Learning to Cluster under Domain Shift

Willi Menapace[1], Stphane Lathuilire[3], and Elisa Ricci[1,2]

[1] University of Trento, Trento, Italy
[2] Fondazione Bruno Kessler, Trento, Italy
[3] LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France
willi.menapace@gmail.com

Abstract. While unsupervised domain adaptation methods based on deep architectures have achieved remarkable success in many computer vision tasks, they rely on a strong assumption, i.e. labeled source data must be available. In this work we overcome this assumption and we address the problem of transferring knowledge from a source to a target domain when both source and target data have no annotations. Inspired by recent works on deep clustering, our approach leverages information from data gathered from multiple source domains to build a domain-agnostic clustering model which is then refined at inference time when target data become available. Specifically, at training time we propose to optimize a novel information-theoretic loss which, coupled with domain-alignment layers, ensures that our model learns to correctly discover semantic labels while discarding domain-specific features. Importantly, our architecture design ensures that at inference time the resulting source model can be effectively adapted to the target domain without having access to source data, thanks to feature alignment and self-supervision. We evaluate the proposed approach in a variety of settings*, considering several domain adaptation benchmarks and we show that our method is able to automatically discover relevant semantic information even in presence of few target samples and yields state-of-the-art results on multiple domain adaptation benchmarks.

Keywords: Unsupervised learning, domain adaptation, deep clustering

1 Introduction

The astonishing performance of deep learning models in a large variety of applications must be partially ascribed to the availability of large-scale datasets with abundant annotations. Over the years, several solutions have been proposed to avoid prohibitively expensive and time-consuming data labeling such as transfer learning [25] or domain adaptation [6] strategies. In particular, unsupervised domain adaptation (UDA) methods [17, 18, 22, 26, 2, 16, 28, 10, 41] leverage the knowledge extracted from labeled data of one (or multiple) source domain(s) to learn a prediction model for a different but related target domain where no labeled data are available. This strategy is illustrated in Fig.1-left.

*Code available at github.com/willi-menapace/acids-clustering-domain-shift

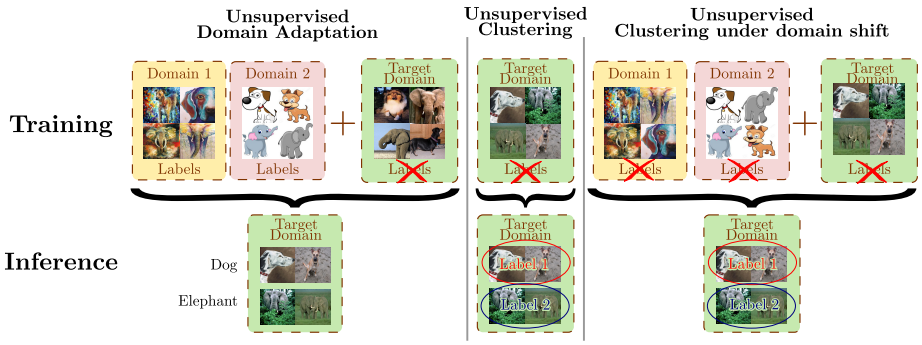


Fig. 1: In the Unsupervised Domain adaptation (UDA) setting, a model is trained combining labeled images from one or several source domains and unlabeled images from the target domain. In the unsupervised clustering setting, unlabeled images from the same domain are grouped into visually similar images. We introduce the Unsupervised Clustering under Domain Shift (UCDS) setting where we leverage unlabeled source domain data to improve target domain clustering.

Over the last decade, increasing efforts have been devoted to develop deep architectures for UDA and promising results have been obtained in several applications such as object recognition [17, 34, 2], semantic segmentation [10], depth estimation [42], etc. While effective in many tasks, current UDA methods rely on a key assumption: annotations associated with data from the source domain(s) must be available. In this paper, we argue that this assumption may hinder the use of UDA in many practical applications. For instance, relaxing the constraints of disposing of labeled source data can broaden the applicability of knowledge transfer methods to tasks and scenarios where gathering annotations is challenging or even impossible (e.g. medical).

A possible alternative to supervised training is unsupervised clustering (Fig.1-center). Clustering is a class of unsupervised learning methods that are designed to group images in such a way that images in the same group contain similar content. Recently, some works [4, 9, 12] have shown that appropriately designed deep architectures can be successfully used to discover clusters in a training set and perform representation learning. By opposition to UDA, clustering does not require any annotation. However, it relies on the assumption that all the data belongs to the same domain. If this condition is not fulfilled, clustering algorithms would tend to group data according to the visual style associated to their domain and not according to their semantic content.

Motivated by these observations, in this paper, we propose a new setting, Unsupervised Clustering under Domain Shift (UCDS), (see Fig.1-right) where we assume that we dispose of data from different known domains but no class labels are available in both source and target domains. Our approach develops under the assumption that, while no annotations from source data are available, still we may benefit from the access to multiple datasets, i.e. to multiple source

domains. This is very reasonable as in many practical applications it is very likely to dispose of several datasets collected under different conditions.

Our method develops from the intuition that, by combining multiple domains with different visual styles, we can obtain clusters based on the semantic content rather than on stylistic or texture features. Importantly, by leveraging multiple source domains, we show that the target domain can be clustered accurately even when target data is limited. Our method is organized in two steps. First, a novel multi-domain deep clustering model is learned which, by seamlessly combining domain-specific distribution alignment layers [2] and an information-theoretic loss permits to discover semantic categories across domains. In a subsequent step, target data are exploited to refine the learned clustering model by simultaneously matching source features distributions with domain-alignment layers and by maximizing the mutual information between the class assignments of pairs of perturbed samples. Recalling these elements, we name our algorithm ACIDS: Adaptive Clustering of Images under Domain Shift.

The major advantage of our two-stage pipeline is that it does not require source and target data to be available simultaneously. Consequently, our setting differs from classical UDA and unsupervised transfer learning scenarios [6, 25] since only the source model is provided to the unlabeled target domain. Discarding the source data at adaptation time can broaden the applicability of our framework to tasks and scenarios that suffer from transmission or privacy issues. Our extensive experimental evaluation demonstrates that our approach successfully discovers semantic categories and outperforms state of the art unsupervised learning models on popular domain adaptation benchmarks: Office-31 [29], PACS [14] and Office-Home [37] dataset.

Contributions. To summarize, the main contributions of this work are: (i) We introduce a new setting, Unsupervised Clustering under Domain Shift (Fig.1-right), where we learn a semantic predictor from unsupervised target samples leveraging from multiple unlabeled source domains; (ii) We propose an information-theoretic algorithm for unsupervised clustering that operates under domain shift. Our method successfully integrates the data-augmentation strategy typically used by deep clustering methods [12] within a feature alignment process; (iii) We evaluate our method on several domain adaptation benchmarks demonstrating that our approach can successfully discover semantic categories even in the presence of domain shift and with few target samples.

2 Related works

In the following we review previous approaches on UDA, discussing both single source and multi-source methods. Since we propose a deep architecture for unsupervised learning under domain shift, we also review related work on deep clustering.

Domain adaptation. Earlier UDA methods assume that only a single source domain is available for transferring knowledge. These methods can be roughly

categorized into three main groups. The first category includes methods which align source and target data distributions by matching the distribution statistical moments of different orders. For instance, Maximum Mean Discrepancy, *i.e.* the distance between the mean of domain feature distributions, is considered in [17, 18, 37, 36], while second order statistics are used [33, 22, 26]. Domain alignment layers derived from batch normalization (BN) [11] or whitening transforms [31] are employed in [2, 16, 21, 28].

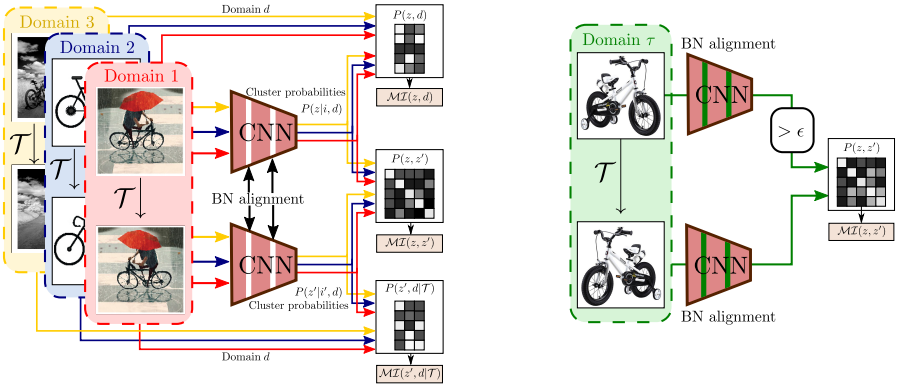
The methods in the second category learn domain-invariant representations considering an adversarial framework. For instance, in [7] a gradient reversal layer is used to learn domain-agnostic representations. Similarly, ADDA [35] introduces a domain confusion loss to align the source and the target domain feature distributions. The third category of methods consider a generative framework (i.e., GANs ([8]) to create synthetic source and/or target images. Notable works are CyCADA [10], I2I Adapt [23] and Generate To Adapt (GTA) [30]. Our method is related to previous works in the first category, as we also leverage domain-alignment layers to perform adaptation. However, we consider a radically different setting where no annotation is provided in the source domain and only the source model (and not the source data) is exploited at adaptation time.

While most previous works on UDA consider a single source domain, recently some works have shown that performance can be considerably improved by leveraging multiple datasets. For instance, in [21] multiple latent source domains are discovered and used for transferring knowledge. Recently, Deep Cocktail Network (DCTN) [40] introduce a distribution-weighted rule for classification which is combined with an adversarial loss. M^3SDA is described in [27]: it reduces the discrepancy between the multiple source and the target domains by dynamically aligning moments of their feature distributions.

Differently from these methods, ACIDS does not assume annotations in the source domain. One related work to ours is [20] where information from multiple source domains is exploited for constructing a domain-dependency graph and then used when the target data are made available. However, in [20] an entropy loss for target model adaptation is considered, which we experimentally observe is less effective than our proposal self-supervised loss. Our method is also related to recent domain generalization (DG) methods [1, 15]. In fact, similarly to DG, we also assume that source and target data are not simultaneously available. However, differently from DG, we make use of target data for model adaptation when they are available.

Deep Clustering. Over the last few years, unsupervised representation learning has attracted considerable attention in the computer vision community. Self-supervised learning approaches mostly differ in the self-supervised losses used to learn feature representations. Notable examples are methods which derive indirect auxiliary supervision from spatial-temporal consistency [38], from solving jigsaw puzzles [24] or from colorization [13].

Recently, some studies have attempted to derive deep clustering algorithms which simultaneously discover groups in training data and perform representation learning. For instance, DEC [39] makes use of an autoencoder to produce



(a) Training on Sources (b) Adaptation on Target

Fig. 2: Illustration of the ACIDS framework for UCDS. In the training stage (Fig.2a), images are clustered by maximizing mutual information between the predictions from the original and transformed images. Domain alignment is addressed combining a Batch Normalization (BN) alignment technique with a novel mutual information minimizing formulation. In the adaptation phase (Fig.2b), domain shift is handled by combining BN alignment with a specific mutual information maximization procedure.

a latent space where cluster centroids are learned. DAC [5] casts the clustering problem into pairwise-classification using a convolutional network to learn feature representations. In [4] DeepCluster, an iterative clustering procedure is devised which adopts k-means to learn representations and uses the subsequent assignments as supervision. Similarly, in [9] an end-to-end clustering approach is proposed where an encoder network is trained with an alternate scheme. Recently, Ji *et al.* propose Invariant Information Clustering (IIC) [12], where a deep network is learned with an information-theoretic criterion in order to output semantic labels, rather than high dimensional representations. Our approach is inspired by this method. However, we specifically address the problem of transferring knowledge from source to target domains.

3 Proposed Method

In this section, we introduce the proposed ACIDS fully unsupervised multi-source domain adaptation framework. The design of the source training framework is guided by two motivations. First, we need to take advantage of the different domains in order to obtain clusters that correspond to semantic labels rather than domain-specific image styles. Second, the network must learn image representations that can be transferred to any unknown target domain.

We assume to observe S source domains I_s composed of images, each depicting an object from C different object categories. In this work, we assume

that C is known a priori but we do not dispose of image labels.We propose to learn image representations allowing to cluster source images according to the unknown category labels. Our goal is to adapt the representation learned on the source domains in order to predict labels on a target domain I_t. In this adaptation stage, we consider that we do not dispose of the images from $I_1, \ldots, I_S$. To this aim, we employ a deep neural network $\phi_\theta : I \to Z$ with parameters θ that predicts cluster assignments probabilities. To obtain network outputs $Z \in [0,1]^C$ that can be interpreted as probability vectors, ϕ_θ is terminated by a layer with a softmax activation function.

In Sec.3.1 we describe the objective used to cluster source images. In order to ensure clustering based on semantic labels and not on domain-specific styles, we introduce in Sec.3.2 a novel information-theoretic alignment mechanism based on the minimization of mutual information between domains and cluster assignments. Here, we also detail our batch normalization alignment layers that complement the framework. Adaptation on the target domain is described in Sec.3.3.

3.1 Multi-domain Clustering with Mutual information

Let $i \in I = \bigcup_{s=1}^{S} I_s$ be an image from the domain $d \in 1..S$. Both i and d are treated here as random variables. We consider that we dispose of a set of image transformations $\mathcal{T}$. After sampling a transformation $t \in \mathcal{T}$, we obtain a transformed version of the image i denoted as i'. Following the approach of [12], we train ϕ_θ in such a way that, first, it returns the same output for both i and i' and, second, it returns different outputs for different images. This double objective can be achieved by maximizing the mutual information between the predictions from i and i' with respect to the network parameters:

$$\max_{\theta} \mathcal{MI}(z, z') \tag{1}$$

where $z = \phi_\theta(x)$ and $z' = \phi_\theta(x')$ are the network cluster assignment predictions. To estimate the mutual information $\mathcal{MI}(z, z')$, we need to compute the joint probability of the cluster assignment $P_{cc'} = P(z{=}c, z'{=}c')$ where $c \in 1..C$ and $c' \in 1..C$ are all the possible cluster indexes. This probability is estimated by marginalization over the current batch. Let us assume to observe a batch composed of N unlabeled images $\{i_s^n\}_{n=1}^N \subset I_s$ from each of S source domains. We have:

$$\begin{aligned} P_{cc'} &= P(z{=}c, z'{=}c') = \sum_{s=1}^{S} P(d{=}s) P(z{=}c, z'{=}c' | d{=}s) \\ &= \frac{1}{SN} \sum_{s=1}^{S} \sum_{n=1}^{N} \phi_\theta(i_s^n) . \phi_\theta(i_s^{n\prime})^\top \end{aligned} \tag{2}$$

Similarly, we estimate the marginal distribution:

$$P_c = P(z{=}c) = \frac{1}{S}\sum_{s=1,n=1}^{S\times N}\phi_\theta(i_s^n) \text{ and } P'_c = P(z{=}c') = \frac{1}{S}\sum_{s=1,n=1}^{S\times N}\phi_\theta(i_s^{n\prime}).$$

From these probability distributions, the mutual information loss is given by:

$$\mathcal{MI}(z,z') = \sum_{c=1}^{C}\sum_{c'=1}^{C} P_{cc'}\ln\frac{P_{cc'}}{P_c.P_{c'}} \tag{3}$$

3.2 Domain Alignment

Feature alignment via mutual information minimization. Training the network ϕ_θ only via the maximization of (3) may lead to solutions where input images are clustered according to the domain information rather than their semantics. To tackle this problem, feature distribution from the different domains should be aligned in such a way that the classifier cannot cluster images according to the domain. We propose to address this domain alignment problem by the combination of two complementary strategies.

First, we propose to formulate domain alignment as a mutual information optimization problem. The key idea of ACIDS alignment strategies is that cluster assignment z should be independent from the domain d of the input image. Consequently, the mutual information between the predicted label z and the image domain d must be minimal. To this aim, we estimate the joint probability distribution $P(z,d)$ by marginalization:

$$P(z,d) = \frac{1}{N}\sum_{s=1}^{S}\sum_{n=1}^{N}\mathbb{1}(s=d)\phi_\theta(i_s^n) \tag{4}$$

Similarly to (3), we can estimate $\mathcal{MI}(z,d)$. Note that this mutual information loss leads to an extremely limited computation overhead compared to alternative solutions such as adversarial approaches.

Even though minimizing $\mathcal{MI}(z,d)$ enforces alignment between the domains, this formulation does not take advantage of the image transformation framework described in Sec.3.1. In order to further use the potential of our data augmentation approach, we propose to use the transformed image to favor domain alignment. More specifically, for every transformation $t \in \mathcal{T}$, the cluster assignment z' should be independent from the domain d of the input image. In other words, the mutual information $\mathcal{MI}(z',d|t)$ should be minimized for every transformation t. Here again, the mutual information $\mathcal{MI}(z',d|t)$ is computed via marginalization similarly to $\mathcal{MI}(z,z')$ and $\mathcal{MI}(z,d)$.

This mutual information minimization is both lightweight and efficient. Nevertheless, it acts only according to a top-down strategy, since alignment is imposed only on the output of the network and not in the early layers. Consequently, we propose to complement our framework with a feature alignment strategy based on batch normalization that acts all over the network.

Feature alignment via batch normalization. We consider that the network ϕ_θ embeds Batch normalization (BN) layers. We adopt the idea of previous works [16, 19, 2] and perform domain adaptation by updating the BN statistics. The main assumption behind this strategy is that the domain-shift can be reduced by aligning the different source feature distributions to a Gaussian reference distribution. We consider that we observe S batches of images, one from each of the S source domains. Assuming a given BN layer, $\{x_s^n\}_{n=1}^N$ and $\{x_s^{n\prime}\}_{n=1}^N$ denote the features, corresponding to domain s, in input to the BN layer for each image and the transformed counterpart respectively. We compute the batch statistics for each domain separately:

$$\forall s \in \{1..S\}, \quad \hat{\mu}_s = \frac{1}{2m}\sum_{n=1}^{N}(x_s^n + x_s^{n\prime}) \quad \hat{\sigma}_s^2 = \frac{1}{2m}\sum_{n=1}^{N}[(x_s^n - \hat{\mu}_s)^2 + (x_s^{n\prime} - \hat{\mu}_s)^2] \tag{5}$$

For a given input x computed from an image of the domain s, the output of the normalization layer is computed as follows:

$$\hat{x} = \gamma\frac{x - \mu_s}{\sqrt{\sigma_s^2 + \epsilon}} + \beta \tag{6}$$

where γ and β are the usual affine transformation parameters of the BN layer, while $\epsilon \in \mathbb{R}$ is a constant introduced for numerical stability. Note that the affine transformation parameters are shared among the different domains. This strategy guarantees that every BN layer outputs feature distributions from every domain with a mean value equal to 0 and a variance equal to 1. The main advantage of ACIDS framework is twofold. First, it does not require any additional loss that would imply more hyper-parameter tuning to obtain good convergence. Second, adaptation on the target data can be performed without accessing the source data.

3.3 Training and adaptation procedures

Overall objective function. In the previous section, we detailed how we estimate three different mutual information terms. The term $\mathcal{MI}(z, z')$ must be maximized while $\mathcal{MI}(z, d)$ and $\mathcal{MI}(z', d|t)$ must be minimized. Consequently the total minimization objective function can be written:

$$\mathcal{L} = -\mathcal{MI}(z, z') + \mathcal{MI}(z, d) + \frac{1}{T}\sum_{t\in\mathcal{T}}\mathcal{MI}(z', d|t). \tag{7}$$

where T is that cardinality of $\mathcal{T}$.

Improving stability. The computation of the mutual information $\mathcal{MI}(z, z')$ is based on the estimation of the marginal probability matrix $P_{cc'} \in \mathbb{R}^{C\times C}$. Following a standard SGD approach, this matrix is computed for every batch.

However, estimating this full probability matrix from a small batch can be inaccurate when the number of classes C is high. In addition, increasing the batch size may lead to memory issues. Practically, we observed in our preliminary experiments, that a large batch size is critical to obtain satisfying convergence of the IIC model. In our context, the issue appears to be even stronger since the images originate from different domains. Our assumption is that the higher variance of the features, despite feature alignment, leads to gradients with higher variance and unstable training. To tackle this issue, we propose to robustify the estimation of the marginal probability matrices using a moving average strategy. Considering that $P_{cc'}$ is the matrix associated to the current batch, the mutual information in Eq.(3) is computed using $\tilde{P}_{cc'}$:

$$\tilde{P}_{cc'} = \alpha P_{cc'} + (1-\alpha)\hat{P}_{cc'} \tag{8}$$

where $\hat{P}_{cc'}$ is the probability matrix $\tilde{P}_{cc'}$ estimated on the previous batch and α is a dynamic parameter. From a probabilistic point of view, this formulation can be understood as a stronger marginalization since the distribution is estimated considering in Eq.(2) not only on the N samples of the current batches but also the past batches. This estimation is correct under the assumption that the network ϕ_θ did not change too much in past SGD steps.

Adaptation to the target domain. At test time, we dispose of images from the target domain $\{i_\tau^n\}_{n=1}^{N_\tau} \in I_\tau$. However, we assume that we do not dispose anymore of the training data from the source domains. Adaptation is performed using two successive procedures. First, in order to align the feature distribution of the target data with the source distributions, we estimate the statistics of the inputs of each BN layer as in Eq.(5). The output of each BN layer is then computed according to Eq.(6). Second, our model is adapted using a variant of the mutual formulation used at training time and described in Sec.3.1 computed only on the target domain I_τ. We argue that in an unsupervised setting it is beneficial to treat samples with high prediction confidence differently from the ones with low confidence [32]. The rationale is to drive low confidence predictions towards certainty represented by high confidence predictions while not altering the latter. We propose to treat images i_τ with a prediction confidence larger than a given threshold ϵ as fixed points whose output class prediction c must be replicated by the corresponding transformed image i'_τ. This differs from the mutual information approach employed in Sec.3.1 where the output correspondence is achieved only implicitly and an incorrect class assignment to image i'_τ may negatively alter the prediction of i_τ as well, causing instability. We define:

$$\tilde{\phi}_\theta(i) = \begin{cases} \mathbb{1}(c = \operatorname{argmax} \phi_\theta(i)) & \text{if} \max \phi_\theta(i) \geq \epsilon \\ \phi_\theta(i) & \text{otherwise} \end{cases}$$

$$P_{cc'} = \frac{1}{N}\sum_{n=1}^{N} \tilde{\phi}_\theta(i_\tau^n).\phi_\theta(i_\tau^{n\prime})^\top \text{ and } P_c = P(z{=}c) = \sum_{n=1}^{N} \tilde{\phi}_\theta(i_\tau^n)$$

Table 1: Ablation results on the PACS dataset: (i) training is performed on a single, merged source domain (ii) training performed on a single source domain, (iii) removed feature alignment via mutual information minimization, (iv) removed BN feature alignment + (iii); (v) no target adaptation, (vi) target adaptation using entropy instead of mutual information. Accuracy (%) on target domain.

Target domain:	A	C	P	S	Avg
Merged source (i)	23.6	32.8	31.2	28.2	29.0
Single source A (ii)	-	31.0	45.8	28.7	-
Single source C (ii)	31.0	-	42.5	35.0	-
Single source P (ii)	33.0	33.8	-	30.5	-
Single source S (ii)	25.0	30.2	37.2	-	-
w/o domain mi loss (iii)	27.4	24.3	50.9	23.0	31.4
w/o BN alignment (iv)	23.7	34.3	38.6	23.0	22.1
w/o target adaptation (v)	34.8	36.5	44.2	40.8	39.1
entropy target adaptation (vi)	29.2	36.6	29.7	41.0	34.1
ACIDS	**42.1**	**44.5**	**64.4**	**51.1**	**50.5**

Then, we compute the mutual information term $\mathcal{MI}(z, z')$ in Eq.3 using the newly defined $P_{cc'}$ and P_c. This newly defined mutual information loss no longer suffers from the wrong i'_τ prediction problem because the argmax operation stops gradient propagation in the high confidence predictions, fixing them and making the model focus on low confidence ones.

4 Experiments

In this section, we evaluate the effectiveness of ACIDS on three widely used domain adaptation datasets and perform an ablation study showing the importance of each component of our method.

Datasets. The PACS [14] dataset is a domain adaptation dataset composed of 9,991 images divided in 7 classes spanning 4 different domains: Photo (P), Art (A), Cartoon (C) and Sketch (S). The different domains of PACS represent a rich variety of visual characteristics, from natural images to sketches, which cause large semantic gaps and make it a challenging domain adaptation dataset.

The Office31 dataset [29] contains 4,110 images divided in 3 different domains and 31 classes, namely: Amazon (A), DSLR (D) and Webcam (W).

The Office-Home [37] dataset is a larger domain adaptation benchmark containing about 15,500 images belonging to 65 different classes across 4 domains: Art (A), Clipart (C), Product (P), RealWorld (R). In addition to containing domains with a large variety of visual characteristics, the dataset presents the challenge of a large number of classes.

Evaluation protocol. We perform multiple evaluations of our model, considering at each time one of the domains as the target and the remaining ones as the

source domains. We train the model until convergence on all the sources. Then, the target domain becomes available and the source domains are discarded. At adaptation time we instantiate the domain-specific BN parameter for the target domain and perform their estimation using the newly available target images. This provides the starting point for the adaptation phase which proceeds until convergence on the target domain. In all our experiments we report the accuracy score computed on the target domain.

Implementation details. We use a randomly initialized ResNet-18 as the backbone of our model. Following [12], we adopt an overclustering strategy that fosters the model to learn more discriminant features. Instead of using only a single head with a number of outputs equal to C, we add an auxiliary overclustering head with a larger number of outputs and train the two in alternating epochs. Joint training was also considered as an alternative, but performance was negatively affected. We use respectively 49, 155 and 130 output units in the auxiliary head for the PACS, Office-Home and Office31 datasets respectively. Moreover, in order to increase robustness to bad head initialization and facilitate convergence, we replicate both the standard and the overclustering head 5 times and compute the losses for the current batch on each of them, using the average loss as the optimization objective. Further implementation details are reported in supplementary material.

4.1 Ablation Study

In this section, we present the results of our ablation study evaluating the impact of each of the components of ACIDS. We produce different variations of our method obtained as follows: (i) Training is performed on a single source domain created by merging all the source domains; (ii) Training is performed only on a single source domain, while the others are discarded; (iii) The feature alignment via mutual information mechanism proposed in Sec.3.2 is removed; (iv) Both the feature alignment via mutual information minimization mechanism and the Batch normalization feature alignment mechanism are removed, relying only on the mutual information clustering loss during training; (v) No target adaptation is performed; (vi) During adaptation the mutual information clustering loss is replaced by a prediction entropy maximization loss with threshold.

We report the quantitative results on the PACS dataset in Table 1. The ablation (iii) confirms the importance of using the mutual information loss for feature alignment during training. An analysis of the produced label assignments which we report in the supplementary material, in fact, shows that without this alignment mechanism the model produces clusters based on the domain rather than the underlying classes. The effect is that the network focuses more on learning style differences between domains rather than on semantic features, resulting in degraded performance. Removing also the BN feature alignment mechanism (iv) exacerbates the alignment problems, producing features that are not representative of the image's semantics. Moreover, training using only a single domain as the source (ii) shows a loss in performance with respect to multi-source training, highlighting that the model acquires stronger generalization capabilities when

given information about the multiple sources. Furthermore, (i) shows that it is beneficial to instantiate different BN parameters for each source domain, otherwise, the domain shift between the multiple source domains would not be mitigated. Lastly, the proposed mutual information procedure for target adaptation outperforms the entropy-based target domain adaptation method (vi) which causes a degradation in time of the performance after a small gain in the first few epochs.

In the supplementary material we report an additional ablation on the α parameter introduced in Sec.3.3.

4.2 Comparison with other methods

We now present a comparison of ACIDS against different baseline methods. We employ two popular deep clustering methods as the first baselines, namely IIC [12] and DeepCluster [4]. In both cases, we train the model using only the target data. The choice of IIC is motivated by its similarity to our method and by its state-of-the-art clustering performance [12]. For fairness, we make use of a ResNet-18 backbone on both methods and train them on the target domain. Besides, we introduce two variations of IIC that include source information: *IIC-Merge*: We train IIC on a dataset obtained by merging all the source and the target domain together; *IIC+DIAL*: Following [3], we insert domain-specific BN layers into IIC and jointly train on all source domains plus the target domain. We also compare our method with a continuous domain adaptation strategy used in [20] where we use ACIDS for training on the sources but adopt an entropy loss term for the target adaptation phase which is performed online. We also provide upper bounds for our method's performance given by SOTA domain adaptation algorithms using labeled source domains.

We report the performance of our method on the PACS dataset in Table 2. Our method performs substantially better than the DeepCluster and IIC baselines in the Art, Cartoon and Sketch domains with accuracy gains in the range from 2.3% to 4.9% with respect to IIC, while on the Photo domain our approach does not reach its performance. Moreover, our adaptation procedure outperforms the continuous domain adaptation baseline whose entropy loss does capture the semantic aspects given by our mutual information approach. The comparison with the upper bounds shows instead the obvious advantage of using supervision on the source domains. Due to the large difference of this setting with the proposed one, we omit these upper bounds from the successive evaluations.

In Table 3 we report the performance of ACIDS on the Office31 dataset. The proposed approach achieves state of the art results, performing better than both DeepCluster and IIC on all domains with accuracy gains from 0.5% to 2.1% with respect to the strongest baseline.

Lastly, Table 4 shows the results obtained on the Office-Home dataset. The proposed approach performs significantly better than the DeepCluster baseline on each domain and performs better than IIC on the Clipart and Product domains. Similarly to the results on the PACS dataset, our target adaptation procedure performs better than the continuous domain adaptation strategy.

Table 2: Comparison of the proposed approach with SOTA on the PACS dataset. Accuracy (%) on target domain. MS denotes multi source DA methods.

	Source supervision	C,P,S→A	A,P,S→C	A,C,S→P	A,C,P→S	Avg
DeepCluster [4]	×	22.2	24.4	27.9	27.1	25.4
IIC [12]	×	39.8	39.6	**70.6**	46.6	49.1
IIC-Merge [12]	×	32.2	33.2	56.4	30.4	38.1
IIC [12] + DIAL [3]	×	30.2	30.5	50.7	30.7	35.3
Continuous DA [20]	×	35.2	34.0	44.2	42.9	39.1
ACIDS	×	**42.1**	**44.5**	64.4	**51.1**	**50.5**
AdaBN [16]	✓	77.9	74.9	95.7	67.7	79.1
DIAL [3]	✓	87.3	85.5	97.0	66.8	84.2
DDiscovery [21] MS	✓	87.7	86.9	97.0	69.6	85.3
Jigsaw [1] MS	✓	84.9	81.07	98.0	79.1	85.7
AutoDIAL [2] MS	✓	90.3	90.9	97.9	79.2	89.6

Table 3: Comparison of the proposed approach with SOTA on the Office31 dataset. Accuracy (%) on target domain.

	D,W→A	A,W→D	A,D→W	Avg
DeepCluster [4]	19.6	18.7	18.9	19.1
IIC [12]	31.9	34.0	37.0	34.3
IIC-Merge [12]	29.1	**36.1**	33.5	32.9
IIC [12] + DIAL [3]	28.1	35.3	30.9	31.4
Continuous DA [20]	20.5	28.8	30.6	26.6
ACIDS	**33.4**	**36.1**	**37.5**	**35.6**

4.3 Limited target data scenario

One of the major advantages of ACIDS is the possibility of extracting semantic features from the source domains that directly transfer to the target domain. This makes it particularly suitable for the task of domain adaptation when few target samples are available. We repeat the same experiments of Sec.4.2 on the Office-Home dataset where the source domains are not altered and we consider a target domain built by randomly sampling a given portion of images in each class of the original target domain. We show the numerical results in Fig.3. We achieve a large performance boost compared to the baselines, in particular, we achieve an average 4.1% and 4.9% increase in accuracy with respect to IIC when 10% and 5% of the target images are available. Note that DeepCluster is not able to operate in the 5% scenario due to an insufficient number of target samples.

Table 4: Comparison of the proposed approach with SOTA on the Office-Home dataset. Accuracy (%) on target domain.

	C,P,R→A	A,P,R→C	A,C,R→P	A,C,P→R	Avg
DeepCluster [4]	8.9	11.1	16.9	13.3	12.6
IIC [12]	**12.0**	15.2	22.5	**15.9**	16.4
IIC-Merge [12]	11.3	13.1	16.2	12.4	13.3
IIC [12] + DIAL [3]	10.9	12.9	15.4	12.8	13.0
Continuous DA [20]	10.2	11.5	13.0	11.7	11.6
ACIDS	**12.0**	**16.2**	**23.9**	15.7	**17.0**

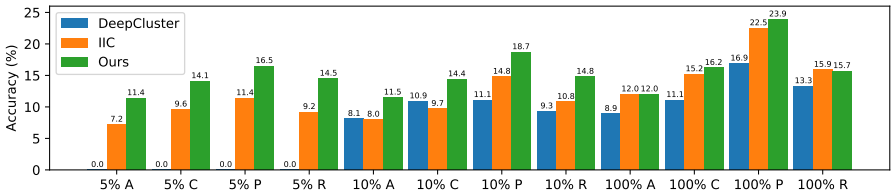


Fig. 3: Comparison of the proposed approach with SOTA on the Office-Home dataset in the limited target data scenario. Labels express the target domain (A,C,P or R) and the percentage of images used in the target domain.

5 Conclusions

In this paper, we propose a novel domain adaptation setting and show it is possible to transfer knowledge from multiple source domains to a target domain when both sources and target data have no annotations. Our method makes use of a novel information-theoretic loss for feature alignment and couples it with domain-alignment layers to discover semantic labels from the source domains. When target data becomes available, we perform adaptation without requiring the availability of source data. We achieve state-of-the-art performance on three widely used domain adaptation datasets and show a clear advantage of the proposed approach under low target data conditions. Future works will consider the adaptation of the approach to the unsupervised segmentation scenario.

Acknowledgements

We acknowledge financial support from H2020 EU project SPRING - Socially Pertinent Robots in Gerontological Healthcare. This work was carried out under the “Vision and Learning joint Laboratory” between FBK and UNITN.

References

1. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2229–2238 (2019)
2. Carlucci, F.M., Porzi, L., Caputo, B., Ricci, E., Bulò, S.R.: Autodial: Automatic domain alignment layers. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5077–5085 (2017)
3. Carlucci, F.M., Porzi, L., Caputo, B., Ricci, E., Bulò, S.R.: Just dial: Domain alignment layers for unsupervised domain adaptation. In: Image Analysis and Processing - International Conference on Image Analysis And Processing (ICIAP) 2017. pp. 357–369 (2017)
4. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Computer Vision – European Conference of Computer Vision (ECCV) 2018. pp. 139–156 (2018)
5. Chang, J., Wang, L., Meng, G., Xiang, S., Pan, C.: Deep adaptive image clustering. In: IEEE International Conference on Computer Vision (ICCV). pp. 5880–5888 (2017)
6. Csurka, G.: Domain adaptation in computer vision applications, vol. 2. Springer (2017)
7. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML) - Volume 37. p. 11801189 (2015)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems (NIPS). pp. 2672–2680 (2014)
9. Haeusser, P., Plapp, J., Golkov, V., Aljalbout, E., Cremers, D.: Associative deep clustering: Training a classification network with no labels. In: German Conference on Pattern Recognition (2018)
10. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. In: Proceedings of the 35th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 80, pp. 1989–1998. PMLR (10–15 Jul 2018), http://proceedings.mlr.press/v80/hoffman18a.html
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (ICML). pp. 448–456 (2015)
12. Ji, X., Vedaldi, A., Henriques, J.F.: Invariant information clustering for unsupervised image classification and segmentation. IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9864–9873 (2019)
13. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: Computer Vision – European Conference of Computer Vision (ECCV) 2016. pp. 577–593 (2016)
14. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5542–5550 (2017)
15. Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.Z., Hospedales, T.M.: Episodic training for domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1446–1455 (2019)

16. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation. In: 5th International Conference on Learning Representations (ICLR). OpenReview.net (2017), https://openreview.net/forum?id=Hk6dkJQFx
17. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: Proceedings of the 32nd International Conference on Machine Learning (ICML) - Volume 37. p. 97105 (2015)
18. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML) - Volume 70. pp. 2208–2217 (2017)
19. Mancini, M., Karaoguz, H., Ricci, E., Jensfelt, P., Caputo, B.: Kitting in the wild through online domain adaptation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (October 2018)
20. Mancini, M., Bulo, S.R., Caputo, B., Ricci, E.: Adagraph: Unifying predictive and continuous domain adaptation through graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6568–6577 (2019)
21. Mancini, M., Porzi, L., Rota Bulò, S., Caputo, B., Ricci, E.: Boosting domain adaptation by discovering latent domains. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3771–3780 (2018)
22. Morerio, P., Cavazza, J., Murino, V.: Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In: International Conference on Learning Representations (ICLR) (2018), https://openreview.net/forum?id=rJWechg0Z
23. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4500–4509 (June 2018)
24. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision (ECCV). pp. 69–84 (2016)
25. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. vol. 22, pp. 199–210 (2011)
26. Peng, X., Saenko, K.: Synthetic to real adaptation with generative correlation alignment networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1982–1991 (2018)
27. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation pp. 1406–1415 (October 2019)
28. Roy, S., Siarohin, A., Sangineto, E., Bulo, S.R., Sebe, N., Ricci, E.: Unsupervised domain adaptation using feature-whitening and consensus loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9463–9472 (2019)
29. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Computer Vision – European Conference of Computer Vision (ECCV) 2010. pp. 213–226 (2010)
30. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks. In: CVPR (2018)
31. Siarohin, A., Sangineto, E., Sebe, N.: Whitening and coloring transform for GANs (2019), https://openreview.net/forum?id=S1x2Fj0qKQ
32. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence (2020)

33. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: European Conference on Computer Vision (ECCV). pp. 443–450 (2016)
34. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4068–4076 (2015)
35. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2962–2971 (2017)
36. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
37. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5018–5027 (2017)
38. Wang, X., He, K., Gupta, A.: Transitive invariance for self-supervised visual representation learning. pp. 1338–1347 (10 2017). https://doi.org/10.1109/ICCV.2017.149
39. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML) - Volume 48. p. 478487. ICML16 (2016)
40. Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L.: Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3964–3973 (2018)
41. Zen, G., Sangineto, E., Ricci, E., Sebe, N.: Unsupervised domain adaptation for personalized facial emotion recognition. In: Proceedings of the 16th international conference on multimodal interaction (2014)
42. Zhao, S., Fu, H., Gong, M., Tao, D.: Geometry-aware symmetric domain adaptation for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9788–9798 (2019)