

Gradient Regularized Contrastive Learning for Continual Domain Adaptation

Shixiang Tang*,^{1,3} Peng Su*,² Dapeng Chen,³ Wanli Ouyang¹

¹ University of Sydney, Australia

² The Chinese University of Hong Kong, Hong Kong

³ Sensetime Group Limited, Hong Kong

{stan3906,wanli.ouyang}@uni.sydney.edu.au, psu@ee.cuhk.edu.hk, chendapeng@sensetime.com

Abstract

Human beings can quickly adapt to environmental changes by leveraging learning experience. However, adapting deep neural networks to dynamic environments by machine learning algorithms remains a challenge. To better understand this issue, we study the problem of continual domain adaptation, where the model is presented with a labelled source domain and a sequence of unlabelled target domains. The obstacles in this problem are both domain shift and catastrophic forgetting. We propose Gradient Regularized Contrastive Learning (GRCL) to solve the obstacles. At the core of our method, gradient regularization plays two key roles: (1) enforcing the gradient not to harm the discriminative ability of source features which can, in turn, benefit the adaptation ability of the model to target domains; (2) constraining the gradient not to increase the classification loss on old target domains, which enables the model to preserve the performance on old target domains when adapting to an in-coming target domain. Experiments on Digits, DomainNet and Office-Caltech benchmarks demonstrate the strong performance of our approach when compared to the state-of-the-art.

Introduction

Generalizing models learned from one domain (source domain) to novel domains (target domains) has been a major challenge of machine learning. The performance of the model learned on one domain may degrade significantly on other domains because of different data distribution (Luo et al. 2019; Ben-David et al. 2010; Moreno-Torres et al. 2012; Storkey 2009; Yu et al. 2020). In this work, we investigate continual domain adaptation (continual DA), where models are trained in multi-steps and only part of training samples are presented in each step. Continual DA considers the real-world setting where target domain data are acquired sequentially. As an example, autonomous driving requires adapting to scenes (target domains) in different weathers and different countries. And these training data are usually collected in different seasons: snowy scenes can only be collected in winter while rainy scenes are mostly in summer. Continual DA also considers efficiency when the model is deployed in the real world. When samples from a novel domain are acquired,

conventional DA, i.e. not continual DA, requires to use all samples collected to train the model from scratch, which is time-consuming. Continual DA can address the problem as it enables the model incrementally adapts to the new domain without losing generalization ability on old domains.

In this paper, we consider keeping the discriminative ability of source features can benefit adapting the model to different target domains. Intuitively, as the labels of source domain samples are given, the discriminative ability learned from labelled source samples can guide the adaptation to all target domains. Existing DA methods (Ganin et al. 2016; Long et al. 2017; Tzeng et al. 2017; Saito et al. 2018; Su et al. 2020; Peng et al. 2019b) cannot retrain such discriminative ability on source features because they adopt multitask learning, *i.e.* one classification loss on the source domain and one domain adaptation loss. When minimizing the multitask loss, the classification loss on the source domain may still increase, meaning the model’s discriminative ability on the source domain is weakened. In contrast, our method constrains the classification loss on the source domain non-increase (*source discriminative constraint*) in every training iteration. In this way, we maintain the discriminative ability of source features and more importantly improve the adaptation ability of the learned model to the target domain.

Furthermore, the model trained with sequential data suffers from *catastrophic forgetting*. Existing method (Bobu et al. 2018) handles *catastrophic forgetting* by incorporating a replay in the adversary training framework. However, this approach assumes that the sequential target domain shift follows some specific patterns and suffers from *catastrophic forgetting* when the assumption breaks. In contrast, when adapting the model to a new target domain, we propose to enforce the classification loss not to increase for every old target domain (*target memorization constraint*). This constraint ensures the model not to lose the generalization ability on old target domains when adapting to a new target domain.

Based on the observations above, we propose gradient regularized contrastive learning (GRCL), to tackle continual DA. GRCL leverages the contrastive loss to learn domain-invariant representations using the samples in the source domain, the old target domains and the new target domain. Two constraints, *i.e.* *source discriminative constraint* and *target memorization constraint*, are proposed when optimizing the network. Specifically, the *source discriminative constraint* is

*denotes equal contribution

formulated to constrain that the gradient of the parameters should be positively correlated to the gradient of classification loss for the source domain. And the *target memorization constraint* constrains that the gradient of the parameters should be positively correlated to the gradient of classification loss for every old target domain. The pseudo-labels involved in the *target memorization constraint* are generated by clustering (Yang et al. 2020, 2019; Guo et al. 2020; Ester et al. 1996; Kanungo et al. 2002). These pseudo-labels are high-quality because features in old target domains are discriminative and we can filter out those samples with low confidence.

To summarize, our contributions are as follows: (1) We propose a *source discriminative constraint* to improve discriminative ability of features in the target domains by preserving the discriminative ability of source features. (2) We propose a *target memorization constraint* to explicitly memorize the knowledge on old target domains. The proposed two constraints consistently improve the continual DA by over 5% compared with the baseline method on three benchmarks.

Related Works

Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) aims to transfer the knowledge from a different but related domain (source domain) to a novel domain (target domain). Various methods have been proposed, including discrepancy-based UDA approaches (Long et al. 2017; Tzeng et al. 2014; Ghifary, Kleijn, and Zhang 2014; Peng and Saenko 2018), adversary-based approaches (Liu and Tuzel 2016; Tzeng et al. 2017; Liu et al. 2018; Su et al. 2020), reconstruction-based approaches (Yi et al. 2017; Zhu et al. 2017; Hoffman et al. 2018; Kim et al. 2017) and contrastive learning based approaches (Ge et al. 2020; Park et al. 2020; Kim et al. 2020). To maintain the discriminative ability of the model to the source domain, these approaches resort to adding the task loss, e.g. classification loss for the image classification task, on the source domain data to the adaptation loss as a multitask objective function. Consequently, the task loss on the source domain data often increases even though the multitask objective function is minimized. In contrast, we explicitly enforce that the task loss on the source domain should not increase, preserving the discriminative ability of the model to the source domain. Besides, instead of setting the trade-off parameter manually in the multi-task learning, our GRCL adaptively updates the ratio of gradients in different tasks by solving an optimization problem. Such adaptive updating of the gradient is the key to maintaining the discriminative ability of source features, which, in turn, improves performance on the target domain.

Continual Learning

Continual learning (Prabhu, Torr, and Dokania 2020; Parisi et al. 2019; Zhao et al. 2020) addresses *catastrophic forgetting* in a sequence of supervised learning tasks. Popular methods can be categorized as regularization-based methods (Kirkpatrick et al. 2017; Zenke, Poole, and Ganguli 2017; Aljundi et al. 2018; Li and Hoiem 2017) and memory-based methods (Rebuffi et al. 2017; Castro et al. 2018; Hou et al. 2018, 2019). In particular, GEM (Lopez-Paz and Ranzato 2017) is

most related to our method. It uses episodic memory to store some training samples of old tasks and conducts constrained optimization to address the *catastrophic forgetting* problem. However, GEM cannot address continual DA very well, it aims to learn different classes continuously, while continual DA needs to recognize the images with the same label space but from different domains. To address the continual DA, our method highlights the importance of the source domain data and explicitly utilizes the constraints from the source domain to guide the learning of all other target domains.

Continual Domain Adaptation

When learning a sequence of unlabeled target domains, continual domain adaptation aims to achieve good generalization abilities on all seen domains (Lao et al. 2020). Massimiliano (Mancini et al. 2019) attempted to solve a specific scenario in continual DA, where no target data is available, but with metadata provided for all domains. Gong (Gong et al. 2019) proposed to bridge two domains by generating a continuous flow of intermediate states between two original domains. Several other papers (Wulfmeier, Bewley, and Posner 2018; Hoffman et al. 2014; Wang, He, and Katabi 2020; Cheung et al. 2019) presented continuous domain adaptation with the emphasis to generalize on a transitioning target domain. The approaches (Mancini et al. 2019; Gong et al. 2019; Wulfmeier, Bewley, and Posner 2018; Hoffman et al. 2014) do not explicitly target the *catastrophic forgetting* problem, while the approaches (Bobu et al. 2018) and our approach target the *catastrophic forgetting* problem. Our approach is different from the method in two aspects. First, existing works aim to address *catastrophic forgetting*, but with an implicit assumption that the domain shift follows a specific pattern, *i.e.* data shift domain gradually, e.g. gradually changing weather or lighting condition. However, our approach does not need this assumption because we explicitly set constraints on every old target domain without relying on the domain relationships. Second, compared with the closest work in (Bobu et al. 2018) which uses multitask learning with a simple replay, our GRCL emphasizes the importance of the discriminative source features and tackles *catastrophic forgetting* by strictly following the constraint that the task loss on the old target domains non-increase in every iteration.

Problem Formulation

Let $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ be the labeled dataset of source domain, where each example (x_i^s, y_i^s) is composed of an image $x_i^s \in \mathcal{X}^s$ and a label $y_i^s \in \mathcal{Y}$. Continual domain adaptation defines a sequence of adaptation tasks $\mathcal{T}_{1:N} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_N\}$. On t -th task \mathcal{T}_t , there is an *unlabeled* target domain dataset $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$. Different domains share a common label space \mathcal{Y} but have distinct data distributions. The goal is to learn a label prediction model f that can generalize well on multiple target domains $\{\mathcal{D}_1, \dots, \mathcal{D}_N\}$.

We propose two metrics to evaluate the model adapting over a stream of target domains, namely average accuracy (ACC) and average backward transfer (BWT). After the model adapts to the target domain \mathcal{D}_t , we evaluate its performance on the testing set of the new and all old target

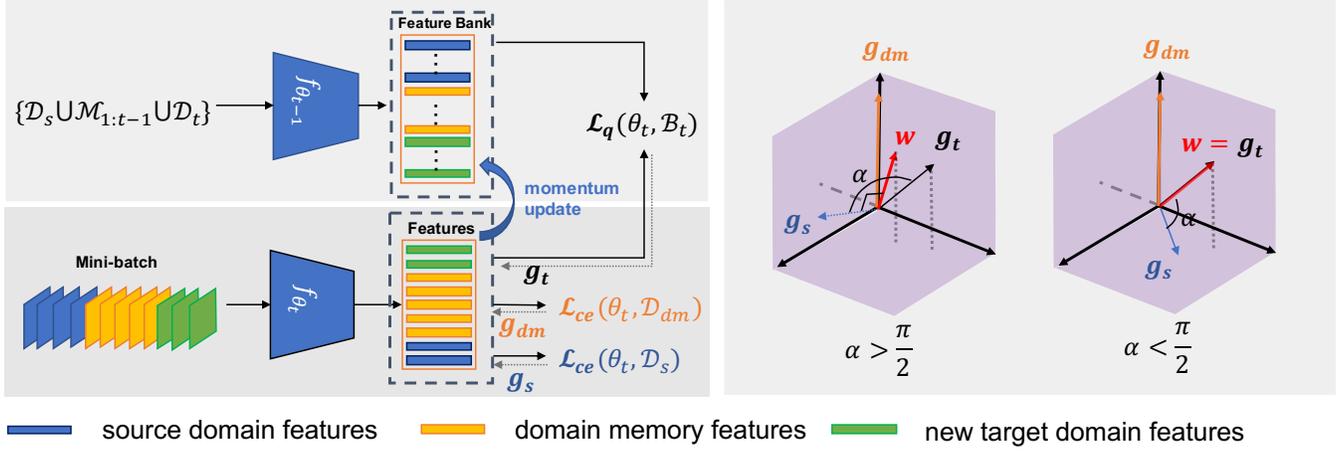


Figure 1: Gradient Regularized Contrastive Learning. Left: Schematic representation of GRCL. The feature bank is initialized by pretrained model $f_{\theta_{t-1}}$. During training, mini-batch samples are sampled from the source domain, domain-episodic memories and the new target domain. Contrastive loss, *source discriminative constraint*, *target memorization constraint* are then imposed on the mini-batch. The features of the mini-batch are involved in updating the feature bank in a momentum way. Right: The gradient regularization is utilized to enforce the gradient of contrastive loss not to increase the cross-entropy loss, which maintains the discriminative ability of feature learned by contrastive learning.

domains $\mathcal{D}_k^{test} (\forall k \leq t)$. Let $R_{t,j}$ denote the test accuracy of the model on the domain \mathcal{D}_j after adapting the model to domain \mathcal{D}_t . We use \mathcal{D}_0 to denote the source domain. ACC and BWT can be calculated as

$$\text{ACC} = \frac{1}{N} \sum_{t=0}^N R_{N,t}, \text{ BWT} = \frac{1}{N-1} \sum_{t=1}^{N-1} R_{N,t} - R_{t,t}. \quad (1)$$

The ACC represents the average performance over all domains when the model finishes all sequential adaptation tasks. BWT indicates the influence on previously observed domains $\mathcal{D}_{k < t}$ when adapting to domain \mathcal{D}_t . The negative BWT indicates that adapting to a new domain decreases the performance on previous domains. The larger these two metrics are, the better the model is.

Methodology

We propose a gradient regularized contrastive learning framework (GRCL) to tackle the challenges in continual unsupervised domain adaptation (Continual DA). When adapting the model to the t -th target domain, the baseline framework is based on contrastive learning with domain-episodic memories $\mathcal{M}_{1:t-1}$ and a feature bank \mathcal{B}_t . The key innovation of GRCL lies in two novel constraints on gradients when jointly training samples in source domain \mathcal{D}_s , domain-episodic memories $\mathcal{M}_{1:t-1}$ and the new target domain \mathcal{D}_t in a contrastive way. The *source discriminative constraint* can maintain the discriminative ability of samples in the source domain and surprisingly, in turn, improves the adaptability on the target domain. The *target memorization constraint* overcomes *catastrophic forgetting* on old target domains when adapting the model to a new target domain.

In greater detail, the training samples in each batch are sampled from the source domain \mathcal{D}_s , episodic memories $\mathcal{M}_{1:t-1}$ and the new target domain \mathcal{D}_t . These samples are trained with

the unified contrastive loss (Eq. 4) with the help of the feature bank \mathcal{B}_t by Eq. 2. *Source discriminative constraint* (Eq. 7) and *target memorization constraint* (Eq. 5) are imposed on samples of the source domain and the episodic memories in the minibatch respectively. The gradient to update the model w is computed by solving the quadratic optimization problem (Eq. 11). The whole pipeline is illustrated in Fig. 1(left).

Baseline Framework with Contrastive Learning

Contrastive learning (Wu et al. 2018; He et al. 2020; Chen et al. 2020) has recently shown the great capability of mapping images to an embedding space, where similar images are close together and dissimilar images are far apart. Inspired by this, we utilize the contrastive loss to push the target instance towards the source instances that have similar appearances with the target input. To better exploit the features from the source domain, old target domains and the new target domain, we unify these features in one feature bank and introduce a unified contrastive loss with the feature bank in detail.

Feature Bank We propose a feature bank \mathcal{B}_t to provide source features, representative old target domain features and new target domain features. We initialize the feature bank as $\mathcal{B}_t = \{k(x), \forall x \in \mathcal{D}_s \cup \mathcal{M}_{1:t-1} \cup \mathcal{D}_t\}$, where $\mathcal{M}_{1:t-1} = \{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_{t-1}\}$ and \mathcal{M}_i stores representative samples in the target domain \mathcal{D}_i . In particular, $k(x)$ is a representation of the input x and can be computed by

$$k(x) = q_{t-1}(f_{\theta_{t-1}}(x)), \quad (2)$$

where $f_{\theta_{t-1}}$ is a CNN-based encoder, and q_{t-1} is a MLP projector after adapting the model to $(t-1)$ -th target domain. All features are normalized by $\|k(x)\|_2^2 = 1$. At each training iteration, the encoded features $q_t(f_{\theta_t}(x))$ in the mini-batch will be used to update the memory bank \mathcal{B}_t by the rule:

$$k(x) \leftarrow mk(x) + (1-m)q_t(f_{\theta_t}(x)), \quad m \in [0, 1] \quad (3)$$

Unified Contrastive Loss For adapting the model to the t -th target domain, the CNN-based f_{θ_t} is initialized by $f_{\theta_{t-1}}$, the MLP projection q_t is initialized by q_{t-1} and the feature bank \mathcal{B}_t is initialized by Eq. 2. Images in each training batch are sampled from the source domain \mathcal{D}_s , the episodic memories $\mathcal{M}_{1:t-1}$ and the new target domain \mathcal{D}_t at a fixed ratio. The contrastive loss (Oord, Li, and Vinyals 2018) computed by each training batch is:

$$\mathcal{L}_{\mathbf{q}}(\theta_t, \mathcal{B}_t) = -\log \frac{\exp(\mathbf{q} \cdot \mathbf{k}^+ / \tau)}{\exp(\mathbf{q} \cdot \mathbf{k}^+ / \tau) + \sum_{\mathbf{k}^- \in \mathcal{B}_t} \exp(\mathbf{q} \cdot \mathbf{k}^- / \tau)}, \quad (4)$$

where \mathbf{q} is a general feature vector $\mathbf{q} = q_t(f_{\theta_t}(x))$ and x denotes the samples in the training batch. \mathbf{k}^+ is the positive key for \mathbf{q} and can be defined as the corresponding feature of sample x stored in \mathcal{B}_t . Features other than \mathbf{k}^+ in \mathcal{B}_t can be used as negative keys \mathbf{k}^- for \mathbf{q} . The temperature τ is empirically set as 0.07.

Source Discriminative Constraint

The contrastive loss can bridge the domain gap by attracting visually similar samples, but it may harm the discriminative ability of features in the source domain. Examples are shown in Fig. 2(b). The orange squares/circles in the purple area are pulled towards their green counterparts, hence the discriminative ability of source features is deteriorated. As labels in the source are reliable, the knowledge learnt from the source domain are valuable for all the target domains. We therefore add the *source discriminative constraint* when minimizing the contrastive loss Eq. 4, which is to regularize the $\mathcal{L}_{ce}(\theta_t, \mathcal{D}_s)$ on the source domain non-increase:

$$\mathcal{L}_{ce}(\theta_t, \mathcal{D}_s) \leq \mathcal{L}_{ce}(\theta_{t-1}, \mathcal{D}_s). \quad (5)$$

We denote w as the vector to update the model and g_s as the gradient of $\mathcal{L}_{ce}(\theta_t, \mathcal{D}_s)$. Inspired by (Lopez-Paz and Ranzato 2017), the Eq. 5 can be rephrased as

$$\langle w, g_s \rangle \geq 0, \quad (6)$$

where $\langle w, g_s \rangle$ means the inner product between the update vector and the original gradient of the classification loss on the source domain.

$$\langle w, g_s \rangle := \langle w, \frac{\partial \mathcal{L}_{ce}(\theta_t, \mathcal{D}_s)}{\partial \theta_t} \rangle \geq 0. \quad (7)$$

As illustrated in Fig.1 (right), if the angle between g_s and g_t is less than $\pi/2$, minimizing Eq. 4 by g_t will not increase the classification loss on the source domain. Therefore, we simply use $w = g_t$ to update the model parameters. If the angle between g_s and g_t is larger than $\pi/2$, updating parameters by g_t will inevitably increase the classification loss the source domain. We therefore enforce w to satisfy Eq. 7 while keeping close to the gradient of the contrastive loss g_t .

Target Memorization Constraint

An essential problem in continual domain adaptation is *catastrophic forgetting*. For this problem, *Target Memorization Constraint* is proposed to keep the classification loss for each domain-episodic memory non-increase:

$$\mathcal{L}_{ce}(\theta_t, \mathcal{M}_i) \leq \mathcal{L}_{ce}(\theta_{t-1}, \mathcal{M}_i) \quad \text{for all } i < t, \quad (8)$$

where \mathcal{M}_i is i -th domain-episodic memory. When the number of target domains increases, the computation burden for Eq. 8 will be large, because we need to solve the $t-1$ constraints of all domain-episodic memory. Alternatively, we leverage a much efficient way to approximate the Eq.8 by

$$\mathcal{L}_{ce}(\theta_t, \mathcal{M}_{1:t-1}) \leq \mathcal{L}_{ce}(\theta_{t-1}, \mathcal{M}_{1:t-1}), \quad (9)$$

where $\mathcal{M}_{1:t-1} = \cup_{i=1}^{t-1} \mathcal{M}_i$. Instead of computing the loss on each individual previous domain, Eq. 9 only computes the loss with the sampled batch of images from $\mathcal{M}_{1:t-1}$. Similarly, Eq. 9 can be rephrased as

$$\langle w, g_{dm} \rangle := \langle w, \frac{\partial \mathcal{L}_{ce}(\theta_t, \mathcal{M}_{1:t-1})}{\partial \theta_t} \rangle \geq 0. \quad (10)$$

Overall Formulation and Solution of GRCL

We combine contrastive learning (Eq. 4) with *source discriminative constraint* (Eq. 7) and *target memorization constraint* (Eq. 10) and then propose the overall objective function (Eq. 11) to obtain the final parameter update vector.

In order to incorporate gradient regularization in contrastive loss minimization, we modify the objective $\min \mathcal{L}_{\mathbf{q}}$ to $\min \|w - g_t\|_2^2$ for each iteration, where g_t is the gradient of contrastive loss and w is the gradient to update the network. The rationality behind is that to efficiently minimize contrastive loss, w need to be as close to g_t as possible under the constrains Eq. 7 and Eq. 9. Mathematically, the overall formulation for GRCL can be defined as

$$\begin{aligned} \min_w \quad & \frac{1}{2} \|w - g_t\|_2^2 \\ \text{subject to} \quad & \langle w, g_s \rangle \geq 0 \\ & \langle w, g_{dm} \rangle \geq 0. \end{aligned} \quad (11)$$

Eq. 11 is essentially a quadratic programming (QP) problem. Directly solving this problem will involve a huge number of parameters (the number of parameters in the neural network). To solve the Eq.11 efficiently, we work in the dual space, resulting in much smaller QP with only 2 variables:

$$\min_u \quad \frac{1}{2} u^\top G G^\top u + g_t^\top G^\top u \quad \text{subject to } u \geq 0, \quad (12)$$

where $G = -(g_s, g_{dm}) \in \mathbf{R}^{2 \times P}$ and we discard the constant term of $g_t^\top g_t$. The formal proof of Eq.12 is provided in Supplementary Materials. Once the solution $u^* = (u_1^*, u_2^*)$ to Eq.12 is found, we can solve the Eq.11 by $w = G^\top u^* + g_t = g_t - u_1^* g_s - u_2^* g_{dm}$. The training protocol of GRCL is summarized in Supplementary Materials.

Discussion In the context of our paper, the previous methods solve DA or continual DA as a multitask learning problem. For example, the loss function of continual DA can be formulated as $\mathcal{L} = \mathcal{L}_{\mathbf{q}}(\theta_t, \mathcal{B}_t) + \lambda_1 \mathcal{L}_{ce}(\theta_t, \mathcal{D}_s) + \lambda_2 \mathcal{L}_{ce}(\theta_t, \mathcal{M})$, where λ_1, λ_2 is the hyper-parameter to trade off the three losses.

Our gradient regularized method differs from multitask learning in two aspects. (1) GRCL ensures that parameters update will not harm the classification loss on both source domain and old target domains. In contrast, multitask learning only minimizes the overall loss, without

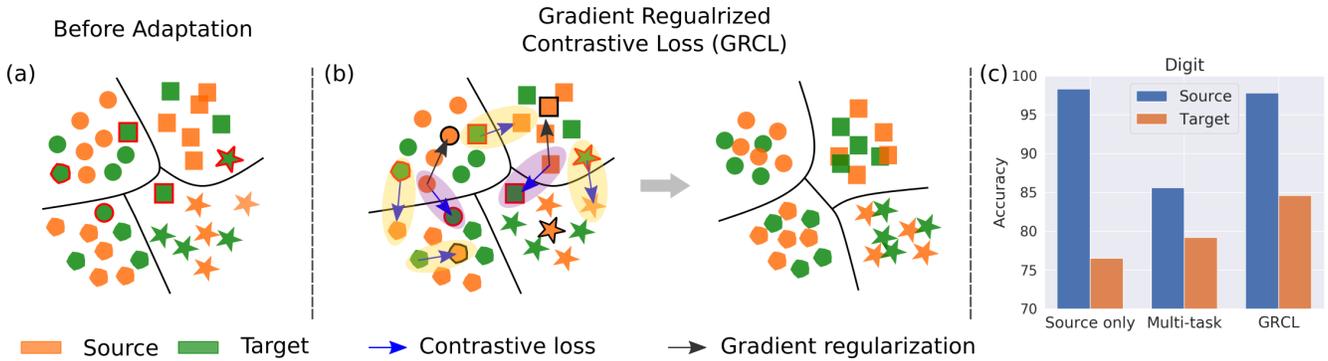


Figure 2: Illustration of GRCL. In Fig.(a,b) the shape denotes class and color denotes domain. The contrastive loss pushes target samples towards its similar ones in the source domain. However, it inevitably pushes some discriminative features towards some less discriminative features (purple area), as the contrastive loss only attracts the visually similar features. Such a problem can be verified by the severe performance degradation on the source domain (Fig.(c)). GRCL regularizes the model update not to increase the loss on source domain (black arrow), that maintains the discriminative ability of learned features.

Methods	Digits		DomainNet		Office-Caltech	
	ACC	BWT	ACC	BWT	ACC	BWT
DANN (Ganin et al. 2016)	74.56 ± 0.14	-11.37 ± 0.09	30.18 ± 0.13	-10.27 ± 0.07	81.78 ± 0.05	-8.75 ± 0.07
MCD (Saito et al. 2018)	76.46 ± 0.24	-10.90 ± 0.11	31.68 ± 0.20	-10.36 ± 0.15	82.63 ± 0.13	-8.70 ± 0.12
DADA (Peng et al. 2019b)	77.30 ± 0.19	-11.40 ± 0.04	32.14 ± 0.14	-8.67 ± 0.09	82.05 ± 0.03	-8.30 ± 0.05
CUA (Bobu et al. 2018)	82.12 ± 0.18	-6.10 ± 0.12	34.22 ± 0.16	-5.53 ± 0.14	84.83 ± 0.10	-4.65 ± 0.08
GRA	84.10 ± 0.15	-0.93 ± 0.10	35.84 ± 0.19	-1.15 ± 0.16	86.53 ± 0.11	-0.03 ± 0.03
GRCL	85.34 ± 0.10	-1.0 ± 0.03	37.74 ± 0.13	-0.67 ± 0.12	87.23 ± 0.06	0.05 ± 0.02

Table 1: ACC and BWT on three continue domain adaptation benchmarks.

source discriminative constraint or *target memorization constraint* guaranteed. (2) Trade-off parameters are different in $w = g_t + \lambda_1 g_s + \lambda_2 g_{dm}$ of multi-task learning and $w = g_t - u_1^* g_s - u_2^* g_{dm}$ of GRCL. Compared with λ_1, λ_2 that are given manually, u_1^*, u_2^* are computed by Eq. 12 and are adaptive in each iteration. Therefore, we conclude that GRCL can better balance the importance of different sub-losses during each iteration.

For the experiments on adapting the model to the first target domain, the results (Fig.2 (c)) show that the multitask loss only brings marginal improvements on the target domain but degrades the performance on the source domain. In contrast, our gradient regularized method can significantly improve the performance on the target while preserving the performance on the source domain.

Experiment

Experimental Setup

Datasets We test our method on three popular datasets.

Digits includes five digits datasets (MNIST (LeCun et al. 1998), MNIST-M (Ganin and Lempitsky 2015), USPS (Hull 1994), SynNum (Ganin and Lempitsky 2015) and SVHN (Netzer et al. 2011)). Each domain has 7,500 images for training and 1,500 images for testing. We consider a continual domain adaptation problem of SynNum → MNIST → MNIST-M → USPS → SVHN.

DomainNet (Peng et al. 2019a) is one of the largest domain adaptation datasets with approximately 0.6 million images distributed among 345 categories. Each domain randomly selects 40,000 images for training and 8,000 images for testing. Five different domains from DomainNet are used to build a continual domain adaptation task as Clipart → Real → Infograph → Sketch → Painting.

Office-Caltech (Gong et al. 2012) includes 10 categories shared by Office-31 (Saenko et al. 2010) and Caltech-256 (Griffin, Holub, and Perona 2007) datasets. Office-31 dataset contains three domains: DSLR, Amazon and Web-Cam. We consider a continual domain adaptation task of DSLR → Amazon → WebCam → Caltech.

Competing Methods We compare GRCL with five alternatives, including (1) DANN (Ganin et al. 2016), a classic domain adversarial training based method; (2) MCD (Saito et al. 2018), maximizing the classifier discrepancy to reduce domain gap; (3) DADA (Peng et al. 2019b), disentangling the domain-specific features from category identity; (4) CUA (Bobu et al. 2018), adopting an adversarial training based method ADDA (Tzeng et al. 2017) to reduce the domain shift and a sample replay loss to avoid forgetting; (5) GRA, replacing the contrastive loss 4 in GRCL with adversary loss in ADDA (Tzeng et al. 2017).

Implementation Details For fair comparison, we adopt LeNet-5 (LeCun et al. 1998) on Digits, ResNet-50 (He et al.

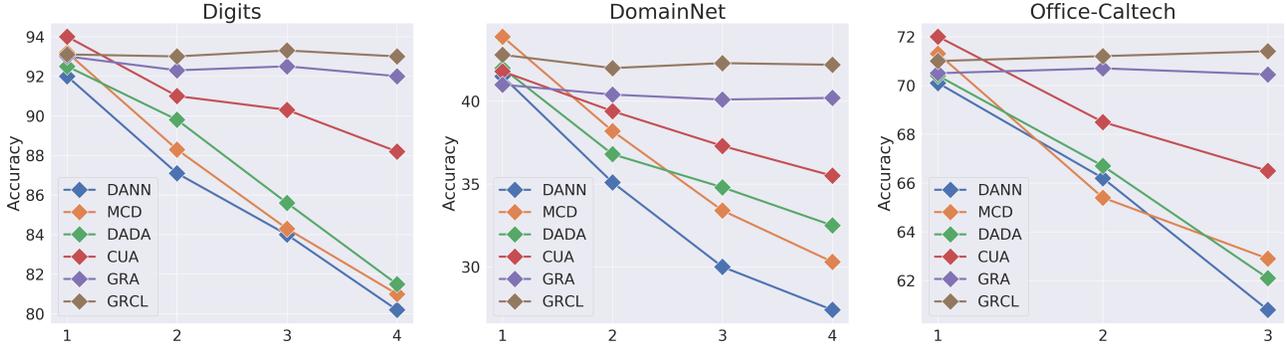


Figure 3: Evolution of classification accuracy on the first target domain as more domains are observed. Existing methods (DANN, MCD, DADA, CUA) exhibit significant performance degradation due to *catastrophic forgetting*.

2016) on DomainNet, and ResNet-18 (He et al. 2016) on Office-Caltech. The number of training and testing images is identical across different domains. For contrastive learning, we set batch size to be 256, feature update momentum to be $m = 0.5$ in Eq. 3, number of negatives to be 1024 and training schedule to be 240 epochs. The MLP head uses a hidden dimension of 2048. Following (Wu et al. 2018; He et al. 2020), the temperature τ in Eq.4 is 0.07. For data augmentation, we use random color jittering, Gaussian blur and random horizontal flip. To ensure the discriminative ability of features in domain-episodic memories, the image samples and pseudo-labels are generated by clustering with top-1024 confidence. For methods using memory, CUA, GRA and GRCL use exactly the same size of domain-episodic memory and k -means algorithm to generate pseudo labels.

Comparison with Competing Methods

Table 1 summarizes the detailed results for competing methods and GRCL on three continual DA benchmarks. Each entry in Table 1 represents the mean value and standard deviation which are computed by five runs in corresponding experiments. The larger average accuracy (ACC) reflects the better performance of our model in continual DA and larger backward transfer (BWT) reflects better ability to overcome *catastrophic forgetting*. After the model has been adapted to the final target domain, we report the ACC and BWT over the whole sequential domains.

As shown in Fig.1, GRCL consistently achieves better ACC across three benchmarks, suggesting that the model trained by GRCL owns the best generalization capability across different domains. Unsurprisingly, most methods exhibit lower negative BWT, as *catastrophic forgetting* exists. The methods using memory (CUA, GRA, GRCL) perform better than other methods without memories (DANN, MCD, DADA) by 2% – 5% on ACC and 3% – 5% on BWT. The great improvement on BWT by memory-based methods highlights the importance of memory in the continual DA to overcome *catastrophic forgetting*.

Among the memory-based methods, GRA and GRCL leverage *memorization constraint* while CUA uses replay

buffer and multitask loss, *i.e.*,

$$\mathcal{L} = \mathcal{L}_q(\theta_t, \mathcal{B}_t) + \lambda_1 \mathcal{L}_{ce}(\theta_t, \mathcal{D}_s) + \lambda_2 \mathcal{L}_{ce}(\theta_t, \mathcal{M}). \quad (13)$$

GRA and GRCL achieve significantly better BWT on three benchmarks, suggesting the effectiveness of gradient constraints for combating *catastrophic forgetting*. GRCL consistently achieves better ACC than GRA across all benchmarks. It is because that GRCL utilizes all the samples from domain memory (cached the samples from all previously observed domains) in contrastive loss to bridge the domain gap, while GRA only uses source domain and current target domain in adversarial loss to learn domain-invariant features.

Fig.3 depicts the evolution of classification accuracy on the first target domain as more domains are adapted to. The accuracy of the first target domain significantly drops in memory-free methods (DANN, MCD, CUA) but the accuracy still maintains in memory-based methods. GRCL consistently exhibits minimal forgetting and even positive backward transfer on *Office-Caltech* benchmark.

Ablation Study

We analyze the effectiveness of the individual component of GRCL in Table 2. *Src.*: the model is trained on the source domain and then tested on different target domains. *Crt.+Src.*: the model is formulated as a multitask problem but without supervision of classification loss on old target domains (Eq. 13, $\lambda_2 = 0$). *Crt.+Src.+Mem.*: the model is formulated as *Crt.+Src.* plus supervision on domain-episodic memories (Eq. 13). *Crt.+SDC*: the model is formulated as a contrastive learning problem (Eq. 4) regularized by *Source Discriminative Constraint* (Eq. 7). *Crt.+SDC+TMC(GRCL)*: the model is formulated as a contrastive learning problem (Eq. 4) regularized by *Source Discriminative Constraint* (Eq. 7) and *Target Memorization Constraint* (Eq. 10).

Importance of Contrastive Learning Contrastive learning aims to bridge the gap between different domains. As is shown in Table 2, *Crt.+Src.* outperforms *Src.* by about 2.5%, which shows contrastive learning can bridge the domain gap by attracting visually similar samples. The BWT in

Methods	DomainNet		Office-Caltech	
	ACC	BWT	ACC	BWT
Src.	29.18 ± 0.05	-	80.03 ± 0.05	-
<i>Multitask Training</i>				
Crt.+Src.	31.53 ± 0.15	-8.27 ± 0.07	82.57 ± 0.03	-7.46 ± 0.07
Crt.+Src.+Mem.	35.23 ± 0.13	-4.03 ± 0.11	85.37 ± 0.03	-3.16 ± 0.06
<i>Gradient Regularized</i>				
Crt.+SDC	33.98 ± 0.13	-7.65 ± 0.05	84.75 ± 0.06	-6.88 ± 0.04
Crt.+SDC+TMC	37.74 ± 0.13	-0.67 ± 0.12	87.23 ± 0.06	0.05 ± 0.02

Table 2: Ablation studies of GRCL on individual components.

memory size	256	512	1024	2048
<i>Digits</i>	83.00	84.12	85.34	85.41
<i>DomainNet</i>	33.28	35.75	37.74	37.83

Table 3: ACC as a function of memory size.

Crt.+Src. is higher than other existing memory-free method, *i.e.*, DANN, MCD and DADA, which indicates contrastive loss has the same ability to overcome *catastrophic forgetting*. We attribute this to use memory features in \mathcal{B}_t for contrastive learning while no pseudo-labels are involved. Table 3 and Table 4 shows the ACC of GRCL with various sizes of domain-episodic memories and different training epochs. Because contrastive learning naturally benefits from larger memory banks and longer training schedules (Chen et al. 2020), GRCL gets consistent improvements with their conclusions.

Importance of Source Discriminative Constraint *Source Discriminative Constraint* aims to restore the discriminative ability of features in the source domain and to exploit such discriminative ability to better adapt the model to the target domains. Comparing with *Crt.+Src.* and *Crt.+SDC*, we can see the ACC improves ($\approx 2.2\%$) more than BWT does ($\approx 0.8\%$). The result shows that the improvement of ACC by *Source Discriminative Constraint* mainly results from the model’s better adaptation ability to new target domains instead of overcoming *catastrophic forgetting*. Therefore, we conclude *Source Discriminative Constraint* which keeps the discriminative ability of the features in the source domain can benefit domain adaptation to the target domains.

Importance of Target Memorization Constraint *Target Memorization Constraint* aims to remember knowledge in old target domains when the model adapts to a new target domain. Comparing with *Crt.+SDC* and *Crt.+SDC+TMC* in Table 2, we can see BWT improves by around 6% and ACC improves by around 3.5%. Therefore, we conclude that ACC improvements are mainly from overcoming *catastrophic forgetting* by *Target Memorization Constraint*. Comparing with *Crt.+Src.+Mem.*, *Crt.+SDC+TMC* outperforms ACC by over 2% and BWT by about 4%. The improvement on BWT verifies that gradient constraints learning is more effective to restore old knowledge than multitask learning.

training epoch	120	180	240	300
<i>Digits</i>	80.10	83.46	85.34	85.38
<i>DomainNet</i>	34.80	36.50	37.74	38.16
<i>Office-Caltech</i>	80.93	84.70	87.23	87.28

Table 4: ACC as a function of training epoch.

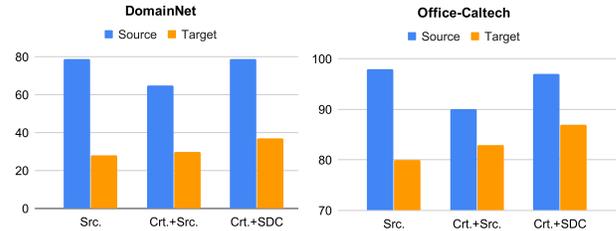


Figure 4: Comparison among *Src.*, *Crt.+Src.* and *Crt.+SDC*. The performance on target represents the averaged accuracy over all different target domains.

Effectiveness of Source Discriminative Constraint on Conventional UDA We show the proposed *Source Discriminative Constraint* outperforms the popular multitask learning in conventional UDA, *i.e.* not Continual DA. We compare three different methods: (1) *Src.*; (2) *Crt.+Src.*; (3) *Crt.+SDC*. The λ_1 in Eq.13 uses the best value obtained via grid search and $\lambda_2 = 0$. The SynNum, Clipart and DSLR are used as the source domain for *Digits*, *DomainNet* and *Office-Caltech* dataset respectively. Rather than ACC and BWT, we evaluate the performance by a different metric for UDA. We report the averaged classification accuracy on adapting the model from the source domain to the different target domains. As shown in Fig.4, *Src.* performs well on the source domain, but worse on the target domain, due to the domain gaps. *Crt.+Src.* improves the performance on the target domain but has a significant adverse effect on the performance on the source domain. *Crt.+SDC* can improve the performance even greater than *Crt.+Src.* on the target domain while maintaining the accuracy on the source domain simultaneously. Therefore, we conclude that because of the *source discriminative constraint*, the discriminative ability of source features is maintained and can further benefit adaptation to the target domain in UDA.

Conclusion

This work studies the problem of continual DA, which is one major challenge in the deployment of deep learning models. We propose Gradient Regularized Contrastive Learning (GRCL) to jointly learn both discriminative and domain-invariant representations. At the core of our method, gradient regularization maintains the discriminative ability of feature learned by contrastive loss and overcomes *catastrophic forgetting* in the continual adaptation process. Our experiments demonstrate the competitive performance of GRCL against the state-of-the-art.

References

- Aljundi, R.; Babiloni, F.; Elhoseiny, M.; Rohrbach, M.; and Tuytelaars, T. 2018. Memory aware synapses: Learning what (not) to forget. In *ECCV*.
- Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning* 79(1-2): 151–175.
- Bobu, A.; Tzeng, E.; Hoffman, J.; and Darrell, T. 2018. Adapting to Continuously Shifting Domains. In *ICLR Workshop*.
- Castro, F. M.; Marín-Jiménez, M. J.; Guil, N.; Schmid, C.; and Alahari, K. 2018. End-to-end incremental learning. In *ECCV*.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.
- Cheung, B.; Terekhov, A.; Chen, Y.; Agrawal, P.; and Olshausen, B. 2019. Superposition of many models into one. In *NeurIPS*.
- Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*.
- Ganin, Y.; and Lempitsky, V. 2015. Unsupervised Domain Adaptation by Backpropagation. In *ICML*.
- Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR*.
- Ge, Y.; Chen, D.; Zhu, F.; Zhao, R.; and Li, H. 2020. Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID.
- Ghifary, M.; Kleijn, W. B.; and Zhang, M. 2014. Domain adaptive neural networks for object recognition. In *Pacific Rim international conference on artificial intelligence*.
- Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*.
- Gong, R.; Li, W.; Chen, Y.; and Gool, L. V. 2019. Dlow: Domain flow for adaptation and generalization. In *CVPR*.
- Griffin, G.; Holub, A.; and Perona, P. 2007. Caltech-256 object category dataset.
- Guo, S.; Xu, J.; Chen, D.; Zhang, C.; Wang, X.; and Zhao, R. 2020. Density-Aware Feature Embedding for Face Clustering. In *CVPR*.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. *CVPR*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.
- Hoffman, J.; Guadarrama, S.; Tzeng, E. S.; Hu, R.; Donahue, J.; Girshick, R.; Darrell, T.; and Saenko, K. 2014. LSDA: Large scale detection through adaptation. In *NIPS*.
- Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*. PMLR.
- Hou, S.; Pan, X.; Change Loy, C.; Wang, Z.; and Lin, D. 2018. Lifelong learning via progressive distillation and retro-spection. In *ECCV*.
- Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a Unified Classifier Incrementally via Rebalancing. In *CVPR*.
- Hull, J. J. 1994. A database for handwritten text recognition research. *PAMI*.
- Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; and Wu, A. Y. 2002. An efficient k-means clustering algorithm: analysis and implementation. *PAMI*.
- Kim, D.; Saito, K.; Oh, T.-H.; Plummer, B. A.; Sclaroff, S.; and Saenko, K. 2020. Cross-domain Self-supervised Learning for Domain Adaptation with Few Source Labels. *arXiv preprint arXiv:2003.08264*.
- Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*.
- Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *PNAS*.
- Lao, Q.; Jiang, X.; Havai, M.; and Bengio, Y. 2020. Continuous Domain Adaptation with Variational Domain-Agnostic Feature Replay.
- LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.
- Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *PAMI*.
- Liu, A. H.; Liu, Y.-C.; Yeh, Y.-Y.; and Wang, Y.-C. F. 2018. A unified feature disentangler for multi-domain image translation and manipulation. In *NIPS*.
- Liu, M.-Y.; and Tuzel, O. 2016. Coupled generative adversarial networks. In *NIPS*.
- Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *ICML*.
- Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. In *NIPS*.
- Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*.
- Mancini, M.; Bulò, S. R.; Caputo, B.; and Ricci, E. 2019. Adagraph: Unifying predictive and continuous domain adaptation through graphs. In *CVPR*.
- Moreno-Torres, J. G.; Raeder, T.; Alaiz-Rodríguez, R.; Chawla, N. V.; and Herrera, F. 2012. A unifying view on dataset shift in classification. *Pattern recognition*.
- Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop*.

- Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* .
- Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks* doi:<https://doi.org/10.1016/j.neunet.2019.01.012>.
- Park, C.; Lee, J.; Yoo, J.; Hur, M.; and Yoon, S. 2020. Joint Contrastive Learning for Unsupervised Domain Adaptation. *arXiv preprint arXiv:2006.10297* .
- Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019a. Moment Matching for Multi-Source Domain Adaptation. In *ICCV*.
- Peng, X.; Huang, Z.; Sun, X.; and Saenko, K. 2019b. Domain Agnostic Learning with Disentangled Representations. In *ICML*.
- Peng, X.; and Saenko, K. 2018. Synthetic to real adaptation with generative correlation alignment networks. In *WACV*.
- Prabhu, A.; Torr, P.; and Dokania, P. 2020. GDumb: A Simple Approach that Questions Our Progress in Continual Learning. In *ECCV*.
- Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *CVPR*.
- Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *ECCV*.
- Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*.
- Storkey, A. 2009. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning* .
- Su, P.; Wang, K.; Zeng, X.; Tang, S.; Chen, D.; Qiu, D.; and Wang, X. 2020. Adapting Object Detectors with Conditional Domain Normalization. In *ECCV*.
- Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*.
- Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* .
- Wang, H.; He, H.; and Katabi, D. 2020. Continuously Indexed Domain Adaptation. *arXiv preprint arXiv:2007.01807* .
- Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*.
- Wulfmeier, M.; Bewley, A.; and Posner, I. 2018. Incremental adversarial domain adaptation for continually changing environments. In *ICRA*.
- Yang, L.; Chen, D.; Zhan, X.; Zhao, R.; Loy, C. C.; and Lin, D. 2020. Learning to Cluster Faces via Confidence and Connectivity Estimation. In *CVPR*.
- Yang, L.; Zhan, X.; Chen, D.; Yan, J.; Loy, C. C.; and Lin, D. 2019. Learning to Cluster Faces on an Affinity Graph. In *CVPR*.
- Yi, Z.; Zhang, H.; Tan, P.; and Gong, M. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*.
- Yu, S.; Li, S.; Chen, D.; Zhao, R.; Yan, J.; and Qiao, Y. 2020. COCAS: A Large-Scale Clothes Changing Person Dataset for Re-identification. In *CVPR*.
- Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *ICCV*.
- Zhao, B.; Tang, S.; Chen, D.; Bilen, H.; and Zhao, R. 2020. Continual Representation Learning for Biometric Identification. *arXiv preprint arXiv:2006.04455* .
- Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.