Explainable Person Re-Identification with Attribute-guided Metric Distillation

Xiaodong Chen[1]* Xinchen Liu[2] Wu Liu[2]† Xiao-Ping Zhang[3] Yongdong Zhang[1] Tao Mei[2]
[1]University of Science and Technology of China, Hefei, China
[2]JD AI Research, Beijing, China [3]Ryerson University, Toronto, Canada
cxd1230@mail.ustc.edu.cn, liuxinchen1@jd.com, liuwu@live.cn, xzhang@ee.ryerson.ca, zyd73@ustc.edu.cn, tmei@live.com

Abstract

Despite the great progress of person re-identification (ReID) with the adoption of Convolutional Neural Networks, current ReID models are opaque and only outputs a scalar distance between two persons. There are few methods providing users semantically understandable explanations for why two persons are the same one or not. In this paper, we propose a post-hoc method, named Attribute-guided Metric Distillation (AMD), to explain existing ReID models. This is the first method to explore attributes to answer: 1) what and where the attributes make two persons different, and 2) how much each attribute contributes to the difference. In AMD, we design a pluggable interpreter network for target models to generate quantitative contributions of attributes and visualize accurate attention maps of the most discriminative attributes. To achieve this goal, we propose a metric distillation loss by which the interpreter learns to decompose the distance of two persons into components of attributes with knowledge distilled from the target model. Moreover, we propose an attribute prior loss to make the interpreter generate attribute-guided attention maps and to eliminate biases caused by the imbalanced distribution of attributes. This loss can guide the interpreter to focus on the exclusive and discriminative attributes rather than the large-area but common attributes of two persons. Comprehensive experiments show that the interpreter can generate effective and intuitive explanations for varied models and generalize well under cross-domain settings. As a by-product, the accuracy of target models can be further improved with our interpreter. [1]

1. Introduction

Person Re-identification (ReID), i.e., retrieval of the same person captured by multiple cameras, has attracted tremendous attention from academia and industry [19, 30, 32, 34, 41, 42]. Although Convolutional Neural Networks (CNNs) have significantly improved the accuracy of person ReID, we still cannot completely trust the results produced by black-box models, especially for critical scenarios [43]. Therefore, this paper is focused on the interpretation of CNN-based person ReID models which is crucial yet rarely studied.

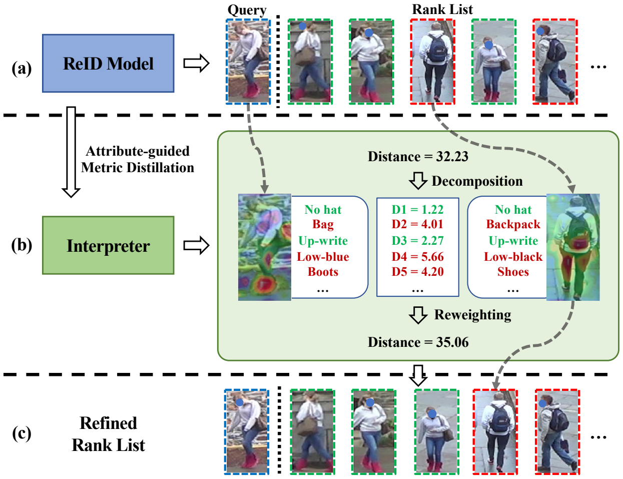

Figure 1. The motivation of attribute-guided metric distillation. (a) Given a query, the ReID model returns a rank list of gallery images based on pairwise metrics. (b) The learned Interpreter can visualize intuitive attention maps of attributes to tell users what attributes make two persons different, and generate contributions of attributes to reflect the impact of each attribute. (c) Refined results by re-weighted distances from Interpreter. (Best viewed in color.)

In recent years, there has been a surge of work in discovering how a target CNN processes input images and makes predictions [10, 24, 46]. These methods usually visualize gradients or salient regions on feature maps w.r.t. the input image and its prediction [4, 8, 24, 25, 26, 46]. Particularly, Chen *et al*. [5] proposed to explain neural networks semantically and quantitatively by decomposing the prediction made by CNNs into semantic concepts by knowledge distillation. However, these methods mainly consider clas-

*This work was done when Xiaodong Chen was an intern at JD AI Research.
†Wu Liu is the corresponding author
[1] See the project on www.xiaodongchen.cn/AMD.github.io/

sification problems. They cannot be directly applied to person ReID, which is an open-set retrieval task and usually solved by metric learning [42, 45].

A CNN-based ReID system usually maps a query image and gallery images into a metric space, then outputs pairwise distances by which a rank list of gallery images is returned, as shown in Figure 1 (a). Although Yang *et al.* [36] proposed Ranking Activation Maps which could visualize related regions of two persons, it still cannot semantically explain why they are similar or not. Attributes, e.g., colors and types of clothes, shoes, etc., are semantically understandable for humans and have been exploited as mid-level features for person ReID [18], but there is no method using attributes for explanations of person ReID. Therefore, we aim to learn an interpreter with the help of semantic attributes for answering two questions: 1) what attributes make two persons different, and 2) how much impact each attribute contributes to the difference, as shown in Figure 1 (b). In real applications, the interpreter not only can help users focus on the most discrepant attributes of two persons but also can assist developers to improve the accuracy of ReID models, as shown in Figure 1 (c).

However, interpretation of ReID models with attributes faces unique challenges. Firstly, since the output of ReID models are distances of pairwise images, it is difficult to use class activation or gradients to visualize salient regions or disentangle semantics as classification [24, 46]. Moreover, persons in the wild can be described by various fine-grained and imbalanced attributes [7, 15, 18], which may bring biases to ReID results as well as the explanations. For example, attributes with large areas, such as coats and pants, always overwhelm small but discriminative ones like hats and shoes. Furthermore, there are only weakly-annotated image-level or ID-level attribute labels without accurate bounding boxes or masks [18], which makes it hard to learn accurate locations and intuitive visualizations for attributes.

To this end, we propose a post-hoc method, named Attribute-guided Metric Distillation, which explores semantic attributes towards explainable person ReID. Specifically, we design a pluggable interpreter network to utilize the knowledge from a target ReID model by metric distillation. The interpreter is grafted on the target ReID model and directly adopts the parameters of the first several CNN stages to exploit the low-level and mid-level features of the target model. The rest layers of the interpreter are equipped with an attribute decomposition head, by which the interpreter can learn to generate a set of attribute-guided attention maps (AAMs) for a pair of input person images. On the one hand, the generated AAMs can be directly used to visualize discriminative attributes for the image pair. On the other hand, the AAMs can be applied to the visual features of the image pair from the target model. By this means, their features and distance can be decomposed into attribute-guided components to quantify the contribution of each attribute to the overall distance. Thus, the interpreter not only can output the quantitative contributions of attributes to the overall distance of two persons but also can generate intuitively visualizations of attributes for users.

To guide learning of the interpreter, we design two loss functions. One is a metric distillation loss which can guarantee the consistency between two distance metrics: 1) the decomposed attribute-guided distances from the interpreter, and 2) the overall distance from the target model. The other is the attribute prior loss. It is designed based on the observation that the difference between two persons mainly comes from the exclusive attributes rather than common ones. Thus, the attribute prior loss makes the interpreter pay more attention to exclusive but discriminative attributes of two persons with only weakly-labeled attributes of persons.

The contributions of this paper are three-fold:

- This is one of the first attempt toward explainable person ReID by attribute-guided metric distillation that can semantically and quantitatively explain the results of existing ReID models;
- We design a pluggable interpreter network with an attribute decomposition head to obtain contributions of attributes to the difference of two persons and generate intuitive visualizations for target ReID models;
- To guide the learning of the interpreter, the metric distillation loss and attribute prior loss are proposed to guarantee consistency during metric distillation and prevent biases of attributes.

To show the effectiveness and compatibility of the interpreter, we apply it to the state-of-the-art ReID models on different datasets with comprehensive experiments. As a by-product, the performance of the state-of-the-art models is further improved with our interpreter.

2. Related Work

Interpretation of predictions made by CNNs. Recent studies on post-hoc methods for the interpretation of CNNs usually adopt visualization of salience maps [24, 25, 46], perturbation of input images [9, 33], decision trees [38], knowledge distillation [5], etc. For example, Simonyan *et al.* [25] first proposed two techniques based on computing the gradient of the class score with respect to the input. Zhou *et al.* [46] proposed the class activation mapping (CAM) to map the predicted class score back to the convolutional layers in CNNs, which could provide an intuitive way to highlight the discriminative regions for a specific class. Zhang *et al.* [38] learned a decision tree to clarify the reasons for predictions made by a CNN via estimating the contributions of object parts. Most recently, Chen *et al.* [5] proposed to semantically and quantitatively explain CNNs

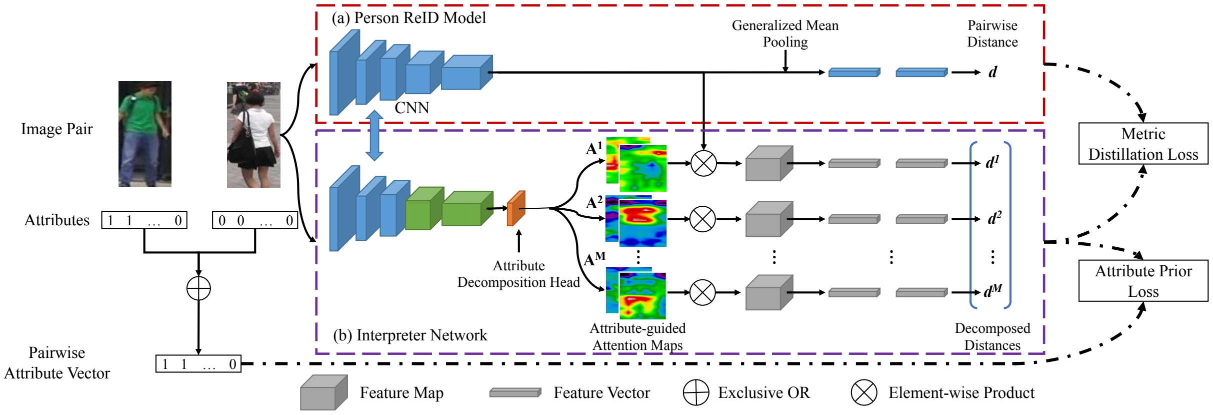

Figure 2. The overall architecture of the attribute-guided metric distillation framework for person ReID. (a) The target ReID model that generates the pairwise distance for an image pair. (b) The interpret network that learns to decompose the pairwise distance into components of attributes and generates attention-guided attention maps for individual attributes. (Best viewed in color.)

by knowledge distillation, which can decompose the prediction into contributions of a group of semantic concepts. Although these methods cannot be directly applied to the interpretation of person ReID models, they inspire us to design a semantic interpreter based on a set of attributes for explainable person ReID.

Explainable Person Re-identification. Recent CNN-based person ReID methods concentrate on two aspects: 1) task-specific modules to learn discriminative features [3, 40] and 2) metric-based loss functions to make the features of different persons more separable in the latent space [13, 27, 31, 44]. Among these methods, various attention modules inspired by the vision system of humans achieve significant performance while making the models more explainable [2, 17, 36, 39]. In particular, Yang *et al*. [36] adopted CAM [46] to discover rich features for person ReID and proposed Ranking Activation Maps (RAMs) to visualize the salient regions based on the similarity a pair of person images. However, RAMs only provide rough and pixel-level visualization on images without any semantic explanations. Therefore, we aims to build a semantic and quantitative interpreter for person ReID models.

3. Attribute-guided Metric Distillation

3.1. Preliminary

This subsection first declares necessary notations and definitions for person ReID. The task of person ReID is, given a query person image, to find images of the same person in a gallery set captured by multiple cameras [16, 42]. CNN-based ReID methods usually follow a paradigm. **1)** We have a dataset $S = \{(x_i, y_i)\}^N$ where x_i and y_i are an image and the ID of a person, and N is the number of samples, while S is divided into a training set and a testing set with non-overlapped IDs. The testing set is further split into a query set Q and a gallery set G. **2)** In the training stage, a CNN $\mathcal{F}(\cdot)$ is trained to embed the image x_i into a latent space as $\boldsymbol{f}_i = \mathcal{F}(x_i)$ by which the features $\boldsymbol{f}_i$ of the same person are close and those of different persons are distant. **3)** In the testing stage, the trained $\mathcal{F}(\cdot)$ takes a pair of images, $x_q \in Q$ and $x_g \in G$, as the input to obtain their features $(\boldsymbol{f}_q, \boldsymbol{f}_g)$ and normalized distance $d_{q,g} = \mathcal{D}(\boldsymbol{f}_q, \boldsymbol{f}_g)$ in the latent space. In this paper, Euclidean distance is used as the distance metric unless otherwise specified. By ranking the distances between a query and all images in G, the most similar one can be matched. Through the above paradigm, the prediction of a ReID system is just $d_{q,g}$ for (x_q, x_g). Given different distance values, *e.g.* $d_{q,g} < d_{q,g'}$, the system cannot semantically and quantitatively explain why x_q and x_g are more similar than x_q and $x_{g'}$, which makes it difficult for users to understand and trust the system.

In this paper, we assume that the person ReID dataset S is weakly labeled with a set of image-level attributes to obtain a new dataset $S_A = \{(x_i, y_i, \boldsymbol{a}_i)\}^N$. For each image $x_i \in S_A$, $\boldsymbol{a}_i = (a_i^1, a_i^2, ..., a_i^M)$ is a binary vector, where a_i^k is the k-th attribute denoted by a Boolean value, and M is number of attribute classes. Our approach aims to semantically and quantitatively explain the distance $d_{i,j}$ with the weakly annotated attributes.

3.2. Attribute-guided Metric Interpreter

This subsection presents design of the interpreter in Attribute-guided Metric Distillation, as shown in Figure 2. Before that, we give the formulation of the target model.

The target person ReID model $\mathcal{F}(\cdot)$ can be an arbitrary off-the-shelf model (e.g., PCB [28], MGN [29], BOT [22], SBS [12], etc) with a CNN (e.g. ResNet [11]) as the backbone. $\mathcal{F}(\cdot)$ is trained on the training data as reviewed in Section 3.1, then we keep it fixed during learning of the interpreter network. As shown in Figure 2 (a), given a pair of images (x_i, x_j), we first extract the feature maps ($\boldsymbol{F}_i$, $\boldsymbol{F}_j$) from the last convolutional layer. Then, we obtain the feature vectors ($\boldsymbol{f}_i$, $\boldsymbol{f}_j$) by generalized mean pooling and compute the distance $d_{i,j}$ in the metric space.

The attribute-guided interpreter network $\mathcal{G}(\cdot)$ is the essential module in the AMD framework. As shown in Figure 2 (b), the interpreter network has the same structure as the target model. Since the low-level and mid-level layers of the CNN capture attribute-related features such as texture and color [1], the first several CNN stages of $\mathcal{G}(\cdot)$ and $\mathcal{F}(\cdot)$ are shared to utilize the attribute-related knowledge learned by $\mathcal{F}(\cdot)$. The high-level layers in $\mathcal{G}(\cdot)$ are learnable to generate the spatial attention maps guided by semantic attributes, which can reflect the contribution of each attribute.

An Attribute Decomposition Head (ADH) is connected after the last convolutional (conv) layer of $\mathcal{G}(\cdot)$. In particular, the ADH contains a $\frac{C}{8} \times 3 \times 3$ conv layer, a $M \times 1 \times 1$ conv layer, and an activation function $\delta(\cdot)$, where C is the channel number of the last conv layer of $\mathcal{G}(\cdot)$. Given an image pair (x_i, x_j), we obtain feature maps from the last conv layer of $\mathcal{G}(\cdot)$. Through ADH we can obtain the Attribute-guided Attention Maps (AAMs) $\boldsymbol{A}_i$ and $\boldsymbol{A}_j \in \mathbb{R}^{M\times w\times h}$. After that, $\boldsymbol{A}_i$ and $\boldsymbol{A}_j$ are sliced into M matrices by channels, i.e., $(A_i^1, A_i^2, ..., A_i^M)$ and $(A_j^1, A_j^2, ..., A_j^M)$, where A_i^k and $A_j^k \in R^{h\times w}$ are the attention maps of k-th attribute, where h and w are the height and width of the attention maps.

To this end, we can apply A_i^k and A_j^k of attribute k to the feature maps $\boldsymbol{F}_i$ and $\boldsymbol{F}_j$ from the target model by

$$\boldsymbol{F_i^k} = \boldsymbol{F_i} \circ A^k, \quad \boldsymbol{F_j^k} = \boldsymbol{F_j} \circ A^k, \tag{1}$$

where $\circ$ is the element-wise multiplication. By this means, each input image can obtain M attribute-guided feature maps in which the pixels activated by attribute k will be highlighted, while other pixels will be depressed. After that, the attribute-guided feature vectors $\boldsymbol{f}_i^k$ and $\boldsymbol{f}_j^k$ can be calculated from $\boldsymbol{F}_i^k$ and $\boldsymbol{F}_j^k$ by generalized mean pooling. Finally, similar to compute $d_{i,j}$ for (x_i, x_j), we can obtain their attribute-guided distances $(d_{i,j}^1, d_{i,j}^2, ..., d_{i,j}^M)$.

It is noteworthy that the activation function $\delta(\cdot)$ in ADH is important for the generation of AAMs. For existing attention modules in ReID methods [17, 36], the attention maps are usually normalized by a sigmoid activation function to constrain the values in $[0, 1]$. However, the sigmoid function will make attention values close to either 0 or 1, which can cause gradient vanishing and failure of model convergence. Besides, when multiplying A^k to $\boldsymbol{F}$, the large-area parts such as upper clothes and pant will dominant the attribute-guided feature vector $\boldsymbol{f}^k$ and make the interpreter learn biased representation for attribute decomposition. Therefore, we design a Positive Exponential Power Unit (PePU):

$$\delta(x) = \begin{cases} \kappa \cdot (x+1)^\tau, & x > 0, \\ \kappa \cdot e^x, & x <= 0, \end{cases} \tag{2}$$

where x is the output of the last conv layer in ADH, κ and τ are growth factor in $(0, 1)$ to smooth attention values and improve propagation of gradients. With PePU, the interpreter network can effectively decompose the salience regions for various attributes by eliminating biases of imbalanced attribute distribution.

In summary, the interpreter network aims to 1) learn the importance of each attribute for the prediction made by a target model via attribute-guided attention maps, and 2) decompose the distance of two persons into a set of attribute-guided distances based on their contributions to the distance, by which it can provide semantically and quantitatively explanations for person ReID. To achieve this goal, we elaborately design two types of loss functions in the next section for learning the interpreter by attribute-guided metric distillation.

3.3. Loss Function

To make the interpreter generate effective and reasonable explanation, we propose two types of objective functions: the metric distillation loss and the attribute prior loss.

Loss function of metric distillation. The main task of $\mathcal{G}(\cdot)$ is to decompose the distance $d_{i,j}$ given by the target model $\mathcal{F}(\cdot)$ into the contributions of attributes, which can be formulated as:

$$d_{i,j} \approx \hat{d}_{i,j} = \sum\nolimits_{k=1}^{M} d_{i,j}^k, \tag{3}$$

where M is the number of attributes, $d_{i,j}^k$ is the attribute-guided distance between x_i and x_j for attribute k, and $\hat{d}_{i,j}$ is reconstructed distance by the interpreter. Therefore, we define the metric distillation loss as

$$L_d = |d_{i,j} - \sum\nolimits_{k=1}^{M} d_{i,j}^k|. \tag{4}$$

Different from conventional knowledge distillation for classification, the metric distillation loss can guarantee the consistency between the distance metrics from the target model and the decomposed components from the interpreter.

Loss function of attribute prior. Only based on the metric distillation loss, the interpreter still cannot decompose the distance in a human understandable way. As discussed in [5], without any prior knowledge, the explainer tends to suffer from the biased representations, which makes the network tend to approximate the overall distance only by a few dominant attributes instead of discriminative attributes. To overcome this problem, we define two groups of constraints for the attribute-guided distances.

The prior constraints are based on the observation that differences of two persons are mainly caused by exclusive attributes like different belongings, rather than common attributes like similar pants. Therefore, given a pair of input images with attributes, i.e., $(x_i, y_i, \boldsymbol{a}_i)$ and $(x_j, y_j, \boldsymbol{a}_j)$, the pairwise attribute vector $\boldsymbol{a}_{i,j}$ is computed by

$$\boldsymbol{a}_{i,j} = \boldsymbol{a}_i \oplus \boldsymbol{a}_j, \tag{5}$$

where $\oplus$ is Exclusive OR. With $\boldsymbol{a}_{i,j}$, we can obtain the common attributes that both x_i and x_j contain or lack, and the exclusive attributes that only one image contains.

Based on $\boldsymbol{a}_{i,j}$, the first group of constraints are applied to the total contribution of exclusive attributes and that of common attributes, which is formulated by:

$$\sum_{e=1}^{M_E} \frac{d^e_{i,j}}{\hat{d_{i,j}}} \geq (\frac{M_E}{M})^{\upsilon}, \quad \sum_{c=1}^{M-M_E} \frac{d^c_{i,j}}{\hat{d_{i,j}}} \leq 1 - (\frac{M_E}{M})^{\upsilon}. \tag{6}$$

where υ is a factor in $(0,1)$ to regulate the proportion of exclusive attributes, $d^e_{i,j}$ is the distance of an exclusive attribute, $d^c_{i,j}$ is the distance of a comment attribute, M_E is the number of exclusive attributes derived from $\boldsymbol{a}_{i,j}$, and M is the number of all attributes. Through Inequation 6, the contributions of exclusive attributes tends to be larger than the linear proportion while those of common attributes tends to be smaller. Based on this prior, we define the first part of the attribute prior loss as:

$$\begin{aligned} L_{p1} = \max(0, (\frac{M_E}{M})^{\upsilon} - \sum\nolimits_{e=1}^{M_E} \frac{d^e_{i,j}}{\hat{d}_{i,j}}) \\ + \max(0, \sum\nolimits_{c=1}^{M-M_E} \frac{d^c_{i,j}}{\hat{d}_{i,j}} - 1 + (\frac{M_E}{M})^{\upsilon}). \end{aligned} \tag{7}$$

The second group of constraints is applied to the contribution of individual attribute. We set a lower bound for each $d^e_{i,j}$ and a upper bound for each $d^c_{i,j}$ formulated by:

$$\frac{d^e_{i,j}}{\hat{d_{i,j}}} \geq e^{-\lambda} \frac{(\frac{M_E}{M})^{\upsilon}}{M_E}, \quad \frac{d^c_{i,j}}{\hat{d_{i,j}}} \leq e^{\lambda} \frac{1 - (\frac{M_E}{M})^{\upsilon}}{M - M_E}. \tag{8}$$

Here, we let the above upper bound and lower bound be equal, so the value of λ can be solved by

$$\lambda = \frac{1}{2} \ln \frac{M - M_E (\frac{M_E}{M})^{\upsilon}}{M_E (1 - (\frac{M_E}{M})^{\upsilon})}. \tag{9}$$

From Equation 9, we can see that the upper bound and lower bound are related to the ratios of exclusive attributes and common attributes to all attributes. Based on these priors, we define the second part of the attribute prior loss as:

$$\begin{aligned} L_{p2} = \sum\nolimits_{e=1}^{M_E} \max(0, e^{-\lambda} \frac{(\frac{M_E}{M})^{\upsilon}}{M_E} - \frac{d^e_{ij}}{\hat{d}_{ij}}) \\ + \sum\nolimits_{c=1}^{M-M_E} \max(0, \frac{d^c_{ij}}{\hat{d}_{ij}} - e^{\lambda} \frac{1 - (\frac{M_E}{M})^{\upsilon}}{M - M_E}). \end{aligned} \tag{10}$$

Through the two attributes prior losses, the interpreter will be more focused on the exclusive attributes that make more contribution to the overall difference of two persons. Finally, the interpreter is optimized by the total loss function:

$$L = L_d + \alpha L_{p1} + \beta L_{p2}, \tag{11}$$

where α and β are the balance factors.

3.4. Training and Inference

Training. Firstly, a ReID model $\mathcal{F}(\cdot)$ trained on person ReID data is used as the target model and fixed during learning the interpreter $\mathcal{G}(\cdot)$. In each training iteration, we take $P \times S$ images as a mini-batch where P is the number of IDs and S is sample number per ID. In a mini-batch, we can obtain $P^2 \times S^2$ pairs of images to train the interpreter with Equation 11. For each pair (x_i, x_j), we use the distance $d_{i,j}$ generated by $\mathcal{F}(\cdot)$ and the attribute vector $\boldsymbol{a}_{i,j}$ as the supervision for training $\mathcal{G}(\cdot)$.

Inference. During testing, given a query image x_q and a gallery image x_g, The interpreter $\mathcal{G}(\cdot)$ can generate the attention maps of attributes $(A^1_q, A^2_q, ..., A^M_q)$ and $(A^1_g, A^2_g, ..., A^M_g)$, and output the attribute-guided distances $\{d^k_{p,q}\}^M$ by forward propagation. The contribution ratio of attribute k is computed by $r^k_{p,q} = d^k_{p,q} / \sum_{k=1}^{M} d^k_{p,q}$.

4. Experiments

To show the effectiveness and compatibility of the interpreter learned by AMD, we first evaluate our method for different target models on individual datasets. Then the cross-domain experiments are conducted to demonstrate the generalization of the interpreter. At last, we incorporate the interpreter with several state-of-the-art ReID models to achieve superior accuracy on different benchmarks.

4.1. Datasets

Our experiments is performed on two large-scale person ReID datasets: **Market-1501** [42] and **DukeMTMC-ReID** [45]. For convenience, we directly use the ID-level attributes labeled by Lin *et al.* [18] to train our interpreter.

Market-1501 contains 751 training IDs with 19,732 images and 750 testing IDs with 13,328 images. We select 26 attributes labeled by [18] including: gender (female/male), hair length (long/short), sleeve length (long/short), length of lower clothing (long/short), type of lower clothing (pants/dress), wearing hat (yes/no), carrying backpack (yes/no), carrying handbag (yes/no), carrying other bags (yes/no), 8 colors of upper clothing, and 9 colors of lower clothing. The statistics of attributes in Market1501 is shown in Figure 3, which reflects the imbalance of attributes.

DukeMTMC-ReID contains 702 training IDs with 16,522 images and 702 testing IDs with 19,889 images. For DukeMTMC-ReID, we select 23 attributes for the interpreter. For more details on the attributes of DukeMTMC-ReID, please refer to the **supplementary material**.

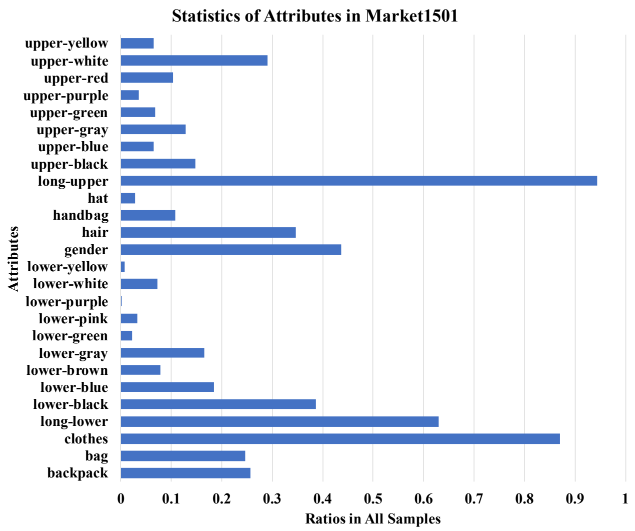

Figure 3. The statistics of attributes on the Market-1501 dataset.

4.2. Implementation Details

This subsection presents the implementation details of our framework and training strategy of the interpreter.

Network Structure. The ReID model $\mathcal{F}(\cdot)$ and the interpreter $\mathcal{G}(\cdot)$ are built as in Section 3.2. For the target models $\mathcal{F}(\cdot)$, we use one of the state-of-the-art ReID models, i.e., Stronger-Baseline (SBS) [12], with different backbones, e.g., ResNet-18/34/50/101. The interpreter $\mathcal{G}(\cdot)$ uses the same backbone with $\mathcal{F}(\cdot)$ and shares the first three CNN stages, i.e., Conv1 to Conv3, from $\mathcal{F}(\cdot)$. The rest stages are initialized by parameters pretrained on ImageNet [6].

Networks Training. The interpreter $\mathcal{G}(\cdot)$ is trained on the training sets of Market-1501 and DukeMTMC-ReID with the attribute number $M = 26$ and 23, respectively. We adopt the Adam [14] optimizer to train $\mathcal{G}(\cdot)$ for 30 epochs with the basic learning rate $lr = 10^{-4}$. The warm-up strategy is used for the first 10 epochs with the initial $lr = 10^{-6}$. For the hyper-parameters in Equation 11, α and β are set to 10.0 and 50.0, respectively. The υ in Equation 6 is set to 0.5. The κ and τ of PePU in Equation 2 is set to $1/M$ and 0.5, respectively. The mini-batch size is 6×4, *et al.*, 6 IDs and 4 samples per ID.

4.3. Evaluation Metrics

Although existing interpretation for CNNs are usually demonstrated by visualization and evaluated by subjective observation, we define a group of objective metrics to evaluate the correctness of our interpreter for person ReID.

Metric for Distillation. Since the interpreter aims to decompose $d_{p,q}$ of (x_q, x_g) generated by a target model into a set of components, we first measure the information loss during metric distillation from $\mathcal{F}(\cdot)$ to $\mathcal{G}(\cdot)$. Given the attribute-guided distances $\{d^k_{q,g}\}^M$ from $\mathcal{G}(\cdot)$, we sum all items to obtain a reconstructed distance $\hat{d}_{q,g} = \sum_{k=1}^{M} d^k_{q,g}$. We report the Average Distance Reconstruction Error (ADRE) over all query-gallery pairs by $\frac{1}{|Q|\cdot|G|}\sum_{q=1}^{|Q|}\sum_{g=1}^{|G|}\frac{|d_{q,g}-\hat{d}_{q,g}|}{d_{q,g}}$. Moreover, we use the reconstructed $\hat{d}_{q,g}$ to perform the ReID task as using $d_{q,g}$. Thus, we can observe the information loss by comparing the ReID performance, *e.g.*, Rank-1 accuracy (Rank-1) and mean Average Precision (mAP), of $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$.

Metric for Attribute Decomposition. As we expect the interpreter to find the most discrepant attributes of x_q and x_g, we measure the ability of $\mathcal{G}(\cdot)$ based on whether it can assign more contributions to the exclusive attributes rather than the common attributes. In traditional explainability literature, measures such as the pointing game [37] or insertion/deletion [23] are not suitable for our task. The "Point Game" requires the bounding boxes for attributes, while ReID datasets only have image-level labels. The "Insertion/Deletion" and the "Blur Integrated Gradients" [35] are mainly designed for the classification task. Therefore, we design two metrics X-mAP$_e$ and X-mAP$_c$ for exclusive attributes and common attributes, respectively.

Given input (x_q, x_g), pairwise attribute vector $\boldsymbol{a}_{q,g}$, and the attribute-guided distances ($d^1_{q,g}, d^2_{q,g}, ..., d^M_{q,g},$), we rank the distances in an descending order as the larger value means more difference. Then we compute the Average Precision (AP) of the ranked list like the retrieval task to measure whether exclusive attributes are ranked at the top positions in the list. The X-mAP$_e$ is the mean value of the AP values over all query and gallery pairs. Similarly, the X-mAP$_c$ is calculated by ranking the distances in an ascending order to measure whether common attributes are ranked the top positions in the list. In our experiments, we evaluate the X-mAP$_e$ and X-mAP$_c$ on the testing set for all query and gallery pairs except the pairs from the same ID.

4.4. Experimental Results of Different Models

This subsection presents experiments for interpreting the SBS models with different backbones (ResNet-34/50/101) on Market-1501 and DukeMTMC-ReID. The metrics including the ReID accuracy of the target model and the interpreter, and X-mAP$_e$, X-mAP$_c$, and ADRE of the interpreter are listed in Table 1. Examples of attribute-guided attention maps (AAMs) and contributions of top-3 attributes generated for SBS (ResNet-50) are shown in Figure 4.

In Table 1, the target model and the corresponding interpreter are grouped for comparison. From ReID accuracy, we can see that interpreters achieve very close accuracy to target ReID models, meanwhile the ADRE between each interpreter and ReID model is also very small. This means that the information loss is very minor during decomposing the distance from the target model into attribute-guided components by distillation. The minor loss is acceptable and reasonable because the attribute-guided interpreter sacrifices some discriminative representations but obtains more

Datasets	SBS Models	Rank-1 (%)	Rank-5 (%)	mAP (%)	X-mAP$_e$ (% ↑)	X-mAP$_c$ (% ↑)	ADRE (% ↓)
Market-1501	ResNet-34	93.94	97.74	83.95	-	-	-
	Interpreter	94.21	97.80	84.12	73.71	96.39	2.31
	ResNet-50	94.77	98.13	87.15	-	-	-
	Interpreter	94.74	98.16	87.11	74.29	96.59	1.99
	ResNet-101	95.94	98.40	88.64	-	-	-
	Interpreter	95.55	98.52	88.29	75.40	96.73	1.87
DukeMTMC-reID	ResNet-34	86.67	92.77	71.71	-	-	-
	Interpreter	86.13	92.91	72.00	69.58	95.79	1.93
	ResNet-50	88.24	94.17	75.54	-	-	-
	Interpreter	87.84	94.34	75.27	70.30	96.03	1.73
	ResNet-101	89.33	95.20	78.41	-	-	-
	Interpreter	89.21	95.15	78.26	70.52	96.11	1.74

Table 1. Evaluation of interpreters for different backbone models on Market-1501 and DukeMTMC-ReID. Each target model and the corresponding interpreter are grouped for comparison. The results show that the interpreters learn consistent knowledge to the target models for effective explanations.

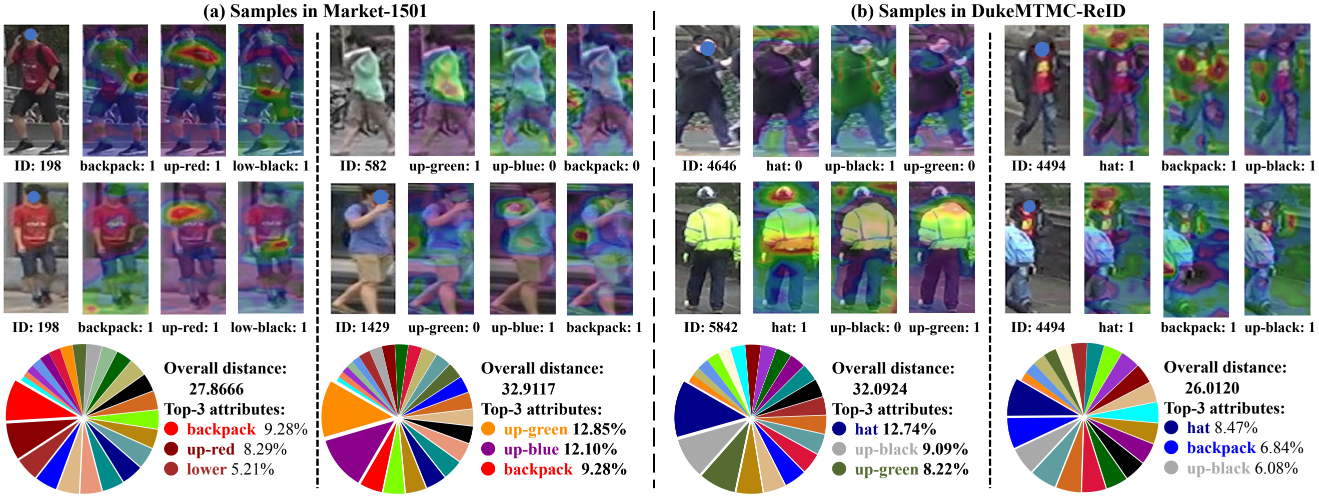

Figure 4. Pairwise examples and explanations for SBS (ResNet-50) on two datasets. For each pair of images, the upper part visualizes the AAMs of the top-3 attributes, which shows that the AAMs are attended to the discriminative attributes. The lower part shows the overall distance and contributions of the top-3 attributes. These figures show the most contributed attributes discovered by the interpreter. (Best viewed in color.)

explainable representations. Moreover, the good performance of X-mAP$_e$ and X-mAP$_c$ objectively reflects the rationality and correctness of interpreter. Furthermore, we can find that the metrics of the interpreters are consistent with the target models on different datasets, which reflects the generalization of the interpreter.

In Figure 4, we show several pairwise images and their explanations for SBS (ResNet-50) models trained on Market-1501 and DukeMTMC-ReID, respectively. For each pair of images, we visualize the attribute-guided attention maps (AAMs) of the top-3 attributes and list the contributions of top-3 attributes to the overall distance. From the AAMs, we can observe that for the positive attributes (labeled as 1), the interpreter can effectively focus on corresponding regions, while for the negative attributes (labeled as 0), the attentions usually spread around persons. Particularly, some small objects like backpacks and hats can also be attended to by the interpreter even though they are partially occluded, which shows the effectiveness of the interpreter.

The effectiveness of the interpreter can also be demonstrated by the top-3 contributory attributes. First of all, the more different the attribute is, the more proportion it contributes to the distance. Taking the first pair of images as an example, they are the same person captured by different cameras. Due to varied viewpoints and illuminations, this person looks different especially on the backpack and the color of upper clothes. The interpreter can effectively assign larger contributions to these discriminative attributes, which can also be observed on other examples.

To further illustrate attentions learned by the interpreter, we show the average attention maps of individual attributes on two datasets in the **supplementary material**.

Datasets	Models	Rank-1 (%)	Rank-5 (%)	mAP (%)	X-mAP$_e$ (% ↑)	X-mAP$_c$ (% ↑)	ADRE (% ↓)
M → D	SBS (ResNet-50) Interpreter	45.02 47.08	61.40 62.61	26.43 28.41	- 59.01	- 91.47	- 2.16
D → M	SBS (ResNet-50) Interpreter	53.15 54.48	71.26 71.59	24.48 25.48	- 59.19	- 92.17	- 1.56

Table 2. Evaluation of the interpreters for SBS (ResNet-50) under the cross-domain setting. M → D means the SBS models and interpreters are trained on Market-1501 and tested on DukeMTMC-ReID, and D → M means the reverse setting. The results demonstrate that the information loss of interpreters is very minor under the cross-domain setting.

Models	Market-1501		DukeMTMC-ReID	
	Rank-1	mAP	Rank-1	mAP
OSNet [47] + Re-weighting	94.7 95.0(**+0.3**)	85.7 86.1(**+0.4**)	87.9 88.5(**+0.6**)	74.1 74.9(**+0.8**)
BOT [22] (R50) + Re-weighting	93.8 94.4(**+0.6**)	84.7 86.1(**+1.4**)	86.9 88.6(**+1.7**)	74.3 75.6(**+1.3**)
CL [27] (R50) + Re-weighting	94.9 95.2 (**+0.3**)	85.7 86.4 (**+0.7**)	87.1 88.3(**+1.2**)	71.9 73.1(**+1.2**)
SBS [12] (R50) + Re-weighting	94.8 95.2(**+0.4**)	87.2 87.9(**+0.7**)	88.2 89.1(**+0.9**)	75.5 75.6(**+0.1**)
SBS [12] (R101) + Re-weighting	95.9 96.1(**+0.2**)	88.6 88.8(**+0.2**)	89.3 90.2(**+0.9**)	78.4 79.1(**+0.7**)

Table 3. Comparison between results of the SOTA methods and refined results by re-weighted distances on Market-1501 and DukeMTMC-ReID. For all compared models, the results are further boosted.

4.5. Cross-domain Evaluation

To validate the generalization of our interpreter, we conduct experiments under the cross-domain setting on Market-1501 and DukeMTMC-ReID. Here we use the SBS (ResNet-50) models trained on Market-1501 and on DukeMTMC-ReID as two target models, then learn two interpreters for these two models. During the evaluation, we apply the SBS model and the interpreter trained on Market-1501 to the testing set of DukeMTMC-ReID (M → D) and vise vase (D → M). The experimental results are listed in Table 2. From the cross-domain results, we can find that the ReID metrics of the interpreters are still consistent with those of the target models, meanwhile the ADRE values are also very small. This also reflects that the interpreters can learn consistent knowledge from the target models under the cross-domain setting. Interestingly, we can find that the results of interpreters are better than those of the target models. This may be because that the generalization of the attribute-based representations is stronger than the visual features learned by the target models for cross-domain ReID. Therefore, attributes may have the potential to bridge the domain gap between different datasets. See the **supplementary material** for more examples.

4.6. Distance Re-weighting by Interpreter

As a by-product, we try a straightforward method to improve the performance of the state-of-the-art models with the explanations generated by the interpreter. As in Section 3.4, our interpreter can output the contributions of exclusive attributes to the overall distance, i.e., $\{d^e_{p,q}\}^{M_E}$ for each pair of query and gallery images. Then we can simply amplify the component of the most contributed attribute to obtain an updated distance. Given an original distance $d_{p,q}$, the updated distance can be computed by linear reweighting as:

$$d^{'}_{i,j} = d_{i,j} + \gamma \cdot \max(\{d^e_{p,q}\}^{M_E}), \qquad (12)$$

where γ is a hyper-parameter and set to 1.0 in experiments.

Comparison between results of the state-of-the-art (SOTA) methods and refined results by re-weighted distances from interpreters are listed in Tabled 3. On both Market-1501 and DukeMTMC-ReID datasets, the performance of all models is improved. Especially on DukeMTMC-ReID, we obtain 0.9% and 0.7% increases in rank-1 and mAP accuracy for the powerful SBS (ResNet-101). These results demonstrate great potential to explore attributes for further improvement of person ReID.

5. Conclusion

This paper presents an Attribute-guided Metric Distillation (AMD) method to use semantic attributes for explainable person ReID. The AMD learns a pluggable interpreter that can be grafted on any target CNN-based ReID model. With the metric distillation guided by attribute priors, the learned interpreter can decompose the distance of two person images into quantitative contributions of attributes by which users can know what attributes make two persons different. Meanwhile, the interpreter can visualize attention maps of discriminative and exclusive attributes to tell users where the most significant attributes are. With such quantitative explanations and intuitive visualizations, the interpreter can help users make decisions more effectively. In future work, the proposed AMD framework can be applied for explanation of other metric-based computer vision tasks like content-based image retrieval [8], vehicle ReID [20, 21], etc.

Acknowledgements. This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0103800.

References

[1] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, pages 3319–3327, 2017. 4

[2] Bryan Bryan, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *ICCV*, pages 3759–3768, 2019. 3

[3] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, pages 371–381, 2019. 3

[4] Lei Chen, Jianhui Chen, Hossein Hajimirsadeghi, and Greg Mori. Adapting grad-cam for embedding networks. In *WACV*, pages 2783–2792, 2020. 1

[5] Runjin Chen, Hao Chen, Ge Huang, Jie Ren, and Quanshi Zhang. Explaining neural networks semantically and quantitatively. In *ICCV*, pages 9186–9195, 2019. 1, 2, 4

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[7] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *ACM MM*, pages 789–792, 2014. 2

[8] Bo Dong, Roddy Collins, and Anthony Hoogs. Explainability for content-based image retrieval. In *CVPR*, pages 95–98, 2019. 1, 8

[9] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, pages 3449–3457, 2017. 2

[10] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *IEEE DSAA*, pages 80–89, 2018. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[12] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *CoRR*, abs/2006.02631, 2020. 3, 6, 8

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. 3

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[15] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *CoRR*, abs/1603.07054, 2016. 2

[16] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 3

[17] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 3, 4

[18] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *PR*, 95:151–161, 2019. 2, 5

[19] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective. *CoRR*, abs/2104.11536, 2021. 1

[20] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *ICME*, pages 1–6, 2016. 8

[21] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. PROVID: progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Trans. Multim.*, 20(3):645–658, 2018. 8

[22] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE TMM*, 22(10):2597–2609, 2020. 3, 8

[23] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. *CoRR*, abs/2006.03204, 2020. 6

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 1, 2

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014. 1, 2

[26] Abby Stylianou, Richard Souvenir, and Robert Pless. Visualizing deep similarity networks. In *WACV*, pages 2029–2037, 2019. 1

[27] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6397–6406, 2020. 3, 8

[28] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In *ECCV*, pages 501–518, 2018. 3

[29] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018. 3

[30] Qi Wang, Xinchen Liu, Wu Liu, An-An Liu, Wenyin Liu, and Tao Mei. Metasearch: Incremental product search via deep meta-learning. *IEEE Trans. Image Process.*, 29:7549–7564, 2020. 1

[31] Di Wu, Si-Jia Zheng, Wenzheng Bao, Xiao-Ping (Steven) Zhang, Chang-An Yuan, and De-Shuang Huang. A novel deep model with multi-loss and efficient training for person *re*-identification. *Neurocomputing*, 324:69–75, 2019. 3

[32] Di Wu, Si-Jia Zheng, Xiao-Ping (Steven) Zhang, Chang-An Yuan, Fei Cheng, Yang Zhao, Yong-Jun Lin, Zhong-Qiu Zhao, Yong-Li Jiang, and De-Shuang Huang. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 337:354–371, 2019. 1

[33] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Towards global explanations of convolutional neural networks with concept attribution. In *CVPR*, pages 8649–8658, 2020. 2

[34] Boqiang Xu, Lingxiao He, Xingyu Liao, Wu Liu, Zhenan Sun, and Tao Mei. Black re-id: A head-shoulder descriptor for the challenging problem of person re-identification. In *ACM MM*, pages 673–681, 2020. 1

[35] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *CVPR*, pages 9677–9686, 2020. 6

[36] 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*, pages 1389–1398, 2019. 2, 3, 4

[37] Jianming Zhang, Zhe L. Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, pages 543–559, 2016. 6

[38] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *CVPR*, pages 6261–6270, 2019. 2

[39] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3183–3192, 2020. 3

[40] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3239–3248, 2017. 3

[41] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *CVPR*, pages 5310–5319, 2021. 1

[42] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 1, 2, 3, 5

[43] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *CoRR*, abs/1610.02984, 2016. 1

[44] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 3

[45] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *ICCV*, pages 3774–3782, 2017. 2, 5

[46] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 1, 2, 3

[47] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3701–3711, 2019. 8