Synthetic18K: Learning Better Representations for Person Re-ID and Attribute Recognition from 1.4 Million Synthetic Images

Onur Can Uner[a], Cem Aslan[b], Burak Ercan[b], Tayfun Ates[b], Ufuk Celikcan[b], Aykut Erdem[c], Erkut Erdem[b,*]

[a] *Department of Computer Engineering, Bilkent University, Ankara, Turkey*
[b] *Department of Computer Engineering, Hacettepe University, Ankara, Turkey*
[c] *Department of Computer Engineering, Koç University, Istanbul, Turkey*

Abstract

Learning robust representations is critical for the success of person re-identification and attribute recognition systems. However, to achieve this, we must use a large dataset of diverse person images as well as annotations of identity labels and/or a set of different attributes. Apart from the obvious concerns about privacy issues, the manual annotation process is both time consuming and too costly. In this paper, we instead propose to use synthetic person images for addressing these difficulties. Specifically, we first introduce Synthetic18K, a large-scale dataset of over 1 million computer generated person images of 18K unique identities with relevant attributes. Moreover, we demonstrate that pretraining of simple deep architectures on Synthetic18K for person re-identification and attribute recognition and then fine-tuning on real data leads to significant improvements in prediction performances, giving results better than or comparable to state-of-the-art models.

Keywords: person re-identification, attribute recognition, synthetic data

*Corresponding author. Email address: erkut@cs.hacettepe.edu.tr

1. Introduction

In developed countries, video surveillance systems have become a vital component of public security, constantly monitoring cameras installed at various different key locations. Person re-identification (person re-ID) is one of the important tasks in video surveillance, which refers to the problem of automatically re-identifying across multiple different cameras with the help of computers. Different from other person-centric classification problems in computer vision such as face recognition, person re-ID systems use identity information only during training and assume the identities are unknown at test time. Person re-ID basically combines three different tasks, which are pedestrian detection, person tracking and person retrieval. Hence, it is extremely challenging due to large variations in lighting conditions, differences in pose and viewpoint changes.

The person re-ID models proposed in the literature can be divided into two main groups: image-based methods and video-based methods. All these approaches are usually evaluated on benchmark datasets annotated with either detected or ground truth human boxes so the problem mostly reduces to the person image retrieval where the aim is to retrieve images of a specific person identity from a large gallery of images involving many different identities. Therefore, success can be defined by comparing the identities of the retrieved images with the identity of the query image. In video-based approaches to person re-ID, the models use multiple bounding boxes for a video query and the videos in the gallery, and additionally try to integrate the temporal information between these boxes. Although research in both groups can benefit from each other, they are considered as different problems. In this work, we tackle the problem of image based re-ID.

As another critical task in video surveillance, person attribute recognition aims at detecting various attributes of a person such as hair color, clothing type and color. Although person attribute recognition has been relatively less studied as compared to person re-ID, these tasks are, in fact, closely related since the mid-level semantic attributes, once identified, provide an intuitive way to describe a specific individual. Hence, a recent direction being explored in recent years is to consider these two challenging tasks in a joint manner in order to improve the performances of each other.

In the past few years, person re-ID research has reached a saturated point where researchers gently enhance the performances on popular benchmark datasets by designing more and more complex architectures or by applying complicated data augmentation schemes. That being said, most of the methods in the literature fail to generalize well to in-the-wild settings because of the fact that existing datasets cover a limited range of samples that could be faced in real life. However, obtaining a comprehensive dataset is very expensive to gather for which you have to use multiple camera sources located in very different environments. Even if you collect the right amount of visual data, the effort required for manual annotation is quite costly. Hence, the existing datasets are not challenging enough in demonstrating different weather and lighting conditions, varieties in person attributes and/or body types. Finally, privacy of the individuals is an important issue for video surveillance. Although these datasets are initially collected for academic purposes, the intention of users may be different when the data become public, leading invasion of privacy of the individual. For example, DukeMTMC-reID dataset [1] has been recently shut down and cannot be downloaded publicly to conform to privacy regulations.

In this study, we deal with the problem of learning simple yet effective representations for the person re-identification and attribute recognition tasks. In particular, in both of these two tasks, the main challenge lies in learning discriminative features which are not sensitive to the changes in the appearance of the person of interest due to illumination variations, scale and viewpoint changes. To overcome these difficulties, in this study, we first introduce a new synthetic dataset called Synthetic18K. The proposed dataset, compared to the existing synthetic datasets mentioned above, is much larger in scale in terms of the number of identities/virtual persons it contains and the number of images that each virtual person has. That is, it contains approximately 1.4 million images of 18K unique virtual persons captured in four synthetic environments (three outdoor and one indoor) as well as with various cubemaps taken in real-life. While generating these virtual persons and obtaining their images, we follow a procedural generation method which gives us the ability to play with both the low and high-level attributes of these synthetic persons and the characteristics of the scenes (weather conditions, times of day, etc.). These aspects are of critical importance for feature learning as the existing real-world data are generally not diverse in various factors like illumination conditions, scenes, clothing, etc. are still considered as challenging tasks. Moreover, covering each one of these factors in a dataset in a balanced manner could be very difficult to achieve, resulting in heavy-tailed data distributions and poor performances for the rare cases. Tackling person re-identification and attribute recognition in a joint manner also introduces certain advantages as these tasks are considered complementary tasks. Yet, no other synthetic datasets handles these two in a combined manner.

Motivated with these, in our work, we also propose pretraining strategies and simple yet effective deep neural architectures for both person re-identification and attribute recognition tasks. we show that our proposed Synthetic18K dataset can be used to learn more robust feature representations for these two tasks. In particular, we proposed three different pretraining schemes, one for solely person re-identification, one for only attribute recognition and one final for a combination of these two tasks. We demonstrate that even simple neural architectures which are pretrained on our synthetically generated images using these strategies and later on fine-tuned on real data, perform competitively compared to complex state-of-the-art models. Our experiments also show handling person re-identification and attribute recognition together gives more accurate results than their single-task counterparts, indicating the importance of the proposed joint pretraining strategy.

Our dataset and models will be available at the project website[1].

2. Related Work

In this section, we briefly review the literature on person re-ID and person attribute recognition, including the methods that perform these two tasks jointly, and mainly focusing on recent deep learning based techniques. Moreover, we look into existing re-ID datasets and their limitations.

2.1. Person Re-Identification

In recent years, person re-ID has emerged as a growing research topic, but its roots can be traced back to multi-camera tracking where the idea is to assign each person a unique latent label and try to re-identify them when they leave the scene and enter again to correctly keep tracking them. Hence, it is usually assumed that people wear the same clothes when they are captured by cameras. The early

[1] `https://hucvl.github.io/synthetic18k`

approaches typically involve dividing the image into multiple image regions and represent each region with some local features such as color histograms and SIFT features and obtain a person representation by the concatenation of these features. One can apply metric learning to further increase the discriminative power of the features. A detailed analysis of these prior works can be found in [2, 3].

With the recent advancements in deep learning and the availability of large scale datasets, the aforementioned shallow models are replaced with their deep counterparts which combine feature learning and metric learning within a single framework. These approaches commonly use a pretrained convolutional neural network (CNN) such as ResNet-50 [4] trained on ImageNet [5] as a backbone and mostly differ from each other in their architectural details and the way their objectives are defined. As for the objectives, the most straightforward approach is to resort to classification loss where each person in the training set is treated as a distinct class [6]. More complicated objectives for Siamese architectures involve pair-wise contrastive loss [7], triplet ranking loss [8, 9] and quadruplet loss [10], or a combination of them [11, 12, 13]. Recent works even go beyond these approaches and consider all the pairwise similarity relations within a batch composed of multiple identities and their images by using graph neural networks [14].

Solving person re-ID task requires features that have some invariance to scale changes [15], illumination and pose variations, viewpoint changes as well as misalignment errors in human detections [2, 3]. Apart from the training objectives mentioned above, some studies directly tackle these issues and design some specialized architectures. Common trends include exploiting pose information by either using off-the-shelf pretrained pose detectors [16], or attaching a pose estimation network to the re-ID network and training them jointly [17]. Other alternatives are learning to localize body parts [18, 19] or employing a human semantic parser network to additionally incorporate the body part features within the person re-ID network [20].

2.2. Person Attribute Recognition

Compared to person re-ID, person attribute recognition is a relatively less studied topic, but it has also gained attraction due to widespread interest in automated visual surveillance systems. Traditional approaches typically employ hand-crafted features such as color histograms and involve classifiers trained independently for each attribute [21, 22]. However, attribute recognition is inherently a multi-label classification problem. Moreover, there are dependencies between some attributes. For example, there is a high probability for a woman wearing high heels to carry her bag in her left or right arm. A way to incorporate this knowledge and improve the prediction performance is to use a graphical model, *e.g.* a conditional Markov random field [23].

In one of the early deep learning based approaches to attribute recognition [24], the authors trained a single CNN model which considers the dependencies between the attributes during training. In particular, the network shares most of its parameters among each attribute classifier and trained based on a KL-divergence based loss. Similarly, in another work, Zhu et al. [25] employ a multi-label loss function but their architecture is a multi-stream CNN model which takes multiple overlapping image regions extracted from the original input person image. In [26], Wang et al. follow a different strategy and employ a recurrent CNN model which sequentially outputs the attribute predictions. Moreover, they utilize similar images in the dataset during training in order to alleviate the issues regarding background clutter and uncontrolled viewing conditions.

2.3. Joint Person Re-ID and Attribute Recognition

In literature, some researchers have also explored how person attribute recognition and person re-ID tasks can promote each other via a joint learning strategy. The idea dates back to [27], where the authors use extracted attributes as additional, mid-level semantic features to improve re-ID performance. The approach is based on training SVM-based attribute classifiers and then applying a greedy strategy to decide the optimum weights of the attributes for re-ID. The follow-up studies, however, approach the problem from a multi-task learning perspective that leverage attribute and identity information to train a single unified model [28, 29].

As for the examples of deep learning based approaches, in [30], Su et al. proposed a semi-supervised strategy, which involves training attribute classifiers on an attribute dataset and exploiting them to extend the annotations of a person re-ID dataset with person-specific attributes. Then, the person re-ID model is trained with a triplet loss defined on top of these attributes, assuming that predicted attribute labels should be similar for the same person. In [31], Lin et al. propose a multi-task learning framework which includes a shared CNN-based encoder and two task-specific branches, one for attribute prediction and another for re-ID. The re-weighted attribute predictions are concatenated to CNN features to incorporate semantic knowledge into person re-ID. In [32], Sun et al. present a deep person re-ID model which incorporates body parts and pose information as well as a secondary attribute classification to improve the discriminative power of the learned features. Liu et al. [33] employ connectionist temporal classification (CTC) loss and self-attention to jointly learn attribute recognition and re-ID. Tay et al. [34] propose a unified architecture that combines attribute features and attribute attention maps with identity and body part classification. In another recent study, Wang et al. [35] suggest to learn and use hidden attributes other than the provided ones in an unsupervised manner to boost the re-identification performance.

2.4. Synthetic Data for Person Re-ID

Several researchers have recently explored the idea of using synthetic data to address various issues in person re-ID. In [36], Barbosa et al. used MakeHuman, a 3D character creation software, to generate 25 male and 25 female bodies which have 8 different outfit types to obtain a synthetic dataset, which they refer to SOMAset. The authors show that this dataset can be used to alleviate a main drawback of real-world datasets that they heavily rely on appearances of clothes, but not much to structural aspects of the human body. In another study [37], Sun et al. focused on issues related to viewpoint changes in re-ID datasets and developed PersonX, a data generation framework which renders hand-crafted clothed human meshes onto several backgrounds in different lighting levels to better understand the role of viewpoint. Lastly, in [38], Bak et al. aimed to address the lack of illumination variances in real re-ID datasets. They rendered 100 different virtual humans in multiple HDR environment maps to simulate different lighting conditions to create the SyRI dataset. They also proposed a domain adaptation technique which makes use of this synthetic data. Xiang et al. [39] proposed another synthetic dataset called GPR for person re-identification which consists of 754 identities and around 440K bounding boxes by using the GTA5 computer game. The identities span a diverse set of persons with different gender, appearance, nationalities, etc. Then, they suggested a domain adaptation technique for unsupervised person re-identification that depends on these synthetically generated person images. Concurrent to our work, Wang et al. [40] proposed to use Unity3D engine to generate a large-scale synthetic dataset called RandPerson that is composed of images from 8K different virtual persons with different races and attributes. In particular, they automatize the clothing of these synthetic persons by generating and using a large number of random UV texture maps. In another recent study, Zeng et al. [41] addressed changes in the illumination conditions by constructing two synthetic datasets containing images with a wide range of illumination variations. These simulated images, however, were not generated by rendering synthetic images but obtained by applying random gamma adjustments to real images.

Our proposed Synthetic18K dataset, while sharing some features of earlier works, departs from them in the manner that it uses a framework which procedurally generates synthetic persons; resulting in a signicantly higher count of unique persons (18K), covering a much more diverse range of looks in terms of body types, clothes, accessories, skin tones and facial features. And since the persons are synthetically generated, we can also provide semantic annotations about them, as well, which can be used to create similarity metrics in the given set of persons. Moreover, while the aforementioned datasets contribute mostly simple illumination changes, Synthetic18K features images of each person at different environments from numerous viewpoints in varying real-life -like environment conditions using a completely procedural atmosphere and weather rendering system.

3. Synthetic18K Dataset

The Synthetic18K dataset is a collection of 1,408,600 synthetically generated images of 18,306 unique persons in varying environmental conditions. The dataset is built for person re-ID and attribute recognition purposes; hence with each image, several annotations are also provided (Table 1). The dataset was generated by re-purposing our procedural generation framework, which is built on Unity graphics engine, specifically for the tasks at hand. Generation process is completely automatized, i.e., it does not need any supervision. Generating around 60 images per second, the whole process took about 8 hours on a system with mid-range specifications (i7-6820HK, NVidia GTX-1070, 16GB DDR4, SSD).

3.1. Synthetic Persons

The synthetic persons in the dataset were procedurally generated at run-time by making use of several content creation layers which consist of predefined set of categorizable, annotatable randomizations as well as procedural, low-level randomizations in order to yield a distinct look in each generated person (Fig. 1).

Table 1: A total of 84 ID-level attributes were used in annotating the Synthetic18K images.

Type	#	Type	#
Body Type	18*	Shoe Color	8
Color of Upper-Body Clothing	10	Hair Color	4
Length of Upper-Body Clothing Sleeve	3	Hair Type	6
Type of Lower-Body Clothing	4	Beard Type	2
Color of Lower-Body Clothing	10	Carrying Bag	Binary
Has Outerwear	Binary	Bag Color	6
Color of Outerwear	11		

*including gender information

Table 2: The numbers showing the attainable variations of facial, clothing and accessory items that can be used in the procedural generation of synthetic persons are given below. Extended variations by color changes are additionally provided inside parentheses.

Facial Items

Item	Male	Female
Hair	4 (48)	3 (32)
Eyebrows	2(24)	2 (24)
Beard	8 (96)	- / -

Clothing and Accessory Items

Item	Male	Female
Upper-Body Clothing	7 (28)	7 (28)
Lower-Body Clothing	6 (240)	13 (520)
Outerwear	2 (80)	3 (120)
Shoes	5 (40)	10 (80)
Bags	3 (12)	3 (12)
Other	2 (4)	3 (18)

Figure 1: An arbitrarily chosen sample of 24 synthetic persons from the Synthetic18K dataset indicating the distinct array of looks prevalent throughout the dataset.

Each person in the dataset has a unique body shape which can be categorized into one of 9 pre-defined major body types per gender. Uniqueness of a body shape is realized by applying a rather small white noise with uniform distribution to pre-defined body blend shapes. Facial attributes of the persons are also affected by these randomizations. The clothing and hair attributes for the persons are generated from a set of several content sets and can be colored at run-time. A shared color system ensures a wide variety of distinct looking persons with randomized colors for their clothing, skin and hair which are then categorized into main groups of colors accordingly for annotation (Table 2).

An arbitrarily chosen sample of 24 generated synthetic persons from the Synthetic18K dataset (Fig. 1) demonstrates that the generated persons are easily distinguishable from one another. Figure 2b presents a comparison of the images of three different synthetic persons from the Synthetic18K dataset to the ones from the real person re-ID datasets Market1501 [42] and DukeMTMC-reID [1].

(a) (b)

Figure 2: (a) Sample images of a synthetic person captured from various viewpoints at different locations, time of day and weather conditions. (b) A comparison of the images of three different synthetic persons from the Synthetic18K dataset (left) to the ones from the real person re-ID datasets Market1501 [42] (middle) and DukeMTMC-reID [1] (right). Real person faces are blurred for privacy concerns.

3.2. Environments

The Synthetic18K dataset contains images that are captured from different 3D environments, of which three are outdoors (a town square, a suburban street and a metropolitan urban district) and one is indoors (a subway station) (Fig. 3a). In addition, Synthetic18K also contains images that use HDR cubemaps captured from real-life as background environments (Fig. 3b). These make up approximately 24% of the entire dataset. However, as the cubemaps are static, these images do not convey background variations in terms of illumination and weather, which are present in the images captured inside the 3D environments.

Each synthetic person has images that are taken at 3 different locations in each scene and cubemap (Fig. 2a). The locations for the scenes are chosen randomly from a set of pre-determined points distributed throughout each scene. At each location, the person's images are captured for each of the time-of-day and weather variations that the framework can simulate (Fig. 3c).

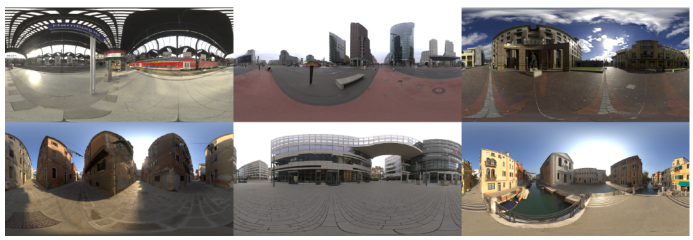
4. Approach

In our work, we consider the following three different tasks to demonstrate the importance of our pretraining strategy with synthetic data: (1) person re-ID, (2) attribute recognition, and (3) joint person re-ID and attribute recognition. The general overview of our approach is given in Fig. 4. In particular, for each aforementioned task, we utilize the same backbone network for feature extraction, and add additional modules to address the specifics of the task. In our experiments, we firstly pretrain the related model parameters on Synthetic18K at first and then fine-tune them on relevant real-world datasets. To validate the effectiveness of our approach, we also evaluate against the widely used strategy of using ImageNet-pretrained models. In the following, we give formal definitions of the tasks and describe the our model architectures together with some training and implementation details.

4.1. Person Re-ID

For person re-ID task, the training data consists of a set of person images $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ where N is the total number of images, and x_i and $y_i \in [1, K]$ refer to i-th person image and its identity label , respectively,

(a)



(b)

(c)

Figure 3: Illustrating the diversity of the environments used to generate the Synthetic18K dataset. (a) Sample images of the 3D environments clockwise from top-left: a metropolitan urban district, a town square, a subway station and a suburban street. (b) Sample HDR cubemaps captured from real-world [43]. (c) Simulation of different times of day and weather conditions at the same environment setting.

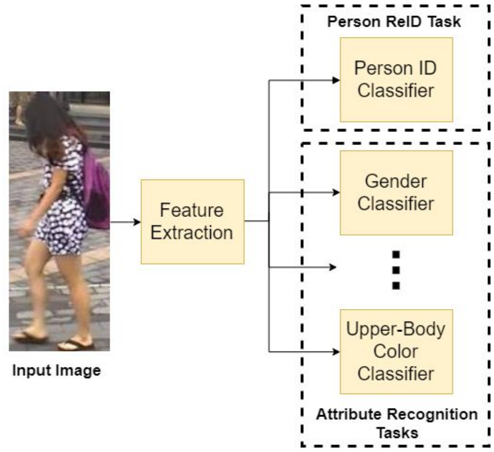


Figure 4: General overview of the proposed framework. In our work, we address person re-ID and attribute recognition by using a common backbone network as the feature extractor and extend this architecture according to the requirements of the task.

with K indicating the number of identities. In our analysis, we use two different network architectures for person re-ID. The first one is a simple network consisting of a basic backbone network and a classifier module. The second one, which we refer to as HUCVReidNet, is a network that employs attention mechanisms to better exploit contextual information and learn more discriminative features. In the test phase, each query and gallery image are represented in terms of feature responses at the last fully-connected layer, and we use the Euclidean distance as the metric to retrieve the nearest neighbors.

Basic Re-ID Network. This network consists of a backbone network and a classifier module attached to the end. Features extracted from the backbone network are first passed through a global average pooling layer, and then passed to classifier module to determine the identity. Classifier module consists of two fully connected (FC) layers, where the second fully connected layer has units as much as the number of identities in the training set. Any standard commonly-used

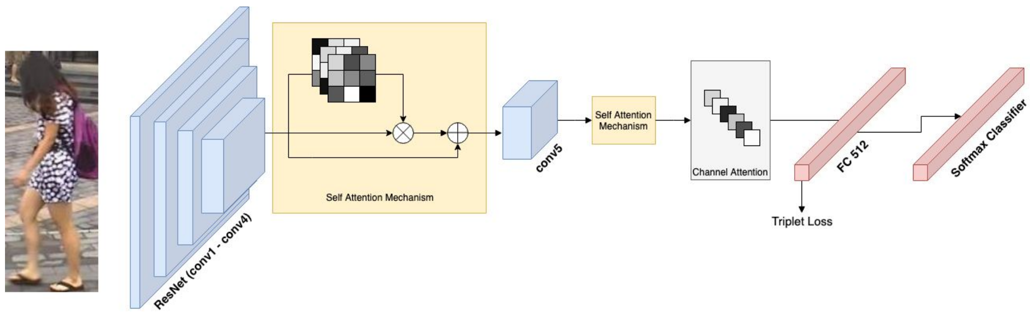


Figure 5: Our proposed HUCVReidNet model for person re-identification.

CNN architecture can be adapted as the backbone network. In our work, we conducted experiments with three different architectures (Section 5.2) and selected DenseNet-121 [44] as it gave the best performance.

HUCVReidNet. Our HUCVReidNet has a novel but simple architecture which employs attentional mechanisms, as shown in Fig. 5. The model uses ResNet-50 [4] as its backbone network. After the last convolutional layer of ResNet-50 (conv5 layer), we apply self attention [45] mechanisms both before and after this layer to refine the feature maps extracted by ResNet-50. Moreover, we apply a channel attention right before the global average pooling layer. Since pooling takes average along spatial dimensions for each channel separately, applying channel attention scales the channel features according to their importance before vectorizing them. These features are then passed to FC block which has linear layer with 1024 units, batch normalization layer and ReLU activation function. After this FC block, a final classification layer whose dimension is equal to the number of identities in the training set.

Loss function. We train the aforementioned models by using a composite loss function that includes an identification loss and a triplet ranking loss functions:

$$\mathcal{L}_{re-ID} = \mathcal{L}_{id} + \mathcal{L}_{tri} \tag{1}$$

Here, the first term $\mathcal{L}_{id}$ treats re-ID as a multi-class classification problem with each person identity representing a distinct class label. In particular, we use cross entropy loss on the final fully connected layer of the classifier module to enforce identity consistency, given as below:

$$\mathcal{L}_{id} = \mathbb{E}\left[-\log p(y_i|x_i)\right] \tag{2}$$

where $p(y_i|x_i)$ denotes the predicted probability of x_i belonging to the identity label y_i based on its extracted deep features. In our implementation, we also apply label smoothing to regularize the trained classifier by adding small constants to ground truth values instead of using 1 and 0s.

The second term in Eqn. (1) is the triplet loss $\mathcal{L}_{tri}$ which casts person re-ID task as a metric learning problem. Specifically, we represent each person image with the deep features extracted by the first fully connected layer of the classifier module. During training, we consider a triplet (x_i, x_j, x_k) consisting of two distinct images of the same person x_i and x_j and an image of a different person x_k. Triplet loss defined as follows enforces the model to learn a feature space in which images of the same person are mapped closer to each other where images of different persons are separated from each other by a large margin:

$$\mathcal{L}_{tri} = \mathbb{E}\left[[d(f(x_i), f(x_j)) + m - d(f(x_i), f(x_k))]_+\right] \tag{3}$$

where $[z]_+ = \max(0, z)$, m is a scalar representing the margin, $f(x_i)$ is the deep feature representation of image x_i, and d denotes the Euclidean distance. In our implementation, to increase the robustness of the learned feature space, we use hard positive and hard negative mining proposed by Hermans et al. [9].

Training details. Our person re-ID networks take 3-channel RGB images of persons that are resized to 128×256 pixels. We train our basic re-ID and HUCVReidNet models for 80 epochs, by using Adam with a batch size of 32. For both of these models, we set the initial learning rate to 0.001 for the newly added layers and 0.0001 for the layers of the backbone network. We do not use any data augmentation during training on our Synthetic18K dataset. Translation and horizontal flip are applied randomly during fine-tuning of real-world datasets.

4.2. Attribute Recognition

The training data for attribute recognition task contains a set of pairs $\mathcal{S} = \{(x_i, \mathbf{a}_i)\}$ where each pair consists of a person image x and a set of attributes $\mathbf{a} = \left(a^1, a^2, \ldots, a^M\right)$. In our work, we use a multi-task learning approach for attribute recognition, where classification of each attribute is considered as a separate classification task. Like our basic architecture in person re-ID, we use DenseNet-121 as our backbone network and define separate classification modules for each attribute. Similarly, each classifier module consist of two FC layers and a classification layer whose dimension of the final is equal to the number of different labels for that attribute. As the backbone network is shared between all classifier modules, it learns to extract features that is useful for all of the attributes.

Loss function. To train our network, we use a weighted cross entropy loss for each attribute classifier module, which results in the following joint loss function:

$$\mathcal{L}_{attr} = \mathbb{E}\left[-\sum_{j=1}^{M} \lambda_j \log p(a_i^j | x_i))\right] \tag{4}$$

Here, λ_j is a scalar denoting the importance of j-th attribute and $p(a_i^j|x_i)$ denotes the predicted probability of x_i having the attribute a_i^j based on the extracted deep features. In our implementation, we set λ_j in accordance with the total number of class samples to avoid class imbalance problem.

Training details. Our attribute recognition network use 128×256 pixels RGB images of persons as input. We train our model for 80 epochs by using Adam algorithm with batches of size 32. We set the initial learning rate to 0.001 for the newly added layers and 0.0001 for the layers of the backbone network. No data augmentation is used during training on synthetic person images. For real data, we apply a similar data augmentation scheme we used as in person re-ID.

4.3. Joint Person Re-ID and Attribute Recognition

In this task, we jointly train for person re-ID and attribute recognition tasks. The intuition is to utilize a shared backbone network for these tasks where the training of this network involves supervision signals from both person id and attribute labels. Since person re-ID and attribute recognition are two closely related tasks, we expect that a joint training scheme would result in better performances for both of these two tasks. The overall system architecture can be seen in Fig. 6. It is similar to the attribute recognition architecture but with an additional classifier module for person re-ID, containing $N + 1$ separate classifier modules, N modules for classifying N distinct attributes and one for the identification. These classifier modules are same as the ones that are used in basic person re-ID and attribute recognition networks.

Loss function. Person re-ID classifier is trained with both cross entropy loss and triplet ranking loss, and attribute classification classifiers are trained with only cross entropy losses. The common backbone network is trained via supervision

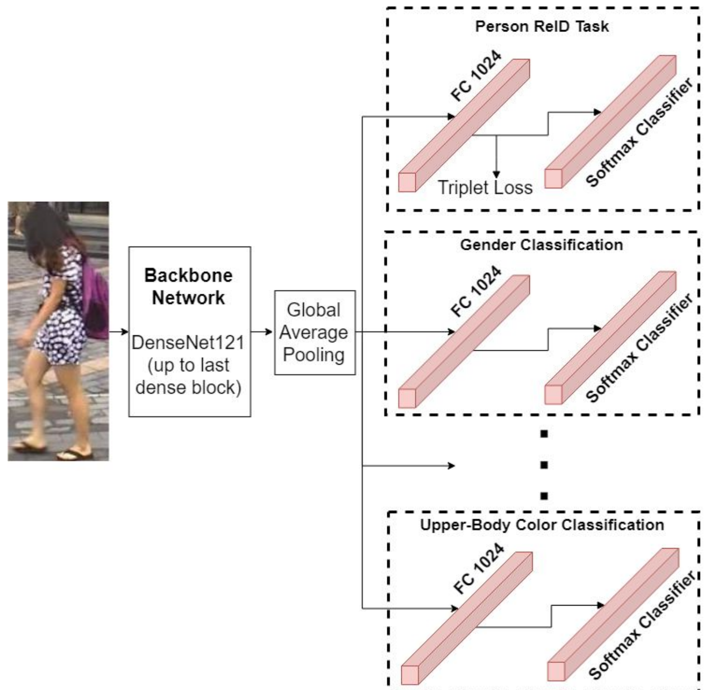


Figure 6: Person Re-ID and Attribute Recognition Multi-Task Network

signals from the combination of these losses as defined below:

$$\mathcal{L}_{joint} = \mathcal{L}_{attr} + \beta \mathcal{L}_{re-ID} \tag{5}$$

where β is a weight factor. In our experiments, we have observed that setting β to 2 gives a good trade-off between to the two tasks.

Training details. Training strategy and hyperparameters for our joint network for the person-reid and attribute recognition tasks is completely same with those for attribute recognition task.

4.4. Pretraining for Feature Learning

The proposed Synthetic18K dataset differs from the existing synthetic datasets proposed for person re-identification in certain aspects as mentioned before. It contains images of larger number of synthetic identities captured under various illumination conditions and backgrounds. But, more importantly, our procedural generation framework allows us to play with the attributes of the generated persons as well. We used this capability to collect a large-scale dataset which can be used for both person re-identification and attribute recognition tasks. Our Synthetic18K dataset can be used for learning feature representations robust for these two tasks. As mentioned in the previous subsections, this is achieved by carefully designing pretraining strategies for the proposed deep models.

In our work, we considered three different pretraining schemes. While our first pretraining strategy involves only the person re-identification task, our second strategy considers the attribution recognition task. For all these settings, we develop simple neural deep models. Finally, our third pretraining scheme employs a combination of these two and involves a multi-task learning setting for joint person re-identification and attribution recognition. In that respect, in our third strategy, we combine our proposed deep models by taking into account their common backbone network architecture and introduce separate heads for each one of these tasks. For all of our pretraining strategies, we follow a similar training scheme. That is, as the first step, we train our proposed model either by using the individual tasks or by utilizing both in a multi-task learning setting. This initial training step lets the deep models learn distinctive features or filter weights for the task(s) under consideration. We then use these weights to initialize the model parameters for the experiments done on the real datasets and perform finetuning the actual real data. While doing so, we set the learning rate to ...

5. Experimental Results

In the following, we first summarize the evaluation metrics used in our experiments. We then provide an analysis on the test set of the Synthetic18K dataset for our deep models for person re-ID and attribute recognition. Next, we test the performances of these pretrained models on several real-life datasets. Finally, we compare our results with the state-of-the-art models proposed for person re-ID, attribute recognition as well as jointly trained ones.

5.1. Evaluation Metrics

To evaluate performance on person re-ID task, we use cumulative matching characteristics (CMC) and mean average precision (mAP). To compute these, gallery images are sorted by their similarity to the query image for each query. CMC curve represents the expectation to include true person identity in the first k images of the sorted gallery images. mAP is the mean value of the precision scores for all queries, where average precision for a single query is the area under the precision-recall curve. Since CMC curve considers only the first match of the true identity within k images, mAP metric is also used for person re-ID, which rewards retrieving multiple true identities. For attribute recognition task, we report classification accuracy for each attribute as well as their averages (mA).

5.2. Validation on Synthetic Images

We first analyze how does the choice of backbone network and loss functions affect the performances of our basic re-ID model. We split our Synthetic18K dataset into training and test sets according to person identities, each containing 12K and 6K different persons, respectively. Table 3 shows the results of our

Table 3: Person Re-ID Performance on Synthetic18K.

	$\mathcal{L}_{id}$				$\mathcal{L}_{id} + \mathcal{L}_{tri}$			
Backbone	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
DenseNet-121	98.9	99.2	100.0	100.0	99.4	99.9	100.0	100.0
ResNet-50	96.6	98.2	99.7	100.0	97.4	99.1	100.0	100.0
MobileNetV2	94.5	96.7	98.9	99.6	96.1	97.1	99.0	99.7

analysis. In particular, we consider DenseNet-121 [44], ResNet-50 [4], and MobileNetV2 [46] models as our backbone network and train them by using solely the identity loss ($\mathcal{L}_{id}$) and the joint loss function containing both the identity and the triplet loss ($\mathcal{L}_{id} + \mathcal{L}_{tri}$). We observe that our model with DenseNet-121 as its backbone achieves slightly higher scores than the other two models. Moreover, training the models with the joint loss function improves the performances. Hence, as we mentioned before, for the rest of the experiments we use DenseNet-121 model in our basic network models for person re-ID and attribute recognition.

In Table 4, we provide the performance of our attribute recognition model on our Synthetic18K dataset. Synthetic person images introduce certain challenges as compared to the aforementioned analysis regarding person re-ID that the individual attribute prediction scores are not very high, especially predicting color attributes seems more difficult. This demonstrates that Synthetic18K dataset could be used as a test-bed for attribute recognition approaches.

5.3. *Using Synthetic Data for Pretraining*

In this section, we provide several experiments regarding how pretraining on our Synthetic18K dataset can help improving performances on real datasets, espe-

Table 4: Attribute Recognition Performance on Synthetic18K.

Gender	Age	Hair	Beard	Weight	Sleeve Len.	L.Body Clth	
99.5	94.8	97.4	95.0	93.0	82.2	97.3	
L.Body Clth Col	U.Body Clth Col	Over Clth	Hand Bag	Hand Bag Col	Hair Col	Shoe Col	mA
58.8	84.1	76.1	85.8	85.2	77.1	93.4	87.1

cially compared to commonly used strategy of using ImageNet pretrained models.

Person Re-ID. In our experiments, we use Market1501 [42] and DukeMTMC-reID [1] datasets containing 32,668 and 34,183 real person images, respectively. In Market1501, 751 identities are allocated for training and the rest 750 identities for testing. In DukeMTMC-reID contains 1404 identities, of which 702 identities are selected for training and the rest for testing. In our analysis, we fine-tune our basic re-ID and HUCVReidNet models, which were pretrained on our Synthetic18K dataset, on the training sets of these datasets and report their performances on the corresponding test sets accordingly. As a comparison, we also provide the results of our models which instead utilize ImageNet pretrained backbones. Table 5 reports these comparisons. As can be seen, pretraining on Synthetic18K improves re-ID performances on both Market1501 and DukeMTMC-reID datasets. For our basic re-ID model, pretraining on Synthetic18K results in 2.9 and 0.6 increase in mAP, and 1.7 and 0.8 increase on Rank-1 scores on Market1501 and DukeMTMC-reID datasets, respectively. Again, for HUCVReidNet, pretraining gives much better results in terms of mAP, Rank-1 and Rank-5 scores. Moreover, with its inherent attention mechanisms, our proposed HUCVReidNet model gives better results than our basic re-ID model.

Attribute Recognition. We conduct our pretraining analysis on Market1501-Attributes [31] dataset, an extended version of Market1501 [42] where each per-

Table 5: Analysis of pretraining on Synthetic18K for person ReID.

		Synthetic18K			ImageNet		
Model	Dataset	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
Basic ReidNet	Market1501	77.3	91.6	96.6	74.4	89.9	96.1
	DukeMTMC-reID	63.3	81.3	90.6	62.7	80.5	89.4
HUCVReidNet	Market1501	78.8	92.4	97.9	78.3	91.9	97.1
	DukeMTMC-reID	63.7	83.0	91.9	63.7	81.7	91.0

Table 6: Analysis of pretraining on Synthetic18K for attribute recognition

Pretrain	gender	age	hair	L.slv	L.low	S.clth	B.pack	H.bag	bag	hat	C.up	C.low	mA
ImageNet	88.7	84.8	85.7	92.5	92.7	93.1	86.2	87.6	73.8	95.2	77.0	70.1	85.6
Synthetic18K	**91.4**	**85.6**	**86.6**	**93.5**	**93.6**	**94.0**	**87.8**	**88.1**	**76.9**	**97.5**	**78.1**	**71.0**	**87.0**

son image is annotated with 27 different attributes. We used the same training and testing splits as in person re-ID. Table 6 shows accuracy scores for each attribute as well as the average accuracy (mA). We observe that the model that has pretrained on Synthetic18K outperforms the ImageNet pretrained model for all attributes, resulting in 1.4 increase in the mean accuracy. This demonstrates that our synthetic data pretraining approach is also effective for attribute recognition task.

Joint Person Re-ID and Attribute Recognition. We use Market1501 and Market1501-Attributes datasets and follow a similar strategy and fine-tune our joint model pretrained on our Synthetic18K dataset on the training set of these datasets using both person attributes and identities, and subsequently evaluate it on the corresponding test set. Table 7 reports prediction accuracies of the person attributes together with mAP and Rank-1 for re-ID. We again observe that our joint model pretrained on

Table 7: Analysis of pretraining on Synthetic18K for Joint Re-ID and Attribution Recognition

Pretrain	gender	age	hair	L.slv	L.low	S.clth	B.pack	H.bag	bag	hat	C.up	C.low	mA	mAP	Rank-1
ImageNet	89.9	85.7	86.1	93.7	92.9	92.9	86.1	88.2	76.3	97.1	76.7	70.3	86.3	76.4	88.9
Synthetic18K	**91.5**	**86.2**	**88.2**	**93.9**	**94.0**	**94.0**	**88.0**	**88.8**	**78.9**	**97.2**	**78.4**	**71.4**	**87.5**	**78.4**	**90.3**

Synthetic18K outperforms the ImageNet pretrained model by a large margin. Pretraining on Synthetic18K gives 1.2 increase in mA, 2.0 increase in mAP and 1.4 increase in Rank-1 score. Moreover, these results demonstrate that jointly training a model for person re-ID and attribute recognition model improves the model performances for the individual tasks (cf. Table 5 and 6).

5.4. Comparison with the state-of-the-art

In this section, we compare the results of our person re-ID, attribute recognition models and their joint version against those of the state-of-the-art approaches. **Person Re-ID.** Table 8 and 9 show the comparison between our basic re-ID and HUCVReidNet models that we pretrained on our Synthetic18K dataset and the state-of-the-art algorithms on Market-1501 and DukeMTMC-reID datasets, respectively. We observe that our models outperforms most of the recently proposed re-ID models. There are a few methods that give relatively better results than ours but these models have highly complex network architectures. For instance, the models in [20, 47, 19, 48] consider part structures of the human to extract local features from images, the method proposed in [14] employs graph neural network instead of the commonly use Siamese networks to compare query-gallery pairs, or the work in [49] uses self-distillation learning along with a novel code pyramid based coarse-to-fine (CtF) hashing code search strategy. In addition to these, the model by Chen et al. [50] utilizes text descriptions of the person images to guide

Table 8: Person Re-ID performances on Market1501.

Method	mAP	Rank-1	Rank-5	Method	mAP	Rank-1	Rank-5
OIM Loss [52]	60.9	82.1	-	HA-CNN [53]	75.7	91.2	-
SpindleNet [17]	-	76.9	91.5	Pose-transfer [54]	58.0	79.8	-
MSCAN [18]	57.5	80.3	-	MLFN [55]	74.3	90.0	-
SSM [56]	68.8	82.2	-	DML [57]	70.5	89.3	-
k-reciprocal [58]	63.6	77.1	-	Suh [47]	79.6	91.7	96.9
Point 2 Set [12]	44.3	70.7	-	SPReID [20]	83.4	93.7	97.6
CADL [59]	47.1	73.8	-	SGGNN [14]	82.8	92.3	96.1
DPFL [60]	73.1	88.9	-	GLILA [50]	81.8	93.3	-
VI+LSRO [61]	66.1	84.0	-	Mancs [51]	82.3	93.1	-
SVDNet [62]	62.1	82.3	92.3	PCB+RPP [19]	81.6	93.8	97.5
OL-MANS [63]	-	60.7	-	IID [41]	71.5	88.5	-
Pose Driven [16]	63.4	84.1	92.7	Wang et al. [40]	70.9	87.2	-
Part Aligned [64]	63.4	81.0	92.0	Xiang et al [39]	50.8	76.2	89.2
HydraPlus-Net [65]	-	76.9	91.3	Generalizing-Reid [66]	71.5	88.1	94.4
TriNet [9]	69.1	84.9	94.2	RGA-SC [67]	88.4	96.1	-
DarkRank [68]	74.3	89.8	-	ISP [48]	88.6	95.3	98.6
PN-GAN [69]	72.6	89.4	-	CtF [49]	84.9	93.7	-
DuATM [70]	76.6	91.4	97.1	Zhuang et al. [71]	77.3	91.3	-
HAP2S_E [72]	69.8	84.2	-	Ours (Basic ReidNet)	77.3	91.6	96.6
HAP2S_P [72]	69.4	84.6	-	Ours (HUCVReid)	78.8	92.4	97.9

learning of the local and global visual features. The model suggested by Wang et al. [51] has many attention blocks at different layers of the backbone network and during its training separate classification loss functions are attached to output of each of these blocks.

Table 9: Person Re-ID performances on DukeMTMC-reID.

Method	mAP	Rank-1	Rank-5	Method	mAP	Rank-1	Rank-5
BoW+KISSME [42]	12.2	25.1	-	SPReID [20]	73.3	86.0	93.0
LOMO+XQDA [73]	17.0	30.8	-	SGGNN [14]	68.2	81.1	88.4
APR [31]	51.9	70.7	-	Suh [47]	69.3	84.4	92.2
ACRN [74]	52.0	72.6	84.8	Mancs [51]	71.8	84.9	-
DPFL [60]	60.6	79.2	-	AANet-152 [34]	74.3	87.7	-
OIM Loss [52]	47.4	68.1	-	PCB+RPP [19]	69.2	83.3	90.5
Basel.+LSRO [61]	47.1	67.7	-	IID [41]	60.6	78.1	-
SVDNet [62]	56.8	76.7	86.4	Wang et al. [40]	60.6	79.4	-
CamStyle [75]	57.6	78.3	-	Xiang et al [39]	51.9	71.2	82.7
Pose-transfer [54]	48.1	68.6	-	Generalizing-Reid [66]	65.2	79.5	88.3
MLFN [55]	62.8	81.0	-	ISP [48]	80.0	89.6	95.5
DuATM [70]	64.6	81.8	90.2	CtF [49]	74.8	87.6	-
PN-GAN [69]	53.2	73.6	-	Zhuang et al. [71]	67.3	82.5	-
HAP2S_P [72]	60.6	75.9	-	Ours (Basic ReidNet)	63.3	81.3	90.6
HAP2S_E [72]	59.6	76.1	-	Ours (HUCVReid)	63.7	83.0	91.9

Attribute Recognition. Table 10 provides a comparison between our model against the recent attribute recognition models on the Market1501-Attributes dataset. Similar to person re-ID, our proposed networks that have been pretrained on our Synthetic18K dataset give better or comparable accuracies as compared to the state-of-the-art models. Only the average accuracies of the methods proposed in [76, 33] are a bit higher than ours, but these models achieve these results either by considering training attribute recognition jointly with person re-ID [33] or by explicitly learning the importance of each attribute on a validation set [76].

Table 10: Attribute Recognition Performances on Market1501-Attributes

Method	gender	age	hair	L.slv	L.low	S.clth	B.pack	H.bag	bag	hat	C.up	C.low	mA
ARN [31]	87.5	85.8	84.2	93.5	93.6	93.6	86.6	88.1	78.6	97.0	72.4	71.7	86.0
APR [31]	88.9	88.6	84.4	93.6	93.7	92.8	84.9	90.4	76/4	97.1	74.0	73.8	86.6
Sun et al. [32]	88.9	84.8	78.3	93.5	92.1	84.8	85.5	88.4	67.3	97.1	87.5	87.2	86.3
AWMDN [76]	-	-	-	-	-	-	-	-	-	-	-	-	88.5
MLFN [55]	-	-	-	-	-	-	-	-	-	-	-	-	85.3
PANDA [77]	-	-	-	-	-	-	-	-	-	-	-	-	86.8
JCM [33]	89.7	87.4	82.5	93.7	93.3	89.2	85.2	86.2	86.9	97.2	92.4	93.1	89.7
AANet-152 [34]	92.3	88.2	86.6	94.5	94.2	94.8	87.8	89.6	79.7	98.0	77.0	70.8	87.8
Ours (Basic)	91.4	85.6	86.6	93.5	93.6	94.0	87.8	88.1	76.9	97.5	78.1	71.0	87.0
Ours (Joint)	91.5	86.2	88.2	93.9	94.0	94.0	88.0	88.8	78.9	97.2	78.4	71.4	87.5

Joint Attribute Prediction and Person Re-ID. In Table 11, we show the results of our joint model along with those of the recent approaches which also consider joint training of a model on both person re-ID and attribute recognition tasks. We find that our model achieves much better re-ID and recognition performances than most of the state-of-the-art approaches. The model JCM-57344 in [33] outperforms our model but it achieves this by using a 57344 dimensional feature embedding. In fact, the re-ID performance of its second version with a 1024 dimensional representation is much lower than ours. Moreover, the AANet models in [34] give a bit better predictions than ours. However, it is important to mention that AANet employs body part locations to extract local features while we only consider a global representation of person images.

6. Conclusion

In this work, we have introduced Synthetic18K dataset that consists of synthetically generated photo-realistic person images. Each image in our dataset is annotated with both the person identity label and the relevant person attributes. Particu-

Table 11: Joint Attribute Prediction and Person Re-ID Performances on Market1501

Method	mA	mAP	Rank-1	Method	mA	mAP	Rank-1
ACRN [74]	-	62.6	83.6	AANet-50 [34]	-	82.4	93.9
JCM-1024 [33]	-	75.7	84.9	AANet-152 [34]	87.8	83.4	93.9
JCM-57344 [33]	89.7	81.2	91.3	Wang et al. [35]	-	76.0	91.3
Sun et al. [32]	87.0	70.1	87.0	Ours (Joint)	87.5	78.4	90.3
APR [31]	86.6	66.9	87.0				

larly, we have addressed person re-ID and attribute recognition tasks and demonstrated that large-scale pretraining of simple deep models on our Synthetic18K dataset greatly improves the model performances on the real-life datasets. Moreover, we have demonstrated that joint training of a basic deep model for person re-ID and attribute recognition on Synthetic18K outperforms the individual model performances and gives better or comparable results than the state-of-the-art methods. As a future work, we plan to investigate the use of synthetic data to boost the performance of video re-ID using computer generated video sequences.

Acknowledgments

This work was supported in part by TUBA GEBIP fellowship awarded to E. Erdem and by TUBITAK-1001 Program Award No. 217E029.

References

[1] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: Proc. ECCV Workshop on Benchmarking Multi-Target Tracking, 2016.

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

[3] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R. J. Radke, A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets, IEEE Trans. Pattern Anal. Mach. Intell. 41 (3) (2019) 523–536.

[4] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. CVPR, 2016, pp. 770–778.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Proc. CVPR, 2009.

[6] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, Mars: A video benchmark for large-scale person re-identification, in: Proc. ECCV, 2016.

[7] D. Chung, K. Tahboub, E. J. Delp, A two stream siamese convolutional neural network for person re-identification, in: Proc. ICCV, 2017.

[8] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognition 48 (10) (2015) 2993–3003.

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

[10] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: Proc. CVPR, 2017.

[11] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, in: Proc. CVPR, 2016, pp. 1288–1296.

[12] S. Zhou, J. Wang, J. Wang, Y. Gong, N. Zheng, Point to set similarity based deep feature learning for person re-identification, in: Proc. CVPR, 2017, pp. 3741–3750.

[13] W. Chen, X. Chen, J. Zhang, K. Huang, A multi-task deep network for person re-identification, in: Proc. AAAI, 2017, pp. 3988–3994.

[14] Y. Shen, H. Li, S. Yi, D. Chen, X. Wang, Person re-identification with deep similarity-guided graph neural network, in: Proc. ECCV, 2018, pp. 486–504.

[15] Z. Wang, R. Hu, Y. Yu, J. Jiang, C. Liang, J. Wang, Scale-adaptive low-resolution person re-identification via learning a discriminating surface, in: Proc. IJCAI, 2016, p. 26692675.

[16] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in: Proc. ICCV, 2017, pp. 3960–3969.

[17] 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: Proc. CVPR, 2017, pp. 1077–1085.

[18] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: Proc. CVPR, 2017, pp. 384–393.

[19] 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: Proc. ECCV, 2018, pp. 480–496.

[20] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in: Proc. CVPR, 2018, pp. 1062–1071.

[21] G. Sharma, F. Jurie, C. Schmid, Expanded parts model for human attribute and action recognition in still images, in: Proc. CVPR, 2013, pp. 652–659.

[22] Y. Deng, P. Luo, C. C. Loy, X. Tang, Pedestrian attribute recognition at far distance, in: Proc. ACM-MM, 2014, pp. 789–792.

[23] H. Chen, A. Gallagher, B. Girod, Describing clothing by semantic attributes, in: Proc. ECCV, 2012, pp. 609–623.

[24] P. Sudowe, H. Spitzer, B. Leibe, Person attribute recognition with a jointly-trained holistic CNN model, in: Proc. CVPRW, 2015, pp. 87–95.

[25] J. Zhu, S. Liao, Z. Lei, S. Z. Li, Multi-label convolutional neural network based pedestrian attributeclassification, Image and Vision Computing 58 (2017) 224–229.

[26] J. Wang, X. Zhu, S. Gong, W. Li, Attribute recognition by joint recurrent learning of context and correlation, in: Proc. ICCV, 2017, pp. 531–540.

[27] R. Layne, T. Hospedales, S. Gong, Person re-identification by attributes, in: Proc. BMVC, 2012, pp. 24.1–24.11.

[28] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, W. Gao, Multi-task learning with low rank attribute embedding for person re-identification, in: Proc. ICCV, 2015, pp. 3739–3747.

[29] P. Peng, Y. Tian, T. Xiang, Y. Wang, T. Huang, Joint learning of semantic and latent attributes, in: Proc. ECCV, 2016.

[30] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Deep attributes driven multi-camera person re-identification, in: Proc. ECCV, 2016.

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

[32] C. Sun, N. Jiang, L. Zhang, Y. Wang, W. Wu, Z. Zhou, Unified framework for joint attribute classification and person re-identification, in: Proc. ICANN, 2018, pp. 637–647.

[33] H. Liu, J. Wu, J. Jiang, M. Qi, R. Bo, Sequence-based person attribute recognition with joint ctc-attention model, arXiv preprint arXiv:1811.08115 (2018).

[34] C.-P. Tay, S. Roy, K.-H. Yap, Aanet: Attribute attention network for person re-identifications, in: Proc. CVPR, 2019, pp. 7134–7143.

[35] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, S. Satoh, Learning sparse and identity-preserved hidden attributes for person re-identification, IEEE Trans. Image Process. 29 (2020) 2013–2025.

[36] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, T. Theoharis, Looking beyond appearances: Synthetic training data for deep CNNs in re-identification, Comput. Vis. Image Und. 167 (2018) 50–62.

[37] X. Sun, L. Zheng, Dissecting person re-identification from the viewpoint of viewpoint, in: Proc. CVPR, 2019.

[38] S. Bak, P. Carr, J.-F. Lalonde, Domain adaptation through synthesis for unsupervised person re-identification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 189–205.

[39] S. Xiang, Y. Fu, G. You, T. Liu, Unsupervised domain adaptation through synthesis for person re-identification, in: Proc. ICME, 2020.

[40] Y. Wang, S. Liao, L. Shao, Surpassing real-world source training data: Random 3d characters for generalizable person re-identification, in: Proc. ACM-MM, 2020, p. 34223430.

[41] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, S. Satoh, Illumination-adaptive person re-identification, IEEE Trans. Multimedia 22 (12) (2020) 3064–3074.

[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proc. ICCV, 2015, pp. 1116–1124.

[43] G. Zaal, HDRI Haven, `https://hdrihaven.com/hdris`, online; accessed: 2019-02-20 (2019).

[44] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proc. CVPR, 2017, pp. 4700–4708.

[45] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, arXiv preprint arXiv:1805.08318 (2018).

[46] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proc. CVPR, 2018, pp. 4510–4520.

[47] Y. Suh, J. Wang, S. Tang, T. Mei, K. Mu Lee, Part-aligned bilinear representations for person re-identification, in: Proc. ECCV, 2018, pp. 402–419.

[48] K. Zhu, H. Guo, Z. Liu, M. Tang, J. Wang, Identity-guided human semantic parsing for person re-identification, in: Proc. ECCV, 2020.

[49] G. Wang, S. Gong, J. Cheng, Z. Hou, Faster person re-identification, in: Proc. ECCV, 2020.

[50] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, X. Wang, Improving deep visual representation for person re-identification by global and local image-language association, in: Proc. ECCV, 2018, pp. 54–70.

[51] C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs: A multi-task attentional network with curriculum sampling for person re-identification, in: Proc. ECCV, 2018, pp. 365–381.

[52] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proc. CVPR, 2017, pp. 3415–3424.

[53] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: Proc. CVPR, 2018, pp. 2285–2294.

[54] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, J. Hu, Pose transferrable person re-identification, in: Proc. CVPR, 2018, pp. 4099–4108.

[55] X. Chang, T. M. Hospedales, T. Xiang, Multi-level factorisation net for person re-identification, in: Proc. CVPR, 2018, pp. 2109–2118.

[56] S. Bai, X. Bai, Q. Tian, Scalable person re-identification on supervised smoothed manifold, in: Proc. CVPR, 2017, pp. 2530–2539.

[57] Y. Zhang, T. Xiang, T. M. Hospedales, H. Lu, Deep mutual learning, in: Proc. CVPR, 2018, pp. 4320–4328.

[58] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proc. CVPR, 2017, pp. 1318–1327.

[59] J. Lin, L. Ren, J. Lu, J. Feng, J. Zhou, Consistent-aware deep learning for person re-identification in a camera network, in: Proc. CVPR, 2017, pp. 5771–5780.

[60] Y. Chen, X. Zhu, S. Gong, Person re-identification by deep learning multi-scale representations, in: Proc. ICCV, 2017, pp. 2590–2600.

[61] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: Proc. ICCV, 2017, pp. 3754–3762.

[62] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: Proc. ICCV, 2017, pp. 3800–3808.

[63] J. Zhou, P. Yu, W. Tang, Y. Wu, Efficient online local metric adaptation via negative samples for person re-identification, in: Proc. ICCV, 2017, pp. 2420–2428.

[64] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, in: Proc. ICCV, 2017, pp. 3219–3228.

[65] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, Hydraplus-net: Attentive deep features for pedestrian analysis, in: Proc. ICCV, 2017, pp. 350–359.

[66] C. Luo, C. Song, Z. Zhang, Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup, in: Proc. ECCV, 2020.

[67] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, Relation-aware global attention for person re-identification, in: Proc. CVPR, 2020.

[68] Y. Chen, N. Wang, Z. Zhang, Darkrank: Accelerating deep metric learning via cross sample similarities transfer, in: Proc. AAAI, 2018.

[69] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, X. Xue, Pose-normalized image generation for person re-identification, in: Proc. ECCV, 2018, pp. 650–667.

[70] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, G. Wang, Dual attention matching network for context-aware feature sequence based person re-identification, in: Proc. CVPR, 2018, pp. 5363–5372.

[71] Z. Zhuang, L. Wei, L. Xie, T. Zhang, H. Zhang, H. Wu, H. Ai, Q. Tian, Re-

thinking the distribution gap of person re-identification with camera-based batch normalization, in: Proc. ECCV, 2020.

[72] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, X. Bai, Hard-aware point-to-set deep metric for person re-identification, in: Proc. ECCV, 2018, pp. 188–204.

[73] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Proc. CVPR, 2015, pp. 2197–2206.

[74] A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information, in: Proc. CVPRW, 2017, pp. 20–28.

[75] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camstyle: A novel data augmentation method for person re-identification, IEEE Trans. Image Process. 28 (3) (2018) 1176–1190.

[76] K. He, Z. Wang, Y. Fu, R. Feng, Y.-G. Jiang, X. Xue, Adaptively weighted multi-task deep network for person attribute classification, in: Proc. ACM-MM, 2017, pp. 1636–1644.

[77] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, Panda: Pose aligned networks for deep attribute modeling, in: Proc. CVPR, 2014, pp. 1637–1644.