MINIMIZING DISTRIBUTION SHIFT BY USING THE DEEP ADVERSARIAL NEURAL NETWORK

Lê Văn Tùng[1*], Đặng Ngọc Huy[1], Nguyễn Văn Tiềm[1]

[1]Trường Đại học Thành Đông

**Tác giả liên hệ: levantungdktd@gmail.com*

ABSTRACT

Minimizing distribution shift is a critical challenge in domain adaptation (DA), as models trained on a source domain often experience degraded performance when applied to a different target domain. To address this issue, deep adversarial neural networks have emerged as a powerful approach to reduce domain discrepancies by leveraging adversarial learning. These networks employ a domain discriminator that encourages the feature extractor to learn domain-invariant representations, thereby aligning the distributions of source and target domains. By minimizing distribution shift, deep adversarial neural networks enable better generalization of deep learning models across diverse applications such as image classification, object recognition, semantic segmentation, and person re-identification. The integration of adversarial training with feature alignment techniques significantly improves model adaptability without requiring extensive labeled data in the target domain. However, challenges such as mode collapse, instability in adversarial training, and the selection of optimal feature representations remain key areas for further research. In this work, we explore deep adversarial neural networks as a solution for minimizing distribution shift and provide an in-depth analysis of their effectiveness, limitations, and potential improvements.

Keywords: *Adversarial neural network, classification, domain adaptation.*

GIẢM THIỂU SỰ THAY ĐỔI PHÂN PHỐI BẰNG MẠNG ĐỐI NGHỊCH TẠO SINH

TÓM TẮT

Giảm sự chênh lệch phân phối dữ liệu là một thách thức trong thích ứng miền (Domain Adaptation - DA), vì các mô hình được huấn luyện trên một miền nguồn thường bị giảm hiệu suất khi áp dụng vào một miền đích. Để giải quyết vấn đề này, mạng nơ-ron đối kháng tạo sinh (Deep Adversarial Neural Network) đã được phát triển như một phương pháp hiệu quả nhằm giảm sự khác biệt giữa các miền thông qua học đối kháng. Mô hình này sử dụng một bộ phân biệt miền, giúp bộ trích xuất đặc trưng học được các biểu diễn không phụ thuộc vào miền, từ đó làm cho phân phối của miền nguồn và miền đích trở nên giống nhau hơn. Nhờ vậy, mạng nơ-ron đối kháng tạo sinh giúp các mô hình học sâu tổng quát tốt hơn, đặc biệt trong các bài toán như phân loại hình ảnh, nhận diện đối tượng, phân đoạn ngữ nghĩa và nhận dạng danh tính cá nhân. Việc kết hợp học đối kháng với căn chỉnh đặc trưng giúp cải thiện khả năng thích ứng của mô hình mà không cần dữ liệu có nhãn trong miền đích. Tuy nhiên, vẫn còn một số thách thức như sự mất ổn định trong huấn luyện đối kháng, nguy cơ sụp đổ chế độ học (mode collapse), và việc lựa chọn đặc trưng phù hợp nhất. Trong nghiên cứu này, chúng tôi phân tích cách mạng nơ-ron đối kháng tạo sinh có thể giúp giảm chênh lệch phân phối, đồng thời đánh giá hiệu quả, hạn chế và hướng cải tiến của phương pháp này.

Từ khóa: *Đáp ứng miền, mạng nơ-ron đối kháng tạo sinh, phân loại.*

Ngày nhận bài: 17/02/2025 Ngày nhận bài sửa: 28/02/2025 Ngày duyệt đăng bài: 05/03/2025

1. INTRODUCTION

Domain adaptation (DA) is a sub-field of transfer learning aims to train a network using source samples that obtain well the performance on the target domain. The main goal of domain adaptation is to approximate the joint distribution of source domain and target domain that is to predict the target labels with minimum expected error. This technique is applied in various real-world applications such as image classification following Azarbarzin and Afsari (2018), Ge (2020), Karimpour (2020), object recognition following Yu (2019), face recognition following Sohn (2017), object detection, semantic segmentation, style translation, person re-identification, and so forth.

• This paper presents a review for domain adaptation techniques in computer vision with deep learning. We cover various current techniques to solve the domain adaptation problem. The current techniques can be categorized into three parts as shown in the Figure 1.

• The first group is the adversarial-based method, which includes techniques for minimizing the distribution difference across domains by using an adversarial manner with a domain discriminator through making confusion domain labels.

• The second group consists of techniques for making more similarities between data distributions in source and target domains by utilizing statistical techniques called the discrepancy-based method.

In the third group, by mapping both the source and target domain into a shared representation domain, the reconstruction-based methods reduce the difference between the domains.

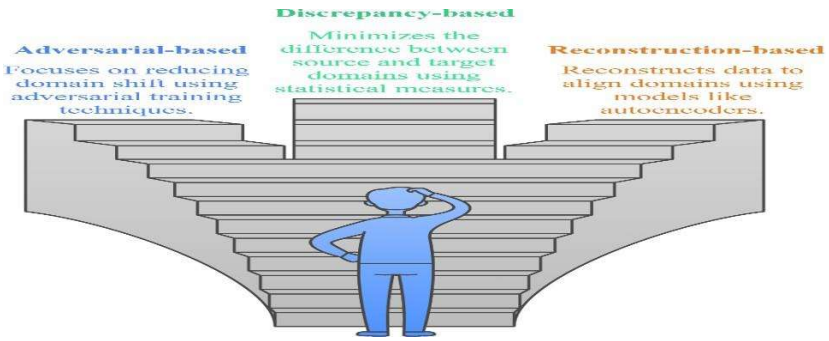


Figure 1. Catergorization of deep domain adaptation

Source: Compiled by the authors

2. ADVERSARIAL- BASED METHOD

Adversarial-based methods following Chadha and Andreopoulos (2019), Long et al. (2018) are an essential type of domain adaptation method to deal with domain adaptation problems, which utilize an adversarial manner for obtaining domain confusion concerning a domain discriminator.

2.1. Discriminative Adversarial Netwo

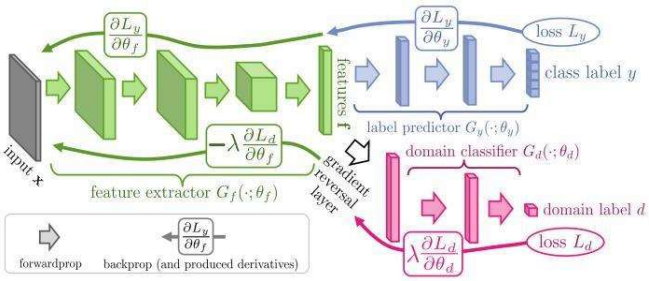


Figure 2. The network architecture of domain adversarial networks

Source: Ganin et al. (2016)

Domain-adversarial training of neural networks following Ganin et al. (2016) is a popular training method in unsupervised domain adaptation, which transfers both domain data to a common feature space and shares weights between the source and target domains. Dann proposes a gradient reversal layer to alleviate the discrepancy between source and target domain. In contrast to the dann method, the adversarial discriminative domain adaptation method following Tzeng et al. (2017) (unshared weights between the two streams) considers independent source and target mappings allowing domain- specific feature extraction to be learned, where the network initializes the target weights pre-trained on the source. By using the adversarial learning method of adda, m-adda following Laradji and Babanezhad (2020) is proposed with a novel metric-learning framework that uses the triplet loss to cluster the source dataset for the task of domain adaptation and a new loss function that regularizes the embeddings of the target dataset to encourage them to form clusters. Improved techniques for adversarial discriminative domain adaptation Chadha and Andreopoulos (2019) is the improved version of adda, which extends the discriminator output over the source classes, in order to additionally incorporate task knowledge into the adversarial loss functions and leveraging on the fixed distribution over source encoder posteriors in order to propose a maximum mean discrepancy (mmd) and reconstruction-based loss function for training a target encoder and discriminator.

2.2. Feature Matching Adversarial Network

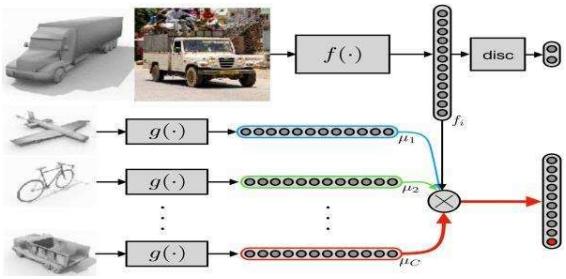


Figure 3. The network architecture of similarity network

Source: Pinheiro (2018)

Based on DANN following Ganin et al. (2016), SimNet method proposed by Pinheiro (2018) is proposed with a different approach that makes some of prototypes from the source data set, then the network is trained to match an embedding of target with the prototypes. An adversarial learning method for domain adaptation is proposed by applying the decision boundaries that are specific for each task to increase the distance between the classifiers. This method utilizes the task-specific classifiers as a discriminator for the

relationship between boundaries of classes and samples of the target domain. Two classifiers work as discriminators and are trained to maximize the discrepancy to detect target samples outside the support by the training knowledge form source domain. In contrast, a feature extractor is trained to minimize the disparity between source domain and target domain by generating target features near the support.

3. DISCREPANCY-BASED METHOD

Discrepancy-based methods try to make more similarity between data distributions. These methods following Tzeng et al. (2014), are implemented by minimizing the difference between the features of source and the target domains.

3.1. Entropy Minimization

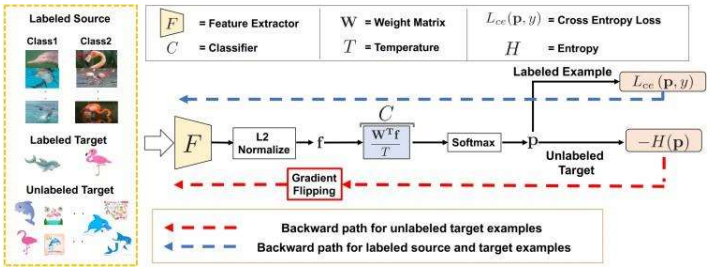


Figure 4. The network architecture of minimax entropy

Source: Saito et al. (2019)

The principle of meca following Morerio et al. (2017) is to connect between correlation alignment and entropy regularization by combining the geodesic correlation alignment with the entropy-based criterion in a unique pipeline that calls minimal-entropy correlation alignment. This method verifies the effectiveness of the proposed approach in terms of systematic improvements over former alignment methods and state-of-the-art techniques for unsupervised domain through an extensive experimental analysis on publicly available benchmarks for transfer object categorization adaptation. In the advent following Vu et al. (2019), they proposed a novel entropy-based adversarial training approach targeting not only the entropy minimization objective but also the structure adaptation from source domain to target domain. This method further improves the performance in specific settings, such as training on particular entropy ranges and incorporating class-ratio priors. Semi-supervised domain adaptation via minimax entropy is the state-of- art method using the minimax entropy. Firstly, they create a prototype representation for each class by maximizing entropy between source data and a few-shot target data. Secondly, reduce the distance between each prototype and its nearby unlabeled samples by minimizing entropy.

3.2. Maximum Mean Discrepancy (MMD)

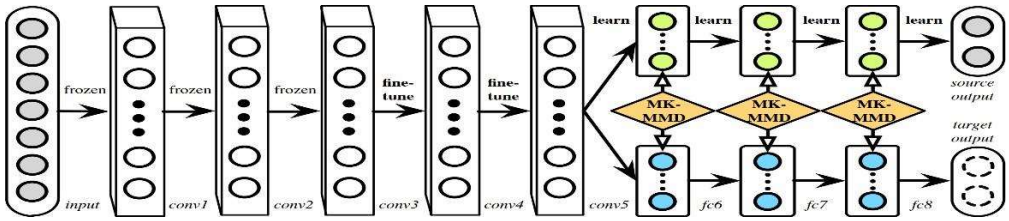


Figure 5. The network architecture of deep adaptation network

Source: Long et al (2015)

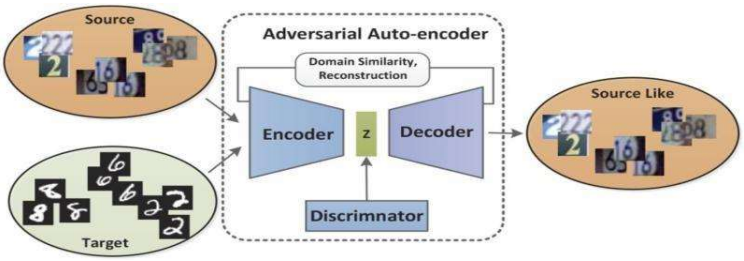


Figure 6. The framework of adversarial autoencoder

Source: Shao and Lan (2019)

Several methods Tzeng et al. (2014), are developed based on MMD, one of the earliest methods is the deep domain confusion Tzeng et al. (2014), where the layer to be considered for the discrepancy and its dimension is automatically selected amongst a set of fine-tuned networks based on linear MMD between the source and the target. The deep adaptation network model following Long et al. (2015) is proposed by using a single layer and linear MMD, which considers the sum of MMDs between several layers. They also introduced an additional work joint adaptation networks. However, JAN Long et al. (2017) is developed to focus on the joint distribution discrepancies of these features instead of the sum of marginal distributions defined between different layers. The deep coral is the extended version of the coral method. The main idea is to learn a nonlinear transformation that aligns the correlations of activation layers between the two streams.

4. RECONSTRUCTION- BASED METHOD

The domain adaptation approaches based on autoencoder reconstruction typically learn the domain-invariant representation by a reconstruction loss in the source and target domains. Graph-based models are usually considered a source graph and a target graph with samples drawn from data manifolds. This method solves the problem of estimating the unknown class labels of the target graph by utilizing the similarity among the two graphs through the weights of graph edges and the label information of the source graph.

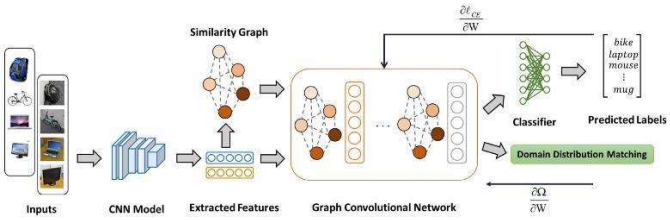


Figure 7. Learning procedure of geometric knowledge embedding

Source: Wu et al. (2020)

4.1. Autoencoder-based Models

A deep autoencoder-based network adaptation framework was presented (Peng, Wang, & Lu, 2012). Autoencoders are typically trained to minimize the reconstruction error by optimizing both the encoder and decoder parameters. The transfer learning with deep autoencoders following Zhu et al. (2019) is proposed to minimize the distance in distributions between domains by using Kullback Leibler (KL) divergence in the embedding encoding layer, and label information of the source domain is encoded using a softmax loss in the label encoding layer. Cross Domain Minimization with Deep autoencoder following Jiang, Chen and Jin (2020) is presented for unsupervised domain

adaptation; the classifier is trained by predicting labels in the source domain and input target domain is reconstructed by using shared features aligned with coral Sun and Saenko (2017) as a regularizer in a unified scheme. The novel autoencoder-based method named adversarial autoencoder (AA) following Shao and Lan (2019) is proposed to incorporate with a feature domain discriminator, so that the encoder tries to encode features that are domain indistinguishable to the feature discriminator for feature alignment. In the AA, a decoder is designed under a reconstruction constraint and a domain similarity constraint based on MMD metrics.

Table 1. Results on the domainnet dataset

Method	Source	DANN		DAN		MCD		JAN		MME	
#Step	-	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot
2,000	46.1	54.4	58.2	54.3	59.5	52.6	56.1	61.1	64.1	62.4	65.2
4,000	49.7	56.1	59.8	57.1	61.0	54.7	58.2	64.7	67.4	64.6	66.7
6,000	50.1	58.3	60.7	58.6	61.9	54.9	58.4	66.0	68.3	67.1	69.2
8,000	53.2	57.9	61.3	57.6	61.1	55.3	57.8	66.8	68.5	68.0	71.0
10,000	53.4	57.6	60.6	59.0	62.0	55.7	58.0	67.8	69.5	69.6	72.0
12.000	53.5	58.1	61.8	59.2	62.5	55.4	57.0	67.8	70.5	70.1	72.3
14,000	55.4	58.7	61.5	58.0	61.8	53.7	56.7	68.2	70.6	70.3	72.3
16,000	55.9	58.3	61.2	59.0	62.8	54.3	54.6	68.7	70.7	70.4	72.3
18,000	55.6	58.3	60.8	59.4	62.1	53.9	56.6	68.4	71.0	**70.6**	72.3
20,000	53.7	58.6	61.4	59.4	62.5	52.6	54.5	68.3	71.0	70.4	**72.3**

Source: Compiled by the authors

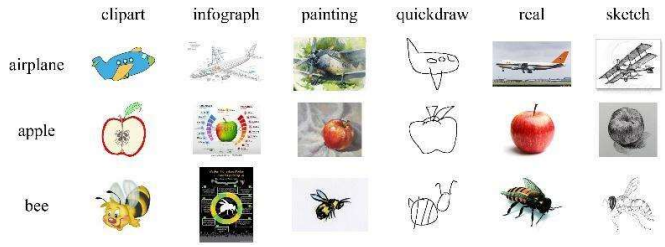


Figure 8. Example images in *domainnet-345* dataset

Source: Compiled by the authors

4.2. Graph-Based Models

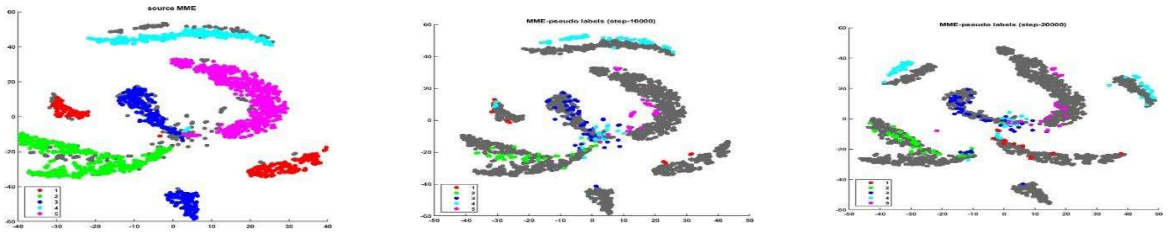
The graph-based models try to solve the problem of estimating the unknown class labels of the target graph utilizing the label information of the source graph and the similarity among the two graphs through the weights of graph edges. It considers a source graph and a target graph with samples drawn from data manifolds. An optimal bayesian transfer learning following Karbalayghareh et al. (2018) classifier combines graph model concept with a Bayesian method for domain adaptation using the prior knowledge of source and target domains. To avoid costly computations, the OBTL classifier is derived based on the Laplace approximated hypergeometric functions. The adaptive bayesian linear regression model following Perrone et al. (2017) is a graph-based method introduced for multi-task applications; a Bayesian linear regression layer models each task in the ABLR on top of common feature space. The method called geometric knowledge embedding following Wu et al. (2020) is introduced in order to learn discriminative and transferable representations to exploit the geometric information of data, which is usually omitted in most existing domain adaptation methods.

They describe the relationship between samples of both source and target data based on their similarities and develop a graph convolutional network to learn discriminative representations based on the constructed graph.

5. EXPERIMENTAL RESULTS

Datasets. We evaluated recent methods using the two benchmark dataset: DomainNet Peng et al. (2019) in Figure 8. DomainNet contains 6 domains of 345 classes each. Among them, we used 2 domains (Real, Clipart) and 126 classes. We chose real images as source domain and clipart images as target domain.

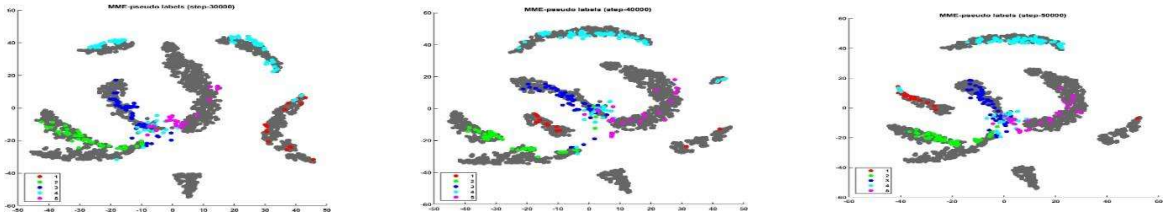
Implementation. We select ResNet-34 following, which is pre-trained on ImageNet following Deng et al. (2009), for the base networks. We compare with the following methods, DANN Ganin et al. (2016), DAN proposed by Long et al. (2015), MCD introduced by, JAN Long et al. (2017), and MME introduced by. All these methods are implemented to address the semi-supervised domain adaptation problem, which includes small number of labeled target images during training. In this case, we implemented two scenarios 1-shot and 3-shot images.



(a) Embedding Space of Source (MME-pseudo labels)

(b) Embedding Space of Target after 10k step (MME-pseudo labels)

(c) Embedding Space of Target after 20k step (MME-pseudo labels)



(d) Embedding Space of Target after 30k step (MME-pseudo labels)

(e) Embedding Space of Target after 40k step (MME-pseudo labels)

(f) Embedding Space of Target after 50k step (MME-pseudo labels)

Figure 9. t-sne visualiation over various the training steps

Source: Compiled by the authors

Figure 9. t-SNE visulalization on DomainNet. (a) Embedding space after training on the Source (Real images), (b) Embedding space on Real→Clipart without domain adaptation and (c) Embedding space on Real→Clipart with domain adaptation (MME method).

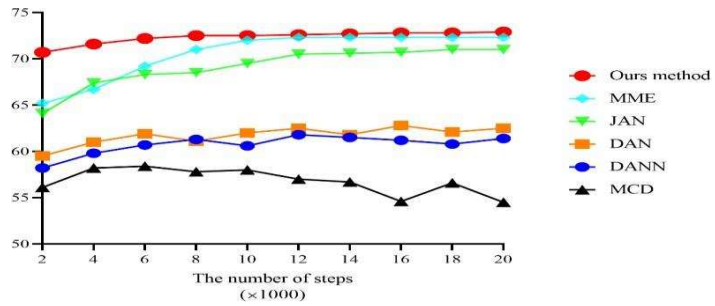


Figure 10. Performance comparison of domain adaptation methods over training steps

Source: Compiled by the authors

As shown in Figure 9, the target features have a tendency to move the source features after applying the domain adaptation method.

The results in Figure 10 compares the performance of domain adaptation (DA) methods based on the number of training steps, with the proposed method achieving the highest accuracy, maintaining a stable performance at around 74%. MME also demonstrates competitive performance, gradually improving and nearly reaching the proposed method, while JAN follows behind with an accuracy of approximately 68% at the final stage. DAN and DANN exhibit moderate performance, fluctuating around 60%, with no significant improvement after a certain number of training steps. In contrast, MCD shows the lowest and most unstable performance, indicating that this method may not be effective in this training scenario. The overall trend suggests that Ours method, MME, and JAN have greater application potential due to their consistent and steady performance improvements over time.

6. CONCLUSION

In this paper, we provided a survey of different methods for domain adaptation with deep learning. We categorized the image classification methods into three main groups based on the technology adopted for domain adaptation: discrepancy-based, adversarial-based, reconstruction-based methods. We reviewed recent papers for different deep visual domain adaptation applications, such as image classification, semantic segmentation, and object detection. Finally, experiment results indicated that the domain adaptation could lead to enormous advancements in classification problems.

7. FUTURE WORKS

In future research, we plan to extend our study by applying adversarial networks to address the distribution shift problem in industrial engines. By leveraging deep adversarial learning, we aim to develop a more robust and adaptive model that can effectively minimize discrepancies between different operating conditions of industrial machinery. This approach will help enhance predictive maintenance, fault detection, and performance optimization by ensuring better generalization across various engine environments. Additionally, we will explore the integration of domain adaptation techniques with real-time monitoring systems to improve the reliability and efficiency of industrial engine operations. Our future work will focus on optimizing the adversarial training process to achieve higher accuracy, stability, and scalability in industrial applications.

REFERENCES

Azarbarzin, S., & Afsari, F. (2018). Domain adaptation by manifold transfer for image classification. *In 2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS).*

Chadha, A., & Andreopoulo, Y. (2019). Improved techniques for adversarial discriminative domain adaptation. *IEEE Transactions on Image Processing*, 29, 2622-2637.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *In 2009 IEEE Conference on Computer Vision and Pattern Recognition,* Miami, FL, USA, 248-255.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research, 17*(1), 2096-2030.

Ge, P., Ren, C. X., Dai, D. Q., & Yan, H. (2020). Domain adaptation and image classification via deep conditional adaptation network. *arXiv, 3*, 1-13.

Ge, W., Chen, H., Cai, D., & He, X. (2018). Improving face recognition with domain adaptation. *Neurocomputing*, 287, 45-51.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *arXiv*, 10, 1-12.

Jiang, B., Chen, C., & Jin, X. (2020). Unsupervised domain adaptation with target reconstruction and label confusion in the common subspace. *Neural Computing and Applications*, 32(9), 4743-4756.

Kan, M., Wu, J., Shan, S., & Chen, X. (2014). Domain adaptation for face recognition: Targetize source domain bridged by common subspace. *International Journal of Computer Vision*, 109(1-2), 94-109.

Karbalayghareh, A., Qian, X., & Dougherty, E. R. (2018). Optimal Bayesian transfer learning. *IEEE Transactions on Signal Processing, 66*(14), 3724-3739.

Karimpour, M., Saray, S. N., Tahmoresnezhad, J., & Aghababa, M. P. (2020). Multi-source domain adaptation for image classification. *Machine Vision and Applications, 31*(6), 1-19.

Laradji, I. H., & Babanezhad, R. (2020). M-ADDA: Unsupervised domain adaptation with deep metric learning. *arXiv*, *6*, 1-14.

Long, M., Cao, Y., Wang, J., & Jordan, M. I. (2015). Learning transferable features with deep adaptation networks. *arXiv, 27,*1-9.

Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2016). Unsupervised domain adaptation with residual transfer networks. *arXiv, 16,*1-9.

Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2017). Deep transfer learning with joint adaptation networks. *arXiv*, 17,1-10.

Long, M., Cao, Z., Wang, J., & Jordan, M. I. (2018). Conditional adversarial domain adaptation. *arXiv, 29,* 1- 11.

Morerio, P., Cavazza, J., & Murino, V. (2018). Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *arXiv, 28,* 1- 14.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., & Wang, B. (2019). Moment matching for multi-source domain adaptation. arXiv, *27*, 1- 24.

Peng, Y., Wang, S., & Lu, B.-L. (2012). Marginalized denoising autoencoders for domain adaptation. *ArXiv, 23*, 1-8.

Perrone, V., Jenatton, R., Seeger, M., & Archambeau, C. (2017). Multiple adaptive Bayesian linear regression for scalable Bayesian optimization with warm start. *arXiv, 8*, 1-6.

Pinheiro, P. O. (2018). Unsupervised domain adaptation with similarity learning. *In Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), 8004-8013.

Saito, K., Kim, D. H., Sclaroff, S., Darrell, T., & Saenko, K. (2019). Semisupervised domain adaptation via minimax entropy. *arXiv, 14*, 1-6.

Shao, R., & Lan, X. (2019). Adversarial auto-encoder for unsupervised deep domain adaptation. *IET Image Processing, 13*(14), 2772- 2777.

Sohn, K., Liu, S., Zhong, G., Yu, X., Yang, M. H., & Chandraker, M. (2017). Unsupervised domain adaptation for face recognition in unlabeled videos. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3210- 3218.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., & Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *ArXiv, 10,* 1-9.

Vu, T. H., Jain, H., Bucher, M., Cord, M., & Perez, P. (2019). Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. *ArXiv, 17, 1-9.*

Wu, H., Yan, Y., Ye, Y. Z., Ng. M. K. (2020). Geometric knowledge embedding for unsupervised domain adaptation. *Knowledge-Based Systems*, *191*, 105155.

Yu, F., Wang, D., Chen, Y., Karianakis, N., Shen, T., Yu, P., Lymberopoulos, D., Lu, S., Shi, W., & Chen, X. (2019). Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning. *ArXiv, 5,* 1-14.

Zhu, Y., Wu, X., Li, P.., Zhang, Y., & Hu, X. (2019). Transfer learning with deep manifold regularized auto-encoders. *Neurocomputing*, 369, 145-154.