trained on ImageNet as the backbone network, respectively, where IBN-ResNet50 shows a better performance by integrating both IN and BN modules.

Implementation details. We implement our method on the pytorch platform. All person images are normalized and resized to 384×128. Data are augmented by horizontal flipping, random erasing, and cropping. As the backbone network, fully connected layer in Resnet50 is replaced by 1×1 fully convolutional layer to reduce parameters while retaining the intrinsic spatial structure of image, and network is pre-trained on the source set. With the labeled source set, mini-batch is randomly sampled with 8 instances of 16 identities during the supervised pre-training process, and with the unlabeled target set, size of mini-batch is set to 64 composed of 4 instances of 16 clustering during the unsupervised model fine-tuning. We set 0.5% as the mining ration ρ and 1500 as the number of reference persons according to the experimental analysis of MAR, the SGD optimizer is adopted with 0.9 for the momentum and 5×10^{-4} for the weight decay. The learning rate is set to 0.2×10^{-4} and multiplied by 0.1 every 100 epochs. We evaluate our proposed method using the cumulative matching characteristics(CMC) and mean average precision (mAP).

4.2 Ablation study

The proposed Consistency-aware Unsupervised Label Learning (CULL) consists of two consistency constraints: hetero-generated label consistency and cross-granularity attention consistency. We verify the effect of consistency constraint by performing ablation studies on Market and Duke sets. During each ablation experiment, we change its related setting while keeping the other architectures invariant, and the corresponding results are as follows.

I. Consistency-aware with different settings

on Market and Duke, we first investigate the effect of single consistency and joint consistency. Results are reported in Table 1. In this table, ResNet50 is used as the backbone network. "Within-domain" means that model is trained and evaluated within-domain labeled set. "Direct cross-domain" means that model is trained on the labeled source set and verified on the target set without any domain adaptation. "BaseLine" means that the pre-trained Resnet50 on source set is fine-tuning on the target set with its pseudo labels generated via DBSCAN clustering. There is a drastic performance drop of "Direct cross-domain" compared with the "Within-domain", which demonstrates the indispensable effect of domain adaption strategy on the cross-domain Re-ID task.

We further analyse the contribution of our consistency constraint. In this table, ResNet50 is used as the backbone network. "HGLCC" or "CGACC" respectively refers to the cross-domain performance fine-tuned only with hetero-generated label consistency constraint or with cross-granularity attention consistency constraints in our CULL method. "Baseline+HGLCC" represents using the proposed HGLCC method for pseudo labels refinement during the fine-tuning process, and "Baseline+CGACC" represents performing the proposed CGACC method to refine noisy pseudo labels during the fine-tuning process. Performance of the "Baseline" outperforms that of the "Direct cross-domain". In particular, mAP is improved to 67.2%/56.3% on Duke-to-Market and Market-to-Duke tasks, respectively. "Baseline+HGLCC" gets the performance gain by 2.4%/2.5% in rank-1 and 4.6%/4.8%

Table 1 Evaluation of consistency-aware in terms of different settings

	Duke→ Market				Market→Duke			
method	R1	R5	R10	mAP	R1	R5	R10	mAP
Within-domain	92.2	97.1	98.3	79	87.9	93.1	96.1	76.4
Direct cross-domain	58.7	75.7	81.5	29.3	49.4	64.8	70.3	32.6
Baseline	84.7	92.6	94.7	67.2	72.5	83.2	86.6	56.3
Baseline+HGLCC	87.1	**93.2**	**96.7**	71.8	75.4	85.7	87.1	61.6
Baseline+CGACC	**88.3**	92.7	96.5	**73.2**	**76.9**	**87.6**	**90.1**	**63.9**
CULL	**90.7**	**95.7**	**97.6**	**78.2**	**82.7**	**90.1**	**92.4**	**68.7**

$1^{st}/2^{st}$ best results are in **red**/**blue**

in mAP on two sets compared to that of "Baseline", it confirms that generated pseudo labels are more reliable with an effective clustering and ; A similar performance improvement is achieved with the setting of "Baseline+CGACC", the performance improvement verifies the contribution of our multi-granularity reference similarity on generating high-level pseudo labels in the proposed CGACC method. Finally, our "CULL" obtains the best performance by integrating the benefit of the two consistency constraints in rank-1 (90.7%/82.7%) and mAP(73.2%/63.9%) accuracy.

The above comparisons demonstrate that robustness of the generated labels can be effectively improved by mining and utilizing the potential consistency of unlabeled samples in the unsupervised Re-ID task.

II. Hetero-generated label consistency with different settings

Ablation studies are further carried out to demonstrate the effect of the hetero-generated label consistency (HGLCC).

In our work, HGLCC is used to mine hard negative pairs and hard positive pairs simultaneously, and hard triplet instances are formed to learn a discriminative representation by minimizing the hard triplet loss. Table 2 shows the results of our HGLCC on different sets. In this table, ResNet50 is used as the backbone network. "Baseline+LHGLCC w/o H_P" means that model is fine-tuned by HGLCC loss with just consideration of the hard negative pair H_N. "Baseline+LHGLCC w/o H_P" achieves a better cross-domain re-ID performance than by "Baseline" in rank 1 and mAP, which means that HGLCC can effectively mine noisy pseudo labels (i.e., negative sample) in the same cluster while improving the ability to distinguish different people with more similar appearances. We refer to the soft multilable

Table 2 Evaluation of consistency-aware with different settings

	Duke→ Market				Market→Duke			
method	R1	R5	R10	mAP	R1	R5	R10	mAP
Direct cross-domain	58.7	75.7	81.5	29.3	49.4	64.8	70.3	32.6
MAR [35]	67.7	81.9	87.3	40.0	67.1	79.8	84.2	48.0
Baseline+SA [21]	**86.5**	**94.8**	**96.8**	68.2	72.7	**85.6**	**89.4**	56.7
Baseline+DA [60]	78.7	86.4	89.2	57.1	**75.9**	84.5	87.1	**61.2**
Baseline	84.7	92.6	94.7	67.2	72.5	83.2	86.6	56.3
Baseline+HGLCC w/o H_P	85.1	92.2	95.1	**70.1**	74.3	83.7	86.7	58.9
Baseline+HGLCC	**87.1**	**93.2**	**96.7**	**71.8**	**75.4**	**85.7**	**87.1**	**61.6**

$1^{st}/2^{st}$ best results are in **red**/**blue** agreement proposed in MAR for the hard negative sample mining. Since pseudo labels are more reliable, generated by DBSCAN cluster in our method than by the feature similarity in MAR. (i.e., MAR), "Baseline+LHGLCC w/o H_P" still outperforms MAR by 17.4% /7.2% in R1 on Market and Duke. Moreover, data augmentation is always used to improve generalization ability by adding sample diversity. Data is also augmented in our training phase by simple operations such as image flipping, random erasing. In contrast, "Baseline+DA" [60] uses more complex data augmentation for the model fune-tuning by changing the channel order of original images, and it shows a clear performance gain than its "Baseline" operation. "Baseline+LHGLCC w/o H_P" also shows advantages by improving 2.4% /13% in R1/mAP on the Market than that of "Baseline+DA" [60], though there is a slight performance drop by 1.4% in R1 than "Baseline+SA" [21], which offers high diversity in model fune-tuning by data augmentation on target domain through StarGAN, but "Baseline+SA" performs a detrimental effect directly with its "Baseline" due to noise generation according to its ablation experiment. By joint hard positive pseudo labels and hard negative pseudo labels during the model fine-tuning, performance is improved of "Baseline+HGLCC" by 2% /1.1% in R1 and 1.7%/2.2% in mAP than that of "Baseline+LHGLCC w/o H_P" on Market and Duke set. This is mainly because hard positive sample plays the same important role as the hard negative samples in identifying the same instances with less visual similarity. The above results demonstrate that our HGLCC method can effectively refine noisy pseudo labels and learn a discriminative feature representation during the fine-tuning process.

III. Cross-granularity attention consistency with different settings

Effectiveness of spatial attention model. To demonstrate the effect of our spatial attention on localizing available salient regions, we compare feature maps of sample images with different settings in Figure 3. In the first line, we randomly sample 2 identities and 4 instances for each identity. Each instance presents obvious appearance variation caused by various noise, such as complex background, occlusion and pose change.

The second line exhibits feature maps without the spatial attention map, and feature maps with the spatial attention maps are shown in the third line. In which, different colors represent different salient degrees, that is, salience gradually decreases with the color from red to blue. It is observed that salient regions are wrongly located on the background regions or occluded regions for most of the instances without the attention mode, and we mark some of them by red dashed areas to make the phenomenon clearer. By comparison, spatial attention model can locate most of the salient regions in the foreground than those of the non-attention map, for example, those wrongly located salient regions are mostly excluded from our corresponding marked red dashed regions in the third row. In addition, though some salient regions are correctly located on the body-part without the attention mode, locations of these salient regions are always inconsistent for different instance of each identity. In contrast, spatial attention maps in the third line have more consistent salient regions located in torso and legs, especially for the first identity. Though instances for the second identity have more serious occlusion and view change, there are a few wrong located salient regions in the background for the attention map (such as the salient region in the sixth column). Spatial attention model still localizes most of the salient regions in the foreground than those of the non-attention map. Spatial attention is added into our cross-granularity consistency constraint to make the consistency work even under the part occlusion.

Effectiveness with different granular pseudo labels. Keeping the other phases invariant, we investigate the effect of our Cross-Granularity Attention Consistency Constraint (CGACC) with different granular pseudo labels. Table 3 shows the results of CGACC with different settings. ResNet50 is used for the first time as the backbone network. Compared with the "Direct cross-domain", there is an obvious performance improvement on

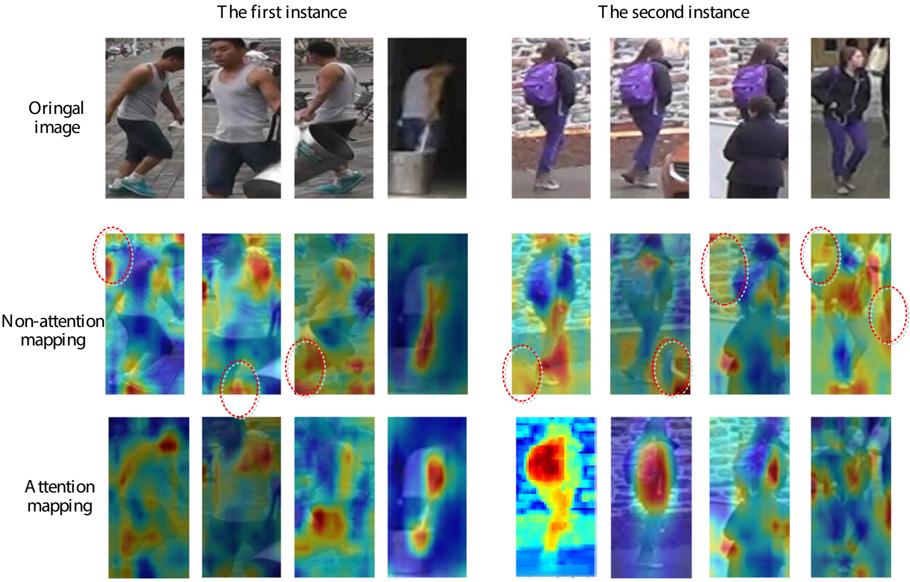

Fig. 3 Feature map visualization of our CGACC about whether with spatial attention or not. (Deep red indicates more salient, light blue means less salient)

the two sets, which verifies the contribution of our CGACC with fine-grained local cues. "Baseline+CGACCC_*g*" means the global feature triplets assigned only according to the local pseudo labels. "Baseline+CGACCC_*l*" means the local feature triplets assigned only according to the global pseudo labels. Performance of L_{cgc} gains a small margin than those of "Baseline+CGACCC_*g*" and "Baseline+CGAC_*l*" on the Duke set, since noisy pseudo labels in the L_{cgc} can be refined by reaching the features consistency with holistic discriminative cues and the local finer ones. Best performance is yielded to 88.3%/76.9% in R1 and 73.2%/63.9% in mAP with our CGACC by sharing the pseudo labels between the global and local features in our CGACC, What's more, it benefits from the discriminative feature extraction of the spatial attention model in our CGACC. IBN-ResNet50 is further adopted as the backbone as the setting in "Baseline(IBN-ResNet50)+NII", which applies the fine-tuned

Table 3 Evaluation of cross-granularity attention consistency with different settings

	Duke→ Market				Market→Duke			
method	R1	R5	R10	mAP	R1	R5	R10	mAP
Direct cross-domain	58.7	75.7	81.5	29.3	49.4	64.8	70.3	32.6
Baseline	84.7	92.6	94.7	67.2	72.5	83.2	86.6	56.3
L_{cgc} [39]	83.2	90.5	93.7	64.3	**76.5**	**87.2**	**90.4**	59.7
Baseline(IBN-ResNet50)+NII [30]	90.1	96.2	97.6	78.0	82.0	89.7	92.3	66.9
Baseline+CGACC_*g*	85.9	**93.1**	**95.0**	**68.7**	74.5	85.4	88.1	60.8
Baseline+CGACC_*l*	**86.8**	**92.8**	95.2	67.8	75.2	84.9	89.0	**61.6**
Baseline+CGACC	**88.3**	92.7	**96.5**	**73.2**	**76.9**	**87.6**	**90.1**	**63.9**
Baseline(IBN-ResNet50)+CGACC	90.8	95.7	98.1	78.8	83.1	88.7	93.1	68.6

$1^{st}/2^{st}$ best results are in **red**/**blue** for the comparisons with ResNet50
$1^{st}/2^{st}$ best results are in red/blue for the comparisons with IBN-ResNet50

model on the target domain with NII(neighborhood information integration), NII is aided to generate reliable pseudo labels by measuring sample similarity with both samples themselves and their neighborhoods. Compared to "Baseline(IBN-ResNet50)+NII", our "Baseline(IBN-ResNet50)+CGACC" can effectively improve the Rank-1 and mAP accuracy on Market and Duke set.

This result shows that, compared to just considering the mono-granular global similarity between samples in "Baseline(IBN-ResNet50)+NII", we consider the cross-granular similarity consistency both whole and corresponding part between samples, it is more credible in the reliable pseudo labels generation.

4.3 Comparisons to the state of the art

We evaluate the performance of our CULL method on the Duke → Market and Market→Duke by comparing with some state-of-the-art unsupervised methods. Comparisons are carried out mainly with the four kinds of cross-domain Re-ID methods: Non Machine Learning (NML), Unsupervised Domain Adaptation (UDA), Pseudo-label Generation (PLG) and Part Alignment Constraint (PAC). The details are as follows.

Table 4 summarizes the performance comparisons on the Duke → Market task, and analogous comparisons are performed on the Market→Duke task in Table 5. From both of them, we can observe that our CULL approach has achieved superior performance in all comparative experiments, the details are as follows.

1) Comparison With NML and UDA methods: For LOMO [45] and BoW [57] of NML methods, performance is obviously inferior to that of machine learning-based method due to the poor universal hand-crafted features. There are remarkable performance gains of UDA by style transformation from source domain to target domain in SPGAN [47] and TJ-AIDL [61], or style transformation within the target domain in HHL [48] ,or domain-invariant features extraction in NE [23], but performances are less satisfactory than PLG-based methods, because they do not benefit from pseudo-labels, which can be used to serve as supervised information for model training.

2) Comparison With PLG methods: Our CULL method belongs to the PLG mothod, it can reach an approximate supervised Re-ID performance with refined pseudo labels through the model fine-tuning on the target domain. It is obvious that our CULL surpasses most of the PLG methods, apart from a slight performance drop compared to that of ACL [21] on the Duke → Market. Compared to the use of a single network in our CULL method, mutual learning is used in ACL to extend their learning ability by transferring knowledge between two-stream models, our method also achieves a more significant performance gain than ACL on the Market → Duke, which verifies the advantage of our cross-granularity attention consistency constraint(CGACC) in coping with the dress resemblance problem of Duke, considering that Duke is more challenging than Market, imputed to its obvious resolution difference, dress resemblance, pedestrian cluster, serious occlusion, and so on. HQP(high-quality pseudo labels) [30] shares a similar idea with us on the design of soft label similarity to guide the clustering. For a fair comparison, we use IBN-ResNet50 as the backbone as HQP. Our method also outperforms the HQP method by a large margin in both target domains. Two factors should be devoted to performance improvement: (1) different from the mining of only hard negative samples in MAR, hard positive samples are considered as the same important and are mined to contribute to the hard triplet sample for a discriminative representation learning; (2) besides the discriminative global features, local cues are joined to further refine the noisy pseudo labels, simultaneously enriching the

Table 4 Comparisons with the state-of-the-art on the Duke → Market task

Methods		R1	R5	R10	mAP
NML	LOMO [45]	27.2	41.6	49.1	8
	BoW [57]	35.8	52.4	60.3	14.8
UDA	PTGAN [19]	38.6	57.3	66.1	15.7
	SPGAN [47]	51.5	70.1	76.8	27.1
	TJ-AIDL [61]	58.2	74.8	81.1	26.5
	HHL [48]	62.2	78.8	84.0	31.4
	NE [23]	65.5	80.2	85.1	34.8
PLG	PUL [37]	45.5	60.7	–	20.5
	CAMEL [62]	54.5	73.1	–	26.3
	DECAMEL [38]	60.2	76.0	–	32.4
	SSG [36]	76.0	85.8	89.3	60.3
	MAR [35]	67.7	81.9	87.3	40.0
	MPR(learn by guess)[63]	89.1	95.8	97.2	76.3
	UDAP [51]	74.7	86.9	90.3	53.7
	AD-cluster [52]	86.7	94.4	96.5	68.3
	QGML [12]	84.1	–	–	68.5
	HQP(IBN-Resnet50) [30]	92.3	96.9	97.9	80.3
	MDJL [60]	80.3	87.4	89.7	59.8
	RDA [31]	86.0	–	–	66.7
	ACL [21]	**91.5**	**96.7**	**97.9**	**77.6**
	MMT [64]	87.7	94.9	96.9	71.2
PAC	EANet [22]	67.7	–	–	48.0
	ICE [39]	**90.8**	**95.8**	97.2	73.8
ours	CULL(Resnet50)	90.7	95.7	**97.6**	**78.2**
	CULL(IBN-Resnet50)	92.8	96.1	98.2	82.7

$1^{st}/2^{st}$ best results are in **red**/**blue** for the comparisons with ResNet50
$1^{st}/2^{st}$ best results are in red/blue for the comparisons with IBN-ResNet50

representation of the features. In contrast, the generated pseudo labels in HQP method lack the reference consistency constraint between corresponding part of partner, they are easily noisy by the intra-domain variation.

3) Comparison With PAC methods: Among the PAC methods compared, our method refers to the cross-granularity consistency constraint in ICE, though ICE outforms our method by 0.1% Rank-1 on the Market set, our method can also produce satisfactory results on the Duke set. Reasons are as follows: (1) Spatial attention mechanism is introduced into CGACC to make the consistency work even under the part occlusion in our method. (2) By the generated discriminative labels, the fine-tuned model overcomes interference of intra-domain variation by pulling the intra-person features closer while pushing the inter-person features away. but we present obvious performance boosting even in the serious occlusion Duke set, which confirms the consistency hypothesis of our discriminative regions between the corresponding parts in any case.

Table 5 Comparisons with the state-of-the-art on the Market→Duke task

Methods		R1	R5	R10	mAP
NML	LOMO [45]	12.3	21.3	26.6	4.8
	BoW [57]	17.1	28.8	34.9	8.3
UDA	PTGAN [19]	27.4	–	50.7	–
	SPGAN [47]	41.1	56.6	63.0	22.3
	TJ-AIDL [61]	44.3	59.6	65.0	23.0
	HHL [48]	46.9	61.0	66.7	27.2
	NE [23]	57.0	70.6	75.3	34.7
PLG	PUL [37]	30.0	43.4	48.5	16.4
	CAMEL [62]	40.3	57.6	–	19.8
	SSG [36]	73.0	80.6	83.2	53.4
	MAR [35]	67.1	79.8	84.2	48.0
	MPR(learn by guess)[63]	82.7	90.5	93.5	69.3
	UDAP [51]	68.4	80.1	93.5	49.0
	AD-cluster [52]	72.6	82.5	85.5	54.1
	QGML [12]	76.0	–	–	60.9
	HQP(IBN-Resnet50) [30]	82.6	90.2	92.4	68.0
	MDJL [60]	78.6	86.6	88.7	62.8
	RDA [31]	75.2	–	–	59.4
	ACL [21]	78.1	**89.2**	**92.6**	63.3
	MMT(ResNet50) [64]	76.8	88.0	92.2	63.1
PAC	EANet [22]	78.0	–	–	51.6
	ICE [39]	**80.2**	88.5	91.6	**66.4**
ours	CULL(ResNet50)	**82.7**	**90.1**	**92.4**	**68.7**
	CULL(IBN-Resnet50)	83.1	90.8	94.7	70.1

$1^{st}/2^{st}$ best results are in **red**/**blue** for the comparisons with ResNet50
$1^{st}/2^{st}$ best results are in red/blue for the comparisons with IBN-ResNet50

5 Conclusion

In this paper, we constructed a deep network of consistency constraint for the unsupervised cross-domain discriminative representation learning. In this paper, a hetero-generated label consistency constraint was first introduced to refine noisy pseudo labels by mining similarity consistency of underlying target pairs on absolute visual features and relative comparative characteristics. A crossgranularity attention consistency constraint was investigated to further improve the robustness against noisy pseudo labels by mining fine-grained similarity consistency of matching pairs on discriminative regions of the corresponding part. Our Re-ID model was fine-tuned by the refined pseudo labels to reduce the domain gap while coping with the intra-domain variation. Our approach significantly outperformed the state-of-the-art methods, which verified the effectiveness of our consistency constraints on improving the cross-domain unsupervised learning.

Although experimental comparisons have validated the effectiveness of the proposed approach, compared to ACL, our method performs the optimal results in the

DukeMTMC→Market1501 task. Noisy pseudo labels are still impacting its efficiency, since DukeMTMC has more challenging within-domain variation, the generated pseudo labels are noise only by our reference consistency constraint from this source domain. In future research, we will explore the relationship between the source domain and the target domain by jointly considering some advance methods, such as adversarial training, mutual learning, and semantic segmentation.

Funding This work was supported in part by the National Natural Science Fund of Shanxi Province under grant 202103021224192; in part by the National Natural Science Fund of Shanxi Province under grant 202303021222084; in part by the Key Projects of Basic Strengthening Plan under Grant 2022-JCJQ-ZD-018-11; in part by the Chongqing Talent under Grant cstc2024ycjh-bgzxm0082; in part by the Joint Equipment Pre Research and Key Fund Project of the Ministry of Education under Grant 8091B012207; in part by the Human Resources and Social Security Bureau Project of Chongqing under Grant cx2020073; in part by the Central University Operating Expenses under Grant 2024CDTGF-044; and in part by the Guangdong Oppo Mobile Telecommunications Corporation Ltd. under Grant H20240164.

Data Availability The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the data provider's request for confidentiality of the data and results.

Declarations

Conflicts of interest The authors declare no conflict of interest.

Informed consent Informed consent was obtained from all subjects involved in the study

References

1. Guo J, Yuan Y, Huang L, Zhang C, Yao J-G, Han K (2019) Beyond human parts: Dual part-aligned representations for person re-identification. In: Proceedings of the IEEE/CVF International conference on computer vision, pp 3642–3651
2. Bai X, Yang M, Huang T, Dou Z, Yu R, Xu Y (2020) Deep-person: Learning discriminative deep features for person re-identification. Pattern Recognit 98:107036
3. Geng Y, Lian Y, Zhou M, Kong Y, Zhu Y (2020) Exploiting multigranular salient features with hierarchical multi-mode attention network for pedestrian re-identification. J Vis Commun Image Represent 73:102914
4. Zheng L, Huang Y, Lu H, Yang Y (2019) Pose-invariant embedding for deep person re-identification. IEEE Trans Image Process 28(9):4500–4509
5. Fan Z, Huang Z, Chen Z, Xu T, Han J, Kittler J (2024) Lightweight multiperson pose estimation with staggered alignment self-distillation. IEEE Transactions on Multimedia
6. Huang L, Zhang W, Nie J, Wei Z (2021) Person re-identification based on multi-appearance model. Multimed Tools Appl 80:16413–16423
7. Xu Y, Guo J, Huang Z, Qiu W (2018) Sparse coding with cross-view invariant dictionaries for person re-identification. Multimed Tools Appl 77(9):10715–10732
8. Jiang K, Zhang T, Zhang Y, Wu F, Rui Y (2020) Self-supervised agent learning for unsupervised cross-domain person re-identification. IEEE Trans Image Process 29:8549–8560
9. Wang Z, Li X, Duan H, Zhang X (2022) A self-supervised residual feature learning model for multifocus image fusion. IEEE Trans Image Process 31:4527–4542
10. Zhang J, Ge Y, Gu X, Hua B, Xiang T (2022) Self-supervised pre-training on the target domain for cross-domain person re-identification. In: Proceedings of the 29th ACM international conference on multimedia, pp 4268–4276
11. Kang G, Jiang L, Yang Y, Hauptmann AG (2019) Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4893–4902
12. Zhang L, Li H, Liu R, Wang X, Wu X (2024) Quality guided metric learning for domain adaptation person re-identification. IEEE Transactions on Consumer Electronics 14(8)

13. Long M, Zhu H, Wang J, Jordan MI (2017) Deep transfer learning with joint adaptation networks. In: International conference on machine learning, pp 2208–2217. PMLR
14. Zhang X, Li S, Jing XY, Ma F, Zhu C (2020) Unsupervised domain adaption for image-to-video person re-identification. Multimed Tools Appl 79(45):33793–33810
15. Genc A, Ekenel HK (2019) Cross-dataset person re-identification using deep convolutional neural networks: effects of context and domain adaptation. Multimed Tools Appl 78(5):5843–5861
16. Jin C (2023) Cross-database facial expression recognition based on hybrid improved unsupervised domain adaptation. Multimed Tools Appl 82(1):1105–1129
17. Yang Z, Liu G, Xie X, Cai Q (2020) Efficient dynamic domain adaptation on deep cnn. Multimed Tools Appl 79(2)
18. Long M, Cao Y, Wang J, Jordan M (2015) Learning transferable features with deep adaptation networks. In: International conference on machine learning, pp 97–105. PMLR
19. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 79–88
20. Chong Y, Peng C, Zhang J, Pan S (2021) Style transfer for unsupervised domain-adaptive person re-identification. Neurocomputing 422:314–321
21. Yao L, Lin B-Y, Haq QMU, Islam IU (2023) Unsupervised cross-domain adaptation through mutual mean learning and gans for person re-identification. In: 2023 3rd International conference on artificial intelligence (ICAI), pp 122–128. IEEE
22. Huang H, Yang W, Chen X, Zhao X, Huang K, Lin J, Huang G, Du D (2018) Eanet: Enhancing alignment for cross-domain person re-identification. arXiv e-prints, 1812
23. Jia Z, Wang W, Li Y, Zeng Y, Wang Z, Yin G (2023) Cross-domain person re-identification with normalized and enhanced feature. Multimed Tools Appl, pp 1–25
24. Liu Y, Cheng D, Zhang D, Xu S, Han J (2024) Capsule networks with residual pose routing. IEEE Transactions on Neural Networks and Learning Systems
25. Shao Z, Han J, Debattista K, Pang Y (2023) Textual Context-Aware Dense Captioning With Diverse Words. IEEE Trans Multimed 25:8753–8766
26. Liu Y, Zhang D, Zhang Q, Han J (2021) Part-object relational visual saliency. IEEE Trans Pattern Anal Mach Intell 44(7):3688–3704
27. Ji Z, Zou X, Lin X, Liu X, Huang T, Wu S (2020) An attention-driven two-stage clustering method for unsupervised person re-identification. In: Computer Vision–ECCV 2020: 16th european conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16, pp 20–36. Springer
28. Zhou S, Lei N, Zhou J, Xiong J, Zhang J (2024) The triple refinement of self-paced learning style for unsupervised cross-domain person re-identification. Image and Vision Comput 141:104870
29. Wang H, Yang M, Liu J, Zheng W-S (2023) Pseudo-label noise prevention, suppression and softening for unsupervised person re-identification. IEEE Trans Inf Forensic Secur 18:3222–3237
30. Li Y, Zhu X, Sun J, Chen H, Li Z (2023) Unsupervised person re-identification based on high-quality pseudo labels. Appl Intell 53(12):15112–15126
31. Wei P, Zhang C, Tang Y, Li Z, Wang Z (2023) Reinforced domain adaptation with attention and adversarial learning for unsupervised person re-id. Appl Intell 53(4):4109–4123
32. Fu Y, Wei Y, Zhou Y, Shi H, Huang G, Wang X, Yao Z, Huang T (2019) Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol:33, pp 8295–8302
33. Chen C, Han J, Debattista K (2024) Virtual Category Learning: A Semi-Supervised Learning Method for Dense Prediction with Extremely Limited Labels. IEEE Trans Pattern Anal & Mach Intell 4:1–17
34. Lin Y, Dong X, Zheng L, Yan Y, Yang Y (2019) A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8738–8745
35. Yu H-X, Zheng W-S, Wu A, Guo X, Gong S, Lai J-H (2019) Unsupervised person re-identification by soft multilabel learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2148–2157
36. Fu Y, Wei Y, Wang G, Zhou Y, Shi H, Huang TS (2019) Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6112–6121
37. Fan H, Zheng L, Yan C, Yang Y (2018) Unsupervised person re-identification: Clustering and fine-tuning. ACM Trans Multimed Comput Commun Appl (TOMM) 14(4):1–18
38. Yu H-X, Wu A, Zheng W-S (2018) Unsupervised person re-identification by deep asymmetric metric embedding. IEEE Trans Pattern Anal Mach intell 42(4):956–973
39. Li Y, Yao H, Xu C (2022) Intra-domain consistency enhancement for unsupervised person re-identification. IEEE Trans Multimed 24:415–425

40. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arxiv 2017. arXiv:1703.07737 4
41. Wenbai C, Lu Y, Ma H, Chen Q, Xibao W, Peiliang W (2022) Self-attention mechanism in person re-identification models. Multimed Tools Appl 81(4):4649–4667
42. Ester M, Kriegel H-P, Sander J, Xu X et al (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd 96:226–231
43. Chen HP, Shen XJ, Long JW (2016) Histogram-based colour image fuzzy clustering algorithm. Multimed Tools Appl 75(18):11417–11432
44. Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European conference on computer vision, pp 262–275. Springer
45. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2197–2206
46. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 2360–2367. IEEE
47. Deng W, Zheng L, Ye Q, Kang G, Yang Y, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 994–1003
48. Zhong Z, Zheng L, Li S, Yang Y (2018) Generalizing a person retrieval model hetero-and homogeneously. In: Proceedings of the european conference on computer vision (ECCV), pp 172–188
49. He Z, Yang B, Chen C, Mu Q, Li Z (2020) Clda: an adversarial unsupervised domain adaptation method with classifier-level adaptation. Multimed Tools Appl 79:33973–33991
50. Zhuang Z, Wei L, Xie L, Zhang T, Zhang H, Wu H, Ai H, Tian Q (2020) Rethinking the distribution gap of person re-identification with camera-based batch normalization. In: European conference on computer vision, pp 140–157. Springer
51. Song L, Wang C, Zhang L, Du B, Zhang Q, Huang C, Wang X (2020) Unsupervised domain adaptive re-identification: Theory and practice. Pattern Recognit 102:107173
52. Zhai Y, Lu S, Ye Q, Shan X, Chen J, Ji R, Tian Y (2020) Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9021–9030
53. Duan H, Long Y, Wang S, Zhang H, Willcocks C, Shao L (2023) Dynamic Unary Convolution in Transformers. IEEE Trans Pattern Anal & Mach Intell 45(11):12747–12759
54. Wang Z, Li X, Duan H, Su Y, Zhang X, Guan X (2021) Medical image fusion based on convolutional neural networks and non-subsampled contourlet transform. Expert Syst Appl 171:114574
55. Shao Z, Han J, Marnerides D, Debattista K (2022) Region-object relation-aware dense captioning via transformer. IEEE Transactions on Neural Networks and Learning Systems, pp 1–12
56. Shao Z, Han J, Debattista K, Pang Y (2024) DCMSTRD: End-to-end dense captioning via multi-scale transformer decoding. IEEE Transactions on Multimedia
57. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
58. Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE international conference on computer vision, pp 3754–3762
59. Pan X, Luo P, Shi J, Tang X: Two at once: Enhancing learning and generalization capacities via ibn-net. In: Proceedings of the european conference on computer vision (ECCV), pp 464–479 (2018)
60. Chen F, Wang N, Tang J, Yan P, Yu J (2023) Unsupervised person re-identification via multi-domain joint learning. Pattern Recognit 138:109369
61. Wang J, Zhu X, Gong S, Li W (2018) Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2275–2284
62. Yu H-X, Wu A, Zheng W-S: Cross-view asymmetric metric learning for unsupervised person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 994–1002 (2017)
63. Pereira TdC, Campos TE (2021) Learn by guessing: Multi-step pseudo-label refinement for person re-identification. arXiv:2101.01215
64. Ge Y, Chen D, Li H (2020) Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv preprint arXiv:2001.01526

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yanbing Geng[1] · Yongjian Lian[1] · Fangshu Cui[1] · Xiaowei Zhang[2] · Mingliang Zhou[3] · Geao Zhang[4]

✉ Mingliang Zhou
zml-0913yy@163.com

Yanbing Geng
gyb@nuc.edu.cn

Yongjian Lian
lyj@nuc.edu.cn

Fangshu Cui
fscui@nuc.edu.cn

Xiaowei Zhang
xiaowei19870119@sina.com

Geao Zhang
786625012@qq.com

[1] School of Data Science and Technology, North University of China, Taiyuan, Shanxi 030051, China

[2] Shandong Key Laboratory of Intelligent Information Processing, School of Computer Science and Technology, Qingdao University, Qingdao 266071, China

[3] School of Computer Science, Chongqing University, 174 Shazheng Street, Shapingba District 400044, Chongqing, China

[4] College of Information Science and Engineering, Northeastern University, 11, Lane 3, Wenhua Road, Heping District, Shenyang, Liaoning 110819, China