SADA: Say Less
Sparse Auto-encoder for Transformer Based Unsupervised Domain Adaptation

Muhammad Aarash Abro
Department of Computer Science
Lahore University of Management Sciences
25100330@lums.edu.pk

Muhammad Usman Ahmad
Department of Computer Science
Lahore University of Management Sciences
25100160@lums.edu.pk

Shaheer Ahmad
Department of Computer Science
Lahore University of Management Sciences
25100184@lums.edu.pk

Abstract

Unsupervised Domain Adaption (UDA) aims to extract domain invariant knowledge from labeled source domains to improve the performance of unlabeled target domains. Previously, Convolution Neural Network (CNN) based architecture dominated the field; recent findings have shown a significant improvement in using a Vision Transformer for this task. Another recent field includes using Sparse Auto-encoder in Language-related Tasks, showing promising results in feature extraction [Cunningham et al. [2023]]. Our study proposes a novel transformer-based approach coupled with a Sparse Autoencoder to tackle the UDA problem. Our method introduces two key factors: First, the integration of Sparse Auto-encoder to generate a sparse representation of intermediate feature representation for extracting domain-invariant features. Second, we coupled the Sparse auto-encoders with Invariant Risk Management (IRM), introducing a gradient penalty for better domain invariant learning. These two components work in synergy to enhance domain-invariant feature representation learning—detailed experiments were done on Office-31, PACS, and Office-Home. In addition, we tried different transformer-based backbones to prove generality and demonstrate that our method shows significant UDA improvement, achieving very promising results. For more details, the implementation is available on GitHub: https://github.com/mani-ahmad/project-say-less.git

1 Introduction

In the evolving field of universal domain adaptation, generalising machine learning models across diverse data distributions remains a significant challenge. Datasets in the real world suffer from domain shifts, where the statistical properties of training data differ from those encountered during deployment. These discrepancies can degrade model performance, limiting its applicability in real-world tasks. Traditionally, models have been trained using the ERM paradigm, which minimises average loss on training data, assuming that training and test distributions are identical. However, ERM fails under domain shifts, as it tends to capture superficial correlations rather than casual relationships in the data

CS-6304 Advanced Topics in Machine Learning
Template from NeurIPS 2024

Invariant Risk Minimization (IRM) was introduced to address this issue, aiming to uncover stable, domain-invariant features across diverse environments. While IRM is a theoretically sound framework for learning invariant representations, it faces multiple practical challenges. For example, high computational overhead and difficulties in fully disentangling domain-specific noise from meaningful invariant features. As a result, enhancing feature extractors to prioritize domain-invariant information remains the primary focus in domain adaptation. Another approach, **Domain-Adversarial Training**, employs adversarial objectives to align feature distributions between source and target domains, effectively reducing domain discrepancy [Ganin et al. [2016]]. **Transfer Learning** techniques focus on transferring knowledge from a related source domain to a target domain, often through fine-tuning pre-trained models on target data [Pan and Yang [2010]]. Recent advancements in **Meta-Learning** have also been applied to domain adaptation, enabling models to rapidly adapt to new domains with minimal data by learning to learn across various tasks. Each method offers unique advantages and challenges in the quest for effective domain adaptation [Finn et al. [2017]].

Recent advancements in **Sparse Autoencoders (SAEs)** have demonstrated their effectiveness in text-based tasks, particularly in uncovering interpretable representations from transformer-based language models such as GPT-4 [Gao et al. [2024]]. By leveraging sparsity, SAEs reduce redundancy, improve feature selectivity, and enhance interpretability, making them a powerful tool for analyzing high-dimensional activations. Inspired by these successes, we hypothesize that applying SAEs to image-based intermediate feature representations could offer similar benefits. Specifically, we propose integrating SAEs with multi-scale image feature extractors, such as **Convolutional Neural Networks (CNNs)** and **Vision Transformers (ViTs)**, to enforce sparsity and guide the models toward domain-invariant feature learning.

Our proposed methodology involves a multi-step process. First, we extract multi-scale feature representations from input images using backbone architectures, including **VGG**, **ResNet**, **ViT Base**, and **SWIN-ViT**. These intermediate features are then passed through a **Sparse Autoencoder (SAE)** layer, where sparsity is enforced using **L1 regularization** and reconstruction losses. Next, the SAE outputs are optimized using a **unified loss function**, combining **Invariant Risk Minimization (IRM)**, **L1 sparsity constraints**, and SAE reconstruction objectives. Additionally, we explore multiple **aggregation strategies** for these sparse representations, including **concatenation**, **attention-based aggregation**, and **weighted averaging**, to identify the most effective way to combine features across layers.

For evaluation, we will conduct experiments on publicly available benchmark datasets commonly used for domain adaptation tasks, such as **Office-31** [Saenko et al. [2010]], **PACS** [Li et al. [2017]], and **Office-Home** [Venkateswara et al. [2017]]. These datasets provide diverse domain pairs (e.g., webcam-to-DSLR, art-to-photo) that will allow us to assess the robustness and generalization of our approach under varying domain shifts.

The anticipated outcomes of our work include:

- **Improved Accuracy Under Domain Shifts:** Sparse representations are expected to improve model robustness across diverse domains.
- **Reduced Training Time and Computational Overhead:** Our integration of sparsity and IRM is designed to converge faster compared to traditional domain adaptation frameworks.
- **Enhanced Interpretability of Learned Features:** Sparse representations provide clearer, more interpretable insights into the invariant features learned by the model.

By addressing computational efficiency and domain-invariant feature extraction challenges, our framework aims to bridge the gap between theoretical advancements in sparsity and their practical applications in domain adaptation tasks. This research improves domain adaptation performance and advances our understanding of how sparse representations can enhance cross-domain generalization in computer vision models. This work builds on insights from text and vision-based sparse representation learning, bridging the gap between theoretical advancements in sparsity and their practical applications in domain adaptation tasks.

2 Related Work

There have been significant advancements in the field of domain adaptation in computer vision with the rise of robust feature extraction backbones, such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). These architectures have entirely revamped how image representations are learned, offering unique feature extraction and downstream generalization strengths. Additionally, techniques like Sparse Autoencoders (SAEs) have shown remarkable promise in enforcing feature sparsity and improving representation interpretability. This section discusses these three pillars—ViTs, CNNs, and SAEs—in the context of domain adaptation and highlights their contributions, limitations, and relevance to our work.

Convolutional Neural Networks (CNNs) have long been the cornerstone of image feature extraction, owing to their ability to learn hierarchical spatial features through convolutional filters. Introducing architectures like VGG [Simonyan and Zisserman [2015]] and ResNet marked significant milestones. ResNet, introduced by He et al. in 2015 [2015 He et al. [2016]], addressed the degradation problem in deep networks by implementing residual learning, allowing for the training of substantially deeper models. However, CNNs only have a limited receptive field, which restricts their capacity to model long-range dependencies within images [Romero et al. [2023]]. This limitation is more significant in domain adaptation tasks, where global context is highly important for extracting invariant features across domains. While fine-tuning and domain adversarial training techniques have been applied to CNNs to bridge domain shifts, the local receptive field often constrains their ability to fully disentangle domain-specific artefacts from invariant features. Consequently, CNNs remain effective but gradually outperform architectures capable of capturing global relationships.

In contrast, **Vision Transformers (ViTs)** have introduced a paradigm shift in image feature extraction by leveraging self-attention mechanisms, allowing them to capture global dependencies across an entire image. Dosovitskiy introduced the Vision Transformer, demonstrating that transformers could achieve state-of-the-art performance on image recognition tasks without convolutional layers. ViTs treat image patches as sequences and model their relationships through multi-headed self-attention, enabling a more holistic understanding of visual information [Dosovitskiy et al. [2021]]. Architectures like ViT Base and SWIN-ViT have demonstrated superior performance across visual benchmarks, including domain adaptation tasks. SWIN-ViT in particular, incorporates a hierarchical transformer design and shifted window mechanisms to effectively balance global and local contexts [Liu et al. [2021]]. However, ViTs are challenging—they often require large-scale datasets and significant computational resources to achieve optimal performance [Chen et al. [2021]]. Moreover, while ViTs are highly effective in extracting global features, they may still inadvertently encode domain-specific biases in their representations, limiting their robustness in cross-domain tasks [Alijani et al. [2024]].

Sparse Autoencoders (SAEs) have emerged as powerful tools for learning compact, interpretable, and domain-agnostic representations, especially in high-dimensional data settings. SAEs have demonstrated their ability to reduce redundancy, encourage feature selectivity, and improve interpretability in language models such as GPT-4 [Gao et al. [2024]]. Their capacity to enforce sparsity on latent representations minimizes noise and focuses on the most informative features. Despite their success in text-based domains, the application of SAEs to vision tasks remains underexplored Chen et al. [2023]. Our work builds upon the intuition that SAEs, when integrated with robust backbone architectures such as CNNs and ViTs, can encourage the extraction of domain-invariant features by minimizing irrelevant domain-specific variations [Chen et al. [2021]]. Furthermore, sparse representations are computationally efficient and often converge faster, addressing some of the limitations of traditional domain adaptation methods such as IRM and domain adversarial training [Ahmad and Scheinkman [2019], Chen et al. [2023]].

The evolution of feature extraction architectures in computer vision has transitioned from localized to global representation learning, with **Convolutional Neural Networks (CNNs)** excelling at capturing hierarchical spatial patterns and **Vision Transformers (ViTs)** introducing self-attention mechanisms to model global dependencies effectively. While CNNs, exemplified by architectures like **ResNet** [Romero et al. [2023]] In contrast, ViTs, introduced by Dosovitskiy et al. [2021]. Addressing these biases and resource constraints remains an ongoing challenge [Chen et al. [2021]]. Alongside these advancements, Sparse Autoencoders (SAEs) have demonstrated immense promise in enforcing sparsity on latent representations, reducing redundancy, and improving computational efficiency [Chen et al. [2023], Ahmad and Scheinkman [2019]]. However, their integration with image-

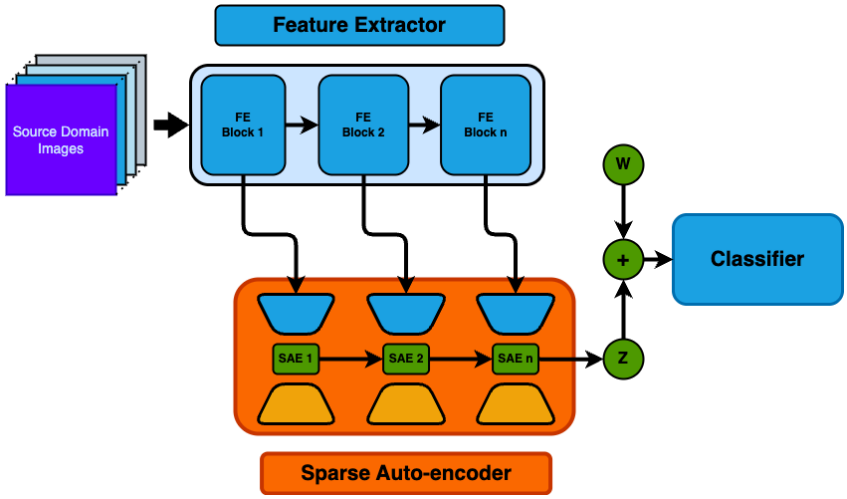

Figure 1: Architecture of the proposed domain adaptation model. The framework begins with a Feature Extractor Backbone (e.g., ViT, SWIN-ViT, or CNN) that generates Intermediate Feature Representations. Sparse Autoencoders (SAEs) enforce sparsity, reducing redundancy and enhancing domain-invariant feature selectivity. These refined features are passed to a Classifier for final predictions. The model is optimized using a Combined Loss Function, integrating Invariant Risk Minimization (IRM), L1 Sparsity, and SAE Reconstruction Loss, ensuring robust performance across diverse domain shifts.

based architectures, especially in the context of domain adaptation, remains underexplored [Chen et al. [2023]]. By combining SAEs with both CNN and ViT backbones, we leverage sparsity to enforce domain invariance while retaining the unique strengths of each architecture. In this work we present a framework to improve robustness and generalization across domain shifts by addressing the constraints of CNNs, ViTs, and lack of implementation of SAEs in vision tasks.

3 Methodology

Let source domain $\mathcal{D}_s = \{(X_{s_i}, Y_{s_i})\}_{i=1}^{n}$ be a dataset containing labelled data where X_{s_i} represent images and Y_{s_i} there respective labels and n denotes the number of sample in the source domain. Meanwhile, $\mathcal{D}_t = \{X_{t_i}\}_{i=1}^{m}$ represents the target domain with X_{t_i} representing the images in the target domain, containing a total of m samples. Unsupervised Domain Adaption (UDA) aims to minimize the domain shift between source and target, achieving the desired performance on the unlabeled target domain. Our project combines CNN, Transformer Feature extractor, and Sparse Auto-encoder to learn domain invariance through Invariant Risk Minimization (IRM). Figure 1 shows the model for our method.

3.1 Feature Extraction

The feature extractor plays the most vital role in any downstream task, meaningfully representing our input data. We have used two types of backbone, Convolution Neural Networks (CNNs) and Vision Transformers, to study the impact of different architectures on our method.

For CNN-based backbones, we extracted intermediate and final feature representations. For Vision Transformers, we utilized only the output of the final layer as it provides refined global features that are most representative of the input.

Formally, let $x \in \mathbb{R}^d$ represent an input sample from either the source or target domain. A backbone network processes x to produce intermediate and final representations:

$$F = \{f_1, f_2, \ldots, f_k\}, \quad f_i \in \mathbb{R}^{d_i} \tag{1}$$

where f_i is the feature representation extracted at the i-th layer, and k denotes the total number of selected layers.

By comparing the performance across backbones, we demonstrated that our methodology aligns with the different capabilities of different architectures.

3.2 Sparse Feature Representations

To enhance domain-invariant feature learning, the intermediate and final representations $F = \{f_1, f_2, \ldots, f_k\}$ were passed through a Sparse Auto-encoder (SAE). The SAE comprises an encoder-decoder structure designed to produce sparse representations while maintaining the embedding space dimensions of the input features. The encoder maps f_i to a latent space z_i, and the decoder reconstructs $\hat{f}_i$ from z_i, ensuring that the input embedding space is preserved.

The forward pass of the SAE can be represented as:

$$z_i = \text{Encoder}(f_i), \quad \hat{f}_i = \text{Decoder}(z_i) \tag{2}$$

To train the SAE, a combination of reconstruction loss and ℓ_1- sparse penalty was employed:

$$\mathcal{L}_{\text{SAE}} = \|f_i - \hat{f}_i\|_2^2 + \lambda \|z_i\|_1 \tag{3}$$

where λ controls the sparsity of the latent representations. We theorize that this sparse representation preserves only the most essential and domain-invariant features are preserved, reducing noise and improving generalization.

3.3 Feature Aggregation

After obtaining sparse representations $Z = \{z_1, z_2, \ldots, z_k\}$, we aggregated these features using different strategies tailored to the type of backbone:

Weighted aggregation was done for CNN-based backbones (ResNet50, VGG19). Learnable weights were initialized to each feature representation, and these weights were passed through a softmax function to ensure stability:

$$z_{\text{agg}} = \sum_{i=1}^{k} \alpha_i z_i, \quad \alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{k} \exp(w_j)} \tag{4}$$

where α_i are the normalized weights and w_i are learnable parameters.

For Transformer-based backbones (ViT Base, Swin Base, DeiT Base), we observed that the α of the final layer becomes much greater than the rest. Hence, only the final layer's output was used for aggregation.

3.4 Invariant Risk Minimization

We incorporated Invariant Risk Management (IRM) to reinforce our approach further. IRM presents a gradient penalty on top of ERM loss, preventing the model from overfitting on the source domain. Using the gradient penalty, IRM enforces domain invariance.

The IRM loss has two parts: First, the Empirical Risk Minimization (ERM) term, which computes the cross-entropy loss over source data:

$$\mathcal{L}_{\text{ERM}} = -\frac{1}{n} \sum_{i=1}^{n} Y_{s_i} \log \hat{Y}_{s_i} \tag{5}$$

where Y_{s_i} and $\hat{Y}_{s_i}$ denote the true and predicted labels, respectively. Second, the IRM penalty, which imposes a gradient constraint to reduce sensitivity to domain-specific features:

$$\mathcal{L}_{\text{IRM}} = \|\nabla_\theta \mathcal{L}_{\text{ERM}}\|^2 \tag{6}$$

where θ represents the model parameters.

The final loss function combines these terms, weighted by empirically determined coefficients α and β:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{ERM}} + \beta \mathcal{L}_{\text{IRM}} \tag{7}$$

The coefficients α and β were selected based on empirical evidence to provide the best accuracy across all domains within the same dataset.

3.5 Evaluation Protocol

During training, only the source domain is used. The model was later evaluated on the target domain to study its domain generalization abilities. We evaluated our model on current standard UDA datasets, which include PACS, Office-31, and Office-Home.

4 Experimental Design

To assess the robustness and generality of our method, we evaluated our methodology and research questions by analyzing the performance across different datasets, backbones, and configurations.

4.1 Research Questions

Four key research questions guided our experiments:

RQ1: What is the impact of Sparse Autoencoders (SAEs) in preserving critical domain-invariant features across layers?

RQ2: How do different model families compare as feature extraction backbones for domain adaptation (e.g., CNN vs. Vision Transformers)?

RQ3: How efficient is our approach in terms of training time and convergence compared to state-of-the-art (SOTA) methods?

RQ4: What roles do IRM and SAE play individually and in combination in improving domain invariance?

4.2 Experimental Setup

We evaluated our method on three widely used UDA benchmark datasets: Office-31, PACS, and Office-Home. Office-31 comprises 31 categories across three domains (Amazon, DSLR, and Webcam) and represents a classical UDA benchmark. With its four domains (Photo, Art, Cartoon, Sketch) and seven categories, PACS challenges generalization across diverse styles. Office-Home features 65 categories across four domains (Art, Clipart, Product, Real-World) and introduces higher variability. All images were resized to 224×224 and normalized.

We used CNN-based backbones (ResNet50 and VGG19) and Vision Transformers (ViT Base, Swin Base, DeiT Base) to test generality across architectures. For CNNs, intermediate and final feature representations were extracted since earlier layers capture low-level patterns like edges and textures crucial for domain adaptation. In contrast, only the final layer was used for Vision Transformers, as its attention mechanism aggregates refined global features.

The sparse feature representations extracted from these backbones were aggregated using three strategies to determine the best method: naive concatenation, weighted aggregation (with learnable weights normalized via softmax), and an attention-based approach where latent representations were passed through an attention module before concatenation. Weighted aggregation was ultimately selected for CNN-based models, while the last layer's output sufficed for Vision Transformers due to their inherent ability to integrate global information.

4.3 Invariant Risk Minimization and Ablation Studies

To address **RQ4**, we conducted ablation studies to evaluate the contributions of IRM and SAE to domain invariance. Four configurations were tested: (1) a baseline Vision Transformer (ViT Base) without IRM or SAE, (2) ViT Base with IRM only, (3) ViT Base with SAE only, and (4) ViT Base with both IRM and SAE (our proposed method). The IRM component penalized gradients to reduce domain-specific sensitivity, while the SAE ensured sparse, domain-invariant representations. The combined use of IRM and SAE was hypothesized to provide complementary advantages.

Table 1: Comparison of Methods on Office-31 Dataset

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
Base ViT	79.31	98.39	96.60	79.12	54.27	52.89	76.76
Base DeIT	80.92	**100**	98.61	78.74	62.40	57.18	79.64
Base Swin	88.35	**100**	99.14	87.79	70.74	70.50	86.09
Base Resnet-50	61.11	95.83	97.92	72.92	51.79	57.32	72.82
SSRT (ViT)	48.48	70.26	94.18	72.90	58.60	74.12	69.76
SADA-Resnet-50 (Ours)	72.09	95.58	95.09	61.13	51.05	53.46	71.40
SADA-DeIT (Ours)	81.33	98.59	91.45	81.38	68.83	56.12	79.62
SADA-Swin (Ours)	92.97	99.80	98.99	**91.70**	77.64	76.93	89.67
SADA-ViT (Ours)	**93.37**	**100**	**99.75**	88.69	**80.37**	**82.00**	**90.70**
SSRT (ViT-GRL)	**97.70**	99.20	**100**	**98.60**	**83.50**	**82.20**	**93.53**

Table 2: Comparison of Methods for PACS Dataset

	P→A	P→C	P→S	A→P	A→S	A→C	C→A	C→S	C→P	S→A	S→P	S→C	Avg
Base ViT (%)	70.7	30.72	21.5	96.28	53.32	68.38	75.29	**46.42**	85.56	50.1	49.94	50.51	58.23
SADA-ViT (%)	**73.39**	**41.08**	31.89	**99.58**	**59.12**	**74.7**	86.47	45.93	**96.71**	70.8	45.21	**63.91**	**65.73**
SSRT (%)	73.17	39.2	**32.85**	98.56	24.73	64.8	**90.2**	23.74	92.69	**73.14**	**69.4**	58.95	61.79

4.4 Efficiency and Evaluation Metrics

To answer the **RQ3**, we measured the average training time for convergence of our method against the state-of-the-art approaches, such as SSRT [Sun et al. [2022]], under the same GPU (A100) system. Optimal hyperparameters for both methods were used to ensure further fairness.

Model performance for domain invariance was tested based on accuracy on the target domain. Results from the ablation study show the contribution of IRM and SAE in our novel approach. Aside from these, the model efficiency was measured by taking the average across multiple training samples.

4.5 Plan for Analysis

The results were analyzed to answer the research questions comprehensively. We compared performance across CNN and Transformer-based backbones to evaluate generalizability (**RQ2**), assessed aggregation strategies for their impact on feature preservation (**RQ1**), and measured the efficiency of our method relative to SSRT (**RQ3**). Ablati n studies quantified the individual and combined contributions of IRM and SAE to domain invariance (**RQ4**). Finally, by testing on three datasets (Office-31, PACS, and Office-Home), we ensured that our findings were robust and not dataset-specific.

5 Results and Findings

Through our experiments, we find that using sparse autoencoders along with IRM presents itself as a strong candidate for unsupervised domain adaptation (UDA). Though SADA is still an immature technique, it exceeds all baseline measures and competes with state-of-the-art (SOTA) methods in some benchmarks. Our initial results, as shown in Table 1, indicate that the SADA framework with a pre-trained ViT outperforms conventional domain adaptation techniques such as DAN and DANN. Moreover, SADA also outperforms SOTA techniques such as SSRT; however, this is only under the condition that the SSRT backbone is also ViT, instead of ViT-GRL (adversarially trained backbone).

Tables 2 and 3 show the performance of SADA on the PACS and Office-Home datasets, respectively. These results demonstrate the consistent performance of our technique across datasets, displaying strong potential for scaling and generalization, and confirming the feasibility of this approach as per our first research question.

Furthermore, by comparing the performance of different model families as the backbone of our framework, we can clearly identify the strengths and weaknesses of each architecture. When using ResNet-50 as the backbone, SADA produces promising results. However, it reveals a large gap in

Table 3: Comparison of Models on Office-Home Dataset

	A→C	A→P	A→R	C→P	C→R	C→A	P→R	P→A	P→C	R→A	R→C	R→P	Avg
Base SWIN (%)	55.37	75.82	81.68	75.71	77.27	71.07	82.51	67.90	50.76	75.81	55.71	85.55	71.26
SADA-SWIN (%)	**69.51**	**83.4**	**86.32**	**84.61**	**85.91**	**79.93**	**87.81**	**76.39**	**64.65**	**81.95**	**69.44**	**89.32**	**79.94**

performance from the baseline whenever it fails to generalize to a domain, indicating low capability for generalization.

On the other hand, using models from the ViT family as a backbone produces far more promising results. Using ViT-Base and SWIN as the backbone yields especially consistent results in outperforming the baseline, making it a far more robust model family for UDA. We the rize that this is due to the global receptive field of the attention mechanism, which facilitates learning more general feature representations, thereby translating to domain-invariant feature representations.

We have also visualized the alignment between source and target domain features using t-SNE diagrams, as presented in Appendix A.2. These diagrams demonstrate that our models are competent at capturing domain-invariant information and eliminating domain shifts from the feature space.

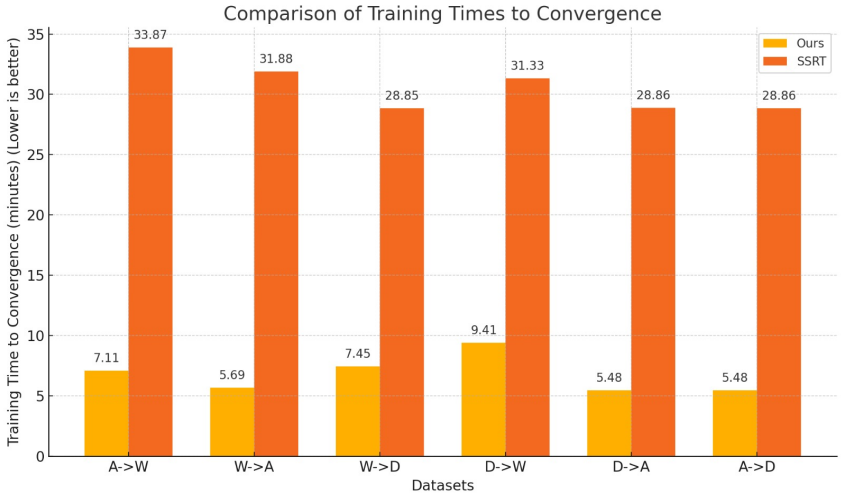

Figure 2: Comparison of training times

We also compared the training time/cost and computational overhead incurred during the training of SADA against SOTA techniques such as SSRT. We found that SADA almost always converges within 10 epochs or less and incurs, on average, 77.9% less training time to reach convergence with a batch size of 32 on an NVIDIA A100 GPU. Figure 2 clearly illustrates the drastic difference in training times between the two methods, leading us to conclude that SADA is a far more efficient and low-cost framework for UDA.

Due to time constraints, we could not use an adversarially trained ViT-GRL backbone in SADA. However, this could significantly improve current performance, enabling real competition with SOTA techniques such as FFTAT [Yu et al. [2024]].

In conclusion, while SADA is still not a mature technique, it shows abundant room for improvement to become an efficient replacement for current SOTA techniques.

6 Discussion

Unsupervised domain adaptation is an open problem in machine learning that has a far more practical impact than almost any other problem area in fields such as autonomous driving, medical imaging, and remote sensing. We designed SADA as an efficient technique to obtain promising results in terms of adapting to different domains. Through our experiments, we show that sparse auto-encoders, leveraging IRM and a robust feature extractor backbone, can succesfully learn to filter out noise and

domain-dependent information from a feature vector. This i made especially clear and prominent due to the fact that the latent dimension for the autoencoder remains the same as the input dimension. Therefore, the model cannot rely on dimensionality reduction to discard noise and domain-dependent features. Instead, by enforcing sparisty on such a latent forces the model to only keep relevant and important information, leading to domain transferable representations that generalize well across multiple domains.

This sparsity driven mechanism has proven to be robust in pushing a feature extractor to highlight integral patterns in the data while ignoring/discarding domain dependent information. All of this can be achieved without having to depend on target domain input data (unsupervised domain adaptation assumes target domain labels are not present). Current feature alignment techniques such as DAN (Domain Adaptation Network) [Long et al. [2015]] use measures such as MMD (maximum mean discrepancy) to align source and target domains. Furthermore, SOTA techniques like SSRT heavily rely on data manipulation techniques. In particular, they add perturbations to the input data in order to amplify the model's performance by aligning the feature representations of two perturbed variants of the same image. However, SSRT also has to rely on their safe training mechanism due to this, since perturbation beyond a certain point could lead to a model collapse, and this threshold could change during training, which would lead to severly unpredictable results. Their safe-training mechanisms mitigates this quite well, however it may lead to inefficient training. Lastly, SSRT and other SOTA all utilize a ViT-GRL backbone, that is trained adversarially through a domain discriminator. Again, normally this may lead to unstable results (as is the training of GANs).

At the moment, SADA does not employ any of the above-mentioned techniques. In particular, we don't utilize target domain data, neither inputs nor labels. However, going forward, we'd like to implement a mechanism that utilizes target domain input data to align feature space. The t-SNE diagrams of SADA models show that there is some discrepancy present between source and target domain distributions. It may be possible to further refine and align source and target domains using discrepancy-based feature alignment techniques. Moreover, we could also utilize data perturbation to a certain extent (as to avoid model collapse) to further improve performance.

Another interesting aspect of SADA is when we use SAEs and IRM in isolation. The gradient term in IRM loss acts as a regularizer, ensuring slower and less drastic weight updates so as to not overfit to training data. This does lead to general improvements over the baseline performance. Similarly, sparse autoencoders, when used in isolation, lead to significant performance gains, which hints that the sparsity-driven approach does learn domain invariant features well. However, when combined with IRM, this leads to further performance gains. This could perhaps mean that sparse autoencoders on their own are too keen to adapt to input data, and need a regularizer to slow them down and fit well to the overall distributions.

Another promising avenue for improvement is the integration of adversarially trained backbones into our framework [Ganin and Lempitsky [2015]]. Such backbones have become a standard component of many leading domain adaptation techniques and are supported by strong theoretical foundations. An adversarially sparse hybrid architecture, combining sparsity enforcement with adversarial training mechanisms, could further enhance the robustness of domain-invariant features while preserving computational efficiency. Additionally, newer variants of Sparse Autoencoders (e.g., Top-K Sparse Autoencoders) warrant investigating whether different sparsity mechanisms can refine feature representations more effectively.

Resource constraints prevented us from scaling our experiments to larger datasets like DomainNet. Larger and more diverse datasets could provide stronger empirical evidence of SADA's consistency, scalability, and robustness across varying levels of domain shifts. Addressing this limitation will be a priority in future work.

From a broader perspective, SADA's contributions go beyond just experimental benchmarks, providing significant practical benefits through its ability to efficiently extract domain-invariant features without heavily depending on labelled target data. In real-time autonomous systems, where obtaining labelled target domain data is often impractical, methods like SADA can help close the performance gap between training and deployment environments. Its computational efficiency also makes it ideal for resource-constrained applications, such as those running on edge devices. However, it is essential to recognize the ethical and societal implications of unsupervised domain adaptation techniques. Models used in diverse real-world settings must be thoroughly evaluated for potential biases and fairness issues, especially with datasets gathered from various geographic, cultural, or

socioeconomic backgrounds. Future research should include an ethical evaluation framework to ensure the responsible use of domain adaptation models. SADA significantly advances domain adaptation research by effectively merging sparsity-driven feature refinement with strong backbone architectures. While our results are encouraging, future efforts—like integrating target domain data, investigating adversarial backbones, utilizing advanced sparse autoencoder variants, and expanding experiments to larger datasets—will further enhance the framework's impact on the wider field of domain adaptation.

In conclusion, SAD represents a meaningful step forward in domain adaptation research by successfully combining sparsity-driven feature refinement with robust backbone architectures. While our findings are promising, future directions—such as integrating target domain data, exploring adversarial backbones, adopting advanced sparse autoencoder variants, and scaling experiments to larger datasets—will further solidify the framework's contributions to the broader field of domain adaptation.

7 Conclusion

We tackled significant challenges in unsupervised domain adaptation by exploring how Sparse Autoencoders (SAEs) can improve the extraction of domain-invariant features. Our findings indicate that SAEs effectively maintain essential domain-invariant features across different layers (RQ1). Additionally, we found that Vision Transformers surpass CNNs in capturing global dependencies for domain adaptation (RQ2), and our method demonstrates competitive training efficiency and convergence when compared to leading techniques (RQ3). Moreover, we discovered that while IRM aligns features between domains, SAEs enhance them, and using both together yields synergistic advantages, boosting overall domain invariance (RQ4). These results underscore the promise of merging sparsity-driven feature refinement with strong feature extraction frameworks for scalable and efficient domain adaptation. We plan to incorporate target domain data, utilize adversarial training frameworks and assess larger datasets to further improve the adaptability and robustness of our proposed approach.

7.1 Contributions

- **Muhammad Aarash Abro:** Original ideation and implementation of SADA architecture + thorough analysis and experimentation of further ideas, pulled benchmarks for
- **Muhammad Usman Ahmad:** Worked on improvement and testing of the implementation. Gave the idea of weighted aggregation and found the hyperparameters. Did the main testing on ViT, Swin, and Diet across Office-31, Office-Home, and PACS and the ablation study.
- **Shaheer Ahmad:** Address any limitations of the project, such as computational restrictions, data constraints, or assumptions that might limit generalizability. Discussing these limitations shows critical reflection and contextualizes the findings accurately + Worked on Experimentation of backbone + Implemented our backbone with state of the art technique + experimented with Adverserial Approach.

References

Subutai Ahmad and Scott Scheinkman. How can we be so dense? the benefits of using highly sparse representations. *arXiv preprint arXiv:1903.11257*, 2019. URL `https://arxiv.org/abs/1903.11257`.

Shadi Alijani, Jamil Fayyad, and Homayoun Najjaran. Vision transformers in domain adaptation and domain generalization: a study of robustness. *Neural Computing and Applications*, 36:17979–18007, 2024. URL `https://link.springer.com/article/10.1007/s00521-024-10353-5`.

Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, and Zhangyang Wang. Chasing sparsity in vision transformers: An end-to-end exploration. In *Advances in Neural Information Processing Systems*, 2021. URL `https://arxiv.org/abs/2106.04533`.

Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer. *arXiv preprint arXiv:2303.17605*, 2023. URL https://arxiv.org/abs/2303.17605.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. URL http://proceedings.mlr.press/v70/finn17a.html.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation, 2015. URL https://arxiv.org/abs/1409.7495.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL http://jmlr.org/papers/volume17/15-239/15-239.pdf.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

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*, pages 770–778, 2016.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks, 2015. URL https://arxiv.org/abs/1502.02791.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

David W. Romero, David M. Knigge, Albert Gu, Efstratios Gavves, Erik J. Bekkers, Jakub M. Tomczak, Mark Hoogendoorn, and Jan-Jakob Sonke. Modelling long range dependencies in n-d: From task-specific to a general purpose cnn. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZW5aK4yCRqU.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pages 213–226. Springer, 2010.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2015.

Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation, 2022. URL https://arxiv.org/abs/2204.07683.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

Xiaowei Yu, Zhe Huang, and Zao Zhang. Feature fusion transferability aware transformer for unsupervised domain adaptation, 2024. URL https://arxiv.org/abs/2411.07794.

A Appendix

A.1 Ablation Study on Office-31 Dataset

Table 4: Comparison of Base ViT, IRM ViT, SAE ViT, and SADA ViT on the Office-31 dataset across all combinations.

Combinations	Base ViT (%)	IRM ViT (%)	SAE ViT (%)	SADA ViT (%)
A→W	79.31	83.53	88.96	**93.37**
A→D	79.12	79.5	**93.57**	91.47
W→A	52.89	64.5	74.94	**77.85**
W→D	96.6	97.99	98.87	**98.99**
D→A	54.27	56.76	74.94	**75.75**
D→W	98.39	99.6	**100**	**100**
Average	76.76	80.31	88.55	**89.57**

To understand the contribution of each component in our proposed model, we conducted an ablation study on the **Office-31 dataset** using four configurations: **Base ViT**, **IRM ViT** (using only IRM loss), **SAE ViT** (using only sparsity enforcement), and **SADA-ViT** (a combination of IRM loss and sparsity enforcement). This analysis isolates the effects of domain alignment (IRM), sparsity enforcement (SAE), and their combined impact on domain adaptation performance.

The **Base ViT** is a baseline model, utilizing the Vision Transformer (ViT) architecture without additional constraints tailored for domain adaptation. In comparison, tbfIRM ViT, which introduces Invariant Risk Minimization (IRM), aims to learn domain-invariant features by aligning feature distributions across domains. Results reveal moderate improvements in IRM ViT compared to Base ViT, particularly in tasks requiring strong domain alignment. For example, performance on the **A→W** task increased from **79.31%** to **83.53%**, while improvements on **W→A** rose from **52.89%** to **64.5%**. However, IRM ViT shows limited improvements in tasks such as **D→A**, indicating that domain alignment alone is insufficient for addressing sparsity and redundancy in feature representations.

In contrast, **SAE ViT**, which focuses on enforcing sparsity through Sparse Autoencoders (SAEs), demonstrates substantial gains across several adaptation tasks. Notably, performance on **A→D** increased from **79.12%** in Base ViT to **93.57%**, while **W→A** improved from **52.89%** to **74.94%**. These results suggest that sparsity-enforced feature representations effectively minimize redundant domain-specific information, allowing the model to focus on critical invariant features. However, while SAE ViT excels in tasks requiring fine-grained feature alignment, it lacks the explicit domain alignment mechanism provided by IRM, as evidenced by the marginal improvement in tasks such as **W→D**.

Comparing **IRM ViT** and **SAE ViT**, we observe that sparsity enforcement in SAE ViT offers more consistent improvements across challenging adaptation tasks. For instance, on t e **A→D** task, SAE ViT outperforms IRM ViT by a significant margin (**93.57%** vs. **79.5%**). Similarly, on bfW→A, SAE ViT achieves **74.94%**, compared to **64.5%** in IRM ViT. These results emphasize that sparsity is more effective in capturing invariant features than domain alignment alone.

When sparsity enforcement and domain alignment mechanisms are combined in **SADA-ViT**, the model consistently performs better across most adaptation tasks. On **A→W**, S DA-ViT achieves **93.37%**, surpassing SAE ViT's **88.96%**. Similar improvements are observed in tasks like **W→A**, where OUR-ViT records **77.85%**, outperforming SAE ViT's **74.94%**. These results suggest that the combination of sparsity and domain alignment mechanisms in OUR-ViT creates a synergistic effect, leveraging the strengths of both approaches to improve generalization across diverse domain shifts.

The findings from our ablation study indicate that while IRM focuses on aligning feature distributions, SAE effectively reduces feature redundancy and emphasizes invariant representations. SADA-ViT successfully combines these strengths, achieving the highest average accuracy across all tasks (**89.57%**) compared to **Base ViT (76.76%)**, **IRM ViT (80.31%)**, and **SAE ViT (88.54%)**. A summary of these results is provided in Table 4. SADA-ViT demonstrates superior performance across all tasks compared to other configurations, as shown in Figure 3,

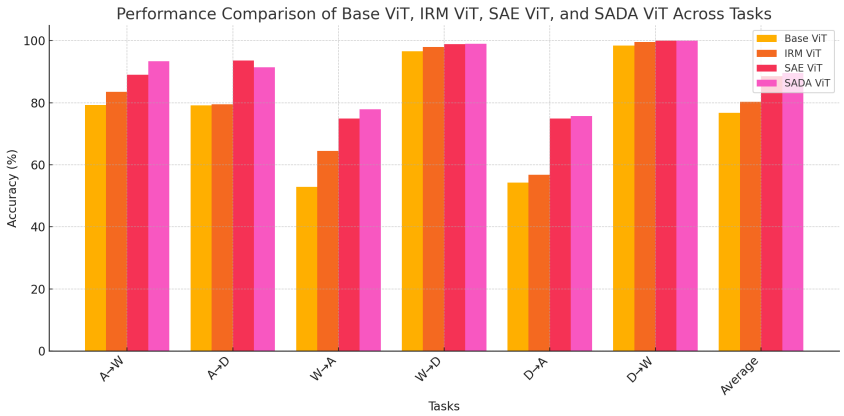

Figure 3: Performance Comparison of Base ViT, IRM ViT, SAE ViT, and OUR-ViT Across Tasks.

In conclusion, this ablation study demonstrates that sparsity and domain alignment are complementary mechanisms for improving domain adaptation performance. The superior performance of **SADA-ViT** highlights the importance of combining these mechanisms to achieve robust and generalized feature extraction across diverse domain shifts.

A.2 t-SNE diagrams for SADA-ViT

Following are visualizations of the alignment between source and target domains for our SADA-ViT model.

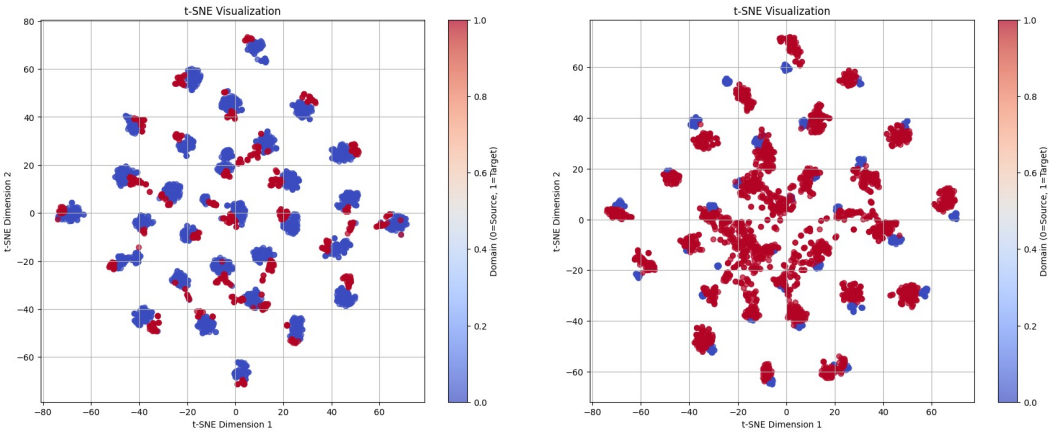

Figure 4: t-SNE for tasks A→W and A→D.

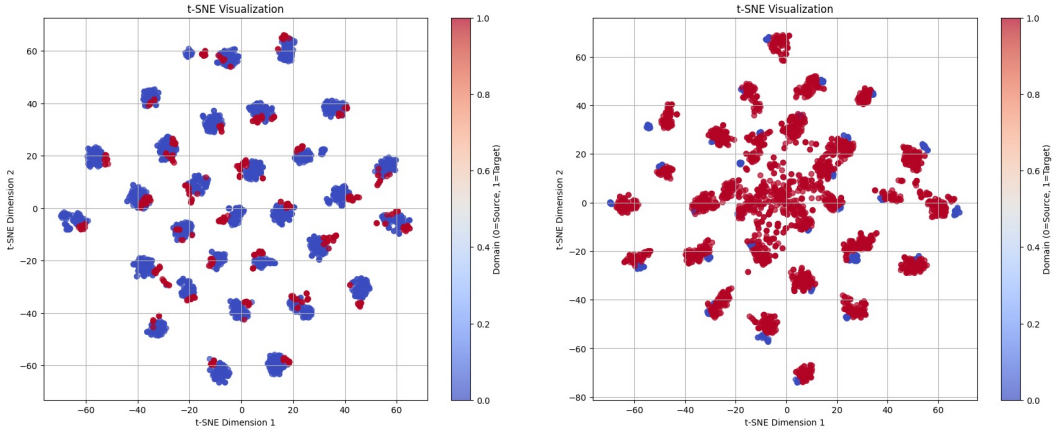

Figure 5: t-SNE for tasks W→A and W→D.

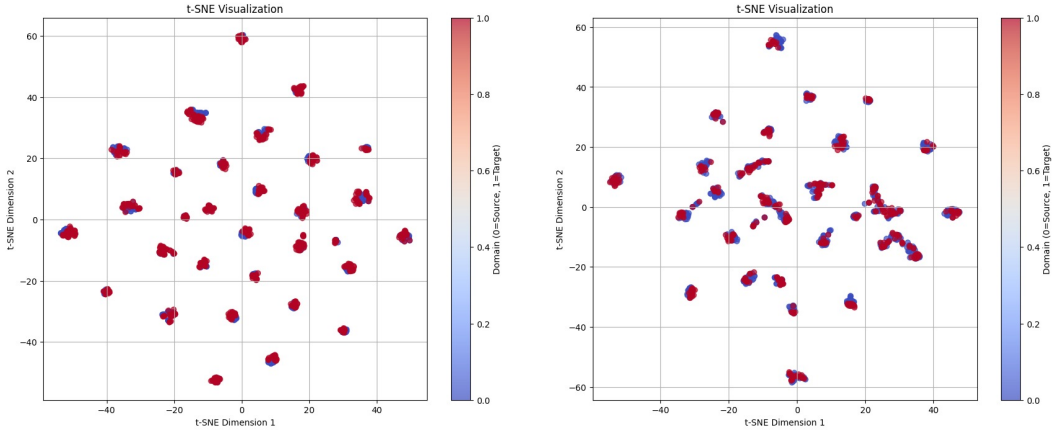

Figure 6: t-SNE for tasks D→A and D→W.