An Attention-driven Two-stage Clustering Method for Unsupervised Person Re-Identification

Zilong Ji[1], Xiaolong Zou[2], Xiaohan Lin[2], Xiao Liu[3], Tiejun Huang[2], and Si Wu[2,3]

[1] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, China. jizilong@mail.bnu.edu.cn

[2] School of Electronics Engineering & Computer Science, Peking University, Beijing, China.

[3] IDG/McGovern Institute for Brain Research,Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China.

{xiaolz,Lin.xiaohan,xiaoliu23,tjhuang,siwu}@pku.edu.cn

Abstract. The progressive clustering method and its variants, which iteratively generate pseudo labels for unlabeled data and per form feature learning, have shown great process in unsupervised person re-identification (re-id). However, they have an intrinsic problem of modeling the in-camera variability of images successfully, that is, pedestrian features extracted from the same camera tend to be clustered into the same class. This often results in a non-convergent model in the real world application of clustering based re-id models, leading to degenerated performance. In the present study, we propose an attention-driven two-stage clustering (ADTC) method to solve this problem. Specifically, our method consists of two strategies. Firstly, we use an unsupervised attention kernel to shift the learned features from the image background to the pedestrian foreground, which results in more informative clusters. Secondly, to aid the learning of the attention driven clustering model, we separate the clustering process into two stages. We first use kmeans to generate the centroids of clusters (stage 1) and then apply the k-reciprocal Jaccard distance (KRJD) metric to re-assign data points to each cluster (stage 2). By iteratively learning with the two strategies, the attentive regions are gradually shifted from the background to the foreground and the features become more discriminative. Using two benchmark datasets Market1501 and DukeMTMC, we demonstrate that our model outperforms other state-of-the-art unsupervised approaches for person re-id.

Keywords: Attention, Clustering, Unsupervised Learning, Person Re-id

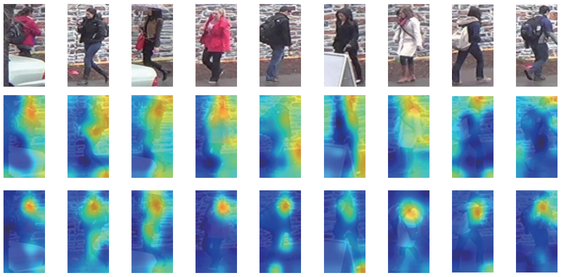

Fig. 1. Examples of class activation maps (CAMs) of pedestrians extracted from the same camera. From top to bottom are the original images, the CAMs without attention, and the CAMs with attention (the attcention mechanism is described in Sec.3.1). Without attention, the CAMs highlight more on the background, leading to that images from the same camera are likely to be assigned to the same cluster. With attention, the CAMs focus more on the informative features of pedestrians.

1 Introduction

The difficulties faced by supervised learning have motivated people to develop unsupervised person re-id models which is more applicable in the real world setting. One promising approach is the clustering-based method. The idea is to train a clustering model for the unlabeled data points and a feature learning model from the pseudo-labeled dataset in a iterative manner. However, in a real world re-id system, pedestrian images detected in the same camera often share similar background. This results in a clustering model which assigns pedestrian features extracted from the same camera into the same cluster. Such model shows great attention to the image background and fails to capture the in-camera variability of images (Fig. 1). Therefore, it is necessary to shift the foci from the background to the foreground during the implementation of the clustering based model. Under the setting of supervised person re-id, it is often done by introducing an attention kernel to highlight the informative features of pedestrians (e.g., logos on clothes, backpacks) and suppresses uninformative ones (e.g., the background) [23, 38, 41, 14]. However, due to the lack of supervisory signals under the setting of unsupervised person re-id, it is hard for the attention model to learn correct attentive regions. An alternative way is to use the off-the-shelf pose estimation model to propose hard attentive local regions [34], but this introduces local network branches which increases computational complexity of the model.

In the present study, to solve the aforementioned challenges, we propose an Attention-Driven, Two-stage Clustering method, referred to as ADTC hereafter (Fig. 2A), for unsupervised person re-id task. Specificaly, we adopt a voxel attention kernel to highlight the features of images that are informative for pedestrian discrimination. This attention mechanism enhances the informative spatial regions for pedestrians and recalibrates the channel-wise feature information adaptively according to the inter-dependencies between channels. As a result, it

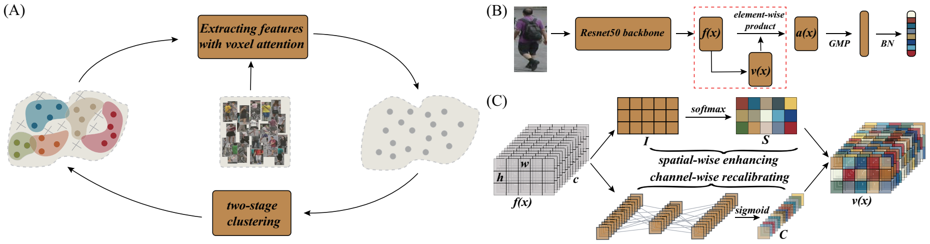

Fig. 2. The scheme of our method ADTC. (A) Our model consists of two iterative operations, the voxel attention and the two-stage clustering. The gray shadow denotes the manifold of feature representations at the current round, and different colors represent different clusters. (B) The feature extractor of our model. GMP denotes global max-pooling and BN batch normalization. (C) The detail of the attention kernel in the red dotted box in (B).

enlarges the separations between the negative and positive image pairs with respect to a query. Moreover, this voxel attention kernel only has a small number of trainable parameters, avoiding the overfitting problem during the iterative training. Furthermore, to improve the training of the attention-related parameters under the unsupervised setting, we adopt a two-stage clustering process to generate pseudo-labels for data points. We first use kmeans++ [1] to generate the centroids of clusters and then apply the k-reciprocal Jaccard distance (KRJD) metric [45] to re-assign data points to each cluster. Due to the appealing property of KRJD, data points belonging to the same class are more likely to be aggregated together, and the clustering quality of images is significantly improved, which in return facilitates the training of the model parameters. Overall, in our model, data clustering (generating pseudo-labels) and model training (optimizing feature representations with attention) are executed iteratvely (Fig. 2A), and they promote each other to achieve good performances. Using benchmark datasets, we demonstrate that the proposed model can largely correct the mistakes made by the previous clustering based models (Fig. 4) and outperform other state-of-the-art unsupervised models for person re-id. The main contributions of this paper include:

- We propose to use an unsupervised voxel attention strategy to correct the mistakes made by the clustering based re-id models.
- We propose to use a two-stage clustering strategy to generate pseudo-labels for data points, which improves the clustering quality and stabilizes the progressive training.
- Our model achieves the state-of-the-art performances under the unsupervised setting for person re-id on a number of benchmark datasets.

2 Related Work

2.1 Unsupervised Person Re-ID

Traditional unsupervised person re-id studies have mainly focused on feature engineering [9, 7, 42, 13], which created hand-craft features using human prior knowledge that can be applied directly to the unsupervised learning paradigm. These methods are efficient for a small dataset, but often fail to deal with a large dataset, since they can not fully exploit the data distribution to extract the appropriate semantic features. Recently, the domain adaptation strategy has been widely used for unsupervised person re-id [24, 18, 25, 33, 32], which attempts to reduce the discrepancy between the source and target data domains. During training, the knowledge learned from the source domain is continuously transferred to the target domain to facilitate the learning process. For example, Lin et al. [20] developed a feature alignment method to align the source and target data in the feature space by jointly optimizing the classification and alignment losses. Deng et al.[5] proposed a SPGAN model to preserve the similarity between two domains and integrate image translation and model learning. However, these approaches rely heavily on the assumption that the two domains have similar distributions. When the discrepancy between two domains is large, there is no guarantee that these methods will work well. Another direction for unsupervised person re-id is the clustering-based method [6, 28, 40, 21, 39, 8], which generates pseudo-labels by clustering data points in the feature space and then use these pseudo-labels to train the model as if in the supervised manner. Fan et al. [6] proposed a progressive clustering method to transfer the pre-learned deep representations to an unseen domain, where feature clustering and representation learning are performed iteratively like the EM-style algorithm. Lin et al. [21] proposed a bottom-up clustering approach to jointly optimize a convolutional neural network and the relationship between the individual samples. Recently, Yang et.al [39] introduced the asymmetric co-teaching startegy in the clustering based method. For a clustering-based unsupervised model, the clustering quality of data is crucial. Compared to the existing clustering-based models, our method has two differences: 1) we use an attention mechanism to drive the clustering process, and 2) we cluster data points in two stages using a more appropriate distance metric. It turns out that our method improves the clustering quality significantly, which further leverages the model performances (see the details in Sec. 3.1 and Sec. 3.2).

2.2 Attention in Person Re-ID

The attention in a person re-id model aims to highlight the informative features of images to avoid the mis-alignments due to pose variance, occlusion, or body parts missing in a bounding box [36, 27, 3, 49, 4]. The attention mechanisms proposed in the literature can be divided into two main categories: hard-attention and soft-attention. The former typically uses a pose estimation model to locate coarse regions and then exploit these local features for discrimination [34, 43, 30, 15]. However, these hard region-level attentions rely heavily on the pose estimation, which is often inaccurate and does not consider the pixel-level information within the selected regions that are potentially important for the identification task. A soft-attention mechanism typically inserts trainable layers into the main body of the model to mask the convolutional feature maps, so that the informative regions are highlighted [16, 38, 31, 2]. Two main soft-attention mechanisms are widely used: the spatial attention and the channel attention. The former enables the model to pay attention to the valuable features at different spatial locations, and the latter enables the model to improve the representational power by performing channel-wise recalibration. There are also works combining the two soft-attention mechanisms. For example, Li et al. [17] proposed a Harmonious Attention Convolutional Neural Network (HA-CNN) which combines the pixel-level spatial information and the scale-level channel information to jointly learn the attentive regions and feature representations. Notably, so far the attention mechanism has only been used under the supervised setting; here we apply it under the unsupervised setting which is much harder to optimize.

3 Our Approach

3.1 Voxel attention (VA)

We first introduce the voxel attention strategy[4]. Given an input image $\boldsymbol{x}$ in the unlabeled dataset X, denote the output of the backbone model as the corresponding feature map $\boldsymbol{f}^{w\times h\times c}$, where w, h, c are the values of width and height, and the number of channels, respectively. The attention feature map $\boldsymbol{a}^{w\times h\times c}$ is defined as (for clearance, we omit the superscript hereafter),

$$\boldsymbol{a} = \boldsymbol{v} \odot \boldsymbol{f}, \tag{1}$$

where $\boldsymbol{v}$ is the voxel attention kernel having the same size as $\boldsymbol{f}$, and $\odot$ denotes the element-wise product. $\boldsymbol{v}$ is composed of two complementary parts: the spatial and the channel attentions (Fig. 2C). For the spatial attention part, we first calculate the mean intensity of activation at each spatial location along the corresponding channel, which is given by $I(i,j) = \sum_{l=1}^{c} \frac{f(i,j,l)}{c}$; afterwards we apply softmax to calculate the probability of $I(i,j)$, which is $\boldsymbol{S}(i,j) = e^{I(i,j)} / \left[\sum_{i,j} e^{I(i,j)}\right]$. Here, the divisive normalization makes the spatial filters competitive (acts like global inhibition) to highlight the most active (informative) ones. Note that no trainable parameter is introduced for the spatial attention branch. For the channel attention branch, we adopt the idea of [11] and apply a squeeze-and-excitation block to improve the quality of representations. Firstly, we perform global average pooling on $\boldsymbol{f}$ to squeeze the global spatial information into a channel descriptor $\boldsymbol{C}_{in}^{c}$, with each element $c_{in}^{l} = \sum_{i=1,j=1}^{h,w} \frac{f(i,j,l)}{(h\times w)}$ aggregating the feature information distributed across the spatial space in channel l. Secondly,

[4] The term of voxel attention comes from that it is a 3D attention mask combining the spatial and channel attentions.

to capture the inter-dependencies between different channels in $\boldsymbol{f}$, we employ a gating function on $\boldsymbol{C}_{in}^{c}$ by forming a bottleneck with two fully connected layers, i.e.,

$$\boldsymbol{C} = \sigma\left[W_2 ReLU(W_1 \boldsymbol{C}_{in})\right], \tag{2}$$

where σ represents the sigmoid function, $W_1 \in \boldsymbol{D}^{d\times c}$, $W_2 \in \boldsymbol{D}^{c\times d}$, $\boldsymbol{C} \in \boldsymbol{D}^{c}$, with $d \ll c$. The total number of parameters in the channel attention part is only $2cd$, which is computationally efficient. Eventually, the voxel attention kernel $\boldsymbol{v}$ can be written as tensor multiplication between $\boldsymbol{S}$ and $\boldsymbol{C}$,

$$\boldsymbol{v} = \boldsymbol{S} \times \boldsymbol{C}, \tag{3}$$

i.e., each voxel v_i in $\boldsymbol{v}$ at the location (i, j, l) is calculated as $\boldsymbol{S}(i, j) \times \boldsymbol{C}(l)$ (see Fig. 2C).

The above voxel attention kernel can be regarded as a self-attention function, which not only enhances the quality of spatial encoding by attending to active spatial locations in the feature map $\boldsymbol{f}$, but also recalibrates the channel-wise feature responses adaptively by capturing the inter-dependencies between channels. Compared to the harmonious attention (HA) [17], the voxel attention has a few differences: 1) it has a much simpler form with a much smaller number of trainable parameters; 2) it is only applied after the backbone model, while HA is inserted between several building blocks; 3) it includes a normalization operation in the spatial attention to highlight the informative spatial locations. It turns out that these differences contribute to improve the model performances significantly (see Sec. 4.3,4.7).

3.2 Two-stage clustering (TC)

We now introduce the two-stage clustering strategy. The choice of the distance metric is crucial for clustering. Although an off-the-shelf clustering algorithm operating in the feature space, rather than in the raw pixel space, can alleviate the problem of "curse of dimensionality" [29] to some extent, it may still lead to an unsatisfactory clustering quality. Here we adopt a two-stage procedure to improve the clustering performance. Firstly, we use the conventional kmeans++ to get the centroids of clusters, denoted as $\{c_m\}_{m=1}^{M}$, with M the predefined number of clusters. Secondly, we re-assign data points to each cluster according to their k-reciprocal Jaccard distances (KRJDs) [45] to the cluster centroids. The k-reciprocal nearest neighbours of a feature point are defined as,

$$R(\boldsymbol{g}, k) = \{\boldsymbol{g_j} \; |(\boldsymbol{g_j} \in N(\boldsymbol{g}, k)) \cap (\boldsymbol{g} \in N(\boldsymbol{g_j}, k))\}, \tag{4}$$

where $\boldsymbol{g}$ is a feature point for clustering, which is obtained by performing max-pooling and 1-D batch normalization on the re-weighted attention feature map $\boldsymbol{a}$. $N(\boldsymbol{g}, k)$ denotes the k nearest neighbours of $\boldsymbol{g}$. $R(\boldsymbol{g}, k)$ indicates that $\boldsymbol{g}$ and each element in its neighbourhood are the mutually k nearest neighbours of each other. The KRJD distance between two feature points is then defined as

$$\boldsymbol{J}(\boldsymbol{g_i}, \boldsymbol{g_j}) = 1 - \frac{|R(\boldsymbol{g_i}, k) \cap R(\boldsymbol{g_j}, k)|}{|R(\boldsymbol{g_i}, k) \cup R(\boldsymbol{g_j}, k)|}. \tag{5}$$

Compared to Euclidean distance, KRJD takes into account the reciprocal relationship between data points, and is a stricter rule measuring whether two feature points match or not (see Fig. 5 and more examples in SI.6). KRJD can also be seen as a refinement of the k-nearest neighbour in the Euclidean space which is more accurate for sorting feature points. Then we obtain a refined cluster $\mathcal{C}_m^p$ by selecting the top p closest feature points to c_m with the KRJD metric. Some of the refined clusters may share some data points due to noises or variances of input images, especially when feature points are intertwined with each other at the first few rounds of training. To alleviate this problem, we remove data points having ambiguous pseudo-labels, and obtain the final pseudo-labeled training set $\{(x_j, y_j)\}_{j=1}^{N_r}, y_j \in [1, 2, ..., M]$, where N_r is the number of remaining data.

3.3 Progressive Training

In our model, the voxel attention (in combination with model training and feature extraction) and two-stage clustering (generating pseudo-labels) are performed iteratively. At each training round t, we optimize the model parameters using the pseudo-labelled train set. When choosing the loss function, we note that the clustering assignments of two adjacent training rounds can be completely different, even if the same set of training samples are used. We therefore adopt the metric learning loss, rather than the softmax loss, as the latter will lead to the failure of model learning. In other words, we only impose that the difference of (dis-)similarities between the positive and negative pairs with respect to a query is larger than a predefined margin, such that the absolute values of assignments are irrelevant. Specifically, we adopt the triplet loss with in-batch hard example mining [10] to optimize the model parameters, which is written as

$$\begin{aligned} &L_{tri}^m\left(\boldsymbol{g}, \boldsymbol{g}^+, \boldsymbol{g}^-; \boldsymbol{\theta}\right) = \max(0, \| \boldsymbol{g} - \boldsymbol{g}^+ \|_2^2 - \| \boldsymbol{g} - \boldsymbol{g}^- \|_2^2 + m), \\ &\text{where} \quad \boldsymbol{g}^+ = \arg\max_{\{\boldsymbol{g}^p\}} \|\boldsymbol{g} - \boldsymbol{g}^p\|_2^2, \text{and} \quad \boldsymbol{g}^- = \arg\min_{\{\boldsymbol{g}^n\}} \|\boldsymbol{g} - \boldsymbol{g}^n\|_2^2. \end{aligned} \tag{6}$$

Here $\{\boldsymbol{g}^p\}$ and $\{\boldsymbol{g}^n\}$ denote the positive and negative sets with respect to $\boldsymbol{g}$ in the mini-batch, respectively, m is the margin between feature pairs, $\boldsymbol{\theta}$ denotes the model parameters. In order to avoid overfitting on the current pseudo labeled set, we only train $\mathcal{M}^t$ in each round for a few gradient update steps to get $\mathcal{M}^{t+1}$. $\mathcal{M}^T$ denotes the final model when the stopping criterion is reached. The two steps of attention-driven clustering and feature learning are performed iteratively, and they facilitate each other to achieve the final well-performing model. The detail of our method ADTC is summarized in Algorithm 1.

4 Experiments

4.1 Datasets

Market-1501 is a dataset containing 32668 images with 1501 identities captured from 6 cameras [44]. The dataset is split into three parts: 12936 images with 751

Algorithm 1 Attention-driven Two-stage Clustering (ADTC) method for unsupervised person re-id

```
Input: The unlabeled dataset $X$, the model $\mathcal{M}^0$.
Output: Final model $\mathcal{M}^T$.
t=0.
repeat
    Attention Step:
    Extracting feature point $\boldsymbol{f_i}$ of each data point $x_i \in X$ before the global max-pooling layer.
    Applying the voxel attention kernel $\boldsymbol{v_i}$ on $\boldsymbol{f_i}$ to get the attention feature point $\boldsymbol{a_i}$.
    Applying global max-pooling and 1-D batch normalization on $\boldsymbol{a_i}$ to get the final feature point $\boldsymbol{g_i}$.
    Clustering Step:
    Performing kmeans++ clustering on $\{\boldsymbol{g_i}\}_{i=1}^{N}$ and obtaining centroids $\{c_m\}_{m=1}^{M}$.
    For each centroid $c_m$, computing its $p$-nearest neighbours $\mathcal{C}_m^p$ based on the KRJD metric, and assigning the pseudo-label $m$ to all data points in $\mathcal{C}_m^p$.
    Removing ambiguous data points belonging to more than one clusters and obtaining the pseudo-labelled train set $\{(x_j, y_j)\}_{j=1}^{N_r}$.
    Parameter Updating Step:
    Training $\mathcal{M}^t$ with the triplet loss on $\{(x_j, y_j)\}_{j=1}^{N_r}$ to get $\mathcal{M}^{t+1}$.
    t = t+1;
until $t = T$
```

identities forming the training data, 19732 images with 750 identities forming the testing gallery, and another 3368 images from the testing gallery forming the query data.

DukeMTMC contains 36411 images with 1812 identities captured from 8 cameras [26]. The dataset is split into three parts: 16522 images with 702 identities forming the training data, 17661 images with 1110 identities forming the testing gallery, and another 2228 images with 702 identities from the testing gallery forming the query data. Note that the evaluation protocol on two dataset are the same.

4.2 Implementation Details

We use a Resnet-50 pretrained on Imagenet as the backbone model. Following [37], we add a batch normalization layer after the global pooling layer to prevent overfitting and directly use the batch-normalized global pooling features to execute identity classification (for the performance of the model architecture on supervised dataset, see SI.4). The output channels are set as 800 in the voxel attention kernel. During clustering, we set the number of clusters M to be 1000 (for the effect of M, see SI.2) and the neighbour size p is 20. All input images are resized to 256×128. Except random horizontal flipping, no other data augmentation strategy is used. 32 pseudo-classes and 4 examples per class are randomly sampled to form a mini-batch. The margin m between negative pairs and positive pairs is 0.3. The total training rounds is set to be 20. To prevent overfitting, the model is fine-tuned for 10 epochs in each round. The Adam optimizer is used for optimization with an initial learning rate of 0.0001 which exponentially decays after epoch 5 (for more detailed setting of hyper-parameters, see SI.1).

4.3 Model Performances on Benchmark Datasets

We compare our model with other state-of-the-art unsupervised person re-id methods on two benchmark datasets Market1501 and DukeMTMC. These methods include: 1) two hand-crafted features: LOMO [19], BoW [44]; 2) four feature alignment methods, MMFA [20], TJ-AIDL [32], ARN [18], and EANet [12]; 3) four GAN-based domain adaptation methods, IPGAN [22], eSPGAN+LMP [5], CamStyle [47], and HHL [46]; 4) two clustering-based methods, PUL [6] and DAR [28]. Note that when training on Market1501, we first initialize our model on DukeMTMC and vice versa (domain adaptation).

source to target	DukeMTMC to Market1501				Market1501 to DukeMTMC			
	mAP	rank1	rank5	rank10	mAP	rank1	rank5	rank10
Directly transfer	18.8	44.0	62.1	69.4	18.2	34.0	49.1	55.9
LOMO [19]	8.0	27.2	-	-	4.8	12.3	-	-
BoW [44]	14.8	35.8	-	-	8.3	17.2	-	-
MMFA [20]	24.7	45.3	59.8	66.3	27.4	56.7	75.0	81.8
TJ-AIDL [32]	26.5	58.2	74.8	81.1	23.0	44.3	59.6	65.0
ARN [18]	39.4	70.3	80.4	86.3	33.4	60.2	73.9	79.5
EANet [12]	51.6	78.0	-	-	48.0	67.7	-	-
IPGAN [22]	25.6	56.4	76.0	82.5	26.7	46.8	62.0	67.9
eSPGAN+LMP [5]	30.4	52.6	66.3	71.7	31.7	63.6	80.1	86.1
CamStyle [47]	27.4	58.8	78.2	84.3	25.1	48.4	62.5	68.9
HHL [46]	31.4	62.2	78.8	84.0	27.2	46.9	61.0	66.7
PUL [6]	20.1	44.7	59.1	65.6	16.4	30.4	44.5	50.7
DAR [28]	53.7	75.8	89.5	93.2	49.0	68.4	80.1	83.5
SSG [8]	58.3	**80.0**	90.0	92.4	**53.4**	**73.0**	80.6	83.2
ADTC w/o DA	38.8	59.5	71.6	76.9	37.9	59.4	70.0	74.1
ADTC (Ours)	**59.7**	79.3	**90.8**	**94.1**	52.5	71.9	**84.1**	**87.5**

Table 1. Comparison of different unsupervised learning methods. DukeMTMC to MarKet1501 means model initialized on DukeMTMC and trained on Market1501. Market1501 to DukeMTMC means model initialized on Market1501 and trained on DukeMTMC. ADTC w/o DA means we trained our model directly on the unlabeled dataset without initialization on the source domain dataset. Note that the LOMO, BoW and PUL also don't use the source domain data to initialize models.

The results are summarized in Table 1. We observe that: 1) our model achieves 59.7%/79.3% on Market1501 and 52.5%/71.9% on DukeMTMC on the mAP/rank1 accuracy, which is one of the state-of-the-art (SOTA) models. Note that we only initialized the model on the source labeled domain and then trained it without any auxiliary label information in the unlabeled domain; whereas most of the aforementioned methods keep using the auxiliary label information in the source domain during the domain transfer learning. 2) Compared to the feature alignment methods which implicitly make an assumption that the data distributions of the source and target domains are similar, our model learns directly from the unlabeled target dataset and achieves better performances. 3) Compared to the GAN-based models which aim at translating the style of labeled images from the source domain to the target domain, our model achieves better performances even without the voxel attention or two-stage clustering (see Table.2). 4) Although the clustering-based SSG model achieves a slightly better performance on DukeMTMC (mAP/rank1) than ours, they use multi learning branches and the DBSCAN clustering method while our model only consists of only one learning branch and adopts the simple kmeans clustering method. Notably, the main concern in our paper is to enhance the in-camera variability so as to improve the accuracy of unsupervised person ReID model rather than introduce other strategies to boost the performance. Overall, our model achieves the state-of-art performances on the two benchmark datasets. In below, we inspect how different elements of the model contribute to its superior peformances.

4.4 Contribution of the Voxel Attention

Fig. 3A&B present the class activation maps (CAMs) [48] of a few example images, which display the spatial regions where the model pays attention to. We see that without the voxel attention, the model pays more attention to the background than to the foreground, resulting in wrong cluster assignments. Indeed, such a degenerate performance often occurs in a clustering-based method without attention, since pedestrian images extracted from the same camera, especially those from the same location, tend to have less variability than those from different cameras (also see Fig. 1). Consequently, the model will assign clusters based on the overall image appearances, rather than the details of pedestrians, and thus fail to capture the in-camera variability of images crucial for the re-id task. Fig. 3A&B also show that the voxel attention helps to increase the margin of the negative pair $(\boldsymbol{g}, \boldsymbol{g}^-)$ and decrease the margin of the positive pair $(\boldsymbol{g}, \boldsymbol{g}^+)$ in a triplet. We calculate the margin difference $\delta = \|\boldsymbol{g} - \boldsymbol{g}^-\|_2^2 - \|\boldsymbol{g} - \boldsymbol{g}^+\|_2^2$ of 10000 triplets randomly sampled from DukeMTMC, and find that by applying the voxel attention, δ increases significantly across the whole dataset (Fig. 3C). This implies that the images belonging to the same identity have a more compact aggregation in the feature space, which makes the retrieval task easier than that without the voxel attention (see SI.5).

To further unveil the role of the voxel attention, we differentiate the wrongly retrieved rank1 images to a query into the in-camera errors (ICE), i.e., those in the same camera as the query, and the cross-camera errors (CCE), i.e., those in different cameras with the query. Fig. 4 compares the results of our model with that of the progressive clustering method without attention. It shows that

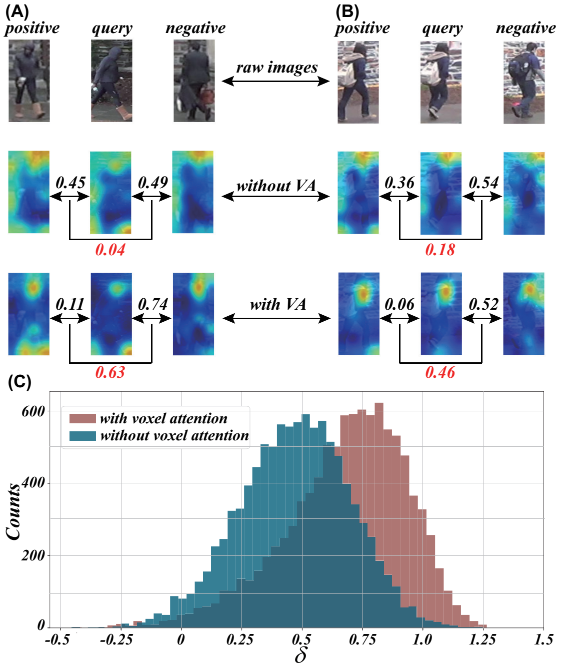

Fig. 3. The voxel attention highlights the informative parts of images and makes them more discriminatable. (A-B) Two examples from DukeMTMC with/without the voxel attention. From top to bottom are the raw images, the CAMs without the voxel attention, and the CAMs with the voxel attention. The value in black stands for the euclidean distance between two feature maps, and the value in red for the margin difference defined in Sec. 4.4. (C) The statistical result of the margin difference δ from 10000 triplets randomly sampled from DukeMTMC.

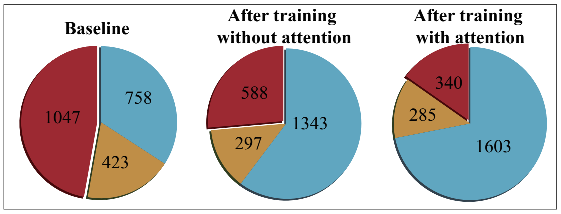

Fig. 4. The voxel attention enhances the in-camera discrimination. From left to right are the results of the initialized model without training (baseline), the model with progressive clustering but no attention, and the model with progressive clustering and attention. The total number of query images is 2228. Blue, red, and orange: the number of query images having the correct rank1, the number of in-camera error (ICE), and the number of cross-camera errors (CCE). DukeMTMC is used.

without attention, the progressive clustering method can improve the rank1 accuracy from 34.0% to 60.3% compared to the baseline (i.e., the result of the model initialized via the label data); but our model can improve the rank1 accuracy further to 71.9%. Notably, this further improvement is mainly attributed to the decrease of ICE, from 588 to 340 out of 2228 queries. This supports our idea that the voxel attention helps to capture the in-camera variability of images; whereas the progressive clustering method without attention is lack of this capability and hence makes more mistakes in-camera identifications.

4.5 Contribution of Two-Stage Clustering

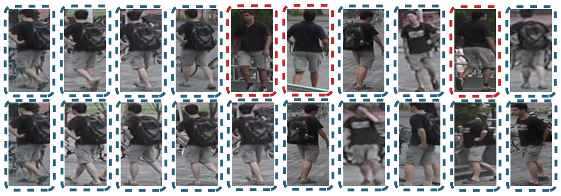

Fig. 5. Example clusters with top 10 nearest neighbours after training with/without two-stage clustering. Market1501 is used. Upper: ranking by the Euclidean distance to the cluster centroid. Lower: ranking by KRJD to the cluster centroid with two-stage clustering. Blue, Red: the correctly, the wrongly assigned images.

We continue to inspect the contribution of two-stage clustering. Fig. 5 shows that when two-stage clustering is used during training, more positive (correct) examples appear in the neighbourhood of a given cluster centroid, compared to that of using only the Euclidean distance based Kmeans++ algorithm. This indicates that KRJD indeed serve as a better metric to compute the neighbourhood relationship between feature points, which improves the clustering quality and leverage the model performances (see SI.6 for more examples).

4.6 Contribution of Progressive Training

We further inspect how the voxel attention and two-stage clustering are executed iteratively to generate good feature representations. To measure the clustering quality, we adopt the normalized mutual information (NMI), which is given by

$$NMI\left(\mathcal{C},\mathcal{L}\right)=\frac{I(\mathcal{C},\mathcal{L})}{\sqrt{H(\mathcal{C})H(\mathcal{L})}}, \tag{7}$$

where $\mathcal{C}=\{\mathcal{C}^p_1,\mathcal{C}^p_2,...,\mathcal{C}^p_M\}$ denote M clusters, $\mathcal{L}$ the corresponding ground truth label set, and I the mutual information between $\mathcal{C}$ and $\mathcal{L}$. $H(\mathcal{C})$ and $H(\mathcal{L})$ are the entropies of $\mathcal{C}$ and $\mathcal{L}$, respectively. The value of NMI is between 0 and 1,

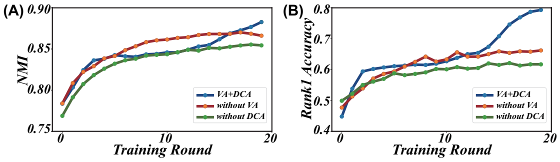

Fig. 6. (A) The clustering performance NMI vs. the training round. (B) The rank1 accuracy vs. the training round.

with 1 standing for the perfect labeling of data points. The larger the NMI, the closer the pseudo-labels to the ground truth[5]. Fig. 6A shows how the clustering performance increases along with the training round. Initially, the assignment of clusters is unsatisfactory (NMI $\approx$ 0.77), as data points are intertwined with each other. Along with the training, data points belonging to the same class are gradually grouped together, and the assigned pseudo-clusters become more similar to the ground truth (NMI $\approx$ 0.90). Fig. 6B further shows that the rank1 accuracy of the model increases in the same pace as the clustering performance. This suggests that in our model, data clustering and model training promote each other during progressive training, in the sense that the improved assignments by two-stage clustering will select more reliable samples to facilitate the learning of the voxel attention, which in return will highlight more informative features to further improve cluster assignments.

4.7 Component Analysis of ADTC

Source to Target	DukeMTMC to Market1501				Market1501 to DukeMTMC			
	mAP	rank1	rank5	rank10	mAP	rank1	rank5	rank10
Only TC	41.1	66.2	84.2	88.9	28.2	49.8	68.2	74.2
Only VA	35.5	61.7	74.3	79.1	32.6	52.0	65.3	69.4
TC + channel attention	42.8	68.9	87.1	91.2	30.1	52.2	71.5	78.9
TC + spatial attention	41.3	66.6	85.0	89.2	28.8	50.7	69.1	75.2
TC + HA	50.6	76.2	88.1	92.0	48.9	69.2	81.5	85.1
TC + CABM	55.2	77.3	88.8	93.5	49.1	69.8	82.0	85.9
Full model	**59.7**	**79.3**	**90.8**	**94.1**	**52.5**	**71.9**	**84.1**	**87.5**

Table 2. Component analysis of the performances of our model. Except the ablating part, all other hyper-parameters are fixed.

We carry out component analysis of our method. Table 2 shows that both the voxel attention and two-stage clustering are indispensable to our model, in the

[5] Note that NMI is independent of the absolute values of labels, in term of that a permutation of cluster labels does not change its value.

sense that when either of them is ablated, the model performance is degraded. Moreover, we check that for the voxel attention, both the channel attention and the spatial attention are indispensable, in the sense that when either of them is ablated, the model performance is degraded. We also replace the proposed voxel attention module with the Harmonious Attention (HA) kernel [17] and the CABM attention kernel [35] (Table 2). It shows that the proposed attention kernel is superior and leads to better performance under the unsupervised setting. Besides, we also carry out robustness analysis of our model to hyper-parameters, e.g., the number of clusters, the margin m the updating epochs in each training round (see SI.2) and the balance level of the original dataset (SI.3). All these results indicate that our model is potentially feasible in real-world applications.

5 Conclusion

In this study, we have proposed an Attention-Driven Two-stage Clustering (ADTC) method for learning an unsupervised model for person re-id. It captures the in-camera variability of images and reduce the noisy labels when clustering(which has been ignored in current unsupervised ReID methods). The method has two indispensable components. Firstly, we use the voxel attention strategy to highlight the informative parts of pedestrian images, which captures the in-camera variability of images crucial for the re-id task. Secondly, we adopts a two-stage clustering strategy, which uses the KRJD metric to improve the clustering quality and stabilizes the progressive training. Through progressive training, the two strategies facilitate with each and enables our model to outperform other unsupervised approaches for person re-ID and achieve the state-of-the-art performances on two benchmark datasets. We also empirically show that our model is robust to a number of varying conditions, making it potentially feasible in real-world applications.

Acknowledgments

ZLJ designed the study and carried out the experiments. XLZ, XHL and XL helped with integrating algorithms and conducting experiments. TJH and SW contributed to the conception and design of the study and revision. ZLJ and SW wrote the manuscript. This work was supported by Huawei Technology Co., Ltd(YBN2019105137) and Guangdong Province with grant (No. 2018B030338001, SW). This work was also supported by BMSTC(Beijing municipal science and technology commission) with grant (No. Z161100000216143, SW) and the National Natural Science Foundation of China (No. 61425025, T.J. Huang).

References

1. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027–1035. Society for Industrial and Applied Mathematics (2007)
2. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person reidentification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
3. Chen, G., Lin, C., Ren, L., Lu, J., Zhou, J.: Self-critical attention learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9637–9646 (2019)
4. Dai, Z., Chen, M., Gu, X., Zhu, S., Tan, P.: Batch dropblock network for person re-identification and beyond. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3691–3701 (2019)
5. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 994–1003 (2018)
6. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **14**(4), 83 (2018)
7. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person reidentification by symmetry-driven accumulation of local features. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2360–2367. IEEE (2010)
8. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person reidentification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6112–6121 (2019)
9. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European conference on computer vision. pp. 262–275. Springer (2008)
10. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. arXiv preprint arXiv:1703.07737 (2017)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
12. Huang, H., Yang, W., Chen, X., Zhao, X., Huang, K., Lin, J., Huang, G., Du, D.: Eanet: Enhancing alignment for cross-domain person re-identification. arXiv preprint arXiv:1812.11369 (2018)
13. Kodirov, E., Xiang, T., Gong, S.: Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In: BMVC. vol. 3, p. 8 (2015)
14. Lan, X., Wang, H., Gong, S., Zhu, X.: Deep reinforcement learning attention selection for person re-identification. In: BMVC (2017)
15. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 384–393 (2017)
16. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 369–378 (2018)

17. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2285–2294 (2018)
18. Li, Y.J., Yang, F.E., Liu, Y.C., Yeh, Y.Y., Du, X., Frank Wang, Y.C.: 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. pp. 172–178 (2018)
19. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
20. Lin, S., Li, H., Li, C.T., Kot, A.C.: Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. arXiv preprint arXiv:1807.01440 (2018)
21. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8738–8745 (2019)
22. Liu, J.: Identity preserving generative adversarial network for cross-domain person re-identification. arXiv preprint arXiv:1811.11510 (2018)
23. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: Proceedings of the IEEE international conference on computer vision. pp. 350–359 (2017)
24. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
25. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1306–1315 (2016)
26. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision. pp. 17–35. Springer (2016)
27. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1179–1188 (2018)
28. Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X.: Unsupervised domain adaptive re-identification: Theory and practice. arXiv preprint arXiv:1807.11334 (2018)
29. Steinbach, M., Ertöz, L., Kumar, V.: The challenges of clustering high dimensional data. In: New directions in statistical physics, pp. 273–309. Springer (2004)
30. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3960–3969 (2017)
31. Wang, H., Fan, Y., Wang, Z., Jiao, L., Schiele, B.: Parameter-free spatial attention network for person re-identification. arXiv preprint arXiv:1811.12150 (2018)
32. Wang, J., Zhu, X., Gong, S., Li, W.: 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 (2018)
33. Wei, L., Zhang, S., Gao, W., Tian, Q.: 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 (2018)

34. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: Glad: Global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 420–428. ACM (2017)
35. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
36. Xia, B.N., Gong, Y., Zhang, Y., Poellabauer, C.: Second-order non-local attention networks for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3760–3769 (2019)
37. Xiong, F., Xiao, Y., Cao, Z., Gong, K., Fang, Z., Zhou, J.T.: Towards good practices on building effective cnn baseline model for person re-identification. arXiv preprint arXiv:1807.11042 (2018)
38. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2119–2128 (2018)
39. Yang, F., Li, K., Zhong, Z., Luo, Z., Sun, X., Cheng, H., Guo, X., Huang, F., Ji, R., Li, S.: Asymmetric co-teaching for unsupervised cross-domain person re-identification. In: AAAI. pp. 12597–12604 (2020)
40. Zhang, X., Cao, J., Shen, C., You, M.: Self-training with progressive augmentation for unsupervised cross-domain person re-identification. arXiv preprint arXiv:1907.13315 (2019)
41. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: 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. pp. 1077–1085 (2017)
42. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3586–3593 (2013)
43. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose invariant embedding for deep person re-identification. IEEE Transactions on Image Processing (2019)
44. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: The IEEE International Conference on Computer Vision (ICCV) (December 2015)
45. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1318–1327 (2017)
46. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–188 (2018)
47. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camstyle: A novel data augmentation method for person re-identification. IEEE Transactions on Image Processing **28**(3), 1176–1190 (2018)
48. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
49. Zhou, S., Wang, F., Huang, Z., Wang, J.: Discriminative feature learning with consistent attention regularization for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8040–8049 (2019)