Knowledge-Based Systems 252 (2022) 109354

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Semantic driven attention network with attribute learning for unsupervised person re-identification

Simin Xu [a], Lingkun Luo [a], Jilin Hu [b], Bin Yang [b], Shiqiang Hu [a,*]

[a] School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai, 200240, China
[b] Department of Computer Science, Aalborg University, Aalborg, 9220, Denmark

ARTICLE INFO

Article history:
Received 14 January 2022
Received in revised form 27 June 2022
Accepted 27 June 2022
Available online 2 July 2022

Keywords:
Person re-identification
Domain adaptation
Semantic driven attention
Attribute learning

ABSTRACT

Unsupervised domain adaptation (**UDA**) person re-identification (**re-ID**) aims to transfer knowledge from a labeled source domain to guide the task proposed on the unlabeled target domain, in which people share different identifications and cross multiple camera views within two different domains. Consequently, traditional **UDA re-ID** techniques generally suffer due to the negative transfer caused by the inevitable noise generated by variant **backgrounds**, while the **foregrounds** also lack sufficient reliable identification knowledge to guarantee the qualified cross-domain **re-ID**. To remedy the raised negative transfer caused by variant backgrounds, we propose a novel *body structure estimation* **(BSE)** mechanism enforced semantic driven attention network (**SDA**), which enables the designed model with semantic effectiveness to distinguish the foreground and background. In searching for the reliable feature representations as in the foreground areas, we propose a novel label refinery mechanism to dynamically optimize the traditional attribute learning techniques for the strengthened personal attribute features and thus resulting the qualified **UDA-re-ID**. Extensive experiments demonstrate the effectiveness of our method in solving unsupervised domain adaptation person re-ID task on three large-scale datasets including Market-1501, DukeMTMC-reID and MSMT17.

1. Introduction

Person re-identification intends to identify the person of interest across non-overlapping cameras [1]. Due to its remarkable contributions to video surveillance and criminal investigation, person re-ID has received enormous research attentions in recent years. Recently, thanks to the impressive development of deep learning **(DL)** [2–4], the accuracy of supervised person re-ID has been significantly lifted via borrowing the merits of the highly discriminative feature representation enforced deep models [5–9]. However, these **DL**-based approaches crucially rely on sufficient annotated labels which are labor-intensive to obtain, thus restricting their application in real-world scenarios when confronting with the newly generated huge quantity data. To remedy this issue, increasing efforts have been made on the effective domain adaptation techniques which aim at transferring the knowledge from the well-labeled source domain to the unlabeled target domain despite the large distribution divergence among the sample distributions. For this purpose, unsupervised domain adaptation techniques greatly encourage the development of the cross-domain person re-ID.

Specifically, unsupervised domain adaptation (**UDA**) person re-ID boosts the accuracy on a fully unlabeled image dataset (target domain) by leveraging the knowledge from an existing labeled image dataset (source domain) through domain adaptation techniques. However, traditional **UDA** approaches for generic classification tasks [10–13] implicitly assume that the source domain and target domain share the same label space, while in person re-ID tasks the source and target domain are constructed by different people thus different labels. To solve this issue, pseudo label based methods [14–18] utilize the discriminative effectiveness of the source model to assign pseudo labels for recognizing the unlabeled target domain, which has received more attentions due to its simple yet effective rationale.

Previous **UDA**-based person re-ID techniques significantly lift the performance via simply enjoying the fruits of **UDA**, *i.e.*, it explicitly reduces the domain shift to ensure the qualified knowledge transferring across the different domains. However, it still falls short due to the following two challenges which are visualized in Fig. 1.

- **Challenge 1: (Varying backgrounds induced negative transfer)** We argue that existing **UDA**-based person re-ID

[*] Corresponding author.
E-mail addresses: siminxu0613@sjtu.edu.cn (S. Xu), lolinkun@gmail.com (L. Luo), hujilin@cs.aau.dk (J. Hu), byang@cs.aau.dk (B. Yang), sqhu@sjtu.edu.cn (S. Hu).

https://doi.org/10.1016/j.knosys.2022.109354
0950-7051/© 2022 Elsevier B.V. All rights reserved.

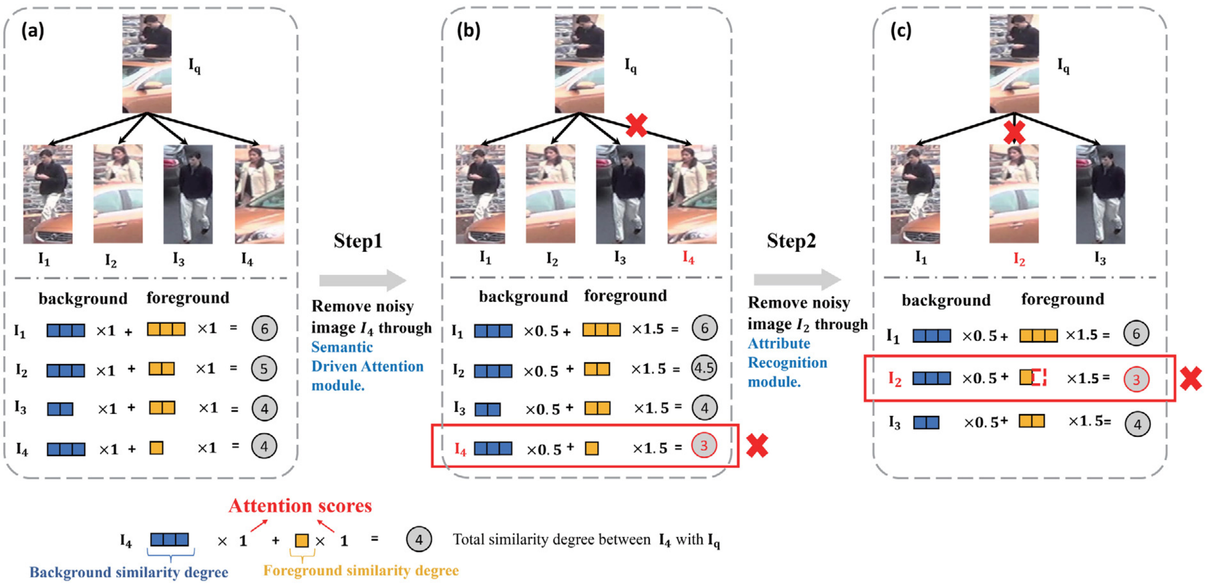


Fig. 1. Illustration of two challenges solved by our proposed **SDAAL**. In Fig. 1(a), the top part illustrates an example query image I_q with its neighbor images I_1, I_2, I_3, I_4 searched by a baseline model according to feature similarities. The bottom part provides an abstract calculation of the total similarities between each neighbor image and the query according to their background similarity and foreground similarity. Different quantities of blue and yellow rectangles denote the background and foreground feature similarity degrees respectively and larger quantity corresponds to higher feature similarity. The multipliers "1" mean the attention scores which denote how much importance to be attached to the background and foreground similarity. In Fig. 1(b), our proposed Semantic Driven Attention module re-allocates the importance assigned to the background and foreground similarities for reducing the impact of the background, thus removing the noisy image I_4. In Fig. 1(c), we strengthen the discriminability of the foreground features with the qualified attribute features according to the Attribute Recognition Module, thus decreasing the foreground similarity degree of the noisy image I_2 from two yellow rectangles denoted in Fig. 1(b) to one rectangle.

methods fail to distinguish the parts of foreground and background in the training model. Therefore, as the unavoidable noise, the varying backgrounds are generally trained during the model training, which increases the burden for effective functional learning as well as induces the negative transfer. Motivated by this, it is beneficial to enable the model to distinguish the foreground and background, so it can avoid the brutally forced knowledge transfer among the varying backgrounds.

- **Challenge 2: (Unreliable attribute feature learning caused indiscriminative feature representation)** Due to the lack of sufficient identification knowledge of the foreground people in the target domain, the discriminability of the learned model is reduced significantly. To address this problem, previous researches further explore the effectiveness of the *attribute learning* by searching the newly generated *attribute features*. Unfortunately, these methods generally suffered due to the unreliable *attribute features* extracted by the fixed *attribute learning* model, which is unable to well adapt to the unlabeled target domain without model *fine-tuning*.

To remedy the raised challenges, in this paper, we propose a novel **S**emantic **D**riven **A**ttention network with **A**ttribute **L**earning **(SDAAL)**. As shown in Fig. 1(a), previous **UDA** methods fail to distinguish the background and foreground, thus attaching equal importance to the two parts of images when training the model. However, our proposed **SDAAL** discriminatively learns the foreground/background parts to reduce the negative transfer and strengthens the attribute learning techniques to enhance the discriminative feature representation. To remedy **Challenge 1**, we first propose the semantic driven attention (**SDA**) based network, which can distinguish the foreground and background and re-weight the importance of the different foreground/background parts respectively (Fig. 1(b)). It is also interesting to note that, the proposed **SDA** is hybridized with the body structure estimation (**BSE**) mechanism, so our attention model can enjoy the high-level semantic body part features rather than the low-level pixel features. As a result, the proposed mechanism reduces the mentioned negative transfer and achieves the low computational efficiency simultaneously. To remedy **Challenge 2**, we introduce an attribute recognition module (**ARM**) with a novel label refinery mechanism, which can dynamically optimize the **ARM** for strengthening the discriminability of the foreground descriptions on the target dataset (Fig. 1(c)). Specifically, we semantically divide the image into four sub-images, of which the corresponding attribute features are extracted to formalize the required *personal attribute*. Finally, the fine-tuning of the model is implemented by the label refinery mechanism.

The main contributions of this paper are summarized as follows:

- We introduce a novel **S**emantic **D**riven **A**ttention network (**SDA**) which enables the trained model with semantic view-sight to distinguish the background and foreground, thus reducing the potentially existed negative transfer in solving the **UDA**-based reID. Moreover, by making use of the **BSE** mechanism, **SDA** enjoys the high-level semantic body part features as well as achieves high training efficiency.
- We propose a novel label refinery mechanism to improve the reliability of the *attribute features* in the foreground people of the unlabeled target domain. In this mechanism, it can dynamically optimize the traditional attribute learning techniques by well adapting the unseen target domain and extracting the qualified *attribute features*, thus yielding the qualified **UDA-re-ID**.
- Extensive experiments on three large-scale datasets demonstrate that our proposed method achieves competitive results in solving unsupervised domain adaptation person re-ID task.

2. Related works

2.1. Unsupervised domain adaptation approaches

We categorize unsupervised domain adaptation person re-ID approaches into three branches.

Learning domain-invariant feature based methods [19–22] intend to narrow the feature distribution discrepancy across the source and target domain into the common feature space using some metric measurements, *e.g.*, Maximum Mean Discrepancy (**MMD**) [23] or Earth Mover's Distance (**EMD**) [24]. Although these techniques have achieved significant progresses, they require the strict consistency of label spaces between the source domain and target domain. Unfortunately, the pedestrians of the two domains in **UDA** re-ID tasks share different identifications, thus falling short to simply embrace previous experiences to solve the **UDA** re-ID tasks.

Style transfer based methods [25–27] transfer source labeled images to match the style of target unlabeled images by applying generative adversarial networks (**GANs**) [28]. **CycleGAN** [25] proposes the approach for translating an image from a source domain to a target domain without paired examples using the adversarial loss and cycle consistency loss. To further improve the quality of translation, **PTGAN** [26] introduces an extra constraint to preserve the content similarity of images during translation by leveraging the consistency of the foreground. In addition, **SPGAN** [27] combines the Siamese network with the **CycleGAN** and proposes to perform the image translation with the constraints of two types of unsupervised similarities, *i.e.*, self-similarity and domain-dissimilarity. However, to the best of our knowledge, the mentioned style transfer methods generally care about the global distribution alignment while ignoring to explore the specific sub-domain discriminability, thus being unable to reduce the potentially existed *negative transfer*. **Pseudo label based methods** [14–17] typically consists of three main steps: 1. pre-training the feature extraction model on the labeled source domain; 2. assigning pseudo labels to the target domain according to the pre-trained model; and 3. fine-tuning the model on the target domain with pseudo labels. Generally, the reliability of pseudo labels is influenced by the domain gap between the source domain and target domain. To remedy the raised issue, Zheng et al. [29] propose to evaluate the reliability of pseudo labels by calculating the inconsistency of two models according to their predictions and then incorporate the uncertainty of samples to the objective losses. To avoid introduce extra parameters or modules, Zheng et al. [30] allow the model to output the semantic segmentation prediction as well as the uncertainty of the prediction via the prediction variance in unsupervised semantic segmentation adaptation. Inspired by the reliable sample selection methods in unsupervised learning [31,32], several popular person re-identification researches also propose to exploit how to make the best use of the target pseudo-labeled candidates through various sampling strategies. Different from the previous static sampling strategies, Wu et al. [33] propose a dynamic sampling strategy to increase the number of the selected pseudo-labeled candidates step by step for sake of model robustness. In [34], they further utilize the unselected data whose pseudo labels are not reliable for jointly fine-tuning the initial CNN model through the exclusive loss. However, they adopt the Nearest Neighbors (NN) classifier to define the confidence of label estimation as the distance between the unlabeled data and its nearest labeled neighbor, while the rationale is that the labeled and unlabeled data share the same distribution. Some other researches related to our work attempt to fully exploit the similarities between unlabeled target samples. Fu et al. [17] introduce the Self-similarity Grouping (**SSG**) method by mining the potential similarity of the global body and local parts to build multiple independent clusters. Yang et al. [16] design an asymmetric co-teaching framework which cooperates two models to select more reliable samples with pseudo labels for each other. Through the alternative training of these two models, the clustering accuracy can be guaranteed with training samples both clean and miscellaneous.

Despite the remarkable performance achieved by previous researches, inadequate disentanglement of the varying backgrounds and the foreground is still an intractable problem to solve. Our proposed framework explores the contributions of attention mechanism to distinguish the background and foreground with a semantic view-sight and employs the specifically proposed attribute learning to further strengthen the foreground feature descriptions, thus yielding a qualified **UDA** re-ID.

2.2. Attention mechanism

Inspired by the human perception scheme, attention mechanism has been witnessed superior effectiveness in natural language processing [35–37] and computer vision fields [38–40]. Starting from the Transformer [35] proposed by Vaswani et al., the effectiveness of the attention mechanism is well explored to leverage the global knowledge among the input and output dialogs for improving the machine translation tasks. Lately, Fan et al. [37] design a novel recurrent attention network to yield the attention-enhanced spatial context for Visual Dialog. In computer vision tasks, Dosovitskiy et al. [39] argue that the reliance of attention on CNN is not necessary and apply a novel Vision Transformer (ViT) to sequential image patches for further lifting the performance of image classification tasks. Fan et al. [40] firstly apply the Transformer for point cloud video modeling and design the P4 Transformer for spatio-temporal modeling by using the raw point cloud videos.

In person re-ID, attention mechanism aims to enforce more attention for identifying the informative foreground areas [41–43]. Recently, Chen et al. [5] propose a joint spatial–temporal attention model (**STAL**) to learn the quality scores of multiple spatial–temporal units. Wang et al. [44] extend the concept of self-attention [35] by calculating the interaction between pixel-pairs to obtain the global pixel-level attention respectively, thus promoting the development of attention mechanism in person re-identification. Following this research line, Chen et al. [45] segment person sequences into multiple snippets and then calculate the self-attention within each snippet for feature embedding. Liu et al. [46] apply the non-local attention module to incorporate video characteristics into the representation and validate the effectiveness of non-local attention in solving person re-ID tasks. In [47], Li et al. categorize the attention mechanism into hard region-level attention as well as the soft pixel-level attention and combine them to form a unified attention block for the optimized feature representations. However, the hard regional attention is learned simply by searching the candidate transformation matrix without considering the relationship between each regions. In this research, we explore the effectiveness of non-local attention as proposed in unsupervised domain adaptation re-ID tasks and seamlessly embed the body structure estimation into attention generation to strengthen the discriminability of re-ID model with semantic power.

2.3. Attribute learning

Attribute learning has received significant attention in person re-identification in terms of its high reputation in providing additional discriminative effectiveness through the extracted invariant property of attributes, *e.g.*, gender or age. Su et al. [48] propose a three-stage semi-supervised deep attribute learning

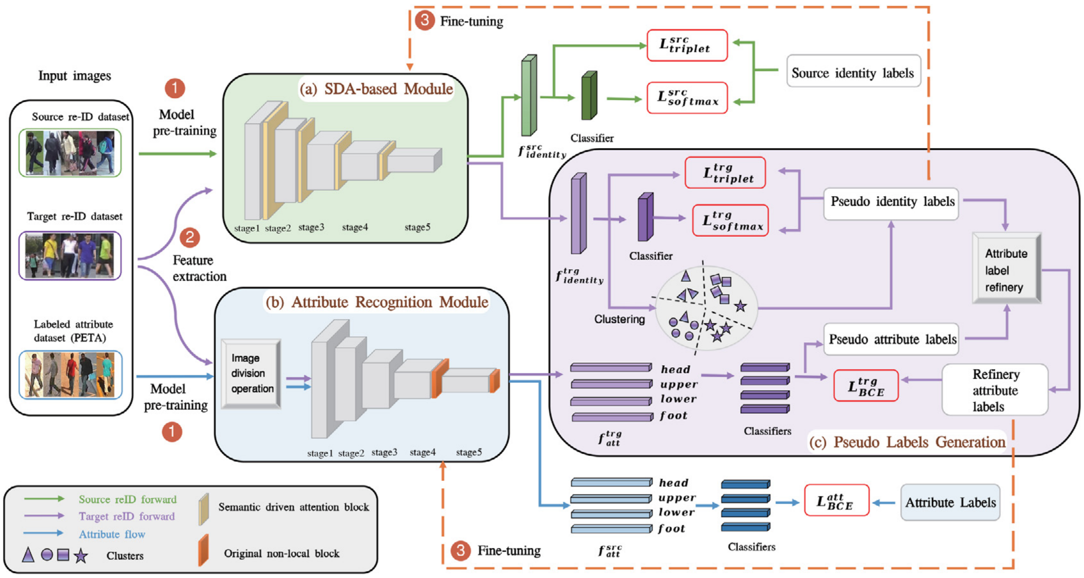


Fig. 2. The overall architecture of our proposed **SDAAL**, which consists of three key parts: (a) **Semantic Driven Attention (SDA) based Module** embedded with four **SDA** blocks enable the model to re-weight the importance of the background and the foreground, thus obtaining more discriminative identity features efficiently. (b) **Attribute Recognition Module** extracts four sub-features to predict four sub-groups of attributes associated with different body parts for strengthening the discriminability of the foreground. (c) **Pseudo Labels Generation** introduces the clustering algorithm to generate pseudo identity labels and a label refinery mechanism to optimize the initial pseudo attribute labels.

algorithm, which achieves promising performance through recognizing a large set of human attributes from a limited number of labeled attribute data. Similarly, in **ACRN** [49], the attribute classifier is pre-trained on separate data and then collaborated with the fine-tuning process of the person re-id model by using identity labels. Thereafter, Lin et al. [50] provide the attribute labels of the Market-1501 and DukeMTMC-reID datasets and demonstrate the effectiveness of multi-task learning in lifting the re-ID accuracy. In addition, attribute features also achieve potential abilities in unsupervised learning. Wang et al. [51] introduce the transferable joint attribute-identity deep learning (**TJ-AIDL**) method for simultaneously learning the attribute-semantic and identity-discriminative feature representations without any supervised knowledge in the target domain. However, previous attribute learning-based approaches generally rely on quite an amount of labeled attribute data, which is costly and sometimes impossible to obtain. Therefore, considering the experimental setting of **UDA** person re-ID, the proposed attribute learning is unable to be further optimized on the unlabeled target domain. To approach the raised issue, in this paper, we propose a novel label refinery mechanism to strengthen the feature extraction model on the unlabeled target domain via the dynamic optimizing strategy, thus ensuring the qualified attribute feature learning.

3. Methodology

To address the negative transfer caused by varying backgrounds and the insufficient identification knowledge in unsupervised domain adaptation person re-ID tasks, we propose a novel **S**emantic-**D**riven **A**ttention network with **A**ttribute **L**earning (**SDAAL**) framework which jointly unifies the semantic driven aggregation features with personal attribute information within the proposed framework. For clarification, the overall network architecture of the proposed method is shown in Fig. 2. Specifically, the proposed **SDAAL** consists of three key parts: (a) **S**emantic **D**riven **A**ttention based module (**SDA**), (b) **A**ttribute **R**ecognition **M**odule (**ARM**) and (c) Pseudo Labels Generation. In this section, we first define the problem and give a brief introduction of our proposed framework in Section 3.1. Then, we detail our proposed **SDA** and **ARM** in Sections 3.2 and 3.3 respectively. Section 3.4 describes the proposed pseudo labels generation mechanisms based on the aforementioned two feature extraction modules.

3.1. Problem definition and overall framework

Problem definition: In this paper, we intend to solve the unsupervised domain adaptation person re-identification (**UDA re-ID**) task. The research intention is to learn both the identity and attribute feature extraction models to recognize the unlabeled target dataset by using the provided source labeled re-ID dataset and attribute dataset. Specifically, each label in the attribute dataset is a $4P$-dimensional vector which indicates the attribute labels of four sub-groups and each sub-group contains P attributes. Meanwhile, one half of the target dataset will serve as the training part to fine-tune the models and the rest will be used for testing. Then our objective is that given a specific query image, following the overall feature similarities generated by the two feature extraction models, the learned identity and attribute feature extraction models retrieve all the images belonging to the same identity as the query image.

Overall process: To leverage the knowledge in the labeled source domain, the first step is to pre-train the identity feature extraction model and attribute feature extraction model on the two source datasets respectively. By using the constraints of the provided loss functions, we obtain the updated models based on the source domain. Then we apply the obtained models to the training part of the target re-ID data to extract identity features and attribute features as the second step. Lately, we perform the Density-Based Spatial Clustering of Applications with Noise (**DBSCAN**) clustering algorithm [52] on the identity features and assign pseudo identity labels according to the calculated clustering results. For the attribute recognition module, four models with the same structure are utilized to predict four sub-groups of attributes labels and we also propose a novel label refinery mechanism to optimize the pseudo attribute labels with the guidance of identity features. Subsequently, we propose to fine-tune the semantic driven attention based module in order to ensure the pre-trained model well suits the target dataset.

Through performing the feature extraction and fine-tuning iteratively to dynamically optimize the model parameters, we obtain the final two models to evaluate the performance on the testing part of the target dataset. For clarification, we denote the identity features of testing images extracted by **SDA**-based module as $\{\mathbf{f}_{iden}\}$ and the attribute features extracted by **ARM** as $\{\mathbf{f}_{att}\}$ respectively. Then we calculate the total similarity of the identity features and attribute features between the ith query image and the jth gallery image as below:

$$d(i,j) = d(f^i_{iden}, f^j_{iden}) + \alpha \cdot d(f^i_{att}, f^j_{att}), \tag{1}$$

$$d(f^i_{att}, f^j_{att}) = \frac{1}{4}\sum_{part} d(f^i_{part}, f^j_{part}), \tag{2}$$

where $d(\cdot)$ can be determined by the dot product of the two feature vectors, $\left\{\mathbf{f}_{part}\right\}$ denotes attribute features of different body parts and $part \in \{head, upper, lower, foot\}$. α is the parameter to balance the contributions of the proposed two losses. The determination of α will be discussed in Section 4.4.

For each query image, we rank the feature similarities between this image with all gallery images and calculate the re-identification accuracy of our model according to the ranking results. For more clarification of the evaluation protocol, readers are suggested to refer our Section 4.1.

3.2. Semantic driven attention based module

In order to strengthen identity feature descriptions for the unsupervised domain adaptation person re-ID task, we equip the backbone ResNet-50 with semantic driven attention modules. Thus the trained model with semantic view-sight can distinguish the background and foreground and dynamically re-weight the importance of different body parts by hybridizing the body structure estimation mechanism. In this section, we first revisit the original non-local attention block and then illustrate the details of our proposed semantic driven attention block.

3.2.1. Revisit non-local attention block

The basic non-local attention block aims at aggregating information from all positions via a pixel-level attention map. Fig. 3(a) illustrates the whole process of the original non-local attention block. We denote $\mathbf{X} \in \mathbb{R}^{C\times N}$ as the feature map of the input, where C is the dimension of features and N is the number of positions in the feature map (*e.g.* $N = H \times W$ for images, $N = H \times W \times T$ for videos). Given an input feature $\mathbf{x}_i \in \mathbb{R}^C$ sampled from $\mathbf{X}$, the corresponding output $\mathbf{z}_i$ of non-local operation can be expressed as:

$$\mathbf{z}_i = \mathbf{x}_i + W_z \sum_{j=1}^{N} \frac{e^{\theta(\mathbf{x}_i)^T\phi(\mathbf{x}_j)}}{\mathcal{C}(\mathbf{x})}(W_v \cdot \mathbf{x}_j), \tag{3}$$

where $\mathcal{C}(\mathbf{x}) = \sum_{\forall j} e^{\theta(\mathbf{x}_i)^T\phi(\mathbf{x}_j)}$ is a normalization factor, i is the index of a given query position and j enumerates all positions in the feature map. W_z and W_v are all transform matrices which are implemented as, *e.g.,* 1 × 1 × 1 convolutions. The number of channels represented by W_v is set to be half of the number of channels in $\mathbf{x}_i$, which reduces approximate 50% computation efficiency via comparing with the non-local attention block enforced version. Then the weight matrix W_z projects the aggregated feature to the original dimensional embedding space from $\mathbb{R}^{C'}(C' = \frac{C}{2})$ to $\mathbb{R}^C$, thus matching the number of channels with the given input feature $\mathbf{x}_i$.

As for the pairwise function $e^{\theta(\mathbf{x}_i)^T\phi(\mathbf{x}_j)}$ which calculates the relationship between position i and j, Wang et al. [44] propose four instantiations to meet various needs in practical applications, *i.e.,* Gaussian, Embedded Gaussian, Dot product and Concatenation: (1) Gaussian function is defined as $e^{\mathbf{x}_i^T\mathbf{x}_j}$, where $\mathbf{x}_i^T\mathbf{x}_j$ is dot-product similarity. (2) Embedded Gaussian is a simple extension of Gaussian and defined as $e^{\theta(\mathbf{x}_i)^T\phi(\mathbf{x}_j)}$, where $\theta(\mathbf{x}_i) = W_\theta\mathbf{x}_i$ and $\phi(\mathbf{x}_i) = W_\phi\mathbf{x}_i$ are two embeddings. (3) Dot product is defined as a dot-product similarity $\theta(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. (4) Concatenation is defined as $ReLU(\mathbf{w}_f^T[\theta(\mathbf{x}_i), \phi(\mathbf{x}_j)])$, where $\mathbf{w}_f$ is a weight vector that projects the concatenated vector to a scalar. In this paper, we adopt the most widely-used instantiation, Embedded Gaussian.

3.2.2. Semantic driven attention block

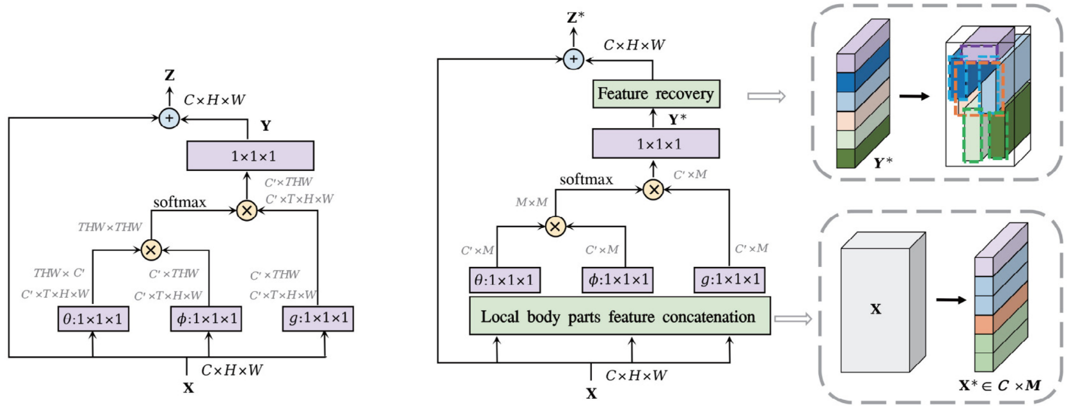
As we introduced in Section 3.2.1, the original non-local operation requires heavy computation and memory cost due to the $\mathcal{O}(N^2)$ complexity of dense affinity calculation between features of all positions. Directly embedding the non-local attention module into the backbone for feature extraction increases the training difficulty, thus preventing the potential benefit from practical application. We consider a better trade-off between computation complexity and performance and introduce a semantic driven attention block by exploring the spatial redundancy with a body structure estimation mechanism. Fortunately, our proposed **SDA** enjoys comparative performance as well as high training efficiency. Compared to the original non-local attention block, we introduce two additional components which are elaborated as follows:

Local body parts feature concatenation. We first pre-train a human pose estimation network [53] with the MPII human pose dataset [54] and then apply this network to our re-ID images to predict 14 joints of pedestrians for generating 6 salient body parts, which correspond to head, torso, right arm, left arm, right leg and left leg as illustrated in Fig. 4(a). We extract corresponding local features of all the body parts according to their positions from the input feature of the semantic driven attention module. After applying average pooling to each local feature, we obtain six feature vectors and concatenate them for the subsequent softmax attention calculation. In Fig. 4(a), the original input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ is replaced by $\mathbf{X}^* \in \mathbb{R}^{C\times M}$ ($M = 6$). It is worth noting that HW is always much larger than M, thus naturally reducing the computation cost from $\mathcal{O}(H^2W^2)$ to $\mathcal{O}(M^2)$. We then perform the non-local operation on $\mathbf{X}^*$, given an input feature $\mathbf{x}^*_i \in \mathbb{R}^C$ sampled from $\mathbf{X}$, the intermediate output $\mathbf{y}^*_i$ before feature recovery operation can be expressed as:

$$\mathbf{y}^*_i = W_z \sum_{j=1}^{M} \frac{e^{\theta(\mathbf{x}^*_i)^T\phi(\mathbf{x}^*_j)}}{\mathcal{C}(\mathbf{x}^*)}(W_v \cdot \mathbf{x}^*_j), \tag{4}$$

where $\mathbf{y}^*_i$ denotes the sum information aggregated from all body features for each $\mathbf{x}^*_i$. The 'Softmax attention aggregation' and '1 × 1 × 1' convolution between Fig. 4(a) and Fig. 4(b) are implemented by the $\sum_{j=1}^{M} \frac{e^{\theta(\mathbf{x}^*_i)^T\phi(\mathbf{x}^*_j)}}{\mathcal{C}(\mathbf{x}^*)}(W_v \cdot \mathbf{x}^*_j)$ and W_z in Eq. (4) respectively.

Note: To enrich the information of feature representations, many researchers propose to consider the similarity between part features for assisting the similarity measurement between the global features. One direct idea is to divide the feature map into equal horizontal stripes [7,55–58]. However, the uniform partition ignores to handle the semantic misalignment caused by the variations of poses within different images. To remedy this issue, the popular methods [5,59,60] take advantage of off-the-shelf pose estimation models to ensure the divided body part with pose estimation awareness, thus potentially enforcing the semantic effectiveness enhanced body part methods. Our work belongs to the latter branch but goes one step further in providing a global view-sight to see the relationships among different body parts via



(a) Original non-local attention block

(b) Our proposed semantic driven attention block

Fig. 3. Details of the original non-local attention block and our proposed semantic driven attention block. Compared to the original non-local block illustrated in Fig. 3(a), we add two components in Fig. 3(b): local body parts feature concatenation and feature recovery, which are illustrated specifically in Fig. 4.

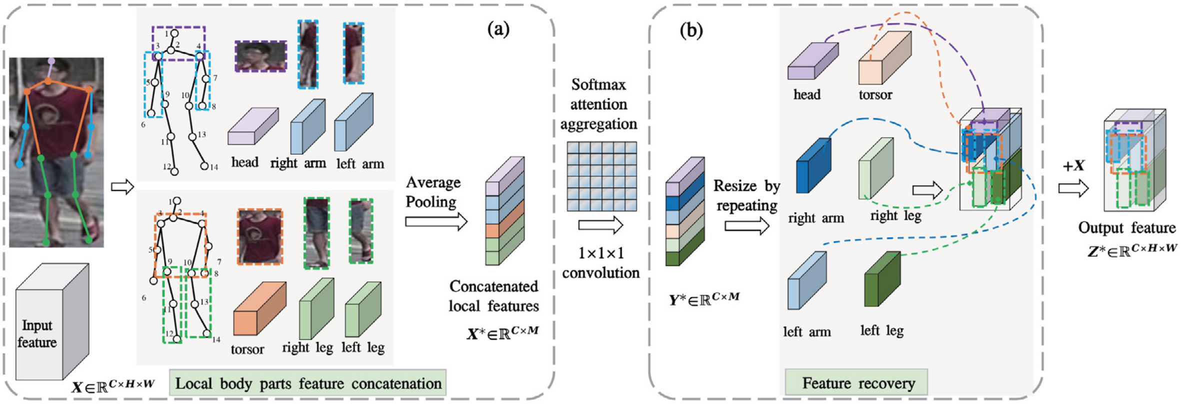


Fig. 4. Details of the local body parts feature concatenation layer and feature recovery layer in our proposed **SDA** module. We first extract the corresponding features of six salient body parts (head, right arm, left arm, torsor, right leg and left leg) from the input feature of the **SDA** according to the 14 joints predicted by the human pose estimation network. Then we apply average pooling to these six features and concatenate them for further softmax attention calculation. After updating the concatenated features, we repeat the elements of each feature vector to recover its original height and weight.

borrowing the merits of the non-local neural network [44]. Consequently, our proposed **SDAAL** enhances the discriminability of the trained model to distinguish the foreground and background for more effective cross-domain person re-identification.

Feature recovery. We repeat the element of $\mathbf{y}_i^*$ to recover the size of each aggregated local body feature according to its original size before average pooling. As illustrated in Fig. 4, we add each local body feature to the original input feature according to the corresponding position predicted by the previous body part generation, thus the final output feature of our semantic driven attention module is composed of the updated local body feature and the original background feature.

On the one hand, the intuition behind this strategy is that the pixels within the same body part are supposed to possess the similar characteristics. It is reasonable to utilize the average feature as a representative and only perform affinity calculation between average features of all the local body parts. On the other hand, since the attention computation only involves the local features of salient body parts of pedestrians, the influence of cluttered background is decreased, thus strengthening the feature descriptions when encountering background variations in unsupervised domain adaptation re-ID.

Loss function. As illustrated in Fig. 2, we utilize both the cross-entropy loss and the batch-hard triplet loss to pre-train the **SDA**-based module on the source re-ID dataset with ground truth identity labels and then fine-tune on the training part of the target dataset with pseudo identity labels.

Our total loss function for optimizing the **SDA**-based module is the combination of the two losses mentioned above:

$$L_{SDA} = L_{cross-entropy} + L_{triplet}. \qquad (5)$$

During each iteration, we extract the target features by the pre-trained source model and apply the **DBSCAN**-based clustering algorithm [52] to assign pseudo identity labels to the target images for further fine-tuning. More experimental details can be found in Section 4.2.

3.3. Attribute recognition module

Previous researches demonstrate the effectiveness of attribute learning in person re-ID tasks under an obvious assumption that images of the same person tend to share the same semantic attributes. Fig. 5 illustrates three images of two different pedestrians and the bottom two images belong to the same person. We notice that in Fig. 5(a) the feature similarities between these three images are very close due to the similar appearance with only identity labels applied while in Fig. 5(b) the bottom two images become closer to each other and far away from the top one with more detailed descriptions provided by attribute labels. Unfortunately, the lack of sufficient annotated attribute labels restricts the application of attribute learning on **UDA** person re-ID. Therefore, we introduce a novel label refinery mechanism to

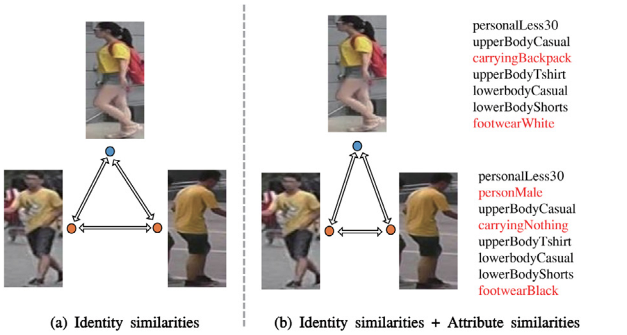


Fig. 5. The three pedestrians are of different identities. Guided with ID and attribute labels, the bottom two identities are getting closer to each other in target space while the top one is pushed far away.

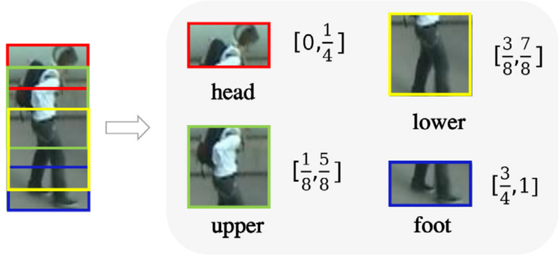


Fig. 6. The detailed image dividing strategy. Each pedestrian image is divided into four parts in height direction according to a fixed ratio.

Table 1
The partitioned attribute groups. We select 32 attributes from the PETA dataset and divide into four groups according to their associated locations.

Group name	Attribute names
Head	personalLess30, personalLess45, personalLess60, personalLarger60, accessoryHat, hairLong, personalMale, accessorySunglasses
UpperBody	carryingBackpack, carryingOther, upperBodyCasual, upperBodyFormal, upperBodyJacket, carryingNothing, upperBodyShortSleeve, upperBodyTshirt
LowerBody	lowerBodyCasual, lowerBodyFormal, lowerBodyJeans, lowerBodyShorts, lowerBodyShortSkirt, lowerBodyTrousers, lowerBodyBrown, lowerBodySuits
Foot	footwearLeatherShoes, footwearSandals, footwearShoes, footwearSneaker, footwearBlack, footwearBrown, footwearWhite, footwearStocking

dynamically optimize the attribute learning on the target domain for providing qualified personal attribute information.

We first pre-train a simple convolutional neural network on the PETA attribute dataset [61] and then apply this network to predict attribute labels of the target re-ID dataset. PETA dataset is organized by 10 publicly available small-scale datasets, including more than 60 attributes on 19,000 images of different pedestrians. In order to avoid the attributes which rarely appear in the target dataset, we in this paper only select 32 attributes which are further divided into 4 groups according to their associated locations, *i.e.* head, upperbody, lowerbody and foot. Table 1 lists the details of partitioned attribute groups. We assume that each attribute group is associated with its corresponding local part, thus each pedestrian image will be divided into 4 parts in a certain proportion along height direction. Fig. 6 clarifies the detailed experimental processing of our approach. We adopt ResNet-50 for attribute feature extraction in source pre-training phase and two non-local attention blocks are embedded into the model to enhance its ability of concentrating on informative parts during the fine-tuning phase on the target dataset. Since the size of input for attribute recognition model has been decreased due to the image division operation, we directly adopt the original non-local attention block in this module. We utilize four attribute recognition models under the same model architecture for learning the specific attribute feature of each local part.

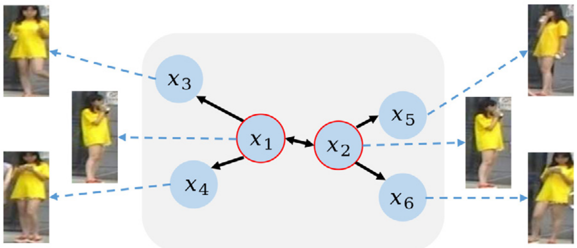
It is worth noting that Sun et al. [7] adopt a similar strategy by dividing pedestrians into different squares to describe pedestrian samples for person retrieval. They leverage Refined Part Pooling (RPP) modules to segment the divided squares into more tiny pieces for additionally capturing the global relationship among the unconnected divided squares and thereby improving the uniform partition. Our proposed **SDAAL** enjoys more efficient model training since it ignores the RPP module refined distribution measurement, while still achieving competitive performance due to the explicitly moduled attention for avoiding the negative knowledge transferring.

Loss function. We use the Binary Cross-Entropy (**BCE**) loss to train the attribute recognition module. We calculate the attribute loss of each attribute sub-group and then add them together as the final attribute recognition loss.

3.4. Pseudo labels generation

In this section, we detail the pseudo labels generation mechanisms for the two feature extraction models, *i.e.*, (1) identity feature extraction model (Fig. 2(a)) and (2) attribute feature extraction model (Fig. 2(b)), respectively.

- *Pseudo labels generation for identity feature extraction model*: In searching the pseudo identity labels, we directly apply the **DBSCAN**-based clustering algorithm [52] to identity features and generate the pseudo labels according to the clustering results.
- *Pseudo labels generation for attribute feature extraction model*: In this experimental setting, we specifically introduce the pseudo attribute label refinery mechanism in order to enhance the reliability of the calculated pseudo labels. Specifically, we perform the label refinery mechanism to resist the noisy pseudo attribute labels through borrowing the identity prediction results from the semantic driven attention module. According to the identity features extracted by the **SDA** module, we first calculate the feature similarities between every two images and rank the similarities between other images for each image. Given an image x_i and its nearest neighbor image x_j, if x_i is also the nearest neighbor of x_j, we assume that they belong to the same person and denote x_i and x_j as a reliable pair. After finding all these reliable pairs, we select the top p percent of them according to the feature similarities. In our experiments, we set $p = 70$ according to the accuracy of identity prediction. Then we revise the attribute prediction results of the pre-trained attribute recognition model according to the assumption that images belong to the same person should possess the same attributes. For each selected reliable pair, we join the second and third nearest neighbors of each image to form a group. If the predicted attribute labels of images in the group differ to each other, we revise the labels of the minority to subject to the majority. The label refinery mechanism is detailed in Fig. 7. Finally, we fine-tune the attribute recognition module with the revised attribute labels.



images	Predicted attribute labels of head part							
x_1	1	0	0	0	0	~~0~~ 1	0	0
x_2	~~0~~ 1	0	0	0	0	1	~~1~~ 0	0
x_3	1	0	0	0	0	1	0	0
x_4	1	0	0	0	0	1	0	0
x_5	1	0	0	0	0	1	0	0
x_6	~~0~~ 1	0	0	0	0	~~0~~ 1	0	0

Fig. 7. Pseudo attribute label refinery mechanism. We take the attributes of the head part for an example and the numbers of the table in this figure denote the predicted labels for the attributes listed in the first row of Table 1. Image x_1 is the nearest neighbor of image x_2 and vice versa, we denote them as a reliable pair. Image x_3, x_5 are the second nearest neighbors of image x_1 and image x_2, respectively. Image x_4, x_6 are the third nearest neighbors of image x_1 and x_2. We compare the predicted attribute labels of the six images mentioned above and revise the possible wrong labels. Take the first attribute of head part for an example, since the labels of more than a half images in the group are 1, we then assign all the images with label 1.

4. Experiments

4.1. Datasets and evaluation metrics

To evaluate the effectiveness of our proposed method, we conduct experiments on three standard datasets: DukeMTMC-reID [62], Market-1501 [63] and MSMT17 [26]. The evaluation statistics are summarized in Table 2 with some samples illustrated in Fig. 8. All three datasets are divided into two parts for training and testing respectively, whenever possible we directly borrow the experimental settings as reported in the previous researches [17,26,64] for fair comparison.

DukeMTMC-reID [62] dataset is a subset of the multi-target multi-camera tracking dataset which contains eight 85-minutes' high-resolution videos captured from eight different cameras. This dataset includes 36,411 images of 1,404 pedestrians, which are further divided into three parts: 16,522 images of 702 pedestrians as training set, 17,661 images of 1,110 pedestrians as the gallery and 2,228 images of 702 pedestrians from the initial selection of the gallery as the query.

Market-1501 [63] dataset consists 32,668 images from 1,501 pedestrians captured by six different cameras on the campus of Tsinghua University. All these pedestrian images are automatically detected by the DPM detector. Similar to the division of the DukeMTMC-reID dataset, we use 751 pedestrians with totally 12,936 images as the training set, 750 pedestrians with totally 19,732 images as the gallery and 3,368 images selected from the same 750 pedestrians in the gallery as the query.

MSMT17 [26] dataset is the largest re-ID dataset which includes 126,441 bounding boxes of 4,101 identities from 15 cameras during 4 days. These 15 cameras include 12 outdoor and 3 indoor ones. The whole dataset is divided into 32,621 images of 1,041 identities for training and 93,820 images of 3,060 identities for testing. To our best knowledge, the MSMT17 dataset is the most challenging re-ID dataset with large-scale images and multiple cameras.

Table 2
The evaluation setting statistics of three datasets. DukeMTMC-reID, Market-1501 and MSMT17 are abbreviated as Duke, Market and MSMT, respectively.

Benchmark	Train ID	Test ID	Image
Duke	702	702	36,411
Market	751	750	32,668
MSMT	1,041	3,060	126,441

Evaluation protocol. In this work, experimental results are evaluated by the standard Cumulative Match Characteristic (CMC) and mean average precision (mAP). We measure the performance of our proposed model in terms of Rank-1, Rank-5 and Rank-10 with CMC, where Rank-n indicates the average matching correct rate among the top-n images with the highest confidence. The mAP denotes the mean of different hit probabilities. Following the settings of state-of-the-art unsupervised re-ID methods, we evaluate our proposed method on the above three datasets and under four benchmark protocols, including Market→Duke, Duke→Market, Market→MSMT and Duke→MSMT.

4.2. Implementation details

In our experiments, all input images are uniformly resized to 256 × 128 and synchronously augmented with random erasing to ensure each pedestrian with more than 8 images. Then we randomly select 4 identities and sample 8 images for each identity to form the mini-batch for training. We adopt the ImageNet pretrained ResNet-50 as our backbone network and modify *conv5_1* to stride 1 instead stride 2 to better adapt the re-ID task. For our semantic driven attention based module, we insert 4 semantic driven attention block after *conv1_1, conv2_2, conv3_3 and conv4_4* respectively during fine-tuning. For our attribute recognition module, we insert one original non-local attention block after *con3_3* and another one after *conv4_4*. We train our **SDA**-based feature extraction network for 200 epochs with both the cross-entropy loss and the batch-hard triplet loss and choose Adam optimizer with an initial learning rate of 10^{-4} and decay it by 10 every 50 epochs. As for the attribute recognition module, we train the network for 150 epochs with binary cross-entropy loss and choose Adam optimizer with an initial learning rate of 10^{-3} and decay it by 10 every 50 epochs. For the **DBSCAN**-based clustering algorithm applied in identity pseudo labels assignment, we constrain the minimum size of a cluster to 4 and set density radius p = 35. Other parameters are kept the same as in [14]. After a clustering step, we fine-tune the model on the target dataset with pseudo identity labels for 15 epochs, and iterate this procedure for 10 rounds to obtain the final **SDA**-based model.

4.3. Comparison with state-of-the-art methods

We compare our method with multiple unsupervised state-of-the-art methods using three large-scale datasets including Market-1501, DukeMTMC-reID and MSMT17 datasets.

4.3.1. Performance on DukeMTMC-reID and Market-1501 dataset

Table 3 reports the comparison results on DukeMTMC and Market-1501 datasets under fully unsupersived setting (**FU**) and unsupervised domain adaptation setting (**UDA**). We test the performance of seven different **FU** methods including **UMDL** [65], **CAMEL** [66], **PUL** [67], **BUC** [64], **CrossCamera** [55], **JVTC** [68] and **HCT** [69]. From the results we can see that our proposed **SDAAL** surpasses most of these **FU** methods by a large margin via borrowing the valuable information from labeled source dataset. For instance, **BUC** achieves a rank-1 accuracy of 66.2% on the Market-1501 dataset and 47.4% on the DukeMTMC re-ID dataset, which is 16.4% and 25.4% lower than our **SDAAL**.

Fig. 8. Some samples in Market-1501, DukeMTMC-reID and MSMT17 datasets.

Table 3
Unsupervised person re-id performance comparison with state-of-the-art methods on Market-1501 and DukeMTMC-reID datasets. We mark the second-best results by underline and the best results by **bold** text.

Methods	Reference	Market-1501				DukeMTMC-reID			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
UMDL [65]	CVPR16	34.5	52.6	59.6	12.4	18.4	31.4	37.6	7.3
CAMEL [66]	ICCV17	54.5	–	–	26.3	–	–	–	–
PUL [67]	TOMM18	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
BUC [64]	AAAI19	66.2	79.6	84.5	38.3	47.4	62.6	68.4	27.5
CrossCamera [55]	TIP20	73.7	84.0	87.9	38.0	56.1	66.7	71.5	30.6
JVTC [68]	ECCV20	79.5	89.2	91.9	47.5	74.6	82.9	85.3	50.7
HCT [69]	CVPR20	80.0	91.6	95.2	56.4	69.6	83.4	87.4	50.7
PTGAN [26]	CVPR18	38.6	–	66.1	–	27.4	–	50.7	–
SPGAN [27]	CVPR18	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
TJ-AIDL [51]	CVPR18	58.8	74.8	81.1	26.5	44.3	59.6	65.0	23.0
MMFA [70]	BMVC18	56.7	75.0	81.8	27.4	45.3	59.8	66.3	24.7
ARN [19]	CVPR18	70.3	80.4	86.3	39.4	60.2	73.9	79.5	33.4
CamStyle [71]	CVPR18	58.8	78.2	85.3	27.4	48.4	62.5	68.9	25.1
HHL [72]	ECCV18	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
ATNet [73]	CVPR19	55.7	73.2	79.4	25.6	45.1	59.5	64.2	24.9
ECN [74]	CVPR19	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
SSG [17]	ICCV19	80.0	90.0	92.4	**58.3**	**73.0**	80.6	83.2	53.4
DAAM [20]	AAAI20	77.8	89.9	93.7	53.1	71.3	82.4	**86.3**	48.8
IPL [14]	PR20	75.8	89.5	93.2	53.7	68.4	80.1	83.5	49.0
MMCL [75]	CVPR20	80.3	89.4	92.3	45.5	65.2	75.9	80.0	40.2
ACT [16]	AAAI20	80.5	–	–	60.6	72.4	–	–	**54.5**
AE [76]	TOMM20	81.6	**91.9**	94.6	58.0	67.9	79.2	83.6	46.7
SDAAL	This paper	**82.6**	91.7	**94.7**	56.7	72.8	**82.5**	86.1	52.3

Table 4
Unsupervised person re-id performance comparison with state-of-the-art methods on MSMT17 datasets. 'Market → MSMT' represents the source domain is Market-1501 and the target domain is MSMT17. 'Duke → MSMT' represents the source domain is DukeMTMC-reID and the target domain is MSMT17. We mark the second-best results by underline and the best results by **bold** text.

Methods	Reference	Source	MSMT17			
			R-1	R-5	R-10	mAP
PTGAN [26]	CVPR18	Market	10.2	–	24.4	2.9
ECN [74]	CVPR19	Market	25.3	36.3	42.1	8.5
SSG [17]	ICCV19	Market	31.6	–	49.6	13.2
SDAAL	This paper	Market	**40.1**	**51.5**	**56.8**	**17.4**
PTGAN [26]	CVPR18	Duke	11.8	–	27.4	3.3
ECN [74]	CVPR19	Duke	30.2	41.5	46.8	10.2
SSG [17]	ICCV19	Duke	32.2	–	51.2	13.3
SDAAL	This paper	Duke	**47.0**	**58.1**	**63.7**	**20.4**

To demonstrate the effectiveness of our method, we also evaluate the performance of fifteen **UDA** methods: **PTGAN** [26], **SPGAN** [27], **TJ-AIDL** [51], **MMFA** [70], **ARN** [19], **CamStyle** [71], **HHL** [72], **ATNet** [73], **ECN** [74], **SSG** [17], **DAAM** [20], **IPL** [14], **MMCL** [75], **ACT** [16] and **AE** [76]. In this setting, when tested on Market-1501 dataset, DukeMTMC-reID is used as the source and vice versa. The experimental results report that the performance of **GAN** based methods are much lower than pseudo label based methods. For example, **TJ-AIDL** obtains a rank-1 accuracy of 58.8% when using DukeMTMC-reID as a source dataset and tested on Market-1501, exceeding **SPGAN** by 7.3%. Another pseudo label based method **SSG** exploits both global and local similarities to build clusters, thus achieving a comparative result of 80.0% on Market-1501 dataset and 73.0% on DukeMTMC-reID dataset. Specifically, our proposed **SDAAL** achieves a rank-1 accuracy of 82.6% on the Market-1501 and 72.8% on DukeMTMC-reID, which mainly thanks to the implement of semantic-based spatial relation within pedestrian features for more reliable clustering. When compared to **TJ-AIDL** which also considers attribute information, our proposed method outperforms it by 23.8% and 28.5% respectively on rank-1 accuracy. In addition, although the performance of our **SDAAL** is slightly inferior to **SSG** on DukeMTMC-reID dataset, we obtain an improvement of 2.6% on Market-1501 dataset.

4.3.2. Performance on MSMT17 dataset

To further verify the effectiveness of our algorithm, we conduct experiments on a larger and more challenging dataset MSMT17. Following the experimental setting as the state of the art methods, we take Market-1501 and DukeMTMC-reID datasets as the source domain respectively and MSMT17 as the target domain. Considering that MSMT17 is a newly released dataset, only three unsupervised domain adaptation methods **PTGAN**, **ECN** and **SSG** are reported in Table 4. From the table we can see that our proposed **SDAAL** also achieves comparative performance on MSMT17, especially for taking DukeMTMC-reID dataset as the source domain. We achieve 47.0% in rank-1 accuracy and 20.4% in mAP, exceeding the **SSG** by 14.8% and 7.1% respectively. Those experimental results clearly demonstrate the superior performance of the proposed method.

4.3.3. Comparison with different losses

Table 5 compares the performance of our proposed **SDAAL** with different loss function designs on three datasets under the four benchmark protocols, including Duke→Market, Market→ Duke, Duke→MSMT and Market→MSMT. From the table we can see that the results exhibit slight fluctuations when combining the cross-entropy loss with other losses including the triplet loss [77], sphere loss [78], lifted loss [79], instance loss [80], contrastive loss [9] and circle loss [81]. Among these losses, the instance loss provides a proper initialization for ranking loss and further regularizes the training process, thus achieving the best results on Rank-1 accuracy. Furthermore, considering the remarkable contribution of the contrastive loss in verification problem, we also investigate the performance of adding the contrastive loss as well as another loss to the cross-entropy loss for improving the final retrieval results. We notice that increasing the quantity of different losses properly can lead to better performance. As visualized in Table 5, the best performance is achieved in using CE+Constrast+Sphere loss, which generally reaches the best Rank-1 accuracy and mAP across all transfer tasks.

4.4. Parameters analysis

In this section, we investigate the effect of different values of the hyper-parameter α which balances the identity similarity and attribute similarity obtained by the **SDA**-based module and the **ARM** respectively. We utilize the original pre-trained

Table 5
Evaluation of different loss functions on cross-domain re-ID tasks with our proposed **SDAAL**. We report Rank-1 accuracy (%) and mAp (%) and mark the best results by **bold** text. 'CE' represents the cross-entropy loss.

Loss function	Duke to Market		Market to Duke		Duke to MSMT		Market to MSMT	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
CE+Triplet	82.6	56.7	72.8	52.3	47.0	20.4	40.1	17.4
CE+Sphere	82.2	56.3	72.4	51.8	46.6	19.9	39.7	17.0
CE+Lifted	81.9	56.7	72.1	52.3	46.2	20.3	39.8	17.4
CE+Instance	82.8	56.8	73.0	52.4	47.2	20.5	40.3	17.6
CE+Contrast	82.5	57.0	72.6	52.6	46.9	20.6	40.0	17.8
CE+Circle	82.6	57.1	72.8	52.7	47.0	20.8	40.1	18.0
CE+Contrast+Triplet	82.7	57.2	72.8	52.8	47.1	20.9	40.2	18.1
CE+Contrast+Circle	82.4	57.2	72.6	52.8	46.8	20.9	40.0	18.1
CE+Contrast+Sphere	**82.9**	**57.3**	**73.1**	**52.9**	**47.3**	**21.1**	**40.4**	**18.2**

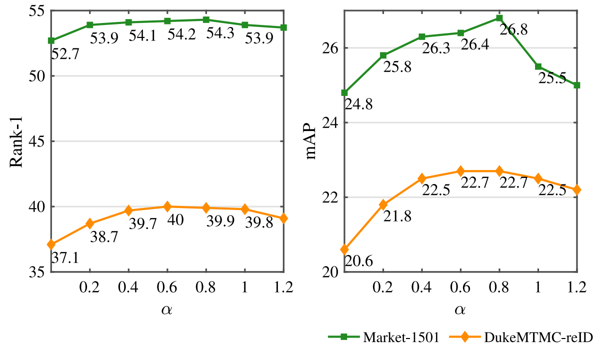


Fig. 9. Evaluation of different values of parameter α which balances the identity similarity and attribute similarity.

Table 6
Evaluations of different fine-tuning strategies on two datasets with ResNet-50 baseline. 'Direct transfer' means directly applying the pre-trained source model to the target dataset for inference. '**Res** fine-tune' means fine-tuning the model with original ResNet-50 structure. '**NL** fine-tune' and '**SDA** fine-tune' add the original non-local block and the proposed semantic driven attention block to the ResNet-50 respectively for fine-tuning on the target dataset.

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
Direct transfer	52.7	24.8	37.1	20.6
Res fine-tune	79.7	50.5	69.3	46.9
NL fine-tune	81.5	54.8	71.4	50.4
SDA fine-tune	**81.9**	**56.1**	**72.0**	**51.5**

Table 7
Computation and memory statistics comparison between the original non-local block and our **SDA** block. We report the GPU memory, training time and FLOPs when processing an input mini-batch of 32 images. (FLOPS: Floating-point operations per second.).

Methods	Memory (MB)	Time (ms)	FLOPs (G)
Res fine-tune	6868	87.6	130.24
NL fine-tune	8676	142.3	190.54
SDA fine-tune	**6919**	**97.5**	**132.42**

source models to extract identity features and attribute features. Then we calculate the final accuracy of person re-identification according to the total similarities under different values of α following (7). Experimental results are shown in Fig. 9, which analyze the Rank-1 accuracy and mAP. The value of α ranges from 0.2 to 1.2 and the step size is 0.2. From the results, we can see that for any value of $\alpha \geq 0$, our strategy systematically improves the results of direct transfer. More specifically, when $\alpha \in [0.2, 1.2]$, the performance is affected only slightly and the optimal result is obtained when α is set to 0.8. This confirms that our approach is insensitive to small variations of α. It is worth noting that the DPM detector enforced attribute similarity on the Market-1501 dataset is less reliable than the quality of carefully labeled attribute similarity on the DukeMTMC-reID dataset. As a result, the performance of mAP on the DukeMTMC-reID dataset expresses more flat curvature in re-weighting the importance of the attribute similarity and the identity similarity.

4.5. Ablation studies

In this section, comprehensive ablation evaluations are conducted to investigate the contribution of individual components in our proposed approach.

Semantic driven attention module. To demonstrate the superiority of our improved **SDA**-based module, we adopt three different model structures for identity feature extraction: ResNet-50 as the baseline (**Res** fine-tune), ResNet-50 embedded with the original non-local attention layers (**NL** fine-tune) and ResNet-50 embedded with our proposed **SDA** module (**SDA** fine-tune). Table 6 compares the results on target datasets after fine-tuning. As the experimental results show, by directly applying the pre-trained source model to the target dataset, the rank-1 accuracy on two datasets are 52.7% and 37.1% respectively. After fine-tuning with the original ResNet-50 structure, the rank-1 accuracy reaches 79.7% on the Market-1501 dataset and 66.7% on the DukeMTMC-reID dataset. With the original non-local layers added, the rank-1 accuracy is improved by 1.8% on the Market-1501 dataset and 2.1% on the DukeMTMC-reID dataset compared to the **Res** fine-tune. Then we measure the accuracy of fine-tuning with our proposed **SDA**-module. The results show that our proposed model structure has a improvement of 2.2% and 2.7% on the two datasets than the baseline **Res** fine-tune.

We also compare the computation and memory statistics between the original non-local block (**NL** fine-tune) with our proposed semantic driven attention block (**SDA** fine-tune) in Table 7. From the table we can see that the GPU memory, training time and FLOPs are largely reduced by using our proposed **SDA** block rather than the original non-local block. As shown in Table 7, our proposed **SDA** fine-tune increases limited computational burden comparing to the baseline setting (Res fine-tune) while significantly reducing the computational burden as reported in **NL** fine-tune. Therefore, through exploring the merits of the body structure estimation mechanism, our proposed **SDA** network can enjoy comparative performance as well as high training efficiency.

Attribute recognition module. Attributes are utilized to provide additional information for further confirming whether two images belong to the same person. In Table 8, the baseline with direct transfer yields only 52.7% and 37.1% rank-1 accuracy on the Market-1501 dataset and DukeMTMC re-ID dataset. When

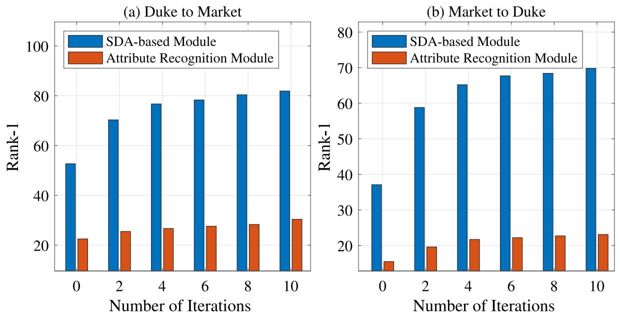


Fig. 10. Evaluation of re-ID accuracies for two networks.

Table 8
Ablation studies of the proposed framework on two datasets with ResNet-50 baseline. The analysis shows the influences by different components and design choices on Rank-1 and mAP (%).

Component	Design choice				
Direct transfer		√	√	√	√
Attribute recognition			√	√	√
Identity fine-tune				√	√
Attribute fine-tune					√
Duke to Market	Rank-1	52.7	54.3	81.9	**82.6**
	mAP	24.8	26.8	56.1	**56.7**
Market to Duke	Rank-1	37.1	39.9	72.0	**72.8**
	mAP	20.6	22.7	51.5	**52.3**

attribute similarities are added to the identity similarities, improvements of 2.6% and 2.9% are achieved on these two datasets respectively. Such improvements show that attribute recognition plays a certain role in assisting person re-identification. After several iterations for fine-tuning, the final results achieve 82.6% and 72.8% in rank-1 accuracy on the Market-1501 dataset and DukeMTMC re-ID dataset respectively, exceeding the results without the attribute fine-tuning process by 0.7% and 0.8%, which demonstrate the effectiveness of the label refinery mechanism.

Fine-tuning. Several works demonstrate that fine-tuning is a powerful strategy in unsupervised domain adaptation tasks. We fine-tune the **SDA**-based feature extraction network with the pseudo identity labels. The experimental results show that the rank-1 accuracy is increased by 29.2% on the Market-1501 dataset and 32.7% on the DukeMTMC re-ID dataset. As for the attribute recognition module, we also assign pseudo attribute labels according to the attribute prediction results. After fine-tuning for the attributes recognition with the label refinery mechanism, our final framework achieves 82.6% rank-1 accuracy on the Market-1501 dataset and 72.8% on the DukeMTMC re-ID dataset. Fig. 10 illustrates the accuracy of two models during the iterations of fine-tuning process. We can observe from the results that both two models perform better as the number of iterations increases, which verifies the effectiveness of all the sub-networks in our proposed method.

4.6. Visualization

To further investigate the discriminative ability of the iterative clustering strategy, we randomly select 10 identities with their images from the target dataset and extract their feature embeddings during iterations. We use t-SNE [82] to visualize the embeddings by plotting their 2-dimension feature representations in Fig. 11. Each point represents one image and points with the same color indicate pedestrians with the same identity. From the visualization results, we can observe that our model is better than direct transfer after the first stage of iteration.

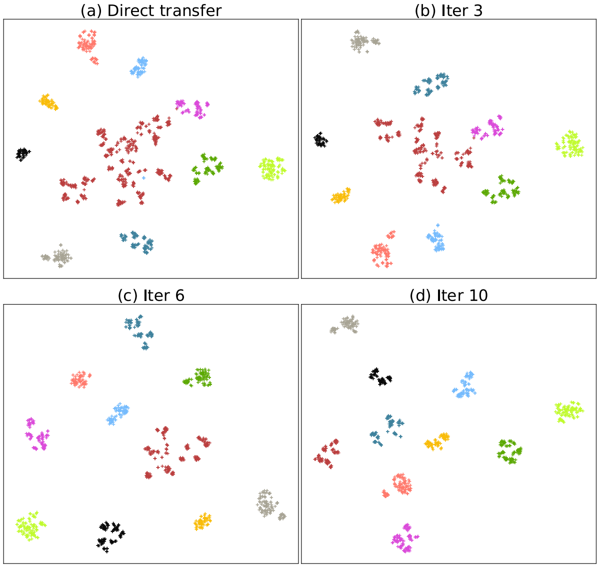


Fig. 11. Feature distributions visualized by t-SNE. Fig. 11(a): The model is trained on Market-1501 and directly transferred to DukeMTMC-reID. Fig. 11(b)(c)(d): The results of our proposed SDA-based model after fine-tuning 3, 6 and 10 iterations respectively.

By increasing the number of iterations, we observe a clear and constant gathering of points with the same color, which indicates that the model has gradually learned more discriminative feature representations. This visualization demonstrates that our proposed **SDAAL** effectively strengthen the discriminability of feature representations, thus enforcing the target images with the same identity to gather together based on their similarities after some stage of iterations.

5. Conclusion

In this research, we propose a novel **S**emantic **D**riven **A**ttention network with **A**ttribute **L**earning method (**SDAAL**) in solving the existing challenges of traditional **UDA**-based person re-ID techniques. In order to remedy the varying backgrounds induced negative transfer, we introduce the body structure estimation enforced Semantic Driven Attention network, which effectively reduces the negative impacts caused by the varying backgrounds as well as enjoys high training efficiency. Additionally, we propose a novel label refinery mechanism in order to properly optimize the attribute feature learning model for extracting reliable attribute feature representations, and thus yielding the qualified **UDA** re-ID. Extensive experimental results demonstrate that our proposed framework achieves very competitive re-ID accuracies to the state-of-the-art approaches. Future work includes hybridizing the META learning techniques into the paradigm of **SDAAL** for searching the best candidate hyper-parameters to accelerate the global optimization and lift the accuracy simultaneously.

CRediT authorship contribution statement

Simin Xu: Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Lingkun Luo:** Writing – review & editing. **Jilin Hu:** Writing – review & editing. **Bin Yang:** Writing – review & editing. **Shiqiang Hu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61773262, 62006152) and the Aeronautical Science Fundation of China (20142057006).

References

[1] L. Zheng, Y. Yang, A.G. Hauptmann, Person re-identification: Past, present and future, 2016, arXiv preprint arXiv:1610.02984.

[2] T. Kieu, B. Yang, C. Guo, C.S. Jensen, Distinguishing trajectories from different drivers using incompletely labeled trajectories, in: CIKM, 2018, pp. 863–872.

[3] C. Guo, B. Yang, J. Hu, C.S. Jensen, L. Chen, Context-aware, preference-based vehicle routing, VLDB J. 29 (5) (2020) 1149–1170.

[4] P. Yuan, C. Sha, X. Wang, B. Yang, A. Zhou, S. Yang, XML structural similarity search using MapReduce, in: WAIM, 2010, pp. 169–181.

[5] G. Chen, J. Lu, M. Yang, J. Zhou, Spatial-temporal attention-aware learning for video-based person re-identification, IEEE Trans. Image Process. 28 (9) (2019) 4192–4205, http://dx.doi.org/10.1109/TIP.2019.2908062.

[6] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint discriminative and generative learning for person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2138–2147.

[7] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 480–496.

[8] Y. Li, X. Jiang, J.-N. Hwang, Effective person re-identification by self-attention model guided feature learning, Knowl.-Based Syst. 187 (2020) 104832.

[9] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person reidentification, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 14 (1) (2017) 1–20, http://dx.doi.org/10.1145/3159171.

[10] Z. Wang, X. Shu, C. Chen, Y. Teng, L. Zhang, J. Xu, A semi-symmetric domain adaptation network based on multi-level adversarial features for meningioma segmentation, Knowl.-Based Syst. (2021) 107245.

[11] J. Li, G. Li, Y. Shi, Y. Yu, Cross-domain adaptive clustering for semi-supervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2505–2514.

[12] Y. Zhang, M. Ye, Y. Gan, W. Zhang, Knowledge based domain adaptation for semantic segmentation, Knowl.-Based Syst. 193 (2020) 105444.

[13] S.B. Yang, C. Guo, J. Hu, J. Tang, B. Yang, Unsupervised path representation learning with curriculum negative sampling, in: IJCAI, 2021, pp. 3286–3292.

[14] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, X. Wang, Unsupervised domain adaptive re-identification: Theory and practice, Pattern Recognit. 102 (2020) 107173.

[15] J. Lv, W. Chen, Q. Li, C. Yang, Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7948–7956.

[16] F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng, X. Guo, F. Huang, R. Ji, S. Li, Asymmetric co-teaching for unsupervised cross-domain person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 12597–12604.

[17] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, T.S. Huang, 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, 2019, pp. 6112–6121.

[18] J.-P. Ainam, K. Qin, J.W. Owusu, G. Lu, Unsupervised domain adaptation for person re-identification with iterative soft clustering, Knowl.-Based Syst. 212 (2021) 106644.

[19] Y.-J. Li, F.-E. Yang, Y.-C. Liu, Y.-Y. Yeh, X. Du, Y.-C. Frank Wang, Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 172–178.

[20] Y. Huang, P. Peng, Y. Jin, J. Xing, C. Lang, S. Feng, Domain adaptive attention model for unsupervised cross-domain person re-identification, 2019, arXiv preprint arXiv:1905.10529.

[21] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, W. Zuo, Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2272–2281.

[22] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, PMLR, 2015, pp. 97–105.

[23] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A. Smola, A kernel method for the two-sample-problem, Adv. Neural Inf. Process. Syst. 19 (2006) 513–520.

[24] Y. Rubner, C. Tomasi, L.J. Guibas, The Earth Mover's Distance As a Metric for Image Retrieval, The Earth Mover's Distance as a Metric for Image Retrieval, 2000.

[25] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[26] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 79–88.

[27] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, J. Jiao, 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, 2018, pp. 994–1003.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[29] K. Zheng, C. Lan, W. Zeng, Z. Zhang, Z.-J. Zha, Exploiting sample uncertainty for domain adaptive person re-identification, 2020, arXiv preprint arXiv:2012.08733.

[30] Z. Zheng, Y. Yang, Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation, Int. J. Comput. Vis. 129 (4) (2021) 1106–1120.

[31] H. Fan, X. Chang, D. Cheng, Y. Yang, D. Xu, A.G. Hauptmann, Complex event detection by identifying reliable shots from untrimmed videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 736–744.

[32] H. Fan, P. Liu, M. Xu, Y. Yang, Unsupervised visual representation learning via dual-level progressive similar instance selection, IEEE Trans. Cybern. PP (99) (2021) 1–11, http://dx.doi.org/10.1109/TCYB.2021.3054978.

[33] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, Y. Yang, Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5177–5186.

[34] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, Y. Yang, Progressive learning for person re-identification with one example, IEEE Trans. Image Process. 28 (6) (2019) 2872–2881, http://dx.doi.org/10.1109/TIP.2019.2891895.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[36] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, 2018, CoRR abs/1810.04805 arXiv:1810.04805.

[37] H. Fan, L. Zhu, Y. Yang, F. Wu, Recurrent attention network with reinforced generator for visual dialog, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 16 (3) (2020) 1–16, http://dx.doi.org/10.1145/3390891.

[38] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[40] H. Fan, Y. Yang, M. Kankanhalli, Point 4D transformer networks for spatio-temporal modeling in point cloud videos, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021.

[41] M. Zhang, M. Xin, C. Gao, X. Wang, S. Zhang, Attention-aware scoring learning for person re-identification, Knowl.-Based Syst. 203 (2020) 106154.

[42] Y. Xu, L. Zhao, F. Qin, Dual attention-based method for occluded person re-identification, Knowl.-Based Syst. 212 (2021) 106554.

[43] M. Liu, K. Wang, R. Ji, S.S. Ge, J. Chen, Pose transfer generation with semantic parsing attention network for person re-identification, Knowl.-Based Syst. 223 (2021) 107024.

[44] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[45] D. Chen, H. Li, T. Xiao, S. Yi, X. Wang, Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1169–1178.

[46] C.-T. Liu, C.-W. Wu, Y.-C.F. Wang, S.-Y. Chien, Spatially and temporally efficient non-local attention network for video-based person re-identification, 2019, arXiv preprint arXiv:1908.01683.

[47] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2285–2294.

[48] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Deep attributes driven multi-camera person re-identification, in: European Conference on Computer Vision, Springer, 2016, pp. 475–491.

[49] A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 20–28.

[50] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recognit. 95 (2019) 151–161.

[51] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2275–2284.

[52] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, vol. 96, (34) 1996, pp. 226–231.

[53] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with Human Body Region guided feature decomposition and fusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.

[54] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D human pose estimation: New benchmark and state of the art analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2014.

[55] Y. Lin, L. Xie, Y. Wu, C. Yan, Q. Tian, Unsupervised person re-identification via softened similarity learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3390–3399.

[56] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 274–282.

[57] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, T. Huang, Horizontal pyramid matching for person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, (01) 2019, pp. 8295–8302.

[58] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, J. Sun, Alignedreid: Surpassing human-level performance in person re-identification, 2017, arXiv preprint arXiv:1711.08184.

[59] M.M. Kalayeh, E. Basaran, M. Gökmen, M.E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1062–1071.

[60] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, W. Gao, Attention driven person re-identification, Pattern Recognit. 86 (2019) 143–155.

[61] Y. Deng, P. Luo, C.C. Loy, X. Tang, Pedestrian attribute recognition at far distance, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 789–792.

[62] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3754–3762.

[63] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.

[64] Y. Lin, X. Dong, L. Zheng, Y. Yan, Y. Yang, A bottom-up clustering approach to unsupervised person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8738–8745.

[65] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, Y. Tian, Unsupervised cross-dataset transfer learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1306–1315.

[66] H.-X. Yu, A. Wu, W.-S. Zheng, Cross-view asymmetric metric learning for unsupervised person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 994–1002.

[67] H. Fan, L. Zheng, C. Yan, Y. Yang, Unsupervised person re-identification: Clustering and fine-tuning, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 14 (4) (2018) 1–18, http://dx.doi.org/10.1145/3243316.

[68] J. Li, S. Zhang, Joint visual and temporal consistency for unsupervised domain adaptive person re-identification, in: European Conference on Computer Vision, Springer, 2020, pp. 483–499.

[69] K. Zeng, M. Ning, Y. Wang, Y. Guo, Hierarchical clustering with hard-batch triplet loss for person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13657–13665.

[70] S. Lin, H. Li, C.-T. Li, A.C. Kot, Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification, 2018, arXiv preprint arXiv:1807.01440.

[71] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camera style adaptation for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5157–5166.

[72] Z. Zhong, L. Zheng, S. Li, Y. Yang, Generalizing a person retrieval model hetero-and homogeneously, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 172–188.

[73] J. Liu, Z.-J. Zha, D. Chen, R. Hong, M. Wang, Adaptive transfer network for cross-domain person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7202–7211.

[74] Z. Zhong, L. Zheng, Z. Luo, S. Li, Y. Yang, Invariance matters: Exemplar memory for domain adaptive person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 598–607.

[75] D. Wang, S. Zhang, Unsupervised person re-identification via multi-label classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10981–10990.

[76] Y. Ding, H. Fan, M. Xu, Y. Yang, Adaptive exploration for unsupervised person re-identification, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 16 (1) (2020) 1–19, http://dx.doi.org/10.1145/3369393.

[77] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, 2017, arXiv preprint arXiv:1703.07737.

[78] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, SphereFace: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.

[79] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.

[80] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embeddings with instance loss, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 16 (2) (2020) 1–23, http://dx.doi.org/10.1145/3383184.

[81] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, Y. Wei, Circle loss: A unified perspective of pair similarity optimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020.

[82] P.E. Rauber, A.X. Falcao, A.C. Telea, et al., Visualizing time-dependent data using dynamic t-SNE, 2016.