A Review of Single-Source Deep Unsupervised Visual Domain Adaptation

Seri Lee
Computer Science and Engineering
Seoul National University
Seoul, Republic of Korea
sally20921@snu.ac.kr

***Abstract*—Large-scale labeled training datasets have enabled deep neural networks to excel across a wide range of benchmark vision tasks. However, in many applications, it is prohibitively expensive and time-consuming to obtain large quantities of labeled data. To cope with limited labeled training data, many have attempted to directly apply models trained on a large-scale labeled source domain to another sparsely labeled or unlabeled target domain. Unfortunately, direct transfer across domains often performs poorly due to the presence of *domain shift* or *dataset bias*. Domain adaptation is a machine learning paradigm that aims to learn a model from a source domain that can perform well on a different (but related) target domain. In this paper, we review the latest single-source deep unsupervised domain adaptation methods focused on visual tasks and discuss new perspectives for future research. We begin with the definitions of different domain adaptation strategies and the descriptions of existing benchmark datasets. We then summarize and compare different categories of single-source domain adaptation methods, including discrepancy-based methods, adversarial discriminative methods, adversarial generative methods, and self-supervision-based methods. Finally, we discuss future research directions with challenges and possible solutions.**

***Index Terms*—Domain adaptation, discrepancy-based methods, adversarial learning, self-supervised learning, transfer learning**

I. Introduction

In the last decade, deep neural networks (DNNs) have achieved significant progress in various computer vision tasks where large-scale labeled training data are available. However, in many applications, it is difficult to obtain a large amount of labels, as manual annotation is expensive and time-consuming.

An alternative solution is to train a model on another related large-scale source domain with labels (*e.g.* a simulation domain) and apply it to the unlabeled target domain (*e.g.* a real-world domain). However, due to the presence of *domain shift* or *dataset bias* [1], such a direct transfer might not perform well.

One may argue that the pre-trained source models can be fine-tuned in the target domain. However, fine-tuning still requires considerable quantities of labeled training data, which may not be available for many applications.

A. Domain Adaptation in context of other sample-efficient learning methods

Domain adaptation techniques were introduced to address the domain shift between source and target domains and for this reason, they have recently attracted significant interest in both academia and industry. *Domain adaptation* (DA), also known as *domain transfer*, is a specialized form of transfer learning that aims to learn a model from a labeled source domain that can generalize well to a different (but related) unlabeled or sparsely labeled target domain. It belongs to sample efficient learning class, together with zero-shot learning, few-shot learning, and self-supervised learning.

We briefly compare domain adaptation with other methods in this category. While the unsupervised domain adaptation (UDA) does not require the annotations of the target data, it usually needs a sufficient number of unlabeled target samples to train the model. Compared to UDA, zero-shot learning does not need either the target data annotations or the unlabeled target samples. However, existing methods often require some auxiliary information, such as the attributes of the images, or the description of the classes. Further, zero-shot learning is trained on known classes and tested on unknown classes, which demands the model to generalize from known classes to unknown classes. Since the known classes and the unknown classes are from different distributions, there is no concept of domain shift in zero-shot learning. Different from zero-shot learning, DA deals with the same learning tasks on different domains.

Self-supervised learning (SSL) is a learning paradigm that captures the intrinsic patterns and properties of input data without using human-provided labels. The basic idea of SSL is to construct some auxiliary tasks solely based on the data itself without using human-annotated labels and force the network to learn meaningful representations by performing the auxiliary task well. Typical self-supervised learning approaches generally involve two aspects: constructing auxiliary tasks and defining loss functions. The auxiliary tasks are designed to encourage the model to learn meaningful representations of input data. The loss functions are defined to measure the difference between a model's prediction and a fixed target, the similarities of sample pairs in a representation space (*e.g.* contrastive loss), or the difference between probability distributions (*e.g.* adversarial loss). Compared with domain adaptation, SSL does not specifically address the domain shift problem between different domains.

B. Domain Adaptation Challenges

Albeit DA is a very effective method, it is not without blemish. The main challenge for single-source UDA is domain shift, *i.e.*, the difference between the source and target distributions that leads to unreliable predictions on the target domain. Typically, three types of domain shift are considered: covariate shift, label shift, and concept drift.

The presence of domain shift causes the direct transfer of models trained on the source domain to perform poorly on the target domain.

II. Domain Adaptation Taxonomy

We introduce a standard definition of the variables and models to enable effective comparisons and classification. Let x and y respectively denote the input data and output label, drawn from a specific domain probability distribution $P(x, y)$. In typical domain adaptation, there is one source domain and one target domain. Suppose the source data and corresponding labels drawn from the source distribution $P_S(x, y)$ are X_S and Y_S respectively, and the target data and corresponding labels drawn from the target distribution $P_T(x, y)$ are X_T and Y_T respectively. The corresponding marginal distributions include $P_S(x)$, $P_S(y)$, $P_T(x)$, $P_T(y)$, and conditional distributions include $P_S(x|y)$, $P_S(y|x)$, $P_T(x|y)$, $P_T(y|x)$. Three typical sources of variation between the two domains considered in the literature include:

- covariate shift, where $P_S(y|x) = P_T(y|x)$ for all x, but $P_S(x) \neq P_T(X)$;
- label shift, where $P_S(x|y) = P_T(x|y)$ for all y, but $P_S(y) \neq P_T(y)$;
- concept drift, where $P_S(x, y) \neq P_T(x, y)$.

In addition, the source dataset is $D_S = \{X_S, Y_S\} = \{(x_S^i, y_S^i)\}_{i=1}^{N_S}$, the target dataset is $D_T = \{X_T, Y_T\} = \{(x_T^j, y_T^j)\}_{j=1}^{N_T}$, where N_S and N_T are the number of source samples and target samples respectively, $x_S^i \in \mathbb{R}^{d_S}$ and $x_T^j \in \mathbb{R}^{d_T}$ are referred as observations in the source domain and the target domain, and y_S^i and y_T^j are the corresponding class labels.

Suppose the number of labeled target samples is N_{TL}; then, the DA problem can be classified into different categories:

- unsupervised DA, when $N_{TL} = 0$;
- fully supervised DA, when$N_{TL} = N_T$;
- semi-supervised DA, otherwise.

Suppose the number of labeled source samples is N_{SL}; then, DA can be classified into:

- strongly supervised DA, when $N_{SL} = N_S$;
- weakly supervised DA, otherwise.

Based on the representations, d_S and d_T, of the source and target domains (*e.g.* images vs text), we can classify DA into:

- homogeneous DA, when $d_S = d_T$;
- heterogeneous DA, otherwise.

Although without labels, the target data is usually available during training in typical DA. If the target data is also unavailable, we often denote this task as domain generalization or zero-shot DA. Therefore, we have:

- *domain adaptation*, when X_T is available during training;
- *domain generalization* or *zero-shot* DA, when X_T is unavailable during training.

We focus on the review of recent unsupervised domain adaptation (UDA) methods under homogeneous, single-source, single-target, strongly-supervised, and closed-set settings, *i.e.* $N_{TL} = 0, d_S = d_T, N_S = 1, N_T = 1, N_{SL} = N_S, C_S = C_T$. The goal is to learn a model f that can correctly predict a sample from the target domain based on labeled $\{X_S, Y_S\}$, and unlabeled $\{X_T\}$.

III. Single-Source DUDA

In this section, we first introduce a theoretical view for domain adaptation. Second, we summarize different categories of single-source DUDA. Finally, we compare the advantages and disadvantages of these methods.

A. Theory Brief

[1] formalizes the intuitive notion that reducing the two distributions while ensuring a low error on the source domain, yields accurate results in the target domain. Further, the theory justifies the basis of many recent DA algorithms that learn domain-invariant representations, using either domain adversarial classifier or discrepancy-based approaches.

B. Discrepancy-based Methods

Discrepancy-based methods explicitly measure the discrepancy between the source and target domains on corresponding activation layers of the two network streams. Long et al. designed a Deep Adaptation Network, where the discrepancy is defined as the sum of the multiple kernel variant of maximum mean discrepancies (MK-MDD) betweenthe fully connected (FC) layers. Sun et al. proposed correlation alignment (CORAL) to minimize domain shift by aligning the second-order statistics of the source and target features of the last FL layer. Apart from the CORAL loss on the last FL layer, Zhuo et al. also incorporated the CORAL loss on the last convolutional (conv) layer. To deal with the high dimension of convolutional layer activations, activation-based attention mapping is employed to distill it into low dimensional representations. The CORAL losses on both the last convolutional layer and the last FC layer are combined.

Wu et al. studied the UDA problem for 3D liDAR point cloud segment from synthetic data to real world data. At every batch of training, in addition to the focal loss to learn semantics from the point cloud on the synthetic batch, they employed the geodesic distance to penalize discrepancies between batch statistics from two domains. In recent papers, Zellinger et al. proposed to match the higher order central moments of probability distributions by means of order-wise moment differences. They utilized the equivalent representation of probabiliby distributions by moment sequences to define a new distance function, called Central Monment Discrepancy (CMD). Chen et al. explored the benefits of using higher-order statistics (in this case mainly third-order and fourth-order statistics) for domain matching. They proposed a Higher-order

Moment Matching (HoMM) method, and further extended the HoMM into reproducing kernel Hilbert spaces (RKHS). Some other types of divergence are also designed to align the source and target domains. Lee et al. desinged sliced Wasserstein discrepancy (SWD) to capture the natural notion of dissimilarity between the outputs of task-specific classifiers. It provides a geometrically meaningful guidance to detect target samples that are far from the support of the source and enables efficient distribution alignment in an end-to-end trainable fashion. Roy et al. proposed domain alignment layers which implement feature whitening for the purpose of matching source and target feature distributions. Additionally, they leveraged the unlabeled target data by proposing the Min-Entropy Consensus loss, which regularizes training while avoiding the adoption of many user-defined hyper-parameters.

Instead of explicitly modeling the discrepancy between the source and the target domains, some papers implicitly minimize domain discrepancy by aligning the Batch Normalization (BN) statistics. Li et al. proposed to adopt domain specific normalization for different domains. The proposed Adaptive BN (AdaBN) replaces the moving average mean and variance of all BN layers in the task network trained on the source domain with the mean and variance estimated from the target mini-batches. AdaBn and other DUDA methods define a prior on which layers should be adapted. Instead, Cariucci et al. propsed to learn automatically which layers of the network should be aligned and the corresponding alignment degree. The Auto-Domain Alignment Layer (AutoDIAL) is embedded multiple times to align the learned feature representations of the source and target domains at different levels. These BN-based mehtods have fewer parameters to tune, higher computational efficiency, and competitive performance.

The methods described above measure the domain discrepancy at the domain level, which neglects the information concerning the classes from which the samples are drawn and thus may lead to misalignment and poor performance. Kang et al. proposed Contrastive Adaptation Network, which optimizes a new metric, Contrastive Domain Discrepancy (CDD), by explicitly minimizing the intra-class domain discrepancy. The source and target samples of the same underlying class are drawn closer, while the samples from different classes are pushed apart. Pan et al. recently proposed Transferrable Prototypical Networks, which perform domain alignment such that prototypes for each class in the source and target domains are close in the embedding space and the predictions from prototypes separately on source and target data are similar.

Most of the papers mentioned above consider aligning the marginal distributions in the feature space. When confronted with complex tasks, these approaches would fail when the label distributions are drastically different between source and target domains. The joint alignment of distributions $D_S = (X_S, Y_S)$ and $D_T = (X_T, Y_T)$ is considered under the assumptions that $P_S(x) \neq P_T(x)$ and $P_S(y|x) \neq P_T(y|x)$. The joint distribution across domains is projected to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ and MMD is used as the distance metric. During the joint distribution alignment, the distribution shift $P_S(x)$ and $P_T(x)$, $P_S(y|x)$ and $P_T(y|x)$ are significantly reduced.

The above methods all adopt weight-sharing between the two streams of the Siamese architecture that attempts to reduce the impact of domain shift by learning domain-invariant features. However, domain invariance may be detrimental to discriminative power. On the contrary, Rozantsev et al. proposed to explicitly model the domain shift and relaxed the weight-sharing constraint to a linear correlation. They jointly optimized a weight regularizer, representing the loss between corresponding layers of the two streams, and an unsupervised regularizer, encoding the MMD measure and favoring similar distributions of the source and target representations.

C. Adversarial Discriminative Models

Adversarial discriminative models usually employ an adversarial objective with respect to a domain discriminator to encourage domain confusion. In the early-stage of adversarial discriminative models, the domain adversarial training of neural networks is proposed to learn domain invariant and task discriminative representations. It is directly derived from the seminal theoretical works of Ben et al. and directly optimizes the $\mathcal{H}$-divergence between source and target. By deriving the generalization bound on the target risk and obtaining an empirircal formulation of the objective, Ganin et al. proposed the Domain-Adversarial Neural Networks (DANN) algorithm. From this point of view, the adversarial discriminative models are originally similar to the discrepancy-based models. Recently, a couple of adversarial discriminative models were proposed with different algorithms and network architectures, thus differing from the discrepancy-based methods.

Suppose m_S and m_T are the representation mappings of the source and target domains, respectively, and d is a domain discriminator, which classifies whether a data point is drawn from the source or the target domain. The adversarial discriminator is trained typically based on adversarial loss L_{a_d}. The loss L_{a_m} used to train representation mapping is different in existing methods. The Domain-Adversarial Neural Network optimizes the mapping to minimize the discriminator loss directly $L_{a_m} = -L_{a_d}$, which might be problematic, since the discriminator converges quickly during training, causing the gradients to vanish. A gradient *reversal* layer was proposed to achieve domain adversarial training with a single feed-forward network with standard backpropagation and stochastic gradient descent. Tzeng et al. proposed Adversarial Discriminative Domain Adaptation (ADDA), using an inverted label GAN loss to split the optimization process into two independent objectives for generator and discriminator.

Besides aligning marginal distributions, several methods also align conditional or joint distributions. Long et al. considered aligning conditional distribution across domains, and proposed Conditional Domain Adversarial Network (CDAN). Based on DANN, they used conditional discriminator D $(f \cdot g)$ with imporved discriminability, where f is feature extractor and g is classifier, to capture the cross-covariance between feature representations and classifier predictions. To extend

joint distribution alignment, Du et al. used dual adversarial strategy to train a dual-discriminator to pit against each other. Cicek et al. also aimed for joint distribution $P(d, y)$ alignment over domain d and label y by a joint preoictor and aligned its output with classifier's prediction. After analyzing the drawbacks of feature-level alignment methods, Liu et al. proposed Transferable Adversarial Training (TAT), not only adapting feature representations from different domains, but also generating transferable examples to make the classifier learn a more robust decision boundary.

Xu et al. explored two common limitations in current adversarial methods. Sampling from source and target domains separately is insufficient to ensure domain-invariance at the whole latent space, and does not give the discriminator a hard label to judge real and fake samples. They proposed a mixed version of the discriminator to guarantee domain-invariance in a more continuous latent space, thus improving the robustness of the models performance. Chen et al. adopted the concept of self-training. They analyzed the noise of pseudo-labels in the confusion matrix and proposed correspondingly an adversarial-learned loss to accurately estimate the confusion matrix. In this way, their propsoed method inherits the strength of both adversarial learning and self-training paradigm.

Hoffman et al. made the very first effort for domain adaptation in semantic segmentation. They employed a pixel-level adversarial loss to enforce the network to extract domain-invariant features for semantic segmentation and further applied category-specific constraints, *e.g.* pixel percentage histograms. Instead of only performing domain adversarial globally, Chen et al. propsoed to perform feature alignment jointly at the global and class-wise levels by leveraging soft labels from source and target-domain data. Hong et al. propsoed to learn a conditional generator to transform features of synthetic images to real-image like features. However, the proposed method is network-specific and only applied to the FCN model structure. While previous works mostly perform feature alignment in the middle of a network, Tsai et al. adopted adversarial learning in the output space. To further enhance the adapted model, they constructed a multi-level adversarial network to effectively perform output space domain adaptation at different feature levels. To address DA in object detection, they applied multi-level domain alignment with adversarial training, and Chen et al. performed domain alignment on both image level and instance level. Weak alignment model was introduced which focused the adversarial alignment loss on images that are globally similar, putting less emphasis on aligning images that are globally dissimilar. Zhu et al. instead proposed to perform adversarial learning on region level for domain alignment. Recently, Zhent et al. proposed a coarse-to-fine feature adaptation approach for object detection. Different from image level or instance level feature alignment, foreground regions are extracted by attention mechanism, and aligned through multi-layer adversarial learning. Based on prototypical representations, Hu et al. recently proposed a Prototypical Adversarial Learning scheme to align both feature representations and intermediate prototypes across domains.

D. Adversarial Generative Models

Adversarial generative models combine the domain discriminative model with a generative component generally based on generative adversarial nets (GANs), which include a generator g and a discriminator d. g takes random noise z as input to generate a virtual image, while d takes the output of g and real images x as input to classify whether an image is real or generated. The learning process is that d tries to maximize the probability of correctly classifying real and generated images, while g tries to generate images to maximize the probability of d making a mistake. The Coupled Generative Adversarial Networks (CoGAN) is composed of a tuple of GANs, and each is responsible for synthesizing images in one domain. CoGAN corresponds to a constrained min-max game of two teams, each with two players.

Instead of taking random noise as input, the generator of more recent GAN based methods is usually conditioned on the source data. Shrivastava et al. proposed simulated and unsupervised learning (SimGAN) to improve the realism of a simulator's output suing unlabled real data. The discriminator's loss in SimGAN is the same as is used in a traditional GAN, while a self-regularization loss is added in teh refiner (generator) loss to ensure that the refined data do not change much, which aims to preserve the annotation information. THe generator in the pixel-level DA is conditioned on both a noise vector and an image from the source domain. To penalize large low-level differences between the source and generated images for foreground pixels only, the model learns to minimize a masked Pairwise Mean Squared Error (PMSE) which only calculates the masked pixels (foreground) of the source and teh generated images. Sankaranarayanan et al. proposed to learn a mutual feature embedding for source and target images, and to generate intermediate domain images from source and target embeddings. They also designed a multi-class discriminator to encourage the model to extract more class-discriminative features.

To overcome the under-constrained nature of GAN, they proposed CycleGAN with a cycle-consistency constraint. Based on the CycleGAN loss, some effective adaptation methods were introduced. Hoffman et al. proposed discriminatively-trained Cycle-Consistent Adversarial Domain Adaptation (CyCADA), which adapts representations at both the pixel-level and feature-level, enforces cycle-consistency, and leverages a task loss to perform semantic segmantation adaptation. Similarly, Russo et al. introduced bi-directional image translation mapping and proposed class-consistency loss. While CycleGAN can only translate low-level appearance, *e.g.* texture, they realized multiple view-point transformation combining with key-point detection network. Similarly, Tzeng et al. perfomred domain adaptation on object detection using pixel-level allignment and feature-level alignment. Extending previous CycleGAN-based works, Li et al. proposed cycle-consistent conditional adverasarial transfer networks (CATN) to improve adversarial training and feature generation process by conditioning on the classifier prediction. Instead of using

a discriminator, Wu et al. explored channel-wise statistics alignment of CNN features to guide the generation process. Liu et al. combined CoGAN with Variational Autoencoder (VAE) to perform unsupervised image-to-image translation. A shared latent space between source and target domains is inferred to align the joint distributions of different domains. And then training data closer to the target domain can be sampled from the shared latent space. Besides the CycleGAN loss, Kang et al. proposed to impose the attention alignment penalty to reduce the discrepancy of attention maps across domains. To make the attention mechanism invariant to domain shift, the target network is trained with a mixture of real and synthetic data from both source and target domains. Hsu et al. leveraged CycleGAN together with feature-level alignment for object detection adaptation. Kim et al. further propsoed to generate diversified intermediate domains to help domain-invariant representation learning for object detection. A multi-domain discriminator is leveraged to encourage the feature to be indistinguishable among the domains.

E. Self-supervision-based Methods

Self-supervision based methods incorporate auxiliary self-supervised learning task(s) into the original task network. Training the self-supervision task jointly with the original task network is helpful to bring the source and target domains closer. Ghifary et al. designed a three-layer Multi-task Autoencoder (MTAE) architecture to transform the original image into analogs in multiple related domains. The hidden-input and hidden-output weights represent shared and domain-specific parameters, respectively. The learned features are then used as input to a classifier. The category-level correspondence across domains is required. Self-domain and between-domain reconstruction task are introduced as the self-supervision task and are performed during training. Deep reconstruction classification network (DRCN) combines a convolutional supervised network for source label prediction with a deconvolutional unsupervised network for target reconstruction. The feature mapping parameters of the two streams are shared, while the labeling parameters of the supervised network and the decoding parameters of the unsupervised network are learned individually. MTAE requires that the number of samples of corresponding categories in the two domains should be the same. After the sample selection procedure, some important information may be missing. Further, the output of the algorithm is learned features, based on which a classifier (multi-class Vector Machine with a linear kernel in this paper) needs to be trained. DRCN employs an end-to-end strategy, without the requirement of aligned pairs. The above two methods use the same encoder to extract domain-invariant features, ignoring the individual characteristics of each domain. The above two methods use the same encoder to extract domain-invariant features, ignoring the individual characteristics of each domain. Bousmalis et al. explicitly learned to extract image representations that are partitioned into two subspaces. One component is private to each domain, which aims to capture domain-specific properties, such as background. The other is shared across domain with the goal of capturing shared representations by using autoencoders and explicit loss functions, *i.e.* scale-invariant mean square error (SIMSE).

Except for the reconstruction task, more recent self-supervision tasks (*e.g.* image rotation prediction and jigsaw prediction) have been used for DA. Xu et al. suggested using self-supervision pretext tasks (*e.g.* image rotation, patch location prediction) over a feature extractor. Feng et al. proposed to use self-supervision pretext tasks as part of their framework for domain generalization. Carlucci et al. proposed to solve domain adaptation/generalization by introducing a jigsaw puzzle as a self-supervision task. Images are decomposed into 9 patches which are then randomly shuffled and used to form images of the same dimension of the original ones. The Maximal Hamming distance algorithm is used to define a set of patch permutations and assign an index to each of them. The convolutional network is optimized to satisfy two objectives: object recognition on the ordered images and jigsaw classification, namely the permutation index recognition on the shuffled images. Sun et al. further proposed to perform domain adaptation by jointly learning multiple self-supervision tasks. Source and target images share the same convolutional feature encoder, and the extracted features are then fed into different self-supervision task heads: image rotation prediction, patch location prediction, and flip prediction. Since images from different domains normally have many low-level visual differences, *e.g.* brightness, texture, etc., self-supervision tasks that predict high-level structural labels are more favorable for domain adaptation. Kim et al. proposed a cross-domain self-supervised learning approach for DA. It captures apparent visual similarities with both in-domain and across-domain self-supervision. Consequently, they could perform DA with only few source labels. Self-supervised learning has also been introduced into point-cloud adaptation, in which region construction is introduced as a new pretext task.

F. Qualitative Comparison

To thoroughly review the various single-source DUDA methods, we use the following qualitative criteria:

1) *Theory guarantee*: if the target risk has upper bound; and if the upper bound can be minimized by the algorithm.
2) *Efficiency*: the computation cost of the training and inference of the algorithm.
3) *Task scalability*: if the algorithm is applicable to complex tasks, such as semantic segmentation and object detection.
4) *Data scalability*: if the algorithm is applicable to large and complex datasets with rather diversified images.
5) *Data dependency*: if the algorithm can be well trained with small datasets.
6) *Optimizability*: if the algorithm is easy to train and requires less hyper-parameter tuning.
7) *Performance*: how well the algorithm performs.

Discrepancy-based methods usually define a distance measurement between the source and target distributions. Based on this definition, an upper bound of the target risk can

be derived and domain adaptation and domain adaptation algorithms can be designed to minimize this upper bound. Compared with other DUDA categories, many of the existing discrepancy-based methods have better theoretical guarantees. Since most discrepancy-based methods do not add significantly large blocks onto the backbone network, the whole network architectures are usually not very complicated. On the other hand, the computation efficiency of the discrepancy-based methods is usually higher than other categories and the training of the network does not highly rely on large datasets. On the other hand, these methods are not as applicable to large and complex datasets with more diversified images as other categories. In terms of optimizability, since the networks are not very complicated, they are easier to train and require less hyperparameter tuning. Most of the discrepancy-based methods learn image-level representations, instead of pixel-level ones, thus they are not as applicable to complex tasks, such as semantic segmantation, as other categories. It is difficult for most discrepancy-based methods to achieve satisfying performance on complex datasets and tasks.

Adversarial discriminative approaches are the most widely used methods to solve DA problems and achieve remarkable results. Several theoretical studies on these methods focus on the investigation of generalization bound and risk analysis. These methods have competitive computational efficiency and task scalability. In terms of data scalability, they work well across different kinds of datasets. Due to the reliance on the convergence of a min-max game between the discriminator and the feature extractor, they do not always work well on small datasets and are also relatively difficult to optimize.

There is usually no good theoretical support behind adversarial generative approaches since they mainly leverage GAN or other kinds of generative models to reduce the visual gap between source and target domains. However, they usually perform well on many complex tasks with high dimensional solution space, such as semantic segmentation and object detection. It is also because of their reliance on the generative models that they usually require the source and target domains to have homogeneous visual patterns and cannot easily scale to more complex datasets. Since they rely on generative models to build pattern transformation between source and target domains, they require large-scale datasets to robustly train the generative methods. Correspondingly, these approaches also require more computing resources and a more complicated optimization process. Despite the apparent difference, both discrepancy-based methods and adversarial methods can be understood as approaches that attempt to align the marginal feature distributions of both domains. While both methods are intuitive and have seen empirical success in several cases, fundamental limitation exists for both lines of work.

In a recent paper, the authors proved an information-theoretical lower bound on the joint error of methods based on learning domain-invariant representations, showing that when the label distributions of the two domains differ, any algorithm has to achieve a large error on at least one of the two domains. Since only source error can be minimized due to the availability of labeled samples, this implies an increasing error on the target domain. Furthermore, the better the distribution alignment, the worse the joint error. In a concurrent work, they also identified the insufficiency of learning domain-invariant representation for successful adadtation. They further analyzed the information loss of non-invertible transformations, and proposed a generalization upper bound that directly takes it into account.

While most of the work we discussed so far focuses on learning domain-invariant representations, methods based on estimating the importance ratio of density functions between target and source domains are also abundant in the literature. Most of these approaches exhibit provable generalization guarantees under the covariate shift assumption. An interesting avenue for future research is combining the distribution alignment method using deep networks for feature learning with importance ratio reweighting. Note that, different from traditional methods where the importance ratio is estimated between the data density functions, recent work has explored the alternative direction where the importance ratio between the marginal label distributions of the two domains is estimated instead. The fundamental limitation of domain-invariant representations is the potential discrepancy between the marginal label distributions. To overcome such lower bound, one could use the importance ratio between label distributions to compensate for such label discrepancy, as explored in several recent work.

Compared with other methods, self-supervision-based methods do not have a strong theoretical guarantee since these methods are mostly based on the intuition that by forcing the CNN encoder to perform the desired task on the source domain and the pretext tasks on the target domain, the CNN encoder could extract domain-invariant features for both. In terms of computational cost, self-supervision-based methods perform the self-supervision tasks with additional heads, which are normally light-weight CNNs. They normally have more computation cost than discrepancy-based methods, while having less computation cost than adversarial generative methods. Self-supervision-based methods do not have assumptions on the downstream task, and are applicable to complex tasks. In terms of data scalability, self-supervision-based methods are robust and applicable to complex datasets. The self-supervision tasks are normally simple tasks which are easy to train along with the downstream task network. Finally, self-supervision-based methods usually have better performance than discrepancy-based methods, but are less performant than adversarial discriminative and generative methods.

IV. Future Directions

Existing DUDA methods have achieved promising performance on many computer vision tasks, such as object classification and semantic segmentation. However, there is still a large performance gap between existing methods and the upper bound (train and test both on target domain). To help address the remaining challenges, we provide some possible improvements over the state-of-the-art methods. In addition,

we present more practical settings of DA, new applications of DA and brave new perspectives in DA.

A. New Methodologies of DA

Incorporating Previous Knowledge As domain shift is usually caused by imaging process, such as illumination changes, incorporating prior knowledge into the adaptation process may lead to a performance increase. For adversarial methods, imposing multi-level constraints jointly in adaptation, such as low-level appearances, mid-level features, and high-level semantics, can better preserve the structure and attributes of the source data and thus perform better. Designing an effective and direct metric to evaluate the quality of adaptation, instead of testing the performance on the target domain, would accelerate the training process of GANs.

Meta Learning Across Domains Meta learning algorithms provides a learning to learn paradigm that is effective at learning meta models with the capability of solving new tasks for meta model training and the optimized model can only solve new tasks similar to the training ones. These limitations make them suffer from performance decline in the presence of insufficiency of training tasks in the target domains and task heterogeneity, where the source tasks present different characteristics from the target tasks. Besides the above challenge, there may be data distribution shift between source tasks and target tasks, exposing more severe challenges to existing meta learning algorithms. Cross-domain meta learning provides promising solutions to address these challenges by essentially learning more transferable representations.

Contrastive Learning for DA DUDA methods are recently focusing on the disentanglement of the features into domain-invariant and domain-specific ones based on data variations. Domain-invariant features play an important part in reducing the noisy information from each domain, thus making learned features discriminative of the category. Current approaches of contrastive learning for domain addaptation are highly dependent on the design of specific tasks. For example, different DA tasks may rely on different pretext tasks. Therefore, a potential research direction is to design a common pretext task. These methods are often criticized for their computational cost since a large number of negative samples have to be selected for comparison with every single positive sample. Thus, an approach to decrease computational complexity is needed.

REFERENCES

[1] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, and Sanjit A Seshia. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems 2162-237X*, 2020.