Instance-level Heterogeneous Domain Adaptation for Limited-labeled Sketch-to-Photo Retrieval

Fan Yang, *Member, IEEE,* Yang Wu, *Member, IEEE,* Zheng Wang, *Member, IEEE,* Xiang Li, Sakriani Sakti, *Member, IEEE,* and Satoshi Nakamura, *Fellow, IEEE*

***Abstract*—Although sketch-to-photo retrieval has a wide range of applications, it is costly to obtain paired and rich-labeled ground truth. Differently, photo retrieval data is easier to acquire. Therefore, previous works pre-train their models on rich-labeled photo retrieval data (i.e., source domain) and then fine-tune them on the limited-labeled sketch-to-photo retrieval data (i.e., target domain). However, without co-training source and target data, source domain knowledge might be forgotten during the fine-tuning process, while simply co-training them may cause negative transfer due to domain gaps. Moreover, identity label spaces of source data and target data are generally disjoint and therefore conventional category-level Domain Adaptation (DA) is not directly applicable. To address these issues, we propose an Instance-level Heterogeneous Domain Adaptation (IHDA) framework. We apply the fine-tuning strategy for identity label learning, aiming to transfer the instance-level knowledge in an inductive transfer manner. Meanwhile, labeled attributes from the source data are selected to form a shared label space for source and target domains. Guided by shared attributes, DA is utilized to bridge cross-dataset domain gaps and heterogeneous domain gaps, which transfers instance-level knowledge in a transductive transfer manner. Experiments show that our method has set a new state of the art on three sketch-to-photo image retrieval benchmarks without extra annotations, which opens the door to train more effective models on limited-labeled heterogeneous image retrieval tasks.**

***Index Terms*—Domain Adaptation, Cross-modal Image Retrieval, Sketch, Person Re-identification**

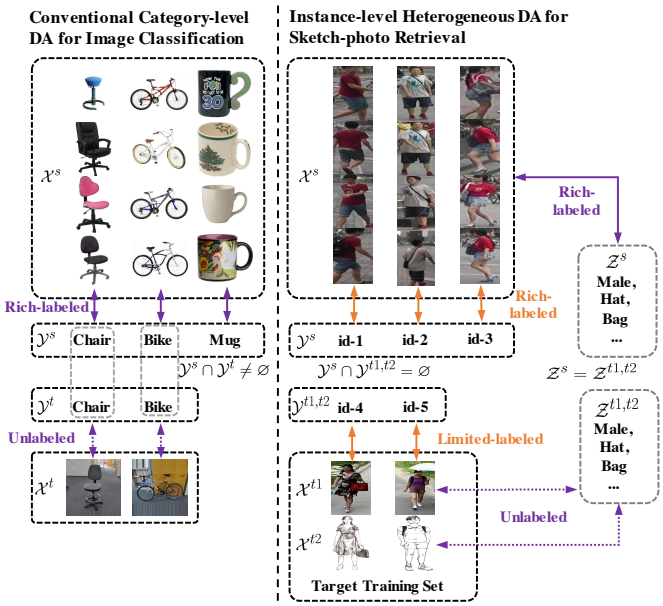


Fig. 1. Comparison of Category-level DA (left) and IHDA (right). The solid two-way arrow connector and the dotted two-way arrow connector represent labeled and unlabeled, resp. The orange two-way arrow connector and the purple two-way arrow connector represent inductive transfer flow and transductive transfer flow, resp.

I. Introduction

Domain Adaptation (DA) [1] is used to transfer the knowledge of the rich-labeled source domain ($\mathcal{D}^s$) to a unlabeled/limited-labeled target domain ($\mathcal{D}^t$) and improve the effectiveness of a given task in the target domain. Conventional DA studies mainly focus on the category-level cases, such as image classification and segmentation tasks [2]–[6]. In category-level DA, the target label space ($\mathcal{Y}^t$) and the source label space ($\mathcal{Y}^s$) are either fully shared or partially shared, which can be represented as $\mathcal{Y}^s \cap \mathcal{Y}^t \neq \varnothing$ (see Fig. 1 and the notations used in this section are summarized in TABLE I).

Manuscript received on February 4, 2020, revised on May XX, 2020, and accepted on July XX, 2020. *(Corresponding author: Yang Wu, Zheng Wang)*

Fan Yang, Sakriani Sakti, and Satoshi Nakamura are with Nara Institute of Science and Technology & RIKEN, Center for Advanced Intelligence Project, Nara, Japan (e-mail: yang.fan.xv6@is.naist.jp; ssakti@is.naist.jp; s-nakamura@is.naist.jp). Zheng Wang is with the Digital Content and Media Sciences Research Division, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan (e-mail: wangz@nii.ac.jp). Xiang Li is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: me@shawnli.me). Yang Wu is with the Computer Vision Lab., Kyoto University, Kyoto, Japan, and also with Nara Institute of Science and Technology, Nara 630-0192, Japan (e-mail: wu.yang.8c@kyoto-u.ac.jp).

In real-world applications, the needs of DA are not limited to category-level tasks. For instance, the instance-level sketch-to-photo retrieval, which has been investigated in many studies [7]–[14], is suffering lack of training data as it is costly to draw sketches. Meanwhile, plenty of rich-labeled photo retrieval datasets are available. In many existing photo retrieval datasets [15]–[19], even instance attributes are annotated [20], [21]. Accordingly, majority of sketch-to-photo retrieval methods [7], [10], [22]–[26] choose to pre-train their models on a rich-labeled photo retrieval data (i.e., source domain) and then fine-tune them on the limited-labeled sketch-to-photo retrieval data (i.e., target domain). However, without co-training source and target data, source domain knowledge might be forgotten during the fine-tuning process, while simply co-training them may cause negative transfer due to domain gaps.

DA methods are commonly used to reduce domain gaps on co-training source and target datasets. sketch-to-photo retrieval data consists of two heterogeneous modalities as sketch and photo. Accordingly, we divide the target data into two sub-domains as $\mathcal{D}^{t1}$ (i.e., photos) and $\mathcal{D}^{t2}$ (i.e., sketches), where $\mathcal{D}^t = \mathcal{D}^{t1} \cup \mathcal{D}^{t2}$. The feature spaces of $\mathcal{D}^s$, $\mathcal{D}^{t1}$, and $\mathcal{D}^{t2}$ are $\mathcal{X}^s$, $\mathcal{X}^{t1}$ and $\mathcal{X}^{t2}$, respectively. If we regard each identity as

one class, the identity label space of $\mathcal{X}^s$ is $\mathcal{Y}^s$ while $\mathcal{X}^{t1}$ and $\mathcal{X}^{t2}$ share the same identity label space $\mathcal{Y}^{t1,t2}$. Nonetheless, the source data is annotated by other contributors and identity label spaces of the source and target domains are generally disjoint (i.e., $\mathcal{Y}^s \cap \mathcal{Y}^{t1,t2} = \varnothing$). It is challenging for category-level DA methods to figure out "what to transfer and where to transfer" in this case. It shows that coarsely reducing domain gaps may incur a target error [27]. It is insufficient to simply match the marginal distributions, conditional distributions of both domains should also be matched.

To address this issue, we raise an Instance-level Heterogeneous Domain Adaptation (IHDA) framework, which breaks through limitations of simple fine-tuning and category-level DA. We select attributes that can jointly characterize instance-level properties for all domains to form a shared label space as $\mathcal{Z}^s = \mathcal{Z}^{t1,t2}$, where $\mathcal{Z}^s$ and $\mathcal{Z}^{t1,t2}$ are proposed attribute spaces for source domain and target domain, respectively. For instance-level retrieval problems, since all of the instances belong to the same category, the existence of shared attributes is guaranteed. We assume that the attributes of the source domain are observed. For some photo-sketch retrieval applications, a lot of related photo retrieval datasets already offer attributes [16], [28]. When attributes of the source dataset are not available, it is still possible to collect it from large web resources at an acceptable cost. As many objects are accompanied by attributes on website resources, both photos and attributes can be downloaded to construct the source data. For instance, the UT-Zap50K dataset [15], a source data in our experiment, was collected from a shopping website.

After $\mathcal{Z}^s$ and $\mathcal{Z}^{t1,t2}$ are formed, our IHDA framework seeks to maximally transfer domain knowledge from source data to target data. Due to $\mathcal{Y}^s \cap \mathcal{Y}^{t1,t2} = \varnothing$, $\mathcal{Y}^s$ and $\mathcal{Y}^{t1,t2}$ are distinct but related. Learning to identify $\mathcal{Y}^s$ is inductive to identify $\mathcal{Y}^{t1,t2}$. Thus, pre-training identify labels on $\mathcal{Y}^s$ and then fine-tuning identity labels on $\mathcal{Y}^{t1,t2}$ can transfer the instance-level knowledge from an inductive transfer manner. Since $\mathcal{Z}^s = \mathcal{Z}^{t1,t2}$, referring to them, cross-dataset domain gaps (i.e., $\mathcal{D}^s$ to $\mathcal{D}^{t1}$ and $\mathcal{D}^{t2}$) and heterogeneous domain gaps (i.e., $\mathcal{D}^s$ and $\mathcal{D}^{t1}$ to $\mathcal{D}^{t2}$) can be reduced via an unsupervised Domain Adaptation [2]–[5], which transfers instance-level knowledge from a transductive transfer manner.

Although the ground truth of $\mathcal{Z}^{t1,t2}$ is unknown, it is leveraged to associate source and target domains by using an entropy minimization criterion [29]. The entropy minimization criterion does penalize low-confident predictions of $\mathcal{Z}^{t1,t2}$. In order to minimize it, target domain features are forced to match source domain features at an instance level other than roughly reducing the cross-domain gaps. Such an approach remarkably improves the IHDA performance.

In summary, our work has **three contributions**:

- We highlight the challenges for improving limited-labeled sketch-to-photo retrieval performance from the transfer learning perspective.
- We propose a novel IHDA framework to maximally utilize the source domain knowledge to benefit the target task at instance-level image retrieval. It overcomes the limitations of applying simple fine-tuning and conventional DA.
- Our IHDA framework contributes new state-of-the-art results on instance-level sketch-to-photo retrieval task. It opens the door to train more effective cross-modal image retrieval models by using related rich-labeled single-modal image data.

TABLE I
SUMMARY OF NOTATIONS.

Symbol	Description
$\mathcal{D}^s$	Rich-labeled source domain
$\mathcal{D}^t$	Limited-labelled target domain, $\mathcal{D}^t = \mathcal{D}^{t1} \cup \mathcal{D}^{t2}$
$\mathcal{D}^{t1}$	Target photo domain, $\mathcal{D}^{t1} \in \mathcal{D}^t$
$\mathcal{D}^{t2}$	Target sketch domain, $\mathcal{D}^{t2} \in \mathcal{D}^t$
$\mathcal{X}^t$	Target data space, $\mathcal{X}^t = \mathcal{X}^{t1} \cup \mathcal{X}^{t2}$
$\mathcal{X}^{t1}$	Target photo data space, $\mathcal{X}^{t1} \in \mathcal{X}^t$
$\mathcal{X}^{t2}$	Target sketch data space, $\mathcal{X}^{t2} \in \mathcal{X}^t$
$\mathcal{Y}^s$	Source identity label space
$\mathcal{Y}^{t1,t2}$	Target photo-sketch identity label space, which is shared for $\mathcal{D}^{t1}$ and $\mathcal{D}^{t2}$
$\mathcal{Z}^s$	Source attribute label space, which is shared for $\mathcal{D}^s$, $\mathcal{D}^{t1}$, and $\mathcal{D}^{t2}$
$\mathcal{Z}^{t1,t2}$	Target photo-sketch attribute label space, which is shared for $\mathcal{D}^s$, $\mathcal{D}^{t1}$, and $\mathcal{D}^{t2}$

The rest of this paper is organized as follows. In Section II, a brief review of related works is given. In Section III, we illustrate the details of the proposed IHDA framework. Section IV reports experimental results to prove the effectiveness of our IHDA framework. In Section V, we discuss the application scope of our method. Finally, Section VI concludes this paper.

II. RELATED WORKS

We summarize related works and clarify their differences and relationships in TABLE II.

TABLE II
A COMPARISON OF RELATED WORKS.

Methods	#Modalities	DA & #Domains	Target Attributes	Identities $\mathcal{Y}^t \cap \mathcal{Y}^s = \varnothing$
[23] [24] [25] [26]	2	w/o DA	-	-
[10] [22] [8] [30]	2	w/o DA	Annotated	-
[2] [3] [4] [5]	1, 2	w/ DA, 2	-	No
[31] [32]	1	w/ DA, 2	Unknown	No
[33] [34] [35] [36]	1	w/ DA, 2	Unknown	Yes
Ours	**2**	**w/ DA, 3**	**Unknown**	**Yes**

Owing to the shortage of training data, studies [10], [22]–[26] take rich-labeled photo retrieval data for pre-training and then fine-tune their model on the sketch-to-photo retrieval data. However, there could be a dilemma in their approaches: performing a long-term fine-tuning may forget the source domain knowledge, which is known as catastrophic forgetting [37], [38], while a short-term fine-tuning may under-fit on the target data.

Alternatively, both the source and target data could be co-trained together. Due to the across-domain discrepancy, simply co-train both source and target data may lead to negative transfer [1] and impair the performance. To alleviate this problem, Domain Adaptation (DA) is desired. Unlike coarse category-level DA [2]–[5], or fine-grained category-level DA [31], [32], applying DA in instance-level heterogeneous retrieval faces the challenge that $\mathcal{Y}^s \cap \mathcal{Y}^{t1,t2} = \varnothing$. To alleviate this challenge, some instance-level domain adaptation methods [35], [36]

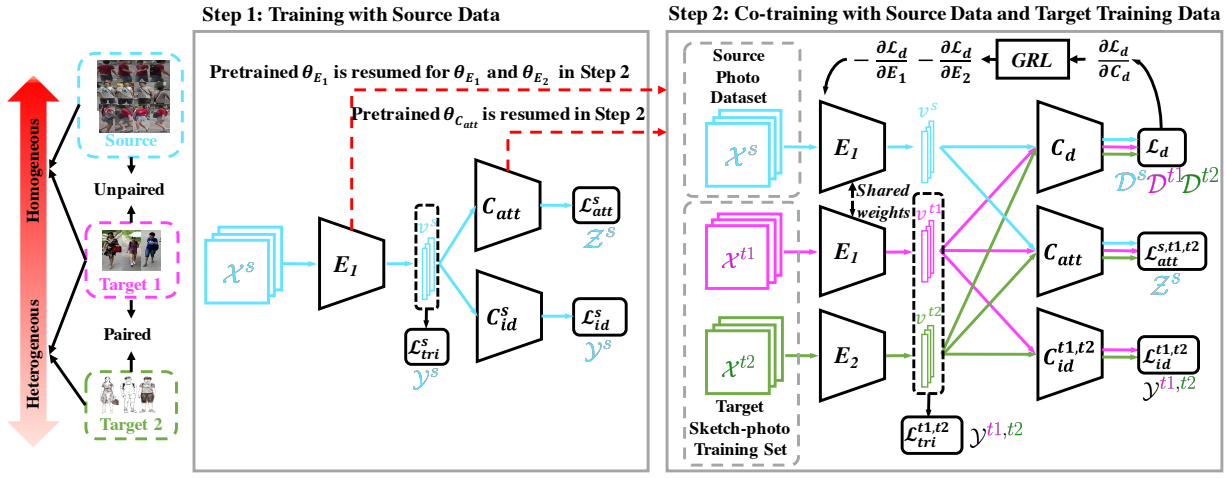


Fig. 2. The framework of Instance-level Heterogeneous Domain Adaptation (IHDA). E_1 and E_2 are encoders for photos (i.e., $\mathcal{X}^s$ and $\mathcal{X}^{t1}$) and sketches (i.e., $\mathcal{X}^{t2}$), resp, and v^s, v^{t1}, and v^{t2} denotes their embedding features; C^s_{id} and $C^{t1,t2}_{id}$ stand for the identity classifier of source and target domains, resp; C_{att} and C_d denote the shared-attribute classifier and domain classifier, resp; GRL is the Gradient Reversal Layer. Best viewed in color.

transferred the source domain knowledge to the target domain by using pseudo labels created by clustering algorithms on the target domain. However, clustering algorithms can easily generate noise in pseudo labels and impair the model performance. In our sketch-to-photo task, the noise label issue is more serious due to the heterogeneous domain gap. Therefore, instead of using pseudo labels, we specifically propose an IHDA framework, which selects annotated attributes in source data to form a shared label space for all domains.

Using attributes to learn the instance feature is particularly prominent in instance-level retrieval tasks. Before us, studies [8], [10], [22], [30] had annotated attributes on both source and target data to improve retrieval performance. Compared with their methods, our IHDA is more practical since no labeled attribute is required for the target data, which mitigates the burden of manual labeling. Moreover, IHDA even outperforms their methods in our experiments. Previous works [33], [34] also applied instance-level DA without annotated target attributes. However, their approaches focus on homogeneous DA scenario while our IHDA can tackle the more challenging instance-level heterogeneous DA problem. Besides, [34] only adapts the marginal distribution while [33] and ours jointly adapts the marginal distribution and conditional distribution, which can learn feature embedding that is robust for substantial distribution difference [39], [40].

We align conditional distribution by applying conditional entropy minimization [40], [41], which can take the task-specific boundaries into account. IHDA forces source and target domain distributions to be matched at an instance level with the guidance of attributes. We further demonstrate such an approach does help to improve the model performance in our ablation studies.

III. Methodology

We propose an Instance-level Heterogeneous Domain Adaptation (IHDA) framework as Fig. 2 shows. Since $\mathcal{X}^s$ and $\mathcal{X}^{t1}$ are heterogeneous to $\mathcal{X}^{t2}$, we construct two separated encoder neural networks E_1 and E_2 respectively for $(\mathcal{X}^s, \mathcal{X}^{t1})$ and $\mathcal{X}^{t2}$. ResNet-50 [42] is used as the backbone for both them and we denote their network parameters as θ_{E_1} and θ_{E_2}. The inputs are represented as x^s, x^{t1} and x^{t2}, where $x^s \in \mathcal{X}^s$, $x^{t1} \in \mathcal{X}^{t1}$ and $x^{t2} \in \mathcal{X}^{t2}$. IHDA takes two steps to transfer domain knowledge from the source domain to target domains. In the step 1, the encoder E_1, along with an identity classifier C^s_{id} and an attribute classifier C_{att}, are pre-trained on the source data. In the step 2, the weights of the pre-trained attribute classifier are resumed, and the weights of E_1 are reloaded for both E_1 and E_2. The source and target training data are used to co-train the whole network by coupling an adversarial domain adaptation. We implement the adversarial domain adaptation by the Gradient Reversal Layer [43]. *Related codes are available at https://github.com/fandulu/IHDA.*

A. Identity Learning

We build identity classifiers C^s_{id} and $C^{t1,t2}_{id}$ to perform identity classification. Each of them is a single dense layer, where the dimension of the output is equal to the number of identities. We denote the number of identities as N^s_{id} and $N^{t1,t2}_{id}$ for the source and target training data, respectively. The network parameters of C^s_{id} and $C^{t1,t2}_{id}$ are denoted as $\theta_{C^s_{id}}$ and $\theta_{C^{t1,t2}_{id}}$, respectively.

The identify classification losses $\mathcal{L}^s_{id}$ and $\mathcal{L}^{t1,t2}_{id}$ are defined as follows:

$$
\begin{aligned}
\hat{\psi}^s &= Softmax\Big(C^s_{id}\big(E_1(x^s)\big)\Big),\\
\hat{\psi}^{t1} &= Softmax\Big(C^{t1,t2}_{id}\big(E_1(x^{t1})\big)\Big),\\
\hat{\psi}^{t2} &= Softmax\Big(C^{t1,t2}_{id}\big(E_2(x^{t2})\big)\Big),\\
\mathcal{L}^s_{id} &= -\sum_{i=1}^{N^s_{id}} y^s_i \log(\hat{\psi}^s_i),\\
\mathcal{L}^{t1,t2}_{id} &= -\sum_{i=1}^{N^{t1,t2}_{id}} y^{t1}_i \log(\hat{\psi}^{t1}_i) - \sum_{i=1}^{N^{t1,t2}_{id}} y^{t2}_i \log(\hat{\psi}^{t2}_i),
\end{aligned}
\tag{1}
$$

where y^s, y^{t1} and y^{t2} are one-hot ground truth labels of x^s, x^{t1} and x^{t2}, respectively. The subscript i indicates the i_{th} element. $\hat{\psi}^s$, $\hat{\psi}^{t1}$ and $\hat{\psi}^{t2}$ are the corresponding predicted identity probabilities. Here, $y^s, \hat{\psi}^s \in \mathcal{Y}^s$ and $y^{t1}, y^{t2}, \hat{\psi}^{t1}, \hat{\psi}^{t2} \in \mathcal{Y}^{t1,t2}$.

In the step 1, $\mathcal{L}_{id}^{s}$ is used for the source data. In the step 2, since $\mathcal{Y}^s \cap \mathcal{Y}^{t1,t2} = \varnothing$, $\mathcal{L}_{id}^{t1,t2}$ is solely used for the target data. Apart from the identity classification, we also use a triplet loss to simultaneously learn joint embedding features that can better represent similarity and difference of identities. The joint embedding feature is a global average pooling feature with 2048 dimensions. The triplet loss $\mathcal{L}_{tri}^{s}$, used in step 1, is the conventional triplet loss [44] and we will not give its details here.

We extend the conventional triplet loss to be a *heterogeneous triplet loss* in step 2. For simplicity, let $v^{t1} = E_1(x^{t1})$ and $v^{t2} = E_2(x^{t2})$ stand for embedding features of photos and sketches, respectively. We represent the learned embedding feature space as $\mathcal{V}$, where $v^{t1}, v^{t2} \in \mathcal{V}$. For feature-label pairs in a mini-batch of target training data, we sample a sketch anchor (v_a^{t2}, y_a^{t2}), a photo positive sample (v_p^{t1}, y_p^{t1}) and a photo negative sample (v_n^{t1}, y_n^{t1}) with

$$y_p^{t1} = y_a^{t2} \text{ and } y_n^{t1} \neq y_a^{t2}, \tag{2}$$

where $y_a^{t2}, y_p^{t1}, y_n^{t1} \in \mathcal{Y}^{t1,t2}$.

The heterogeneous triplet loss function is as follows,

$$\mathcal{L}_{tri}^{t1,t2} = \left[\left\| v_a^{t2} - v_p^{t1} \right\|_2^2 - \left\| v_a^{t2} - v_n^{t1} \right\|_2^2 + \alpha \right]_+, \tag{3}$$

where α indicates the margin between the positive and negative sketch-to-photo pairs. We empirically set $\alpha = 0.3$.

B. Attribute Learning

Since selected attributes are shared among three domains, an attribute classifier C_{att} is used for these domains in both step 1 and step 2. C_{att} consists of three dense layers. For the first two layers, each has 512 dimensions. The dimension of the output layer is N_{att}, which is the number of attributes. The network parameters of C_{att} is represented as $\theta_{C_{att}}$.

Generally, each identity can be assigned to multiple attributes at the same time, we treat it as a multi-label classification task. Therefore, we apply a Binary Cross-Entropy loss for each kind of attribute and then sum them up as the multi-label classification loss. On the source domain, the attribute classification loss is

$$\begin{aligned} \hat{\phi}^s &= Sigmoid\Big(C_{att}\big(E_1(x^s)\big)\Big), \\ \mathcal{L}_{att}^{s} &= \sum_{i=1}^{N_{att}} -z_i^s \log(\hat{\phi}_i^s) - (1 - z_i^s) \log(1 - \hat{\phi}_i^s), \end{aligned} \tag{4}$$

where z^s is one-hot ground truth label of x^s and the subscript i indicates the i_{th} element; $\hat{\phi}^s$ is the estimated attributes. Here, $z^s, \hat{\phi}^s \in \mathcal{Z}^s$.

Without increasing annotations, attributes are unknown for the target data and therefore the loss function above cannot be applied directly. Inspired by previous works [5], [29], [40], [45], a multi-label entropy minimization criterion is proposed to associate the attribute learning between the source domain and target domains. It does penalize low-confident predictions of target attributes. In order to minimize it, target domains are forced to match with the source domain at an instance level other than roughly reducing domain discrepancies, which leads to a better DA performance. On the target domains, the attribute classification loss is

$$\begin{aligned} \hat{\phi}^{t1} &= Sigmoid\Big(C_{att}\big(E_1(x^{t1})\big)\Big), \\ \hat{\phi}^{t2} &= Sigmoid\Big(C_{att}\big(E_2(x^{t2})\big)\Big), \\ \mathcal{L}_{att}^{t1,t2} &= -\sum_{i=1}^{N_{att}} \hat{\phi}_i^{t1} \log(\hat{\phi}_i^{t1}) - \sum_{i=1}^{N_{att}} \hat{\phi}_i^{t2} \log(\hat{\phi}_i^{t2}), \end{aligned} \tag{5}$$

where $\hat{\phi}^{t1}$ and $\hat{\phi}^{t2}$ are estimated attributes of x^{t1} and x^{t2}, respectively. Here, $\hat{\phi}^{t1}, \hat{\phi}^{t2} \in \mathcal{Z}^{t1,t2}$.

Moreover, when the photo and sketch are paired in target data, their predicted attributes should be identical. Using this property, we further construct an attribute-consistent loss for paired target data as

$$\mathcal{L}_{con}^{t1,t2} = \mathbb{1}_{y^{t1}=y^{t2}} \left\| \hat{\phi}^{t1} - \hat{\phi}^{t2} \right\|_2, \tag{6}$$

where $\mathbb{1}_{y^{t1}=y^{t2}}$ is 1 when $y^{t1} = y^{t2}$ (i.e., sketch and photo are paired), and 0 elsewhere.

Overall, in step 1, we only use $\mathcal{L}_{att}^{s}$ for attribute classification; in step 2, the entire attribute classification loss is

$$\mathcal{L}_{att}^{s,t1,t2} = \mathcal{L}_{att}^{s} + \mathcal{L}_{att}^{t1,t2} + \mathcal{L}_{con}^{t1,t2}. \tag{7}$$

C. Adversarial Domain Adaptation Learning

Coupling adversarial training with deep learning introduces a powerful tool to harness domain adaptation. It has been successfully applied to plenty of tasks [4], [5], [43], [46], [47].

In IHDA, $\mathcal{X}^s$ and $\mathcal{X}^{t1}$ are heterogeneous to $\mathcal{X}^{t2}$. Although $\mathcal{X}^s$ and $\mathcal{X}^{t1}$ are closed to homogeneous, there are differences in their illuminations, texture styles and image resolutions. As a result, we accommodate adversarial domain adaptation from the traditional two domains to three domains in this task.

We build a domain classifier C_d to distinguish domains s, $t1$ and $t2$. C_d consists of three dense layers. For the first two layers, each has 512 dimensions. The dimension of the output layer is N_d, which is the number of domains. The network parameters of C_d is represented by θ_{C_d}. Considering the adversarial domain adaptation is applied, a **Reverse** Categorical Cross-Entropy Loss is applied to form the loss function:

$$\begin{aligned} \hat{\rho}^s &= Softmax\Big(C_d\big(E_1(x^s)\big)\Big), \\ \hat{\rho}^{t1} &= Softmax\Big(C_d\big(E_1(x^{t1})\big)\Big), \\ \hat{\rho}^{t2} &= Softmax\Big(C_d\big(E_2(x^{t2})\big)\Big), \\ \mathcal{L}_d &= \sum_{i=1}^{N_d} d_i^s \log(\hat{\rho}_i^s) + \sum_{i=1}^{N_d} d_i^{t1} \log(\hat{\rho}_i^{t1}) + \sum_{i=1}^{N_d} d_i^{t2} \log(\hat{\rho}_i^{t2}), \end{aligned} \tag{8}$$

where N_d stands for the number of domains used for heterogeneous domain adaptation, and $N_d = 3$ in this work. d^s, d^{t1} and d^{t2} are one-hot domain labels of x^s, x^{t1} and x^{t2}, respectively; the subscript i indicates the i_{th} element; $\hat{\rho}^s$, $\hat{\rho}^{t1}$ and $\hat{\rho}^{t2}$ are estimated domains of x^s, x^{t1} and x^{t2}, respectively.

The adversarial domain adaptation procedure is a two-player game. One player C_d intends to maximize $\mathcal{L}_d$ so that d^s, d^{t1} and d^{t2} could be distinguished. Another player is the combination of encoders E_1 and E_2, which attempts to minimize $\mathcal{L}_d$ and confuse C_d by reducing the cross-domain gaps.

The Gradient Reversal Layer (GRL) [43] is applied between the encoder and the domain classifier, which simplifies the two-player adversarial optimization by one step in an iteration.

D. Optimization Goals of IHDA Framework

By integrating aforementioned Equations, we have the loss functions of step 1 and step 2 as

$$\begin{aligned}\mathcal{L}_{step1} &= \mathcal{L}^s_{id} + \lambda_1 \mathcal{L}^s_{tri} + \lambda_2 \mathcal{L}^s_{att},\\ \mathcal{L}_{step2} &= \mathcal{L}^{t1,t2}_{id} + \lambda_1 \mathcal{L}^{t1,t2}_{tri} + \lambda_2 \mathcal{L}^{s,t1,t2}_{att} + \lambda_3 \mathcal{L}_d,\end{aligned} \tag{9}$$

where λ_1, λ_2 and λ_3 are trade-off parameters.

In step 1, the optimization goal is to find the network parameters $\hat{\theta}_{E_1}$, $\hat{\theta}_{C^s_{id}}$ and $\hat{\theta}_{C_{att}}$ that satisfy

$$\left(\hat{\theta}_{E_1}, \hat{\theta}_{C^s_{id}}, \hat{\theta}_{C_{att}}\right) = \underset{\theta_{E_1}, \theta_{C^s_{id}}, \theta_{C_{att}}}{\operatorname{argmin}} \mathcal{L}_{step1}. \tag{10}$$

In step 2, pretrained $\hat{\theta}_{E_1}$ and $\hat{\theta}_{C_{att}}$ are resumed. Then E_1, E_2, $C^{t1,t2}_{id}$ and C_{att} are simultaneously optimized by minimizing $\mathcal{L}_{step2}$. While C_d is optimized by maximizing $\mathcal{L}_d$.

The optimization goals are to find the network parameters $\hat{\theta}_{E_1}$, $\hat{\theta}_{E_2}$, $\hat{\theta}_{C^{t1,t2}_{id}}$, $\hat{\theta}_{C_{att}}$ and $\hat{\theta}_{C_d}$ that satisfy

$$\begin{aligned}\left(\hat{\theta}_{E_1}, \hat{\theta}_{E_2}, \hat{\theta}_{C^{t1,t2}_{id}}, \hat{\theta}_{C_{att}}\right) &= \underset{\theta_{E_1}, \theta_{E_2}, \theta_{C^{t1,t2}_{id}}, \theta_{C_{att}}}{\operatorname{argmin}} \mathcal{L}_{step2},\\ \left(\hat{\theta}_{C_d}\right) &= \underset{\theta_{C_d}}{\operatorname{argmax}}\ \mathcal{L}_d.\end{aligned} \tag{11}$$

These two optimization goals are jointly updated in each training step.

E. Insight of Optimization Goals

In IHDA framework, we have $\mathcal{X} \mapsto \mathcal{V}$ (by E_1 and E_2), $\mathcal{V} \mapsto \mathcal{Y}$ (by $C^s_{id}, C^{t1,t2}_{id}$), and $\mathcal{V} \mapsto \mathcal{Z}$ (by C_{att}). We denote $P(v^s)$ and $P(v^t)$ as the marginal distributions of embedding features, $P(y^s|v^s)$ and $P(y^t|v^t)$ as the conditional distributions based on identity labels, and $P(z^s|v^s)$ and $P(z^t|v^t)$ as the conditional distributions based on attribute labels, in the source and target domain, respectively.

As theoretically and experimentally analyzed in [27], [39], [48]–[51], it is insufficient to obtain good embedding representations $\mathcal{V}$ by simply reducing the marginal domain distribution discrepancy. Conditional domain distribution discrepancy should also be considered.

Therefore, other than coarsely minimizing the marginal domain distribution discrepancy, which is represented by

$$\textbf{min}\{\left\|P(v^s) - P(v^t)\right\|\}, \tag{12}$$

we aim to jointly reduce marginal domain discrepancy and conditional domain distribution discrepancy:

$$\begin{cases} \textbf{min}\{\|P(v^s) - P(v^t)\|\}, \\ \textbf{min}\{\|P(y^s|v^s) - P(y^t|v^t)\|\}. \end{cases} \tag{13}$$

Ideally, we should minimize $\|P(y^s|v^s) - P(y^t|v^t)\|$, however, as we have analysed, in instance-level retrieval tasks, we have $\mathcal{Y}^s \cap \mathcal{Y}^t = \varnothing$, it is challenging to minimize $\|P(y^s|v^s) - P(y^t|v^t)\|$.

Since attribute learning has a main training objective on attributes but also learn representations that can identify individuals as a side effect, we substitute $\|P(y^s|v^s) - P(y^t|v^t)\|$ with $\|P(z^s|v^s) - P(z^t|v^t)\|$, where z is the attribute and we have $\mathcal{Z}^s = \mathcal{Z}^t$. Then we obtain the following optimization goal:

$$\begin{cases} \textbf{min}\{\|P(v^s) - P(v^t)\|\} \Leftrightarrow \mathcal{L}_d, \\ \textbf{min}\{\|P(z^s|v^s) - P(z^t|v^t)\|\} \Leftrightarrow \mathcal{L}^{s,t1,t2}_{att}. \end{cases} \tag{14}$$

They are identical to the last two items of $\mathcal{L}_{step2}$ (see Equation 9). We minimize the conditional distribution discrepancy $\mathcal{L}^{s,t1,t2}_{att}$ by a semi-supervised learning approach [40], [41]. Be the same as balance factors in Balanced Distribution Adaptation [50], λ_2 and λ_3 are used to adjust the importance reducing the marginal distribution discrepancy and conditional distribution discrepancy.

Consequently, we approach jointly optimize the marginal distribution and conditional distribution in our IHDA. We also experimentally show the effectiveness of jointly optimize the marginal distribution and conditional distribution in DA in ablation studies.

IV. Experiments

A. Datasets

In our experiments, source datasets include **UT-Zap50K** [15], **CelebFaces** [16], and **Market-1501** [17], their corresponding target datasets are **QMUL-Shoes** [10], **IIIT-D Viewed Sketch** [52] and **PKU-Sketch** [7], respectively. Since sketches do not contain color information, we select non-color attributes from the annotated source to form the shared label space $\mathcal{Z}^s = \mathcal{Z}^{t1,t2}$.

1) The UT-Zap50K dataset is a large-scale shoe photo dataset consisting of 50K images and around 20K identities. We select all non-color attributes to form the shared label space, such as "Wide Toe", "Snip Toe", "Toggle Closure", and so on. The QMUL-Shoes dataset contains 419 sketch-to-photo pairs of shoes, in which 304 pairs are for training and 115 pairs are for testing. Since sketches are collected from non-professional drawers, it encounters more challenges to recognize sketch-to-photo pairs than others. Although the target QMUL-Shoes also includes shoes attributes, they are different from the ones defined in UT-Zap50K dataset. To perform supervised learning on target attributes (e.g., [30]), we need to add new attributes to the target, which is laborious. Our method takes advantage of unsupervised DA and does not use labeled attributes in the target dataset.

2) The CelebFaces dataset is a large-scale face photo dataset with around 10K identities, 202K face images, and 40 binary

attributes. We select all non-color attributes to form the shared label space, such as "Wearing-Hat", "Bald", "Eyeglasses", and so on. The IIIT-D Viewed Sketch dataset includes 238 sketch-to-photo pairs of faces, in which the sketches are drawn by professional artists. Unlike our source dataset CelebFaces, attribute annotation is not available in IIIT-D Viewed Sketch dataset. We take the same evaluation protocols as [24], [26] proposed.

3) The Market-1501 dataset is a large-scale person photo dataset with around 1.5K identities, 32K person images. In work [28], [53], 27 binary attributes are annotated for Market-1501 dataset. We select all non-color attributes to form the shared label space, as "gender", "hair", "length of sleeve", "length of lower-body clothing", "style of clothing", "hat", "age", and "bag" ("hand bag" and "pack bad" are regarded as "bag"). The PKU-Sketch dataset consists of samples of 200 persons, in which each person has one sketch and two photos taken from disjoint cameras. Besides, there is no attribute annotation for it. We take the evaluation protocols [7], and 150 identities are randomly chosen for training while the rest 50 are used for testing. We report average values of 10 times experiments as the final results.

Note that, we utilize the whole source dataset and a small portion of the annotated target training set for the training, while the target testing set is untouched in the training process.

B. Experimental Settings

Following the procedure in Fig. 2, we start with pre-training the E_1 branch on the source data with 60 epochs in step 1. The batch size is 64. Then, in step 2, by reloading the pre-trained θ_{E_1} and $\theta_{C_{att}}$ weights, we co-train the whole IHDA network on both source and target training data with another 60 epochs. The batch size is 96 (32 sketch-to-photo pairs from the target dataset and 32 samples from the source dataset). We choose Adam optimizer [54] for the training and the initial learning rate is set to be 1×10^{-4}. To properly utilize pre-trained network parameters, we apply a warming-up learning rate schedule, which is modified from [55], to adjust the learning rate. We set $\lambda_1 = 1$, $\lambda_2 = 0.1$ and $\lambda_3 = 0.1$ to balance the whole network. Note that the input image size for shoe, face, and person samples are 96×96, 144×128, and 384×128, respectively. In the inference stage, a L_2 Normalization is applied to each embedding feature before calculating their relative distances.

C. Comparison with Methods of Sketch-to-photo Retrieval

We compare our method with state-of-the-art methods of sketch-to-photo retrieval on three types of target datasets, respectively for the sketch-to-photo retrieval of shoes, faces, and persons.

For the QMUL-Shoes dataset, we compare our method with Triplet SN [10], MAR-FBIR [22], CD-AFL [7], and DSSA Triplet [23]. They are pre-trained on external datasets TU Berlin Sketch [56] and Edge-style ImageNet [57]. Our model takes the UT-Zap50K dataset as the source dataset since shoe attributes are annotated in it. The rank-1 and rank-10 re-identification accuracy of each method are reported in TABLE III.

TABLE III
THE PERFORMANCE OF SKETCH-TO-PHOTO RETRIEVAL ON THE QMUL-SHOES DATASET.

Model	w/ external Data	R1	R10
Triplet SN [10] (pre-train w/o att + fine-tune)	TU Berlin Sketch &Edge-style ImageNet	39.1	87.8
MAR-FBIR [22] (pre-train w/o att + fine-tune)	TU Berlin Sketch &Edge-style ImageNet	50.4	91.3
CD-AFL [7] (pre-train w/o att + fine-tune)	TU Berlin Sketch &Edge-style ImageNet	56.4	92.6
DSSA Triplet [23] (pre-train w/o att + fine-tune)	TU Berlin Sketch &Edge-style ImageNet	61.7	94.8
Our IHDA (Instance-level heterogeneous DA w/ att)	UT-Zap50K	**68.7**	**95.7**

For the IIIT-D Viewed Sketch dataset, we compare our method with Deep Face [58], Light CNN [24], CDL [25], and RCN [26]. The CelebFaces dataset is used in their pre-training. Our model takes the CelebFaces dataset as the source dataset and uses the IHDA framework. The rank-1 re-identification accuracy of each method is listed in TABLE IV.

TABLE IV
THE PERFORMANCE OF SKETCH-TO-PHOTO RETRIEVAL ON THE IIIT-D VIEWED SKETCH DATASET.

Model	w/ external Data	R1
Deep Face [58] (pre-train w/o att + fine-tune)	CelebFaces	80.9
Light CNN [24] (pre-train w/o att + fine-tune)	CelebFaces	84.0
CDL [25] (pre-train w/o att + fine-tune)	CelebFaces	85.4
RCN [26] (pre-train w/o att + fine-tune)	CelebFaces	90.3
Our IHDA (Instance-level heterogeneous DA w/ att)	CelebFaces	**95.7**

For the PKU-Sketch dataset, we compare our method with Triplet SN [10], GN Siamese [59], and CD-AFL [7]. Among them, GN Siamese and CD-AFL are pre-trained on the Market-1501 dataset. Note that CD-AFL only applied adversarial supervised DA on PKU-Sketch dataset. When considering Market-1501 dataset together, it simply applied pre-training and fine-tuning. Our model takes the Market-1501 dataset as the source dataset and uses the IHDA framework. The rank-1, rank-5, rank-10 and rank-20 re-identification accuracy of each method are reported in TABLE V.

TABLE V
THE PERFORMANCE OF SKETCH-TO-PHOTO RETRIEVAL ON THE PKU-SKETCH DATASET.

Model	w/ external Data	R1	R5	R10	R20
Triplet SN [10] (pre-train w/o att + fine-tune)	×	9.0	26.8	42.2	65.2
GN Siamese [59] (Ppe-train w/o att + fine-tune)	Market-1501	28.9	54.0	62.4	78.2
CD-AFL [7] (pre-train w/o att + fine-tune)	Market-1501	34.0	56.3	72.5	84.7
Our IHDA (Instance-level heterogeneous DA w/ att)	Market-1501	**85.6**	**94.8**	**98.0**	**100.0**

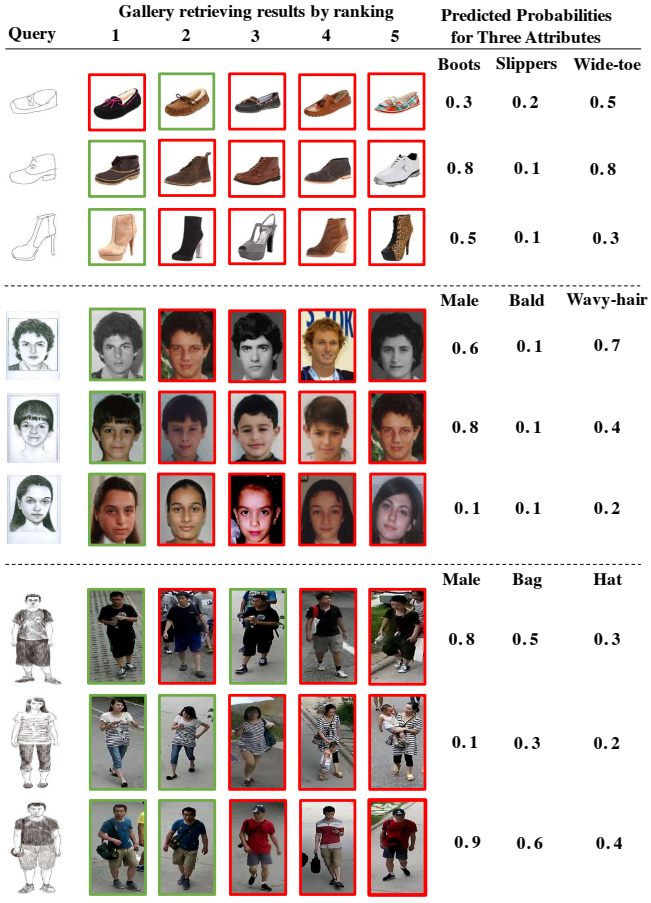


Fig. 3. Visualization of retrieving results on QMUL-Shoes, IIIT-D, and PKU-Sketch datasets. In the same row, gallery images with green bounding boxes are identical to the query image, while others with the red bounding box are different identities. Predicted probabilities are shown for three attributes here.

On QMUL-Shoes dataset, using IHDA framework outperforms the previous state-of-the-art method by 7.0% and reaches 68.7% in rank-1 retrieving accuracy. A 5.4% gain of rank-1 retrieving accuracy is made on IIIT-D Viewed Sketch dataset, with a value as high as 95.7%. On PKU-Sketch dataset, IHDA surpasses the previous state-of-the-art method up to 51.6% and obtains 85.6% in rank-1 retrieving accuracy. This is not happened by coincidence and we will go deep into the essential mechanism of IHDA in ablations studies.

The improvement in PKU-Sketch dataset is more significant than the other two datasets. We consider there are two main reasons: (1) Sketches of QMUL-Shoes dataset are drawn by amateurs while sketches of IIIT-D and PKU-Sketch datasets are drawn by artists with high qualities and more details (see Fig. 3), we can achieve high performance on IIIT-D and PKU-Sketch datasets. (2) The body poses within a body sketch-to-photo pair could be remarkably different, which causes larger distortion than face sketch-to-photo pairs. Therefore, the fine-tuning strategy can reach a good performance on the IIIT-D dataset, but on the more challenging PKU-Sketch dataset, our IHDA shows its superiority.

In Fig. 3, we illustrate several retrieving results on target testing data. It can be observed that each retrieved photo has a similar appearance and structure as the query sketch. This proves that the joint heterogeneous embedding is well learned in our IHDA framework. For some abstract sketches, it is even difficult to identify which is the correct corresponding photo by eyes. This explains why our IHDA framework is hindered to further improve sketch-to-photo retrieving performance.

D. Comparison of Using Different Source Datasets and Training Strategies

Due to the lack of annotations, most of sketch-photo retrieval works [7], [22]–[26], [58], [59] utilize external datasets to pre-train and then fine-tune their model on the target dataset. However, there two questions worth to be further explored: 1) though unrelated sketch-photo source data (e.g., Sketchy [60]) cannot fit our complete IHDA framework, could an incomplete IHDA framework obtain satisfactory results by using it as the source dataset? 2) since attribute learning is applied in our IHDA framework, could it also improve the pre-training and fine-tuning framework? To answer these questions, we construct three baselines by using part of our IHDA framework components and perform the same training strategy on three experimental datasets.

Specifically, in Baseline A, we conduct experiments by using Sketchy dataset [60] as the source dataset, and QMUL-Shoes, IIIT-D Viewed Sketch, and PKU-Sketch as the target datasets. Referring to the training process in Fig. 2, we first pre-train two encoders (for photo and sketch separately) using Sketchy in Step 1, and then fine-tune both encoders using target datasets in Step 2. In Baseline B and C, whose components and source dataset are the same as IHDA framework, excepting that attribute learning is removed or only used in Step 1. The results are shown in TABLE VI, TABLE VII, and TABLE VIII.

TABLE VI
THE SOURCE DATASET SETTING AND SKETCH-TO-PHOTO RETRIEVAL PERFORMANCE ON THE QMUL-SHOES DATASET.

Model	w/ external Data	R1	R10
Our Baseline A (pre-train w/o att + fine-tune)	Sketchy	59.2	93.3
Our Baseline B (pre-train w/o att + fine-tune)	UT-Zap50K	57.5	92.8
Our Baseline C (pre-train w/ att + fine-tune)	UT-Zap50K	59.0	93.1
Our IHDA (Instance-level heterogeneous DA w/ att)	UT-Zap50K	**68.7**	**95.7**

The results of Baseline A consistently demonstrate when the naive pre-training and fine-tuning strategy is applied, using Sketchy dataset for pre-training leads to better performance than using a homogeneous dataset. Since Sketchy dataset contains sketch and photo pairs, using it in pre-training might improve low-level feature representation in both encoders. However, samples in Sketchy dataset and our target datasets may not belong to the same category, not matter the same identity. Such a content divergence hinders the model's capability on transferring the high-level semantic knowledge to target datasets. To some extent, an ideal source dataset might satisfy: 1) including sketch-photo pairs (Sketchy dataset); 2) including

TABLE VII
THE SOURCE DATASET SETTING AND SKETCH-TO-PHOTO RETRIEVAL PERFORMANCE ON THE IIIT-D VIEWED SKETCH DATASET.

Model	w/ external Data	R1
Our Baseline A (pre-train w/o att + fine-tune)	Sketchy	90.5
Our Baseline B (pre-train w/o att + fine-tune)	CelebFaces	88.3
Our Baseline C (pre-train w/ att + fine-tune)	CelebFaces	88.6
Our IHDA (Instance-level heterogeneous DA w/ att)	CelebFaces	**95.7**

TABLE VIII
THE SOURCE DATASET SETTING AND SKETCH-TO-PHOTO RETRIEVAL PERFORMANCE ON THE PKU-SKETCH DATASET.

Model	w/ external Data	R1	R5	R10	R20
Our Baseline A (pre-train w/o att + fine-tune)	Sketchy	61.0	82.6	93.6	95.2
Our Baseline B (pre-train w/o att + fine-tune)	Market-1501	58.8	78.0	90.0	94.2
Our Baseline C (pre-train w/ att + fine-tune)	Market-1501	59.4	79.2	91.2	95.8
Our IHDA (Instance-level heterogeneous DA w/ att)	Market-1501	**85.6**	**94.8**	**98.0**	**100.0**

different identities but the same category as the target dataset (our selected source datasets). However, it is difficult to find such an ideal source dataset for all target datasets, we have to make trade-offs. Luckily, our proposed IHDA framework significantly improves the performance of target datasets by assuming the availability of attributes but dropping the need for category-matched sketch-photo data.

The performances of Baseline B and C are similar, which indicates that *solely including attribute learning in existing approaches [7], [23], [26] may not remarkably improve the retrieval performance on the target data*. More than using attributes in pre-training, our IHDA uses attributes to form a shared label space and perform semi-supervised attribute learning to intermediately reduce the domain gaps in terms of marginal distributions and conditional distributions. As a result, our IHDA can significantly improve the performance in three target datasets.

E. Comparison with Methods of Photo-to-sketch Retrieval

Although the sketch-to-photo retrieval (query: sketch; gallery: photo) is a dominant application scenario, we perform an inverse experiment to investigate the performance of photo-to-sketch retrieval (query: photo; gallery: sketch) on three benchmarks used in our paper. The results are shown in Tables IX, X, and XI (Note, some unpublished codes are re-implemented by us, bias could exist).

The photo-to-sketch retrieval performance is similar to sketch-to-photo retrieval performance, which indicates that distinguishable embedding is obtained for sketches and photos in a symmetrical way.

TABLE IX
THE PERFORMANCE OF PHOTO-TO-SKETCH RETRIEVAL ON THE QMUL-SHOES DATASET.

Model	w/ external Data	R1	R10	R20
DSSA Triplet [23] (pre-train w/o att + fine-tune)	TU Berlin Sketch &Edge-style ImageNet	61.9	95.1	98.6
Our IHDA (Instance-level heterogeneous DA w/ att)	UT-Zap50K	**69.6**	**97.4**	**99.1**

TABLE X
THE PERFORMANCE OF PHOTO-TO-SKETCH RETRIEVAL ON THE IIIT-D VIEWED SKETCH DATASET.

Model	w/ external Data	R1	R10	R20
RCN [26] (pre-train w/o att + fine-tune)	CelebFaces	90.8	95.2	97.8
Our IHDA (Instance-level heterogeneous DA w/ att)	CelebFaces	**96.2**	**98.6**	**99.2**

TABLE XI
THE PERFORMANCE OF PHOTO-TO-SKETCH RETRIEVAL ON THE PKU-SKETCH DATASET.

Model	w/ external Data	R1	R10	R20
CD-AFL [7] (pre-train w/o att + fine-tune)	Market-1501	37.6	76.2	92.8
Our IHDA (Instance-level heterogeneous DA w/ att)	Market-1501	**88.2**	**100.0**	**100.0**

F. Comparison with Methods of Instance-level Domain Adaptation on Sketch-to-photo Retrieval Tasks

Existing instance-level domain adaption methods mainly perform their experiments on cross-dataset photo (i.e., homogeneous) retrieval tasks. In this part, we compare our IHDA framework with two open-source works [35], [36] on PKU-Sketch dataset. The results are shown in Tables XII.

TABLE XII
THE PERFORMANCE OF SKETCH-TO-PHOTO RETRIEVAL ON THE PKU-SKETCH DATASET (COMPARED WITH DA METHODS).

Model	w/ external Data	R1	R5	R10	R20
MMT [35] (Instance-level homogeneous DA w/o att)	Market-1501	42.8	62.4	71.2	88.2
UDA-Reid [36] (Instance-level homogeneous DA w/o att)	Market-1501	40.4	60.0	72.8	90.4
Our incomplete IHDA (Instance-level heterogeneous DA w/o att)	Market-1501	70.4	82.0	92.6	96.2
Our complete IHDA (Instance-level heterogeneous DA w/ att)	Market-1501	**85.6**	**94.8**	**98.0**	**100.0**

Though MMT [35] and UDA-Reid [36] have achieved great successes on cross-dataset homogeneous retrieval tasks (*e.g.*, synthetic/real photo), heterogeneous (*e.g.*, sketch-to-photo) domain gaps may be out of the capabilities of their models. Even without attribute learning, our incomplete IHDA still can significantly outperform MMT and UDA-Reid on the PKU-Sketch dataset. In particular, their domain adaptation is specifically designed to improve feature representation in cross-dataset photo (homogeneous) retrieval tasks, and thus a single encoder is applied. However, we consider that applying

domain adaption on sketch-to-photo (heterogeneous) retrieval task is different. To customize instance-level domain adaptation to the heterogeneous case, our IHDA framework applies two encoders for sketch and photo separately. Besides, our domain adaptation consists of three domains, as $\mathcal{D}^s$, $\mathcal{D}^{t1}$ and $\mathcal{D}^{t2}$, and corresponding objective functions are designed for jointly reducing domain gaps within them. Moreover, unlike MMT and UDA-Reid that need to include the target testing set in their training, our IHDA framework keeps the target testing set untouched during the training process, which demonstrates its better generalization.

G. Attribute Selection

Since the source datasets are not originally designed for the target datasets, some attributes of source datasets may not suitable for the target dataset. We would like to highlight that task-irrelevant attributes, which is against our assumption $\mathcal{Z}^s = \mathcal{Z}^{t1,t2}$, should be excluded. in terms of sketch-to-photo retrieval task, color-related attributes are task-irrelevant. As all attributes are defined at the semantic level, excluding color attributes of the source dataset is **deterministic** rather than heuristic.

What will happen if we select color attributes to apply IHDA framework on the sketch-to-photo retrieval task? Taking Market-1501 and PKU-Sketch datasets as an example, we do experiments to prove that using color attributes can cause a negative transfer. As TABLE XIII shows, the sketch-photo retrieval performance significantly decreases after adding color attributes. The color attributes may work as a noise, make the model has no incentive to learn the real relevant semantic information between source and target domains. Compared with manually annotating sketch-photo datasets, it worth spending a few seconds to exclude color attributes at the beginning.

TABLE XIII
COMPARISON OF USING COLOR ATTRIBUTES ON PKU-SKETCH DATASET.

Model	R1	R5	R10	R20
IHDA (w/o color attributes)	85.6	94.8	98.0	100.0
IHDA (w/ color attributes)	54.4	76.2	88.6	90.2

Aside from color attributes, there are several non-color attributes could be selected. Since we may not have the ground-truth attributes on the target dataset, it is challenging to apply traditional feature selection methods [61] to select the most useful ones in our task. As an alternative approach, we assume that if the semantic information of selected attributes can be well learned in the target domain, a new attribute classifier, which is trained from scratch using predicted target attributes, should correctly recognize the attribute in the source domain. Based on this assumption, we propose a novel procedure to select attributes as Fig. 4 shows. First, we utilize all available non-color attributes to run our IHDA framework with the default training setting. Then, we estimate attributes for target testing photos, which are untouched during the training process. After that, target testing photos and corresponding predicted attributes are used to train the photo encoder (i.e., E_1) and attribute classifier (i.e., C_{att}) from the scratch. Finally, we estimate the source attributes and compare them with the ground truth. The comparison results are used for attribute selection.

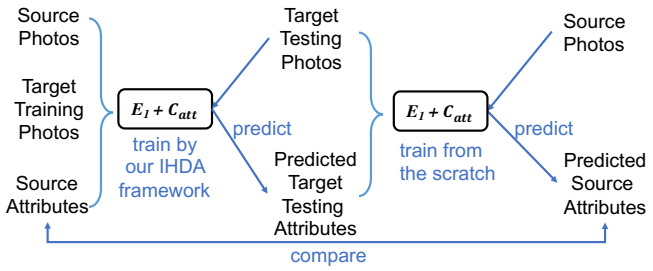


Fig. 4. The procedure of selecting attributes for IHDA framework.

We use the PKU-Sketch dataset to demonstrate such an attribute selection procedure. The attribute recognition accuracy on the source dataset is reported in TABLE XIV.

TABLE XIV
ATTRIBUTE RECOGNITION ACCURACY ON SOURCE DATASET (TRAINED ON TARGET PHOTO TRAINING SET). "L.SLV", "L.LOW", "S.CLTH" DENOTE "LENGTH OF SLEEVE", "LENGTH OF LOWER-BODY CLOTHING", "STYLE OF CLOTHING", RESP.

Attributes	L.slv	L.low	bag	gender	hair	S.clth	hat	age
Accuracy (%)	81.6	75.9	73.7	72.1	70.5	68.4	62.1	37.5

Based on our assumption, the attribute with a higher recognition accuracy might contribute more in our IHDA framework. To verify this assumption, we further select two groups of attributes, as "L.slv, L.low, bag" and "S.clth, hat, age", to run IHDA framework. The results in TABLE XV agree with our assumption.

TABLE XV
THE SKETCH-TO-PHOTO RETRIEVAL PERFORMANCE BY USING DIFFERENT ATTRIBUTES ON PKU-SKETCH DATASET.

Model	R1	R5	R10	R20
IHDA (w/ L.slv, L.low, bag)	84.2	94.4	97.8	100.0
IHDA (w/ S.clth, hat, age)	80.0	93.2	96.6	98.0

Although we can select important attributes through the above method, it is inefficient. Considering that we have chosen closely related source datasets for target datasets, non-color attributes may generally be shared between them and the total number of attributes should not be a large number. Could we just use all non-color attributes?

We do another experiment on PKU-Sketch dataset to explore this question. Within all non-color attributes (8 in total), we randomly select k (range from 1 to 8) of them and repeat for 10 times to obtain average rank-1 value and uncertainty. As Fig. 5 shows, in the beginning, with more attributes, the rank-1 value is steadily increasing while the uncertainty is decreasing. When more than 6 attributes are used, there is a small difference in the rank-1 value and uncertainty. It suggests that, within a certain range, increasing the number of attributes

helps to improve the performance of IHDA framework. However, above a certain number, adding more shared attributes may not significantly affect the model performance. Therefore, for the sake of convenience, we may just utilize all of the non-color attributes of the source dataset.

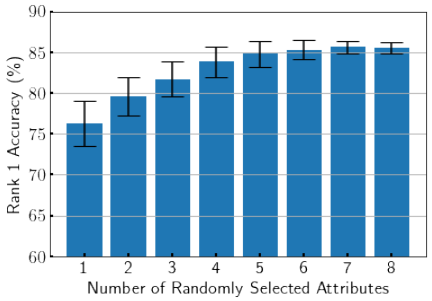


Fig. 5. Analysis on attribute selection.

Since each non-color attribute is defined as a binary label, it has both positive and negative samples in the source dataset. In the domain adaptation process, the framework will assign an attribute all to be negative if positive samples of this attribute do not exist in the target dataset. For such reason, within the same category, any non-color attributes of the source dataset can be used in forming the shared-label space, although some attributes defined in the source data might all be negative in the target data.

H. Trade-off Parameter Selection

In Fig. 6, we explore the sensitivity of trade-off parameters defined in Equation 9. From value 0.001 to 10, we vary one of them by fixing others as our default setting. It shows that λ_1 is not sensitive but very useful in IHDA framework. When λ_2 and λ_3 are assigned to 0.001, it yields a poor retrieval performance compared with assigning them to 0.1. This implies that attribute learning and domain adaptation are greatly beneficial to IHDA framework.

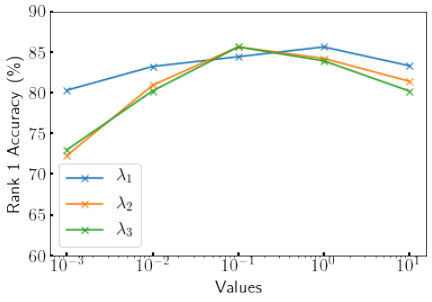


Fig. 6. Analysis on trade-off parameter selection.

I. Ablation Studies

To confirm the effectiveness of our IHDA framework, we conduct more detailed ablation studies by using PKU-Sketch and Market-1501 datasets. The experimental results are shown in TABLE XVI. In TABLE XVI, there are totally 11 variant settings for ablation studies. Among them, variant settings from (1) to (5) are used to investigate the IHDA performance at the different proportions of source and target data; variant settings from (6) to (12) are designed to study the effectiveness of each component in IHDA framework. We have the complete setup (i.e., components and datasets) of IHDA framework in (13), which considerably surpasses others.

Only using the pre-trained weights on ImageNet generates results that are almost identical to random guessing (see (1)). Due to the huge domain gaps between sketches and photos, solely using the pre-trained weights on Market-1501 (even with attributes) leads to poor performance (see (2)). Nevertheless, without external data, only training the model on PKU-Sketch also cannot obtain satisfactory results (see (3)) since lacking training data. Therefore, it is highly desirable to properly transfer the domain knowledge from Market-1501 to PKU-Sketch data. With complete components, the IHDA framework serves for such a purpose and generates promising results, even with 50% and 80% of PKU-Sketch training data (see (4) and (5)).

It is critical to design the proper framework to transfer the domain knowledge from Market-1501 to PKU-Sketch data. Satisfactory retrieval accuracy cannot be approached through pre-training our model on Market-1501 data and then fine-tuning it on PKU-Sketch data (see (6)). In (7), although we apply attribute learning on source data, it does not significantly improve the target retrieval performance by a fine-tuning on target data.

The alternative approach is to co-train the source and target data and reduce the domain gap between them. Without forming a shared label space via attributes, simply applying domain adaptation can roughly alleviate cross-domain discrepancies. Nonetheless, the instance-level source domain knowledge may not be properly transferred and leads to limited improvement (see (8)). On the other hand, without the domain adaptation, solely guided by shared attributes also cannot accomplish promising results (see (12)). These results are consistent with the analysis in Fig. 5. In (9), we show that target unsupervised attribute learning plays an important role in IHDA. The entropy-minimization loss and attribute-consistent loss work jointly to achieve optimal attribute-guided domain adaptation. Specifically, we verify the effectiveness of entropy-minimization and attribute-consistent in setting (10) and (11), respectively. The $\mathcal{L}_{att}^{t1,t2}$ contributes dominant effects while $\mathcal{L}_{con}^{t1,t2}$ can further improve the effectiveness.

Fig. 8 shows the Rank-1 accuracy on the PKU-sketch target set as the number of training epochs goes up. By using source data $\mathcal{X}$, The retrieval performance can increase faster than only using target data. Although there are many potential ways to utilize the source data, which includes aligned sketch-photo pairs and attributes, we propose the most effective one in the full setting proposal.

Furthermore, we analyze how embedding feature distributions change in different learning settings. For visualization purposes, we only select the gender attribute, as 'Male' and 'Female' in IHDA. By randomly picking up samples of 10 identities from the testing set of PKU-Sketch, we project their embedding features into a 2D space and show them in Fig. 7. In simple fine-tuning strategy (i.e., TABLE XVI.(6)),

TABLE XVI
THE EXPERIMENTAL SETTINGS AND RESULTS OF ABLATION STUDIES. IN THE TABLE, 'MARK.' STANDS FOR THE MARKET-1501 DATASET, AND 'PKU.' STANDS FOR THE PKU-SKETCH DATASET.

Index	Description	Modules						Losses								Training Data	R1	R5	R10	R20
		E_1	E_2	C_{id}^{s}	$C_{id}^{t1,t2}$	C_{att}	C_d	$\mathcal{L}_{id}^{s}$	$\mathcal{L}_{id}^{t1,t2}$	$\mathcal{L}_{tri}^{s}$	$\mathcal{L}_{tri}^{t1,t2}$	$\mathcal{L}_{att}^{s}$	$\mathcal{L}_{att}^{t1,t2}$	$\mathcal{L}_d$	$\mathcal{L}_{con}^{t1,t2}$					
(1)	w/o $\mathcal{X}^s$&$\mathcal{X}^{t1,t2}$	✓	✓													ImageNet	2.0	8.0	20.0	40.0
(2)	w/o $\mathcal{X}^{t1,t2}$	✓	✓	✓		✓		✓		✓		✓				Mark.	8.0	30.0	44.0	58.0
(3)	w/o $\mathcal{X}^s$	✓	✓		✓				✓		✓					PKU.(100%)	32.0	58.6	68.2	82.6
(4)	w/ 50% $\mathcal{X}^{t1,t2}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Mark.+PKU.(50%)	75.4	90.6	92.0	94.8
(5)	w/ 80% $\mathcal{X}^{t1,t2}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Mark.+PKU.(80%)	82.2	92.6	95.2	98.0
(6)	w/o $\mathcal{L}_{att}^{s,t1,t2}$&$\mathcal{L}_d$	✓	✓	✓	✓			✓	✓	✓	✓					Mark.+PKU.(100%)	58.8	78.0	90.0	94.2
(7)	w/o $\mathcal{L}_{att}^{t1,t2}$&$\mathcal{L}_{con}^{t1,t2}$&$\mathcal{L}_d$	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓				Mark.+PKU.(100%)	59.4	79.2	91.2	95.8
(8)	w/o $\mathcal{L}_{att}^{s,t1,t2}$	✓	✓	✓	✓		✓	✓	✓	✓	✓			✓		Mark.+PKU.(100%)	70.4	82.0	92.6	96.2
(9)	w/o $\mathcal{L}_{att}^{t1,t2}$&$\mathcal{L}_{con}^{t1,t2}$	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓		Mark.+PKU.(100%)	70.6	83.2	92.4	95.8
(10)	w/o $\mathcal{L}_{con}^{t1,t2}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		Mark.+PKU.(100%)	82.8	94.2	96.8	98.4
(11)	w/o $\mathcal{L}_{att}^{t1,t2}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	Mark.+PKU.(100%)	73.2	85.6	92.6	96.8
(12)	w/o $\mathcal{L}_d$	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓	Mark.+PKU.(100%)	72.4	84.2	93.2	96.6
(13)	IHDA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Mark.+PKU.(100%)	**85.6**	**94.8**	**98.0**	**100.0**

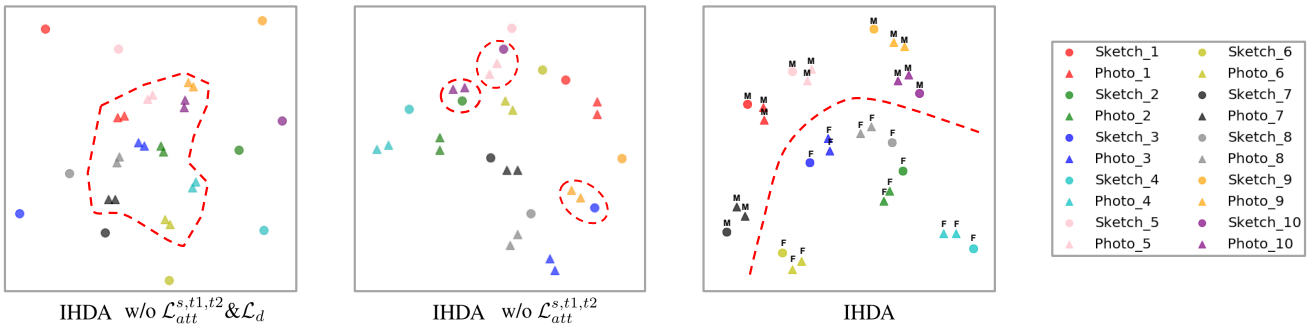


Fig. 7. Visualization of embedding features for 10 identities in PKU-Sketch testing set by t-SNE [62]. Each color represents a specific identity by ground-truth annotation. Triangles and circles denote the photo and sketch embedding, respectively. Character "M" and "F" indicate predicted "gender" attribute as "Male" and "Female", resp. Red dot lines highlight where to look at.

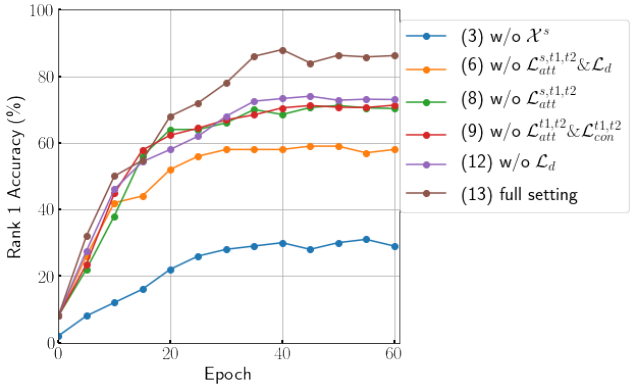


Fig. 8. Rank 1 accuracy on PKU-Sketch target set as the number of training epochs goes up.

although both Market-1501 dataset and PKU-Sketch training set are utilized, photo embedding has a boundary to sketch embedding. By applying a DA without attribute guidance (i.e., TABLE XVI.(7)), the boundary between photo and sketch is eliminated but the relative distance between some sketch-to-photo pairs could be incorrect. By using the complete IHDA, sketch-to-photo pairs have become more distinguishable with the attribute guidance.

V. DISCUSSION

It is hard to judge "pros" or "cons" in our methods without considering the application scenario. Our IHDA framework is specifically designed for **instance-level heterogeneous image retrieval tasks**, it may not be applicable for category-level retrieval tasks. In instance-level retrieval tasks, the shared attributes are guaranteed since all instances belong to the same category. In contrast, it is challenging to find shared attributes for all instances in category-level retrieval tasks, such as Sketchy dataset [60], PACS dataset [63] and M3SDA dataset [64], Nonetheless, as the key contribution, our IHDA framework has overcome the limitations of applying simple fine-tuning strategy and conventional DA. It enables using rich-labeled photo retrieval data to remarkably improve the retrieval performance on limited-labeled sketch-photo retrieval tasks, which opens the door to train more effective cross-modal image retrieval models.

VI. CONCLUSION

Nowadays, although a large amount of rich-labeled datasets have been made available to the public, how to utilize them

to benefit a related limited-labeled task may vary case by case and is waiting for further exploration. Under such background, we analyze challenges for limited-labeled sketch-to-photo retrieval from the transfer learning perspective and raise an Instance-level Heterogeneous Domain Adaptation framework to tackle them. It demonstrates an important insight: attributes, which are only annotated in the source domain, can be used to form a shared label space for both source and target domains in instance-level retrieval task. With such a shared-label space, instance-level domain knowledge can be well transferred across datasets and heterogeneous modalities. It opens the door to train more effective cross-modal image retrieval models by using related rich-labeled single-modal image data.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers JP17H06101, and a MSRA Collaborative Research 2019 Grant Awarded by Microsoft Research Asia.

REFERENCES

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 3934–3941.

[3] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 801–814, 2019.

[4] Z. Cao, L. Ma, M. Long, and J. Wang, "Partial adversarial domain adaptation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 139–155.

[5] H. Zou, Y. Zhou, J. Yang, H. Liu, H. P. Das, and C. J. Spanos, "Consensus adversarial domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 5997–6004.

[6] Y. Zou, Z. Yu, B. V. K. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 297–313.

[7] L. Pang, Y. Wang, Y.-Z. Song, T. Huang, and Y. Tian, "Cross-domain adversarial feature learning for sketch re-identification," in *Proceedings of the ACM international Conference on Multimedia*, 2018, pp. 609–617.

[8] S. Ouyang, T. M. Hospedales, Y. Song, and X. Li, "Forgetmenot: Memory-aware forensic facial sketch matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5571–5579.

[9] S. Wang, J. Zhang, T. X. Han, and Z. Miao, "Sketch-based image retrieval through hypothesis-driven object boundary selection with hlr descriptor," *IEEE Transactions on Multimedia*, vol. 17, no. 7, pp. 1045–1057, 2015.

[10] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy, "Sketch me that shoe," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 799–807.

[11] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1604–1615, 2016.

[12] S. D. Bhattacharjee, J. Yuan, Y. Huang, J. Meng, and L. Duan, "Query adaptive multiview object instance search and localization using sketches," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2761–2773, 2018.

[13] J. Choi, H. Cho, J. Song, and S. M. Yoon, "Sketchhelper: Real-time stroke guidance for freehand sketch retrieval," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 2083–2092, 2019.

[14] Z. Wang, Z. Wang, Y. Zheng, Y. Wu, W. Zeng, and S. Satoh, "Beyond intra-modality: A survey of heterogeneous person re-identification," *arXiv preprint arXiv:1905.10048*, 2019.

[15] A. Yu and K. Grauman, "Semantic jitter: Dense supervision for visual comparisons via synthetic images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5570–5579.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[17] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[18] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2376–2386, 2019.

[19] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Transactions on Multimedia*, 2020. [Online]. Available: http://dx.doi.org/10.1109/TMM.2020.2969782

[20] Z. Wang, R. Hu, Y. Yu, C. Liang, and W. Huang, "Multi-level fusion for person re-identification with incomplete marks," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1267–1270.

[21] Z. Wang, X. Bai, M. Ye, and S. Satoh, "Incremental deep hidden attribute learning," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 72–80.

[22] J. Song, Y. Song, T. Xiang, T. M. Hospedales, and X. Ruan, "Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval," in *Proceedings of the British Machine Vision Conference*, 2016.

[23] J. Song, Q. Yu, Y. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5552–5561.

[24] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[25] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 1679–1686.

[26] Z. Deng, X. Peng, and Y. Qiao, "Residual compensation networks for heterogeneous face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 8239–8246.

[27] K. Saito, Y. Ushiku, and T. Harada, "Asymmetric tri-training for unsupervised domain adaptation," in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 2988–2997.

[28] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[29] Z. Luo, Y. Zou, J. Hoffman, and L. F. Fei-Fei, "Label efficient learning of transferable representations acrosss domains and tasks," in *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 165–177.

[30] D. Liu, N. Wang, C. Peng, J. Li, and X. Gao, "Deep attribute guided representation for heterogeneous face recognition." in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2018, pp. 835–841.

[31] T. Gebru, J. Hoffman, and L. Fei-Fei, "Fine-grained recognition in the wild: A multi-task domain adaptation approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1349–1358.

[32] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4109–4118.

[33] S. Lin, H. Li, C.-T. Li, and A. C. Kot, "Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification," in *Proceedings of the British Machine Vision Conference*, 2018, pp. 19–32.

[34] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2275–2284.

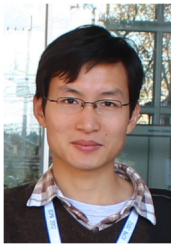
[35] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rJlnOhVYPS

[36] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognition*, p. 107173, 2020.

[37] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6467–6476.

[38] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 139–154.
[39] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2200–2207.
[40] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," *arXiv preprint arXiv:1904.06487*, 2019.
[41] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proceedings of the Advances in Neural Information Processing Systems*, 2004, pp. 529–536.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[43] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the International Conference on International Conference on Machine Learning*, 2015, pp. 1180–1189.
[44] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *Journal of Machine Learning Research*, vol. 11, no. Mar, pp. 1109–1135, 2010.
[45] Y. Shu, Z. Cao, M. Long, and J. Wang, "Transferable curriculum for weakly-supervised domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 4951–4958.
[46] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.
[47] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training." in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2018, pp. 677–683.
[48] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 2208–2217.
[49] J. Zhang, W. Li, and P. Ogunbona, "Jdeep face recognitionoint geometrical and statistical alignment for visual domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5150–5158.
[50] J. Wang, Y. Chen, S. Hao, W. Feng, and Z. Shen, "Balanced distribution adaptation for transfer learning," in *Proceedings of the IEEE International Conference on Data Mining*, 2017, pp. 1129–1134.
[51] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon, "On learning invariant representations for domain adaptation," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 7523–7532.
[52] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized MCWLD for matching sketches with digital face images," *IEEE Trans. Information Forensics and Security*, vol. 7, no. 5, pp. 1522–1535, 2012.
[53] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh, "Learning sparse and identity-preserved hidden attributes for person re-identification," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 2013–2025, 2019.
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[55] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.
[56] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, and M. Alexa, "Sketch-based shape retrieval," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 31:1–31:10, 2012.
[57] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[58] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference*, vol. 41, 2015, pp. 1–12.
[59] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 119:1–119:12, 2016.
[60] ——, "The sketchy database: learning to retrieve badly drawn bunnies," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–12, 2016.
[61] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.
[62] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
[63] D. Li, Y. Yang, Y. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the International Conference on Computer Vision*, 2017, pp. 5543–5551.
[64] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1406–1415.

Fan Yang received a B.S. degree and an M.S degree from Nanjing University and Nara Institute of Science and Technology in 2012 and 2018, respectively. He is currently a Ph.D. candidate at the Nara Institute of Science and Technology. His research focuses on video processing.



Yang Wu (M'19) received a BS degree and a Ph.D degree from Xi'an Jiaotong University in 2004 and 2010, respectively. He is currently a program-specific senior lecturer with Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. He is also a guest associate professor of Nara Institute of Science and Technology (NAIST), where he was an assistant professor of the NAIST International Collaborative Laboratory for Robotics Vision, from Dec.2014 to Jun. 2019. From 2011 to 2014, he was a program specific researcher with the Academic Center for Computing and Media Studies, Kyoto University. His research is in the fields of computer vision, pattern recognition, and image/video search and retrieval.

Zheng Wang (M'19) received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2006 and 2008, respectively, and the Ph.D. degree from the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, China, in 2017. He is currently a JSPS Fellowship Researcher at Shin'ichi Satoh's Lab, National Institute of Informatics, Japan. His research interests focus on person re-identification and instance search. He received the Best Paper Award at the 15th Pacific-Rim Conference on Multimedia (PCM 2014) and the 2017 ACM Wuhan Doctoral Dissertation Award.

Xiang Li received his B.S. degree from Sun Yat-sen University in 2019 (Distinguished Graduate). He is currently a Research Assistant at Nanyang Technological University. His research focus on machine learning and computer vision.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATRNICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. In 2011-2017, she was an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation. Currently, she is a research associate professor at NAIST, as well as a research scientist at RIKEN, the Center of for Advanced Intelligent Project AIP, Japan. She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. She is also the officer of ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL) and a board Member of Spoken Language Technologies for Under-Resourced Languages (SLTU). Her research interests include statistical pattern recognition, graphical modeling framework, deep learning, multilingual speech recognition & synthesis, spoken language translation, affective dialog system, and cognitive communication.

Satoshi Nakamura is Director of Data Science Center and Professor at the Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan, Team Leader of Riken AIP Tourism Information Analytics Team, and Honorarprofessor of Karlsruhe Institute of Technology, Germany. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was an Associate Professor of the Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director-General of Keihanna Research Laboratories and the Executive Director of Knowledge-Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor Nara Institute of Science and Technology, Japan. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving various speech-to-speech translation research projects in the world including C-STAR, IWSLT, and A-STAR. He received the Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affairs and Communications. He also received the LREC Antonio Zampolli Award 2012. He has been elected Board Member of International Speech Communication Association, ISCA, in 2011-2019, IEEE Signal Processing Magazine Editorial Board Member since April 2012-2015, IEEE SPS Speech and Language Technical Committee Member since 2013-2016. He is ATR Fellow, IPSJ Fellow, ISCA Fellow, and IEEE Fellow.