Pose-Transformation and Radial Distance Clustering for Unsupervised Person Re-identification

Siddharth Seth
ssiddharth@iisc.ac.in

Akash Sonth
akashsonth@vt.edu

Anirban Chakraborty
anirban@iisc.ac.in

Visual Computing Lab,
Department of Computational and Data Sciences,
Indian Institute of Science,
Bangalore, India

Abstract

Person re-identification (re-ID) aims to tackle the problem of matching identities across non-overlapping cameras. Supervised approaches require identity information that may be difficult to obtain and are inherently biased towards the dataset they are trained on, making them unscalable across domains. To overcome these challenges, we propose an unsupervised approach to the person re-ID setup. Having zero knowledge of true labels, our proposed method enhances the discriminating ability of the learned features via a novel two-stage training strategy. The first stage involves training a deep network on an expertly designed pose-transformed dataset obtained by generating multiple perturbations for each original image in the pose space. Next, the network learns to map similar features closer in the feature space using the proposed discriminative clustering algorithm. We introduce a novel radial distance loss, that attends to the fundamental aspects of feature learning - compact clusters with low intra-cluster and high inter-cluster variation. Extensive experiments on several large-scale re-ID datasets demonstrate the superiority of our method compared to state-of-the-art approaches.

1 Introduction

Person re-identification deals with matching the identity of a query image from a dataset consisting of person images taken across a disjoint set of cameras. There has been considerable research on this problem using supervised [3, 11, 36, 51, 55, 57], one-shot [45, 47], transfer learning [5, 9, 52], unsupervised domain adaptation [12, 17, 18, 22, 49, 50, 56], and unsupervised learning [6, 23, 24, 25, 38, 44] frameworks. Supervised approaches require a labeled dataset having both identity and camera information, making them inherently biased towards the dataset they are trained on and are therefore not generalizable. The unsupervised counterpart does not assume the availability of any such labeled information making it more deployment friendly in real-world scenarios.

Unsupervised domain adaptation (UDA) methods [12, 17, 18, 22, 49, 50, 56] tackle the problem of domain generalizability by training on some available labeled source dataset and

then adapting to the target unlabeled dataset. Though such methods have shown to yield state-of-the-art results, they raise two important concerns: 1) labeled dataset is expensive to obtain, restricting the scalability to real-world environments, and 2) there is no way to ascertain the domain gap between the source and target datasets. Thus, one cannot ascertain the ideal dataset that will yield the most superior domain transfer on target. *Purely unsupervised* methods [6, 23, 24, 25, 38, 44] differ from UDA techniques as they delve one step further into the complexity of the problem by being completely unaware of any identity information, either in source or target dataset. In this work, we propose a novel discriminative clustering approach for the purely unsupervised re-ID setup.

Learning without labels is challenging for any problem in general. For the re-ID task, this involves learning discriminative features from the unlabeled dataset to identify query images in the gallery set. Initially, we create pseudo labels by assigning each image its own unique label. However, a simple re-ID training using pseudo labels as the supervisory signal fails to learn any useful features. We thus seek to enhance the network's discriminative ability using a two-stage training procedure. We first take each image and generate similar looking, pose-transformed images in the identity space. These are then assigned labels corresponding to their original images. This helps project the image identity into different poses, thereby yielding a set of more realistic augmentations, thus giving the model an insight into different camera angles inherently present in the dataset. The base CNN is then trained on these samples using the aforementioned pseudo labels as supervision. This is done in conjunction with a metric learning loss to learn a latent space where data points belonging to different identities are farther apart than those belonging to the same identity.

The second stage builds on the first stage by using the proposed discriminative clustering algorithm. Plugging out the pose-transformed samples from the dataset, this step only lets the network access the original unlabeled dataset. This not only eases the burden of training with a large dataset but also helps in learning true distribution features. Our novel loss function aims to explicitly move points in different clusters farther apart and the points inside a cluster within a certain distance, thereby increasing the inter-cluster variation and decreasing the intra-cluster variation. This helps the network achieve powerful discriminability.

Our contributions are summarized as follows:

- We design a pose-transformed dataset that does not require any identity information, is non-parametric, and can be adopted to any person re-identification method.

- We introduce a novel *radial distance* loss that explicitly attends to inter and intra cluster variations by maintaining a minimum *radial distance* between clusters while pushing dissimilar samples away. We also show that it helps discerning similar looking identities.

- We achieve state-of-the-art results for the unsupervised person re-identification task on the Market-1501 [53], DukeMTMC-reID [54], and MSMT17 [42] datasets.

2   Related Works

Traditional approaches using hand-crafted features [10, 23, 26, 28, 29], dictionary learning [19], and saliency analysis [39, 52] did not implicitly extract relevant features invariant to multiple factors of variations (viewpoint, illumination, background clutter).

Unsupervised Person re-ID. Unlike traditional approaches, deep methods [6, 24, 25, 58] instead give the network the freedom to learn discriminative features. [57] use camera-based batch normalization that disassembles re-ID datasets and aligns each camera independently. [24] propose a cross-camera encouragement term that relieves the issue of camera variance. However, intra-camera identity or camera IDs knowledge is essential for such approaches, which may not always be feasible. [58] use multi-label classification for label prediction and a memory bank to store the updated image features. [40] assign camera-aware proxies for dealing with intra-ID variance, thus assuming camera ID information. [44] use tracklet data thereby assuming the availability of temporal information. In light of the above observation, we formulate the solution to the unsupervised person re-ID problem in the simplest form with no assumption of any form of identity related supervision.

Generative approaches. With the emerging popularity of generative adversarial networks, several recent works [8, 11, 16, 27, 33, 54, 55, 58] adopt a procedure to augment the dataset by generating realistic looking training samples. The generative module in [55] generates high quality cross-id images that are used to augment the dataset online. [27] disentangle pose and appearance information using a labeled source dataset and adapt to target dataset. However, all the above methods are either supervised in the target domain or in the source domain for UDA approaches and are therefore unsuitable in an unsupervised person re-ID scenario. On the contrary, we present a simple method for generating image-level samples by spatially transforming the images without leveraging any image identity information, making the approach completely unsupervised.

Clustering-based methods. Clustering algorithms have been used to identify and group similar features together for better discriminability. [6, 25, 58] initially assign each image to belong to its own cluster and then progressively merge clusters. [25] use bottom up clustering to automatically merge similar images into one cluster. However, they make a strong assumption that images are evenly distributed between the identities which will almost never be true for any dataset, and will fail to work when deployed to a real world scenario. [6] explicitly take care of cluster compactness and inter-cluster separability using dispersion based clustering by measuring the distance between all members between two clusters. [58] propose to use multilabel classification to identify similarity between images. The method inherently optimizes the inter- and intra- class distances. However, simply measuring the distance between two points does not properly ensure cluster separability and compactness.

Unsupervised learning. There are several works that focus on learning discriminative features. Center loss [43], for instance, tries to make every cluster compact by projecting the cluster features at the center. DeepCluster [10] and SwAV [2] have been shown to be effective for image classification tasks where both shape and appearance of an object play a role, such as in large-scale datasets like ImageNet [4]. Recently, [14] highlight that reducing bias of deep networks on texture helps improve ImageNet accuracy. However, learning discriminative features for person re-ID requires the model to discern between different identities of humans that frequently appear same due to similar apparel and/or occlusion. Though such unsupervised techniques still work well for re-ID, they often do not explicitly tackle the problem of inter and intra cluster variation. Those that do [6, 58, 43], fail to address the challenging problem of distinguishing between similar looking identities.

Different from such works, we use clustering with the proposed *radial distance* loss that addresses the twin goals of 1) maintaining high inter and low intra cluster variation, and 2) discerning similar appearing identities.
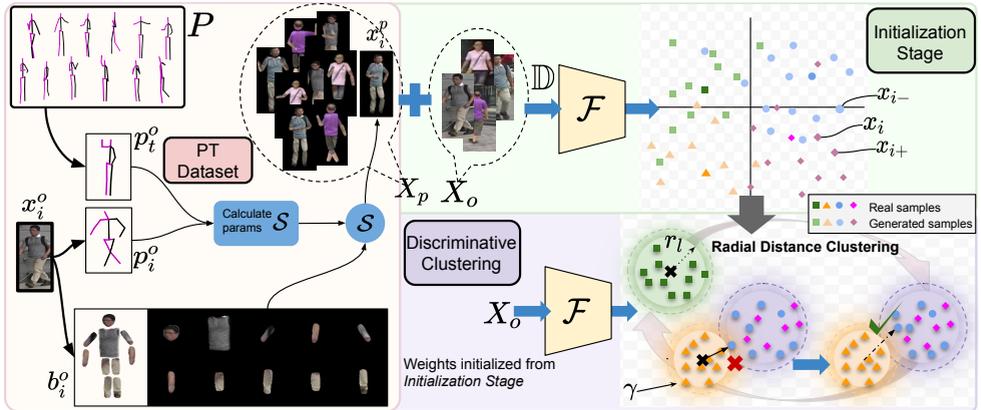
Figure 1: Pipeline of the proposed person re-id framework. PT Dataset: Poses and parts corresponding to the original dataset are obtained. Parts are transformed spatially as per the sampled poses, p_t^o to generate images, x_i^p. PT dataset = original dataset + generated samples is used to train the initialization stage. Initialization Stage: Trained using triplet and classification losses. Discriminative Clustering: Weights for $\mathcal{F}$ are taken from the Initialization Stage. $\mathcal{F}$ is trained only on the original dataset using triplet, classification losses and the proposed radial distance loss. γ represents the minimum radial distance. The blue point violates γ for Orange cluster, and thus, is pushed away.

3 Approach

In this section, we describe our proposed unsupervised learning framework for person re-ID. Let the training set be denoted as $X_o = \{x_i^o\}_{i=1}^N$, where $x_i^o \in \mathbb{R}^{H \times W \times 3}$ with no associated labels. Our goal is to learn a feature representation using a network $\mathcal{F}$ on this unlabeled dataset. Initially each image is assigned its own unique label. The set of pseudo labels are then defined as $Y = \{y_i^o\}_{i=1}^N$ where $y_i^o \in \{1, 2, .., N\}$. We create an auxiliary dataset $X_p = \{x_j^p\}_{j=1}^{K \times N}$, where K is a constant integer scaling factor. The set X_p is constructed via K different pose transformations applied on each image in X_o, where K is a hyperparameter. For each x_i^o, the corresponding K pose-transformed images are assigned the label y_i^o. The pseudo labeled Pose-Transformed (PT) dataset thus becomes $\mathbb{D} = \{X_o \cup X_p, Z\}$, where $Z = \{y_i\}_{i=1}^{(K+1)N}$ and $y_i \in \{1, 2, .., N\}$. For simplicity and future references, we denote the samples in the Pose-Transformed dataset $\mathbb{D}$ as $\{x_i, y_i\}$. The pseudo labels, thus defined, enable us to transform the unsupervised re-ID setting into a supervised one. We initially train a deep network $\mathcal{F}$ on $\mathbb{D}$ in a supervised fashion. However, unlike the trivial augmentation functions, such as flip, tilt, random erase, etc., the pose transformations result in a much larger diversity for the data samples per class, akin to same exhibited by the target images observed across a camera network. As a result, the proposed PT dataset can be utilized to provide a strong supervision to the network and lay the groundwork for a better discriminative training.

Next, we seek to improve the discriminative power of $\mathcal{F}$ by further training it via clustering and the proposed *radial distance* loss. The aim is to pull similar features closer and push dissimilar features further apart, the fundamental goal of any metric learning framework. Evaluation is performed by measuring the Euclidean distance between the extracted features

of query and gallery images using $\mathcal{F}$.

3.1 Pose-Transformed dataset

Pose-Transformed (PT) dataset is modeled by spatially transforming the body parts of a person in a re-ID image to a different random pose. We choose to transform each image into K different poses. To achieve this, we extract the poses, $P = \{p_i^o\}_{i=1}^N \in \mathbb{R}^{15 \times 3}$ corresponding to each person in X_o using the pretrained pose estimation model [20]. The aim now is to transform the person in image x_i^o from pose p_i^o, to a different randomly sampled pose, p_t^o, $t \neq i$ from the set of poses P. We utilize the same set of poses extracted from the target images to ensure that PT dataset adheres to a re-ID setup.

Body part masks. We propose to use two techniques for extracting body parts, represented by $b_i^o \in \mathbb{R}^{10 \times H \times W \times 3}$, 1) using the aforementioned pose estimation method [20], and 2) manually annotating the body parts of a *single* image and utilizing them throughout the dataset. The latter has the advantage of not being restricted to use a method that estimates human poses and body parts simultaneously. Through a quantitative comparison in Sec 4.4, we show that the overall strategy is robust to using any of these techniques.

Parts transformation. The final step performs the alignment of the detected body parts to the randomly sampled target pose, p_t^o. This is done by computing the parameters (rotation, scaling and translation) of the spatial transformation, $\mathcal{S}_t$, required for transforming the detected pose, p_i^o to the target pose, p_t^o. We then get the transformed sample as, $x_i^p = \mathcal{S}_t \circ b_i^o$.

Our aim here is to learn representative features from the dataset for each identity. During spatial transformation, minute details from the face and body accessories may get distorted or removed. But most importantly, we retain the body texture from the original image. This is unlike the deep generative models [30, 41, 55] that attempt to reconstruct a person in different poses. Even though they successfully reconstruct the person, the texture is comparatively different than the original image. We thus seek to generate "a set of perturbations for each original target", that does not focus solely on image quality but on retaining most of the features present in the original image. This is particularly useful for the re-ID task where two different identities may look quite similar. Please refer Supp. for more details.

3.2 Initialization stage

The objective of the initialization stage is to build a groundwork for learning strong discriminative features in the discriminative clustering stage.

Re-identification objective. Using the aforementioned pseudo labeled dataset, $\mathbb{D}$, the network $\mathcal{F}$ optimizes for identity classification using a cross entropy loss,

$$\mathcal{L}_{cls} = \frac{1}{(K+1)N} \sum_{i=1}^{(K+1)N} \log p(y_i | x_i) \tag{1}$$

Metric learning objective. We use the softmax-triplet loss for learning discriminative features early in the training. The loss is defined as,

$$\mathcal{L}_{tri} = \sum_{i=1}^{(K+1)N} \frac{e^{||\mathcal{F}(x_i) . \mathcal{F}(x_{i+}) / \tau||}}{e^{||\mathcal{F}(x_i) . \mathcal{F}(x_{i+}) / \tau||} + e^{||\mathcal{F}(x_i) . \mathcal{F}(x_{i-}) / \tau||}} \tag{2}$$

where, for any target image x_i, x_{i+} is a positive sample taken from the same cluster, and x_{i-} is a negative sample taken from a different cluster (Initialization Stage, Fig. 1). This

step helps the network learn discriminative features for each image, x_i^o. Also, τ denotes temperature parameter and $||.||$ implies L2 distance unless specified.

3.3 Discriminative clustering

Fig. 1 shows the discriminative clustering stage. Although the pose transformed dataset provides a strong supervisory signal, the learned features may not be discriminative enough for accurate re-identification. We next describe the discriminative clustering strategy which strongly boosts the discriminative power of the network.

At this stage, all the pose-transformed samples used in the initialization phase are dropped from the training data, $\mathbb{D}$. This not only reduces the computational overhead while computing clusters during training but also lets the network concentrate on learning features present in the original dataset, X_o. We perform hierarchical clustering [48] to partition the dataset X_o into $\mathcal{C}$ clusters. We also find the centroid, c_l and radius, r_l for each cluster C_l obtained from the clustering algorithm. While a centroid is estimated as the mean of all the features in the cluster, a cluster radius is defined as the Euclidean distance between features of data-sample from the same cluster farthest from its centroid, and that of the centroid itself. Thus, a cluster centroid is computed as $c_l = mean\{\mathcal{F}(x_j^o) \,|\, x_j^o \in C_l\}$ and a cluster radius as $r_l = \max\{||c_l - x_j^o||, \forall x_j^o \in C_l\}$. These estimated cluster parameters are further used in enforcing the *radial distance*, discussed below.

Radial distance loss. At this stage we want to encourage large inter-cluster distance while maintaining high similarity among features of data points belonging to the same cluster. We propose to maintain a minimum radial distance for each image from all but the parent cluster in the space. The *radial distance* between a cluster C_l and a feature from any sample $x_i^o \in C_k (l \neq k)$ is defined as, $d_i^l = ||c_l - \mathcal{F}(x_i^o)||$. To ensure strict separation between any two clusters, this distance must satisfy the following condition: $d_i^l \geq r_l + \gamma$, where γ is a margin that ensures a minimum expected separation in terms of the radial distance.

We define the radial distance loss as

$$\mathcal{L}_{rd} = \sum_{i=1}^{N}\sum_{l \neq k}^{\mathcal{C}} \max(0, r_l + \gamma - d_i^l) \tag{3}$$

where γ is the minimum radial distance or margin that needs to be maintained. Thus, if the point $\mathcal{F}(x_i^o)$ lies inside the minimum radial distance, i.e. breaches the minimum margin, for any cluster C_l, except its own, it is assigned a penalty. Cluster assignments r_l and c_l are calculated once every epoch and assumed to be the same throughout the epoch. This prevents the network from training on adversely changing cluster radii and centroids, and makes the cluster sizes robust to changing features. For instance, the cluster radius, r_l may actually shrink if the corresponding $\mathcal{F}(x_j^o)$ comes closer to the centroid, c_l. However, the minimum radial distance will still prevent any point from coming closer to C_l as r_l is kept constant for the epoch. Similarly, the cluster radius may increase if the corresponding point $\mathcal{F}(x_j^o)$ moves away from c_l. Even though the outlier point may now flout the minimum radial distance, it may become a part of a cluster it is closer to, at the next assignment. Thus, the cluster assignment for the data points to C_l, and the corresponding cluster radius, r_l is updated in the epoch.

On similar lines, triplet loss aims at maximizing the positive sample distance from the negative sample distance, i.e., it works at point level. On the other hand, *sd* loss works at cluster level, where the distance of the centroid (anchor) to the positive sample (radius) does

not change throughout the epoch. All samples therefore inside the radius are considered as belonging to the positive class for one learning epoch. Unlike in triplet loss, a slight movement in the feature space, inside the radius, ensures that cluster assignment does not change for a positive sample. A drastic change signifies an incorrect assignment which is corrected during re-assignment at the next epoch. This makes the algorithm robust to changing cluster assignments and cluster sizes.

Finally, the network is trained using the triplet loss, the classification loss, and the radial distancing loss. The overall loss function is thus given by

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{tri}\mathcal{L}_{tri} + \lambda_{rd}\mathcal{L}_{rd} \tag{4}$$

where λ_{tri} and λ_{sd} are balancing hyperparameters. Triplet loss is now computed using positive samples from the same and negative samples from a different cluster while the number of samples for both classification and triplet loss change to N.

Table 1: Comparison of our proposed method with state-of-the-art unsupervised person re-ID methods on Market-1501, DukeMTMC-reID and MSMT17. * number taken from [25].

Methods	Market-1501				DukeMTMC-reID				MSMT17			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
LOMO [23]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8	-	-	-	-
BOW [53]	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3	-	-	-	-
OIM [46]*	38.0	58.0	66.3	14.0	24.5	38.8	46.0	11.3	-	-	-	-
BUC [25]	61.9	73.5	78.2	29.6	40.4	52.5	58.2	22.1	-	-	-	-
TAUDL [21]	63.7	-	-	41.2	61.7	-	-	43.5	28.4	-	-	12.5
DBC [9]	69.2	83.0	87.8	41.3	51.5	64.6	70.1	30.0	-	-	-	-
TSSL [42]	71.2	-	-	43.3	62.2	-	-	38.5	-	-	-	-
SSL [24]	71.7	83.8	87.4	37.8	52.2	63.5	68.9	28.6	-	-	-	-
HCT [48]	80.0	91.6	95.2	56.4	69.6	83.4	87.4	50.7	-	-	-	-
MMCL [38]	80.3	89.4	92.3	45.5	65.2	75.9	80.0	40.2	-	-	-	-
CycAs [41]	84.8	-	-	64.8	77.9	-	-	60.1	50.1	-	-	26.7
Ge et al. [13]	88.1	95.1	97	73.1	-	-	-	-	42.3	55.6	61.2	19.1
CAP [40]	91.4	96.3	97.7	79.2	81.1	89.3	91.8	67.3	67.4	78.0	81.4	36.9
Ours	**93.6**	**97.2**	**98.3**	**81.6**	**86.3**	**92.4**	**94.7**	**70.1**	**69.9**	**80.3**	**85.4**	**40.7**

4 Experiments

4.1 Datasets and Evaluation Metrics

Experiments are conducted on three large-scale datasets - Market1501 [53], DukeMTMC-ReID and MSMT17 [42] datasets. The Cumulative Matching Characteristic (CMC) curve and mean Average Precision (mAP) evaluation metrics are reported on these datasets.

Market1501 [53] dataset consists of 32,668 images from 1,501 identities captured using 6 cameras. The training set consists of 12,936 images from 751 identities. The gallery and query set consist of 19,732 images and 3,368 images respectively from 750 identities.

DukeMTMC-ReID [34] [45] dataset is a subset of the DukeMTMC [34] tracking dataset. It consists of 36,411 images from 1,404 identities captured using 8 cameras. The dataset is split into 16,522 images from 702 identities for the training set, 17,661 images from 702 identities for the gallery set, and 2,228 images from 702 identities for the query set.

MSMT17 [42] dataset consists of 126,411 images from 4,101 identities captured using 15 cameras (12 outdoor and 3 indoor). The training set consists of 32,621 images from 1,041 identities. The gallery and query set consist of 82,161 and 11,659 images respectively.

4.2   Implementation details

We adopt ResNet-50 [15] pre-trained on ImageNet [4] as the backbone network for a fair comparison with other works. It is then trained on PT dataset for 40 epochs using classification and triplet losses. The Adam optimizer is adopted with a weight decay of 0.0005, and the learning rate is set to 3.5×10^{-4}. For each identity from the generated training set, a mini-batch of 64 is sampled with $P = 16$ randomly selected identities and $M = 4$ randomly sampled images for computing the softmax triplet loss. Next, before each epoch, we perform hierarchical clustering [48] on the extracted features from the initialization stage, followed by the computation of cluster centroids and radii. The labels obtained from the clusters are used to train the model along with the classification, triplet, and radial distancing losses. We use the PyTorch library for implementation and train the model on a single RTX 2080 GPU.

4.3   Comparison with State-of-the-Arts

We compare our proposed approach with prior arts including LOMO [23], BOW [53] that are based on hand-crafted features, TAUDL [21], TSSL [44], and CycAs [41] that use tracklets, among other unsupervised learning methods [7, 13, 24, 25, 58]. Table 1 shows the Rank-1 and mAP results on Market-1501 [53], DukeMTMC-reID [54], and MSMT17 [42] datasets. We achieve state-of-the-art performance on Market1501 dataset with 93.6% R-1 and 81.6% mAP, improving on the prior art [40] by 2.2% in R-1 and 2.4% in mAP. Similarly, for the DukeMTMC-reID dataset, we report 86.3% R-1 and 70.1% mAP, an improvement of 5.2% in R-1 and 2.8% in mAP over [40]. Our method also outperforms the state-of-the-art [40] by **2.5%** in Rank-1 accuracy and **3.8%** in mAP on the more challenging MSMT17 dataset.

Table 2: Ablation study comparing traditional augmentation techniques *AUG* with PT dataset and the importance of radial distance loss ($\mathcal{L}_{rd}$) on Market-1501 dataset.

X_o	AUG	PT	$\mathcal{L}_{sd}$	R-1	mAP
✗	✗	✗	✗	7.8	2.6
✓	✗	✗	✗	28.2	10.5
✓	✓	✗	✗	34.9	14.6
✓	✓	✗	✓	42.2	19.9
✓	✗	✓	✗	58.7	27.7
✓	✗	✓	✓	90.3	76.1
✓	✓	✓	✓	93.6	81.6

Table 3: Ablation study showing the effectiveness of the PT dataset when integrated with recent state-of-the-arts for Market-1501 dataset.

Method	R-1	mAP
MMCL [58]	84.2	51.6
Ge et al. [13]	90.1	77.4
CAP [40]	92.7	80.3

Table 4: Ablation on using intuitive and easy to obtain single image body parts annotation (SIA) vs body parts cum pose estimation (PE [21]) method on Market-1501.

SIA	PE	$\mathcal{L}_{sd}$	R1	mAP
✓	✗	✗	59.1	28.5
✗	✓	✗	58.7	27.7
✓	✗	✓	90.1	76.8
✗	✓	✓	90.3	76.1

4.4   Ablation Study

Effectiveness of PT dataset. We perform an extensive evaluation on Market-1501 dataset under different settings as shown in Table 2. We refer to traditional image-level augmentation techniques such as flip, tilt, jitter, random crop etc. used in re-ID methods as *AUG*. The first row shows direct evaluation on ImageNet pre-trained ResNet50 which fails miserably. Second row shows results for training only on original dataset, X_o. Next, employing *AUG* provides the network a better discriminative ability. Using PT dataset as a replacement to *AUG* shows a significant boost in performance of the re-ID model. This establishes the effectiveness and importance of using highly diverse and non-trivial augmentations which we
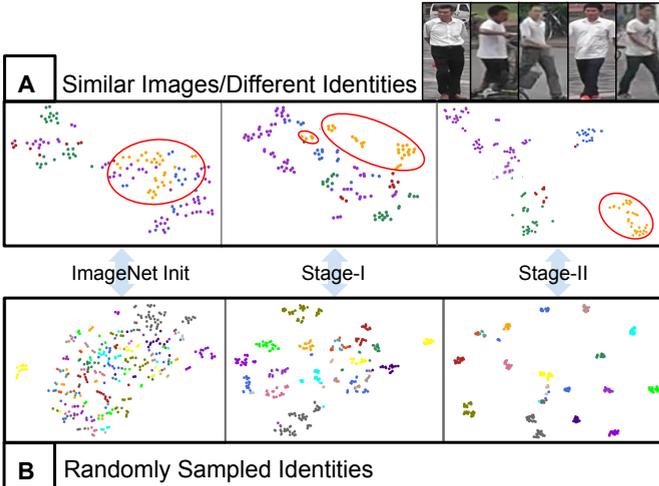
Figure 2: t-SNE visualization for **A.** 5 clusters with similar appearing but different identities, represented by images in the top row. **B.** 20 randomly chosen clusters
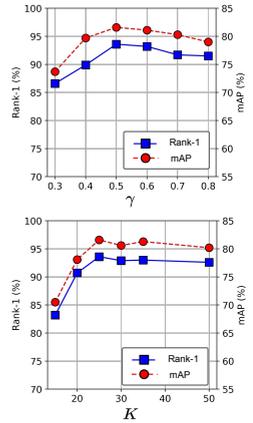

Figure 3: Ablation on the Market-1501 by varying the margin γ, and sample scaling factor K.

perform using pose space transformations. We also demonstrate this in Table 3 by using PT dataset with state-of-the-art methods, reporting an improved performance on the respective original baselines. Lastly, Table 4 compares the performance of our approach when using the two techniques of extracting body parts of a person as described in Sec. 3. One can easily obtain body parts from a single image annotation (SIA) which can then be used for the complete dataset. This method fares favorably to the one used for obtaining both poses and body parts [20] but has the advantage of not requiring a body part estimator. One can then freely choose from any of the abundantly available off-the-shelf pose estimation models.

Effectiveness of radial distance loss. We hypothesize, a well-learned feature space will learn similar features closer together than dissimilar features. However, in some cases, even the similar features may belong to different identities and therefore must be strictly separable. This would enable the method to distinguish between the small yet relevant differences often existing between similar identities. The sd loss, $\mathcal{L}_{sd}$ maintains a strict separability of clusters owing to the cluster radius and margin distances. This ensures that points from neighboring clusters do not overlap, making them well separable. Furthermore, because the penalty outside the minimum distance is zero, this lets features from similar-looking identities reside close by in the feature space. We illustrate this in the next subsection. Table 2, shows that $\mathcal{L}_{sd}$ enables the model to learn powerful discriminative features on both *AUG* and PT datasets, yielding state-of-the-art results on benchmark datasets. We obtain best results (see last row Table 2) when using augmentations along with our PT dataset for our discriminative learning strategy.

Effectiveness of our discriminative learning method. To verify that the proposed method does indeed maximize inter-clusters distances and minimize intra-cluster distances, we plot the t-SNE [37] visualization shown in Fig. 2. Panel *A* shows the plot for 5 different identities that appear similar due to their clothing patterns. The first plot represents the clusters when plotted using ImageNet initialized weights. The highlighted ID in yellow shows that images of the same identity are not separable. The second plot depicts the results from

our initialization stage. The network has started to identify the discriminative features. Images from the highlighted ID are seen to have come together, whereas a small number of images of the same ID have clustered at some distance. Finally, the last plot depicts the clusters obtained after training finishes. The clusters have become more distinguishable showing the high discriminative nature of the learned features. Though the final goal is not to obtain perfect clusters, nevertheless, the model is able to distinguish between different similar looking identities. Similarly, Panel B illustrates results on 20 randomly chosen identities.

Analysis on minimum radial distance margin γ. The minimum radial distance margin ensures that clusters are at least a minimum distance γ from one another. A small margin may lead to images being close in the feature space, and thus being misclassified. A high margin on the other hand may lead to clusters themselves having a large radius. This would mean that images of the same cluster can move far away from each other within the cluster. An optimum margin keeps the clusters compact enough while keeping the adjacent clusters at least preset distance away. Fig. 3 shows the ablation on selecting the best γ value.

Analysis on sample scaling factor K. It may intuitively seem that a high value of K will give better results. Fig. 3 gives the evidence that the performance improvement with increasing K is only up to a certain value and post that the performance stagnates. We provide a simple reason for this. PT dataset is created using randomly sampled poses which are all plausible. However, a pose estimation model may not always be effective against occlusion, missing body parts or background clutter. This leads to inaccurate pose and body part detection. Any random pose transformation applied to this image will become contaminated with source errors, leading to the generation of noisy samples.

5 Conclusion

We present an expertly designed, identity independent pose-transformed dataset. This constitutes spatially transforming images in the pose space for generating a realistic and diverse set of images. We show this can be easily adopted to other person re-ID methods. Furthermore, the novel radial distancing criterion introduced attends to the fundamental goal of feature learning, i.e., low intra-cluster and high inter-cluster variation, thereby learning strong discriminative features as established by extensive experimental analysis.

6 Acknowledgements

References

[1] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[3] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, 2019.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018.

[6] Guodong Ding, Salman Khan, Zhenmin Tang, Jian Zhang, and Fatih Porikli. Dispersion based clustering for unsupervised person re-identification. In *BMVC*, 2019.

[7] Guodong Ding, Salman H Khan, Zhenmin Tang, J Zhang, and F Porikli. Dispersion based clustering for unsupervised person re-identification. In *BMVC*, page 264, 2019.

[8] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. In *NeurIPS*, 2019.

[9] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2018.

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[11] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*, 2018.

[12] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020.

[13] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020.

[14] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Yan Huang, Jingsong Xu, Qiang Wu, Zhedong Zheng, Zhaoxiang Zhang, and Jian Zhang. Multi-pseudo regularized label for generated data in person re-identification. *Transactions on Image Processing*, 2018.

[17] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Sbsgan: Suppression of inter-domain background shift for person re-identification. In *ICCV*, 2019.

[18] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, 2020.

[19] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, 2015.

[20] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[21] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018.

[22] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *ICCV*, 2019.

[23] S. Liao, Y. Hu, Xiangyu Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[24] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian. Unsupervised person re-identification via softened similarity learning. In *CVPR*, 2020.

[25] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019.

[26] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo. Person re-identification by iterative re-weighted sparse ranking. *Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[27] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *CVPR*, 2018.

[28] Bingpeng Ma, Yu Su, and Frederic Jurie. Bicov: A novel image representation for person re-identification and face verification. In *BMVC*, 2012.

[29] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV Workshop*, 2012.

[30] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, 2017.

[31] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018.

[32] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, 2016.

[33] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018.

[34] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCVW*, 2016.

[35] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. *arXiv preprint arXiv:2007.09278*, 2020.

[36] C. Tay, S. Roy, and K. Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, 2019.

[37] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[38] D. Wang and S. Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, 2020.

[39] Hanxiao Wang, Shaogang Gong, and Tao Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014.

[40] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[41] Zhongdao Wang, Jingwei Zhang, Liang Zheng, Yixuan Liu, Yifan Sun, Yali Li, and Shengjin Wang. Cycas: Self-supervised cycle association for learning re-identifiable descriptions. In *ECCV*, 2020.

[42] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[43] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, 2016.

[44] Guile Wu, Xiatian Zhu, and Shaogang Gong. Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI*, 2020.

[45] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018.

[46] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[47] Mang Ye, Xiangyuan Lan, and Pong C. Yuen. Robust anchor embedding for unsupervised video person re-identification in the wild. In *ECCV*, 2018.

[48] Kaiwei Zeng, Munan Ning, Wang Yaohua, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *CVPR*, 2019.

[49] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *CVPR*, 2020.

[50] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. In *ECCV*, 2020.

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

[52] R. Zhao, W. Oyang, and X. Wang. Person re-identification by saliency learning. *Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[53] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[54] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

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

[56] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019.

[57] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *ECCV*, 2020.

[58] Yang Zou, Xiaodong Yang, Zhiding Yu, B. V. K. Vijaya Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*, 2020.