

Adversarial Attribute-Image Person Re-identification

Zhou Yin¹, Wei-Shi Zheng^{1,3*}, Ancong Wu¹, Hong-Xing Yu¹, Hai Wan¹,
Xiaowei Guo², Feiyue Huang², Jianhuang Lai¹

¹ School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

² YouTu Lab, Tencent

³ Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

Abstract

While attributes have been widely used for person re-identification (Re-ID) which aims at matching the same person images across disjoint camera views, they are used either as extra features or for performing multi-task learning to assist the image-image matching task. However, how to find a set of person images according to a given attribute description, which is very practical in many surveillance applications, remains a rarely investigated cross-modality matching problem in person Re-ID. In this work, we present this challenge and leverage adversarial learning to formulate the attribute-image cross-modality person Re-ID model. By imposing a semantic consistency constraint across modalities as a regularization, the adversarial learning enables to generate image-analogous concepts of query attributes for matching the corresponding images at both global level and semantic ID level. We conducted extensive experiments on three attribute datasets and demonstrated that the regularized adversarial modelling is so far the most effective method for the attribute-image cross-modality person Re-ID problem.

1 Introduction

Pedestrian attributes, e.g., age, gender and dressing, are searchable semantic elements to describe a person. In many scenarios we need to search person images in surveillance environment according to specific attribute descriptions provided by users, as depicted in Figure 1. We refer to this problem as the *attribute-image person re-identification (attribute-image Re-ID)*. This task is significant in finding missing people with tens of thousands of surveillance cameras equipped in modern metropolises. Compared with conventional image-based Re-ID [Zhao *et al.*, 2014; Yang *et al.*, 2016], attribute-image Re-ID has the advantage that its query is much easier to be obtained, e.g., it is more practical to search for criminal suspects when only verbal testimony about the suspects' appearances is given.

*Corresponding author, email: wszheng@ieee.org

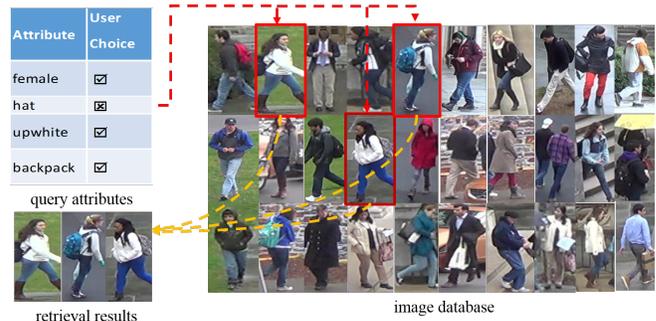


Figure 1: The attribute-image Re-ID problem. The query is an attribute vector labeled by users, and the corresponding target person images that are matched with the query are retrieved.

Despite the great significance, the attribute-image Re-ID is still a very open problem and has been rarely investigated before. While a lot of attribute person Re-ID models [Lin *et al.*, 2017; Layne *et al.*, 2012a; Su *et al.*, 2016; Layne *et al.*, 2012b; 2014b; 2014a; Su *et al.*, 2015a] have been developed recently, they are mainly used either for multi-task learning or for providing extra features so as to enhance the performance of image-image person Re-ID model. The most intuitive solution to attribute-image Re-ID might be to predict attributes for each person image, and search within the predicted attributes [Siddiquie *et al.*, 2011; Vaquero *et al.*, 2009; Scheirer *et al.*, 2012]. If we can reliably recognize the attributes of each pedestrian image, this might be the best way to find the person corresponding to the query attributes. However, recognizing attributes from a person image is still an open issue, as pedestrian images from surveillance environment often suffer from low resolution, pose variations and illumination changes. The problem of imperfect recognition limits the intuitive methods in bridging the gap between the two modalities (attribute and image), which are heterogeneous from each other. In addition, very often in a large-scale scenario, the predicted attributes from two pedestrians are different but very similar, leading to a very small inter-class distance in the predicted attribute space. Therefore, the imperfect prediction deteriorate the reliability of these existing models.

In this paper, we propose an adversarial attribute-image Re-ID framework. Intuitively, when we hold some attribute description in mind, e.g., “dressed in red”, we generate an

obscure and vague imagination on how a person dressed in red may look like, which we refer to as a concept. Once a concrete image is given, our vision system automatically processes the low-level features (e.g., color and edge) to obtain some perceptions, and then try to judge whether the perceptions and the concept are consistent with each other.

More formally, the goal of our adversarial attribute-image Re-ID framework is to learn a semantically discriminative joint space, which is regarded as a concept space (Figure 2), for generative adversarial architecture to generate some concepts that are very similar to the concepts extracted from raw person images. However, the generic adversarial architecture is still hard to fit the match between two extremely large discrepant modalities (attribute and image). For this problem, we impose a semantic consistency regularization across modalities in order to regularize the adversarial architecture, enhancing the learned joint space to build a bridge between the two modalities.

In a word, our framework learns a semantically discriminative structure of low-level person images, and generate a correspondingly aligned image-analogous concept for high-level attribute towards image concept. By the proposed strategy, directly estimating the attributes of a person image is averted, and the problems of imperfect prediction and low semantic discriminability are naturally solved, because we learn a semantically discriminative joint space, rather than predicting and matching attributes.

We have conducted experiments on three large-scale benchmark datasets, namely Duke Attribute [Lin *et al.*, 2017], Market Attribute [Lin *et al.*, 2017] and PETA [Deng *et al.*, 2014], to validate our model. By our study, some interesting findings are:

- (1) Compared with other related cross-modality models, we find the regularized adversarial learning framework is so far most effective for solving the attribute-image Re-ID problem.
- (2) For achieving better cross-modality matching between attribute and person image, it is more effective to use adversarial model to generate image-analogous concept and get it matched with image concept rather than doing this in reverse.
- (3) The semantic consistency as regularization on adversarial learning is important for the attribute-image Re-ID.

2 Related Works

2.1 Attribute-based Re-ID

While pedestrian attributes in most research are side information or mid-level representation to improve conventional image-image Re-ID tasks [Lin *et al.*, 2017; Layne *et al.*, 2012a; Su *et al.*, 2016; Layne *et al.*, 2012b; 2014b; 2014a; Su *et al.*, 2015a; 2015b; 2018], only a few work [Vaquero *et al.*, 2009] has discussed attribute-image Re-ID problem. The work in [Vaquero *et al.*, 2009] is to form attribute-attribute matching. However, despite the improvement on performance achieved by attribute prediction methods [Li *et al.*, 2015], directly retrieving people according to their predicted attributes is still challenging, because the attribute prediction methods are still not robust to the cross-view condition changes like different lighting conditions and viewpoints.

In this work, for the first time, we present extensive investigation on the attribute-image Re-ID problem under an adversarial framework. Rather than directly predicting attributes of an image, we cast the cross-view attribute-image matching as cross-modality matching by an adversarial learning problem.

2.2 Cross-modality Retrieval

Our work is related to cross-modality content search, which aims to bridge the gap between different modalities [Hotelling, 1936; Mineiro and Karampatziakis, 2014; Andrew *et al.*, 2013; Kiros *et al.*, 2014]. The most traditional and practical solution to this task is Canonical Correlation Analysis (CCA) [Hotelling, 1936; Mineiro and Karampatziakis, 2014; Andrew *et al.*, 2013], which projects two modalities into a common space that maximizes their correlation. Ngiam *et al.* and Feng *et al.* also applied autoencoder-based methods to model the cross-modality correlation [Ngiam *et al.*, 2011; Feng *et al.*, 2014], and Wei *et al.* proposed a deep semantic matching method to address the cross-modality retrieval problem with respect to samples annotated with one or multiple labels [Wei *et al.*, 2017]. Recently, A. Eisenschat and L. Wolf have designed a novel model of two tied pipelines that maximize the projection correlation using an Euclidean loss, which achieves state of the art results in some datasets [Eisenschat and Wolf, 2017]. Two most related works to ours are proposed in [S.Li *et al.*, 2017; Li *et al.*, 2017], which retrieve pedestrian images using language descriptions. Compared with this setting, our attribute-image Re-ID has its own strength in embedding more pre-defined attribute descriptions for obtaining better performance.

2.3 Distribution Alignment Methods

The adversarial model employed in this work is in line with the GAN methods [Eric *et al.*, 2017; Reed *et al.*, 2016; Goodfellow *et al.*, 2014], which has its strength in distribution alignment by a two player min-max game. As different modalities follow different distributions, our cross-modality attribute-image Re-ID problem is also related to the distribution alignment problem. For performing distribution alignment, there are other techniques called domain adaptation techniques [Tzeng *et al.*, 2014; Long and Wang, 2015; Ganin and Lempitsky, 2015]. In domain adaptation, to align the distribution of data from two different domains, several works [Tzeng *et al.*, 2014; Long and Wang, 2015] apply MMD-based loss, which minimize the norm of difference between means of the two distributions. Different from these methods, the deep Correlation Alignment (CORAL) [Sun and Saenko, 2016] method proposed to match both the mean and covariance of the two distributions. Our work is different from these methods as our framework not only bridges the gap between the two largely discrepant modalities, but also keeps the semantic consistency across them.

3 Attribute-image Person Re-ID

Given an attribute description A_i , attribute-image Re-ID aims at re-identifying the matched pedestrian images I_i from an image database $\mathcal{I} = \{I_i\}_{i=1}^N$ captured under real surveillance

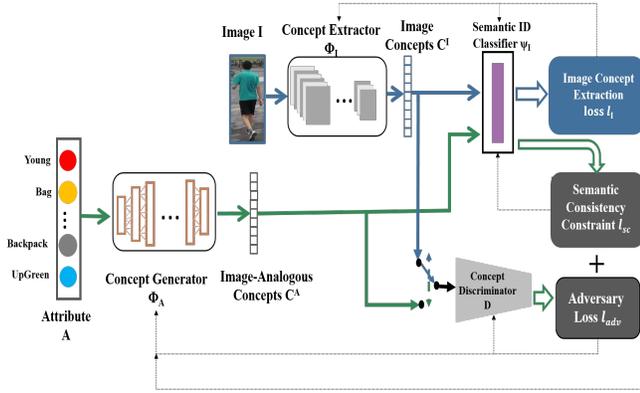


Figure 2: The structure of our network. It consists of two branches: the image branch (top blue branch) learns to extract semantic concepts from images, and the attribute branch (bottom green branch) learns to generate image analogous concepts from attributes. Dash lines represent the gradient flow of the three objectives that we propose. See Sec. 3.3 for details about the network architecture.

environment, where N is the size of \mathcal{I} . Since different person images could have the same attribute description, the attribute-image Re-ID uses Semantic ID (i.e., attribute description) to group person images. That is, people with the same attribute description are of the same semantic ID.

The goal of our method is to learn two mappings Φ_I and Φ_A that respectively map person images and high-level semantic attributes into a joint discriminative space, which could be regarded as the concept space as mentioned. That is, $C_i^I = \Phi_I(I_i)$ and $C_i^A = \Phi_A(A_i)$, where C_i^I and C_i^A are the mid-level concept that is generated from the image I_i and attribute A_i , respectively. To achieve this, we form an image embedding network by CNN and an attribute-embedding network by a deep fully connected network. We parameterize our model by Θ , and obtain Θ by optimizing a concept generation objective $L_{concept}$. Given training pairs of images and attributes (I_i, A_i) , the optimization problem is formulated as

$$\min_{\Theta} L_{concept} = \frac{1}{N} \sum_{i=1}^N l_{concept}(\Phi_I(I_i), \Phi_A(A_i)). \quad (1)$$

In this paper, we design $l_{concept}$ as a combination of several loss terms, each of which formulates a specific aspect of consideration to jointly formulate our problem. The first consideration is that the concepts we extract from the low-level noisy person images should be semantically discriminative. We formulate it by image concept extraction loss l_I . The second consideration is that image-analogous concepts C^A generated from attributes and image concepts C^I should be homogeneous. Inspired by the powerful ability of generative adversary networks to close the gap between heterogeneous distributions, we propose to embed an adversarial learning strategy into our model. This is modelled by a concept generating objective l_{CG} , which aims to generate concepts not only discriminative but also homogeneous with concepts extracted from images. Therefore, we have

$$l_{concept} = l_I + l_{CG}. \quad (2)$$

In the following, we describe each of them in detail.

3.1 Image Concept Extraction

Our image concept extraction loss l_I is based on softmax classification on the image concepts $\Phi_I(I)$. Since our objective is to learn semantically discriminative concepts that could distinguish different attributes rather than specific persons, we re-assign semantic IDs y_i for any person image I_i according to its attributes rather than real person IDs, which means different people with the same attributes have the same semantic ID. We define the image concept extraction loss as a softmax classification objective on semantic IDs. Denoting Ψ_I as the classifier and I as the input image, the image concept extraction loss is the negative log likelihood of predicted scores $\Psi_I(\Phi_I(I))$:

$$l_I = \sum_i -\log \Psi_I(\Phi_I(I_i))_{y_{I_i}}, \quad (3)$$

where I_i is the i^{th} input image, y_{I_i} is the semantic ID of I_i and $\Psi_I(\Phi_I(I_i))_k$ is the k^{th} element of $\Psi_I(\Phi_I(I_i))$.

3.2 Semantic-preserving Image-analogous Concept Generation

Image-analogous Concept Generation. We regard Φ_A as a generative process, just like the process of people generating an imagination from an attribute description. As the semantically discriminative latent concepts could be extracted from images, they can also provide information to learn the image-analogous concepts $\Phi_A(A)$ for attributes as a guideline.

Mathematically, the generated image-analogous concepts should follow the same distribution as image concepts, i.e., $P_I(C) = P_A(C)$, where C denotes a concept in the joint concept space of $\Phi_I(I)$ and $\Phi_A(A)$ and P_I, P_A denote the distribution of image concepts and image-analogous concepts, respectively. We learn a function \hat{P}_I to approximate image concept distribution P_I , and force the image-analogous concepts $\Phi_A(A)$ to follow distribution \hat{P}_I . It can be achieved by an adversarial training process of GAN, in which discriminator D is regarded as \hat{P}_I and the generator G is regarded as image-analogous concept generator Φ_A .

In the adversary training process, we design a network structure (see Sec. 3.3) and train our concept generator Φ_A with a goal of fooling a skillful concept discriminator D that is trained to distinguish the image-analogous concept from the image concept, so that the generated image-analogous concept is aligned with the image concept. We design the discriminator network D with parameters θ_D and denote the parameters of Φ_A as θ_G . The adversarial min-max problem is formulated as

$$\min_{\theta_G} \max_{\theta_D} V(D, G) = \mathbb{E}_{I \sim p_I} [\log D(\Phi_I(I))] + \mathbb{E}_{A \sim p_A} [\log(1 - D(\Phi_A(A)))]. \quad (4)$$

The above optimization problem is solved by iteratively optimizing θ_D and θ_G . Therefore, the objective can be decomposed into two loss terms l_{adv}^G and l_{adv}^D , which are for training the concept generator Φ_A and the discriminator D , respectively. Then the whole objective during adversary training l_{adv} could be formed by:

$$l_{adv} = \lambda_D l_{adv}^D + \lambda_G l_{adv}^G, \quad (5)$$

where

$$\begin{aligned}
 l_{adv}^G &= -\mathbb{E}_{A \sim p_A} [\log D(\Phi_A(A))], \\
 l_{adv}^D &= -\mathbb{E}_{I \sim p_I} [\log D(\Phi_I(I))] \\
 &\quad - \mathbb{E}_{A \sim p_A} [\log(1 - D(\Phi_A(A)))].
 \end{aligned}$$

Semantic Consistency Constraint. The adversarial learning pattern l_{adv} is important for generator Φ_A to generate image-analogous concept with the same distribution of image concept $\Phi_I(I)$. Furthermore, we should generate meaningful concepts preserving the semantic discriminability of the attribute modality, i.e., $P_I^{sid}(C) = P_A^{sid}(C)$, where P_I^{sid} and P_A^{sid} denote the distributions of image concepts and image-analogous concepts of semantic ID sid . If we analyze the image concept extraction loss l_I in Equation (3) independently, Ψ_I can be regarded as a function to approximate a set of distributions $P_I^{sid}(C)$ for each semantic ID sid . With the assumption that the generated image-analogous concepts should be in the same concept space as image concepts, Ψ_I is shared by image concept extraction and image-analogous concept generation, so as to guarantee identical distribution of two modalities in semantic ID level. We integrate a semantic consistency constraint l_{sc} using the same classifier for image concept Ψ_I :

$$l_{sc} = \sum_i -\log \Psi_I(\Phi_A(A_i))_{y_{A_i}}, \quad (6)$$

where A_i is the i^{th} input attribute, y_{A_i} is the semantic ID of A_i and $\Psi_I(\Phi_A(A_i))_k$ is the k^{th} element of $\Psi_I(\Phi_A(A_i))$. Thus the overall concept generating objective for attributes l_{CG} becomes the sum of l_{adv} and l_{sc} :

$$l_{CG} = l_{adv} + l_{sc}. \quad (7)$$

By this way, we encourage our generative model to generate a more homogeneous image-analogous concept space, while at the same time correlate image-analogous concepts with semantically matched image concepts by maintaining semantic discriminability in the learned space.

3.3 The Network Architecture

Our network architecture is shown in Figure 2. Firstly, the concept generator is particularly designed to have multiple fully connected layers in order to obtain enough capacity to generate image-analogous concepts which are highly heterogeneous from the input attribute. Details are shown in Table 1. Secondly, our concept discriminator is also a combination of fully connected layers, each followed by batch normalization and leaky ReLU, except for the output layer, which is processed by the Sigmoid non-linearity. Finally, the concept extractor is obtained by removing the last Softmax classification layer of Resnet-50 and adding a 128-D fully connected layer. We regard the feature produced by the FC layer as the image concept. Note that the dimension of the last layer in the concept generator is also set to 128.

As introduced above, we impose the semantic consistency constraint on attribute and thus we pass image-analogous concepts into the same Semantic ID classifier as that for

Structure	Size
fc1	$attributeSize \times 128$
BatchNormalization	128
ReLU	128
fc2	128×256
BatchNormalization	256
ReLU	256
fc3	256×512
BatchNormalization	512
ReLU	512
fc4	$512 \times embeddingSize$
Tanh	$embeddingSize$

Table 1: The structure of our network’ attribute part. Fc means fully connected layers. 128 is set to be the embedding size in our work.

image concepts. At the inference stage, we rank the gallery pedestrian image concepts C^I according to their cosine distances to the query image-analogous concepts C^A in the latent embedding space.

Implementation Details. We first pre-trained our image network for 100 epochs using the semantic ID, with an adam optimizer [Kingma and Ba, 2015] with learning rate 0.01, momentum 0.9 and weight decay $5e-4$. After that, we jointly train the whole network. We set λ_G in Eq. (2) as 0.001, and λ_D as 0.5, which will be discussed in Section 4.2. The total epoch was set to 300. During training, we set the learning rate of the attribute branch to 0.01, and set the learning rate of the image branch to 0.001 because it had been pre-trained. The batch size of training is 128 and the setting of optimizer is the same as that of pre-training. Hyper-parameters are fixed in comparisons across all the datasets.

4 Experiments

4.1 Datasets and Settings

Datasets. We evaluate our approach and compare with related methods on three benchmark datasets, including Duke Attribute [Lin *et al.*, 2017], Market Attribute [Lin *et al.*, 2017], and PETA [Deng *et al.*, 2014]. We tried to follow the setting in literatures. The Duke Attribute dataset contains 16522 images for training, and 19889 images for testing. Each person has 23 attributes. We labelled the images using semantic IDs according to their attributes. As a result, we have 300 semantic IDs for training and 387 semantic IDs for testing. Similar to Duke Attribute, the Market Attribute also has 27 attributes to describe a person, with 12141 images and 508 semantic IDs in the training set, and 15631 images and 484 semantic IDs in the test set. For PETA dataset, each person has 65 attributes (61 binary and 4 multi-valued). We used 10500 images with 1500 semantic IDs for training, and 1500 images with 200 semantic IDs for testing.

Evaluation Metrics. We computed both Cumulative Match Characteristic (CMC) and mean average precision (mAP) as metrics to measure performances of the compared models.

4.2 Evaluation on the Proposed Model

Adversarial vs. Other Distribution Alignment Techniques. For our attribute-image Re-ID, we employ the adversarial technique to make the image-analogous concepts generated from attribute aligned with the image concepts.

Method	Market				Duke				PETA			
	rank1	rank5	rank10	mAP	rank1	rank5	rank10	mAP	rank1	rank5	rank10	mAP
DeepCCA [Wang <i>et al.</i> , 2015]	8.12	23.97	34.55	9.72	33.28	59.35	67.64	14.95	14.24	22.09	29.94	14.45
DeepCCA [Andrew <i>et al.</i> , 2013]	29.94	50.70	58.14	17.47	36.71	58.79	65.11	13.53	14.44	20.77	26.31	11.49
2WayNet [Eisenschat and Wolf, 2017]	11.29	24.38	31.47	7.76	25.24	39.88	45.92	10.19	23.73	38.53	41.93	15.38
DeepMAR [Li <i>et al.</i> , 2015]	13.15	24.87	32.90	8.86	36.60	57.70	67.00	14.34	17.80	25.59	31.06	12.67
CMCE [Li <i>et al.</i> , 2017]	35.04	50.99	56.47	22.80	39.75	56.39	62.79	15.40	31.72	39.18	48.35	26.23
ours w/o adv	33.83	48.17	53.48	17.82	39.30	55.88	62.50	15.17	36.34	48.48	53.03	25.35
ours w/o sc	2.08	4.80	4.80	1.00	5.26	9.37	10.87	1.56	3.43	4.15	4.15	5.80
ours w/o adv+MMD	34.15	47.96	57.20	18.90	41.77	62.32	68.61	14.23	39.31	48.28	54.88	31.54
ours w/o adv+DeepCoral	36.56	47.61	55.92	20.08	46.09	61.02	68.15	17.10	35.62	48.65	53.75	27.58
ours	40.26	49.21	58.61	20.67	46.60	59.64	69.07	15.67	39.00	53.62	62.20	27.86

Table 2: Comparison results on the three benchmark datasets. Performances are measured by the rank1, rank5 and rank10 matching accuracy of the cumulative matching curve, as well as mAP. The best performances are indicated in **red** and the second indicated in **blue**.

While CCA is also an option and will be discussed when comparing our method with DCCA later, we examine whether other widely used alignment methods can work for our problem. We consider the MMD objective, which minimize difference between means of two distributions, and DeepCoral [Sun and Saenko, 2016], which matches both mean and covariance of two distributions, as traditional and effective distribution alignment baselines. Since their original models cannot be directly applied, we modify our model for comparison, that is we compare with 1) our model without the adversary learning but with an MMD objective(ours w/o adv+MMD); 2) our model without the adversary learning but with Coral objective(ours w/o adv+DeepCoral). We also provide the baseline that adversarial learning is not presented, denoted as “ours w/o adv”.

Compared with the model that does not use adversarial learning (ours w/o adv), all the other baselines including our adversary method perform clearly better. Among all, the adversary learning framework generally performs better (with the best and second best performance) as shown in Table 2.

With vs. Without Semantic Consistency Constraint. In our framework, we tested our performance when the semantic consistency constraint is not used, denoted as “ours w/o sc”. As reported in Table 2, without semantic consistency constraint the performance drops sharply. This is because although the distributions of two modalities are aligned, the corresponding pair is not correctly matched. Hence, the semantic consistency constraint actually regularizes the adversarial learning to avert this problem. As shown, with semantic consistency constraint but without adversarial learning (i.e., “ours w/o adv”) our model clearly performed worse than our full model. All the observations suggest the generic adversarial model itself does not directly fit the task of aligning two modalities which are highly discrepant, but the regularized one by semantic consistency constraint does.

A2Img vs. Img2A. In our framework, we currently use the adversarial loss to align the generated image-analogous concept of attribute towards image concept, we call such case generation from attributes to image (A2Img). We now provide comparative results on generation from image to attributes (Img2A). As reported in Table 3, we find that Img2A is also effective, which even outperforms A2Img on the PETA dataset. But on larger datasets Market and Duke, A2Img performs better. The reason may be that the distribution of

Method	Market	Duke	PETA
Ours (i.e. A2Img)	40.3	46.6	39.0
Img2A (reverse of the proposed)	36.0	43.7	43.6
Real Images	8.13	20.01	19.85

Table 3: The rank1 matching accuracy of some variants of our model. “A2Img” denotes our model which generates concepts from attributes. “Img2A” does the reverse of “A2Img”. “Real Images” denotes the model which generates images (rather than concepts) for attributes.

semantic IDs is much sparser than the distribution of images. Thus, estimating the manifold of images from the training data is more reliable than estimating that of attributes. But in PETA, the number of images is relatively small while semantic IDs are relatively abundant compared with the other two datasets. Moreover, PETA also has more complicated sceneries and larger number of attribute descriptions, which are more challenging for images to learn discriminative concepts. Thus learning generated attribute-analogous concepts and aligning with attribute concepts provides more discriminative information, and Img2A performs better on PETA.

Generation in Concept Space or in Image Space. What if our model generates images instead of concepts, according to the attributes? We study how the generated image-analogous pattern (whether concepts or images) affects the effectiveness of our model. To this end, we use the conditional GAN in [Reed *et al.*, 2016] to generate fake image, which have aligned structure with real images, from our semantic attributes and a random noise input. We have modified some input dimension and added some convolution and deconvolution layers in [Reed *et al.*, 2016] to fit our setting. Firstly we train the generative models for 200 epochs, and then the classification loss is added for another training of 200 epochs.

We find the retrieval performance is worse than our original model, as shown in Table 3. This is probably because generating the whole pedestrian image introduces some noise, which is harmful in discriminative tasks like attribute-image Re-ID. In contrast, generating concepts which are relatively “clean” can avoid introducing unnecessary noise. Thus, generating image-analogous concepts in the discriminative concept space is more effective.

4.3 Comparison with Related Work

Comparing with Attribute Prediction Method. As mentioned above, an intuitive method of attribute-image Re-ID

is to predict attributes from person images and perform the matching between predicted attributes and query attributes. We compare our model with a classical attribute recognition model DeepMAR [Li *et al.*, 2015], which formulates attribute recognition as a multi-task learning problem and acts as an off-the-shelf attribute predictor in our experiment. As shown in Table 2, our model outperforms DeepMAR, and it is because DeepMAR still suffers from the problem of indiscriminate predicted attributes. Different from DeepMAR, we choose to learn latent representations as the bridge between the two modalities, where we successfully avert the problem caused by attribute prediction and learn more discriminative concepts using adversary training.

Comparing with Cross Modality Retrieval Models. Since our problem is essentially a cross-modality retrieval problem, we compare our model with the typical and commonly used Deep canonical correlation analysis (DCCA) [Andrew *et al.*, 2013], Deep canonically correlated autoencoders (DCCA) [Wang *et al.*, 2015] and a state-of-the-art model 2WayNet [Eisenschat and Wolf, 2017]. Deep CCA applies the CCA objective in deep neural networks in order to maximize the correlation between two different modalities. DCCA [Wang *et al.*, 2015] jointly models the cross-modality correlation and reconstruction information in the joint space learning process. 2WayNet is a recently proposed two-pipeline model which maximizes sample correlations.

We show the comparative results in Table 2. From Table 2, we can observe that our model outperforms all the cross modality retrieval baselines on all three datasets by large margins. This is partially because our model not only learns to close the gap between the two modalities in the joint concept space, but also keeps the semantic consistency of the extracted and generated concepts.

In addition, we compare our model with the most related one, i.e., the cross modality cross entropy (CMCE) model [Li *et al.*, 2017], which achieved a state-of-the-art result in text-based person retrieval. We train the CMCE model with semantic ID for fair comparison. The results in Table 2 show that our model is comparable (on Market) or more effective (on Duke and PETA) for the attribute-image Re-ID problem.

4.4 Further Evaluations

Finally we present some further evaluations of our model. We first evaluate the effects of two important hyper-parameters λ_D and λ_G . We present the results on the Duke Attribute dataset in Figure 3. The trends are similar on other datasets, and therefore Figure 3 might be useful for determining the hyper-parameters on other datasets.

Secondly, we conduct qualitative evaluations on our proposed model. Figure 4 shows examples of top-10 ranked images according to a query attribute description from the Market Attribute dataset. We find that fine-grained features of pedestrian images (e.g. stride of a backpack) are the main reasons that cause mistakes in our baseline (see ours w/o adv in the second row of Figure 4). But with the adversarial objective, our model could get an intuition and generate the concept of what a person wearing a backpack would look like, and then concentrate more on possible fine-grained features.

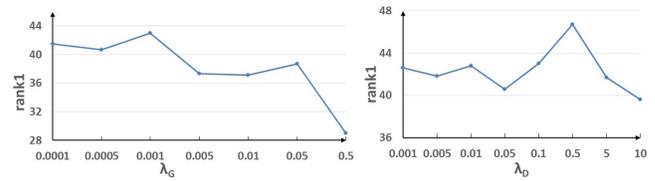


Figure 3: Results of experiment on the trade-off parameters λ_G and λ_D . We firstly set λ_D to 1 and change the value of λ_G , and get the results in the left image. Then we chose our best $\lambda_G=0.001$ in our experiments and change λ_D on the right.



Figure 4: Qualitative example in Market Attribute Dataset. The first row shows the results of our proposed method and the second are about a baseline. To save space, we only list 6 attribute items among all the 27 ones in Market Attribute in the third row. The inaccurately retrieved samples are marked by red rectangles in the figure.

5 Conclusion

The attribute-image Re-ID problem is a cross-modal matching problem that is realistic in practice, and it differs notably from the previous attribute-based person Re-ID problem that is still essentially an image-image Re-ID problem. In this work, we have identified its challenge through the experiments on three datasets. We have shown that an adversarial framework regularized by a semantic consistency constraint is so far the most effective way to solve the attribute-image Re-ID problem. Also, by our learning, we find that under the regularized adversarial learning framework, it is more useful to learn image-analogous concept from inquired attributes and make it aligned with the corresponding real image's concept, as compared with its reverse.

Acknowledgements

This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002), NSFC (61522115, 61661130157, 61472456, U1611461, 61573386), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), the Royal Society Newton Advanced Fellowship (NA150459), Guangdong NSF (2016A030313292), Guangdong Program (2016B030305007, 2017B010110011).

References

[Andrew *et al.*, 2013] Galen Andrew, Raman Arora, Karen Livescu, and Jeff Bilmes. Deep canonical correlation analysis. In *ICML*, 2013.

- [Deng *et al.*, 2014] Y. Deng, P. Luo, C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACMMM*, 2014.
- [Eisenschat and Wolf, 2017] A. Eisenschat and L. Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017.
- [Eric *et al.*, 2017] Tzeng Eric, Hoffman Judy, Saenko Kate, and Darrell Trevor. Adversarial discriminative domain adaptation. In *CVPR*, 2017.
- [Feng *et al.*, 2014] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *ACMMM*, 2014.
- [Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.
- [Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.
- [Hotelling, 1936] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 1936.
- [Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [Kiros *et al.*, 2014] R. Kiros, R. Salakhutdinov, and R.S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *arXiv*, 2014.
- [Layne *et al.*, 2012a] Ryan Layne, Tim Hospedales, and Shaogang Gong. Person re-identification by attributes. In *BMVC*, 2012.
- [Layne *et al.*, 2012b] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Towards person identification and re-identification with attributes. In *ECCV*, 2012.
- [Layne *et al.*, 2014a] Ryan Layne, Tim Hospedales, and Shaogang Gong. Re-id: Hunting attributes in the wild. In *BMVC*, 2014.
- [Layne *et al.*, 2014b] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. *Attributes-Based Re-identification*. 2014.
- [Li *et al.*, 2015] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, 2015.
- [Li *et al.*, 2017] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *ICCV*, 2017.
- [Lin *et al.*, 2017] Yutian Lin, Liang Zheng, and Wu Yu and Yang Yi Zheng, Zhedong and. Improving person re-identification by attribute and identity learning. In *arXiv*, 2017.
- [Long and Wang, 2015] Mingsheng Long and Jianmin Wang. Learning transferable features with deep adaptation networks. 02 2015.
- [Mineiro and Karampatziakis, 2014] Paul Mineiro and Nikos Karampatziakis. A randomized algorithm for cca. In *CoRR*, 2014.
- [Ngiam *et al.*, 2011] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, 2011.
- [Reed *et al.*, 2016] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.
- [Scheirer *et al.*, 2012] W. Scheirer, N. Kumar, P. Belhumeur, and T. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *CVPR*, 2012.
- [Siddiquie *et al.*, 2011] B. Siddiquie, R. S. Feris, and L. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.
- [S.Li *et al.*, 2017] S.Li, T.Xiao, H.Li, and X.Wang et al. Person search with natural language description. In *CVPR*, 2017.
- [Su *et al.*, 2015a] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV*, 2015.
- [Su *et al.*, 2015b] Chi Su, Shiliang Zhang, Fan Yang, Guangxiao Zhang, Qi Tian, Wen Gao, and Larry S. Davis. Tracklet-to-tracklet person re-identification by attributes with discriminative latent space mapping. In *ICMS*, 2015.
- [Su *et al.*, 2016] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, 2016.
- [Su *et al.*, 2018] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Multi-type attributes driven multi-camera person re-identification. In *PR*, 2018.
- [Sun and Saenko, 2016] Baochen Sun and Kate Saenko. Deep coral: correlation alignment for deep domain adaptation. In *ICCV*, 2016.
- [Tzeng *et al.*, 2014] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, 2014.
- [Vaquero *et al.*, 2009] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *WACV*, 2009.
- [Wang *et al.*, 2015] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *ICML*, 2015.
- [Wei *et al.*, 2017] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE T. CYBERN*, 2017.
- [Yang *et al.*, 2016] Y. Yang, Z. Lei, S. Zhang, H. Shi, and S. Z. Li. Metric embedded discriminative vocabulary learning for high-level person representation. In *AAAI*, 2016.
- [Zhao *et al.*, 2014] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014.