Self-Supervised Representation Learning From Multi-Domain Data

Zeyu Feng Chang Xu Dacheng Tao
UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering,
The University of Sydney, Darlington, NSW 2008, Australia
zfen2406@uni.sydney.edu.au, {c.xu, dacheng.tao}@sydney.edu.au

Abstract

We present an information-theoretically motivated constraint for self-supervised representation learning from multiple related domains. In contrast to previous self-supervised learning methods, our approach learns from multiple domains, which has the benefit of decreasing the build-in bias of individual domain, as well as leveraging information and allowing knowledge transfer across multiple domains. The proposed mutual information constraints encourage neural network to extract common invariant information across domains and to preserve peculiar information of each domain simultaneously. We adopt tractable upper and lower bounds of mutual information to make the proposed constraints solvable. The learned representation is more unbiased and robust toward the input images. Extensive experimental results on both multi-domain and large-scale datasets demonstrate the necessity and advantage of multi-domain self-supervised learning with mutual information constraints. Representations learned in our framework on state-of-the-art methods achieve improved performance than those learned on a single domain.

1. Introduction

Unsupervised visual representation learning algorithms using deep convolutional neural networks (CNNs) have led to breakthroughs in relieving the burden of massive manual annotation [47, 5, 11, 48, 19]. They are capable of learning high-level semantic image representation transferable to various downstream tasks without using expensive annotated labels, which greatly expend the scope of applications for CNNs. Among many unsupervised learning methods, the recently emerged self-supervised learning (SSL) techniques produce excellent representations, achieving state-of-the-art performance on standard computer vision benchmarks [34, 20, 18, 7, 36, 51]. SSL discovers supervisory signals directly from the input data itself and defines a pretext task from this supervision. CNNs trained to accomplish such objectives have to understand the input data, for instance salient objects and surrounding backgrounds for object-oriented images. Intermediate layers of the CNN will hence gain the ability of extracting high-level semantic representations for this type of data, which are useful for solving different downstream tasks like image recognition.

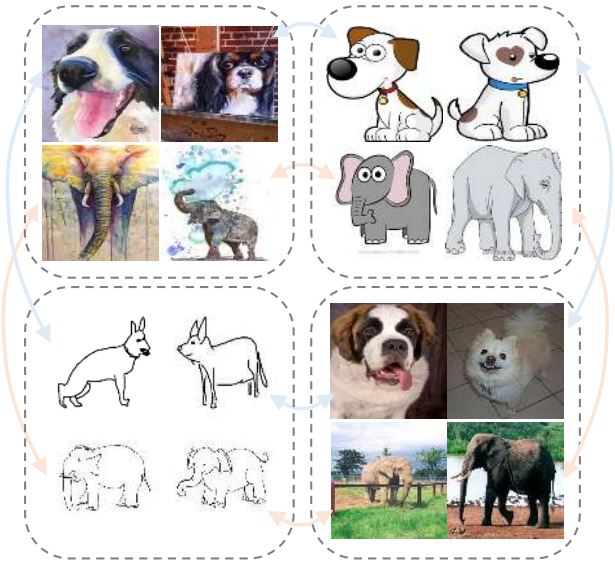

Figure 1: We propose to perform self-supervised learning using data from multiple related domains. (Images are selected from the PACS dataset [26].)

Although the efficient training on unlabeled data largely alleviates the burden of human labeling, the properties of unlabeled training data themselves are not investigated thoroughly for image based SSL. Most of the prior work focuses on proposing novel pretext tasks to improve the learned representation. Few methods investigate what is the influence of training data used or in which way should training data be chosen for SSL. In computer vision community, it has long been recognized that datasets collected for vision tasks are often individually biased and deviated from the goal of representing the visual world [46], if application on such real world images is our aim. Moreover, total available images for some domains can also be fundamentally constrained

like art images and sketches [26]. CNNs trained on one single dataset will likely lead to biased representations, paying excessive attention on unwanted variations in images induced by sources like viewing angle, illumination condition and imaging system [46]. SSL algorithms generally train CNNs on one dataset (ImageNet). As a result, such restricted training on only one dataset will less likely lead to good unbiased general-purpose representations of images we interest.

Training using multiple datasets has achieved great success in compensating the training set bias [12, 21, 14]. However, such considerations have not been appreciated for SSL. Toward the goal of learning unbiased representations that filter away unwanted variations, we propose to train SSL models on data from multiple related domains. Given a dataset on which we want to learn representations, we can exploit existing datasets from other related domains that contain semantically overlapping but non-identical information (See Figure 1 for illustration), and perform multi-domain learning (MDL) for SSL. This has the benefit of enriching the data variety, decreasing the build-in bias of individual dataset, leveraging shared information and allowing knowledge transfer across multiple domains, making discoveries that could not have been obtained from any individual domain alone [43]. The learned CNN is expected to extract superior representations on the images of interest than training solely on them.

As has been discovered in supervised learning, learning from multiple domains by just naïvely concatenating more datasets is not the best policy, which can even lead to reduced performance on the dataset of interest [46, 21]. We also observed this phenomenon for SSL. This fact suggests that the presence of domain difference can impact performance when left untreated and cross-domain relationship has to be considered.

Based on this observation, in this paper we present mutual information (MI) based criteria for SSL from multi-domain data. MI is used as an indicator of how much inter- and intra-domain information the model captures. In order to capture semantically shared information across different domains, we minimize the MI between the representation and the domain label of images. Under this objective, domain-invariant information that excludes unwanted variations will be encoded to the high-level representations. On the other hand, the enforced domain invariance and the existence of dataset imbalance may let the model overlooks or overfits some domains and hence lose their information. Regarding persevering specific information of each domain, we introduce constraints on the value of the MI between input images and their CNN representations for every domain, so that the representation will maintain certain level of information on every domain. To make these two information theoretic constraints computable, an adversarial approximation of the variational upper bound and a contrastive lower bound of MI are applied to approximately optimize the objectives. Therefore, the learned representation will result in a controllable trade-off between learning domain-invariant and domain-specific information.

To demonstrate the effectiveness of our proposed MI criteria on MDL for SSL, we conduct experiments on the multi-domain dataset PACS [26], as well as on large-scale datasets ILSVRC 2012 [42] and Places [55] following the SSL benchmark. We perform ablation studies to examine the effectiveness of each component in our model. Experimental results demonstrate the advantages of our approach.

2. Related work

This work relates to several topics in computer vision and machine learning: self-supervised learning (SSL), multi-domain learning (MDL), domain generalization (DG) and mutual information (MI) criterion, which we briefly review here.

Self-supervised learning. SSL constructs pretext tasks by discovering supervisory signals directly from the input data itself. CNNs trained to predict this supervisory information will encode high-level semantic representations of the input. Notable types of pre-text tasks for images include constructing relationship between image patches like patch position prediction [9, 31], solving jigsaw puzzle [32, 6] and counting [33], and reconstructing part of the image like image completion [37], colorization [52, 24, 25] and channel prediction [53].

Some other important aspects beyond the form of pretext tasks have also been studied. For example, Ren and Lee [41] studied the effects of synthetic images for representation learning and the influence of the domain gap between synthetic images and real-world images. It is relied on the free ground truth from synthetic images. Doersch and Zisserman [10] investigated the effect of combining multiple pretext tasks together. They conclude that deeper networks outperform shallow networks and combining tasks always improves performance over the tasks alone. Kolesnikov *et al*. conducted a thorough large-scale study on the choice of modern CNNs for self-supervised learning by revisiting several pretext tasks [22]. They discover many crucial insights related to the CNNs architecture including skip-connections and the number of filters. Our work also focuses beyond designing pretext task. We explore the influence of using multiple related datasets and propose two strategies for learning with multi-domain data.

Multi-domain learning and domain generalization. MDL aims to solve the shortcomings of a single dataset by using the data from multiple domains [12]. Several methods in supervised learning setting design specific network to handle domain-related feature, such as encoding domain descriptor [50] and using domain-specific parame-

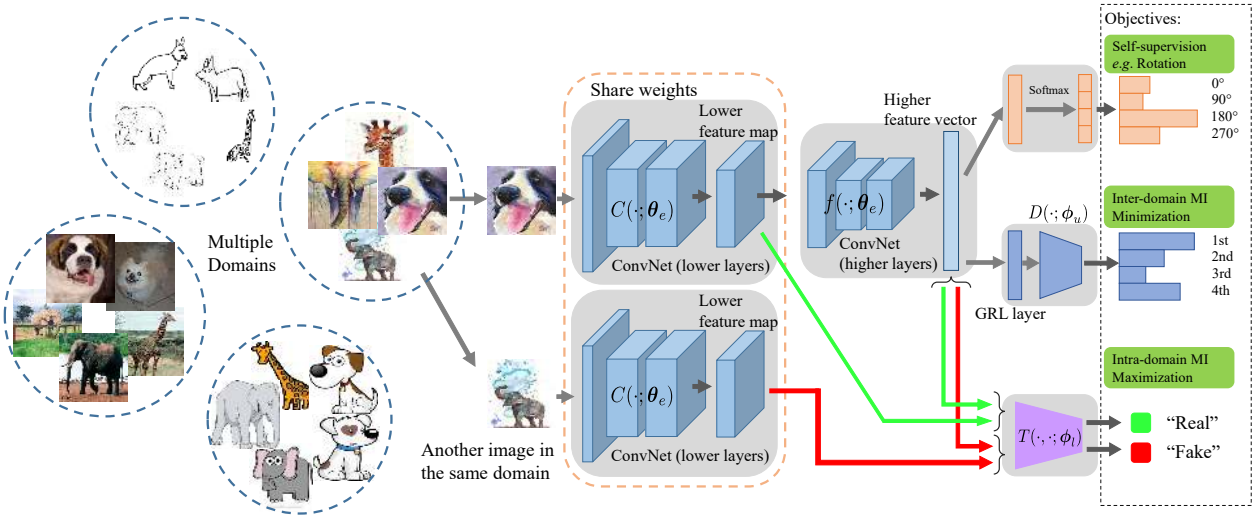

Figure 2: Illustration of the proposed method. We leverage image data from multiple related domains to perform a self-supervised learning task. MI acts as a proxy for domain-related information and is used as constraints for the main SSL task.

ters [39, 40, 8]. Some other work seek a common feature extractor, for example discovering domain-related neurons by domain guided dropout [49] or learning domain-invariant representation [43]. Since SSL expect the encoder to be able to extract better representation for certain input images or similar ones, we do not reply on specific-parameters and focus on a common feature extractor.

Another line of research that aims to improve the generalization ability of a supervised learning algorithm is DG [16, 26, 30, 27, 28, 29]. Note that our approach is essentially different from methods that cope with DG problem, where they care more about building a domain-agnostic classifier that is effective when applied in an unseen target domain. While for SSL, the aim is learning better representation for input images, so that they can be used to extract representations of these images or similar ones. We argue that it is demanding to require the learned representation transfer to a dramatically different domain, where the pre-text task could even be unsuitable (*e.g.* transferring representations learned by RotNet [18] to a rotation-invariant image domain seems unreasonable). We aim at taking more related unlabeled datasets into training to boost the performance on similar images given one dataset, even though this dataset exist in small amounts. Most of the DG methods only conduct experiments on small-scale datasets like VLCS [14] and PACS [26]. It is unclear whether they are able to scale to large-scale datasets like ImageNet and Places. We aim at improving SSL with images from large-scale datasets. Data from each domain are expected to help each other, and we mainly evaluate the learned representation based on the performance on each individual domain.

Mutual information criterion. MI criterion has been explored before to model the relationship of data from different domains. Shi and Sha [44] examined the objectives in the form of both MI between all data and their binary domain labels and MI between the target data and estimated class labels for unsupervised domain adaptation. However, their model and the corresponding computation of MI build upon discriminative clustering and metric formulation, which can not be scaled to deep neural networks. Gholami *et al.* [17] use MI for multi-target domain adaptation with labeled source domain data. Its optimization of MI objective is based on Barber & Agakov lower bound [3] of MI. MI has achieved wide applications and successes in deep learning [2, 1, 4]. It has also been used to establish connections between structure in data [36, 19]. We use tractable bounds of MI in this paper to establish connections between multiple domains.

3. Multi-domain learning

In this section, we first introduce the problem setting and present the proposed information-theoretic constraints. Then we describe in detail of the tractable approximation of MI minimization and maximization. Our model is summarized in Figure 2.

Our goal is to transform each image example $\mathbf{x} \in \mathcal{X}$ from a certain domain, where $\mathcal{X}$ denotes an input space for images, into a high-level semantic representation $\mathbf{z} \in \mathcal{Z}$ that is transferable to a variety of downstream tasks in an unsupervised way. To achieve this goal, we employ a parametric encoder function $E(\cdot;\boldsymbol{\theta}_e) : \mathcal{X} \to \mathcal{Z}$ with parameters $\boldsymbol{\theta}_e$ (*e.g.* a neural network).

We are interested in learning with data from multiple domains. Assume the number of related domains available at hand is M. For $i = 1, \ldots, M$, the i-th domain has N_i

training images: $S_i = \{(\mathbf{x}_i^{(j)}, d_i^{(j)})\}_{j=1}^{N_i}$, where d is the discrete domain label. We denote the empirical probability distribution of $\mathbf{x}_i$ on the i-th domain by $p_i(\mathbf{x})$. The representation $\mathbf{z}$ for an image $\mathbf{x}$ are obtained by sampling from a conditional probability distribution $p_{\boldsymbol{\theta}_e}(\mathbf{z}|\mathbf{x})$ parameterized by $\boldsymbol{\theta}_e$. There are several possible choices for the encoder distribution $p_{\boldsymbol{\theta}_e}(\mathbf{z}|\mathbf{x})$. In this paper we assume that $p_{\boldsymbol{\theta}_e}(\mathbf{z}|\mathbf{x})$ is defined by the deterministic function of $\mathbf{x}$, which is $E(\cdot;\boldsymbol{\theta}_e)$. The marginal distribution of $\mathbf{z}$ on each domain is then

$$p_{i,\boldsymbol{\theta}_e}(\mathbf{z}) = \sum_{\mathbf{x}\in S_i} p_{\boldsymbol{\theta}_e}(\mathbf{z}|\mathbf{x}) p_i(\mathbf{x}). \tag{1}$$

First of all, we want to encode the representation $\mathbf{z}$ with semantic information by training under an SSL objective. Let $F(\cdot;\boldsymbol{\theta}_f)$ denotes the head network for SSL that takes $\mathbf{z}$ as input. The loss function for an SSL method is defined as $l(F(E(\mathbf{x};\boldsymbol{\theta}_e);\boldsymbol{\theta}_f))$ for simplicity. Many state-of-the-art SSL methods can be used here to learn representations. For example, if we choose Rotation [18] as the SSL task, then $l(\cdot,\cdot)$ is the cross-entropy loss for rotation classification. Given data from all domains available, the objective for multi-domain SSL is

$$\min_{\boldsymbol{\theta}_e,\boldsymbol{\theta}_f} \mathcal{L}_f = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i} l(F(E(\mathbf{x}_i^{(j)};\boldsymbol{\theta}_e);\boldsymbol{\theta}_f)). \tag{2}$$

Learning under this objective alone is equal to naïvely combining datasets. Our experimental results reveal that the performance of MDL is sometimes no better than learning on the single domain, which shows that naïvely adding additional training examples is not always beneficial. we next introduce MI based constraints to address this issue.

3.1. Mutual information constraints

As previously discussed, we have to explicitly model cross-domain relationship so that the resulting representation could learn cross-domain semantic knowledge. We now discuss in detail of our desiderata.

3.1.1 Domain-invariant information

Regarding leveraging information across domains, our desired properties of the representation are that they capture the common semantic knowledge in the input data across different domains, although they may appear differently. Variations in appearance of images may include viewing angle, illumination condition, image style, imaging system, place where datasets are collected and even preference of dataset collectors [14, 43]. For some downstream tasks (*e.g.* object-oriented image recognition), these variations are harmful for representation learning since they are unrelated to the decision of the task most of the time. Hence, we hope the $\mathbf{z}$ of images with similar objects from different domains will be similar as much as possible and reveal the information of their specific form of variation as less as possible.

Let $p(\mathbf{x})$ and $p_{\boldsymbol{\theta}_e}(\mathbf{z})$ denote the empirical mixture distributions derived from the collection of distributions from every domain, and $\mathbf{x} \sim p(\mathbf{x})$ and $\mathbf{z} \sim p_{\boldsymbol{\theta}_e}(\mathbf{z})$ are random variables. We express our desideratum for learning similar concepts from related domains as limiting the maximum value of MI $I(\mathbf{z}, d)$ between the image representation $\mathbf{z}$ from all domains and the corresponding domain label d of the original image. Conceptually, this objective is similar to the idea in existing works to make marginal distributions of the representation similar across domains [15, 43]. If $I(\mathbf{z}, d)$ is small, then given a $\mathbf{z}$, it is hard to tell which domain the input image $\mathbf{x}$ is from. As a result, the learned representation will discard unwanted domain-related variations and form a domain-invariant representation space, where every domain has a similar marginal distribution.

3.1.2 Domain-specific information

Although domain-related variations are discarded, enforcing the similarity in marginal distributions bear no direct consequence on useful information capture on each domain. The domain-invariant representation space can also be created by projecting input images to a random invariant space without semantic correspondence. Maintenance of specific information for each domain is necessary. Furthermore, the MDL strategy introduced so far does not take dataset imbalance into consideration. Images in some domains could be abundant while in other domains they might be scarce. Domains with only small amount of data will be either overlooked by the domain-invariant objective or overfitted by the SSL objective. Specific information about these domains should be preserved to ensure an intra-domain performance.

Formally, let $\mathbf{x}_i \sim p_i(\mathbf{x})$ and $\mathbf{z}_i \sim p_{i,\boldsymbol{\theta}_e}(\mathbf{z})$ denote the random variables of input images and representations from i-th domain, respectively. Our desideratum is limiting the minimum value of MI $I(\mathbf{z}_i, \mathbf{x}_i)$ for every domain so that domain-specific information is retained in the representation to a certain level for every domain.

Rewriting the objective function (2) with these two desiderata, we have the following constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}_e,\boldsymbol{\theta}_f} \quad & \mathcal{L}_f = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i} l(F(E(\mathbf{x}_i^{(j)};\boldsymbol{\theta}_e);\boldsymbol{\theta}_f)) \\ \text{s.t.} \quad & I(\mathbf{z}, d) < \epsilon_u \\ & I(\mathbf{z}_i, \mathbf{x}_i) > \epsilon_l, \forall i \in \{1, \ldots, M\}, \end{aligned} \tag{3}$$

which is different from (2) as the introduced MI constraints

allow $\mathbf{z}$ to be more semantically representative by excluding unwanted variations while retaining specific information in each $\mathbf{x}_i$. The hyper-parameters ϵ_u and ϵ_l control the amount of MI between $\mathbf{z}$ and $\mathbf{x}$. Approximating the Lagrangian dual of problem (3) using Lagrangian multipliers λ_u and λ_l, the objective becomes:

$$\min_{\boldsymbol{\theta}_e,\boldsymbol{\theta}_f} \mathcal{L}_f + \lambda_u I(\mathbf{z}, d) - \lambda_l \sum_{i=1}^{M} I(\mathbf{z}_i, \mathbf{x}_i). \tag{4}$$

Both MI terms in (4) are difficult to compute and optimize. We provide tractable approximations by using upper and lower bounds of MI in the following two sections.

3.2. Upper bound of $I(\mathbf{z}, d)$ via adversarial training

Mutual information is upper bounded by replacing one of the marginal distributions with a variational posterior distribution [1, 2, 54, 38]. Formally, for any distribution $q(d)$, we can have an upper bound of $I(\mathbf{z}, d)$:

$$\begin{aligned} I(\mathbf{z}, d) &= \mathbb{E}_{p_{\boldsymbol{\theta}_e}(\mathbf{z},d)}\big[\log p_{\boldsymbol{\theta}_e}(d|\mathbf{z}) - \log p(d)\big] \\ &= \mathbb{E}_{p_{\boldsymbol{\theta}_e}(\mathbf{z})} D_{\mathrm{KL}}(p_{\boldsymbol{\theta}_e}(d|\mathbf{z}) \| q(d)) - D_{\mathrm{KL}}(p(d) \| q(d)) \\ &\leq \mathbb{E}_{p_{\boldsymbol{\theta}_e}(\mathbf{z})} D_{\mathrm{KL}}(p_{\boldsymbol{\theta}_e}(d|\mathbf{z}) \| q(d)) := C. \end{aligned} \tag{5}$$

However, the $p_{\boldsymbol{\theta}_e}(d|\mathbf{z})$ in Eq. (5) is intractable. We can instead approximate $p_{\boldsymbol{\theta}_e}(d|\mathbf{z})$ with a parameterized model $q_{\boldsymbol{\phi}_u}(d|\mathbf{z})$, thus this upper bound has a lower bound [45]:

$$\begin{aligned} C &\geq \mathbb{E}_{p_{\boldsymbol{\theta}_e}(\mathbf{z})}\big[D_{\mathrm{KL}}(p_{\boldsymbol{\theta}_e}(d|\mathbf{z}) \| q(d)) - \\ &\qquad D_{\mathrm{KL}}(p_{\boldsymbol{\theta}_e}(d|\mathbf{z}) \| q_{\boldsymbol{\phi}_u}(d|\mathbf{z}))\big] \\ &= \mathbb{E}_{p_{\boldsymbol{\theta}_e}(\mathbf{z},d)}\big[\log q_{\boldsymbol{\phi}_u}(d|\mathbf{z}) - \log q(d)\big] := \hat{C}. \end{aligned} \tag{6}$$

Maximizing $\hat{C}$ with respect to $\boldsymbol{\phi}_u$ will decrease $D_{\mathrm{KL}}(p_{\boldsymbol{\theta}_e}(d|\mathbf{z}) \| q_{\boldsymbol{\phi}_u}(d|\mathbf{z}))$, making $\hat{C}$ a good approximate toward the upper bound C. The $q(d)$ in Eq. (5) can be chosen as a kernel density estimate based on all datasets [45]. By making $D_{\mathrm{KL}}(p(d) \| q(d))$ as small as possible, C gets closer to $I(\mathbf{z}, d)$. Therefore, minimization of $I(\mathbf{z}, d)$ can be achieved through the following adversarial objective:

$$\min_{\boldsymbol{\theta}_e} \max_{\boldsymbol{\phi}_u} \mathcal{L}_u = \mathbb{E}_{p_{\boldsymbol{\theta}_e}(\mathbf{z},d)}\big[\log q_{\boldsymbol{\phi}_u}(d|\mathbf{z}) - \log q(d)\big]. \tag{7}$$

Practically, we model $q_{\boldsymbol{\phi}_u}(d|\mathbf{z})$ as a function $D(\cdot; \boldsymbol{\phi}_u) : \mathcal{Z} \to \Delta = \{\boldsymbol{\alpha} \in \mathbb{R}^M : \alpha_1 + \ldots + \alpha_M = 1, \alpha_d \geq 0, d = 1, \ldots, M\}$ (*e.g.* a neural network with softmax output) with parameters $\boldsymbol{\phi}_u$ that outputs a probability vector for an input $\mathbf{z}$, where Δ is the probability simplex. The value of the d-th component is denoted by $D^{(d)}(\cdot; \boldsymbol{\phi}_u)$. Modeled with empirical distributions, $\mathcal{L}_u$ can be further expressed as

$$\mathcal{L}_u = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{N_i} \sum_{j=1}^{N_i} \log\big[D^{(d_i^{(j)})}(E(\mathbf{x}_i^{(j)}; \boldsymbol{\theta}_e); \boldsymbol{\phi}_u) / q(d_i^{(j)})\big]. \tag{8}$$

Interestingly, this formulation is equal to the cross-entropy loss used in multi-class classification. The network $D(\cdot; \boldsymbol{\phi}_u)$ classifies input $\mathbf{z}$ into correct domain while $E(\cdot; \boldsymbol{\theta}_e)$ tries to confuse $D(\cdot; \boldsymbol{\phi}_u)$. In practice, the $q(d)$ is a constant value and can be omitted during optimization.

3.3. Lower bound of $I(\mathbf{z}_i, \mathbf{x}_i)$

It is able to maximize the MI $I(\mathbf{z}_i, \mathbf{x}_i)$ in objective (4) by just maximizing one of its tractable lower bounds. MI can have a lower bound formulation based on Noise Contrastive Estimation [36]:

$$\begin{aligned} I(\mathbf{z}_i, \mathbf{x}_i) &\geq \hat{I}^{(NCE)}(\mathbf{z}_i, \mathbf{x}_i) \\ &:= \mathbb{E}_{p_i(\mathbf{x})}\Big[T(E(\mathbf{x}_i; \boldsymbol{\theta}_e), \mathbf{x}_i; \boldsymbol{\phi}_l) - \\ &\qquad \mathbb{E}_{\tilde{p}_i(\mathbf{x})}\Big[\log \sum_{\mathbf{x}_i'} \mathrm{e}^{T(E(\mathbf{x}_i; \boldsymbol{\theta}_e), \mathbf{x}_i'; \boldsymbol{\phi}_l)}\Big]\Big], \end{aligned} \tag{9}$$

where $\mathbf{x}_i'$ is the random variable of input images sampled from the distribution $\tilde{p}_i(\mathbf{x}) = p_i(\mathbf{x})$. We can also maximize MI by maximizing the Jensen-Shannon divergence (JSD) [35] formulation of MI, which is capable of providing stable approximation results [19]. To be specific, the JSD formulation is

$$\begin{aligned} \hat{I}^{(JSD)}(\mathbf{z}_i, \mathbf{x}_i) := \mathbb{E}_{p_i(\mathbf{x})}\big[-\mathrm{sp}(-T(E(\mathbf{x}_i; \boldsymbol{\theta}_e), \mathbf{x}_i; \boldsymbol{\phi}_l))\big] - \\ \mathbb{E}_{p_i(\mathbf{x}) \times \tilde{p}_i(\mathbf{x})}\big[\mathrm{sp}(T(E(\mathbf{x}_i; \boldsymbol{\theta}_e), \mathbf{x}_i'; \boldsymbol{\phi}_l))\big], \end{aligned} \tag{10}$$

where $\mathrm{sp}(x) = \log(1 + e^x)$ is the softplus function. As suggested in [19], the function $T(\cdot, \cdot; \boldsymbol{\phi}_l)$ can share lower layers with $E(\cdot; \boldsymbol{\theta}_e)$ so that $E(\cdot; \boldsymbol{\theta}_e) = f(\cdot; \boldsymbol{\theta}_e) \circ C(\cdot; \boldsymbol{\theta}_e)$ and $T(\cdot, \cdot; \boldsymbol{\phi}_l) = D(C(\cdot; \boldsymbol{\theta}_e), E(\cdot; \boldsymbol{\theta}_e); \boldsymbol{\phi}_l)$. Maximizing Eq. (10) with respect to $\boldsymbol{\theta}_e$ and $\boldsymbol{\phi}_l$ will maximize the MI $I(\mathbf{z}_i, \mathbf{x}_i)$ in objective (4).

Our complete model comprises three core modules: multi-domain self-supervised learning (Eq. (2)), domain-invariant representation constraint (Eq. (8)) and domain-specific information preservation (Eq. (10)), and can be written as the following minimax objective:

$$\min_{\boldsymbol{\theta}_e,\boldsymbol{\theta}_f,\boldsymbol{\phi}_l} \max_{\boldsymbol{\phi}_u} \mathcal{L}_f + \lambda_u \mathcal{L}_u - \lambda_l \sum_{i=1}^{M} \hat{I}^{(JSD)}(\mathbf{z}_i, \mathbf{x}_i). \tag{11}$$

Solving this objective requires adversarial training of the CNN. We connect a Gradient Reversal Layer (GRL) [15] after $E(\cdot; \boldsymbol{\theta}_e)$, so that maximizing $\mathcal{L}_u$ w.r.t. $\boldsymbol{\phi}_u$ will give rise to the minimization of $\mathcal{L}_u$ w.r.t. $\boldsymbol{\theta}_e$.

4. Experiments

In this section, we conduct experiments on three types of dataset to demonstrate the effectiveness of our approach. These datasets are:

- PACS dataset [26]: A small-scale multi-domain dataset containing 4 sub-datasets, where the image styles are different. This mainly aims to examine how our approach performs in scenarios where the total number of data available is limited.
- ImageNet (ILSVRC 2012) [42] and Places [55]: We combine these two large-scale datasets and each dataset is viewed as a domain. The former mainly contains object-oriented images while the latter contains scene-oriented images. We test the performance of our apporach under large-scale datasets through this setting.
- PASCAL VOC 2007 [13]: We test how our approach performs when we use more diverse data available (ImageNet and Places) to help the learning on a rather small dataset (PASCAL).

We investigate the behavior of our mutual information constraints in comparison to both the standard single-domain SSL model and the strategy of naïvely combining datasets (marked as *DeepAll* in tables) as proof-of-principle.

Linear classification is a common procedure for feature evaluation [53]. Its rationality has also been confirmed recently by a study through thorough experiments, where it is shown that a linear model is adequate for evaluating the quality of a representation [22]. Therefore, we evaluate the learned representations by training a linear multi-class classifier on top of them. High performance on this task requires high-level semantic image understanding from the learned representation. Following previous procedure [36, 19], for all experiments we evaluate representations from the last convolutional layer (`conv5`) and the output of the encoder $E(\cdot;\boldsymbol{\theta}_e)$ (last fully-connected) layer (`fc7`).

4.1. Implementation details

We choose predicting image rotation (RotNet) [18] and AET [51] as the running examples of SSL since they are efficient methods and achieve state-of-the-art results on many downstream tasks. The proposed multi-domain solution can be integrated with mainstream SSL methods. As several transformation copies of an image are created in every batches in RotNet and AET, we apply the MI constraints separately on each copy of a minibatch. The encoder function $E(\cdot;\boldsymbol{\theta}_e)$ is implemented as a standard AlexNet architecture [23] following the setting of [18]. It consists of five convolutional layers and two fully-connected layers. The prediction function $F(\cdot;\boldsymbol{\theta}_f)$ of SSL is implemented as a one-layer linear network. For functions $T(\cdot,\cdot;\boldsymbol{\phi}_l)$ and $D(\cdot;\boldsymbol{\phi}_u)$ used in MI approximation, we use a three-layer multilayer perceptron (MLP) with the number of hidden layers being 512. The feature map $C(\cdot;\boldsymbol{\theta}_e)$ are taken from the `conv4` layer of the encoder $E(\cdot;\boldsymbol{\theta}_e)$. For all experiments, we set the Lagrangian multipliers λ_u and λ_l as 0.1, except on PACS λ_l is 1. In order to prevent the network from seeing different levels of total images for each domain, we divide each data batch equally to each domain. Our model is trained with momentum of 0.9, a batch size of 128 and an l_2 penalization of all weights with $5 \cdot 10^{-4}$. The learning rate is set to 0.01 initially and then decayed by a factor of 10 when loss on validation set reaches plateau.

On PACS dataset, due to the low number of total images in the dataset, the number of channels on each convolutional layer of $E(\cdot;\boldsymbol{\theta}_e)$ are scaled to $1/4$ of the original size. The last convolutional layer is followed by 2 fully connected layers with output size of 512 and 64, respectively. The `conv5` feature is pooled to a size of 64 by global average pooling for linear evaluation. The number of hidden layers in $D(\cdot;\boldsymbol{\phi}_u)$ is also scaled to 64. For experiments involving ImageNet, the output of $E(\cdot;\boldsymbol{\theta}_e)$ is first linearly projected to a 128-dimensional feature vector before it is fed into $T(\cdot,\cdot;\boldsymbol{\phi}_l)$ according to the practice of [19] for the purpose of reducing memory consumption. Feature map on `conv5` are spatially resized (with adaptive max pooling) so as to have around 9,000 elements [52] for linear evaluation.

4.2. PACS dataset

PACS [26] consists of images from photo (P), art painting (A), cartoon (C) and sketch (S) domains. Although it is originally proposed for the purpose of evaluating DG methods, the four domains in PACS are closely related and share same object-level semantics (same seven classes), while are seemingly dissimilar (different image style). Training on any one of these domains alone will not guarantee good comprehensive semantics for objects.

The numbers of images in each domain are 1,670, 2,048, 2,344 and 3,929, respectively. The total number of images is 9,991. We use the original train-validation split on each domain in PACS, and train our model on training set and report the representation evaluation results on each validation set. The linear classifier is chosen as support vector machine (SVM). Experimental results are summarized in Table 1.

From the results, we can see that *DeepAll* (training on all sub-datasts together) is slightly better than training an SSL algorithm on a single domain on average. But the performance on some domains get decreased. This suggests that SSL from multi-domain data without considering cross-domain relationship will hurt the representation. Our method (*DeepAll+MI*) outperforms *DeepAll* and single domain training on most domains. The average accuracies on `conv5` and `fc7` are improved by 1.1% and 3.0% under RotNet, respectively. Information loss on some domain are successfully saved back. These results confirm the advantage of utilizing the proposed mutual information constraints. Our method is effective in boosting the SSL on multiple domain by leveraging information across domains.

Training domain\Test domain	Art painting		Cartoon		Photo		Sketch		*Average*	
	conv5	fc7	conv5	fc7	conv5	fc7	conv5	fc7	conv5	fc7
DeepAll-labels	64.7	58.0	85.4	89.2	83.2	85.2	76.1	74.7	77.4	76.8
Art painting (RotNet)	50.4	44.6	61.6	53.9	77.7	75.4	61.8	58.0	62.9	58.0
Cartoon (RotNet)	49.9	**51.6**	65.9	57.1	76.2	74.2	66.7	63.0	64.7	61.5
Photo (RotNet)	45.1	37.7	65.5	53.2	80.9	73.8	64.0	59.8	63.9	56.1
Sketch (RotNet)	45.4	38.7	56.3	45.0	71.9	64.1	**72.6**	59.9	61.6	51.9
PACS (DeepAll, RotNet)	54.3	43.0	**68.5**	58.7	80.9	73.5	60.9	61.9	66.2	59.3
PACS (DeepAll, AET)	53.1	43.4	67.6	42.2	77.3	72.1	66.3	50.6	66.1	52.1
Ours (DeepAll+MI, RotNet)	**55.5**	49.7	**68.5**	**66.3**	**81.6**	**77.1**	63.4	**64.7**	**67.3**	**64.5**
Ours (DeepAll+MI, AET)	**56.9**	**46.7**	**69.6**	**56.9**	**80.9**	**73.9**	**67.9**	**59.7**	**68.8**	**59.3**

Table 1: Top-1 linear classification accuracies on PACS dataset using activations from different pretraining strategies.

Training dom.\Test dom.	ImageNet		Places		*Average*	
	conv5	fc7	conv5	fc7	conv5	fc7
ImageNet-labels	47.9	55.9	37.7	41.3	42.8	48.6
Random	7.2	1.1	11.9	3.5	9.6	2.3
ImageNet	31.9	20.4	32.5	24.7	32.2	22.6
Places	30.1	10.5	**34.1**	19.7	32.1	15.1
ImageNet+Places (DeepAll)	31.6	21.4	33.2	28.5	32.4	25.0
Ours (DeepAll+MI)	**32.5**	**26.0**	33.7	**31.8**	**33.1**	**28.9**

Table 2: Top-1 linear classification accuracies on ImageNet and Places validation set using activations from different pretraining strategies.

Training domain\Test domain	PASCAL Classification	
	conv5	fc7
ImageNet-labels	80.3	83.5
Random	55.6	45.2
ImageNet	74.3	72.7
ImageNet+Places	74.5	73.8
ImageNet+PASCAL	74.8	73.2
Ours (ImageNet+Places+MI)	75.0	75.6
Ours (ImageNet+PASCAL+MI)	74.8	75.3

Table 3: Mean average precision on PASCAL VOC 2007 using activations from different pretraining strategies.

4.3. ImageNet and Places

ImageNet and Places are two large-scale image datasets with 1,281,167 and 2,448,873 images in training set, respectively, and 3,730,040 in total. As usual, we train logistic regression on top of the representations on the training set and report accuracies on the validation set [53]. We precompute the visual representations for all training images and train the logistic regression by SGD for 50 epochs. This is inspired by [22], enabling a fast evaluation and comparison between different scenarios. Table 2 shows the linear classification accuracies with the representations learned in RotNet.

Either of these two datasets has enough images to let CNNs get a reasonably good representation in SSL. When integrating more data into training, we can see that performance does not improve much. This is reflected by comparing *ImageNet* entry with *ImageNet+Places* entry on ImageNet performance, as well as comparing *Places* entry with *ImageNet+Places* entry on Places performance. The result even decreases on `conv5` layer (from 31.9 to 31.6 and from 34.1 to 33.2). Transfer learning results get large improvements (comparing *ImageNet* entry with *ImageNet+Places* entry on Places performance, and vice versa on ImageNet), which is possibly due to the explicit use of target domain images. Again, our method (*DeepAll+MI*) further outperforms naïve combination (*DeepAll*). The improvement on `fc7` layer is most significant, which represents that learning from multi-domain data and the proposed constraints are able to improve the representation outputted by $E(\cdot;\boldsymbol{\theta}_e)$ and mitigate its over-fitting toward the SSL task.[1]

4.4. PASCAL VOC

Transfer a CNN pretrained in a pre-text task on ImageNet to PASCAL dataset is a standard test in SSL experimental benchmark. The relatively small size of the training sets on PASCAL makes it a good proxy toward real-world applications. In order to show the effect of multi-domain learning, we first pre-train RotNet on ImageNet and Places, and test on PASCAL by training a linear logistic regression (a multi-label cross entropy loss) on top of the features. This is somewhat similar to domain generalization setting and will show the generalization ability of our method. We then pre-train RotNet on ImageNet and PASCAL, which evaluate the effect when we combine the target data at hand with a large dataset.

[1]These results of linear evaluation do not use data augmentation, and they are lower than those reported in RotNet [18]. For consistency and comparison with RotNet, when trained with data augmentation, our method is able to improve the performance on `conv5` from 37.3% [18] to 38.2%, and from 34.8% [18] to 36.0% on ImageNet and Places, respectively (Results of RotNet are reproduced by us and outperform those reported in [18] (36.5 and 33.7)).

Training domain\Test domain		Art painting		Cartoon		Photo		Sketch		*Average*	
	(λ_u, λ_l)	conv5	fc7	conv5	fc7	conv5	fc7	conv5	fc7	conv5	fc7
PACS (DeepAll)	(0, 0)	54.3	43.0	**68.5**	58.7	80.9	73.5	60.9	61.9	66.2	59.3
Ours (DeepInvariance)	(0.1, 0)	55.4	45.0	68.4	60.1	80.4	73.4	63.9	58.0	67.0	59.1
Ours (DeepSpecific)	(0, 0.1)	55.8	48.2	66.9	64.5	80.8	76.5	61.5	61.9	66.3	62.8
Ours (Full)	(0.1, 0.1)	**56.3**	45.4	67.6	62.9	78.1	75.8	**66.7**	**65.7**	67.2	62.5
Ours (Full)	(0.1, 1)	55.5	**49.7**	**68.5**	**66.3**	**81.6**	77.1	63.4	64.7	67.3	**64.5**
Ours (Full)	(0.1, 0.01)	53.2	46.5	67.1	62.7	80.8	75.4	63.4	60.4	66.1	61.3
Ours (Full)	(1, 0.1)	55.3	44.7	67.9	64.1	81.0	**79.3**	65.3	60.4	**67.4**	62.1
Ours (Full)	(0.01, 0.1)	55.5	46.4	66.3	64.2	79.5	76.2	64.5	64.3	66.5	62.8

Table 4: Comparison of different components and different values of parameters λ_u and λ_l in our model on PACS linear classification task.

As shown in Table 3, both these two strategies are better than ImageNet pretrained model on PASCAL. This indicates that leveraging information from large scale datasets is useful for representation learning on PASCAL. Our method further improves simple dataset combination. Pretraining on ImageNet and Places with MI constraints achieve the best results. This suggests that the proposed mutual information constraints improve the generalization ability of the representation.

4.5. Ablation studies

In this section, we further conduct experiments on PACS linear classification task to understand the impact of different components and different hyper-parameter values in our method.

4.5.1 Impact of different components

To investigate the contributions of each component in our framework, we compare the following variants: **DeepAll**: Train SSL on all available domains (naïve combination). **DeepInvariance**: Train SSL on all available domains with constraint in Eq. (7) only. **DeepSpecific**: Train SSL on all available domains with constraint in Eq. (10) only. **Full**: Our full model (Eq. (11)). Results for every variant are summarized in Table 4. We can observe the influence of each individual component:

1. *DeepInvariance* outperforms *DeepAll* mainly on conv5 layer, which can be seen from the *Average* results. This invariance constraint is imposed on the output layer of $E(\cdot;\boldsymbol{\theta}_e)$ (fc7). It seems that the enforced invariance alone does not add additional information toward the output feature on average, but the intermediate layer will encode better representation.
2. *DeepSpecific* outperforms *DeepAll* mainly on fc7 layer, demonstrated by the improved performance on it. This is the result of maintaining domain-specific information on each domain.
3. Ours (*Full*) model achieves a trade-off between *DeepInvariance* and *DeepSpecific*, and outperforms *DeepAll*. Note that the result of using domain-invariant and domain-specific constraints together does not simply equal to linearly add their effects separately. They interact in a complex way and can further improve each of them.

4.5.2 Impact of different λ value

Finally, we also evaluate the influence of the parameters λ_u and λ_l in our model. The last five rows in Table 4 summarize the results on PACS linear classification task with different settings of λ_u and λ_l. We observe that the relative strength of these two MI constraints will have an effect on the final results. Emphasizing each one of them will make the performance follow the effect of *DeepInvariance* or *DeepSpecific*. These results verify the effect of the two MI desiderata and their ability in seeking a controllable trade-off between learning domain-invariant and domain-specific information.

5. Conclusion

In this paper, we have presented an information-theoretic approach to improve the use of training data when combining datasets from multiple domains for self-supervised learning, and demonstrated its effectiveness with RotNet using popular vision datasets. Our proposed mutual information constraints explicitly exploit common, invariant as well as specific information across different domains. The learned representation seeks a trade-off between maximal invariance and maximal information maintenance, which lead to improved performance than previous results. We believe that learning from multiple domains is beneficial to representation and is a promising future direction especially for practical applications of self-supervised learning.

Acknowledgement

This work was supported in part by the Australian Research Council under Project FL-170100117, Project DP-180103424, and Project DE-180101438.

References

[1] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 159–168, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[2] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[3] David Barber and Felix Agakov. The IM algorithm: A variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 201–208, Cambridge, MA, USA, 2003. MIT Press.

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug 2013.

[6] Uta Büchler, Biagio Brattoli, and Björn Ommer. Improving spatiotemporal self-supervision by deep reinforcement learning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 797–814, Cham, 2018. Springer International Publishing.

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 139–156, Cham, 2018. Springer International Publishing.

[8] Xilun Chen and Claire Cardie. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1226–1240. Association for Computational Linguistics, 2018.

[9] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[10] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[11] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations*, 2017.

[12] Mark Dredze, Alex Kulesza, and Koby Crammer. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79(1):123–149, May 2010.

[13] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan 2015.

[14] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR.

[16] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[17] Behnam Gholami, Pritish Sahu, Ognjen (Oggi) Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach, 2019.

[18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

[20] Simon Jenni and Paolo Favaro. Self-supervised feature learning by learning to spot artifacts. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. Undoing the damage of dataset bias. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 158–171, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[22] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[24] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 577–593, Cham, 2016. Springer International Publishing.

[25] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[26] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[27] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[28] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[29] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3915–3924, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[30] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[31] T. Nathan Mundhenk, Daniel Ho, and Barry Y. Chen. Improvements to context based self-supervised learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[32] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing.

[33] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[34] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[35] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 271–279. Curran Associates, Inc., 2016.

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[37] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[38] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[39] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 506–516. Curran Associates, Inc., 2017.

[40] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[41] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015.

[43] Alice Schoenauer-Sebag, Louise Heinrich, Marc Schoenauer, Michele Sebag, Lani Wu, and Steve Altschuler. Multi-domain adversarial learning. In *International Conference on Learning Representations*, 2019.

[44] Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 1079–1086, New York, NY, USA, July 2012. Omnipress.

[45] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 2164–2173, Naha, Okinawa, Japan, 16–18 Apr 2019. PMLR.

[46] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1521–1528, Washington, DC, USA, 2011. IEEE Computer Society.

[47] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM.

[48] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[49] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[50] Yongxin Yang and Timothy M. Hospedales. A unified perspective on multi-domain and multi-task learning. In *International Conference on Learning Representations*, 2015.

[51] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[52] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 649–666, Cham, 2016. Springer International Publishing.

[53] Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[54] Shengjia Zhao, Jiaming Song, and Stefano Ermon. The information autoencoding family: A lagrangian perspective on latent variable generative models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[55] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc., 2014.