Domain Adaptation with DANN: Enhancing Cross-Domain Image Classification from CIFAR-10 to STL-10

Ammar Ahmed Khan
Department of Software Engineering
NED University of Engineering and Technology
Karachi, Pakistan
khan4405225@cloud.neduet.edu.pk

Muhammad Asim
Department of Software Engineering
NED University of Engineering and Technology
Karachi, Pakistan
asim4409005@cloud.neduet.edu.pk

***Abstract*—Domain adaptation is a central machine learning problem whose capability of transferring knowledge from a well labelled source domain to a less labelled or even completely unlabeled target domain is a critical challenge in practice. The problem, known as domain shift, frequently results in models trained on one dataset performing less than optimally when they are applied to another. In this work we study the efficacy of Domain Adversarial Neural Networks (DANN) in mitigating the impact of domain shift by training a pre-trained ResNet-18 model trained on CIFAR-10 on STL-10, which is less labelled. To learn domain-invariant features, we employ adversarial training to drive the feature distributions of the source and target domains to be similar. Finally, our experiments show that fine tuning the ResNet 18 model on the target domain with DANN leads to significant improvement in classification performance, indicating that the idea of adversarial domain adaptation techniques can help increase model generalisation for real world applications.**

***Index Terms*—Domain Adaptation, Domain-Adversarial Neural Network, Cross-Domain Learning, Feature Alignment, Transfer Learning**

I. INTRODUCTION

Recently, domain adaptation has been a very important approach in Machine learning particularly in situations where models trained in one environment need to work well in a different one. Traditionally, machine learning models have typically assumed that the source training data and the target test data come from the same distribution. In reality, however, this assumption often doesn't hold; either because the conditions, environments, or types of data sampled have changed. For example, a model trained to classify images in one domain, such as high-resolution, clear images, may struggle to generalise well to another domain, like low-resolution images captured in different lighting or angles. We refer to this mismatch between data distributions between domains as the 'domain shift' and show that it can substantially degrade performance of a model when tried on new domains.

For this reason, the domain adaptation (DA) techniques have been designed to address this issue, by adapting models to perform well in a target domain even if it is different from the source domain. Domain adaptation is particularly useful in situations where collecting large amounts of labelled data in the target domain is difficult, expensive, or time-consuming. With labelled data from a related source domain, we are able to enhance the model performance in the target domain without having to begin from scratch with no training data at all. [16] [17] The ability to transfer knowledge between domains makes DA a critical component of modern machine learning tasks including image recognition, speech processing, and even autonomous driving, in which data are often highly variable.

Despite the advances in machine learning, traditional models face a significant limitation: they are not robust to domain shifts. [3] [18] This has created a gap in research, as conventional supervised learning methods perform poorly when applied to unseen or new domains with different characteristics. For instance, a model trained on images of objects under controlled lighting might fail when tasked with recognizing those objects in outdoor environments. This gap highlights the need for domain adaptation methods that allow models to generalise across different domains without requiring massive amounts of labelled data for every new domain.

Transfer learning is a broader sub field of domain adaptation in which one task's knowledge is transferred to another. However, in transfer learning, tasks can vary, whereas in domain adaptation, they are identical (e.g., classification of images), but their environments differ (i.e., between domains). [4] One of the many ways of domain adaptation is to fine tune models on the target domain data, feature alignment techniques which make the source and target feature spaces closer, or adversarial training where models learn domain invariant feature. These approaches decrease the discrepancy between source and target domains, and therefore improve generalisation.

The availability of labelled data in the target domain determines the scenario in which domain adaptation can be used. Supervised domain adaptation assumes both domains have labelled data, and unsupervised domain adaptation assumes none of the target domain is labelled. This research is centred around a semi-supervised domain adaptation problem that sits

between these two extremes where some labels are available in the target domain but not enough to train a model from scratch. In this study, we fine tune a pre-trained model on CIFAR 10 (the source domain) using the full STL 10 dataset (the target domain) to show that domain adaptation improves model performance in a real world setting.

This study aims to investigate whether Domain Adversarial Neural Networks (DANN) perform well on a domain shift between a target domain (STL-10) that has less labelling and a source domain (CIFAR-10) that has more labelling. Specifically, we adjust a pre-trained ResNet-18 model to adapt to the target domain and then assess how adversarial training can enhance the model's ability to generalise to the domain adaptation tasks in the real world.

II. Litrarture Review

In machine learning, domain adaptation (DA) is a crucial technique, especially when the training and testing datasets—also known as the source and target domains—do not match. Our project uses advanced DA techniques to improve image classification in remote sensing and medical diagnostics. To understand the current state of the domain and identify gaps that our project can address, it is crucial to review existing literature. With a focus on domain adaptation in image classification, this literature review identifies significant contributions and methodologies that serve as the basis for our work.

One important study by Peng et al. [4] examined DA methods for hyperspectral image classification in remote sensing. The authors explored both traditional shallow methods and recent deep DA approaches, applying them to two datasets from the IEEE GRSS Data Fusion Contests: Houston2013 and Houston2018. These datasets represented hyperspectral images with varying spectral bands and class categories, and the goal was to adapt a model trained on Houston2013 (source domain) to classify images from Houston2018 (target domain). The research compared subspace-based methods, such as GFK and SA, and deep DA models, including adversarial learning approaches like DAAN and MCD. Their results showed that while traditional methods struggled with spectral shifts between domains, deep DA methods—particularly TSTnet, which incorporates Graph Convolutional Networks (GCNs) and Convolutional Neural Networks (CNNs)—achieved the best performance by aligning feature distributions across domains. [13] [20] However, the authors noted the computational intensity of deep learning methods, which may not scale well for larger datasets, and highlighted the challenge of class overlap in certain domains, where subtle spectral differences between classes could lead to misclassifications.

Karimpour et al. [5] have enriched the domain adaptation line of research by exploring Multi-Source Domain Adaptation (MDA) for the image classification tasks. Their study focused on two benchmark datasets: Office+Caltech and PIE. MDA techniques are crafted to use knowledge of source domains to operate in a target domain since there are issues relating to when data distributions differ across sources as well as between the source and target domain. [19] [15] Results of the proposed method in this study demonstrated higher improvement of classification accuracy especially for multi source domains. For instance, the MDA approach reveals an approximately 43.82 percent increase of the average improvement compared to the baseline methods for the PIE dataset and 20.46 percent maximum improvement for the Office+Caltech dataset. Nevertheless, the authors agreed that larger datasets or those where 'd' is larger may be challenging for the method that critical hyperparameters need to be tuned to optimize the result. Further, the idea of constructing domain-invariant clusters may limit its ability to generalize to more diverse real-world domains where the relation between source and target domains may be more intricate.

Yin et al. [6] then proposed the extension of the DA space to Universal multi-source domain adaptation while comparing the latest approach in this field, the Universal Multi-source Adaptation Network (UMAN), to other present day DA techniques. They used such models to analyze single-source, partial, and open-set DA and confirmed that the UMAN yields better performance compared with ResNet, DANN, MDDA, and MDAN, which are well-developed single-source and multi-source DA methods. The most crucial strength of UMAN is that it minimizes negative transfer by learning from multiple source domain and addressing domain and label shifts. The efficiency of UMAN has been evidenced from the results of the study, and by using ablation studies and statistical tests it has been validated that the proposed model performs well in source-combined and multi-source DA. However, the study also revealed that the number of target domains affected the performance of UMAN and it began deteriorating at some threshold value for the number of source domains, highlighting that the efficacy of the model possibly might not be high in other extreme multi-source situations. Moreover, the study suggested that UMAN's performance is less proactive to fluctuating label sets and overlapping domain compared to the other methods, however, its endurance might be improved across more assorted datasets.

Deng et al. [7] investigated the use of multi-kernel learning (MKL) along with active learning (AL) to tackle classification problems in hyperspectral images (HSIs), which frequently have notable domain shifts between source and target domains. In order to reduce the quantity of labelled samples needed for training, their suggested framework, MKL-MS, iteratively chooses the most instructive samples for labelling by fusing the advantages of multi-kernel classifiers with an active learning approach. Their experiments on two widely used hyperspectral datasets—Pavia Center and University Area—showed that MKL-MS outperformed traditional methods, including SVM and random sampling techniques, by achieving higher overall accuracy (OA) and faster convergence. The integration of domain adaptation into their framework allowed for robust performance despite large shifts in data distribution. However, the study also highlighted several limitations, including the

poorer performance of MKL-MS in the early stages of learning, where fewer labeled samples were available. Additionally, while the method was effective for large domain shifts, its advantages in situations with smaller shifts were not well-explored, and its computational cost—due to the use of multiple kernels and the iterative AL process—was noted as a potential drawback.

Wei et al. [8] presented a novel domain adaptation method, CenterDA, which was evaluated on three benchmark datasets: Office-31, Office-Home and ImageCLEF-DA. The most significant aspect of CenterDA is the center alignment approach of CenterDA that maps the target domain data against the source domain data learnt from the common class center. To achieve transfer learning alongside of label smoothing and Maximum Mean Discrepancy (MMD), this strategy was employed. Based on three transfer tasks, CenterDA demonstrated superior performance comparing to other state of the art methods. For instance, it demonstrated the mean recognition rate of 89.5 percent on the Office-31 database and outperformed most basic approaches. However, some limitations were also observed in the study, for example the method relies on the characteristics of the domain and it is slower with small datasets like ImageCLEF-DA limited by the number of samples (600). The authors also mentioned the possibility of further improvement in terms of time while explaining that due to the fine-tuning of the model, the time complexity becomes rather high.

The FFSC (Feature Fusion and Sample Clustering) approach was presented by Wang et al. [9] for unsupervised domain adaptation (UDA) in the classification of images from breast cancer histopathology. This method uses multi-level feature fusion and sample clustering strategies to improve classification performance in cross-dataset domain adaptation scenarios. The authors demonstrated that FFSC outperforms existing methods like MCD, AFN, MCC-UDA, and MDD, particularly in breast cancer image classification tasks. Their key innovation lies in the integration of low- and high-level feature fusion with clustering strategies, which significantly enhanced classification accuracy. [10] However, the study also acknowledged limitations such as the small size of the SNL dataset, which might limit the generalizability of the method, and the reliance on pre-trained deep learning models that may not be readily available in all clinical settings. The authors also suggested that further research is needed to explore the clinical applicability of their model in real-world diagnostic environments.

III. Methodology

This section explains the Domain-Adversarial Neural Network (DANN) method used for domain adaptation between CIFAR-10 and STL-10. Model architecture, evaluation metrics, dataset preparation, and domain adaptation through adversarial training comprise the four stages of the process.

Our source domain data for this is the CIFAR-10 dataset, which consists of 10,000 test and 50,000 training images, each of which is labelled as being related to one of ten object classes (truck, aeroplane, car, bird, cat, deer, dog, frog, horse, ship, or truck). 5,000 labelled training images and 8,000 test images, each covering 10 object classes but with varying distributions, resolutions, and image qualities, make up the STL-10 dataset that was used to supply target domain data.

To align the input dimensions between these datasets, STL-10 images (96x96) were resized to match CIFAR-10's 32x32 resolution. Normalization was performed using dataset-specific means and standard deviations for each channel (RGB) to standardize pixel values between 0 and 1. The following transformations were applied:

- CIFAR-10: Horizontal flipping, random cropping, and color jitter.
- STL-10: Similar augmentations as CIFAR-10 after resizing to 32x32.

Both datasets were split into mini-batches for training, with a batch size of 64.

A. *Model Architecture*

We employed a ResNet-18 model as the backbone for feature extraction. This architecture was chosen due to its proven capacity to extract meaningful hierarchical features from images, especially for complex classification tasks. [11] ResNet-18, with its 18 convolutional layers and skip connections, provides a balance between performance and computational efficiency.

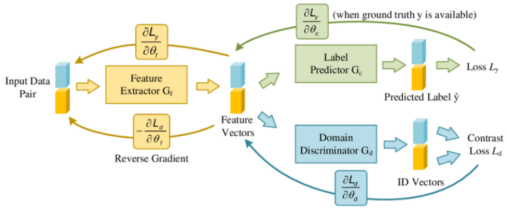

Fig. 1. DANN Architecture [1]

Feature Extractor *:* The ResNet-18 model was modified to exclude its final fully connected layer, producing a feature representation $f(x)$ of each input image x. These features serve as input to two components:

- The classifier, which predicts object categories for the source and target domains.
- The domain discriminator, which attempts to distinguish between source and target domain samples.

Domain-Adversarial Training *:* The DANN model learns domain invariant features using adversarial learning, by aligning the source and target domains feature distributions. The domain discriminator is trying to classify the domain of each input, and the feature extractor tries to fool it by making domain labels indistinguishable.

The adversarial training is guided by a gradient reversal layer (GRL). During backpropagation, the GRL inverts the gradients from the domain discriminator, forcing the feature extractor to produce domain-invariant features.

Mathematically, the total loss is a combination of two components:

- Classification Loss: Cross-entropy loss for predicting object classes from the source domain $\mathcal{L}_{\text{cls}}$.
- Domain Loss: Binary cross-entropy loss for distinguishing between source and target domain samples $\mathcal{L}_{\text{dom}}$.

The total loss function for the model is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{dom}}$$

where λ is a trade-off parameter controlling the importance of domain adaptation relative to classification accuracy.

B. *Adaptation Methodology*

*1) **Optimization Strategy** :* The Adam optimizer was used with a learning rate of 1×10^{-4}. A gradual decay schedule was applied, reducing the learning rate by a factor of 0.1 every 20 epochs. The model was trained for 50 epochs, and we used early stopping to prevent overfitting. Gradient clipping was also applied to stabilise the adversarial training.

*2) **Evaluation Metrics**:* The performance of the DANN model was evaluated using the following metrics:

- Classification Accuracy: Accuracy on the target domain (STL-10) measures how well the model classifies images after domain adaptation.
- Domain Classification Accuracy: This measures the accuracy of the domain discriminator. A low accuracy for the domain discriminator indicates that the feature extractor has successfully aligned the feature distributions of the source and target domains.
- Confusion Matrix: The confusion matrix was computed on the target domain to visualize class-wise classification performance and identify misclassifications between similar classes.

*3) **Baseline Comparisons**:* To assess the impact of domain adaptation, a baseline model was trained on CIFAR-10 without any domain adaptation techniques and evaluated on STL-10. This baseline provides insight into the domain shift between the datasets, showing how poorly a source-only model performs without adaptation. The comparison of this baseline with the DANN model demonstrates the improvement brought by adversarial training.

C. *Experimental Setup*

The model was trained on a Kaggle P100 GPU, which provided the necessary computational power for handling the DANN model efficiently. The experimental process followed this sequence:

- Preprocessing the datasets and resizing the images to match dimensions.
- Implementing the ResNet-18-based feature extractor, classifier, and domain discriminator.
- Training the DANN model with adversarial loss to encourage domain-invariant feature learning.
- Evaluating the model on the target dataset (STL-10) and comparing results with the baseline.

This methodology outlines the complete workflow, from data preparation and model design to the adversarial training process and performance evaluation, providing a comprehensive framework for using DANN in the context of domain adaptation between CIFAR-10 and STL-10.

IV. RESULTS AND DISCUSSION

When the new features were obtained from the source domain CIFAR-10, and adapting to the STL-10 target domain, the improvements achieved through the DANN model were relatively high. The first model where no form of domain adaptation was done performed a classification accuracy of 40

The degree of feature alignment is explained by information obtained from the t-SNE visualisation of domain-invariant features (Figure 1). The figure displays both the similarities and differences between the CIFAR-10 and STL-10 distributions. These areas of overlap suggest that the model has met its objective of finding characteristics shared by the two domains of use and exhibits strong cross-domain use capabilities. They do, however, form discrete clusters, indicating that domain-specific features continue to exist and that the model is completely invariant to domain changes. These results demonstrate the benefits and drawbacks of the feature extraction function that is currently in use, as well as the potential for additional modification to get rid of excessive domain specificity.

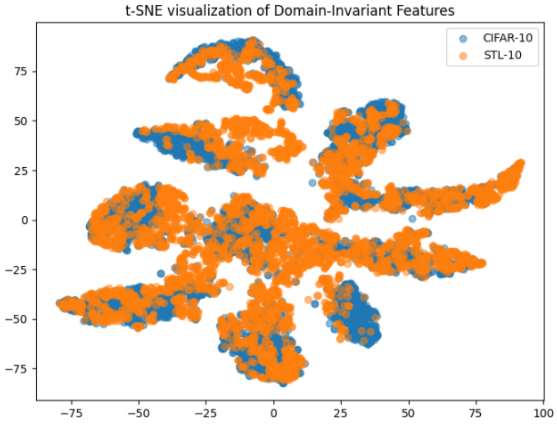

Fig. 2. t-SNE Visualization of Domain-Invariant Features

Additional characteristics of the model's performance in various classes are displayed in the target domain confusion matrix in Figure 2. As has been observed, several classifications, such as automobiles or animals, possess or resemble traits comparable to one another and are frequently misclassified. The model sometimes substitutes the words "cars," "trucks," and "cats" for "dogs," for example. Despite the fact that domain adaptation has improved overall accuracy, plotting these patterns shows that the model is unable to distinguish

between items in the target domain that are occasionally indistinguishable with the naked eye. This highlights the need for improved feature extraction or, at the very least, class-specific fine-tuning in the training phase.

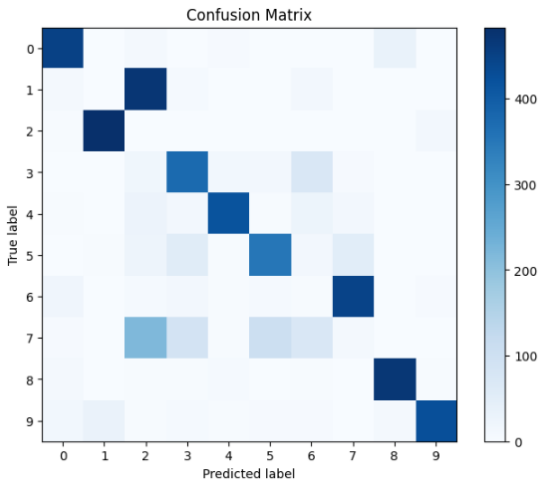

Fig. 3. Confusion Matrix for Target Domain (STL-10)

Additional information can be discovered from the training loss curves shown in Figure 4, with classification loss and domain loss on the vertical axis and epochs on the horizontal axis. The classification loss keep on decreasing consistently proving that the model is learning the trivial classification of the target domain more and more every epoch. Nevertheless, the domain loss still stays high and almost unchanged during the training; this could only mean that the adversarial part (designed that to reduce the domain difference of the features) is ineffective. Perhaps, the domain loss should gradually reduce with time as the feature extractor effectively deceives the domain discriminator, but the relative stability thereof suggests that further optimization of this adversarial learning is possible.

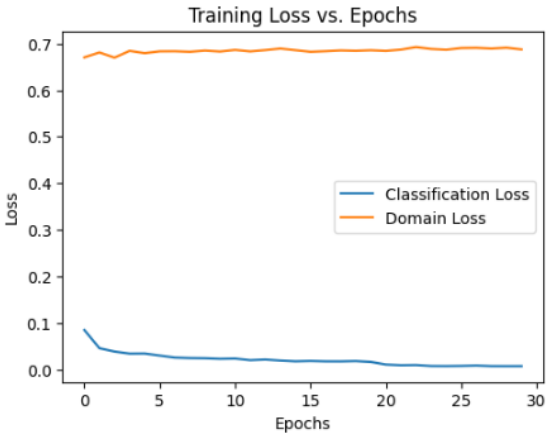

Fig. 4. Training Loss Curves for Classification and Domain Loss

These findings are further echoed in the domain discriminator accuracy plot in figure 5 which show accuracy oscillating around 51 percent as the training progresses. Approximately a 50 percent accuracy rate tells that the domain discriminator is being deceived more often than not, suggesting that the feature extractor accomplishes domain-invariant feature representation to a certain extent. However, the oscillations indicate that there is noise and, thus, the model seems to be only partially picking up highly domain invariant features at any given time. This instability could be a reason for some of the domain-specific clusters which were visible in t-SNE and PCA plots.

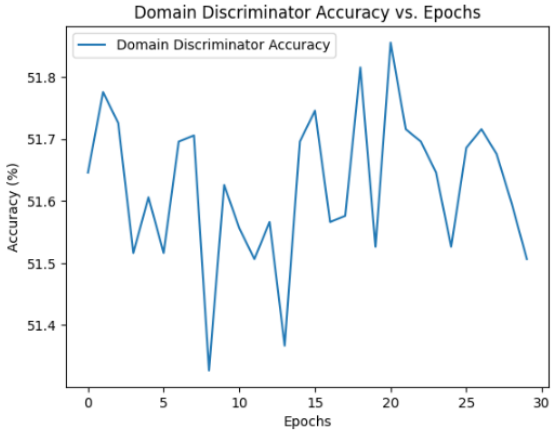

Fig. 5. Domain Discriminator Accuracy Across Epochs

So the model's accuracy improved from 40 percent to 50 percent, which demonstrates the value of introducing adversarial training to reduce domain shift. While this is a meaningful improvement, the gap between the source and target domains suggests there is still more work to be done in terms of aligning features. Refining the model architecture or loss functions may lead to further gains in accuracy and generalization.

V. CONCLUSION

In this study, we examined the use of DANN in transfer learning of ResNet-18 pre-trained for CIFAR-10 source DOM with less labelled images compared to STL-10 target DOM. The domain shift between these datasets due to varying image quality, content, and labelling hindered the study challenge and the possible impact of direct model transfer from these datasets. In order to improve the target set's classification accuracy, we employed DANN in this work to make sure that the feature spaces between the source and target domains match. The outcomes demonstrate that, in comparison to a standard classifier that does not use domain adaptation, our suggested approach, DANN, enhances performance on the target domain by roughly 10 percent. Additionally, it is known through the confusion matrix research that while DANN improves classification outcomes, certain classes in the target domain are still identifiable while others appear similar in the visual image. Therefore, our work shows that DANN can be used in situations when there is limited access to labelled data in the target domain. In this context, it is demonstrated

that domain adaptation techniques like DANN provide a very practical and effective way to extend the applicability of machine learning algorithms onto new, unlabelled domains. In order to overcome the fine-grained classification challenge and enhance performance, future research should focus more on how to better align the domain-aligned features and investigate additional domain adaptation techniques.

VI. FUTURE IMPROVEMENT

Finally, we discussed some areas that require more research after demonstrating that DANN can learn a domain in-variance mapping for the adaptation between CIFAR-10 and STL-10. Future research may concentrate on investigating even more dependable domain adaptation strategies, such as increasing the number of adversarial techniques or other divergence metrics that could improve feature distribution matching and, as a result, produce greater generalisation on the target environment. Additionally, the feature extraction and classification methods need to be further improved in order to distinguish between a few related classes in one of the target domains that are necessary for object detection based on the data of the confusion matrix. Further work could be done when utilising sophisticated architectures or even selecting to use the models in ensemble to better identify classes. Incorporating techniques in a semi-supervised or self-supervised way may also be advantageous when using such a model, especially when learning from unprocessed data from the target domain. To find out if DANN will work for other types of domain adaptation, another line of research might involve using different sets with different domain shifts in different tasks. Under the same guise of practical applicability, we will be able to assess the degree of risk-based utility of these techniques when applied to real-world scenarios like medical imaging and self-driving automobiles, among others.

REFERENCES

[1] G. Xu, R. Qin, Z. Zheng, and Y. Shi, "An adaptive system for wearable devices to detect stress using physiological signals," arXiv preprint arXiv:2407.15252, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2407.15252.

[2] Duan, L., Tsang, I.W., Xu, D., Maybank, S.J.: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on pp. 1375–1381 (2009)

[3] Dinh, C.V., Duin, R.P., Piqueras-Salazar, I., Loog, M.: FIDOS: a generalized Fisher based feature extraction method for domain shift. Pattern Recognit. 46(9), 2510 (2013)

[4] Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J.: Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22(14), e49 (2006)

[5] Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Trans. Neural Netw. 22(2), 199 (2011)

[6] Y. Yin, Z. Yang, H. Hu, and X. Wu, "Universal multi-Source domain adaptation for image classification," Pattern Recognition, vol. 121, pp. 108238–108238, Jan. 2022, doi: https://doi.org/10.1016/j.patcog.2021.108238.

[7] C. Deng, X. Liu, C. Li, and D. Tao, "Active multi-kernel domain adaptation for hyperspectral image classification," Pattern Recognition, vol. 77, pp. 306–315, May 2018, doi: https://doi.org/10.1016/j.patcog.2017.10.007.

[8] G. Wei, Z. Wei, L. Huang, J. Nie, and X. Li, "Center-aligned domain adaptation network for image classification," Expert Systems with Applications, vol. 168, p. 114381, Apr. 2021, doi: https://doi.org/10.1016/j.eswa.2020.114381.

[9] P. Wang, G. Yang, Y. Li, P. Li, Y. Guo, and R. Chen, "Deep sample clustering domain adaptation for breast histopathology image classification," Biomedical signal processing and control, vol. 87, pp. 105500–105500, Jan. 2024, doi: https://doi.org/10.1016/j.bspc.2023.105500.

[10] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in Proc. 32nd Int. Conf. Machine Learning (ICML), Lille, France, Jul. 2015, pp. 1180-1189.

[11] Vernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Proceedings of the IEEE international conference on computer vision pp. 2960–2967 (2013)

[12] Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. Compu. Vis. ECCV 2010, 213–226 (2010)

[13] Duan, L., Xu, D., Chang, S.F.: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on pp. 1338–1345 (2012)

[14] "An In-Depth Analysis of Domain Adaptation in Computer and Robotic Vision." Available: https://irobotics.aalto.fi/manipulation-and-grasping/.

[15] Sensor Fusion, ScienceDirect, Available: https://www.sciencedirect.com/topics/engineering/sensor-fusion.

[16] Chen, M., Weinberger, K.Q., Blitzer, J.: Co-training for domain adaptation. In: Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS), pp. 2456–2464 (2011)

[17] "Domain adversarial neural networks for domain generalization: when it works and how to improve." Available: https://arxiv.org/abs/2102.03924

[18] Onur Tunali, "Maximum Mean Discrepancy in Machine Learning." Available: https://www.onurtunali.com/ml/2019/03/08/maximum-mean-discrepancy-in-machine-learning.html.

[19] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Long Beach, CA, USA, Jun. 2019, pp. 6936–6945.

[20] "Learning to Adapt Structured Output Space for Semantic Segmentation." Available: https://arxiv.org/abs/1802.10349