Cross-Domain Adversarial Feature Learning for Sketch Re-identification

Lu Pang[1,2], Yaowei Wang[3*], Yi-Zhe Song[4], Tiejun Huang[1,2], Yonghong Tian[1,2*]

[1] Peking University Shenzhen Graduate School, Shenzhen, China
[2] National Engineering Laboratory for Video Technology, School of EE&CS, Peking University, Beijing, China
[3] School of Information and Electronics, Beijing Institute of Technology, Beijing, China
[4] SketchX, Queen Mary University of London, London, UK

ABSTRACT

Under person re-identification (Re-ID), a query photo of the target person is often required for retrieval. However, one is not always guaranteed to have such a photo readily available under a practical forensic setting. In this paper, we define the problem of Sketch Re-ID, which instead of using a photo as input, it initiates the query process using a professional sketch of the target person. This is akin to the traditional problem of forensic facial sketch recognition, yet with the major difference that our sketches are whole-body other than just the face. This problem is challenging because sketches and photos are in two distinct domains. Specifically, a sketch is the abstract description of a person. Besides, person appearance in photos is variational due to camera viewpoint, human pose and occlusion. We address the Sketch Re-ID problem by proposing a cross-domain adversarial feature learning approach to jointly learn the identity features and domain-invariant features. We employ adversarial feature learning to filter low-level interfering features and remain high-level semantic information. We also contribute to the community the first Sketch Re-ID dataset with 200 persons, where each person has one sketch and two photos from different cameras associated. Extensive experiments have been performed on the proposed dataset and other common sketch datasets including CUFSF and QUML-shoe. Results show that the proposed method outperforms the state-of-the-arts.

KEYWORDS

Sketch re-identification, Adversarial feature learning, Cross-doman matching, Domain-invariant features

ACM Reference Format:
Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, Yonghong Tian. 2018. Cross-Domain Adversarial Feature Learning for Sketch Reidentification. In 2018 ACM Multimedia Conference (MM' 18), October 22–26, 2018, Seoul, Republic of Korea. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240606

Corresponding author: Yaowei Wang, Yonghong Tian(email: yaoweiwang@bit.edu.cn, yhtian@pku.edu.cn).

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240606

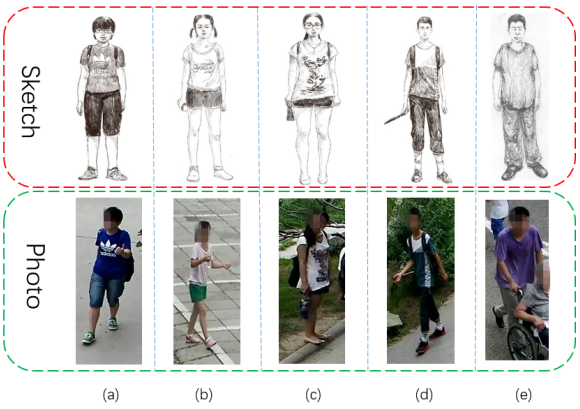

Figure 1: Challenges of sketch Re-ID. Sketches lack color information and contain person's outline information. (a-b) Photos are affected by camera viewpoint. (b-c) Human pose is various. (c-d) Photos are full of cluttered background. (e) Person is occluded.

1 INTRODUCTION

Person re-identification (Re-ID) aims to match a photo query of the suspect within a gallery database. Despite great strides made, a key problem that was often neglected lies with the availability of such a photo query – it is commonplace that usable suspect photos can not be readily obtained. This problem had long been recognized by law enforcements, and have motivated research on forensic facial sketches matching [15, 32] where a facial sketch drawn by a professional artist according to the description of an eyewitness is matched to a dataset of mugshot photos. In this paper, we extend this challenging cross-domain matching problem to person re-identification by proposing for the first time the problem of sketch re-identification (sketch Re-ID), where whole-body sketches (other than just facial ones) are used to match against a photo gallery database.

Sketch Re-ID is of great importance for law enforcement. For example, when a criminal is witnessed but not be photographed by a surveillance camera, a sketch Re-ID system can automatically search all the surveillance videos to locate this criminal according to an artist's drawing. Consequently, police can cut down the number of suspects quickly and focus on those potential suspects. Once a suspect can be captured by a surveillance camera, his behaviors can be analyzed and all the witnesses around him can also be tracked successfully, which help police save lots of manpower and material.

The Sketch Re-ID problem is a challenging task due to the large domain gap between sketches and photos. As shown in Fig. 1, comparing to a photo with cluttered background, a sketch lacks color information and contains person's outline information and little on texture information. Besides, a sketch shows the person with the frontal view while a person photo can come from any camera and can be affected by many factors such as camera view, human pose and occlusion. Communication barriers between eye-witnesses and artists also increase the gap of two domains. As a result, a sketch can have a great discrepancy with the identical person photo. Resembling person Re-ID, there are many similar persons for the gallery set, which makes sketch Re-ID a more challenging problem. Finally, although the sketch Re-ID problem is promising and important for surveillance applications and public security, it is impracticable to study mainly because of the lack of a fit-for-purpose dataset.

We address all above challenges by proposing a model for sketch Re-ID and introducing a sketch Re-ID dataset. For the dataset, it consists of 200 persons, and there are two photos from two different cameras and one sketch for each person. For the model, we propose a deep adversarial learning architecture to jointly learn distinguishable identity features and domain-invariant features. Human visual system distinguishes a sketch from the matched photo mainly by high-level semantic features. This is because the gap between sketches and photos is very large. Some low-level features cannot be regarded as distinguishable features between two domains and even can disturb retrieval results. Namely, high-level semantic features are truly domain-invariant features. Motivated by this, the aim of our model is to filter low-level interfering features which are quite sensitive for two domains. In our model, we introduce a domain discriminator and two feature generators for sketches and photos, respectively. The objective of domain discriminator is to differentiate sketches from photos. The objective of sketch feature generator is to generate features of sketches similar to those of photos. And the objective of photo feature generator is to generate features of photos similar to those of sketches. By playing a min-max game between the domain discriminator and two feature generators, those low-level interfering features especially sensitive to domains are filtered while the generated features of sketches and photos remain domain-invariant. In the training phase, we feed a pair features of a photo/sketch pair into a domain discriminator. Our model learns domain-invariant features by jointly optimizing multiple objectives including classification loss, pairwise loss and adversarial loss. In the testing stage, we use feature generators to extract features and compute Euclidean distance between sketch and photo features. Finally, we introduce the pre-training stage using auxiliary datasets to solve the data insufficiency problem. This is because, despite we providing a large number of annotations for the dataset, it is not sufficient to train a deep learning network.

Our contributions are summarized as follows: (1) We develop a deep adversarial learning architecture to jointly learn identity features and domain-invariable features by filtering low-level features and remaining high-level semantic features. Experiments show that our model obtains the state-of-the-art performance on the sketch Re-ID dataset. And extensive experiments demonstrate that our proposed cross-domain adversarial feature learning method also shows remarkable performance on other two sketch-photo datasets, including CUFSF dataset [32] and QMUL-shoe dataset [29]. (2) We introduce a sketch Re-ID dataset containing 200 persons, in which each person has one sketch and two photos. We believe that this dataset opens up the opportunity for the research of this challenging problem.

2 RELATED WORK

This work is related to person Re-ID with deep models, cross-domain image-to-image translation and fine-grained sketch-based image retrieval. The following three subsections review some works on these parts separately.

2.1 Person Re-ID with Deep Models

Deep learning has become the popular method for person Re-ID [1, 3, 9, 33, 36]. Most deep-learning-based person Re-ID models can be divided into two sub-networks, namely, feature learning sub-network and distance metric learning sub-network. For the feature learning network, some researchers have adopted a classical convolution neural network (CNN) like GoogleNet [25] and ResNet [11] while others have designed their network architectures for particular purposes [34]. Instead of extracting deep features only from the whole body, detailed part cues have recently been considered to distinguish visually-similar persons. Some researchers learnt local features from multiple body parts of persons [26, 27], demonstrating more effective than dividing person images into several fixed-length strips [4]. Meanwhile, some other works [35, 37] focused on tackling the pose variation problem. Zhang et al. [33] aligned persons by dynamically matching the local part features from top to bottom with the minimum total distance. Zheng et al. [37] took person images generation to enhance the performance. Moreover, most of these methods (e.g., [26, 33]) had multiple branches to learn global features and local part features jointly. These branches promoted each other to learn the detailed features. It should be noted that the objective of our feature learning sub-network is not to focus on local detailed features, but to learn high-level semantic domain-invariant features where GoogleNet is used as the base network.

For the distance metric learning network, different losses such as classification loss [26], pairwise verification loss [16] and triplet ranking loss [4, 20] have been adopted. These losses were designed to pull the similarity of those images from the same persons closer and push different persons more distant. The feature learning network optimized by these losses could generate more discriminative identity features. However, some information among identity features might be sensitive to domains such as color and low-level texture information. Thus by introducing adversarial feature learning, our proposed framework can filter these sensitive information and learn domain-invariant features.

2.2 Cross-domain Image-to-Image Translation

A large number of GAN-based models (e.g., [2, 13, 19, 28]) on cross-domain image translation have been proposed recently. GAN [10] plays a min-max game between generator F and discriminator D. It is formulated as follows:

$$\min_F \max_G L = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(F(z)))] \quad (1)$$

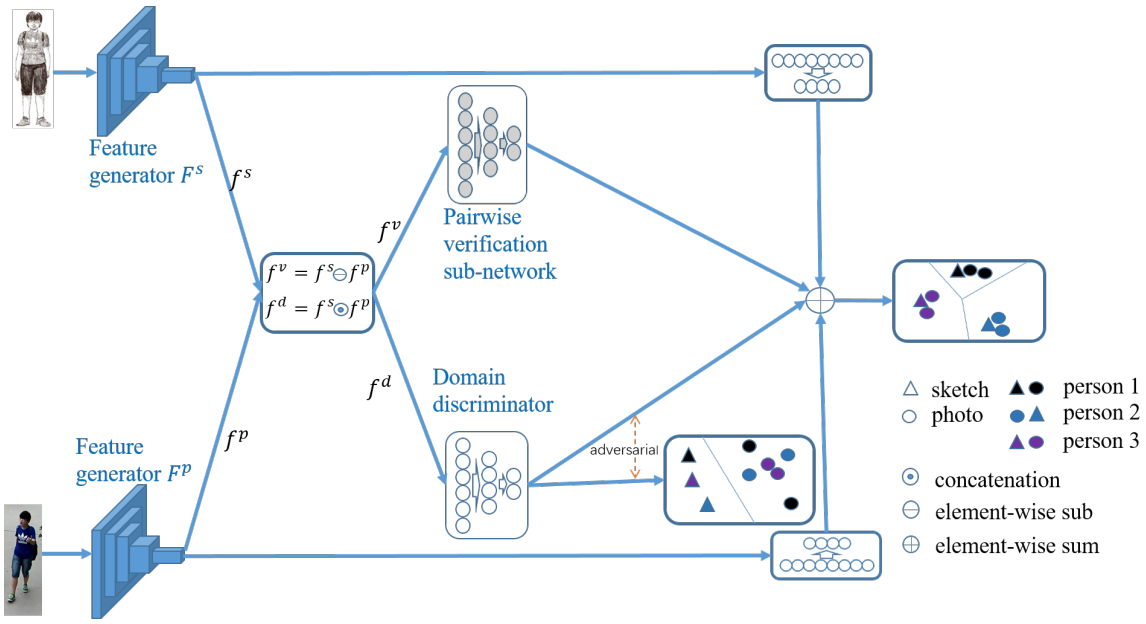

Figure 2: Architecture of the proposed cross-domain adversarial feature learning approach. We have two feature generators for sketches and photos, respectively. Then we feed two features into a domain discriminator. The losses output by the domain discriminator are two opposite objectives. Meanwhile two intra-domain classifiers are trained for person label classification and a pairwise verification network is introduced to judge whether they are the same person or not.

where the task of D is to distinguish training images and generated images and the objective of F is to generate authentic images to fool D. On this basis, many versions of GAN have been developed. Phillip et al. [13] took an image as the condition of the generator and then successfully translate edge images into color images. DualGAN [28] and CycleGAN [38] allowed to mutually learn distributions of images between two domains to address the problem of lacking human annotations. Basically, these GANs were designed for image generating, and consequently features extracted from those generated images were not effective enough for the cross-domain retrieval task because GAN simply mimics data distribution. Instead, CoGAN [18] took discriminator outputs to train a classifier, while BiGAN [6] learnt an inverse mapping from training images to the representation space and could learn useful features for image classification. To learn more effective features for sketch Re-ID which is essentially a cross-domain image-to-image translation problem rather than an image classification problem, our generators try to generate useful features instead of images while the discriminator is to distinguish features from the sketch domain and photo domain. By playing a min-max game between the feature generators and the discriminator, our model can map sketches and photos into a domain-variant space.

2.3 Fine-Grained Sketch-Based Image Retrieval

There are many works [7, 8, 12, 31] focusing on category-level sketch classification and sketch-to-photo retrieval. In [17], Li et al. firstly adopted a deformable part-based model (DPM) representation to address the fine-grained sketch-based image retrieval (FG-SBIR) problem. Three datasets on fine-grained sketch-based image retrieval (FG-SBIR) were released recently. Several FG-SBIR models [21, 24, 29] were developed on these datasets. Yu et al. [29] learnt an embedding space for sketch and photo domains. Song et al. [24] employed the attention model and high order distance function to solve the misalignment problem of two domains. Note that in order to eliminate the gap between two domains, the models in [24, 29] needed an additional preprocessing to extract edgeMaps from photos. Pang et al. [21] employed the generative network to reconstruct sketches and a hybrid model to enhance the performance of their discriminative network. Compared with these works, our model employs adversarial feature learning to reduce domain gap, as well as to extract domain-invariant features.

3 CROSS-DOMAIN ADVERSARIAL FEATURE LEARNING

3.1 Problem Definition

Given a person sketch x^s, the purpose of sketch Re-ID is to match and return all person photos having the same label with x^s from a gallery photos set.

Assume we have a probe-sketch set of S person sketches $X^s = \{x_k^s, y_k^s\}_{k=1}^S$, where x_k^s and y_k^s denote the sketch image and the corresponding sketch label of the k-th person. And we also have a gallery-photo set of P person photos $X^p = \{x_k^p, y_k^p\}_{k=1}^P$, where x_k^p and y_k^p are the photo and the corresponding photo label of the k-th person. In the training stage, a sketch feature extractor F^s and a photo feature extractor F^p are learned. In the test stage, given $(x^s, y) \in X^s$, we extract the sketch representation as $F^s(x^s)$ and all gallery representations as $\{F^p(x_k^p)\}_{k=1}^P$, and then compute

the similarity distance measured with euclidean distance between $F^s(x^s)$ and each gallery representation. The objective function of sketch Re-ID is formulated as follows:

$$\min \sum_{k=1}^{P} \|F^s(x^s) - F^p(x_k^p)\|_2^2 q_k \tag{2}$$

where $q_k = 1$ if $y = y_k^p$ else $q_k = 0$.

Aiming to deal with the large gap between sketches and photos, we propose a cross-domain adversarial feature learning architecture for sketch Re-ID. The architecture contains two parts, including identity feature learning and adversarial feature learning. Identity feature is distinguishable for different persons in one specific domain, while adversarial learning can make the feature distributions from two domains similar. The two parts mutually enhance each other. In the following, we will further describe the details of our proposed architecture.

3.2 Joint Feature Learning Architecture

Our joint feature learning architecture is shown in Fig. 2. Our model has two feature generators and takes a sketch/photo pair as input, then learns features from two domains. We train two label classifiers on the features from different domains, respectively. Besides, a pairwise verification sub-network is trained to judge whether two features belong to one identity. The discriminator receives features from two domains and learns a classifier on domain labels. Meanwhile the adversarial feature learning needs to generate features that are more distinguishable to identities but invariant to different domains.

Existed person Re-ID methods [1, 3, 9, 33, 36] mostly focus on the distinguishable identity feature learning of different persons. However, the learned identity features often contain too much low-level detailed information and thus cannot be directly applied to the sketch domain, since a sketch is only the outline of the query person. To solve this problem, we introduce the cross-domain adversarial feature learning, which can reserve cross-domain high-level features and filter the low-level details. We jointly optimize two kinds of objectives with adversarial learning. Let $L_{C^s}(F^s, C^s)$ and $L_{C^p}(F^p, C^p)$ be the identity classifier losses of the sketch domain and the photo domain, $L_V(F^s, F^p, V)$ be the identity pairwise verification loss and $L_{advF}(F^s, F^p, D)$ be the cross-domain adversarial feature learning loss. The objective of our model is formulated as follows:

$$\begin{aligned} \min_{F^s, F^p} L_{final} &= L_{C^s}(F^s, C^s) \\ &+ L_{C^p}(F^p, C^p) \\ &+ \alpha L_V(F^s, F^p, V) \\ &+ \beta L_{advF}(F^s, F^p, D) \end{aligned} \tag{3}$$

where F^s and F^p are the feature generation networks of sketch and photo, C^s and C^p are classification networks for sketches and photos, respectively, V is the pairwise verification network and D is the domain discriminator network. And α and β are weights of the pairwise verification loss and the adversarial feature learning loss.

In this formulation, the identity feature learning loss includes two classification losses $L_{C^s}(F^s and C^s)$, $L_{C^p}(F^p, C^p)$, and a pairwise verification loss $L_V(F^s, F^p, V)$, whose details are shown on sec. 3.3. The cross-domain adversarial feature learning frame will be shown on sec.3.4.

3.3 Identity Feature Learning

The two popular identity feature learning losses are classification loss and pairwise verification loss. Our model has two feature generators for sketches and photos, separately. For each domain, our model trains an identity classifier. Assume we have a probe-sketch set containing S persons, and a gallery-photo set containing P persons. Then the last FC-layer of C^s outputs $C^s(F^s(x_k^s)) = [e_1^s, e_2^s, ...e_S^s] \in R^S$. The objective of the sketch classifier can be formulated as follows:

$$\min_{F^s, C^s} L_{C^s}(F^s, C^s) = -\sum_{i=1}^{S} \log(z_i^s) q_i^s \tag{4}$$

where $z_k^s = \frac{\exp(e_k^s)}{\sum_{i=1}^{S} \exp(e_i^s)}$ and $q_i^s = 1$ if $i = y_k^s$ else $q_i^s = 0$.

Similarly, the objective of the photo classifier can be formulated as follows:

$$\min_{F^p, C^p} L_{C^p}(F^p, C^p) = -\sum_{i=1}^{P} \log(z_i^p) q_i^p \tag{5}$$

where $z_k^p = \frac{\exp(e_k^p)}{\sum_{i=1}^{P} \exp(e_i^p)}$ and $q_i^p = 1$ if $i = y_k^p$ else $q_i^p = 0$.

Moreover, we feed a generated feature pair into the pairwise verification sub-network to verify whether the given sketch and photo are regarding the same person or not. Assume the last FC-layer of V outputs $V(F^p(x_n^p), F^s(x_n^s)) = [e_1, e_2]$, and we have a set of H sketch/photo pairs, then loss function of the pairwise verification sub-network is summarized as bellow:

$$\min_{F^s, F^p, V} L_V(F^s, F^p, V) = -\sum_{n}^{H} V(F^p(x_n^p), F^s(x_n^s)) q_n \tag{6}$$

where $q_n = 1$ if x_n^s and x_n^p are the identical person, otherwise $q_n = 0$.

We introduce above two sorts of losses to train a representation that is distinguishable for different persons. However, these losses cannot really solve the cross-domain problem. So we introduce adversarial feature learning to address this problem.

3.4 Adversarial Feature Learning

We regard a representation as a domain-invariant feature when a domain classifier cannot discriminate domains they come from. The objective of the domain discriminator can be formulated as follows:

$$\begin{aligned} \max_D L_{advD}(F^s, F^p, D) = &\mathbb{E}_{x^s \sim p_{data}(x^s)}[\log D(F^s(x^s))] \\ &+ \mathbb{E}_{x^p \sim p_{data}(x^p)}[\log 1 - D(F^p(x^p))] \end{aligned} \tag{7}$$

where the label indicates the sketch domain.

At the same time, we train two feature generators to generate robust features to the two domains, as well as to map two domain images into an domain-invariant feature space by imitating the

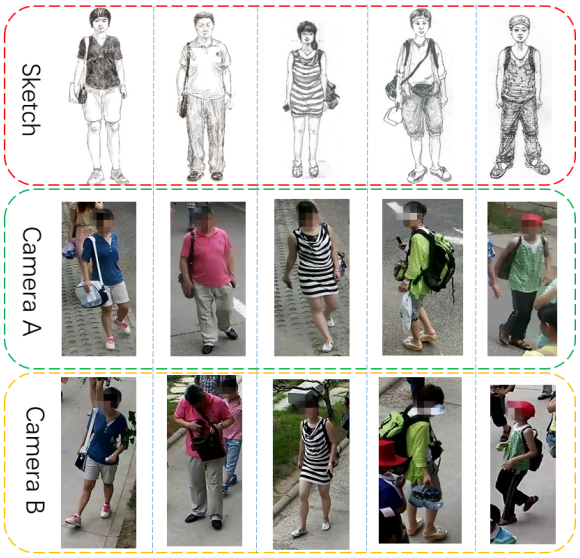

Figure 3: Examples of persons with sketches and photos from two camera views. Sketches are shown on the first row, and the middle and below rows show photos from camera A and camera B, respectively. Three images of each column are taken or drawn for the identical person.

distribution of two domains mutually. The objective for our feature generators can be formulated as follows:

$$\begin{aligned}\min_{F_s,F_p} L_{advF}(F^s,F^p,D) = \mathbb{E}_{x^s\sim p_{data}(x^s)}[\log D(F^s(x^s))] \\ + \mathbb{E}_{x^p\sim p_{data}(x^p)}[\log 1 - D(F^p(x^p))] \\ + \mathbb{E}_{x^p\sim p_{data}(x^p)}[\log D(F^p(x^p))] \\ + \mathbb{E}_{x^s\sim p_{data}(x^s)}[\log 1 - D(F^s(x^s))]\end{aligned} \tag{8}$$

It can be observed that the sketch feature generator is trained to mimic the distribution of photo features and meanwhile the photo feature generator is trained to imitate the distribution of sketch features. The adversarial feature learning loss simultaneously optimize F^s and F^p.

Obviously, the objectives of the domain discriminator and the feature generators are adversarial. It is impossible that the domain discriminator and the feature generators can be both optimal. We actually optimize the domain discriminator and the feature generators alternately.

4 SKETCH RE-ID DATASET

4.1 Dataset Construction

Till now, there are no the sketch Re-ID dataset available publicly. To address this problem, we constructed a sketch Re-ID dataset in this study. The dataset contains 200 persons, each of which has one sketch and two photos. Some examples are shown on Fig. 3. To ensure the dataset created for realistic surveillance system, photos of each person were captured during daytime by two cross-view cameras. We cropped the raw images (or video frames) manually to make sure that every photo contains the one specific person.

Table 1: The numbers of sketches for each painting styles. Labels of the painting styles correspond to Fig. 4. Data split for the training set and testing set is shown on the last column.

Style Category	Number	Training:Testing
(a)	45	34:11
(b)	20	15:5
(c)	80	60:20
(d)	33	25:8
(e)	22	16:6
Total	200	150:50

To embody the forensic sketch-photo matching scene in reality, some volunteers played the role of eyewitnesses and communicated with professional sketch artists. As the eyewitnesses, these volunteers described the suspect appearance after seeing the photo for some time. Before the sketch artist painted the finished product, he/she would modify the sketch several times according to statements of eyewitnesses. We have a total of 5 artists to draw all persons' sketches and every artist has his own painting style. Fig.4 shows some samples of all different styles. Table 1 shows the numbers of sketches from different painting styles. Because the painting style transfer is not considered as a part of the sketch Re-ID problem, we randomly select $\frac{3}{4}$ persons from each painting style for training and $\frac{1}{4}$ for testing to eliminate effects of painting styles. Table 1 also reports the quantities of sketches for each painting style contained in the training set and testing set. It can be observed that the distribution of person sketches from five painting styles is not uniform. Overall, we have 150 persons for training and 50 persons for testing.

We provide identity number (ID) annotations for every sketch and every photo. More specifically, the identity number is unique and all sketches and photos that contain the same person have the same ID.

4.2 Comparison with Other Sketch Datasets

In Fig. 5, we compare our sketch Re-ID dataset with other two sketch datasets including CUFSF dataset [32] and QMUL-shoe dataset [29]. CUFSF is a viewed face sketch-photo dataset, including 1194 persons captured with lighting variations and its sketches were drawn by an artist with shape exaggeration. While QMUL-shoe is a fine-grained sketch-based image retrieval dataset which includes 419 shoes. These photos were collected from on-line shopping websites and its sketches were free-hand drawn by crowd. Our sketch Re-ID dataset and CUFSF dataset were designed for law enforcement and thus their sketches were all drawn by professional artists. Their difference is that CUFSF [32] is a face sketch-photo recognition dataset while our sketch Re-ID dataset focuses on person bodies. On the contrary, QMUL-shoe dataset was designed for the commercial purpose, whose sketches were free-hand and were drawn by amateurs. We can observe that the free-hand sketches are more sparse which needs more pre-training data to avoid the over-fitting.

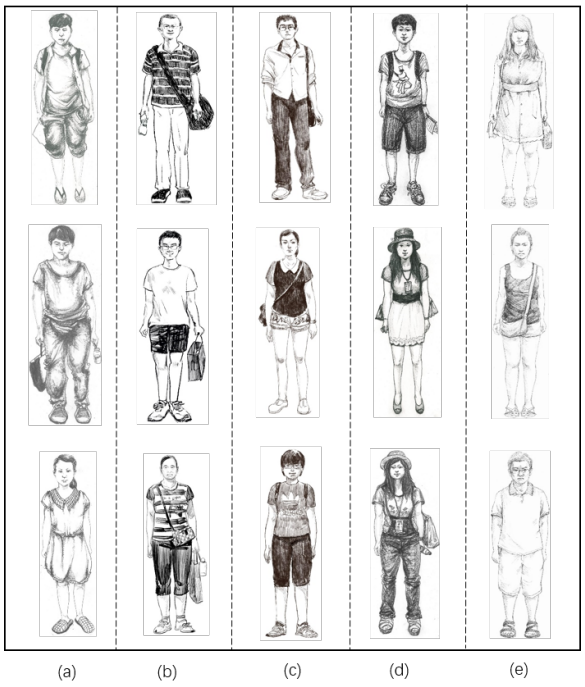

Figure 4: Examples of different painting styles. There are total five styles. The sketches in the same column belong to the same painting style. Here we show three sketches for each style.

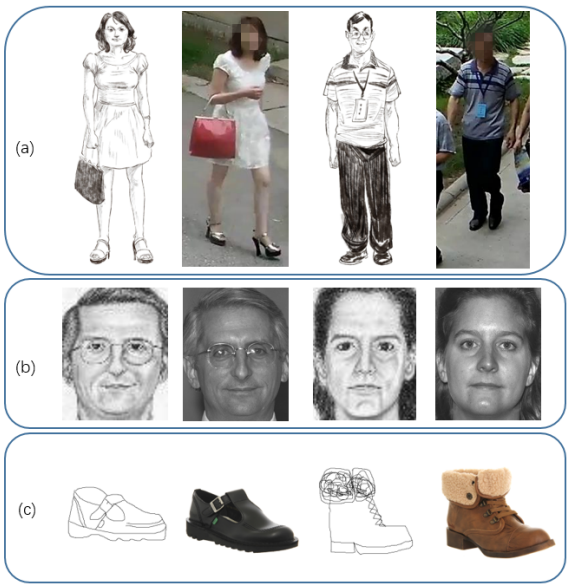

Figure 5: Examples of three sketch datasets. Sketches lack color information and contain the outline information of the give objects. (a) Our sketch Re-ID dataset. (b) CUFSF dataset [32]. (c) QMUL-shoe dataset [29].

5 EXPERIMENTS

The experiments were conducted on the sketch Re-ID dataset to evaluate the performance of our model. To further validate the generalization of our model on similar sketch-photo matching tasks, we also performed the experiments on the other two publicly-available sketch datasets, including CUFSF [32] and QMUL-shoe [29]. For CUFSF dataset, we randomly selected 500 and 694 persons as train and test set according to [32]. For QMUL-shoe dataset, we picked 304 sketch-photo shoe pairs for training and the rest were used for testing.

5.1 Experiments Details

Our model was implemented on Caffe [14]. The two feature generators were modified from GoogleNet [25]. Each feature generator was pre-trained on ImageNet [5]. The 'pool5/7×7' layer of GoogleNet [25] is taken as the last layer of our feature generator. The two classifier sub-networks consist of one fc layer, whose dimension is the category number. The pairwise verification sub-network consists of two fc layers, whose dimensions are 512 and 2, respectively. The domain discriminator contains one convolution layer with 1×1 kernel size and two fc layers with dimensions 512 and 2 separately.

Inspired by [29] and [24], we handled the dataset insufficient problem by pre-training on the public person Re-ID datasets, like Market1501 [36], CUHK03 [16] and Dukemtmc [22]. In our experiments, we pretrained our model on Market1501 [36] without cross-domain adversarial feature learning module. Then we fine-tuned our model on the sketch Re-ID dataset. Besides, we made data augmentation by cropping a 224×224 image as input and we did flipping with 0.5 probability. During the pre-training stage, the learning-rate was initialed as 0.001, and set as 0.0005 in the fine-tuning stage. When testing, we extracted outputs of the 'pool5/7×7' layer of GoogleNet [25] as features and computed the Euclidean distance as the distance score of a sketch-photo pair.

5.2 Results on sketch Re-ID dataset

We have three baselines including one hand-crafted based model and two deep models for this dataset. We perform the experiments ten times and take the average as the final result. Dense-HOG+LBP+rankSVM is the representation of hand-crafted based model. HOG is the classical feature and LBP is an assistant feature for sketch recognition. We concatenate the two features and a rankSVM is used to rank scores. Triplet SN [29] is the model designed for free-hand object sketches. Follow [29], after pre-training the model on edge-maps of ImageNet [5] and TU-Berlin dataset [7], we extract edge-maps of person photos and feed these edge-maps and person sketches into Triplet SN. GN Siamese [23] is a network with two GoogleNet [10] branches optimized by pairwise verification loss. The model is proposed by [23] as a baseline of learning an embedding space of two domains. For fairness the model is pre-trained on the public person re-identification dataset Market1501 [36]. All methods are evaluated by rank-1 accuracy. Performance of these methods is shown on Table 2.

From Table 2, we can see that two GoogleNet-based deep learning models are superior than other baselines and our model displays the state-of-the-art performance. Moreover, our proposed cross-domain

Table 2: Comparative results against other baselines on sketch Re-ID dataset.

sketch Re-ID dataset	rank1	rank5	rank10	rank20
Dense-HOG+LBP+rankSVM	5.1%	16.8%	28.3%	37.9%
Triplet SN [29]	9.0%	26.8%	42.2%	65.2%
GN Siamese [23]	28.9%	54.0%	62.4%	78.2%
our model	**34.0%**	**56.3%**	**72.5%**	**84.7%**

Table 3: Comparative results against other methods on CUFSF dataset.

CUFSF dataset	VR at 0.1% FAR
Single CITP [15]	93.95%
Single GS [15]	96.32%
Fused CITP [32]	98.70%
Fused GS [15]	99.14%
our model	**99.46%**

adversarial learning model is clear at rank1 around 5% increase against GN Siamese without adversarial feature learning. The hand-craft features-based model reports the worst result, which suggests that hand-craft features do not work for this problem. Triplet SN [29] is another inefficient baseline. This is because Triplet SN [29] is specialized for object sketches. These objects are full-frontal, so their edge-maps and sparse free-hand sketches are alike. However, person sketch is distinct to photo edge-map and it needs a method to bridge the gap between two domains.

5.3 Results on CUFSF dataset

For this dataset, we follow [15] and [32], the model is evaluated by Verification Rate(VR) at 0.1% False Acceptance Rate(FAR). According to [15] and [32] we randomly select 500 persons for training and 694 persons for testing. We choose four hand-crafted models which have excellent performance on this dataset. GS is a domain-invariant descriptor proposed by [15] and it is computed by applying Gabor filters to face images. [32] develops a coupled information-theoretic projection(CITP) tree for converting a photo or sketch into discrete codes.

The performances of all methods are shown on Table 3. The results of four baselines are copied from [15] and [32]. We can observe that though they perform well on this dataset, our model outperforms the state-of-the-art method. This demonstrates that our model is universal and can be applied to human body and face sketch-based photo recognition.

5.4 Results on QMUL-Shoe dataset

For this dataset, we follow [24] and [29] and the percentage of correctly predicting the true match at top-1 and top-10 is used as the evaluation metrics. Table 4 reports performance of all comparable methods. According to [24] and [29], we choose HOG-BoW+rankSVM and Dense-HOG+rankSVM as two compared hand-crafted models. HOG features are popular hand-crafted features for sketch recognition before deep features are generally used and many HOG features over grids are concatenated to form Dense-HOG features. A Ranking SVM is a variant of the support vector machine algorithm and is employed to compute the final ranking scores. ISN Deep+rankSVM uses deep features learned from Sketch-a-Net [30], and deep features are feed into a rankSVM to calculate ranking scores. Triplet SN [29] is composed of three identical Sketch-a-Nets and is optimized by triplet ranking loss. DSSA Triplet is the model of [24], which adds three models including attention model, coarse-fine fusion and HOLEF loss to the model of [29].

Table 4: Comparative results against other methods on QMUL-Shoe dataset. DSSA Triplet result is copied from [24] and other four baselines results are copied from [29].

QMUL-Shoe	top-1	top-10
HOG-BoW+rankSVM	17.39%	67.83%
Dense-HOG+rankSVM	24.35%	65.22%
ISN Deep+rankSVM	20.00%	62.61%
Triplet SN [29]	39.13%	87.83%
DSSA Triplet [24]	**61.74%**	**94.78%**
our model	56.35%	92.64%

Table 5: Ablation study of adversarial feature learning on sketch re-id dataset, QMUL-Shoe dataset and CUFSF dataset. AFL is the abbreviation of adversarial feature learning.

sketch Re-ID dataset	rank1	rank5	rank10	rank20
our model without AFL	30.1%	42.2%	60.6%	80.3%
our model	**34.0%**	**56.3%**	**72.5%**	**84.7%**

QMUL-Shoe	top-1	top-10
our model without AFL	51.49%	90.17%
our model	56.35%	92.64%

CUFSF dataset	VR at 0.1% FAR
our model without AFL	99.43%
our model	**99.46%**

The performance of all methods suggests that deep models are more effective than other methods. DSSA model [24] reports the state-of-the-art performance, maybe cause edge-Maps and sparse free-hand drawn sketches being too similar and they being considered as one domain. Besides, DSSA solves the misalignment of two domains. Our model does not handle the misalignment in essence while misalignment is common for the free-hand drawn sketch dataset like QMUL-Shoe, our model performs less well than DSSA. However, our model is designed to learn cross-domain features, which is less useless than [24] for this problem.

5.5 Ablation Study

In our model, we have learned domain-invariant features by jointly optimizing identity feature learning losses and adversarial feature learning loss. We employ adversarial feature learning to filter interfering features that learned by optimizing identity feature learning

Figure 6: The retrieval results of our model and our model without adversarial feature learning(AFL) on three sketch datasets. For each example, the top row is the result of our model and the bottom row is the result of our model without AFL.

losses. In order to evaluate the contribution of adversarial feature learning, we compare our integral model with an elementary version by removing adversarial feature learning(our model without AFL). We conduct experiments on three datasets, respectively. Table 5 reports the results on our sketch re-id dataset, QUML-Shoe dataset and CUFSF dataset. The results show that adversarial feature learning improves our proposed architecture.

5.6 Qualitative Results

We also provide examples of retrieval results of our model on three sketch datasets in Figure 6, compared to our model without adversarial feature learning. We can observe that our proposed model focuses more on the overall outline with adversarial feature learning. For example, on the first example of sketch re-id dataset, the correct person is retrieved as Rank 1 because our model concentrates on person's outline with a pendant on the chest. Similarly on shoes, by attending to the outline with shoelaces and heels, our model can retrieve the correct shoe. About CUFSF dataset, instead of employing the original evaluation criterion(VR at 0.1% FAR), we show two examples by predicting the true match at rank-1. It can be seen that with adversarial feature learning, our model focuses on the high-level outline information.

6 CONCLUSION

In this paper, we introduce sketch Re-ID problem and this problem is more challenging than person Re-ID. For the first time, we address the sketch Re-ID by proposing a cross-domain adversarial feature learning method. Our model can jointly learn identity features and domain invariant features. Compared to other methods, our model reports the state-of-the-art performance on the sketch Re-ID dataset and CUFSF dataset and gains good results on the QMUL-shoe dataset. Experiments demonstrate that our proposed model is universal for person, face and object sketch retrieval problem. For the absence of sketch Re-ID dataset, we present a sketch Re-ID dataset containing 200 persons and each person includes one artist drawn sketch and two photos captured by different cameras.

ACKNOWLEDGMENTS

This work is partially supported by grants from the National Key R&D Program of China under grant 2017YFB1002401, the National Natural Science Foundation of China under contract No. U1611461, No. 61471042, No. 61390515 and No. 61425025, also supported by grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer.

REFERENCES

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3908–3916.

[2] Qifeng Chen and Vladlen Koltun. 2017. Photographic image synthesis with cascaded refinement networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Vol. 1.

[3] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. 2018. Person re-identification by camera correlation aware feature augmentation. *IEEE transactions on pattern analysis and machine intelligence* 40, 2 (2018), 392–408.

[4] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1335–1344.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

[6] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. 2016. Adversarial feature learning. *arXiv preprint arXiv:1605.09782* (2016).

[7] Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM Trans. Graph.* 31, 4 (2012), 44–1.

[8] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. 2011. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on visualization and computer graphics* 17, 11 (2011), 1624–1636.

[9] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. 2016. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244* (2016).

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Rui Hu and John Collomosse. 2013. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding* 117, 7 (2013), 790–806.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. *arXiv preprint* (2017).

[14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 675–678.

[15] Hamed Kiani Galoogahi and Terence Sim. 2012. Face photo retrieval by sketch example. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 949–952.

[16] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 152–159.

[17] Yi Li, Timothy M Hospedales, Yi-Zhe Song, and Shaogang Gong. 2014. Fine-grained sketch-based image retrieval by matching deformable part models. (2014).

[18] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *Advances in neural information processing systems*. 469–477.

[19] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[20] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. 2015. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1846–1855.

[21] Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. 2017. Cross-domain generative learning for fine-grained sketch-based image retrieval. BMVC.

[22] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*.

[23] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 119.

[24] Jifei Song, Yu Qian, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*.

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. 2015. Going deeper with convolutions. Cvpr.

[26] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. 2017. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 420–428.

[27] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. 2017. Deep representation learning with part loss for person re-identification. *arXiv preprint arXiv:1707.00798* (2017).

[28] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint* (2017).

[29] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen Change Loy. 2016. Sketch me that shoe. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 799–807.

[30] Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. 2015. Sketch-a-net that beats humans. *arXiv preprint arXiv:1501.07873* (2015).

[31] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. 2016. Sketchnet: Sketch classification with web images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1105–1113.

[32] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 513–520.

[33] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. 2017. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184* (2017).

[34] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1077–1085.

[35] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. 2017. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732* (2017).

[36] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.

[37] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717* 3 (2017).

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593* (2017).