Graph-based Person Signature for Person Re-Identifications

Binh X. Nguyen[1], Binh D. Nguyen[1], Tuong Do[1], Erman Tjiputra[1], Quang D. Tran[1], Anh Nguyen[2]
[1]AIOZ, Singapore
[2]Imperial College London, UK
{binh.xuan.nguyen,binh.duc.nguyen,tuong.khanh-long.do,erman.tjiputra,quang.tran}@aioz.io
a.nguyen@imperial.ac.uk

Abstract

The task of person re-identification (ReID) is to match images of the same person over multiple non-overlapping camera views. Due to the variations in visual factors, previous works have investigated how the person identity, body parts, and attributes benefit the person ReID problem. However, the correlations between attributes, body parts, and within each attribute are not fully utilized. In this paper, we propose a new method to effectively aggregate detailed person descriptions (attributes labels) and visual features (body parts and global features) into a graph, namely Graph-based Person Signature, and utilize Graph Convolutional Networks to learn the topological structure of the visual signature of a person. The graph is integrated into a multi-branch multi-task framework for person re-identification. The extensive experiments are conducted to demonstrate the effectiveness of our proposed approach on two large-scale datasets, including Market-1501 and DukeMTMC-ReID. Our approach achieves competitive results among the state of the art and outperforms other attribute-based or mask-guided methods. Source available at https://github.com/aioz-ai/CVPRW21_GPS.

1. Introduction

Person re-identification (ReID) aims to retrieve a particular person image in a collection of images captured by multiple cameras from various viewpoints across time. The challenges of the person ReID task come from significant variations of human attributes such as poses, gaits, clothes, as well as challenging environmental settings like illumination, complex background, and occlusions. With the rise of deep learning, most of the recent studies utilize Convolutional Neural Network (CNN) to tackle the person ReID problem. Many approaches have been proposed such as metric learning [14, 30, 3], attention-based [40, 47, 2, 27], GAN-based [32, 9, 57], attribute-based [19, 21, 11, 1, 25, 52], and spatial-temporal-based methods [41].

Recently, attribute-based methods have shown great success in providing semantic features for the deep network [25, 39]. Unlike the person identity label, which offers only coarse information to identify one identity among all other person identities, the attributes are the detailed descriptions that are highly intuitive and mostly unchanged between images captured from different cameras. Therefore, they can be used to explicitly guide the model to learn a robust person representation by defining human characteristics. Furthermore, as shown in [21], attributes can also be used to speed up the retrieval process of the person ReID task by filtering out images from the gallery that do not share the same attributes with the probe image.

In this work, we propose to utilize the person attribute information with its associated body part to encode the visual person signature in one unified framework. We hypothesize that the detailed person descriptions (attributes labels) can be integrated with visual features (body parts and global features) to create a unique signature for a particular person. Since both body parts and attributes provide local representations, by linking them together, the network can have a better understanding of the relationship between visual features and attribute descriptions. Although previous works have investigated how person identity, body parts, and attributes benefit the task of person ReID [21, 52, 35, 39], our key difference is that we utilize Graph Convolutional Networks (GCN) to effectively construct and model the correlation between attributes and body parts with global features. In particular, we treat body part regions and attributes as nodes in a graph and utilize a GCN to learn the topological structure of a person's signatures. The GCN propagates messages on a graph structure. After message traversal on the graph, the node's final representations are obtained from its data and from other node's information. Fig. 1 shows the effectiveness of our approach.

2. Related work

Methods based on deep convolutional networks have dominated in the ReID community. In this section, we thor-

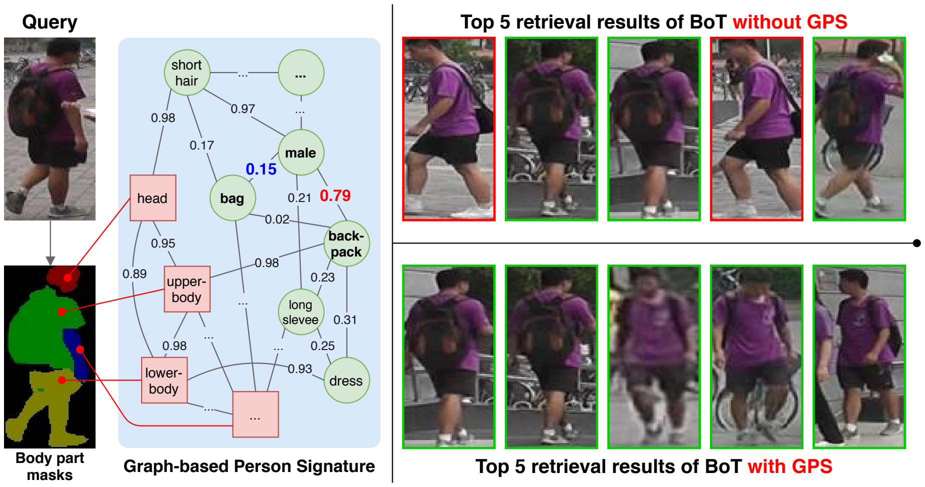

Figure 1. The effectiveness of our GPS in improving retrieval results on Market-1501 dataset [56]. Note that the green/red boxes denote true/false retrieval results, respectively. The retrieval results of the baseline model BoT [24] have some different attributes, e.g., 'bag' and 'backpack'. This leads to false retrieval results at rank-1 and rank-4. However, when integrating our GPS into BoT, these false results are removed. Our GPS gives the correlations between attributes and body parts (the graph in blue background). The correlation between 'male' and 'backpack' is higher than between 'male' and 'bag' (i.e., 0.79 $>$ 0.15). Therefore, the information extracted from GPS makes the person feature more discriminative and consequently improves the results.

oughly review two popular approaches similar to our proposed method: Part-based approach and Attribute-based approach. Other approaches are briefly also mentioned.

Part-based approach. Several efforts [43, 48, 53] learn meaningful features from local parts of a probe image to improve the representativeness of a person by incorporating the body part region information, i.e., the local spatial information. In particular, the local CNN method [48] proposed to split the backbone network into multiple parts horizontally and incorporate the local region information at each layer. Meanwhile, MGN [43] used a multi-branch network; each branch takes responsibility for extracting coarse-to-fine information. In [10], the authors proposed to accurately combine human parts and the coarse non-human parts with a self-attention mechanism. SPReID [16] was proposed to integrate semantic human parsing to person ReID with local visual cues of human parts. Song et al. [37] combined binary segmentation mask with a mask-guided attention model for the person ReID task.

Attribute-based approach. Many studies utilize attribute information to improve the representation of local features in the person ReID task. In [21], the authors proposed an attribute-person recognition (APR) network which learns person ReID and person attribute recognition simultaneously. A^3M [11] introduced a new attention mechanism by learning attribute-guide attention and category-guided attention reciprocally. In [22], the authors proposed a multi-task learning framework with four subtasks for person ReID. In [25], the authors proposed AFFNet, which is a multi-branch model that fuses features from person identity branch and attribute branch. In [52], the authors proposed a multi-branch model levering both identity label and attribute information. The AANet [39] combined the global representation with three tasks, including person ReID, body part localization, and person attribute recognition. The APDR [19] method fused attribute features and body part features to result in the final local features which are then concatenated with the global features for person re-identification.

Other approaches. Deep CNN has been used in various tasks in combining vision, language, and scene attributes [47, 28, 27, 4, 6, 29]. In [55], the authors proposed a pyramid model that can match images at different scales by incorporating local and global information and the gradual cues between them. Considering the distance part-to-part relationship, in [47], the authors proposed an attention mechanism to capture non-local and local correlations directly via second-order feature statistics. Inspired by GAN, Zheng et al. [57] proposed a joint learning framework that couples ReID learning and data generation in an end-to-end manner. More recently, in [41] the authors proposed a two-stream spatial-temporal person ReID (st-ReID) framework that uses both visual semantic information and spatial-temporal information from the camera setting, thus eliminates lots of appearance ambiguity images. Zhou et al.[59] proposed a online joint multi-metric adaptation algorithm which not only takes individual characteristics of testing samples into consideration but also fully utilizes the visual similarity relationships among both query and gallery samples. In [4], the authors proposed the Salience-guided Cascaded Suppression Network which enables the model to mine diverse salient features and integrate these features into the final representation by a cascaded manner. In [49], Yang et al. proposed a Spatial-Temporal Graph Convolutional Network which enables to extract robust spatial-temporal information that is complementary with appearance information for video-based Person Re-identification task. The UnityStyle [23] method was proposed to smooth the style disparities within the same camera and across different cameras. Zhang et al. [54] proposed the Relation-Aware Global Attention module which captures the global structural information for better attention learning. Besides, several methods are proposed to solve the problems of Occluded Person Re-Identification [13, 26, 8, 42].

3. Methodology

The proposed framework is presented in Figure 2. We denote I is a probe person image. This probe image I is first passed through a backbone CNN to get the feature map $\mathbf{F}$. By utilizing a human parsing pretrained model, we extract the body part masks to obtain the visual features of each part. The person attributes are then represented by a lookup word embedding. Given body part features and attribute features, we construct the Graph-based Person Signature which includes attribute nodes and body part nodes conditioned on the correlation matrix. We employ the GCN [17] for reasoning on the person signature graph and encoding the graph into more representativeness features. Our proposed method is a multi-branch multi-task framework for person ReID, where the main branch performs the verification task by optimizing two well-known loss functions: Triplet loss [45] and Center loss [46]. The auxiliary branch performs reasoning on the proposed person signature graph and solves the attribute recognition as well as the person identity classification tasks. The training process is explained in detail in Section 3.3.

3.1. Graph-based Person Signature: Construction

The proposed GPS, denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, consists of nodes $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ with the total number of nodes $N_G = N_A + N_P$, where N_A is the number of attributes and N_P is the number of body parts. Each node denotes either a person attribute or a human body part and is initialized with a D_w-dims feature vector x_v. The graph is represented by an adjacency matrix $\mathbf{M} \in \mathbb{R}^{N_G \times N_G}$ containing weights associated with each edge $(v_i, v_j) \in \mathcal{E}$. The correlation matrix $\mathbf{M}$ has the following form

$$\mathbf{M} = \begin{bmatrix} \mathbf{AA} & \mathbf{AP} \\ \mathbf{PA} & \mathbf{PP} \end{bmatrix},$$

where $\mathbf{AA} \in \mathbb{R}^{N_A \times N_A}$ is the attribute-attribute correlation matrix, $\mathbf{PP} \in \mathbb{R}^{N_P \times N_P}$ is the parts-parts correlation matrix, $\mathbf{PA} \in \mathbb{R}^{N_P \times N_A}$ is the parts-attributes correlation matrix, and $\mathbf{AP} \in \mathbb{R}^{N_A \times N_P}$ is the attribute-parts correlation matrix.

The attributes-attributes matrix. We follow the process as described in [5] to construct the attributes-attributes matrix $\mathbf{AA}$. The element $\mathbf{AA}_{ij}$ denotes the probability of occurrence of attribute j when the attribute i occurs, which is formulated as follow

$$\mathbf{AA}_{ij} = \frac{L_{ij}}{K_i}, \tag{1}$$

where K_i denotes the occurrence times of attribute i in the training set, and L_{ij} denotes the co-occurence of attribute pair i and j.

The parts-parts matrix. We assume that the body parts are always recognizable for every probe image in the training set. Thus we set all elements in the $\mathbf{PP}$ matrix to 1. This can be inferred as if a body part i is recognized, the probability of recognizing the body part j is 1.

The parts-attributes matrix. The element of the matrix $\mathbf{PA}_{ij}$ denotes the probability of the attribute i occurs when the body part j is recognized. We establish a heuristic observation that some attributes only attached to a specific body, e.g., ‘hair length’ is only attached to ‘head’, not ‘lower body’. The body parts and their associate attributes are summarized in Table 1. Based on the recognizable body parts assumption above, given body part i, $\mathbf{PA}_{ij} = 0$ if the attribute j is not attached to the body part i, otherwise $\mathbf{PA}_{ij} = k_j$ where k_j is the percentage of attribute j occurs in the dataset.

The attributes-parts matrix. The element of the matrix $\mathbf{AP}_{ij}$ denotes the probability of the body part i is recognized while the attribute j occurs. Due to our assumption that all body parts are recognizable in the dataset. $\mathbf{AP}_{ij} = 1$ if attribute j is attached to body part i, otherwise, $\mathbf{AP}_{ij} = 0$.

In practice, the attributes are represented by word embedding $\mathbf{Z} \in \mathbb{R}^{N_A \times D_w}$, where N_A is the number of attributes and D_w is the dimensionality of word-embedding vector.

Body parts representation. To obtain the visual presentation of person body part, we utilize the state-of-the-art SCHP pretrained model [18] trained on LIP dataset [20] to predict the body part masks for all the images in advance. Although LIP dataset has 20 labels, in our work, we combine the labels to form the more coarse body part regions, namely *head, upper, lower, arm, and foreground.* Note that the *foreground* is the combination of other body parts, which represents the global attributes of a person such as *age* and *gender*. The parser segments each probe image into N_P body parts represented by a set of masks $\mathcal{H} = \{\mathbf{H}_k\}_{k=1}^{N_P}$, where $\mathbf{H}_k$ is a binary mask with the same size as the probe image. Each mask $\mathbf{H}_k$ is scaled to the same size as feature map $\mathbf{F}$ and is applied L_1 normalization, which results in $\mathbf{H}'_k$. The feature map $\mathbf{F} \in \mathbb{R}^{W \times H \times D}$ has $W \times H$ locations, each location i is associated with a

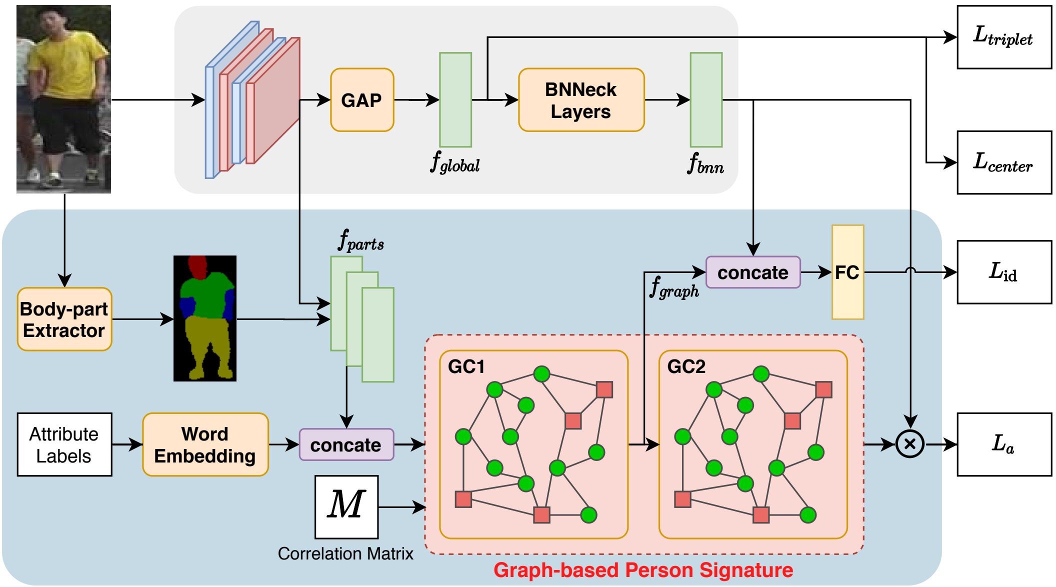

Figure 2. Illustration of our proposed framework including two branches: (1) global branch which extracts person global features; (2) GPS branch which performs reasoning the person attributes and body parts using GCN.

Table 1. Body parts and their associated attributes.

Body part	**Attribute**
Foreground	gender, age
Head	hair length, wearing hat
Upper body	upper clothing's type, upper clothing's color, carrying backpack
Lower body	lower clothing's type, lower clothing's color, lower clothing's length
Arm	sleeve length, carrying bag, carrying handbag

feature vector $\mathbf{f}_i \in \mathbb{R}^D$. The k-th body part $f_{part}^{(k)}$ is computed as below:

$$f_{part}^{(k)} = \sum_{i=1}^{N_P} h_i^{(k)} \mathbf{f}_i, \qquad (2)$$

where $h_i^{(k)}$ is the scalar value at the location i of $\mathbf{H}'_k$. The $f_{part}^{(k)}$ is projected to a D_w-dim vector.

3.2. Graph-based Person Signature: Reasoning

Generally, GCN defines a multi-layer propagation process on a graph $\mathcal{G}$. Precisely, each layer in GCN is formulated as a function $f(\mathbf{X}, \mathbf{M})$ which updates the node representations by propagating information between the input nodes $\mathbf{X} \in \mathbb{R}^{N_G \times D_w}$, where each row represents a node, under the guidance of correlation matrix $\mathbf{M}$. Denoting $\mathbf{H}^{(k)}$ is the feature matrix after passing the input nodes $\mathbf{X}$ to k-th GCN layers. We follow GCN formulation proposed in [17], which takes node features $\mathbf{H}^{(k)} \in \mathbb{R}^{N_G \times d}$ and the corresponding correlation matrix $\mathbf{M}$ as inputs and pass through a GCN layer to transform to $\mathbf{H}^{(k+1)} \in \mathbb{R}^{N_G \times d'}$. According to [17], every GCN layer can be represented as

$$\mathbf{H}^{(k+1)} = \text{LeakyReLU}(\hat{\mathbf{M}}\mathbf{H}^{(k)}\Theta^{(k)}), \qquad (3)$$

where $\Theta^{(k)} \in \mathbb{R}^{d \times d'}$ is a layer-specific trainable weight matrix and $\hat{\mathbf{M}}$ is the normalized version of correlation matrix $\mathbf{M}$. Formally, $\hat{\mathbf{M}}$ is defined as:

$$\hat{\mathbf{M}} = (\mathbf{I} + \mathbf{D})^{-\frac{1}{2}}(\mathbf{M} + \mathbf{I})(\mathbf{I} + \mathbf{D})^{-\frac{1}{2}}, \qquad (4)$$

where $\mathbf{D}$ is the diagonal degree matrix of $\mathbf{M}$, the identity matrix $\mathbf{I} \in \mathbb{R}^{N_G \times N_G}$ is added for forcing the self-loop in $\mathcal{G}$. We aim to learn a set of parameters $\Theta = \{\theta^1, \theta^2, ..., \theta^k\}$ that maps $\mathbf{X}$ to a set of inter-dependent classifier for person multi-attribute recognition.

3.3. Training

3.3.1 Attributes recognition

In training process, we pass the feature map $\mathbf{F}$ to the global average pooling (GAP) layer to get the global feature $\mathbf{f}_{global}$ which is then passed through a BNNeck layer [24] to get $\mathbf{f}_{bnn}$. For graph reasoning, the input node features $\mathbf{X}$ is fed to stacked of k-GCN layers and output node matrix $\mathbf{H}^{(k)} \in \mathbb{R}^{N_G \times D}$. We get a subset nodes $\mathcal{V}_a \subset \mathcal{V}$, where $v_a \in \mathcal{V}_a$ is considered as attribute nodes, and the corresponding output node features of $\mathcal{V}_a$ are stacked to form $\mathbf{W} \in \mathbb{R}^{N_A \times D}$ which is used to parameterized the multiattributes classifier $\mathcal{C}_A$. The feature $\mathbf{f}_{bnn}$ is passed throught the classifier $\mathcal{C}_A$ to get the attribute prediction $\hat{\mathbf{y}}$

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{f}_{bnn}, \tag{5}$$

Given $\hat{\mathbf{y}} \in \mathbb{R}^{N_A}$, prediction score of attribute c is indicated as $\hat{y}_c$. We denote the ground-truth label of an image is $\mathbf{y} \in \{0,1\}^{N_A}$, where y_c indicate whether attribute c appears in the image or not. The cross-entropy loss function is adopted for multi-label recognition. Let $L_a^{(i)}$ be the attribute loss for probe image $I^{(i)}$, which is computed as

$$\mathcal{L}_a^{(i)} = -\frac{1}{N_A}\sum_{c=1}^{N_A} y_c^{(i)} \log(\sigma(\hat{y}_c^{(i)})) + (1 - y_c^{(i)}) \log(1 - \sigma(\hat{y}_c^{(i)})), \tag{6}$$

where $\sigma(.)$ is the sigmoid function. The attribute loss for whole training set is computed as

$$\mathcal{L}_a = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_a^{(i)}, \tag{7}$$

where N is the number of samples in the training set. By optimizing the attribute recognition loss function L_a, the network implicitly models the correlation between person attributes and their associated visual body parts. However, the correlation between attributes and body parts is weakly linked here because it is not aware of the person identity information. To leverage all the information, we add the identity classification objective to the framework.

3.3.2 Person identity classification

The purpose of identity classification is to discriminate the features of a person among all other identities. However, in the person ReID problem, the number of identities is significant. At the same time, the images of each identity are also varied due to the factor of variations (e.g., environment, camera viewpoint, pose, etc.). This intra-variation is a critical challenge for a person ReID system.

However, the attributes of one person do not change significantly when a probe image is captured from different cameras or poses. In GPS, the body part nodes are obtained from the feature map of the probe image; thus, they retain the given person's visual representation. The visual information is propagated to other nodes in the graph after passing through several GCN layers. Denoting the node features of graph $\mathcal{G}$ after passing through l GCN layers is $\mathbf{H}^{(l)}$. We map the whole graph into a graph feature $\mathbf{f}_{graph} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}^{(i)}$, where $\mathbf{h}^{(i)}$ is the representation of node i. By doing that, the graph features $\mathbf{f}_{graph}$ not only can represent the visual information as well as the semantic representation of the person (i.e., correlation of attributes) but also is robust to the addressed variations.

The graph features are then concatenated with person global features $\mathbf{f}_{bnn}$, the resulted features is used for identity classification. The identity prediction logits are computed as follow

$$\mathbf{p} = \mathrm{softmax}(\mathrm{FC}(\mathbf{f}_{bnn} \odot \mathbf{f}_{graph})), \tag{8}$$

where $\odot$ denotes the concatenate operation. Let $\mathbf{q}^{(i)}$ be the one-hot vector indicating the ground-truth identity and $\mathbf{p}^{(i)}$ is the identity prediction logits of image i. We use the Cross-entropy loss as folow

$$\mathcal{L}_{id} = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{q}^{(i)} \log \mathbf{p}^{(i)} \tag{9}$$

3.3.3 Multi-task loss

The network in Figure 2 is trained end-to-end using the following multi-task loss function

$$\mathcal{L} = \alpha_1\mathcal{L}_{id} + \alpha_2\mathcal{L}_{triplet} + \alpha_3\mathcal{L}_{center} + \alpha_4\mathcal{L}_a \tag{10}$$

where L_{id} is the identity loss (9), $L_{triplet}$ is the triplet loss [45], L_{center} is the center loss [46], and L_a is attribute recognition loss (7). Since attribute recognition and person identity classification use global features as input, the Triplet loss and Center loss are used to improve the representativeness of global features and generalize well to an unseen person in the test set.

4. Experiments

4.1. Experimental Setup

Implementation. We integrate our GPS into the recent work BoT [24] as the strong baseline. We employ the ResNet-50 [12] pre-trained on ImageNet as the backbone network in all experiments. To enhance the discriminating power of the backbone, we integrate non-local attention (NLA) [44] into each ResNet block. For each probe image, we resize them into 256 × 128 and pad the resized image 10 pixels with zero values. After that, we randomly crop them into a 256 × 128. For the data augmentation, similar to [24, 40, 2], we use random horizontal flipping and erasing with the probability of 0.5 for both methods. Attribute labels are transformed into N_A × 300 word embedding. Note that N_A corresponds to the number of attributes of the dataset, i.e., 30 and 23 for Market1501 [56] and DukeMTMC-ReID [34] dataset, respectively.

Dataset. To evaluate our proposed method, we conduct our experiments on two large-scale attribute person ReID datasets which are Market-1501 [56] and DukeMTMC [34]. We follow the standard train/test split of each dataset in our experiments.

Table 2. The contribution of losses to the performance of person ReID task on Market1501 [56] dataset. Note that the experiments are conducted with ResNet-50 [12] as backbone CNN network.

$\mathcal{L}_{\mathbf{id}}$	$\mathcal{L}_{\mathbf{triplet}}$	$\mathcal{L}_{\mathbf{center}}$	**without $\mathcal{L}_{\mathbf{a}}$**		**with $\mathcal{L}_{\mathbf{a}}$**	
			mAP	R-1	mAP	R-1
✓			85.5	94.0	87.0	95.1
✓	✓		87.1	94.7	87.6	95.2
✓	✓	✓	87.5	94.9	87.8	95.2

Table 3. The transferable ability of our GPS evaluated on cross-dataset

Models	Market-1501 → DukeMTMC-ReID		DukeMTMC-ReID → Market-1501	
	mAP	R-1	mAP	R-1
BoT [24]	14.6	27.6	21.6	48.6
GPS (our)	21.9	37.0	24.7	52.1

Evaluation. To evaluate the person ReID performance of our GPS and to compare the results with the state-of-the-art methods, we report standard ReID metrics: Cumulative Matching Characteristic (CMC) (as R-1, R-5, and R-10) and mean Average Precision (mAP). Note that, as [21], we ignore the distractor and junks images which are not labelled attributes.

4.2. GPS Analysis

Loss Contribution. In Table 2, we show the contribution of each loss to the final performance on the Market1501 dataset. The person ID classification loss, triplet loss, center loss, and attribute recognition loss are denoted as $\mathcal{L}_{\mathbf{id}}$, $\mathcal{L}_{\mathbf{triplet}}$, $\mathcal{L}_{\mathbf{center}}$, and $\mathcal{L}_{\mathbf{a}}$, respectively. The performance is improved when we incorporate all losses to the framework, which justifies the effectiveness of our proposed method. By using only $\mathcal{L}_{\mathbf{id}}$, we still achieve comparative results with other mask-guided and attribute-based methods. While the triplet loss $\mathcal{L}_{\mathbf{triplet}}$ demonstrates its capability on improving the performance, the center loss $\mathcal{L}_{\mathbf{center}}$ shows a slight impact on the performance. Notably, the attribute loss $\mathcal{L}_{\mathbf{a}}$ shows stability when being incorporated with other loss functions.

Model Interpretability. In this section, we conduct cross-dataset experiments to evaluate the effectiveness of GPS. The model is trained on the source dataset and test directly on the target dataset without finetuning. As shown in Table 3, our GPS archives a significant improvement over the Bag-of-Tricks baseline [24]. This demonstrates the interpretability of our proposed method as well as confirms the effectiveness of learning the attributes for the person ReID task.

Training Parameters. We also provide the number of training parameters of our GPS and the baseline BoT [24] in Table 4 to show the complexity of each method. Overall, our GPS slightly increases about 3M parameters in comparison with the baseline BoT while achieving much better performance.

4.3. Comparison to the State of the Art

Mask-guided and Attribute-based Methods. We compare our method (GPS) to the recent state-of-the-art that used body parts: MGCAM [37], SPReID [16], P^2-Net [10]. For attribute-based approach, we compare our results with ACRN [35], MLFN [1], A^{3}M [11], AANet [39], APR [21], AFFNet [25], PAAN [52], and APDR [19]. Among them, AANet [39], PAAN [52], and APDR [19] are works that use both attributes and body parts to enhance the performance of person ReID task. However, there is no work that leverage the relationship between attributes and body parts to extract person signature embedding as our proposed method.

Other Approaches. We also compare our method with other ReID approaches, including global-based approach: SVDNet [38], TriNet [14]; stripes-based approach: Pyramid [55], Auto-ReID [33], GCP [37]; attention-based approach: Mancs [40], SONA^{2+3} [47], SCAL [2]; GAN-based approach: PN-GAN [32], FD-GAN [9], DG-Net [57]; graph-based approach: SGGNN [36]; spatial-temporal-based approach: st-ReID [41]; other approaches: CAMA [50], DSA [53], FPR [13], SAN [15]. No post-processing such as re-ranking [58] or multi-query fusion [56] is applied to our method.

GPS vs. Baseline. The last two rows of Table 5 show the result of our GPS when being integrated into the baseline BoT. The results clearly show that our GPS significantly improves the performance of BoT in both Market-1501 and DukeMTMC-ReID dataset. This demonstrates the effectiveness of our GPS and confirms the usefulness of learning the attributes in the ReID task.

Evaluation on Market-1501. We evaluate our GPS with other methods on Market-1501 dataset in Table 5. The results show that our method outperforms the state-of-the-art attribute-based methods [39] that use attribute and body part information in all evaluation metrics. Specifically, we outperforms AANet [39] by 5.3% and 1.3% at mAP and R-1, respectively. Our GPS also outperforms the state-of-the-art mask-guided methods, and especially, we outperform

Table 4. The number of parameters of our GPS in comparision with the baseline BoT [24] on Market1501 and DukeMTMC-ReID datasets using ResNet-50 [12] as the backbone network. #nParam indicates the number of parameters and 1K=1000.

Models	DukeMTMC-ReID	Market1501
	#nParam (K)	#nParam (K)
BoT [24]	25,668	25,829
GPS (our)	28,866	28,715

Table 5. Comparison with state-of-the-art methods on Market-1501 [56] and DukeMTMC-ReID [34] datasets. The cyan and yellow boxes are the best results corresponding to mask-guided/attribute-based and other approaches, respectively. Note that no post-processing is applied to our method.

Approach	**Method**	**Market1501**				**DukeMTMC-ReID**			
		mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Others	SVDNet [38]	62.1	82.3	92.3	95.2	56.8	76.7	86.4	89.9
	TriNet [14]	69.1	84.9	94.2	-	-	-	-	-
	AGW-att [51]	86.9	94.9	-	-	77.6	87.5	-	-
	Pyramid [55]	88.2	95.7	98.4	99.0	79.0	89.0	-	-
	HPM [7]	-	-	-	-	74.3	86.6	-	-
	Auto-ReID [33]	85.1	94.5	-	-	-	-	-	-
	GCP [31]	88.9	95.2	-	-	78.6	89.7	-	-
	Mancs [40]	82.3	93.1	-	-	71.8	84.9	-	-
	SONA^{2+3} [47]	88.8	95.6	98.5	99.2	78.3	89.4	95.4	96.6
	SCAL [2]	**89.3**	95.8	98.7	-	78.4	88.6	-	-
	PN-GAN [32]	72.6	89.4	-	-	53.2	73.6	-	88.8
	FD-GAN [9]	77.7	90.5	-	-	64.5	80.0	-	-
	DG-Net [57]	86.0	94.8	-	-	74.8	86.6	-	-
	SGGNN [36]	82.1	92.3	-	-	68.2	81.1	-	-
	st-ReID [41]	87.6	**98.1**	**99.3**	**99.6**	**83.9**	**94.4**	**97.4**	**98.2**
	CAMA [50]	84.5	94.7	-	-	72.9	85.8	-	-
	DSA [53]	87.6	95.7	-	-	74.3	86.2	-	-
	FPR [13]	86.6	95.4	-	-	78.4	88.6	-	-
	SAN [15]	88.0	96.1	-	-	75.5	87.9	-	-
Mask-guided & Attribute-based	MGCAM [37]	74.3	83.8	-	-	-	-	-	-
	SPReID [16]	81.3	92.5	97.2	98.1	71.0	84.4	91.9	93.7
	P^2-Net [10]	85.6	**95.2**	98.2	99.1	73.1	86.5	93.1	95.0
	ACRN [35]	62.6	83.6	92.6	95.3	52.0	72.6	84.8	88.9
	MLFN [1]	74.3	90.0	-	-	62.8	81.0	-	-
	A^{3}M [11]	69.0	86.5	95.2	97.0	-	-	-	-
	AANet [39]	82.5	93.9	-	98.6	72.6	86.4	-	-
	APR [21]	66.9	87.0	95.1	96.4	55.6	73.9	-	-
	AFFNet [25]	81.7	93.7	-	-	70.7	84.6	-	-
	PAAN [52]	77.6	92.4	-	-	65.5	82.6	-	-
	APDR [19]	80.1	93.1	97.2	98.2	69.7	84.3	92.4	94.7
	BoT [24]	85.9	94.5	-	-	76.4	86.4	-	-
	GPS (ours)	**87.8**	**95.2**	**98.4**	**99.1**	**78.7**	**88.2**	**95.2**	**96.7**

P^2-Net [10] by 2.2% at mAP. At the same time, we also get comparative results when comparing with other recent ReID approaches.

Evaluation on DukeMTMC-ReID. Table 5 also summaries the results of our GPS and other methods on DukeMTMC-ReID dataset. Our GPS significantly outperforms other attribute-based methods in all metrics. Specifically, our method outperforms the recent state-of-the-art attribute-based method AANet [39] by 6.1% at mAP and 1.8% at R-1. In addition, we also outperforms ADPR [19] by 9.0%, 3.9%, 2.8%, 2.0% at mAP, R-1, R-5, R-10, respectively. Moreover, our GPS outperforms the state-of-

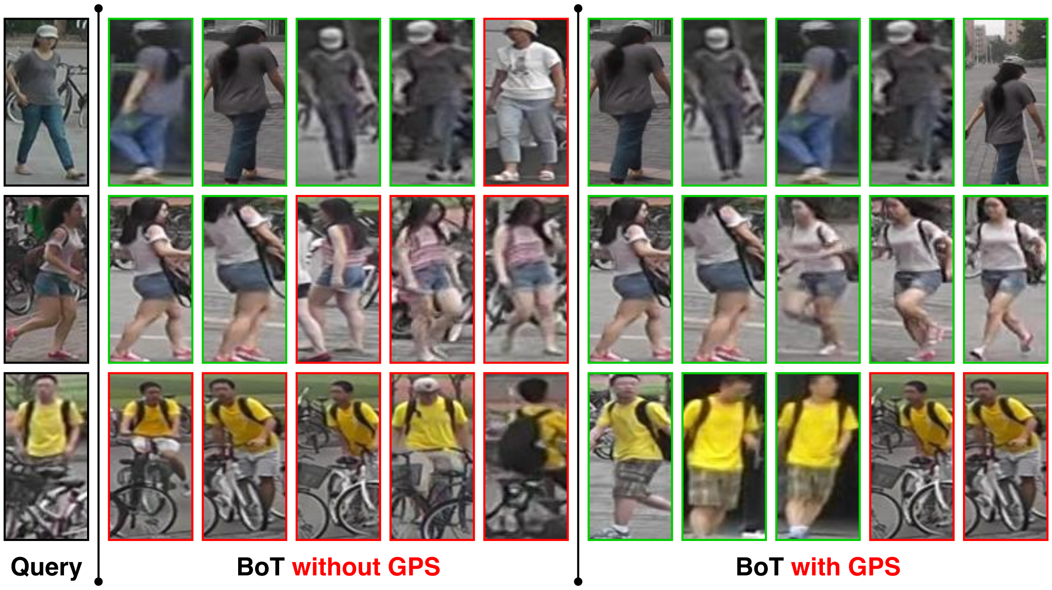

Figure 3. Top 5 retrieval results of some queries on Market-1501 dataset [56]. Note that the green/red boxes denote true/false retrieval results, respectively.

the-art mask-guided method P^2-Net [10] by 4.9%, 1.7%, 2.1%, 1.7% at mAP, R-1, R-5, R-10, respectively. Besides, we also achieve comparative results with other ReID approaches.

Attributed-based and Mask-guided vs. Other approaches. From Table 5, we notice that although our GPS shows a definite improvement over mask-guided and attributed-based methods, it achieves competitive results with methods from other approaches and particularly being outperformed by st-ReID method [41]. Note that the results of st-ReID also completely dominate all methods from all other approaches. The effectiveness of st-ReID comes from the fact that it also uses the spatial-temporal information (i.e., the spatial map of camera setting and temporal information from video timestamp) into the network. This extra information allows the network to encode the person identity from multiple viewpoints, which significantly reduces the effect of different poses, viewpoints, or ambiguity challenges. From experiments, we have observed that our GPS, as well as other attribute-based and mask-guided methods, suffers from the fact that the pretrained body part network cannot provide adequate segmentation masks, so the retrieval results are also affected.

We present some retrieval examples with five retrieved images for each query in Figure 3. As in the visualization, our GPS obtained better retrieval results than the baseline. In the first row of Figure 3, the baseline gets the false retrieval result at Rank-5 due to the similarity of gender, wearing a hat, etc., except the color of the clothes. By leveraging our GPS, the extracted features are more robust to attribute and body part information, then, lead to better retrieval results for ReID model. In the second row, the model with our GPS gives better results by extracting more information about the relationship between 'backpack' attribute and this person identity, thereby eliminating false cases. We also show an example that our GPS does not yet produce entirely correct retrieval results in the third line of the Figure 3. In this case, the lower body of the probe image is partly covered by the bicycle. Thus, the extracted features (i.e., the color of the pants) are not fully captured, which results in the feature misalignment between the probe image and retrieval results.

To conclude, our experiment results demonstrate that our GPS has successfully encoded two sources of local information (i.e., attributes and body parts) and global features, as well as modeling the correlations between them to create a visual signature of person identity. In the future, we would like to combine our approach and spatial-temporal information as in [41] to further improve the results.

5. Conclusion

This paper proposes Graph-based Person Signature (GPS) that effectively captures the dependencies of person attributes and body parts information. We utilize the GCN on the GPS to propagate the information among nodes in the graph and integrate the graph features into a novel multi-branch multi-task network. The experimental results on benchmark datasets confirm the effectiveness of our GPS and demonstrate that our GPS performs better than recent state-of-the-art attribute-based and mask-guided ReID methods.

References

[1] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018. 1, 6, 7

[2] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *ICCV*, 2019. 1, 5, 6, 7

[3] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017. 1

[4] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *CVPR*, 2020. 2

[5] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, 2019. 3

[6] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, 2018. 2

[7] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019. 7

[8] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *CVPR*, 2020. 3

[9] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NIPS*, 2018. 1, 6, 7

[10] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, 2019. 2, 6, 7, 8

[11] Kai Han, Jianyuan Guo, Chao Zhang, and Mingjian Zhu. Attribute-aware attention model for fine-grained representation learning. In *ACMMM*, 2018. 1, 2, 6, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6, 7

[13] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *ICCV*, 2019. 3, 6, 7

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 6, 7

[15] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, 2020. 6, 7

[16] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 2, 6, 7

[17] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017. 3, 4

[18] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *arXiv preprint arXiv:1910.09777*, 2019. 3

[19] Shuzhao Li, Huimin Yu, and Roland Hu. Attributes-aided part detection and refinement for person re-identification. *Pattern Recognition*, 97:107016, 2020. 1, 2, 6, 7

[20] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *TPAMI*, 41(4):871–885, 2018. 3

[21] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019. 1, 2, 6, 7

[22] Hefei Ling, Ziyang Wang, Ping Li, Yuxuan Shi, Jiazhong Chen, and Fuhao Zou. Improving person re-identification by multi-task learning. *Neurocomputing*, 347:109–118, 2019. 2

[23] Chong Liu, Xiaojun Chang, and Yi-Dong Shen. Unity style transfer for person re-identification. In *CVPR*, 2020. 3

[24] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, 2019. 2, 4, 5, 6, 7

[25] Jinghao Luo, Yaohua Liu, Changxin Gao, and Nong Sang. Learning what and where from attributes to improve person re-identification. In *ICIP*, 2019. 1, 2, 6, 7

[26] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, 2019. 3

[27] Anh Nguyen, Thanh-Toan Do, Ian Reid, Darwin G Caldwell, and Nikos G Tsagarakis. V2cnet: A deep learning framework to translate videos to commands for robotic manipulation. *arXiv preprint arXiv:1903.10869*, 2019. 1, 2

[28] Anh Nguyen, Ngoc Nguyen, Kim Tran, Erman Tjiputra, and Quang D Tran. Autonomous navigation in complex environments with deep multimodal fusion network. In *IROS*, 2020. 2

[29] Anh Nguyen, Quang D Tran, Thanh-Toan Do, Ian Reid, Darwin G Caldwell, and Nikos G Tsagarakis. Object captioning and retrieval with natural language. In *CVPRW*, 2019. 2

[30] Binh X Nguyen, Binh D Nguyen, Gustavo Carneiro, Erman Tjiputra, Quang D Tran, and Thanh-Toan Do. Deep metric learning meets deep clustering: An novel unsupervised approach for feature embedding. In *BMVC*, 2020. 1

[31] Hyunjong Park and Bumsub Ham. Relation network for person re-identification. In *AAAI*, 2020. 7

[32] 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 *ECCV*, 2018. 1, 6, 7

[33] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *ICCV*, 2019. 6, 7

[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 *ECCV*, 2016. 5, 7

[35] Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *CVPRW*, 2017. 1, 6, 7

[36] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, 2018. 6, 7

[37] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 2, 6, 7

[38] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017. 6, 7

[39] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, 2019. 1, 2, 6, 7

[40] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018. 1, 5, 6, 7

[41] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. In *AAAI*, 2019. 1, 2, 6, 7, 8

[42] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*, 2020. 3

[43] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACMMM*, 2018. 2

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 5

[45] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006. 3, 5

[46] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 3, 5

[47] Bryan Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *ICCV*, 2019. 1, 2, 6, 7

[48] Jiwei Yang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. Local convolutional neural networks for person re-identification. In *ACMMM*, 2018. 2

[49] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *CVPR*, 2020. 3

[50] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR*, 2019. 6, 7

[51] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv*, 2020. 7

[52] Yan Zhang, Xusheng Gu, Jun Tang, Ke Cheng, and Shoubiao Tan. Part-based attribute-aware network for person re-identification. *IEEE Access*, 7:53585–53595, 2019. 1, 2, 6, 7

[53] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019. 2, 6, 7

[54] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, 2020. 3

[55] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, 2019. 2, 6, 7

[56] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2, 5, 6, 7, 8

[57] 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. 1, 2, 6, 7

[58] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 6

[59] Jiahuan Zhou, Bing Su, and Ying Wu. Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification. In *CVPR*, 2020. 2