Learning Attributes Equals Multi-Source Domain Generalization

Chuang Gan
IIIS, Tsinghua University
ganchuang1990@gmail.com

Tianbao Yang
University of Iowa
tianbao-yang@uiowa.edu

Boqing Gong
CRCV, U. of Central Florida
bgong@crcv.ucf.edu

Abstract

Attributes possess appealing properties and benefit many computer vision problems, such as object recognition, learning with humans in the loop, and image retrieval. Whereas the existing work mainly pursues utilizing attributes for various computer vision problems, we contend that the most basic problem—how to accurately and robustly detect attributes from images—has been left under explored. Especially, the existing work rarely explicitly tackles the need that attribute detectors should generalize well across different categories, including those previously unseen. Noting that this is analogous to the objective of multi-source domain generalization, if we treat each category as a domain, we provide a novel perspective to attribute detection and propose to gear the techniques in multi-source domain generalization for the purpose of learning cross-category generalizable attribute detectors. We validate our understanding and approach with extensive experiments on four challenging datasets and three different problems.

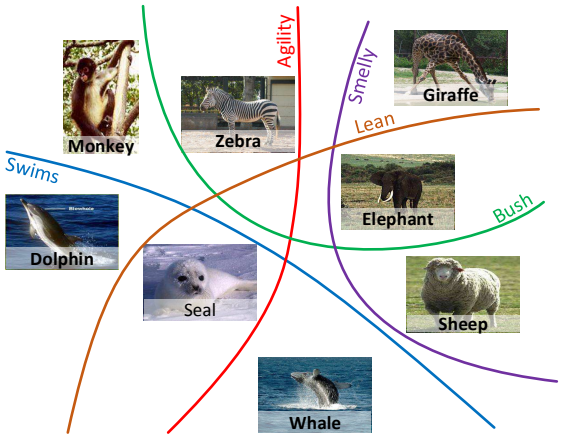

Figure 1. The boundaries between middle-level attributes and high-level object classes cross each other. We thus do not expect that the features originally learned for separating elephant, sheep, and giraffe could also be optimal for detecting the attribute "bush", which is shared by them. We propose to understand attribute detection as multi-source domain generalization and to explicitly break the class boundaries in order to learn high-quality attribute detectors.

1. Introduction

Visual attributes are middle-level concepts which humans use to describe objects, human faces, scenes, activities, and so on (e.g., four-legged, smiley, outdoor, and crowded). A major appeal of attributes is that they are not only human-nameable but also machine-detectable, making it possible to serve as the building blocks to describe instances [18, 42, 57, 55], teach machines to recognize previously unseen classes by zero-shot learning [44, 52], or offer a natural human-computer interactions channel for image/video search [64, 81, 40, 70].

However, we contend that the long-standing pursuit after utilizing attributes for various computer vision problems **has left the most basic problem—how to accurately and robustly detect attributes from images or videos—far from being solved.** Especially, the existing work rarely explicitly tackles the need that **attribute detectors should generalize well across different categories, including those previously unseen ones.** For instance, the attribute detector "four-legged" is expected to correctly tell a giant panda is four-legged even if it is trained from the images of horses, cows, zebras, and pigs (i.e., no pandas).

Indeed, most of the existing *attribute* detectors [44, 18, 42, 9, 32, 35, 8, 10, 12, 75, 31] are built using features engineered or learned for *object* recognition together with off-shelf machine learning classifiers—without tailoring them to reflect the idiosyncrasies of attributes. This is suboptimal; the successful techniques

on object recognition do not necessarily apply to attributes learning mainly for two reasons. First, attributes are in a different semantic space as opposed to objects; they are in the *middle* of low-level visual cues and the high-level object labels. Second, attribute detection can even be considered as an *orthogonal* task to object recognition, in that attributes are shared by different objects (e.g., zebras, lions, and mice are all "furry") and distinctive attributes are present in the same object (e.g., a car is boxy and has wheels). As shown in Figure 1, the boundaries between attributes and between object categories cross each other. Therefore, we do not expect that the features originally learned for separating elephant, sheep, and giraffe could also be optimal for detecting the attribute "bush", which is shared by them.

In this paper, we propose to re-examine the fundamental attribute detection problem and aim to develop an attribute-oriented feature representation, such that one can conveniently apply off-shelf classifiers to obtain high-quality attribute detectors. We expect that the detectors learned from our new representation are capable of breaking the boundaries of object categories and generalizing well across both seen and unseen classes. To this end, we cast **attribute detection as a multi-source domain generalization problem** [50, 80, 51] by noting that the desired properties from attributes are analogous to the objective of the latter.

Particularly, a domain refers to an underlying data distribution. Multi-source domain generalization aims to extract knowledge from several *related* source domains such that it is applicable to different domains, especially to those unseen at the training stage. This is in accordance with our objective for learning cross-category generalizable attributes detectors, if we consider each category as a distinctive domain.

Motivated by this observation, we employ the Unsupervised Domain-Invariant Component Analysis (UDICA) [50] as the basic building block for our approach. The key principle of UDICA is that minimizing the distributional variance of different domains—categories in our context, can improve the cross-domain (cross-category) generalization capabilities of the classifiers. A supervised extension to UDICA was introduced in [50] depending on the inverse of a covariance operator as well as some mild assumptions. However, the inverse operation is both computationally expensive and unstable in practice. We instead propose to use the alternative of centered kernel alignment [13] to account for the attribute labeling information. We show that the centered kernel alignment can be seamlessly integrated with UDICA, enabling us to learn both category-invariant and attribute-discriminative feature representations.

Our approach takes as input the features of the training images, their class (domain) labels, as well as their attribute labels. It operates upon kernels derived from the input data and learns a kernel projection to "distill" category-invariant and attribute-discriminative signals embedded in the original features. The overall output is a new feature vector for each image, which can be readily used in traditional machine learning models like SVMs for training the cross-category generalizable attribute detectors.

The contributions of the paper are summarized below.

- To the best of our knowledge, this work is the first attempt to tackle attribute detection from the multi-source domain generalization point of view. This enables us to explicitly model the need that the attribute detectors should transcend different categories and generalize to previously unseen ones.
- We introduce the centered kernel alignment to UDICA and arrive at an integrated method to strengthen the discriminative power of the learned attributes on one hand, and eliminate the domain differences between categories on the other hand.
- We test our approach to four datasets: Animal With Attributes [44], Caltech-UCSD Birds [76], aPascal-aYahoo [18], and UCF101 [67], and test the learned representations on three tasks: attribute detection itself, zero-shot learning, and image retrieval. Our results are significantly better than those of competitive baselines, verifying the effectiveness of the new perspective for solving attribute detection as domain generalization.

The rest of this paper is organized as follows. In Section 2, we review related work in attribute detection, domain generalization, and domain adaptation. Section 3 and section 4 present the attribute learning framework. The experimental settings and evaluation results are presented in Section 5. Section 6 concludes the paper.

2. Related work and background

Our approach is related to two separate research areas, attribute detection and domain adaptation/generalization. We unify them in this work.

Attributes learning. Earlier work on attribute detection mainly focused on modeling the correlations among attributes [9, 32, 35, 8, 10, 12, 75, 31], localizing some special part-related attributes (e.g., tails of mammals) [4, 6, 37, 3, 83, 59, 14], and the relationship between attributes and categories [79, 48, 32, 54]. Some

recent work has applied deep models to attribute detection [11, 83, 47, 16]. None of these methods explicitly model the cross-category generalization of the attributes, except the one by Farhadi et al. [18] where the authors select features within each category to down-weight category-specific cues. Likely due to the fact that the attribute and category cues are interplayed, their feature selection procedure only gives limited gain. We propose to overcome this challenge by investigating all categories together and employing nonlinear mapping functions.

Attributes possess versatile properties and benefit a wide range of challenging computer vision tasks. They serve as the basic building blocks for one to compose categories (e.g., different objects) [44, 52, 82, 17, 38, 78, 1, 34] and describe instances [18, 42, 57, 55, 77], enabling knowledge transfer between them. Attributes also reveal the rich structures underlying categories and are thus often employed to regulate machine learning models for visual recognition [69, 73, 45, 63, 33, 22, 21, 62]. Moreover, attributes offer a natural human-computer interaction channel for visual recognition with humans in the loop [7], relevance feedback in image retrieval [42, 64, 55, 60, 81, 40, 58, 25, 72], and active learning [41, 56, 5, 46, 43]. In this paper, we test the proposed approach on both attribute detection and its applications to zero-shot learning and image retrieval.

Domain generalization and adaptation. Domain generalization is still at its early developing stage. A feature projection-based algorithm, Domain-Invariant Component Analysis (DICA), was introduced in [50] to learn by minimizing the variance of the source domains. Recently, domain generation has been introduce into computer vision community for object recognition [80, 23] and video recognition [51]. We propose to gear multi-source domain generalization techniques for the purpose of learning cross-category generalizable attribute detectors. Multi-source domain adaptation [49, 29, 24, 71, 15] is related to our approach if we consider a transductive setting (i.e., the learner has access to the test data). While it assumes a single target domain, in attribute detection the test data are often sampled from more than one unseen domain.

2.1. Background on distributional variance

Denote by $\mathcal{H}$ and $k(\cdot,\cdot)$ respectively a Reproducing Kernel Hilbert Space and its associated kernel function. For an arbitrary distribution $P_y(\boldsymbol{x})$ indexed by $y \in \mathcal{Y}$, the following mapping,

$$\mu[P_y(\boldsymbol{x})] = \int k(\boldsymbol{x},\cdot)\mathrm{d}P_y(\boldsymbol{x}) \triangleq \mu_y \tag{1}$$

is injective if k is a characteristic kernel [66, 27, 68]. In other words, the kernel mean map μ_y in the RKHS $\mathcal{H}$ preserves all the statistical information of $P_y(\boldsymbol{x})$.

The distributional variance follows naturally,

$$\mathbb{V}(\mathcal{Y}) = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} \|\mu_y - \mu_0\|^2_{\mathcal{H}},\ \widehat{\mathbb{V}}(\mathcal{Y}) = \mathrm{tr}(KQ), \tag{2}$$

where μ_0 is the map of the mean of all the distributions in $\mathcal{Y}$. In practice, we do not have access to the distributions. Instead, we observe the samples $S_y, y \in \mathcal{Y}$ each drawn from a distribution $P_y(\boldsymbol{x})$ and can thus empirically estimate the distributional variance by $\widehat{\mathbb{V}}(\mathcal{Y}) = \mathrm{tr}(KQ)$. Here K is the (centered)[1] kernel matrix over all the samples, and Q collects the coefficients which depend on only the numbers of samples. We refer the readers to [50] for more details including the consistency between the distributional variance $\mathbb{V}$ and its estimate $\widehat{\mathbb{V}}$.

3. Attribute detection

This section formalizes attribute detection and shows its in-depth connection to domain generalization.

Problem statement. Suppose that we have access to an annotated dataset of M images. They are in the form of $(\boldsymbol{x}_m, \boldsymbol{a}_m, y_m)$ where $\boldsymbol{x}_m \in \mathbb{R}^{\mathsf{D}}$ is the feature vector extracted from the m-th image I_m, $m \in [\mathsf{M}] \triangleq \{1, 2, \cdots, \mathsf{M}\}$. Two types of annotations are provided for each image, the category label $y_m \in [\mathsf{C}]$ and the attribute annotations $\boldsymbol{a}_m \in \{0,1\}^{\mathsf{A}}$. Though we use binary attributes (e.g., the presence or absence of stripes) to in this paper for clarity, it is straightforward to extend our approach to multi-way and continuous-valued attributes. Note that a particular attribute $\boldsymbol{a}^i_m$ could appear in many categories (e.g., zebras, cows, giant pandas, lions, and mice are all furry). Moreover, there may be test images from previously unseen categories $\{\mathsf{C}+1, \mathsf{C}+2, \cdots\}$ for example in zero-shot learning. Our objective is to learn accurate and robust attribute detectors $\mathcal{C}(\boldsymbol{x}_m) \in \{0,1\}^{\mathsf{A}}$ to well generalize across different categories, especially to be able to perform well on the unseen classes.

Attribute detection as domain generalization —A new perspective. In this paper, we understand attribute detection as a domain generalization problem. A

[1] All kernels discussed in this paper have been centered [13].

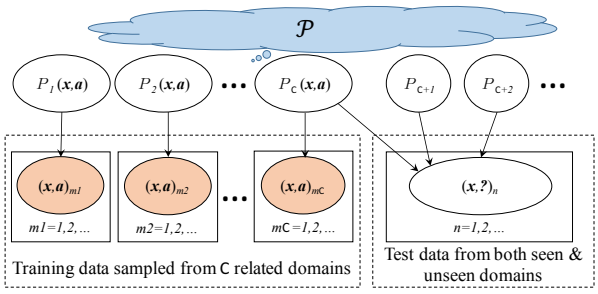

Figure 2. Attribute detection as multi-source domain generalization. Given labeled data sampled from several categories/domains, i.e., distributions $P_y(\boldsymbol{x}, \boldsymbol{a}), y \in [\mathsf{C}]$ over image representations $\boldsymbol{x}$ and attribute labels $\boldsymbol{a}$, we extract knowledge useful for attribute detection and applicable to different domains/categories, especially to previously unseen ones $P_{\mathsf{C}+1}, P_{\mathsf{C}+2}, \cdots$. The domains are assumed related and sampled from a common distribution $\mathcal{P}$.

domain refers to an underlying data distribution. In our context, it refers to the distribution $P_y(\boldsymbol{x}, \boldsymbol{a})$ of a category $y \in [\mathsf{C}]$ over the input $\boldsymbol{x}$ and attribute labels $\boldsymbol{a}$. As shown in Figure 2, the domains/categories are assumed to be related and are sampled from a common distribution $\mathcal{P}$. This is reasonable considering that images and categories can often be organized in a hierarchy. Thanks to the relationships between different categories, we expect to learn new image representations for attribute detection, such that the corresponding detectors will perform well on both seen and unseen classes.

4. Approach

Our key idea is to find a feature transformation of the input $\boldsymbol{x}$ to eliminate the mismatches between different domains/categories in terms of their marginal distributions over the input, whereas ideally we should consider the joint distributions $P_y(\boldsymbol{x}, \boldsymbol{a}), y \in [\mathsf{C}]$. In particular, we use Unsupervised Domain Invariant Component Analysis (UDICA) [50] and centered kernel alignment [13] for this purpose. Note that modeling the marginal distributions $P_y(\boldsymbol{x})$ is a common practice in domain generalization [50, 80, 23] and domain adaptation [30, 53, 26] and performs well in many applications. We leave investigating the joint distributions $P_y(\boldsymbol{x}, \boldsymbol{a})$ for future work.

Next, we present how to integrate UDICA and centered kernel alignment. Jointly they give rise to new feature representations which account for both attribute discriminativeness and cross-category generalizability.

4.1. UDICA

The projection from one RKHS to another results in the following transformation of the kernel matrices, $\mathbb{R}^{\mathsf{M}\times\mathsf{M}} \ni K \mapsto \widetilde{K} = KBB^TK \in \mathbb{R}^{\mathsf{M}\times\mathsf{M}}$ [61]. As a result, one can take (KB) as the empirical kernel map, i.e., consider the m-th row of (KB) as the new feature representations of image I_m and plug them into any linear classifiers. UDICA learns the transformation B by imposing the following properties.

Minimizing distributional variance. The empirical distributional variance (cf. Section 2.1) between different domains/categories becomes the following in our context,

$$\mathbb{V}_B([\mathsf{C}]) = \mathrm{tr}(\widetilde{K}Q) = \mathrm{tr}(B^TKQKB). \tag{3}$$

Intuitively, the domains would be perfectly matched when the variance is 0. Since there are many seen categories, each as a domain, we expect the learned projection to be generalizable and work well for the unseen classes as well.

Maximizing data variance. Starting from the empirical kernel map (KB), it is not difficult to see that the data covariance is $(KB)^T(KB)/\mathsf{M}$ and the variance is

$$\mathbb{V}_B([\mathsf{M}]) = \mathrm{tr}(B^TK^2B)/\mathsf{M}. \tag{4}$$

Regularizing the transformation. UDICA regularizes the transformation by minimizing

$$\mathcal{R}(B) = \mathrm{tr}(B^TKB). \tag{5}$$

Alternatively, one can use the Frobenius norm $\|B\|_F$, as did in [53], to constrain the complexity of B.

Combining the above criteria, we arrive at the following problem,

$$\max_B \quad \frac{\mathrm{tr}(B^TK^2B)/\mathsf{M}}{\mathrm{tr}(B^TKQKB + B^TKB)}, \tag{6}$$

where the nominator corresponds to the data variance and the denominator sums up the distributional variance and the regularization over B.

By solving the above problem, we are essentially blurring the boundaries between different categories and match the classes with each other, due to the distributional variance term in the denominator. This thus eliminates the barrier for attribute detectors to generalize in various classes. Our experiments verify the effectiveness the learned new representations (KB). Nonetheless, we can further improve the performance by modeling the attribute labels using centered kernel alignment.

4.2. Centered kernel alignment

Note that our training data are in the form of $(\boldsymbol{x}_m, \boldsymbol{a}_m, y_m), m \in [\mathsf{M}]$. For each image there are multiple attribute labels which may be highly correlated. Besides, we would like to stick to kernel methods to be consistent with our choice of UDICA—indeed, the distributional variance is best implemented by kernel methods (cf. Section 2.1). These considerations lead to our decision on using kernel alignment [13] to model the multi-attribute supervised information.

Let $L_{m,m'} = \langle \boldsymbol{a}_m, \boldsymbol{a}_{m'} \rangle$ be the kernel matrix over the attributes. Since L is computed directly from the attribute labels, it preserves the correlations among them and serves as the "perfect" target kernel for the transformed kernel $\widetilde{K} = KBB^TK$ to align to. The centered kernel alignment is then computed by,

$$\rho(\widetilde{K}, L) = \frac{\mathrm{tr}(\widetilde{K}L)}{\sqrt{\mathrm{tr}(\widetilde{K}\widetilde{K})}\sqrt{\mathrm{tr}(LL)}} \tag{7}$$

where we abuse the notation L slightly to denote that it is centered [13].

We would like to integrate the kernel alignment with UDICA in a unified optimization problem. To this end, firstly it is safe to drop $\mathrm{tr}(LL)$ in eq. (7) since it has nothing to do with the projection B we are learning. Moreover, note that the role of $\mathrm{tr}(\widetilde{K}\widetilde{K})$ duplicates with the regularization in eq. (6) to some extent, as it is mainly to avoid trivial solutions for the kernel alignment. We thus only add $\mathrm{tr}(\widetilde{K}L)$ to the nominator of UDICA,

$$\max_{B} \quad \frac{\mathrm{tr}(\gamma B^T K^2 B/\mathsf{M} + (1-\gamma)B^T KLKB)}{\mathrm{tr}(B^T KQKB + B^T KB)}, \tag{8}$$

where $\gamma \in [0, 1]$ balances the data variance and the kernel alignment with the supervised attribute labeling information. We cross-validate γ in our experiments. We name this formulation **KDICA**, which couples the centered kernel alignment and UDICA. The former closely tracks the attribute discriminative information and the latter facilitates the cross-category generalization of the attribute detectors to be trained upon KDICA.

Optimization. By writing out the Lagrangian of the formalized problem (eq. (8)) and then setting the derivative with respect to B to 0, we arrive at a generalized eigen-decomposition problem,

$$\begin{aligned} &\left(\gamma K^2/\mathsf{M} + (1-\gamma)KLK\right) B \\ &= \left(KQK + K\right) B\Gamma, \end{aligned} \tag{9}$$

Algorithm 1 Kernel-alignment Domain-Invariant Component Analysis (KDICA).

Input: Parameters γ and $\mathsf{b} \ll \mathsf{M}$. Training data $\mathcal{S} = \{(\boldsymbol{x}_m, y_m, \boldsymbol{a}_m)\}_{m=1}^{\mathsf{M}}$

Output: Projection $B_{\mathsf{M}\times\mathsf{b}}$

1: Calculate gram matrix $[K_{ij}] = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $[L_{ij}] = l(\boldsymbol{a}_i, \boldsymbol{a}_j)$

2: Solve: $(\gamma K^2/\mathsf{M} + (1-\gamma)KLK)B = (KQK + K)B\Gamma$.

3: Output B and $\widetilde{K} \leftarrow KBB^{\mathrm{T}}K$

4: Use (KB) as if they are the features to learn linear classifiers and $\widetilde{K}$ for kernelized classifiers

where Γ is a diagonal matrix containing all the eigenvalues (Lagrangian multipliers). We find the solution B as the Leading eigen-vectors. The number of eigen-vectors is cross-validated in our experiments. Again, we remind that (KB) serves as the new feature representations of the images for training attribute detectors. The details of our proposed framework has been shown in algorithm 1.

5. Experiment

This section presents our experimental results on four benchmark datasets. We test our approach for both the immediate task of attribute detection and two other problems, zero-shot learning and image retrieval, which could benefit from high-quality attribute detectors.

5.1. Experiment setup

Dataset. We use four datasets to validate the proposed approach; three of them contain images for object and scene recognition and the last one contains videos for action recognition. (a) The **Animal with attribute (AWA)** [44] dataset comprises of 30,475 images belonging to 50 animal classes. Each class is annotated with 85 attributes. Following the standard split by the dataset, we divide the dataset into 40 classes (24,295 images) to be used for training and 10 classes (6,180 images) for testing. (b) **Caltech-UCSD Birds 2011 (CUB)** [76] is a dataset with fine-grained objects. There are 11,788 images of 200 different bird classes in CUB. Each class is annotated with 312 binary attributes. We split the dataset as suggested in [1] to facilitate direct comparison (150 classes for training and 50 classes for testing). (c) **aPascal-aYahoo** [18] consists of two attribute datases: the a-PASCAL dataset, which contains 12,695 images (6,340 for training and 6,355 for testing) collected for the Pascal VOC 2008 challenge, and a-Yahoo including 2,644 test images. Each images is annotated with 64 attributes. There are 20 object classes in a-

Approcahes	AWA	CUB	a-Yahoo	UCF101
IAP [44]	74.0/79.2*	74.9*	–	–
ALE [1]	65.7	60.3	–	–
HAP [12]	74.0/79.1*	68.5/74.1*	58.2*	72.1 ± 1.1
CSHAP_G [12]	74.3/79.4*	62.7/74.6*	58.2*	72.3 ± 1.0
CSHAP_H [12]	74.0/79.0*	68.5/73.4*	65.2*	72.4 ± 1.1
DAP [44]	72.8/78.9*	61.8/72.1*	77.4*	71.8 ± 1.2
UDICA (Ours)	**83.9**	**76.0**	**82.3**	**74.3 ± 1.3**
KDICA (Ours)	**84.4**	**76.4**	**84.7**	**75.5 ± 1.1**

Table 1. Average Attribute Prediction Accuracy (%, in AUC.)

Pascal and 12 in a-Yahoo and they are disjoint. Following the settings of [64, 81], we use the pre-defined training images in a-Pascal as the training set and test on a-Yahoo. (d) **UCF101 dataset** [67] is a large dateset for video action recognition. It contains 13,320 videos of 101 action classes. Each action class comes with 115 attributes. The videos are collected from YouTube with large variations in camera motion, object appearance, viewpoint, cluttered background, and illumination conditions. We run 10 rounds of experiments each with a random split of 81/20 classes for the training/testing sets, and then report the averaged results.

Features. For the first three image datasets, we use the Convolutional Neural Network (CNN) implementation provided by Caffe [36], particularly with the 19-layer network architecture and parameters from Oxford [65], to extract 4,096-dimensional CNN feature representations from images ($i.e.$, the activations of the first fully-connected layer fc6). For the UCF101 dataset, we use the 3D CNN (C3D) [74] pre-trained on the Sport 1M dataset [39] to construct video-clip features from both spatial and temporal dimensions. We then use average pooling to obtain the video-level representations. We ℓ_2 normalize the feature representations in the following experiments.

Implementation details. We choose the Gaussian RBF kernel and fix the bandwidth parameter as 1 for our approach to learning new image representations. After that, to train the attribute detectors, we input the learned representations into standard linear Support Vector Machines (see the empirical kernel map in Setcion 4.1). There are two free hyper-parameters when we train the detectors using the representations learned through UDICA, the hyper-parameter C in SVM and the number b of leading eigen-vectors in UDICA. We use five-fold cross-validation to choose the best values for C and b respectively from $\{0.01, 0.1, 1, 10, 100\}$ and $\{30, 50, 70, 90, 110, 130, 150\}$. We use the same C and b for KDICA and only cross-validate γ in equation (9) from $\{0.2, 0.5, 0.8\}$ to learn the SVM based attribute detectors with KDICA.

Evaluation. We first test our approach to attribute detection on all the four datasets (AWA, CUB, aPascal-aYahoo, and UCF101). To see how much the other tasks, which involve attributes, can gain from higher-quality attribute detectors, we further conduct zero-shot learning [52, 44] experiments on AWA, CUB, and UCF101, and multi-attribute based image retrieval on AWA. We evaluate the results of attribute detection and image retrieval by the averaged Area Under ROC Curve (AUC), the higher the better, and the results of zero-shot learning by classification accuracy.

5.2. Attribute prediction

Table 1 presents the attribute prediction performance of our approaches and several competitive baselines. In particular, we compare with four state-of-the-art attribute detection methods: Directly Attribute Prediction (DAP) [44], Indirectly Attribute Prediction (IAP) [44], Attribute Label Embedding (ALE) [1], and Hypergraph-regularized Attribute Predictors (HAP) [12]. Note that we can directly contrast our methods with DAP to see the effectiveness of the learned new representations, because they share the same input and classifiers and only differ in that we learn the new attribute-discriminative and category-invariant feature representations. The IAP model first maps any input to the seen classes and then predicts the attributes on top of those. The ALE method unifies attribute prediction with object class prediction instead of directly optimizing with respect to attributes. We thus do not expect it to perform quite well on the attribute prediction task. HAP explores the correlations among attributes explicitly by hyper-graphs, while we achieve this implicitly in the kernel alignment. Additionally, we also show the results of CSHAP_G and CSHAP_H, two variations of HAP to model class labels.

We include in Table 1 both the results of these methods reported in the original papers, when they are available, and those we obtained (marked by '*') by running the source code provided by the authors. We use

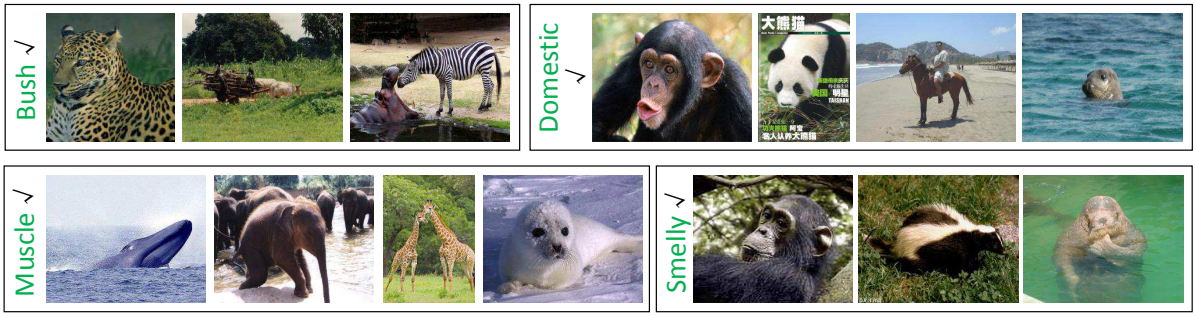

Figure 3. Some attributes on which the proposed KDICA significantly improves the performance of DAP.

Approcahes	AWA	CUB	UCF101
ALE [1]	37.4	18.0	–
HLE [1]	39.0	12.1	–
AHLE [1]	43.5	17.0	–
DA [35]	30.6	–	–
MLA [19]	41.3	–	–
ZSRF [34]	48.7	–	–
SM [20]	66.0	–	–
Embedding [2]	60.1	29.9	-
IAP [44]	42.2/49.4*	4.6/34.9*	–
HAP [12]	45.0/55.6*	17.5/40.7*	–
CSHAP_G [12]	45.0/54.5*	17.5/38.7*	–
CSHAP_H [12]	45.6/53.3*	17.5/36.9*	–
DAP [44]	41.2/58.9*	10.5/39.8*	26.8 ± 1.1
UDCIA (Ours)	**63.6**	**42.4**	$\mathbf{29.6 \pm 1.2}$
KDCIA (Ours)	**73.8**	**43.7**	$\mathbf{31.1 \pm 0.8}$

Table 2. Zero-shot recognition performances. (%, in accuracy)

the same CNN features (for AWA, CUB, and aPascal-aYahoo) and C3D features (for UCF101) we extracted for the baselines and our approach.

Overall results. From Table 1, we can find that UDCIA and KDICA outperform all the baselines on all the four datasets. More specifically, the relative accuracy gains of UDCIA over DAP are 6.3% on the AWA dataset and 5.4% on the CUB dateset, respectively, under the same feature and experimental settings. These clearly validate our assumption that blurring the category boundaries improves the generalizabilities of attribute detectors to previously unseen categories. The KDICA with centered kernel alignment is slightly better than the UDICA approach by incorporating attribute discriminative signals into the new feature representations. Delving into the per-unseen-class attribute detection result, we find that our KDICA-based approach improves the results of DAP for 71 out of 85 attributes on AWA and 272 out of 312 on CUB.

When domain generalization helps. We give some qualitative analyses using Figure 3 and 4 here. For

Figure 4. Example attributes that KDICA could not improve the detection accuracy over the traditional DAP approach.

the attributes in Figure 3, the proposed KDICA significantly improves the performance of the DAP approach. Those attributes ("muscle", "domestic", etc.) appear in visually quite different object categories. It seems like breaking the category boundaries is necessary in this case in order to make the attribute detectors generalize to the unseen classes. On the other hand, Figure 4 shows the attributes for which our approach can hardly improve DAP's performance. The attribute "yellow" is too trivial to detect with nearly 100% accuracy already by DAP. The attribute "swim" is actually shared by visually similar categories, leaving not much room for KDICA to play any role.

5.3. Zero-shot learning

As the intermediate representations of images and videos, attributes are often used in high-level computer vision applications. In this section, we conduct experiments on zero-shot learning to examine whether the improved attribute detectors could also benefit this task.

Given our UDICA and KDICA based attribute detection results, we simply input them to the second layer of the DAP model [44] to solve the zero-shot learning problem. We then compare with several well-known zero-shot recognition systems as shown in Table 2. We run our own experiments for some of them whose source code are provided by the authors. The corresponding results are again marked by '*'.

query	VGG	UDICA	KDICA
single	78.9	83.9	**84.4**
double	77.2	79.5	**81.0**
triple	76.1	78.6	**79.4**

Table 3. Multi-attribute based image retrieval results on AWA by the late fusion of individual attribute detection scores. (%, in AUC)

query	VGG	UDICA	KDICA
single	76.3	78.5	**79.2**
double	75.9	76.1	76.1
triple	75.5	75.6	**75.8**

Table 4. Multi-attribute based image retrieval results on AWA by TagProp. (%, in AUC)

We see that in Table 2 the proposed simple solution to zero-shot learning outperforms the other state-of-the-art methods on the AWA, CUB, and UCF101 datasets, especially its immediate rival DAP. In addition, we notice that our kernel alignment technique (KDICA) improves the zero-shot recognition results over UDICA significantly on AWA. The improvements over UDICA on the other two datasets are also more significant than the improvements for the attribute prediction task (see Section 5.2 and Table 1). This observation is interesting; it seems like implying that increasing the quality of the attribute detectors is rewarding, because the increase will be magnified to even larger improvement for the zero-shot learning. Similar observation applies if we compare the differences between DAP and UDICA/KDICA respectively in Table 2 and Table 1. Finally, we note that our main purpose is indeed to investigate how better attribute detectors can benefit zero-shot learning. We do not expect to have a thorough comparison of the existing zero-shot learning methods.

5.4. Multi-attribute based image retrieval

In this section, we do some experiments on the AWA dataset for the multi-attribute based image retrieval, whose performance depends on the reliabilities of the attribute predictions. We input our learned feature representations to two popular frameworks for multi-attribute based image retrieval: TagProp [28] and the fusion of individual prediction scores [42]. In TagProp, we use its σML variant to compute the ranking scores of the multi-attributes queries. For the fusion of individual classifiers, we directly sum up the confidence scores corresponding to the multiple attributes in a query. The results of the fusion and TagProp are respectively shown in Table 3 and Table 4. We can observe that our attribute-oriented representations improve the fusion technique for image retrieval on a variety of queries (single attribute, attribute pairs, and triplets). Under the TagProp framework, the improvement is marginal on querying by attribute pairs and triples and significant for single-attribute queries.

6. Conclusion

In this paper, we propose to re-examine the fundamental attribute detection problem and develop a novel attribute-oriented feature representation by casting the problem as multi-source domain generalization, such that one can conveniently apply off-shelf classifiers to obtain high-quality attribute detectors. The attribute detectors learned from our new representation are capable of breaking the boundaries of object categories and generalizing well to unseen classes. Extensive experiment on four datasets, and three tasks, validate that our attribute representation not only improves the quality of attributes, but also benefits succeeding applications, such as zero-shot recognition and image retrieval.

Acknowledgement. This work was supported in part by NSF IIS-1566511. Chuang Gan was partially supported by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003. Tianbao Yang was partially supported by NSF IIS-1463988 and NSF IIS-1545995.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 3, 5, 6, 7

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 7

[3] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013. 2

[4] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 2

[5] A. Biswas and D. Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *CVPR*, 2013. 3

[6] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 2011. 2

[7] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010. 3

[8] C.-Y. Chen and K. Grauman. Inferring analogous attributes. In *CVPR*, 2014. 1, 2

[9] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012. 1, 2

[10] L. Chen, Q. Zhang, and B. Li. Predicting multiple attributes via relative multi-task learning. In *CVPR*, 2014. 1, 2

[11] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *CVPR*, 2015. 3

[12] S.-W. Choi, C. H. Lee, and I. K. Park. Scene classification via hypergraph-based semantic attributes subnetworks identification. In *ECCV*, 2014. 1, 2, 6, 7

[13] C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012. 2, 3, 4, 5

[14] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR, pages=3474–3481, year=2012*. 2

[15] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*, 2009. 3

[16] V. Escorcia, J. C. Niebles, and B. Ghanem. On the relationship between visual attributes and convolutional networks. In *CVPR*, 2015. 3

[17] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010. 3

[18] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2, 3, 5

[19] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(2):303–316, 2014. 7

[20] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, 2015. 7

[21] C. Gan, M. Lin, Y. Yang, Y. Zhuang, and A. G. Hauptmann. Exploring semantic inter-class relationships (SIR) for zero-shot action recognition. In *AAAI*, 2015. 3

[22] C. Gan, Y. Yang, L. Zhu, D. Zhao, and Y. Zhuang. Recognizing an action using its name: A knowledge-based approach. *International Journal of Computer Vision*, pages 1–17, 2016. 3

[23] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. *ICCV*, 2015. 3, 4

[24] B. Gong, K. Grauman, and F. Sha. Reshaping visual datasets for domain adaptation. In *NIPS*, 2013. 3

[25] B. Gong, J. Liu, X. Wang, and X. Tang. Learning semantic signatures for 3d object retrieval. *IEEE Transactions on Multimedia*, 15(2):369–377, 2013. 3

[26] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 4

[27] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample problem. In *NIPS*, 2006. 3

[28] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, pages 309–316, 2009. 8

[29] J. Hoffman, B. Kulis, T. Darrell, and K. Saenko. Discovering latent domains for multisource domain adaptation. In *ECCV*, 2012. 3

[30] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *NIPS*, pages 601–608, 2006. 4

[31] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang. Learning hypergraph-regularized attribute predictors. In *CVPR*, 2015. 1, 2

[32] S. J. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *CVPR*, 2011. 1, 2

[33] S. J. Hwang and L. Sigal. A unified semantic embedding: Relating taxonomies and attributes. In *NIPS*, 2014. 3

[34] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *NIPS*, pages 3464–3472, 2014. 3, 7

[35] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014. 1, 2, 7

[36] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, volume 2, page 4, 2014. 6

[37] J. Joo, S. Wang, and S.-C. Zhu. Human attribute recognition by rich appearance dictionary. In *ICCV*, 2013. 2

[38] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa. Online incremental attribute-based zero-shot learning. In *CVPR*, 2012. 3

[39] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 6

[40] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012. 1, 3

[41] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *ICCV*, 2011. 3

[42] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1, 3, 8

[43] S. Lad and D. Parikh. Interactively guiding semi-supervised clustering via attribute-based explanations. In *ECCV*, 2014. 3

[44] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2, 3, 5, 6, 7

[45] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010. 3

[46] L. Liang and K. Grauman. Beyond comparing image pairs: Setwise active learning for relative attributes. In *CVPR*, 2014. 3

[47] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *ICCV*, 2013. 3

[48] D. Mahajan, S. Sellamanickam, and V. Nair. A joint learning framework for attribute models and object descriptions. In *ICCV*, 2011. 2

[49] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *NIPS*, 2009. 3

[50] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. 2, 3, 4

[51] L. Niu, W. Li, and D. Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *CVPR*, 2015. 2, 3

[52] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 1, 3, 6

[53] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. 4

[54] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, pages 1681–1688, 2011. 2

[55] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011. 1, 3

[56] A. Parkash and D. Parikh. Attributes for classifier feedback. In *ECCV*, 2012. 3

[57] G. Patterson and J. Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012. 1, 3

[58] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi. Multi-attribute queries: To merge or not to merge? In *CVPR*, 2013. 3

[59] R. N. Sandeep, Y. Verma, and C. Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *CVPR*, 2014. 2

[60] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *CVPR*, 2012. 3

[61] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, 1997. 4

[62] Z. Shi, Y. Yang, T. M. Hospedales, and T. Xiang. Weakly supervised learning of objects, attributes and their associations. In *ECCV*, 2014. 3

[63] A. Shrivastava, S. Singh, and A. Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *ECCV*, 2012. 3

[64] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011. 1, 3, 6

[65] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 6

[66] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, 2007. 3

[67] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 6

[68] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010. 3

[69] Y. Su, M. Allan, and F. Jurie. Improving object classification using semantic attributes. In *BMVC*, 2010. 3

[70] C. Sun, C. Gan, and R. Nevatia. Automatic concept discovery from parallel text and visual corpora. In *ICCV*, pages 2596–2604, 2015. 1

[71] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye. A two-stage weighting framework for multi-source domain adaptation. In *NIPS*, 2011. 3

[72] R. Tao, A. W. Smeulders, and S.-F. Chang. Attributes and categories for generic instance search from one example. In *CVPR*, 2015. 3

[73] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010. 3

[74] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3d: Generic features for video analysis. *ICCV*, 2015. 6

[75] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, R. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. B. Blaschko, D. Weiss, B. Taskar, K. Simonyan, N. Saphra, and S. Mohamed. Understanding objects in detail with fine-grained attributes. In *CVPR*, 2014. 1, 2

[76] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 2, 5

[77] L. Wang, Y. Qiao, and X. Tang. Motionlets: Mid-level 3d parts for human motion recognition. In *CVPR*, pages 2674–2681, 2013. 3

[78] X. Wang and Q. Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *ICCV*, 2013. 3

[79] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*, 2010. 2

[80] Z. Xu, W. Li, L. Niu, and D. Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV*, 2014. 2, 3, 4

[81] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak attributes for large-scale image retrieval. In *CVPR*, 2012. 1, 3, 6

[82] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*, 2010. 3

[83] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, 2014. 2, 3