BRIDGING DOMAIN SHIFTS THROUGH SELF-CONTRASTIVE LEARNING AND DISTRIBUTION ALIGNMENT

Vinicius Avena, Rodrigo S. Couto, Luís Henrique M. K. Costa, Eduardo A. B. da Silva

Electrical Engineering Program, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

ABSTRACT

Deep learning models excel in computer vision tasks with large labeled datasets but often struggle with performance degradation under domain shifts to unlabeled target domains. Unsupervised Domain Adaptation (UDA) mitigates this challenge by transferring knowledge from a labeled source domain to an unlabeled target domain. In this paper, we propose Self-Contrastive Learning for Domain Adaptation (SCoDA), a novel framework that combines self-supervised contrastive learning with distributional alignment techniques to learn domain-invariant representations. SCoDA jointly optimizes domain alignment and classification by using contrastive loss to reduce feature gaps while preserving task performance through source supervision. Experiments on UDA benchmarks demonstrate that SCoDA outperforms traditional baselines, achieving accuracy improvements without relying on target labels. These results highlight SCoDA's effectiveness in addressing complex domain shifts and its potential for real-world applications. Our code is available at https://github.com/viniavena/SCoDA.

***Index Terms*—** Domain adaptation, contrastive learning, self-supervision, feature alignment, representation learning.

1. INTRODUCTION

Deep learning has transformed computer vision, driving enormous advances in tasks such as image classification, object detection, and semantic segmentation. Nevertheless, these achievements mostly rely on large labeled datasets, which are costly and time-consuming to collect. Meanwhile, the abundance of unlabeled data generated by sensors and devices invites the use of self-supervised learning (SSL) [1], which enables models to extract meaningful representations from unlabeled data through auxiliary pretext tasks. This reduces the reliance on manual labeling and supports scalable solutions.

Despite the advances brought by SSL techniques, a key challenge remains: most models assume that training and testing data are drawn from the same distribution [2]. This assumption rarely holds in practice, as models face distribution shifts, caused by anything from small variations in lighting and resolution to drastic changes in background or environmental conditions, degrading performance [3]. Addressing these shifts without labeling target data is critical for robust and generalizable models.

Unsupervised Domain Adaptation (UDA) [4] seeks to bridge this gap by transferring knowledge from a labeled source domain to an unlabeled target domain. Recent advances in SSL, particularly contrastive learning (CL), have demonstrated promise in learning robust, transferable feature representations [5, 6]. Contrastive learning encourages positive sample pairs (e.g., augmentations of the same image) to cluster together while separating them from negative pairs, fostering domain-invariant features suitable for UDA.

This work presents *Self-Contrastive Learning for Domain Adaptation* (SCoDA), which integrates self-supervised contrastive learning and Maximum Mean Discrepancy (MMD) [7] to align source and target feature distributions while preserving task performance. Unlike prior methods, SCoDA simultaneously optimizes domain alignment and classification, overcoming the limitations of approaches that decouple these processes. Using a unified loss function and simultaneously training the encoder and classifier, SCoDA ensures more robust, domain-invariant representations.

The main contributions of this work are:

- a unified framework, SCoDA, that simultaneously optimizes domain alignment (via MMD) and task-specific learning (via contrastive and classification losses);
- a novel approach to contrastive learning using original and augmented images to enhance feature robustness and domain invariance; and
- validation of the proposal using UDA benchmarks, demonstrating significant improvements over state-of-the-art methods.

The remainder of this paper is structured as follows: Section 2 reviews related work in domain adaptation and contrastive learning. Section 3 details the proposed SCoDA framework, including its architecture and loss functions. Section 4 presents the experimental setup and performance

This study was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, CNPq grants 408255/2023-4, 309304/2021-0, and 310942/2021-7, FAPERJ grants E-26/204.122/2024s, E-26/204.562/2024 and E-26/201.021/2022, FAPESP grants 23/00673-7 and 23/00811-0, and Development and Research Foundation Fundep - Rota 2030.

results. Section 5 concludes the paper and outlines directions for future research.

2. RELATED WORK

Unsupervised Domain Adaptation (UDA) addresses performance degradation caused by distribution shifts between source and target domains. Adversarial UDA methods, such as DANN [8] and ADDA [9], promote domain invariance by balancing a domain classifier to distinguish the source from target features, while a feature encoder is optimized to confuse the classifier; however, they often face training instability. Similarly, approaches based on Generative Adversarial Networks (GANs) [10], which perform image-to-image translation, are computationally expensive and rely on complex min-max optimization, posing practical challenges in real-world scenarios.

Divergence-based methods, such as Maximum Mean Discrepancy (MMD) [11] and Wasserstein Distance [12], minimize feature space discrepancies to enforce alignment. MMD-based approaches, such as DRCN [13], combine feature alignment with target reconstruction to improve domain adaptation. Nonetheless, the success of these methods depends on high-quality embeddings that preserve relevant semantics while discarding domain-specific noise.

Self-supervised learning (SSL) offers a powerful alternative by learning representations without labeled data. The use of pretext tasks like rotation prediction [14] and jigsaw solving [15] has been extended to UDA, demonstrating improved transferability [16]. However, their reliance on carefully crafted pretext tasks limits generalization under severe domain shifts.

Contrastive learning (CL) enhances domain adaptation by promoting intra-class compactness and inter-class separability [17]. Contrastive Adaptation Networks (CAN) [18] use pseudo-labeling for alignment, but their dependence on clustering introduces significant noise. CDA [19] integrates contrastive and MMD-based alignment but does not jointly optimize feature extraction and task-specific training, leading to suboptimal results.

As summarized in Table 1, most frameworks in the literature focus on specific techniques, such as divergence minimization or contrastive learning, often without leveraging the combined strengths of these approaches. In contrast, the SCoDA framework combines MMD for domain alignment with self-supervised contrastive learning into a unified loss function. This enables simultaneous learning of domain-invariant representations and task-specific objectives, ensuring robust adaptation across diverse domains.

3. PROPOSED METHOD

SCoDA consists of three main components: an *encoder* $F(\cdot)$, a *domain alignment head* $G(\cdot)$, and a *classification head* $C(\cdot)$, as illustrated in Figure 1. The encoder extracts features from both source and target domain samples, while the domain alignment head calculates the Maximum Mean Discrepancy (MMD) to ensure statistical alignment between the feature distributions of the domains. The classification head predicts labels for the source samples using supervised learning, leveraging aligned features for improved task-specific performance.

Framework	MMD	SSL	CL	Joint Opt.
DANN [8]				✓
ADDA [9]				✓
Xu et al. [16]		✓		✓
CAN [18]			✓	✓
DRCN [13]	✓			✓
CDA [19]	✓	✓	✓	
SCoDA (ours)	✓	✓	✓	✓

Table 1: Comparison of domain adaptation frameworks. SCoDA combines MMD, self-supervised learning (SSL), and contrastive learning (CL) into a unified framework with joint optimization, addressing limitations of prior methods in domain alignment and task-specific learning.

SCoDA differs from traditional contrastive learning by forming positive pairs between the original image, used for classification, and an augmented version, used to enhance representation robustness. Data augmentations, inspired by the analysis in SimCLR [6] regarding the effectiveness of different transformations, include random cropping, color jitter, Gaussian blur, and grayscale conversion. These transformations preserve semantic content while introducing variability, promoting robust and transferable representations.

3.1. Contrastive learning strategy

Contrastive learning plays a central role in SCoDA, improving feature robustness and domain invariance by maximizing the similarity between positive pairs and minimizing it for negative pairs. In our implementation, a positive pair consists of an original image $x \in \{X_S, X_T\}$ with its augmented version $x^+ \sim \mathcal{T}(x)$. Negative pairs are formed from embeddings of different samples within the same mini-batch, encouraging the model to push apart representations of unrelated data points. This process reinforces the distinction between different features in the embedding space, enhancing the discriminative power of the learned representations.

Feature embeddings for the source and target domains, Z_S and Z_T, are extracted using the shared encoder $F(\cdot)$. For a given sample, each positive pair is represented as $z_i = F(x_i)$ and $z_i^+ = F(x_i^+)$. The contrastive loss for each domain is computed using the normalized temperature-scaled cross-entropy loss (**InfoNCE** [5]):

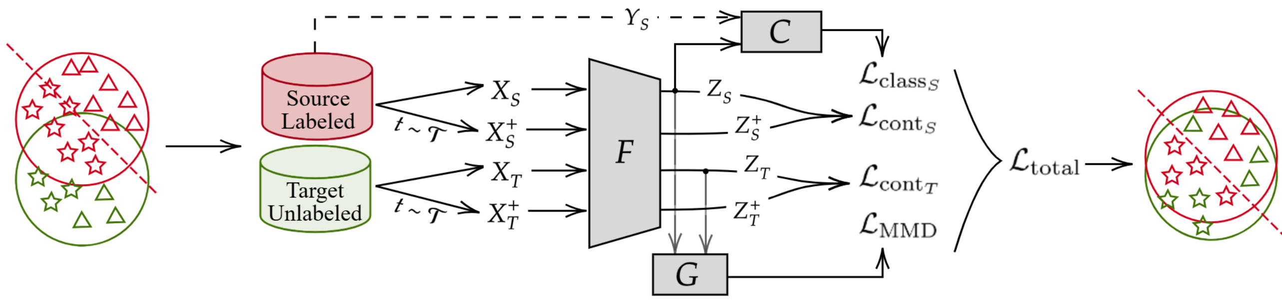

Fig. 1: Overview of the proposed SCoDA framework. For a given input image $x \in \{X_S, X_T\}$, features are extracted using a shared encoder $F(\cdot)$, yielding embeddings $Z_S = F(X_S)$ and $Z_T = F(X_T)$ for source and target domains, respectively. Positive pairs associate the original image x, used in supervised classification on Z_S via the classification head $C(\cdot)$, with its augmented version $x^+ \sim \mathcal{T}(x)$, employed in contrastive learning. The domain alignment head $G(\cdot)$ computes the Maximum Mean Discrepancy (MMD) to minimize the distance between Z_S and Z_T in a Reproducing Kernel Hilbert Space (RKHS) [20], ensuring domain invariance. The overall loss $\mathcal{L}_{\text{total}}$ effectively aligns the source and target domains.

$$\mathcal{L}_{\text{cont}} = -\log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_{j \neq i} \exp(\text{sim}(z_i, z_j)/\tau)}, \quad (1)$$

where z_i^+ is the embedding of the positive pair corresponding to z_i, z_j represents the embedding of a different sample in the batch, τ is the temperature parameter, and $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity, a standard metric to measure the alignment between two normalized vectors. By reinforcing the similarity of positive pairs and pushing apart negative pairs, the model learns a more robust and domain-invariant embedding space, facilitating domain adaptation.

3.2. Domain alignment via MMD

To align source and target feature spaces, we employ Maximum Mean Discrepancy (MMD) [7], a kernel-based method that measures distributional differences by comparing mean embeddings in a Reproducing Kernel Hilbert Space (RKHS). MMD operates as a statistical test to evaluate the similarity between the source and target distributions, under the assumption that aligned distributions share similar statistical properties in the RKHS.

Unlike Kullback-Leibler divergence [21], which requires precise probability density estimates, MMD compares distributions using kernel functions in a high-dimensional space, making it computationally efficient and suitable for scenarios with limited samples.

In SCoDA, MMD is applied to the feature representations extracted by the encoder $F(\cdot)$ to minimize domain discrepancies. Formally, MMD is defined as:

$$|\text{MMD}(Z_S, Z_T)| = \frac{1}{m^2}\sum_{i,j} k(z_i^S, z_j^S) + \frac{1}{n^2}\sum_{i,j} k(z_i^T, z_j^T) - \frac{2}{mn}\sum_{i,j} k(z_i^S, z_j^T), \quad (2)$$

where $Z_S = \{z_1^S, \ldots, z_m^S\}$ and $Z_T = \{z_1^T, \ldots, z_n^T\}$ are the features from the source and target domains, respectively, and $k(\cdot,\cdot)$ is a Gaussian Radial Basis Function (RBF) kernel [20]. By minimizing this metric, SCoDA enforces statistical alignment between the feature spaces, enabling domain-invariant learning.

3.3. Training procedure

The training process in SCoDA integrates supervised and self-supervised learning objectives to align source and target domains while optimizing task-specific performance. As illustrated in Figure 1, the total loss $\mathcal{L}_{\text{total}}$ combines cross-entropy for classification, contrastive loss for feature robustness, and MMD for domain alignment. To train the network with labeled source data, we minimize the cross-entropy loss, ensuring accurate task-specific learning in the source domain:

$$\mathcal{L}_{\text{class}} = -\frac{1}{N}\sum_{i=1}^{N} \log P_\theta(y_i|z_i), \quad (3)$$

where $P_\theta(y_i|z_i)$ denotes the predicted probability for the ground-truth label y_i given input features z_i, and N is the batch size. The embeddings used for classification are further refined through the contrastive objective, ensuring alignment between domain-invariant features and task-specific representations.

The overall training objective is defined as:

$$\min_\theta \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{class}} + \alpha\,(\mathcal{L}_{\text{cont}_S} + \mathcal{L}_{\text{cont}_T}) + \beta\,|\text{MMD}(Z_S, Z_T)|, \quad (4)$$

where θ represents the trainable parameters of the model. The hyperparameter α controls the contribution of the contrastive objective, encouraging robust feature representations, while β regulate domain alignment to mitigate distribution shifts.

Therefore, by processing batches containing labeled source samples and unlabeled target samples simultaneously, the framework effectively trains the encoder, alignment head, and classification head to produce domain-invariant and discriminative representations. Note that in the test phase, only the encoder $F(\cdot)$ and the classification head $C(\cdot)$ are used for label inference, while the contrastive and MMD branches are disabled.

3.4. Implementation details

Our SCoDA implementation employs ResNet-50 [22] as the base encoder, which is fine-tuned during training. The encoder is followed by a bottleneck consisting of a linear layer mapping the encoder's output to a 256-dimensional space, then a ReLU activation and a dropout layer for regularization. The classification head, a single linear layer, predicts class labels using the extracted features. The training was performed for 100 epochs using the AdamW optimizer with a learning rate of 5×10^{-4}, weight decay of 1×10^{-4}, and a batch size of 32, while a Cosine Annealing scheduler dynamically adjusted the learning rate.

4. RESULTS

4.1. Experimental setup

All experiments were conducted on well-established domain adaptation benchmarks based on digit classification datasets. As depicted in Figure 2, the selected benchmarks evaluate SCoDA's ability to handle domain shifts with variations in color, style, and resolution. The evaluated scenarios are ($A \rightarrow B$ indicates that A is the source and B the target domain):

- SVHN $\rightarrow$ MNIST (S$\rightarrow$M): from colorful street-view house numbers [23] to simple grayscale handwritten digits [24];
- MNIST $\rightarrow$ MNIST-M (M$\rightarrow$MM): adapting to images with color patches overlaid on noisy backgrounds [8];
- MNIST $\rightarrow$ USPS (M$\rightarrow$U): shifting to lower-resolution handwritten digits [25].

These benchmarks are widely used in domain adaptation research due to their well-defined challenges, such as significant variations in color, resolution, and style [8]. These diverse shifts help assess the framework's ability to learn domain-invariant representations and generalize to real-world scenarios

To contextualize the improvements achieved by the proposed method, a *source-only baseline* was established as a reference point. This baseline follows the same ResNet-50 architecture and optimization settings as SCoDA but is trained exclusively on the labeled source domain without domain adaptation techniques. It serves as a lower bound, highlighting the degradation in performance when domain shifts are not explicitly addressed.

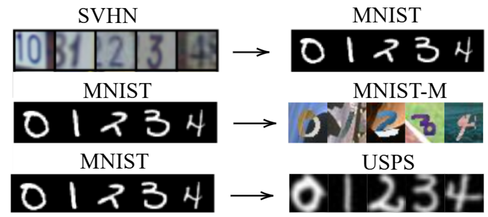

Fig. 2: Examples from the three domain adaptation benchmarks used in this study. On the left, source domains are represented; on the right, target domains.

4.2. Comparison with the state-of-the-art

We compare the performance of SCoDA with state-of-the-art unsupervised domain adaptation (UDA) methods on the selected benchmark datasets. Table 2 presents the accuracy results on the target domain, comparing SCoDA against various methods, including adversarial training (DANN [8], ADDA [9]), divergence minimization (MMD [11]), self-supervised (DRCN [13]), and contrastive learning (CDA [19]) approaches.

Framework	**S→M**	**M→MM**	**M→U**
Source-only	68.9	56.6	82.2
DANN [8]	73.8	76.7	90.8
ADDA [9]	76.0	-	89.4
MMD [11]	71.1	76.9	-
DRCN [13]	81.9	-	91.8
CDA [19]	76.2	60.2	94.2
SCoDA	**92.1**	**85.1**	**96.6**

Table 2: Comparison of related state-of-the-art domain adaptation methods on the selected benchmark datasets. Each result represents the accuracy achieved on the target domain, highlighting the effectiveness of different methods in handling specific domain shifts. A dash (-) indicates that the corresponding method did not report results for that source–target pair.

The results indicate that SCoDA achieves substantial accuracy improvements over the source-only baseline, with absolute accuracy gains of **23.2%**, **28.5%**, and **14.4%** for S$\rightarrow$M, M$\rightarrow$MM, and M$\rightarrow$U, respectively. These improvements confirm the framework's effectiveness in mitigating domain gaps. These improvements confirm that the framework effectively addresses domain shifts.

Notably, SCoDA consistently outperforms previous state-of-the-art methods across all evaluated tasks. For S$\rightarrow$M, it attains **92.1%**, outperforming DRCN (**+10.2%**). For M→U, it achieves **96.6%**, exceeding CDA by **2.4%**. On the challenging M→MM task, SCoDA reaches **85.1%**, significantly outperforming both DANN (76.7%) and MMD (76.9%) by over 8%.

This superior performance can be attributed to its joint optimization of domain alignment (via MMD) and feature discrimination (via contrastive learning). Unlike previous methods that decouple these processes, SCoDA aligns feature distributions while preserving class-discriminative properties, ensuring that domain-invariant representations remain useful for classification.

4.3. Ablation study

We conducted a sensitivity analysis to evaluate the impact of key components in the SCoDA framework, focusing on the SVHN → MNIST benchmark (S→M). In each configuration, exactly one component of the full framework is modified while the others remain fixed. The experiments include analyses of divergence minimization, pretext tasks, and the importance of joint optimization. Table 3 summarizes the results, grouping configurations into distinct themes.

Configuration	**Acc (%)**
SCoDA (full)	**92.1**
Without MMD	77.8
SWD [12] replacing MMD	85.3
Rotation prediction [16] replacing CL	78.7
Separate encoder-classifier optimization	65.1

Table 3: Ablation study showing target domain accuracy (%) results on the SVHN → MNIST domain adaptation task. Configurations are grouped into: divergence minimization – MMD [7] and SWD [12]; pretext tasks – contrastive learning (CL) [6] and rotation prediction [16]; joint vs separate encoder-classifier optimization strategies [19].

Primarily, we analyzed the impact of MMD in the divergence minimization step. Removing MMD caused a 14.3% accuracy drop, while substituting it with the SWD divergence metric [12] led to a reduction of 6.8%. However, SWD incurred a 26% increase in training time, making MMD the best choice for domain alignment.

Next, the impact of Contrastive Learning (CL) as the pretext task for the self-supervised training of the encoder was evaluated. Replacing CL with rotation prediction [16] reduced accuracy by 13.4%, indicating that simple pretext tasks are insufficient for capturing the rich semantic relationships needed for effective domain adaptation.

Lastly, we examined joint optimization, where the encoder and classifier are trained simultaneously. Training them separately, as in CDA [19], caused a sharp accuracy drop to 65.1%, even below the source-only baseline (68.9%). This highlights joint optimization's role in aligning domain-invariant features while preserving their relevance for classification, emphasizing the importance of end-to-end training.

The results of this ablation study confirm that each component of SCoDA plays a crucial role, and their integration is key to achieving state-of-the-art domain adaptation.

5. CONCLUSION

This paper introduced Self-Contrastive Learning for Domain Adaptation (SCoDA), a unified framework that integrates self-supervised contrastive learning with Maximum Mean Discrepancy (MMD) for domain alignment. By jointly optimizing domain adaptation and classification, SCoDA demonstrated significant improvements over existing methods in benchmark tasks, achieving accuracy gains of up to 19% over the source-only baseline. These results highlight the potential of SCoDA to address challenging domain shifts, ensuring robust and transferable representations across diverse scenarios.

Ablation studies provided deeper insights into SCoDA's key components. The incorporation of MMD was shown to significantly reduce domain discrepancies, while the use of contrastive learning as a pretext task enabled the model to learn rich, domain-invariant representations. Additionally, joint optimization of the encoder and classifier proved essential for aligning feature spaces with task-specific objectives.

In future work, we will extend SCoDA to more complex benchmarks and tasks beyond classification, such as semantic segmentation.A key research direction is reducing false negatives in self-supervised contrastive learning to improve feature robustness. In addition, the integration of pseudo-labeling strategies could refine target predictions iteratively, leveraging both labeled source and pseudo-labeled target data to enhance domain adaptation.

6. REFERENCES

[1] Longlong Jing and Yingli Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.

[2] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2, Springer, 2009.

[3] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete,*

Greece, September 5-11, 2010, Proceedings, Part IV 11. Springer, 2010, pp. 213–226.

[4] Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, Jonghye Woo, et al., "Deep unsupervised domain adaptation: A review of recent advances and perspectives," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

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

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[7] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[9] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[10] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 1989–1998.

[11] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[12] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10285–10295.

[13] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," 2016.

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[15] Mehdi Noroozi and Paolo Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*. Springer, 2016, pp. 69–84.

[16] Jiaolong Xu, Liang Xiao, and Antonio M López, "Self-supervised domain adaptation for computer vision tasks," *IEEE Access*, vol. 7, pp. 156694–156706, 2019.

[17] Jeff Z HaoChen, Colin Wei, Ananya Kumar, and Tengyu Ma, "Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations," *Advances in neural information processing systems*, vol. 35, pp. 26889–26902, 2022.

[18] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4893–4902.

[19] Mamatha Thota and Georgios Leontidis, "Contrastive domain adaptation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2021, pp. 2209–2218.

[20] B Schölkopf, "Learning with kernels: support vector machines, regularization, optimization, and beyond," 2002.

[21] Solomon Kullback and Richard A Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al., "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*. Granada, 2011, vol. 2011, p. 4.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] J.J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, 1994.