Museum Exhibit Identification Challenge for the Supervised Domain Adaptation and Beyond

Piotr Koniusz*[1,2], Yusuf Tas*[,1,2], Hongguang Zhang[2,1], Mehrtash Harandi[3],
Fatih Porikli[2], Rui Zhang[4]

[1]Data61/CSIRO, [2]Australian National University,
[3]Monash University, [4]Hubei University of Arts and Science
firstname.lastname@{data61.csiro.au[1], anu.edu.au[2], monash.edu[3]}, renata_zhang@sina.com[4]

Abstract. We study an open problem of artwork identification and propose a new dataset dubbed Open Museum Identification Challenge (Open MIC). It contains photos of exhibits captured in 10 distinct exhibition spaces of several museums which showcase paintings, timepieces, sculptures, glassware, relics, science exhibits, natural history pieces, ceramics, pottery, tools and indigenous crafts. The goal of Open MIC is to stimulate research in domain adaptation, egocentric recognition and few-shot learning by providing a testbed complementary to the famous Office dataset which reaches $\sim$90% accuracy. To form our dataset, we captured a number of images per art piece with a mobile phone and wearable cameras to form the source and target data splits, respectively. To achieve robust baselines, we build on a recent approach that aligns per-class scatter matrices of the source and target CNN streams. Moreover, we exploit the positive definite nature of such representations by using end-to-end Bregman divergences and the Riemannian metric. We present baselines such as training/evaluation per exhibition and training/evaluation on the combined set covering 866 exhibit identities. As each exhibition poses distinct challenges e.g., quality of lighting, motion blur, occlusions, clutter, viewpoint and scale variations, rotations, glares, transparency, non-planarity, clipping, we break down results w.r.t. these factors.

1 Introduction

Domain adaptation and transfer learning are widely studied in computer vision and machine learning [1, 2]. They are inspired by the human cognitive capacity to learn new concepts from very few data samples (cf. training classifier on millions of labeled images from the ImageNet dataset [3]). Generally, given a new (target) task to learn, the arising question is how to identify the so-called *commonality* [4, 5] between this task and previous (source) tasks, and transfer knowledge from the source tasks to the target one. Therefore, one has to address three questions: what to transfer, how, and when [4].

Domain adaptation and transfer learning utilize annotated and/or unlabeled data and perform tasks-in-hand on the target data *e.g.*, learning new categories from few annotated samples (supervised domain adaptation [6, 7]), utilizing available unlabeled data (unsupervised [8, 9] or semi-supervised domain adaptation [10, 7]). Similar is one- and few-shoot learning that trains robust class predictors from one/few samples [11].

*Both authors contributed equally. Our dataset can be found at `claret.wikidot.com`.

Recently, algorithms for supervised, semi-supervised and unsupervised domain adaptation such as *Simultaneous Deep Transfer Across Domains and Tasks* [7], *Second- or Higher-order Transfer (So-HoT)* of knowledge [5] and *Learning an Invariant Hilbert Space* [12], all combined with Convolutional Neural Networks (CNN) [13, 14], have reached state-of-the-art results $\sim$90% accuracy on classic benchmarks such as the Office dataset [15]. Such good results are due to fine-tuning of CNNs on the large-scale datasets such as ImageNet [3]. Indeed, fine-tuning of CNN is a powerful domain adaptation and transfer learning tool by itself [16, 17]. Thus, these works show saturation for CNN features on the Office [15] dataset and its newer Office+Caltech 10 variant [18].

Thus, we propose a new dataset for the task of exhibit identification in museum spaces that challenges domain adaptation/fine-tuning due to its significant domain shifts.

For the source domain, we captured the photos in a controlled fashion by Android phones *e.g.*, we ensured that each exhibit is centered and non-occluded in photos. We prevented adverse capturing conditions and did not mix multiple objects per photo unless they were all part of one exhibit. We captured 2–30 photos of each art piece from different viewpoints and distances in their natural settings.

For the target domain, we employed an egocentric setup to ensure *in-the-wild* capturing process. We equipped 2 volunteers per exhibition with cheap wearable cameras and let them stroll and interact with artworks at their discretion. Such a capturing setup is applicable to preference and recommendation systems *e.g.*, a curator takes training photos of exhibits with an Android phone while visitors stroll with wearable cameras to capture data from the egocentric perspective for a system to reason about the most popular exhibits. Open MIC contains 10 distinct source-target subsets of images from 10 different kinds of museum exhibition spaces, each exhibiting various photometric and geometric challenges, as detailed in Section 5.

To demonstrate the intrinsic difficulty of Open MIC, we chose useful baselines in supervised domain adaptation detailed in Section 5. They include fine-tuning CNNs on the source and/or target data and training a state-of-the-art So-HoT model [5] which we equip with non-Euclidean distances [19, 20] for robust end-to-end learning.

We provide various evaluation protocols which include: (i) training/evaluation per exhibition subset, (ii) training/testing on the combined set that covers all 866 identity labels, (iii) testing w.r.t. various scene factors annotated by us such as quality of lighting, motion blur, occlusions, clutter, viewpoint and scale variations, rotations, glares, transparency, non-planarity, clipping, *etc.*.

Moreover, we introduce a new evaluation metric inspired by the following saliency problem: As numerous exhibits can be captured in a target image, we asked our volunteers to enumerate in descending order the labels of most salient/central exhibits they had interest in at a given time followed by less salient/distant exhibits. As we ideally want to understand the volunteers' preferences, the classifier has to decide which detected exhibit is the most salient. We note that the annotation- and classification-related processes are not free of noise. Therefore, we propose to not only look at the top-k accuracy known from ImageNet [3] but to also check if any of top-k predictions are contained within the top-n fraction of all ground-truth labels enumerated for a target image. We refer to this as a top-k-n measure.

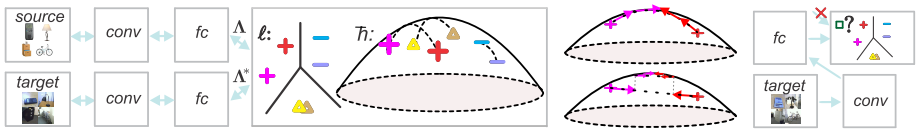

Fig. 1: The pipeline. Figure 1a shows the source and target network streams which merge at the classifier level. The classification and alignment losses ℓ and $\hbar$ take the data $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$ from both streams for end-to-end learning. Loss $\hbar$ aligns covariances on the manifold of $\mathcal{S}_{++}$ matrices. Fig. 1b (top) shows alignment along the geodesic path (ours). Fig. 1b (bottom) shows alignment via the Euclidean dist. [5]. At the test time, we use the target stream and the classifier as in Figure 1c.

To obtain convincing baselines, we balance the use of an existing approach [5] with our mathematical contributions[1] and evaluations. The So-HoT model [5] uses the Frobenius metric for partial alignment of within-class statistics obtained from CNNs. The hypothesis behind such modeling is that the partially aligned statistics capture so-called *commonality* [4, 5] between the source and target domains; thus facilitating knowledge transfer. For the pipeline in Figure 1, we use two CNN streams of the VGG16 network [14] which correspond to the source and target domains. We build scatter matrices, one per stream per class, from feature vectors of the *fc* layers. To exploit the geometry of positive definite matrices, we regularize and align scatters by the Jensen-Bregman LogDet Divergence (*JBLD*) [19] in end-to-end manner and compare to the Affine-Invariant Riemannian Metric (*AIRM*) [20, 21]. However, evaluations of gradients of non-Euclidean distances are slow for large matrices. We show by the use of Nyström projections that, with typical numbers of datapoints per source/target per class being $\sim$50 in domain adaptation, evaluating such distances is fast and exact.

Our contributions are: (i) we collect/annotate a new challenging Open MIC dataset with domains consisting of iamges taken by Android phones and wearable cameras; the latter exhibiting a series of realistic distortions due to the egocentric capturing process, (ii) we compute useful baselines, provide various evaluation protocols, statistics and top-k-n results, as well as include breakdown of results w.r.t. annotated by us scene factors, (iii) we use non-Euclidean JBLD and AIRM distances for end-to-end training of the supervised domain adaptation approach and we exploit the Nyström projections to make this training tractable. To our best knowledge, these distances have not been used before in the supervised domain adaptation due to their high computational complexity.

2 Related Work

Below we describe the most popular datasets for the problem at hand and explain how Open MIC differs. Subsequently, we describe related domain adaptation approaches.

Datasets. A popular dataset for evaluating against the effect of domain shift is the Office dataset [15] which contains 31 object categories and three domains: Amazon, DSLR

[1] We deal with large covariance matrices in a principled manner–the use of Euclidean distance is suboptimal in the light of Riemannian geometry. We make non-Euclidean distances tractable.

and Webcam. The 31 categories in the dataset consist of objects commonly encountered in the office setting, such as keyboards, file cabinets, and laptops. The Amazon domain contains images which were collected from a website of on-line merchants. Its objects appear on clean backgrounds and at a fixed scale. The DSLR domain contains low-noise high resolution images of object captured from different viewpoints while Webcam contains low resolution images. The Office dataset and its newer extension to Caltech 10 domain [18] are used in numerous domain adaptation papers [8, 7, 9, 6, 22, 23, 24, 12].

The Office dataset is primarily used for the transfer of knowledge about object categories between domains. In contrast, our dataset addresses the transfer of instances between domains. Each domain of the Open MIC dataset contains 37–166 specific instances to distinguish from (866 in total) compared to relatively low number of 31 classes in the Office dataset. Moreover, our target subsets are captured in an egocentric manner *e.g.*, we did not align objects to the center of images or control the shutter *etc.*.

A recent large collection of datasets for domain adaptation was proposed in technical report [25] to study cross-dataset domain shifts in object recognition with use of the ImageNet, Caltech-256, SUN, and Bing datasets. Even larger is the latest Visual Domain Decathlon challenge [26] which combines datasets such as ImageNet, CIFAR–100, Aircraft, Daimler pedestrian classification, Describable textures, German traffic signs, Omniglot, SVHN, UCF101 Dynamic Images, VGG–Flowers. In contrast, we target the identity recognition across exhibits captured in egocentric setting which vary from paintings to sculptures to glass to pottery to figurines. Many artworks in our dataset are fine-grained and hard to distinguish from without the expert knowledge.

The Office-Home dataset contains domains such as the real images, product photos, clipart and simple art impressions of well-aligned objects [27]. The Car Dataset [28] contains 'easily acquired' $\sim$1M cars of 2657 classes from websites for the fine-grained domain adaptation. Approach [29] uses 170 classes and $\sim$100 samples per class for attribute-based domain adaptation. Our Open MIC however is not limited to instances of cars or rigid objects. With 866 classes, Open MIC contains diverse 10 subsets with paintings, timepieces, sculptures, science exhibits, glasswork, relics, ancient animals, plants, figurines, ceramics, native arts *etc.*We captured varied materials, some of which are non-rigid, may emit light, be in motion or appear under large scale and viewpoint changes to form extreme yet realistic domain shifts. In some subsets, we also have large numbers[2] of frames for unsupervised domain adaptation.

Domain adaptation algorithms. Deep learning has been used in the context of domain adaptation in numerous recent works *e.g.*, [7, 9, 6, 22, 23, 24, 5]. These works establish the so-called commonality between domains. In [7], the authors propose to align both domains via the cross entropy which 'maximally confuses' both domains for supervised and semi-supervised settings. In [6], the authors capture the 'interpolating path' between the source and target domains using linear projections into a low-dimensional subspace on the Grassman manifold. Method [22] learns the transformation between the source and target by the deep regression network. Our model differs in that our source and target network streams co-regularize each other via the JBLD or AIRM distance

[2]We follow the the traditional domain adaptation paradigm that 'learning quickly from only a few examples is definitely the desired characteristic to emulate in any brain-like system' [30] in contrast to recent big data approaches [28, 29] which take on a complementary adaptation regime.

Dist./Ref.	$d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)$	Invar.	Tr. Ineq.	Geo.	d if $\mathcal{S}_+$	$\nabla_{\boldsymbol{\Sigma}}$ if $\mathcal{S}_+$	$\frac{\partial d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)}{\partial \boldsymbol{\Sigma}}$
Frobenius	$\|\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^*\|\|_F^2$	rot.	yes	no	fin.	fin.	$2(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^*)$
AIRM [20]	$\|\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-\frac{1}{2}}\|\|_F^2$	aff./inv.	yes	yes	∞	∞	$-2\boldsymbol{\Sigma}^{-\frac{1}{2}}\log(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-\frac{1}{2}})\boldsymbol{\Sigma}^{-\frac{1}{2}}$
JBLD [19]	$\log\left\|\frac{\boldsymbol{\Sigma}+\boldsymbol{\Sigma}^*}{2}\right\| - \frac{1}{2}\log\|\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\|$	aff./inv.	no	no	∞	∞	$(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}^*)^{-1} - \frac{1}{2}\boldsymbol{\Sigma}^{-1}$

Table 1: Frobenius, JBLD and AIRM distances and their properties. These distances operate between a pair of arbitrary matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^*$ which are points in $\mathcal{S}_{++}$ (and/or $\mathcal{S}_+$ for Frobenius).

that respects the non-Euclidean geometry of the source and target matrices (other dist. can also be used [31, 32]). We align covariances [5] via a non-Euclidean distance.

For visual domains, the domain adaptation can be applied in the spatially-local sense to target so-called *roots* of domain shift. In [24], the authors utilize so-called 'domainness maps' which capture locally the degree of domain specificity. Our work is orthogonal to this method. Our ideas can be extended to a spatially-local setting.

Correlation between the source and target distributions are often used. In [33], a subspace forms a joint representation for the data from different domains. Metric learning [34, 35] can be also applied. In [8] and [36], the source and target data are aligned in an unsupervised setting via correlation and Maximum Mean Discrepancy (MMD), resp. A baseline we use [5] can be seen as end-to-end trainable MMD with polynomial kernel as class-specific source and target distributions are aligned by the kernelized Frobenius norm on tensors. Our work is somewhat related. However, we first project class-specific vector representations from the last *fc* layers of the source and target CNN streams to the common space via Nyström projections for tractability and then we combine them with the JBLD or AIRM distance to exploit the (semi)definite positive nature of scatter matrices. We perform end-to-end learning which requires non-trivial derivatives of JBLD/AIRM distance and Nyström projections for computational efficiency.

3 Background

Below we discuss scatter matrices, Nyström projections, the Jensen-Bregman LogDet (*JBLD*) divergence [19] and the Affine-Invariant Riemannian Metric (*AIRM*) [20, 21].

3.1 Notations

Let $\boldsymbol{x} \in \mathbb{R}^d$ be a d-dimensional feature vector. $\mathcal{I}_N$ stands for the index set $\{1, 2, ..., N\}$. The Frobenius norm of matrix is given by $\|\boldsymbol{X}\|_F = \sqrt{\sum_{m,n} X_{mn}^2}$, where X_{mn} represents the (m, n)-th element of $\boldsymbol{X}$. The spaces of symmetric positive semidefinite and definite matrices are $\mathcal{S}_+^d$ and $\mathcal{S}_{++}^d$. A vector with all coefficients equal one is denoted by $\mathbb{1}$ and $\boldsymbol{J}_{mn}$ is a matrix of all zeros with one at position (m, n).

3.2 Nyström Approximation

In our work, we rely on Nyström projections, thus, we review their mechanism first.

Proposition 1. *Suppose* $\boldsymbol{X} \in \mathbb{R}^{d \times N}$ *and* $\boldsymbol{Z} \in \mathbb{R}^{d \times N'}$ *store* N *feature vectors and* N' *pivots (vectors used in approximation) of dimension* d *in their columns, respectively. Let* $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *be a positive definite kernel. We form two kernel matrices* $\boldsymbol{K_{ZZ}} \in \mathcal{S}_{++}^{N'}$ *and* $\boldsymbol{K_{ZX}} \in \mathbb{R}^{N' \times N}$ *with their* (i,j)*-th elements being* $k(\boldsymbol{z}_i, \boldsymbol{z}_j)$ *and* $k(\boldsymbol{z}_i, \boldsymbol{x}_j)$*, respectively. Then the Nyström feature map* $\tilde{\boldsymbol{\Phi}} \in \mathbb{R}^{N' \times N}$*, whose columns correspond to the input vectors in* $\boldsymbol{X}$*, and the Nyström approximation of kernel* $\boldsymbol{K_{XX}}$ *for which* $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ *is its* (i,j)*-th entry, are given by:*

$$\tilde{\boldsymbol{\Phi}} = \boldsymbol{K_{ZZ}}^{-0.5} \boldsymbol{K_{ZX}} \quad \text{and} \quad \boldsymbol{K_{XX}} \approx \tilde{\boldsymbol{\Phi}}^T \tilde{\boldsymbol{\Phi}}. \tag{1}$$

Proof. See [37] for details. □

Remark 1. The quality of approximation of (1) depends on the kernel k, data points $\boldsymbol{X}$, pivots $\boldsymbol{Z}$ and their number N'. In the sequel, we exploit a specific setting under which $\boldsymbol{K_{XX}} = \tilde{\boldsymbol{\Phi}}^T \tilde{\boldsymbol{\Phi}}$ which indicates no approximation loss.

3.3 Scatter Matrices

We make a frequent use of distances $d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)$ that operate between covariances $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(\boldsymbol{\Phi})$ and $\boldsymbol{\Sigma}^* \equiv \boldsymbol{\Sigma}(\boldsymbol{\Phi}^*)$ on feature vectors. Therefore, we provide a useful derivative of $d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)$ w.r.t. feature vectors $\boldsymbol{\Phi}$.

Proposition 2. *Let* $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_N]$ *and* $\boldsymbol{\Phi}^* = [\boldsymbol{\phi}_1^*, ..., \boldsymbol{\phi}_{N^*}^*]$ *be feature vectors of quantity* N *and* N^* e.g.*, formed by Eq.* (1) *and used to evaluate* $\boldsymbol{\Sigma}$ *and* $\boldsymbol{\Sigma}^*$ *with* $\boldsymbol{\mu}$ *and* $\boldsymbol{\mu}^*$ *being the mean of* $\boldsymbol{\Phi}$ *and* $\boldsymbol{\Phi}^*$*. Then derivatives of* $d^2 \equiv d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)$ *w.r.t.* $\boldsymbol{\Phi}$ *and* $\boldsymbol{\Phi}^*$ *are:*

$$\frac{\partial d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)}{\partial \boldsymbol{\Phi}} = \frac{2}{N} \frac{\partial d^2}{\partial \boldsymbol{\Sigma}} (\boldsymbol{\Phi} - \boldsymbol{\mu} \mathbb{1}^T), \quad \frac{\partial d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)}{\partial \boldsymbol{\Phi}^*} = \frac{2}{N^*} \frac{\partial d^2}{\partial \boldsymbol{\Sigma}^*} (\boldsymbol{\Phi}^* - \boldsymbol{\mu}^* \mathbb{1}^T). \tag{2}$$

Then let $\boldsymbol{Z}$ *be some projection matrix. For* $\boldsymbol{\Phi}' = \boldsymbol{Z}[\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_N]$ *and* $\boldsymbol{\Phi}'^* = \boldsymbol{Z}[\boldsymbol{\phi}_1^*, ..., \boldsymbol{\phi}_{N^*}^*]$ *with covariances* $\boldsymbol{\Sigma}'$*,* $\boldsymbol{\Sigma}'^*$*, means* $\boldsymbol{\mu}'$*,* $\boldsymbol{\mu}'^*$ *and* $d'^2 \equiv d^2(\boldsymbol{\Sigma}', \boldsymbol{\Sigma}'^*)$*, we obtain:*

$$\frac{\partial d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)}{\partial \boldsymbol{\Phi}} = \frac{2\boldsymbol{Z}^T}{N} \frac{\partial d'^2}{\partial \boldsymbol{\Sigma}'} (\boldsymbol{\Phi}' - \boldsymbol{\mu}' \mathbb{1}^T), \quad \frac{\partial d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)}{\partial \boldsymbol{\Phi}^*} = -\frac{2\boldsymbol{Z}^T}{N^*} \frac{\partial d'^2}{\partial \boldsymbol{\Sigma}'^*} (\boldsymbol{\Phi}'^* - \boldsymbol{\mu}'^* \mathbb{1}^T). \tag{3}$$

Proof. See our supplementary material. □

3.4 Non-Euclidean Distances

In Table 1, we list the distances d with derivatives w.r.t. $\boldsymbol{\Sigma}$ used in the sequel. We indicate properties such as invariance to rotation (*rot.*), affine mainpulations (*aff.*) and inversion (*inv.*). We indicate which distances meet the triangle inequality (*Tr. Ineq.*) and which are geodesic distances (*Geo.*). Lastly, we indicate if the distance d and its gradient $\nabla_{\boldsymbol{\Sigma}}$ are finite (*fin.*) or infinite (∞) for $\mathcal{S}_+$ matrices. This last property indicates that JBLD and AIRM distances require some regularization as our covariances are $\mathcal{S}_+$.

4 Problem Formulation

In this section, we equip the supervised domain adaptation approach So-HoT [5] with the JBLD and AIRM distances and the Nyström projections to make evaluations fast.

4.1 Supervised Domain Adaptation

Suppose $\mathcal{I}_N$ and $\mathcal{I}_{N^*}$ are the indexes of N source and N^* target training data points. $\mathcal{I}_{N_c}$ and $\mathcal{I}_{N_c^*}$ are the class-specific indexes for $c \in \mathcal{I}_C$, where C is the number of classes (exhibit identities). Furthermore, suppose we have feature vectors $\boldsymbol{\phi}$ from an *fc* layer of the source network stream, one per image, and their associated labels y. Such pairs are given by $\boldsymbol{\Lambda} \equiv \{(\boldsymbol{\phi}_n, y_n)\}_{n \in \mathcal{I}_N}$, where $\boldsymbol{\phi}_n \in \mathbb{R}^d$ and $y_n \in \mathcal{I}_C$, $\forall n \in \mathcal{I}_N$. For the target data, by analogy, we define pairs $\boldsymbol{\Lambda}^* \equiv \{(\boldsymbol{\phi}_n^*, y_n^*)\}_{n \in \mathcal{I}_N^*}$, where $\boldsymbol{\phi}^* \in \mathbb{R}^d$ and $y_n^* \in \mathcal{I}_C$, $\forall n \in \mathcal{I}_N^*$. Class-specific sets of feature vectors are given as $\boldsymbol{\Phi}_c \equiv \{\boldsymbol{\phi}_n^c\}_{n \in \mathcal{I}_{N_c}}$ and $\boldsymbol{\Phi}_c^* \equiv \{\boldsymbol{\phi}_n^{*c}\}_{n \in \mathcal{I}_{N_c^*}}$, $\forall c \in \mathcal{I}_C$. Then $\boldsymbol{\Phi} \equiv (\boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_C)$ and $\boldsymbol{\Phi}^* \equiv (\boldsymbol{\Phi}_1^*, ..., \boldsymbol{\Phi}_C^*)$. We write the asterisk in superscript (*e.g.* $\boldsymbol{\phi}^*$) to denote variables related to the target network while the source-related variables have no asterisk. Our problem is posed as a trade-off between the classifier and alignment losses ℓ and $\hbar$. Figure 1 shows our setup. Our loss $\hbar$ depends on two sets of variables $(\boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_C)$ and $(\boldsymbol{\Phi}_1^*, ..., \boldsymbol{\Phi}_C^*)$ – one set per network stream. Feature vectors $\boldsymbol{\Phi}(\boldsymbol{\Theta})$ and $\boldsymbol{\Phi}^*(\boldsymbol{\Theta}^*)$ depend on the parameters of the source and target network streams $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^*$ that we optimize over. $\boldsymbol{\Sigma}_c \equiv \boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi}_c))$, $\boldsymbol{\Sigma}_c^* \equiv \boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi}_c^*))$, $\boldsymbol{\mu}_c(\boldsymbol{\Phi})$ and $\boldsymbol{\mu}_c^*(\boldsymbol{\Phi}^*)$ denote the covariances and means, respectively, one covariance/mean pair per network stream per class. Specifically, we solve:

$$\begin{aligned} \underset{\substack{\boldsymbol{W}, \boldsymbol{W}^*, \boldsymbol{\Theta}, \boldsymbol{\Theta}^* \\ \text{s. t. } ||\boldsymbol{\phi}_n||_2^2 \leq \tau, \\ ||\boldsymbol{\phi}_{n'}^*||_2^2 \leq \tau, \\ \forall n \in \mathcal{I}_N, n' \in \mathcal{I}_N^*}}{\arg\min} \quad & \ell(\boldsymbol{W}, \boldsymbol{\Lambda}) + \ell(\boldsymbol{W}^*, \boldsymbol{\Lambda}^*) + \eta ||\boldsymbol{W} - \boldsymbol{W}^*||_F^2 + \\ & \underbrace{\frac{\sigma_1}{C} \sum_{c \in \mathcal{I}_C} d_g^2 \left(\boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_c^*\right) + \frac{\sigma_2}{C} \sum_{c \in \mathcal{I}_C} ||\boldsymbol{\mu}_c - \boldsymbol{\mu}_c^*||_2^2}_{\hbar(\boldsymbol{\Phi}, \boldsymbol{\Phi}^*)}. \end{aligned} \tag{4}$$

Note that Figure 1a indicates by the elliptical/curved shape that $\hbar$ performs the alignment on the $\mathcal{S}_+$ manifold along exact (or approximate) geodesics. For ℓ, we employ a generic Softmax loss. For the source and target streams, the matrices $\boldsymbol{W}, \boldsymbol{W}^* \in \mathbb{R}^{d \times C}$ contain unnormalized probabilities. In Equation (4), separating the class-specific distributions is addressed by ℓ while attracting the within-class scatters of both network streams is handled by $\hbar$. Variable η controls the proximity between $\boldsymbol{W}$ and $\boldsymbol{W}^*$ which encourages the similarity between decision boundaries of classifiers. Coeffs. σ_1, σ_2 control the degree of the cov. and mean alignment, τ controls the ℓ_2-norm of vectors $\boldsymbol{\phi}$.

The Nyström projections are denoted by $\boldsymbol{\Pi}$. Table 1 indicates that backpropagation on the JBLD and AIRM distances involves inversions of $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\Sigma}^*$ for each $c \in \mathcal{I}_C$ according to (4). As $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\Sigma}^*$ are formed from say 2048 dimensional feature vectors of the last *fc* layer, inversions are too costly to run fine-tuning *e.g.*, $4s$ per iteration is prohibitive. Thus, we show next how to combine the Nyström projections with d_g.

Proposition 3. *Let us choose* $\boldsymbol{Z} = \boldsymbol{X} = [\boldsymbol{\Phi}, \boldsymbol{\Phi}^*]$ *for pivots and source/target feature vectors, kernel* k *to be linear, and substitute them into Eq.* (1)*. Then we obtain* $\boldsymbol{\Pi}(\boldsymbol{X}) = (\boldsymbol{Z}^T\boldsymbol{Z})^{-0.5}\boldsymbol{Z}^T\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{X} = (\boldsymbol{Z}^T\boldsymbol{Z})^{0.5} = (\boldsymbol{X}^T\boldsymbol{X})^{0.5}$ *where* $\boldsymbol{\Pi}(\boldsymbol{X})$ *is a projection of* $\boldsymbol{X}$ *on itself that is isometric* e.g.*, distances between column vectors of* $(\boldsymbol{X}^T\boldsymbol{X})^{0.5}$ *correspond to distances of column vectors in* $\boldsymbol{X}$*. Thus,* $\boldsymbol{\Pi}(\boldsymbol{X})$ *is an isometric transformation w.r.t. distances in Table 1, that is* $d_g^2(\boldsymbol{\Sigma}(\boldsymbol{\Phi}), \boldsymbol{\Sigma}(\boldsymbol{\Phi}^*)) = d_g^2(\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi})), \boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi}^*)))$.

Proof. Firstly, we note that the following holds:

$$\boldsymbol{K}_{\boldsymbol{X}\boldsymbol{X}} = \boldsymbol{\Pi}(\boldsymbol{X})^T\boldsymbol{\Pi}(\boldsymbol{X}) = (\boldsymbol{X}^T\boldsymbol{X})^{0.5}(\boldsymbol{X}^T\boldsymbol{X})^{0.5} = \boldsymbol{X}^T\boldsymbol{X}. \tag{5}$$

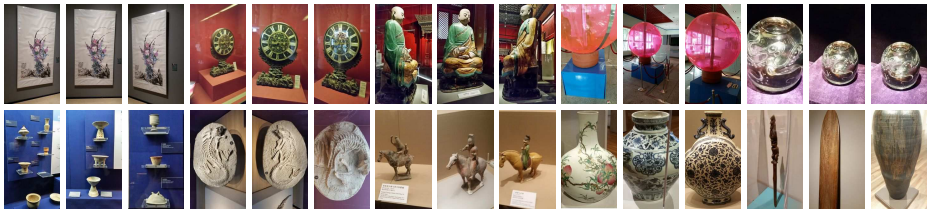

Fig. 2: Source subsets of Open MIC. (Top) Paintings (*Shn*), Clocks (*Clk*), Sculptures (*Scl*), Science Exhibits (*Sci*) and Glasswork (*Gls*). As 3 images per exhibit demonstrate, we covered different viewpoints and scales during capturing. (Bottom) 3 different art pieces per exhibition such as Cultural Relics (*Rel*), Natural History Exhibits (*Nat*), Historical/Cultural Exhibits (*Shx*), Porcelain (*Clv*) and Indigenous Arts (*Hon*). Note the composite scenes of Relics, fine-grained nature of Natural History and Cultural Exhibits and non-planarity of exhibits.

Note that $\boldsymbol{\Pi}(\boldsymbol{X})=\boldsymbol{\mathcal{Z}}\boldsymbol{X}$ projects $\boldsymbol{X}$ into a more compact subspace of size $d'=N+N^*$ if $d' \ll d$ which includes the spanning space for $\boldsymbol{X}$ by construction as $\boldsymbol{Z}=\boldsymbol{X}$. Eq. (5) implies that $\boldsymbol{\Pi}(\boldsymbol{X})$ performs at most rotation on $\boldsymbol{X}$ as the dot-product (used to obtain entries of $\boldsymbol{K}_{\boldsymbol{XX}}$) just like the Euclidean distance is rotation-invariant only *e.g.*, has no affine invariance. As spectra of $(\boldsymbol{X}^T\boldsymbol{X})^{0.5}$ and $\boldsymbol{X}$ are equal, this implies $\boldsymbol{\Pi}(\boldsymbol{X})$ performs no scaling, shear or inverse. Distances in Table 1 are all rotation-invariant, thus $d_g^2(\boldsymbol{\Sigma}(\boldsymbol{\Phi}),\boldsymbol{\Sigma}(\boldsymbol{\Phi}^*))=d_g^2(\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi})),\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi}^*)))$.

A strict proof shows that $\boldsymbol{\mathcal{Z}}$ is a composite rotation $\boldsymbol{VU}^T$ if SVD of $\boldsymbol{Z}=\boldsymbol{U\lambda V}^T$:

$$\boldsymbol{\mathcal{Z}}=(\boldsymbol{Z}^T\boldsymbol{Z})^{-0.5}\boldsymbol{Z}^T=(\boldsymbol{V\lambda U}^T\boldsymbol{U\lambda V}^T)^{-0.5}\,\boldsymbol{V\lambda U}^T=\boldsymbol{V\lambda}^{-1}\boldsymbol{V}^T\boldsymbol{V\lambda U}^T=\boldsymbol{VU}^T. \quad (6)$$

□

In practice, for each class $c \in \mathcal{I}_C$, we choose $\boldsymbol{X}=\boldsymbol{Z}=[\boldsymbol{\Phi}_c,\boldsymbol{\Phi}_c^*]$. Then, as $\boldsymbol{\mathcal{Z}}[\boldsymbol{\Phi},\boldsymbol{\Phi}^*]=(\boldsymbol{X}^T\boldsymbol{X})^{0.5}$, we have $\boldsymbol{\Pi}(\boldsymbol{\Phi})=[\boldsymbol{y}_1,...,\boldsymbol{y}_N]$ and $\boldsymbol{\Pi}(\boldsymbol{\Phi}^*)=[\boldsymbol{y}_{N+1},...,\boldsymbol{y}_{N+N*}]$ where $\boldsymbol{Y}=[\boldsymbol{y}_1,...,\boldsymbol{y}_{N+N*}]=(\boldsymbol{X}^T\boldsymbol{X})^{0.5}$. With typical $N\approx 30$ and $N^*\approx 3$, we obtain covariances of side size $d'\approx 33$ rather than $d=4096$.

Proposition 4. *Typically, the inverse square root $(\boldsymbol{X}^T\boldsymbol{X})^{-0.5}$ of $\boldsymbol{\mathcal{Z}}(\boldsymbol{X})$ can be only differentiated via costly SVD. However, if $\boldsymbol{X}=[\boldsymbol{\Phi},\boldsymbol{\Phi}^*]$, $\boldsymbol{\mathcal{Z}}(\boldsymbol{X})=(\boldsymbol{X}^T\boldsymbol{X})^{-0.5}\boldsymbol{X}^T$ and $\boldsymbol{\Pi}(\boldsymbol{X})=\boldsymbol{\mathcal{Z}}(\boldsymbol{X})\boldsymbol{X}$ as in Prop. 3, and if we consider the chain rule we require:*

$$\frac{\partial d_g^2(\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi})),\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi}^*)))}{\partial\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi}))}\odot\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi}))}{\partial\boldsymbol{\Pi}(\boldsymbol{\Phi})}\odot\frac{\partial\boldsymbol{\Pi}(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}},^{3} \quad (7)$$

then $\boldsymbol{\mathcal{Z}}(\boldsymbol{X})$ can be treated as a constant in differentiation:

$$\frac{\partial\boldsymbol{\Pi}(\boldsymbol{X})}{\partial X_{mn}}=\frac{\partial\boldsymbol{\mathcal{Z}}(\boldsymbol{X})\boldsymbol{X}}{\partial X_{mn}}=\boldsymbol{\mathcal{Z}}(\boldsymbol{X})\frac{\partial\boldsymbol{X}}{\partial X_{mn}}=\boldsymbol{\mathcal{Z}}(\boldsymbol{X})\boldsymbol{J}_{mn}. \quad (8)$$

Proof. It follows from the rotation-invariance of the Euclidean, JBLD and AIRM distances. Let us write $\boldsymbol{\mathcal{Z}}(\boldsymbol{X})=\boldsymbol{R}(\boldsymbol{X})=\boldsymbol{R}$, where $\boldsymbol{R}$ is a rotation matrix. Thus, we have: $d_g^2(\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi})),\boldsymbol{\Sigma}(\boldsymbol{\Pi}(\boldsymbol{\Phi}^*)))=d_g^2(\boldsymbol{\Sigma}(\boldsymbol{R\Phi}),\boldsymbol{\Sigma}(\boldsymbol{R\Phi}^*))=d_g^2(\boldsymbol{R\Sigma}(\boldsymbol{\Phi})\boldsymbol{R}^T,\boldsymbol{R\Sigma}(\boldsymbol{\Phi}^*)\boldsymbol{R}^T)$. Therefore, even if $\boldsymbol{R}$ depends on $\boldsymbol{X}$, the distance d_g^2 is unchanged by any choice of valid $\boldsymbol{R}$ *i.e.*, for the Frobenius norm we have: $||\boldsymbol{R\Sigma R}^T-\boldsymbol{R\Sigma}^*\boldsymbol{R}^T||_F^2=\mathrm{Tr}\left(\boldsymbol{RA}^T\boldsymbol{R}^T\boldsymbol{RAR}^T\right)=$

Fig. 3: Examples of the target subsets of Open MIC. From left to right, each column illustrates Paintings (*Shn*), Clocks (*Clk*), Sculptures (*Scl*), Science Exhibits (*Sci*) and Glasswork (*Gls*), Cultural Relics (*Rel*), Natural History Exhibits (*Nat*), Historical/Cultural Exhibits (*Shx*), Porcelain (*Clv*) and Indigenous Arts (*Hon*). Note the variety of photometric and geometric distortions.

$\text{Tr}\left(\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{A}^T\boldsymbol{A}\right)=\text{Tr}\left(\boldsymbol{A}^T\boldsymbol{A}\right)=||\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^*||_F^2$, where $\boldsymbol{A}=\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^*$. Therefore, we obtain: $\frac{\partial||\boldsymbol{R}\boldsymbol{\Sigma}(\boldsymbol{\Phi})\boldsymbol{R}^T-\boldsymbol{R}\boldsymbol{\Sigma}(\boldsymbol{\Phi}^*)\boldsymbol{R}^T||_F^2}{\partial\boldsymbol{R}\boldsymbol{\Sigma}(\boldsymbol{\Phi})\boldsymbol{R}^T}\odot\frac{\partial\boldsymbol{R}\boldsymbol{\Sigma}(\boldsymbol{\Phi})\boldsymbol{R}^T}{\partial\boldsymbol{\Sigma}(\boldsymbol{\Phi})}\odot\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}}=\frac{\partial||\boldsymbol{\Sigma}(\boldsymbol{\Phi})-\boldsymbol{\Sigma}(\boldsymbol{\Phi}^*)||_F^2}{\partial\boldsymbol{\Sigma}(\boldsymbol{\Phi})}\odot\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}}$ [3] which completes the proof. □

Complexity. The Frobenius norm between covariances plus their computation have combined complexity $\mathcal{O}((d'+1)d^2)$, where $d'=N+N^*$. For non-Euclidean distances, we take into account the dominant cost of evaluating the square root of matrix and/or inversions by SVD, as well as the cost of building scatter matrices. Thus, we have $\mathcal{O}((d'+1)d^2+d^\omega)$, where constant $2<\omega<2.376$ concerns complexity of SVD. Lastly, evaluating the Nyström projections, building covariances and running a non-Euclidean distance enjoys $\mathcal{O}(d'^2d+(d'+1)d'^2+d'^\omega)=\mathcal{O}(d'^2d)$ complexity for $d\gg d'$.

For typical $d'=33$ and $d=2048$, the non-Euclidean distances are $1.7\times$ slower[4] than the Frobenius norm. However, non-Eucldiean distances combined with our projections are $210\times$ and $124\times$ faster than naively evaluated non-Eucldiean distances and the Frobenius norm. This cuts the time of each training from a couple of days to 6–8 hours. Moreover, while unsupervised methods such as CORAL [8] align only two covariances (source and target), our most demanding supervised protocol operates on 866 classes which requires aligning 2×866 covariances. For naive alignment via JBLD, we need 6 days (or much more[4]) to complete. With Nyström projections, JBLD takes $\sim$70 hours.

5 Experiments

Below we detail our CNN setup, discuss the Open MIC dataset and our evaluations.

Setting. At the training and testing time, we use the setting shown in Figures 1a and 1c, respectively. The images in our dataset are portrait or landscape oriented. Thus, we extract 3 square patches per image that cover its entire region. For training, these patches are training data points. For testing, we average over 3 predictions from a group of patches to label image. We briefly compare VGG16 [14] and GoogLeNet [40], and the Eucldiean, JBLD and AIRM distances on subsets of Office and Open MIC. Table 3

[3] For simplicity of notation, $\odot$ denotes the summation over multiplications in chain rules.

[4] For CPU as SVD of large matrices ($d\geq2048$) in CUDA BLAS is close to intractable.

shows that VGG16 and GoogLeNet yield similar scores while JBLD and AIRM beat the Euclidean distance. Thus, we employ VGG16 with JBLD in what follows.

Parameters. Both streams are pre-trained on ImageNet [3]. We set non-zero learning rates on the fully-connected and the last two convolutional layers of each stream. Fine-tuning of both streams takes 30–100K iterations. We set τ to the average value of the ℓ_2 norm of *fc* feature vectors sampled on ImageNet and the hyperplane proximity $\eta = 1$. Inverse in $\mathcal{Z}(\boldsymbol{X}) = (\boldsymbol{X}^T\boldsymbol{X})^{-0.5}\boldsymbol{X}^T$ and matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^*$ are regularized by $\sim$1e-6 on diagonals. Lastly, we set σ_1 and σ_2 between 0.005–1 to perform cross-validation.

Office. It has DSLR, Amazon and Webcam domains. For brevity, we check if our pipeline matches results in the literature on the Amazon-Webcam domain shift ($\mathcal{A}\rightarrow\mathcal{W}$).

Open MIC. The proposed dataset contains 10 distinct source-target subsets of images from 10 different kinds of museum exhibition spaces which are illustrated in Figures 2 and 3, resp.; see also [41]. They include Paintings from Shenzhen Museum (*Shn*), the Clock and Watch Gallery (*Clk*) and the Indian and Chinese Sculptures (*Scl*) from the Palace Museum, the Xiangyang Science Museum (*Sci*), the European Glass Art (*Gls*) and the Collection of Cultural Relics (*Rel*) from the Hubei Provincial Museum, the Nature, Animals and Plants in Ancient Times (*Nat*) from Shanghai Natural History Museum, the Comprehensive Historical and Cultural Exhibits from Shaanxi History Museum (*Shx*), the Sculptures, Pottery and Bronze Figurines from the Cleveland Museum of Arts (*Clv*), and Indigenous Arts from Honolulu Museum Of Arts (*Hon*).

For the target data, we annotated each image with labels of art pieces visible in it. The wearable cameras were set to capture an image every 10s and operated *in-the-wild e.g.*, volunteers had no control over shutter, focus, centering. Thus, our data exhibits many realistic challenges *e.g.*, sensor noises, motion blur, occlusions, background clutter, varying viewpoints, scale changes, rotations, glares, transparency, non-planar surfaces, clipping, multiple exhibits, active light, color inconstancy, very large or small

	Shn	*Clk*	*Scl*	*Sci*	*Gls*	*Rel*	*Nat*	*Shx*	*Clv*	*Hon*	Total
Inst.	79	113	41	37	98	100	111	166	81	40	866
Src+	566	413	225	637	601	775	763	2928	531	1121	8560
Src.	417	650	160	391	575	587	695	2697	503	970	7645
Tgt+	515	323	130	1692	964	1229	868	776	682	417	7596
Tgt.	404	305	112	1342	863	863	668	546 +307K fr	625	364 +73K fr	6092 +380K fr

Table 2: Unique exhibit instances (*Inst.*) and numbers of images of Open MIC in the source (*Src.*) and target (*Tgt.*) subsets plus backgrounds (*Src+*) and (*Tgt+*). We also have $\sim$380K frames (*fr*).

	Alex Net	VGG16	GoogLe Net
S+T	82.4	88.66	88.92
So	84.5	89.45	89.70
JBLD	**85.6**	**90.80**	**91.33**
AIRM	85.2	90.72	91.20

DLID	[6]	51.9
DeCAF$_6$ S+T	[38]	80.7
DaNN	[39]	53.6
Source CNN	[7]	56.5
Target CNN	[7]	80.5
Source+Target CNN	[7]	82.5
Dom. Conf.+Soft Labs.	[7]	82.7

	So	*JBLD*	*AIRM*
sp1	55.8	**57.7**	57.2
sp2	58.9	**58.9**	58.9
sp3	69.6	**71.4**	71.4
sp4	53.8	**57.7**	57.7
sp5	58.3	**60.4**	60.4
acc.	59.3	**61.2**	61.1

Table 3: Verification of baseline setups. (Left) Office ($\mathcal{A}\rightarrow\mathcal{D}$ domain shift) on AlexNet, VGG16 and GoogLeNet streams. We compare baseline fine-tuning on the combined source+target domains (*S+T*), second-order (*So*) Euclidean-based method [5] and our JBLD/AIRM dist. (Middle) State of the art. (Right) Open MIC on (*Clk*) domain shift and VGG16.

exhibits, to name but a few phenomena visible in Figure 3. The numbers and statistics regarding the Open MIC dataset are given in Table 2. Every subset contains 37–166 exhibits to identify and 5 train, val., and test splits. In total, our dataset contains 866 unique exhibit labels, 8560 source (7645 exhibits and 915 backgrounds) and 7596 target (6092 exhibits and 1504 backgrounds including a few of unidentified exhibits) images.

Baselines. We provide baselines such as (i) fine-tuning CNNs on the source subsets (*S*) and testing on the randomly chosen target splits, (ii) fine tuning on target only (*T*) and evaluating on remaining disjoint target splits, (iii) fine-tuning on the source+target (*S+T*) and evaluating on remaining disjoint target splits, (iv) training state-of-the-art domain adaptation So-HoT algorithm [5] equipped by us with non-Euclidean distances.

We include evaluation protocols: (i) training/eval. per exhibition subset, (ii) training/testing on the combined set with all 866 identity labels, (iii) testing w.r.t. scene factors annotated by us (Section 5.2, Challenge III), (iv) unsupervised domain adapt.

5.1 Comparison to the State of the Art

Firstly, we validate that our reference method performs on the par or better than the state-of-the-art approaches. Table 3 shows that the JBLD and AIRM distances outperform the Euclidean-based So-HoT method (*So*) [5] by $\sim 1.6\%$ ($\mathcal{A}\rightarrow\mathcal{D}$, Office, VGG16), 0.9% (*Clk*, Open MIC, VGG16) and recent approaches *e.g.*, [7] by $\sim 2.9\%$ accuracy ($\mathcal{A}\rightarrow\mathcal{D}$, Office, AlexNet). We also observe that GoogLeNet outperforms the VGG16-based model by $\sim 0.5\%$. Having validated our model, we opt to evaluate our proposed Open MIC dataset on VGG16 streams for consistency with the So-HoT model [5].

Supervised vs. unsupervised domain adaptation. The goal of the supervised domain adaptation is to use few source and target training samples per class, all labeled, to mimic human abilities of learning from very few samples. In contrast, the unsupervised

	Shn S	T	S+T	JBLD	*Clk* S	T	S+T	JBLD	*Scl* S	T	S+T	JBLD	*Sci* S	T	S+T	JBLD	*Gls* S	T	S+T	JBLD
sp1	45.3	45.3	59.0	**60.0**	55.8	51.9	55.8	**57.7**	56.5	60.9	65.2	**65.2**	59.3	58.9	65.6	**65.8**	17.6	69.7	73.1	**78.6**
sp2	48.4	52.6	53.7	**62.1**	55.4	44.6	50.0	**58.9**	44.4	50.0	44.4	**50.0**	56.9	57.2	67.1	**69.1**	15.2	75.5	76.2	**76.9**
sp3	46.1	52.7	60.4	**64.8**	58.9	58.9	67.9	**71.4**	55.6	38.9	44.4	**44.4**	69.9	62.0	65.7	**68.2**	17.0	69.1	72.2	**74.0**
sp4	49.5	50.5	54.8	**64.5**	51.9	48.1	46.1	**57.7**	55.0	55.0	55.0	**50.0**	58.1	59.2	64.2	**66.3**	12.2	67.6	69.7	**70.0**
sp5	49.5	57.0	63.4	**69.9**	62.5	41.7	60.4	**60.4**	56.2	56.2	62.5	**62.5**	57.3	53.3	61.5	**64.5**	17.1	69.3	71.7	**72.7**
top-1	47.7	51.6	58.3	**64.3**	56.9	49.1	56.0	**61.2**	53.5	52.2	54.3	**54.4**	58.5	58.1	64.9	**66.8**	15.8	70.2	72.6	**74.4**
top-1-5	48.2	54.2	60.2	**66.4**	58.9	56.3	60.3	**68.9**	54.7	55.4	57.3	**58.4**	60.2	61.7	67.8	**70.2**	19.4	85.1	86.0	**89.0**
top-5	64.5	68.8	76.9	**81.6**	76.7	63.8	78.2	86.9	67.4	66.6	70.0	**70.0**	83.3	82.7	86.0	**88.6**	31.1	90.7	91.0	**92.5**
top-5-5	66.0	73.3	79.5	**84.2**	77.8	75.0	82.7	91.0	69.4	69.8	71.1	**72.0**	85.6	86.3	89.4	**91.3**	42.7	94.5	94.8	**95.8**
Avg_k top-k-k	59.0	63.4	71.0	**76.6**	69.4	65.6	73.6	81.2	63.7	62.5	65.1	**65.1**	75.3	76.0	80.7	**82.5**	31.2	87.9	88.9	**90.5**

	Rel S	T	S+T	JBLD	*Nat* S	T	S+T	JBLD	*Shx* S	T	S+T	JBLD	*Clx* S	T	S+T	JBLD	*Hon* S	T	S+T	JBLD
sp1	18.5	65.0	63.3	**66.3**	38.0	56.2	52.6	**58.8**	33.3	43.2	31.5	**58.6**	47.4	65.8	66.2	**71.4**	65.6	71.1	70.3	**75.8**
sp2	16.5	65.7	63.0	**68.0**	39.9	52.5	52.5	**59.6**	31.8	39.8	27.4	**47.8**	47.0	70.2	65.1	**72.2**	63.9	67.2	70.5	**74.6**
sp3	19.1	70.4	67.4	**70.7**	43.7	56.2	59.4	**59.9**	25.7	47.7	31.2	**47.7**	49.7	64.1	61.5	**67.7**	68.5	70.2	71.8	**79.0**
sp4	18.3	68.5	62.8	**67.1**	41.8	59.8	62.0	**67.9**	33.0	38.8	26.2	**44.7**	48.3	63.0	64.0	**68.5**	67.8	63.6	79.3	**76.9**
sp5	18.1	61.0	59.3	**62.6**	44.6	62.0	63.0	**67.4**	25.7	35.8	28.4	**44.0**	42.3	62.8	54.1	**65.8**	67.5	65.8	75.0	**80.0**
top-1	18.1	66.1	63.2	**67.0**	41.6	57.3	57.9	**62.7**	29.9	41.1	29.0	**48.5**	47.0	65.2	62.2	**69.1**	66.7	67.6	73.4	**77.3**
top-1-5	24.0	76.8	73.2	**79.5**	43.5	62.8	61.9	**67.7**	31.5	47.7	31.9	**56.3**	50.8	69.5	66.6	**73.9**	70.2	70.3	76.3	**79.7**
top-5	26.2	87.1	85.8	**90.3**	60.6	79.3	75.5	84.3	51.6	62.5	51.2	**75.0**	65.3	84.3	79.9	**87.7**	82.1	85.2	88.3	**90.0**
top-5-5	28.7	90.0	89.4	**93.7**	65.3	82.8	80.1	87.0	54.9	67.3	54.8	**77.6**	70.5	89.2	84.4	**91.0**	88.1	88.8	91.7	**92.7**
Avg_k top-k-k	25.2	82.8	80.5	**85.2**	55.7	74.0	72.4	**79.6**	45.1	57.1	44.5	**66.8**	61.5	80.6	76.5	**83.5**	79.7	81.0	84.5	**86.7**

Table 4: Challenge I. Open MIC performance on the 10 subsets for data 5 splits. Baselines (*S*), (*T*) and (*S+T*) are given as well as our JBLD approach. We report top-1, top-1-5, top-5-1, top-5-5 accuracies and the combined scores Avg_k top-k-k. See Section5.2 for details.

case can use large numbers of unlabeled target training samples. We ran our code on the Office-Home dataset [27] which has no supervised protocol. We chose $Cl \rightarrow Ar/Pr \rightarrow Ar$ domain shifts, 20 source and 3 target train images per class (all labeled) which yielded 48.1/49.3 (*So*) and **49.2/50.5%** (JBLD) accuracy. Unsupervised approach [27] that used all available target datapoints yielded 34.69/29.91% accuracy.

5.2 Open MIC Challenge

Below we detail our challenges on the Open MIC dataset and present our results.

Challenge I. Below we run our supervised domain adaptation with the JBLD distance per subset. We prepare 5 training, validation and testing splits. For the source data, we use all samples available per class. For the target data, we use 3 samples per class for training and validation, respectively, and the rest for testing.

We report top-1 and top-5 accuracies. Moreover, as our target images often contain multiple exhibits, we ask a question whether any of top-k predictions match any of top-n image labels ordered by our expert volunteers according to the perceived saliency. If so, we count it as a correctly recognized image. We count these valid predictions and normalize by the total number of testing images. We denote this measure as top-k-n where $k, n \in \mathcal{I}_5$. Lastly, we indicate an *area-under-curve* type of measure Avg_k top-k-k which rewards correct recognition of the most dominant object in the scene and offers some reward if the order of top predictions is wrong (less dominant objects pred. first).

We divided Open MIC into *Shn*, *Clk*, *Scl*, *Sci*, *Gls*, *Rel*, *Nat*, *Shx*, *Clv* and *Hon* subsets to allow short 6–8 hours long runs per experiment. We ran 150 jobs on (*S*), (*T*) and (*S*+*T*) baselines and 300 jobs on JBLD: 5 splits $\times$10 subsets $\times$6 hyperp. choices. Table 4 shows that the exhibits in the Comprehensive Historical and Cultural Exhibits (*Shx*) and the Sculptures (*Scl*) were the hardest to identify given 48.5 and 54.4% top-1 accuracy. This is consistent with volunteers' reports that both exhibitions were crowded, the lighting was dim, exhibits were occluded, fine-grained and non-planar. Moreover, training on the source and testing on target baseline (*S*) scored mere 15.8 and 18.1% top-1

	sp1	sp2	sp3	sp4	sp5	top-1	top-1-5	top-5	top-5-5	Avg_k top-k-k
S	33.9	34.2	34.8	34.2	33.8	34.2	36.0	49.2	53.7	46.0
T	56.9	55.9	58.7	56.0	55.2	56.5	64.1	76.5	80.6	72.5
S+T	56.4	55.2	57.1	56.3	54.4	55.9	62.5	75.8	79.2	71.6
So	64.2	62.4	65.0	62.7	60.0	62.8	70.4	84.0	88.5	79.5
JBLD	**65.7**	**63.8**	**65.7**	**63.7**	**62.0**	**64.2**	**72.0**	**85.7**	**88.6**	**80.8**

Table 5: Challenge II. Open MIC performance on the combined set for data 5 splits. Baselines (*S*), (*T*) and (*S*+*T*) are given as well as second-order (*So*) method [5] and our JBLD approach.

	clp	*lgt*	*blr*	*glr*	*bgr*	*ocl*	*rot*	*zom*	*vpc*	*sml*	*shd*	*rfl*	*ok*
S	41.4	17.0	23.8	27.3	40.3	34.5	29.7	52.7	33.4	14.2	10.4	32.3	65.5
T	56.2	38.2	42.6	56.1	57.9	49.6	58.3	60.4	50.3	29.6	59.2	60.7	64.3
S+T	56.6	34.6	39.8	54.9	56.2	48.3	56.7	65.9	48.7	27.3	56.5	59.0	72.6
JBLD	**65.3**	**48.6**	**51.6**	**64.0**	**65.9**	**56.4**	**65.0**	**70.0**	**58.6**	**34.1**	**70.4**	**67.5**	**81.0**

Table 6: Challenge III. Open MIC performance on the combined set w.r.t. 12 factors detailed in Section 5.2. Top-1 accuracies for baselines (*S*), (*T*), (*S*+*T*), and for our JBLD appr. are listed.

accuracy on the Glass (*Gls*) and Relics (*Rel*) due to extreme domain shifts. The easiest to identify were the Sculptures, Pottery and Bronze Figurines (*Clv*) and the Indigenous Arts (*Hon*) as both exhibitions were spacious with good lighting. The average top-1 accuracy across all subsets on JBLD is 64.6%. Averages over baselines (*S*), (*T*) and (*S*+*T*) are 43.9, 57.8, and 59.2% top-1 acc. To account for uncertainty of saliency-based labeling (classifier confusing which exhibit to label), we report our proposed average top-1-5 acc. as 71.0%. Our average combined score Avg_k top-k-k is 79.8%. The results show that Open MIC challenges CNNs due to *in-the-wild* capture with wearable cameras.

Challenge II. Below we evaluate the combined set covering 866 exhibit identities. In this setting, a single experiment runs 80–120 hours. We ran 15 jobs on (*S*), (*T*) and (*S*+*T*) baselines and 60 jobs on (*So*) and JBLD: 2 distances $\times 5$ splits $\times 6$ hyperp. choices. Table 5 shows that our JBLD approach scores 64.2% top-1 accuracy and outperforms baselines (*S*), (*T*) and (*S*+*T*) by 30, 7.7 and 8.3%. Fine-tuning CNNs on the source and testing on target (*S*) is a poor performer due to the large domain shift in Open MIC.

Challenge III. For this challenge, we break down performance on the combined set covering 866 exhibit identities w.r.t. the following 12 factors: object clipping (*clp*), low lighting (*lgt*), blur (*blr*), light glares (*glr*), background clutter (*bgr*), occlusions (*ocl*), in-plane rotations (*rot*), zoom (*zom*), tilted viewpoint (*vpc*), small size/far away (*sml*), object shadows (*shd*), reflections (*rfl*) and the clean view (*ok*). Table 6 shows results averaged over 5 data splits. We note that JBLD outperforms baselines. The factors most affecting the supervised domain adaptation are the small size (*sml*) of exhibits/distant view, low light (*lgt*) and blur (*blr*). The corresponding top-1 accuracies of 34.1, 48.6 and 51.6% are below our average top-1 accuracy of 64.2% listed in Table 5. In contrast, images with shadows (*shd*), zoom (*zom*) and reflections (*rfl*) score 70.4, 70.0 and 67.5% top-1 accuracy (above avg. 64.2%). Our wearable cameras captured also a few of clean shots scoring 81.0% top-1 accuracy. Thus, we claim that domain adaptation methods need to evolve to deal with such adverse factors. Our suppl. material presents further analysis of combined factors. Figure 4 shows hard to recognize instances.

Moreover, Table 7 present results (left) and the image counts (right) w.r.t. pairs of factors co-occurring together. The combination of (*sml*) with (*glr*), (*blr*), (*bgr*), (*lgt*), (*rot*) and (*vpc*) results in 13.5, 21.0, 29.9, 31.2, 32.6 and 33.2% mean top-1 accuracy, respectively. Therefore, these pairs of factors affect the quality of recognition the most.

Challenge IV. For unsupervised domain adaptation algorithms, we use all source data (labeled instances) for training and all target data as unlabeled input. A previously, we extract 3 patches per image and train *Invariant Hilbert Space* (*IHS*) [12], *Uns. Domain Adaptation with Residual Transfer Networks* (*RTN*) [42] and *Joint Adaptation Networks* (*JAN*) [43]. Table 8 shows results on the Open MIC dataset on the 10 subsets. Unsuper-

Fig. 4: Examples of difficult to identify exhibits from the target domain in the Open MIC dataset.

vised (*IHS*), (*RTN*) and (*JAN*) scored on average 48.3, 49.1 and 52.1%. For split (*Gls*), which yielded 26.0, 30.5 and 34.2% top-1 accuracy, an extreme domain shift prevented algorithms from successful adaptation. On (*Sci*), unsupervised (*IHS*), (*RTN*) and (*JAN*) scored 63.3, 62.2 and 69.8%. On (*Hon*), they scored 67.3, 71.1 and 72.5%. For simple domain shifts, unsupervised domain adaptation yields visible improvements. For harder domain shifts, supervised JBLD from Table 4 works much better. Lastly, for (*Hon*) and (*Shx*) splits and (*JAN*), we added 4.3K and 13K unlabeled target frames (1 photo/s) and got 74.0% and 32.6% accuracy–this is a 1.5 and 0.6% increase over the low number of target images – adding many unsupervised images has only a small positive impact.

6 Conclusions

We have collected, annotated and evaluated a new challenging Open MIC dataset with the source and target domains formed by images from Android and wearable cameras, respectively. We covered 10 distinct exhibition spaces in 10 different museums to collect a realistic *in-the-wild* target data in contrast to typical photos for which the users control the shutter. We have provided a number of useful baselines *e.g.*, breakdowns of results per exhibition, combined scores and analysis of factors detrimental to domain adaptation and recognition. Unsupervised domain adaptation and few-shot learning methods can also be compared to our baselines. Moreover, we proposed orthogonal improvements to the supervised domain adaptation *e.g.*, we integrated non-trivial non-Euclidean distances and Nyström projections for better results and tractability. We will make our data and evaluation scripts available to the researchers.

Acknowledgement. Big thanks go to Ondrej Hlinka and (Tim) Ka Ho from the Scientific Computing Services at CSIRO for their can-do attitude and help with Bracewell.

∩	*clp*	*lgt*	*blr*	*glr*	*bgr*	*ocl*	*rot*	*zom*	*vpc*	*sml*	*shd*	*rfl*
all	65.3	48.6	51.6	64.0	65.9	56.4	65.0	70.0	58.6	34.1	70.4	67.5
clp	65.3	55.1	51.8	67.5	66.8	61.5	67.2	68.1	62.3	45.5	72.7	67.0
lgt	55.1	48.6	41.0	43.6	59.8	43.5	48.3	44.4	46.1	31.2	57.9	80.9
blr	51.8	41.0	51.6	48.7	48.6	37.0	52.3	64.2	43.3	21.0	39.1	59.4
glr	67.5	43.6	48.7	64.0	62.3	47.9	65.1	67.1	60.4	13.5	50.0	64.5
bgr	66.8	59.8	48.6	62.3	65.9	59.6	66.6	76.1	61.2	29.9	79.6	73.2
ocl	61.5	43.5	37.0	47.9	59.6	56.4	55.6	75.4	55.9	40.7	78.8	64.8
rot	67.2	48.3	52.3	65.1	66.6	55.6	65.0	75.5	57.6	32.6	73.4	70.4
zom	68.1	44.4	64.2	67.1	76.1	75.4	75.5	70.0	66.3	n/a	83.3	69.7
vpc	62.3	46.1	43.3	60.4	61.2	55.9	57.6	66.3	58.6	33.2	64.1	61.6
sml	45.5	31.2	21.0	13.5	29.9	40.7	32.6	n/a	33.2	34.1	n/a	46.4
shd	72.7	57.9	39.1	50.0	79.6	78.8	73.4	83.3	64.1	n/a	70.4	80.0
rfl	67.0	80.9	59.4	64.5	73.2	64.8	70.4	69.7	61.6	46.4	80.0	67.5

∩	*clp*	*lgt*	*blr*	*glr*	*bgr*	*ocl*	*rot*	*zom*	*vpc*	*sml*	*shd*	*rfl*
all	5136	335	1728	1346	2290	1529	7344	2278	4571	557	125	2000
clp	5136	216	770	572	1415	873	3401	1803	2549	167	66	1009
lgt	216	335	105	55	92	69	232	9	234	16	38	21
blr	770	105	1728	240	323	235	1348	240	820	152	23	330
glr	572	55	240	1346	183	143	1054	204	640	52	12	155
bgr	1415	92	323	183	2290	565	1604	464	1409	227	49	395
ocl	873	69	235	143	565	1529	1090	183	978	253	33	219
rot	3401	232	1348	1054	1604	1090	7344	1380	3292	405	113	1522
zom	1803	9	240	204	464	183	1380	2278	611	0	18	535
vpc	2549	234	820	640	1409	978	3292	611	4571	370	39	856
sml	167	16	152	52	227	253	405	0	370	557	0	69
shd	66	38	23	12	49	33	113	18	39	0	125	15
rfl	1009	21	330	155	395	219	1522	535	856	69	15	2000

Table 7: Challenge III. (Left) Open MIC performance on the combined set w.r.t. the pairs of 12 factors detailed in Section 5.2. Top-1 accuracies for our JBLD approach are listed. The top row shows results w.r.t. the original 12 factors. Color-coded cells are normalized w.r.t. entries of this row. For each column, intense/pale red indicates better/worse results compared to the top cell, respectively. (Right) Target image counts for pairs of factors.

	Shn	*Clk*	*Scl*	*Sci*	*Gls*	*Rel*	*Nat*	*Shx*	*Clv*	*Hon*	top-1
IHS	47.1	61.9	50.8	63.3	26.0	32.6	51.0	22.0	61.2	67.3	48.3
RTN	54.4	59.0	65.2	62.2	30.5	24.8	44.2	32.1	47.7	71.1	49.1
JAN	51.7	63.6	67.8	69.8	34.2	28.5	47.1	32.0	53.9	72.5	**52.1**

Table 8: Unsupervised domain adaptation: Open MIC performance on the 10 subsets.

Bibliography

[1] Baxter, J., Caruana, R., Mitchell, T., Pratt, L.Y., Silver, D.L., Thrun, S.: Learning to learn: Knowledge consolidation and transfer in inductive systems. NIPS Workshop, http://plato.acadiau.ca/courses/comp/dsilver/NIPS95_LTL/transfer.workshop.1995.html (1995) Accessed: 30-10-2016. 1

[2] Li, W., Tommasi, T., Orabona, F., Vázquez, D., López, M., Xu, J., Larochelle, H.: Task-cv: Transferring and adapting source knowledge in computer vision. ECCV Workshop, http://adas.cvc.uab.es/task-cv2016 (2016) Accessed: 22-11-2016. 1

[3] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. IJCV **115**(3) (2015) 211–252 1, 2, 10

[4] Tommasi, T., Orabona, F., Caputo, B.: Safety in numbers: Learning categories from few examples with multi model knowledge transfer. CVPR (2010) 3081–3088 1, 3

[5] Koniusz, P., Tas, Y., Porikli, F.: Domain adaptation by mixture of alignments of second- or higher-order scatter tensors. CVPR **2** (2017) 1, 2, 3, 4, 5, 6, 10, 11, 12

[6] Chopra, S., Balakrishnan, S., Gopalan, R.: Dlid: Deep learning for domain adaptation by interpolating between domains. ICML Workshop (2013) 1, 4, 10

[7] Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. ICCV (2015) 4068–4076 1, 2, 4, 10, 11

[8] Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. CoRR **abs/1511.05547** (2015) 1, 4, 5, 9

[9] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(1) (2016) 2096–2030 1, 4

[10] Daumé, III, H., Kumar, A., Saha, A.: Frustratingly easy semi-supervised domain adaptation. Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (2010) 53–59 1

[11] L. Fei-Fei; Fergus, R.P.: One-shot learning of object categories. TPAMI **28** (April 2006) 594–611 1

[12] Herath, S., Harandi, M., Porikli, F.: Learning an invariant hilbert space for domain adaptation. CVPR (2017) 2, 4, 13

[13] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. NIPS (2012) 1106–1114 2

[14] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR **abs/1409.1556** (2015) 2, 3, 9

[15] Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. ECCV (2010) 213–226 2, 3

[16] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR (2014) 580–587 2

[17] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., Lecun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. ICLR (2014) 2

[18] Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. CVPR (2012) 2066–2073 2, 4

[19] Cherian, A., Sra, S., Banerjee, A., Papanikolopoulos, N.: Jensen-Bregman LogDet Divergence with Application to Efficient Similarity Search for Covariance Matrices. TPAMI **35**(9) (2013) 2161–2174 2, 3, 5

[20] Pennec, X., Fillard, P., Ayache, N.: A Riemannian Framework for Tensor Computing. IJCV **66**(1) (2006) 41–66 2, 3, 5

[21] Bhatia, R.: Positive definite matrices. Princeton Univ Press (2007) 3, 5

[22] Wang, Y.X., Hebert, M.: Learning to learn: Model regression networks for easy small sample learning. ECCV (2016) 4

[23] Kuzborskij, I., Carlucci, F.M., Caputo, B.: When naïve bayes nearest neighbors meet convolutional neural networks. CVPR (2016) 4

[24] Tommasi, T., Lanzi, M., Russo, P., Caputo, B.: Learning the roots of visual domain shift. ECCV Workshop (2016) 4, 5

[25] Tommasi, T., Tuytelaars, T., Caputo, B.: A testbed for cross-dataset analysis. Technical Report (2014) 4

[26] Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. Part of the PASCAL in Detail Workshop Challenge, http://www.robots.ox.ac.uk/~vgg/decathlon/ (2017) Accessed: 30-10-2017. 4

[27] Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. CVPR (2017) 4, 12

[28] Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Fei-Fei, L.: Fine-grained car detection for visual census estimation. AAAI (2017) 4

[29] Timnit Gebru, Judy Hoffman, L.F.F.: Fine-grained recognition in the wild: A multi-task domain adaptation approach. ICCV (2017) 4

[30] Rajapakse, J.C., Wang, L.: Neural Information Processing: Research and Development. Springer-Verlag Berlin and Heidelberg GmbH & Co. KG (2004) 4

[31] Koniusz, P., Yan, F., Gosselin, P.H., Mikolajczyk, K.: Higher-order occurrence pooling for bags-of-words: Visual concept detection. TPAMI **39**(2) (2017) 313–326 5

[32] Koniusz, P., Zhang, H., Porikli, F.: A deeper look at power normalizations. CVPR (2018) 5774–5783 5

[33] Yeh, Y.R., Huang, C.H., Wang, Y.C.F.: Heterogeneous domain adaptation and classification by exploiting the correlation subspace. Transactions on Image Processing **23**(5) (2014) 5

[34] Kumar Roy, S., Mhammedi, Z., Harandi, M.: Geometry aware constrained optimization techniques for deep learning. CVPR (June 2018) 5

[35] Harandi, M., Salzmann, M., Hartley, R.: Joint dimensionality reduction and metric learning: A geometric take. ICML **70** (2017) 1404–1413 5

[36] Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks (2011) 5

[37] Bo, L., Sminchisescu, C.: Efficient match kernels between sets of features for visual recognition. NIPS (2009) 6
[38] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. ICML (2014) 10
[39] Ghifary, M., Kleijn, W.B., Zhang., M.: Domain adaptive neural networks for object recognition. CoRR **abs/1409.6041** (2014) 10
[40] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. CVPR (2015) 9
[41] Zhang, R., Tas, Y., Koniusz, P.: Artwork identification from wearable camera images for enhancing experience of museum audiences. Museums and the Web (2017) 10
[42] Long, M., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. CoRR **abs/1602.04433** (2016) 13
[43] Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. ICML (2017) 13