Article

Conditional Domain Adaptation with α-Rényi Entropy Regularization and Noise-Aware Label Weighting

Diego Armando Pérez-Rosero * , Andrés Marino Álvarez-Meza , German Castellanos-Dominguez

Posted Date: 11 July 2025

doi: 10.20944/preprints202507.0933.v1

Keywords: Domain adaptation; image classification; Rényi's entropy; class-conditional alignment; noisy labels

Article

Conditional Domain Adaptation with α-Rényi Entropy Regularization and Noise-Aware Label Weighting

Diego Armando Pérez-Rosero *, Andrés Marino Álvarez-Meza and German Castellanos-Dominguez

Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales 170003, Colombia
* Correspondence: dieaperezros@unal.edu.co

Abstract

Domain adaptation is a key approach to ensure that artificial intelligence models maintain reliable performance when facing distributional shifts between training (source) and testing (target) domains. However, existing methods often struggle to simultaneously preserve domain-invariant representations and discriminative class structures, particularly in the presence of complex covariate shifts and noisy pseudo-labels in the target domain. In this work, we introduce Conditional Rényi α-Entropy Domain Adaptation, named CREDA, a novel deep learning framework for domain adaptation that integrates kernel-based conditional alignment with a differentiable, matrix-based formulation of Rényi's quadratic entropy. The proposed method comprises three main components: (i) a deep feature extractor that learns domain-invariant representations from labeled source and unlabeled target data; (ii) an entropy-weighted approach that down-weights low-confidence pseudo-labels, enhancing stability in uncertain regions; and (iii) a class-conditional alignment loss, formulated as a Rényi-based entropy kernel estimator, that enforces semantic consistency in the latent space. We validate CREDA on standard benchmark datasets for image classification, including Digits, ImageCLEF-DA, and Office-31, showing competitive performance against both classical and deep learning-based approaches. Furthermore, we employ nonlinear dimensionality reduction and class activation maps visualizations to provide interpretability, revealing meaningful alignment in feature space and offering insights into the relevance of individual samples and attributes. Experimental results confirm that CREDA improves cross-domain generalization while promoting accuracy, robustness, and interpretability.

Keywords: Domain adaptation; image classification; Rényi's entropy; class-conditional alignment; noisy labels

1. Introduction

A primary challenge in the development of artificial intelligence systems is ensuring that models maintain reliable performance under conditions that differ from those observed during training [1]. Such discrepancies may arise due to changes in the operational environment, variations in acquisition devices, or differences in user population characteristics [2]. These shifts, though often subtle, can significantly impact model behavior and compromise generalization capabilities, even when the underlying task remains unchanged [3]. This vulnerability becomes particularly critical in real-world applications, where it is infeasible to anticipate all possible future scenarios, thus limiting the scalability and trustworthiness of deployed solutions [4]. In this context, validation within the source domain alone proves insufficient to guarantee consistent performance in heterogeneous settings, prompting the development of strategies to mitigate such discrepancies. Among these, domain adaptation has emerged as a key approach, enabling the reuse of pre-trained models in new environments by aligning distributions across domains, thereby reducing the need for extensive data collection and annotation in the target domain [5]. The latter not only enhances the efficiency of knowledge transfer, but also supports the creation of more robust and sustainable systems in dynamic and uncertain environments.

Despite the progress achieved through domain adaptation, the problem of generalizing to unseen domains remains only partially resolved. Domain shifts can take complex forms that go beyond marginal discrepancies, affecting the internal structure of learned representations and leading to systematic performance degradation in the target domain [6]. Consequently, adapted models frequently exhibit degraded or inconsistent performance when deployed in unfamiliar environments, especially under shifts in input distributions that are structural and semantic in nature [7]. This limitation arises primarily from the inability to preserve domain-invariant features under covariate shifts, where noise in input features, biased samples, or insufficient representations can degrade the alignment across domains and compromise the stability of the learned models [8]. Second, generalization is further hindered when the learned features lack discriminative power, particularly in the presence of concept shift and noisy labels. These factors distort latent representations and decision boundaries, making it difficult to maintain semantic clarity in the target domain [9]. Third, the absence of interpretability mechanisms impedes the reliable evaluation of whether predictions are based on meaningful semantic signals or on spurious correlations inherited from the source domain [10]. Collectively, these challenges hinder the development of domain-adaptive systems that are accurate, robust, and interpretable.

In response to the challenges inherent in domain adaptation, numerous classical approaches have been proposed, most of which rely on linear transformations to align source and target distributions. These strategies aim to mitigate distributional discrepancies through statistical alignment techniques. Methods such as Correlation Alignment (CORAL) and Subspace Alignment (SA) reduce marginal discrepancy by aligning covariance matrices or projecting data onto orthonormal subspaces [11,12]. Despite their effectiveness under controlled conditions, their reliance on original feature spaces or linear projections makes them susceptible to distortions, noise, and domain-specific biases, hindering the extraction of invariant representations [13]. To address these limitations, geometrically inspired extensions such as Geometric Transfer Learning (GTL) have been developed, incorporating structural constraints between domains [14]. Nonetheless, they depend on linear subspace representations, which fail to adequately preserve the support of the target domain in the presence of data heterogeneity or limited representational capacity [15]. In addition, techniques such as Transfer Joint Matching (TJM), Transfer Component Analysis (TCA), and Maximum Independence Domain Adaptation (MIDA) seek to align both marginal and conditional distributions via linear projections [16–18]. Yet, they do not guarantee class separability in the latent space, particularly under concept shift or class imbalance, resulting in ambiguous decision boundaries and diminished discriminative performance [19]. A comparable deficiency is noted in Joint Distribution Adaptation (JDA), which, despite modeling joint alignment, assumes uniform relevance across classes and lacks adaptive mechanisms to address intra-class heterogeneity or instance-level significance [20].

Due to the structural constraints of traditional domain adaptation techniques, particularly the decoupling of feature transformation and prediction phases, deep learning methods have emerged as a more cohesive solution for preserving domain-invariant features across the representation space [21]. These approaches leverage the expressive capabilities of deep neural networks to jointly optimize feature extraction and domain alignment, enhancing adaptability under covariate shift [22]. Adversarial training-based models, including Domain-Adversarial Neural Networks (DANN) and its extensions, have demonstrated considerable effectiveness in aligning marginal distributions within a shared latent space [23,24]. Still, while these methods reduce global disparities, they often struggle to maintain class separability, as they do not explicitly model conditional structures or discriminative boundaries [25]. To overcome these limitations, hybrid models have emerged that integrate deep learning architectures with statistical alignment objectives, enabling end-to-end optimization for improved domain adaptation performance [26]. These approaches aim to preserve both predictive accuracy and domain invariance by combining supervised losses with the minimization of statistical discrepancies across multiple network layers [27,28]. However, hybrid methods also face challenges, such as gradient conflicts between classification and alignment objectives, and semantic misalignment caused by noisy pseudo-labels [29]. In summary, deep learning models represent a significant advancement in preserving domain-invariant features, yet their ability to ensure discriminative consistency in the target domain remains limited [30].

Despite significant advancements in deep learning techniques aimed at extracting domain-invariant features, these methods often fall short in preserving a well-defined, discriminative class structure within the target domain. To mitigate this issue, transfer-based strategies include fine-tuning, teacher–student models, meta-learning frameworks, and asymmetric architectures such as Adversarial Discriminative Domain Adaptation (ADDA) have been proposed [21,25,31–33]. These approaches aim to enhance inter-class separation through adaptive training or auxiliary supervision; however, they exhibit notable limitations, including degradation of pretrained features and susceptibility to noise [34,35]. Additionally, ADDA variants lack explicit modeling of class boundaries, often resulting in ambiguous latent representations [36]. In contrast, conditional alignment techniques—particularly Conditional Adversarial Domain Adaptation (CDAN) and its extensions—integrate classifier outputs into the discriminator to capture class-conditional dependencies [37]. While these techniques improve multimodal alignment, they are vulnerable in scenarios with class imbalance or low prediction confidence, potentially distorting decision boundaries and compromising semantic clarity [30].

In addition to generalization and discriminability, interpretability has become a pivotal aspect of domain adaptation, especially in high-stakes applications where understanding model behavior is essential for fostering trust, transparency, and accountability [38]. In this context, latent space analysis has proven valuable for examining the structure of learned representations. Linear techniques such as Principal Component Analysis (PCA) offer computational efficiency but fall short in capturing the nonlinear relationships relevant across multiple domains [39]. In contrast, nonlinear methods like t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are more effective in representing complex inter-domain structures [40]. UMAP, in particular, stands out for its ability to preserve both local and global structures, maintain stability under parameter variation, and scale efficiently—making it especially useful for visualizing semantic alignment across domains [41,42]. Besides, interpretability is especially crucial in sensitive applications. Among post hoc methods, Gradient-weighted Class Activation Mapping (Grad-CAM) generates attention maps that highlight regions influencing model predictions, while its extension, Grad-CAM++, improves spatial resolution through higher-order derivatives, though it remains limited by nonlinear activation functions [43–45]. In domain adaptation, Grad-CAM++ has proven effective not only as an explainability tool but also for visually assessing semantic consistency across domains [46]. Other approaches, such as Layer-wise Relevance Propagation (LRP) and SHapley Additive exPlanations (SHAP), provide quantitative insights by assigning relevance scores to input features, aiding the identification of spurious patterns or conflicting decision rules [47]. The lack of interpretability methods specifically designed for transfer learning and domain adaptation remains a significant limitation, highlighting the need for more robust explanatory tools tailored to cross-domain scenarios [48].

Here, we propose Conditional Rényi α-Entropy Domain Adaptation (CREDA), a novel domain adaptation framework designed to simultaneously preserve domain-invariant representations, enforce class-conditional alignment, and mitigate the effect of noisy pseudo-labels. The core idea of CREDA is to regularize deep feature alignment using a differentiable, matrix-based formulation of Rényi's quadratic entropy, which provides a non-parametric and robust estimate of class-wise distributional similarity. CREDA is implemented as an end-to-end trainable architecture comprising three key stages:

- Deep Feature Extraction: A shared ResNet-18 backbone encodes samples from both source and target domains into a latent representation space.
- Noise-Aware Label Weighting: An entropy-derived confidence score is used to down-weight low-confidence pseudo-labels in the target domain, improving robustness against noisy or ambiguous predictions.
- Class-Conditional Alignment via Rényi-based entropy: A novel entropy-based regularization term is applied over kernel Gram matrices to minimize divergence between class-wise source and target feature distributions.

We evaluate CREDA on three widely used visual domain adaptation benchmarks for image classification: Digits, ImageCLEF-DA, and Office-31. We also compare its performance against state-of-the-art approaches such as DANN, ADDA, and CDAN+E. The results consistently demonstrate that CREDA achieves superior performance in terms of classification accuracy, semantic alignment, and interpretability, with improvements of average accuracy across benchmarks. Qualitative analyses using UMAP and Grad-CAM++ further confirm that CREDA maintains both inter-class separability and cross-domain semantic coherence, highlighting its potential for deployment in real-world, label-scarce environments.

The remainder of this paper is organized as follows: Section 2 introduces the materials and methods. Sections 3 and 4 discuss the experiments and results. Finally, Section 5 outlines the concluding remarks.

2. Materials and Methods

2.1. Kernel Methods Fundamentals

Kernel methods provide a powerful framework for developing non-linear algorithms. The core idea is to implicitly map the input data from its original space $\mathcal{X}$ into a high-dimensional, or even infinite-dimensional, feature space $\mathcal{H}$ via a non-linear mapping $\Phi : \mathcal{X} \to \mathcal{H}$. The space $\mathcal{H}$ is a special type of Hilbert space known as a Reproducing Kernel Hilbert Space (RKHS), and the mapping Φ is chosen such that complex patterns in the data may become simpler, e.g., linearly separable $\mathcal{H}$ [49].

Explicitly computing the coordinates of the mapped data points $\Phi(x)$ is often computationally expensive or infeasible. Then, the kernel trick allows us to bypass this by defining a kernel function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that computes the inner product between two points in the feature space:

$$\kappa(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}. \tag{1}$$

Then, we work directly with the kernel function without ever needing to know the explicit form of Φ or the structure of $\mathcal{H}$. Indeed, an RKHS is uniquely defined by this property, ensuring that all computations can be performed using the kernel [50]. In practice, a common choice for the kernel function is the Gaussian kernel:

$$\kappa_\sigma(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2} \right), \tag{2}$$

which corresponds to an infinite-dimensional feature space, with $\sigma \in \mathbb{R}^+$. Still, its mathematical tractability and intuitive notion of similarity make it a commonly used approach [51].

2.2. Kernel-based α-Rényi's Entropy Estimation

Let X be a continuous random variable with a probability density function (PDF) $f(x)$, $x \in X$, the Rényi's α-order entropy is defined as [52]:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} f(x)^\alpha dx, \tag{3}$$

where $\alpha > 0$ and $\alpha \neq 1$. A primary challenge in applying this definition is that in most practical scenarios, especially with high-dimensional data like deep features, the underlying PDF $f(x)$ is unknown [53]. To circumvent this, a Parzen-window method, also known as Kernel Density Estimation (KDE) can be employed. Namely, given a finite set of N samples $\{x_i \in X\}_{i=1}^N$, the PDF at any point x can be estimated as the average of kernel functions centered at each sample [54]:

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(x, x_i), \tag{4}$$

where the Gaussian kernel is selected for its mathematical simplicity and desirable smoothing behavior (see equation 2). In particular, when $\alpha = 2$ in equation 3, we focus on the special case of Rényi's entropy, known as quadratic entropy. Indeed, the integral term in equation (3), $\int f(x)^2 dx$, is known as the Information Potential (IP) [55], a measure of the average information contained in the distribution. Substituting the KDE estimator $\hat{f}(x)$ into the IP integral, yields:

$$\hat{V}_2(X) = \int \hat{f}(x)^2 dx = \int \left(\frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(x, x_i) \right) \left(\frac{1}{N} \sum_{j=1}^{N} \kappa_\sigma(x, x_j) \right) dx,$$
$$\hat{V}_2(X) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \int \kappa_\sigma(x, x_i) \kappa_\sigma(x, x_j) dx. \tag{5}$$

A significant advantage of using a Gaussian kernel is that the integral in equation (5) has a closed-form solution based on the convolution property of Gaussians [56]:

$$\int \kappa_\sigma(x, x_i) \kappa_\sigma(x, x_j) dx = \kappa_{\sqrt{2}\sigma}(x_i, x_j). \tag{6}$$

The latter simplifies the IP estimator to a practical, sample-based formula that depends only on pairwise interactions between samples, completely bypassing the need for explicit PDF estimation:

$$\hat{V}_2(X) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \kappa_{\sqrt{2}\sigma}(x_i, x_j). \tag{7}$$

Next, let $\mathbf{K} \in \mathbb{R}^{N \times N}$ be a Gram matrix whose elements are the pairwise kernel evaluations, $\mathbf{K}_{ij} = \kappa_{\sqrt{2}\sigma}(x_i, x_j)$. The sum of all elements in this matrix can be computed as $\mathbf{1}^T \mathbf{K} \mathbf{1}$, where $\mathbf{1}$ is a column vector of ones. This gives a matrix-based estimator for the IP:

$$\hat{V}_2(X) = \frac{1}{N^2} \mathbf{1}^\top \mathbf{K} \mathbf{1} \tag{8}$$

Recently, a α -Rényi matrix-based operator extractss from the IP expression in equation 8. More generally, Rényi's entropy can be defined directly over the eigenspectrum of a normalized Gram matrix. If we define a normalized Gram matrix $\mathbf{A} = \mathbf{K}/\mathrm{tr}(\mathbf{K})$, where $\mathrm{tr}(\cdot)$ is the trace operator, the entropy is given by [57]:

$$H_\alpha(\mathbf{A}) = \frac{1}{1-\alpha} \log(\mathrm{tr}(\mathbf{A}^\alpha)) = \frac{1}{1-\alpha} \log\left(\sum_i \breve{\lambda}_i(\mathbf{A})^\alpha \right), \tag{9}$$

where $\breve{\lambda}_i(\mathbf{A})$ are the eigenvalues of $\mathbf{A}$. For our work with $\alpha = 2$, we use a computationally stable form based on the Frobenius norm: $H_2(\mathbf{A}) = -\log(\mathrm{tr}(\breve{\mathbf{A}}^\top \breve{\mathbf{A}}))$, where $\breve{\mathbf{A}} = \mathbf{A}/\mathrm{tr}(\mathbf{A})$, $\mathrm{tr}(\breve{\mathbf{A}}) = 1$, and $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^\top \mathbf{A})$. This matrix-based formulation is essential for deep learning due to several key properties:

- Non-parametric: It makes no prior assumptions about the underlying data distribution, making it highly suitable for the complex and high-dimensional feature spaces learned by neural networks.
- Differentiable: The entropy loss is a function of the Gram matrix elements, which are themselves differentiable functions of the feature vectors produced by a given network. This allows gradients to be backpropagated through the kernel computations to the network's parameters, enabling end-to-end training.
- Robust: The entropy is calculated based on the collective geometric structure of the data, as captured by all pairwise interactions in the Gram matrix. This makes the measure inherently robust to outliers, which would have a limited impact on the overall sum of kernel values.

The matrix-based entropy framework in equation 9 can be extended to measure relationships between two random variables, X and Y, represented by paired feature vectors $\{\mathbf{f}_{X,i}, \mathbf{f}_{Y,i}\}_{i=1}^N$. This is achieved by defining a joint Gram matrix using the Hadamard (element-wise) product, as follows:

- Joint Entropy - (JE). Let $\mathbf{K}_X \in \mathbb{R}^{N\times N}$ and $\mathbf{K}_Y \in \mathbb{R}^{N\times N}$ be the Gram matrices computed from the feature sets of X and Y, respectively. The joint entropy based on the α-Rényi estimator is defined as [58]:

$$H_\alpha(\mathbf{K}_X, \mathbf{K}_Y) = \frac{1}{1-\alpha}\log\big(\mathrm{tr}(\check{\mathbf{K}}_{X,Y})^\alpha\big), \tag{10}$$

  where $\mathbf{K}_{XY} = \mathbf{K}_X \odot \mathbf{K}_Y$, $\check{\mathbf{K}}_{X,Y} = \mathbf{K}_{X,Y}/\mathrm{tr}(\mathbf{K}_{X,Y})$, and $\odot$ denotes the Hadamard product. Of note, the joint matrix $\mathbf{K}_{XY}$ captures the similarity between pairs of samples in the joint feature space.
- Mutual Information - (MI). It quantifies the statistical dependence between two variables. In the matrix-based framework, it is defined in analogy to its classic information-theoretic definition:

$$I_\alpha(\mathbf{K}_X; \mathbf{K}_Y) = H_\alpha(\mathbf{K}_X) + H_\alpha(\mathbf{K}_Y) - H_\alpha(\mathbf{K}_X, \mathbf{K}_Y), \tag{11}$$

  where each entropy term is computed from its respective (normalized) Gram matrix. Maximizing MI is a common objective in representation learning, as it encourages a representation to retain information about a relevant variable.
- Conditional Entropy - (CE). It measures the remaining uncertainty in a variable X given that Y is known. It is defined as:

$$H_\alpha(\mathbf{K}_X|\mathbf{K}_Y) = H_\alpha(\mathbf{K}_X, \mathbf{K}_Y) - H_\alpha(\mathbf{K}_Y) \tag{12}$$

  Minimizing conditional entropy is equivalent to making X more predictable from Y.

2.3. *Domain Adaptation with α-Rényi Entropy-based Label Weighting and Regularization*

Our proposed method, Conditional α-Rényi's Entropy Regularization (CREDA), is designed for end-to-end training in unsupervised domain adaptation. The framework leverages a deep feature extractor $\mathcal{F} : \mathcal{X} \to \mathbb{R}^d$ that maps an input image $\mathbf{x} \in \mathbb{R}^{\check{H}\times\check{W}\times\check{C}}$, with $\mathcal{X} \subseteq \mathbb{R}^{p'}$, $p' = \check{H} \times \check{W} \times \check{C}$, to a d-dimensional feature vector $\mathbf{f} \in \mathbb{R}^d$, as:

$$\mathbf{f} = \mathcal{F}(\mathbf{x}) = (\check{f}_L \circ \check{f}_{L-1} \circ \cdots \circ \check{f}_1)(\mathbf{x}), \tag{13}$$

where $\check{f}_l(\cdot)$ stands for the l-th feature extractor layer ($l \in L$) and $\circ$ is the function composition operator. Besides, a classifier $\mathcal{G} : \mathbb{R}^d \to [0,1]^C$ that predicts class-probability vector $\mathbf{g} \in [0,1]^C$, is defined as follows:

$$\mathbf{g} = \mathcal{G}(\mathbf{f}) = (\check{g}_{\check{L}} \circ \check{g}_{\check{L}-1} \circ \cdots \circ \check{g}_1)(\mathbf{f}), \tag{14}$$

with $\check{g}_{l'}(\cdot)$ as a given classifier layer ($l' \in \check{L}$), $\sum_{c=1}^{C} g_c = 1$ and $g_c \in \mathbf{g}$.

In practice, we are given a labeled source domain $\mathcal{D}^s = \{\mathbf{x}_i^s \in \mathbb{R}^{p'}, \mathbf{y}_i^s \in \{0,1\}^C\}_{i=1}^{N_s}$, with $\sum_{c=1}^{C} y_{i,c}^s = 1$, $y_{i,c}^s \neq y_{i,c'}^s$, $c, c' \in C$, and $y_{i,c}^s, y_{i,c'}^s \in \mathbf{y}_i^s$. Also, an unlabeled target domain is provided as $\mathcal{D}^t = \{\mathbf{x}_j^t \in \mathbb{R}^{p'}\}_{j=1}^{N_t}$. For each class c, we compute the source, target, and source-target kernel-based matrices $\mathbf{K}_c^s \in \mathbb{R}^{n_c^s\times n_c^s}$, $\mathbf{K}_c^t \in \mathbb{R}^{n_c^t\times n_c^t}$, and $\mathbf{K}_c^{st} \in \mathbb{R}^{n_c^t\times n_c^s}$, as follows:

$$\mathbf{K}_c^s = [\kappa_{\sigma_s}(\mathbf{f}_i^s, \mathbf{f}_{i'}^s)], \quad \forall i, i' \in n_c^s : \quad \mathbf{f}_i^s = \mathcal{F}(\mathbf{x}_i^s), \quad y_{i,c}^s = 1 \tag{15}$$

$$\mathbf{K}_c^t = [\kappa_{\sigma_t}(\mathbf{f}_j^t, \mathbf{f}_{j'}^t)], \quad \forall j, j' \in n_c^t : \quad \mathbf{f}_j^t = \mathcal{F}(\mathbf{x}_j^t), \quad \arg\max_{c'} g_{j,c'}^t = c \tag{16}$$

$$\mathbf{K}_c^{st} = [\kappa_{\sigma_{st}}(\mathbf{f}_i^s, \mathbf{f}_j^t)], \quad \forall i \in n_c^s, j \in n_c^t : \quad y_{i,c}^s = 1, \ \arg\max_{c'} g_{j,c'}^t = c, \tag{17}$$

where $g_{j,c}^t \in \mathbf{g}_j^t$ and $\mathbf{g}_j^t = \mathcal{G}(\mathbf{f}_j^t)$. Moreover, n_c^s is the number of samples in $\mathcal{D}^s$ where $y_{i,c}^s = 1$. Likewise, n_c^t holds the number of target inputs satisfying $\arg\max_{c'} g_{j,c'}^t = c$.

To enhance robustness against noisy pseudo-labels in the target set, we propose incorporating a confidence weighting vector $\mathbf{w}^t \in \mathbb{R}^{N_t}$, derived from a quadratic Rényi entropy estimator (see equation 3) that quantifies the uncertainty in the pseudo-label probability prediction, as follows [59]:

$$w_j^t = 1 - \frac{\hat{H}_2(\mathbf{g}_j^t)}{\log(C)}, \tag{18}$$

where $w_j^t \in \mathbf{w}^t$ and:

$$\hat{H}_2(\mathbf{g}_j^t) = -\log\left(\sum_{c=1}^{C}\left(g_{j,c}^t\right)^2\right). \tag{19}$$

Afterward, a target weighting matrix $\tilde{\mathbf{W}}_t^c \in \mathbb{R}^{n_c^t \times n_c^t}$ can be computed, yielding:

$$\tilde{\mathbf{W}}_c^t = \tilde{\mathbf{w}}_c^t(\tilde{\mathbf{w}}_c^t)^\top, \tag{20}$$

where $\tilde{\mathbf{w}}_c^t = \{w_j^t : \arg\max_{c'} g_{j,c'}^t = c\} \in \mathbb{R}^{n_c^t}$.

The core of our CREDA method lies in a novel regularization term that enforces alignment between the class-conditional distributions of the source and target domains. To achieve this, we employ a kernel-based quadratic Rényi entropy mutual information estimator (see Section 2.2):

$$\tilde{\mathrm{I}}_2(\mathbf{K}_c^s;\tilde{\mathbf{K}}_c^t) = \frac{1}{2}\left(H_2(\mathbf{K}_c^s) + H_2(\tilde{\mathbf{K}}_c^t)\right) - H_2(\mathbf{K}_c^{\mathrm{mix}}), \tag{21}$$

being $\tilde{\mathbf{K}}_c^t = \mathbf{K}_c^t \odot \tilde{\mathbf{W}}_c^t$ and:

$$\mathbf{K}_c^{\mathrm{mix}} = \begin{pmatrix} \mathbf{K}_c^s & \mathbf{K}_c^{st} \\ (\mathbf{K}_c^{st})^\top & \tilde{\mathbf{K}}_c^t \end{pmatrix}, \tag{22}$$

which enables the computation of our MI estimator in equation 21 even when the source and target sample sizes differ, namely, $n_c^t \neq n_c^s$.

Finally, the complete CREDA loss integrates the standard supervised cross-entropy on labeled source data with our proposed mutual information regularizer, based on the quadratic Rényi entropy formulation, as follows:

$$\mathcal{L}_{\mathrm{CREDA}} = \sum_{i=1}^{N_s}\sum_{c=1}^{C} y_{i,c}^s \log(\mathcal{G}(\mathcal{F}(\mathbf{x}_i^s))) - \lambda \sum_{c\in C} \tilde{\mathrm{I}}_2(\mathbf{K}_c^s;\tilde{\mathbf{K}}_c^t), \tag{23}$$

where $\lambda \in \mathbb{R}^+$ is a hyperparameter controlling the strength of the domain alignment.

Figure 1 summarizes the core components and training pipeline of our proposed CRERDA model for conditional domain adaptation.

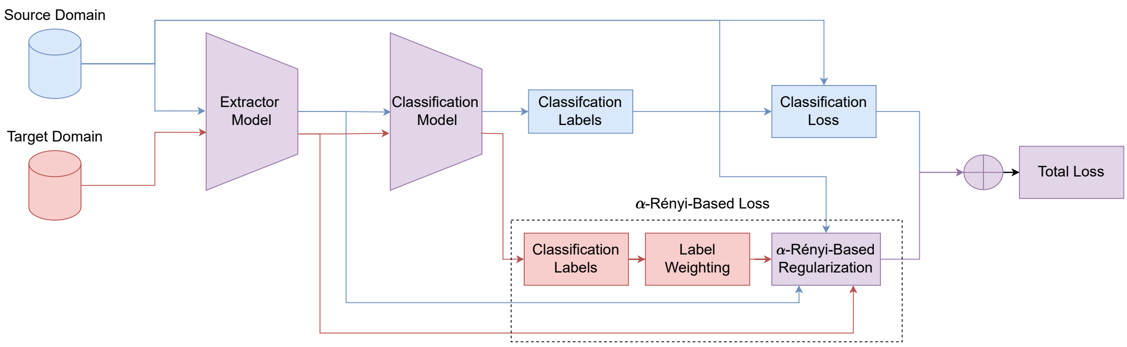

Figure 1. CREDA framework for domain adaptation, incorporating classification loss and α-Rényi Entropy-based label weighting and regularization to attain domain alignment with a class-aware structure. **Blue**: source, **Red**: target, **Purple**: shared.

doi:10.20944/preprints202507.0933.v1

3. Experimental Set-Up

To rigorously evaluate the effectiveness of the proposed CREDA framework for domain adaptation in image classification tasks, we present a comprehensive analysis that includes descriptions of the benchmark datasets, training protocols, comparative baselines, and quantitative and qualitative performance assessments.

3.1. Tested Datasets

To assess the effectiveness and robustness of the proposed domain adaptation method, we conducted extensive experiments on three widely recognized benchmark datasets commonly used in domain adaptation research. Each dataset encompasses visual domains exhibiting substantial distribution shifts, thereby providing a challenging setting for learning domain-invariant representations, as detailed below:

- *Digits:* This benchmark suite is designed for evaluating domain adaptation on digit recognition tasks, spanning both handwritten and natural-scene digits. It comprises three standard datasets: MNIST (M), a large database of handwritten digits; USPS (U), another handwritten digit set characterized by its lower resolution; and SVHN (S), which contains house numbers cropped from real-world street-level images [60]. Notably, the S domain is particularly challenging due to its significant variability in lighting, background clutter, and visual styles compared to M and U (see Figure 2).

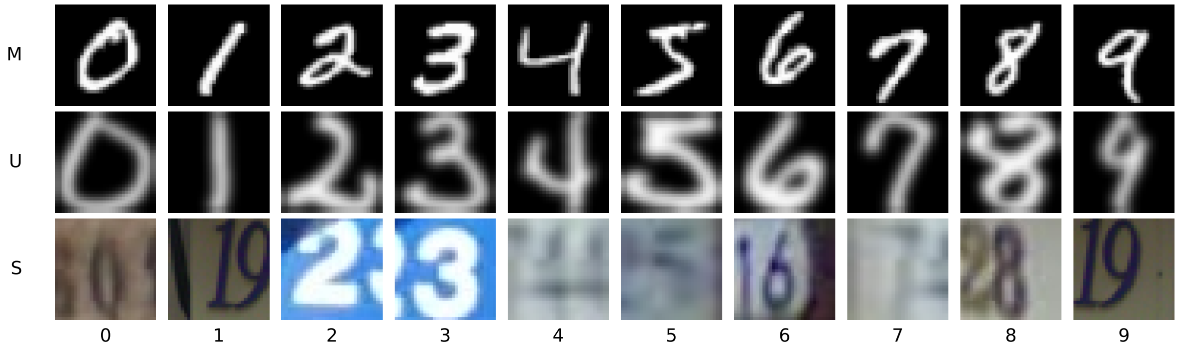

Figure 2. Representative input images for each digit class across source and target domains.

- *ImageCLEF-DA:* This is a standard benchmark for unsupervised domain adaptation, organized as part of the ImageCLEF evaluation campaign. It comprises 12 common object classes shared across three distinct visual domains: Caltech-256 (C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P), see Figure 3. Each domain contains 600 images, with a balanced distribution of 50 images per class [61]. All images are resized to 224 × 224 pixels.

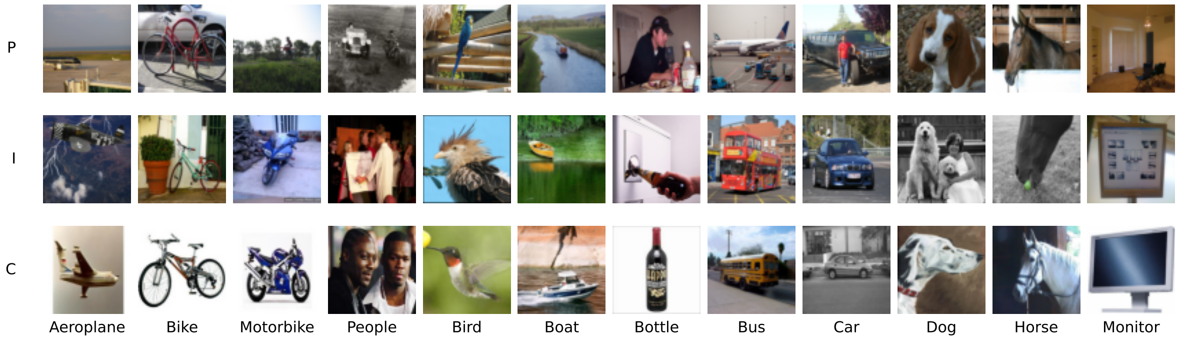

Figure 3. Representative input images for each object class across source and target domains in the ImageCLEF-DA dataset.

- *Office-31:* This is one of the most widely used benchmarks for visual domain adaptation. It consists of 4110 images across 31 object classes, sourced from three domains with distinct visual characteristics: Amazon (A), which features centered objects on a clean, white background under controlled lighting; Webcam (W), containing low-resolution images with typical noise and color artifacts; and DSLR (D), which includes high-resolution images with varying focus and lighting conditions [**?**]. For this study, we selected a subset of ten shared classes, see Figure 4.

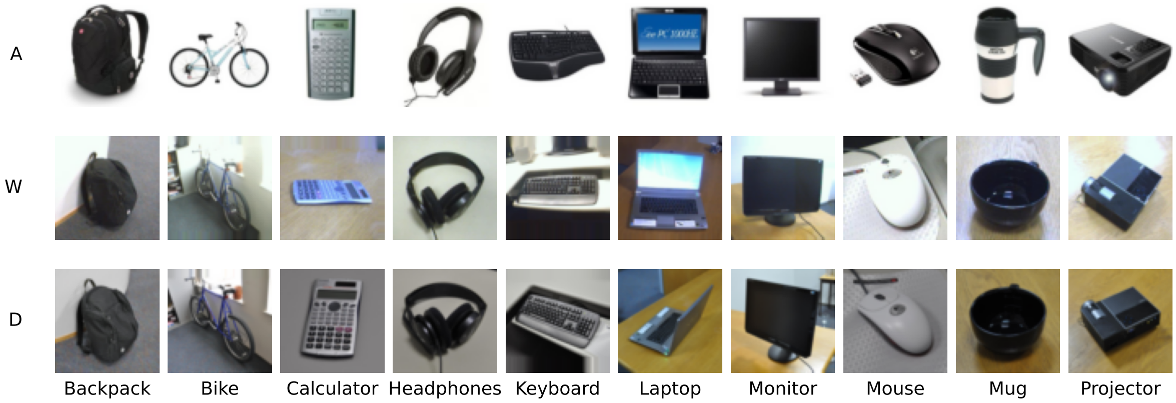

Figure 4. Representative input images for each object class across source and target domains in the Office-31 dataset.

Together, these benchmarks allows evaluating the capacity of domain adaptation methods to generalize across diverse and challenging visual domains.

3.2. Deep Learning Architectures

The architectural components outlined in this section serve as the core modules consistently employed across all experimental configurations. At the heart of the framework lies the feature extractor, which is based on a ResNet-18 convolutional backbone pretrained on ImageNet. To tailor the architecture for domain transfer tasks, the final fully connected layer is removed, while all preceding convolutional and residual blocks are retained. This modification enables the extraction of high-level spatial representations that are robust and transferable across domains [62]. A comprehensive description of the feature extractor's architecture is provided in Table 1.

Table 1. Main feature extractor architecture details.

Layer name	Type	Input shape	Output shape	Param. #
Input	InputLayer	$(3, H, W)$	$(3, H, W)$	0
Conv1	Conv2D + BN + ReLU	$(3, H, W)$	$(64, H/2, W/2)$	9,408
MaxPool	MaxPooling	$(64, H/2, W/2)$	$(64, H/4, W/4)$	0
Layer1	Residual Block ×2	$(64, H/4, W/4)$	$(64, H/4, W/4)$	73,728
Layer2	Residual Block ×2	$(64, H/4, W/4)$	$(128, H/8, W/8)$	230,144
Layer3	Residual Block ×2	$(128, H/8, W/8)$	$(256, H/16, W/16)$	919,040
Layer4	Residual Block ×2	$(256, H/16, W/16)$	$(512, H/32, W/32)$	3,674,112
AvgPool	GlobalAvgPooling	$(512, H/32, W/32)$	$(512, 1, 1)$	0
Flatten	Flatten	$(512, 1, 1)$	(512)	0

The label classifier maps the output of the feature extractor to a vector of class logits corresponding to the C target categories. Its streamlined architecture enables efficient supervised training and facilitates seamless integration into domain adaptation pipelines, as described in Table 2.

doi:10.20944/preprints202507.0933.v1

Table 2. Architecture of the label classifier.

Layer name	Type	Input shape	Output shape	Param. #
Input	InputLayer	(512,)	(512,)	0
FC1	Dense	(512,)	(256,)	131,328
BN1	BatchNorm	(256,)	(256,)	512
ReLU1	Activation	(256,)	(256,)	0
FC2	Dense	(256,)	(128,)	32,896
BN2	BatchNorm	(128,)	(128,)	256
ReLU2	Activation	(128,)	(128,)	0
Output	Dense	(128,)	$(C,)$	$129 \times C$
Softmax	Activation	$(C,)$	$(C,)$	0

To support adversarial training for domain adaptation, the domain discriminator is implemented as a multilayer perceptron. It comprises two fully connected hidden layers with ReLU activations and a final output layer with sigmoid activation. This configuration enables the model to distinguish domain-specific features and promotes the learning of domain-invariant representations. The architecture is summarized in Table 3. The input dimension to this module varies depending on the adversarial method used; for DANN and ADDA, the input is the 512-dimensional feature vector, while for CDAN, it is the $512 \times C$ dimensional joint representation of features and class predictions.

Table 3. Domain discriminator architecture details.

Layer name	Type	Input shape	Output shape	Param. #
Input	InputLayer	(512,)	(512,)	0
FC1	Dense	(512,)	(256,)	131,328
ReLU1	Activation	(512,)	(256,)	0
FC2	Dense	(256,)	(128,)	32,896
ReLU2	Activation	(512,)	(256,)	0
Output	Dense (Sigmoid)	(128,)	(1,)	129

3.3. Assessment and Method Comparison

In all experimental scenarios, we report the classification accuracy and its associated standard deviation in the test set of the target domain. Moreover, during training, model performance is periodically evaluated on validation subsets drawn from both source and target domains to monitor intermediate generalization behavior. In this sense, the Accuracy (ACC) measure is defined as:

$$\mathrm{Acc}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{N} \sum_{i=n}^{N} \mathbb{I}(\hat{y}_n = y_n), \tag{24}$$

where $\hat{y}_n \in \hat{\boldsymbol{y}}$ and $y_n \in \boldsymbol{y}$ denote the predicted and ground truth labels, respectively. $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition is true and 0 otherwise. The standard deviation is estimated from the batch-wise accuracies, serving as a proxy for model stability during inference. The baseline model is trained solely on labeled samples from the source domain and is directly evaluated in the target domain without any adaptation mechanisms. This setting establishes a lower bound for performance under domain shift conditions.

Moreover, the following domain adaptation strategies are considered for comparison:

- DANN: The Domain-Adversarial Neural Network incorporates a domain discriminator linked via a Gradient Reversal Layer (GRL), enabling adversarial training of the feature extractor [63]. This configuration encourages the learning of domain-invariant features while preserving class separability.
- ADDA: The Adversarial Discriminative Domain Adaptation framework separates training into two stages: an initial supervised phase for training the source encoder and classifier, followed by

adversarial fine-tuning of the target encoder (initialized with source weights) to align the source and target distributions [64]. The classifier remains fixed during this second phase.

- CDAN+E: The Conditional Domain Adversarial Network with Entropy Minimization extends adversarial adaptation by feeding the domain discriminator with multilinear features, obtained as the outer product between feature representations and soft classifier predictions [65]. Additionally, an entropy regularization term is applied to the target predictions to encourage confident and well-structured outputs [66].
- CREDA: Our proposed method integrates a supervised classification loss with a conditional divergence term based on Rényi entropy. The objective is to promote intra-class compactness and inter-domain alignment through class-wise kernel-based weighting schemes, yielding representations that are simultaneously discriminative and domain-invariant.

In addition to quantitative measures, we assess the discriminative quality of the learned feature representations using qualitative techniques. Specifically, we employ the well-known Uniform Manifold Approximation and Projection (UMAP) [41], a non-linear dimensionality reduction technique to project high-dimensional features into a two-dimensional latent space, enabling visual inspection of inter-domain and inter-class separability [67]. This technique facilitates an empirical evaluation of how well the feature extractor captures semantically consistent structures across domains. To further complement this analysis, we apply the GradCAM++ method to the classifier module in order to visualize spatial attention regions associated with individual predictions [68]. These attention maps provide insight into the decision-making process of the model and support a comparative interpretation of class activation patterns across source and target domains.

3.4. Training Details

The training procedure follows the standard protocol for unsupervised domain adaptation: all labeled data from the source domain are used along with the entire set of unlabeled data from the target domain. The latter approach aims to learn domain-invariant representations without requiring explicit supervision in the target domain.

All models are trained using the Adam optimizer, with an initial learning rate set to $\eta_0 = 10^{-3}$. To promote stable convergence and mitigate early overfitting of the discriminator, a dynamic scheduling scheme is employed for both the learning rate and the adversarial weighting parameter throughout the training process. Both parameters are updated according to the relative training progress $\check{p} = \frac{\text{epoch}}{\text{total\_epochs}}$, following the expressions:

$$\eta(\check{p}) = \eta_0(1 + \alpha\check{p})^{-\beta}, \quad \lambda(\check{p}) = \frac{1 - e^{-\delta\check{p}}}{1 + e^{-\delta\check{p}}}; \tag{25}$$

where typical hyperparameter values are $\alpha = 10$, $\beta = 0.75$, and $\delta = 10$ [69].

Next, to maintain class balance during model training and evaluation, an initial partition is performed into training (70%), validation (15%), and test (15%) subsets. This process is conducted independently for both the source and target domains. To ensure representative subsets, stratified sampling is applied within each partition, preserving the internal class distributions of each domain. In particular, the independent construction of the validation sets enables consistent and comparable evaluation conditions across domains, which is essential in domain adaptation scenarios where distributional shifts may introduce evaluation bias.

In addition to stratified sampling, the batch size is dynamically adjusted based on the size of the training set (N) in each domain, according to the following empirical rule:

$$\min(\max(16, \lfloor 0.1 \cdot N \rfloor), 64).$$

The lower and upper bounds were established empirically. The lower bound ensures the existence of at least 10 mini-batches per epoch, contributing to optimization stability and preventing prohibitively long training times on small datasets. Conversely, the upper bound avoids excessively large batches that could destabilize learning or exceed GPU memory capacity. This configuration strikes an effective trade-off between gradient stability and computational efficiency, especially when handling domains of different sizes.

It is important to note that, since both dataset partitioning and batch size are determined by the number of available samples in each domain, the number of training instances per epoch is not the same across domains. This asymmetry reflects the inherent scale differences between datasets and allows each domain to contribute proportionally to the learning process without enforcing artificial uniformity.

Regarding the optimization objectives, all models rely on cross-entropy loss for supervised classification on labeled source data. For domain alignment, adversarial approaches such as DANN, ADDA, and CDAN+E use binary cross-entropy loss on a domain discriminator. In CDAN+E, this is enhanced through entropy-aware weighting, where per-sample contributions are modulated by prediction uncertainty—removing the need for an explicit entropy regularization term. All alignment-based methods adopt the same dynamic scheduling function $\lambda(p)$, which progressively increases the weight of the alignment objective throughout training. In the proposed CREDA framework, the adversarial loss is replaced with a divergence-based regularizer that also leverages Rényi entropy for confidence-aware weighting, enabling robust and class-consistent alignment. For all experiments, the kernel bandwidth parameter σ used in the estimation of Rényi's quadratic entropy was fixed to 1. This choice simplifies the training pipeline while maintaining stable alignment behavior across domains.

Besides, to qualitatively assess the discriminative capacity of the learned features, we apply dimensionality reduction using UMAP, leveraging the GPU-accelerated `cuML` implementation. Unless otherwise stated, the default parameters are set as follows: `n_components` = 2, `n_neighbors` = 80, and `random_state` = 42. Prior to projection, features are normalized with `MinMaxScaler`, which facilitates visual inspection of inter-class and inter-domain separability in the latent space. Also, we employ the GradCAM++ technique via the `torchcam` library to visualize class-specific attention regions within the input images. Representative samples for each class are selected from both source and target domains, and the last convolutional layer of the feature extractor is designated as the target layer. The resulting attention masks are normalized and overlaid on the corresponding images, offering a qualitative perspective on the spatial focus of the model during classification.

All experiments are conducted on collaborative computing platforms (Google Colab and Kaggle). The development environment is based on `Python 3.11.11`, using `PyTorch 2.1.2` for model training, `cuML 25.02.01` for GPU-accelerated UMAP visualization, and `torchcam 0.4.0` for `GradCAM++`. All source code and datasets are publicly available at: https://github.com/Daprosero/Domain_Adaptation (accessed on 4 July 2025).

4. Results and Discussion

A fundamental objective in domain adaptation is to learn representations that remain invariant under distributional shifts between domains, commonly referred to as covariate shift. A model's ability to mitigate this challenge is directly reflected in its accuracy on the target domain. To evaluate CREDA's performance quantitatively, we conducted experiments on the three widely adopted benchmark datasets. In the digit adaptation tasks (see Table 4), CREDA achieves the highest average accuracy (67.69%) and performs exceptionally well in challenging tasks such as M→U (98.17%), characterized by significant visual disparities between domains. This evidences its capacity to align distributions with heterogeneous visual features. While ADDA shows competitive performance in specific cases such as M→S and U→M, its overall performance is less stable compared to our proposed method.

Table 4. Accuracy (%) on Digits for unsupervised domain adaptation using ResNet-18.

Method	M → U	M → S	U → M	U → S	S → M	S → U	Avg
Baseline	24.19 ± 15.66	19.59 ± 13.82	76.64 ± 15.49	9.68 ± 10.60	55.82 ± 18.09	72.94 ± 17.30	43.81 ± 15.83
DANN	87.20 ± 12.26	22.39 ± 14.68	82.90 ± 13.07	24.19 ± 15.66	82.11 ± 13.27	76.87 ± 14.64	62.61 ± 13.76
ADDA	94.33 ± 7.11	**40.51 ± 18.16**	**91.66 ± 9.56**	**26.95 ± 15.20**	77.26 ± 14.92	66.91 ± 16.06	66.60 ± 13.83
CDAN	88.57 ± 10.33	19.22 ± 14.07	81.54 ± 12.98	15.48 ± 12.90	**85.33 ± 12.76**	**84.55 ± 12.89**	62.45 ± 12.99
CREDA	**98.17 ± 4.68**	30.48 ± 17.10	89.29 ± 10.79	25.98 ± 15.75	76.48 ± 14.64	83.73 ± 12.53	**67.69 ± 12.92**

Similarly, on the ImageCLEF-DA dataset (see Table 5), CREDA attains the highest average accuracy (67.56%) and outperforms CDAN+E in particularly demanding tasks such as P→C (80.00%) and C→P (54.44%). These results suggest that the Rényi entropy-based regularization employed in our method more effectively balances domain alignment and the preservation of class discriminability.

Table 5. Accuracy (%) on ImageCLEF-DA for unsupervised domain adaptation using ResNet-18.

Method	I → P	I → C	P → I	P → C	C → I	C → P	Avg
Baseline	58.00 ± 21.71	76.83 ± 21.52	68.00 ± 19.30	76.50 ± 19.05	49.83 ± 24.44	38.67 ± 25.60	61.64 ± 20.65
DANN	**61.11 ± 23.13**	62.22 ± 16.39	70.00 ± 16.10	55.56 ± 27.35	62.22 ± 18.04	51.11 ± 23.13	60.70 ± 19.36
ADDA	52.22 ± 21.21	72.22 ± 18.81	**77.78 ± 15.39**	75.56 ± 15.84	58.89 ± 22.82	47.78 ± 22.27	64.07 ± 19.72
CDAN+E	53.33 ± 22.69	**80.00 ± 11.31**	71.11 ± 14.92	76.67 ± 16.96	64.44 ± 22.19	52.22 ± 20.53	66.63 ± 18.43
CREDA	52.22 ± 21.21	76.67 ± 11.92	74.44 ± 13.55	**80.00 ± 19.58**	**65.56 ± 17.54**	**54.44 ± 18.72**	**67.56 ± 17.72**

Lastly, on the Office-31 benchmark (see Table 6), CREDA confirms its superiority with an average accuracy of 81.07% and perfect performance (100.00%) on the D→W and W→D tasks. In contrast, approaches like ADDA show limited robustness, yielding substantially lower performance in tasks such as D→A (20.14%). Our method, in turn, maintains high consistency across all evaluated settings.

Table 6. Accuracy (%) on Office-31 for unsupervised domain adaptation using ResNet-18.

Method	A → W	A → D	D → A	D → W	W → A	W → D	Avg
Baseline	50.51 ± 29.45	55.41 ± 25.37	46.56 ± 34.31	78.98 ± 28.90	54.91 ± 34.50	96.82 ± 7.98	63.03 ± 26.76
DANN	73.33 ± 12.62	66.67 ± 28.87	36.81 ± 18.92	77.78 ± 9.41	51.39 ± 23.83	95.83 ± 7.22	66.30 ± 16.81
ADDA	**84.44 ± 14.61**	**91.67 ± 7.22**	20.14 ± 12.23	71.11 ± 7.36	14.58 ± 10.72	79.17 ± 7.22	60.19 ± 9.56
CDAN+E	80.00 ± 18.96	66.67 ± 26.02	60.42 ± 20.22	93.33 ± 6.85	56.94 ± 16.73	91.67 ± 7.22	74.17 ± 16.50
CREDA	82.22 ± 17.08	79.17 ± 19.09	**70.83 ± 17.15**	**100.00 ± 0.00**	**71.53 ± 11.98**	**100.00 ± 0.00**	**81.07 ± 13.85**

To clarify the reasons for these performance disparities, it is crucial to first examine the inherent complexity of the data domains. Figure 5 presents the 2D UMAP projections of the original feature space, visualized independently for each domain. These plots reveal a fundamental challenge that extends beyond domain shift: the limited class separability within individual domains. This limitation is particularly pronounced in complex datasets such as ImageCLEF-DA and Office-31, where class instances (depicted by distinct colors) exhibit significant entanglement, forming dense and unstructured distributions. Such inherent visual similarity among categories not only complicates classification within the source domain but also serves as a principal source of noisy pseudo-labels in the target domain during unsupervised adaptation. Consequently, a robust domain adaptation strategy must not only align cross-domain distributions but also construct feature representations that enhance inter-class discrimination.

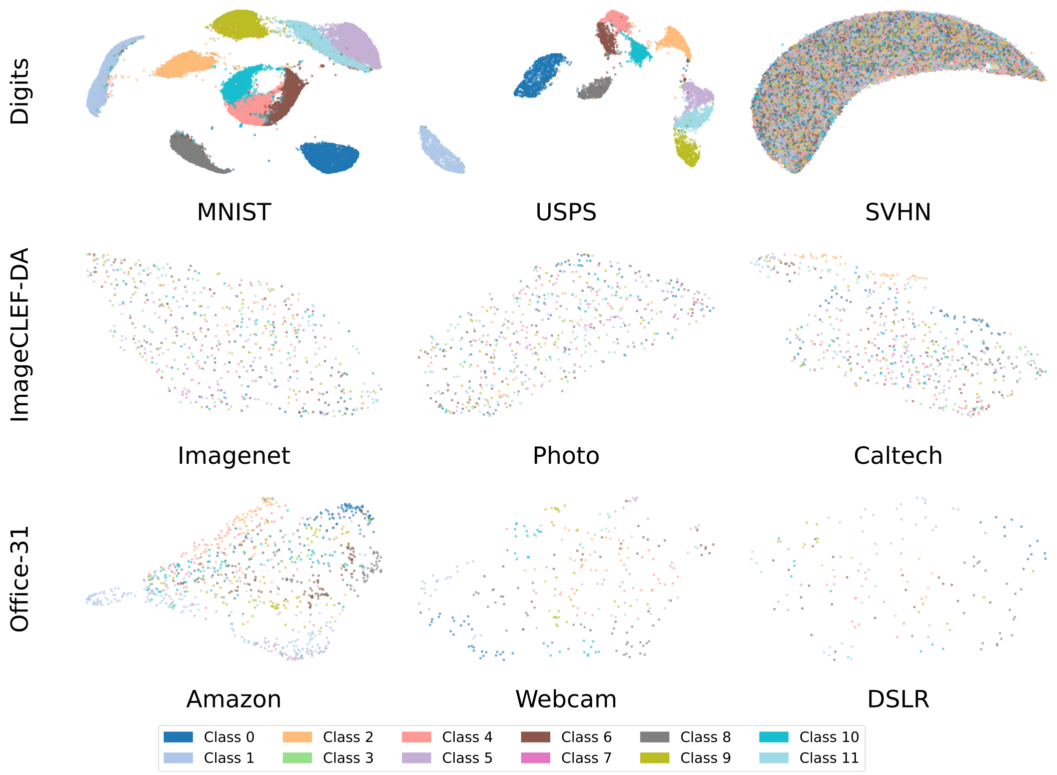

Figure 5. 2D UMAP projections of original feature representations before domain adaptation. **Rows:** evaluated benchmarks. **Columns:** domains within each benchmark.

Building upon this analysis, Figure 6 illustrates how different adaptation techniques address these structural challenges, visualized through UMAP projections of the learned latent spaces. The first column depicts the initial state prior to training, highlighting both the domain gap (e.g., M→U) and the poor semantic organization (e.g., P→C). The baseline model, trained exclusively on source data, fails to bridge this gap, maintaining a clear division between domains. In contrast, adversarial methods like DANN and ADDA achieve partial domain alignment, but often at the expense of class coherence, resulting in fragmented and disordered representations. While CDAN+E introduces a degree of structural consistency, significant inter-class dispersion remains. Ultimately, CREDA yields a markedly superior configuration: it not only facilitates seamless domain integration—evidenced by the homogeneous blending of source and target samples—but also preserves, and in some cases enhances, class-wise separability, as demonstrated by the emergence of compact, well-defined clusters. This outcome provides a visual explanation for CREDA's superior quantitative performance, indicating its ability to balance the removal of spurious domain-specific cues with the preservatio n of underlying semantic structure.

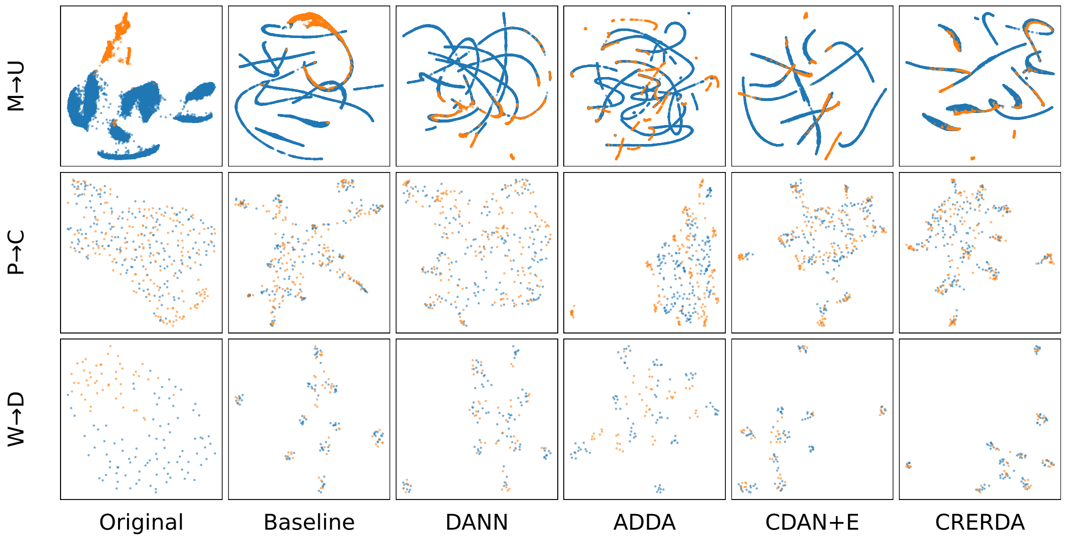

Figure 6. UMAP projections of the learned feature representations across domain adaptation methods, with the source domain shown in blue and the target domain in orange. **Rows:** datasets used in the evaluation. **Columns:** compared adaptation models.

Having established CREDA's capacity to address covariate shift, we next assess whether the learned representations preserve semantic coherence under concept shift, where object appearance changes substantially across domains. In this context, Figure 7 presents UMAP projections with embedded images to qualitatively examine the model's ability to cluster semantically related concepts.

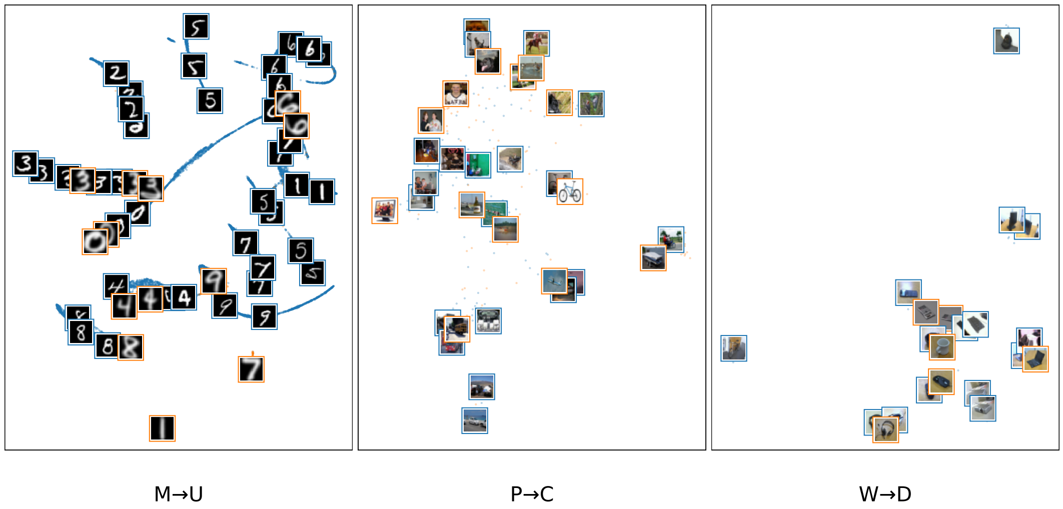

Figure 7. UMAP projections of learned feature representations under the CRERDA model, with input images overlaid, where source domain samples appear in blue and target domain samples in orange. **Left**: Digits. **Middle**: ImageCLEF-DA. **Right**: Office-31.

The results indicate that CREDA learns a semantically rich feature space that transcends superficial variability. For instance, in the M→U task, the model accurately groups digits despite substantial stylistic differences, as seen in the clusters corresponding to digits 6, 0, and 4. In the P→C task, it effectively clusters objects such as bicycles based on defining visual features, despite background and perspective changes. Similarly, in the W→D task, office-related objects are grouped according to their semantic identity, overcoming differences in image quality. Altogether, these visualizations demonstrate that CREDA not only aligns domains but also constructs a feature space in which proximity reflects conceptual similarity—an essential attribute for robust generalization in real-world applications.

Finally, to reinforce the model's reliability, it is essential not only to demonstrate high accuracy and semantic coherence, but also to ensure that its predictions are grounded in interpretable reasoning. In other words, it must be verified that decisions are driven by relevant visual cues rather than spurious correlations.

To address this, we employ Grad-CAM++, with the results shown in Figure 8. The heatmaps reveal strong semantic consistency: regardless of the domain, the model focuses attention on canonical and representative regions of the object, such as the face in a portrait or the main structural components of a vehicle. This confirms that CREDA does not rely on superficial distribution alignment, but rather performs deep and meaningful semantic knowledge transfer. These findings not only enhance trust in the model's predictions but also establish CREDA as a transparent and robust solution for domain adaptation, strengthening the interpretability and reliability of its outputs.

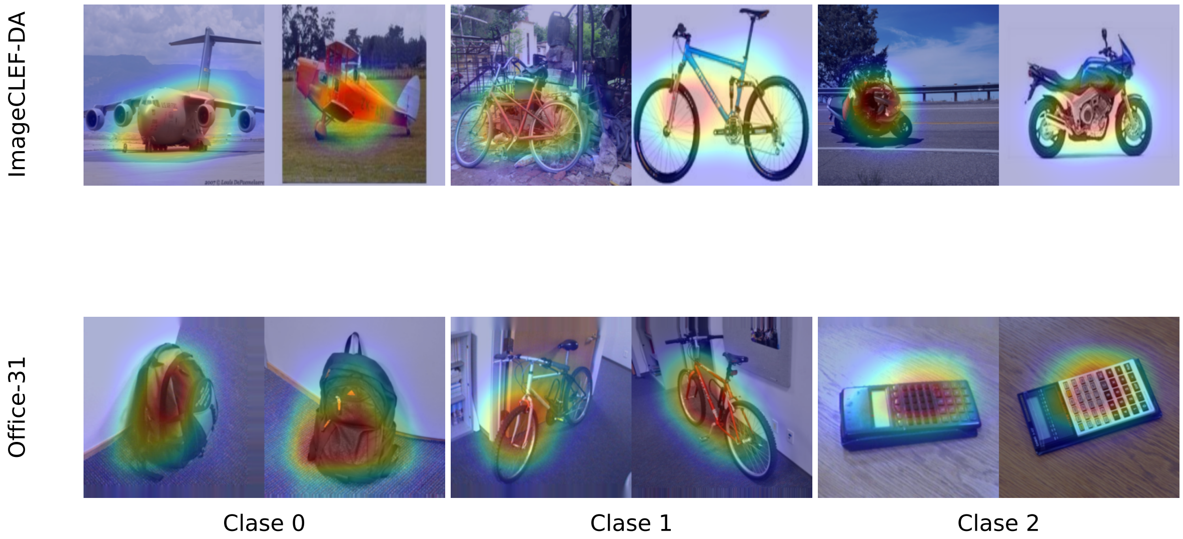

Figure 8. Class-wise visual explanations under the CREDA model. Each pair of images shows the source domain on the left and the corresponding target domain on the right. Heatmaps highlight the most salient regions contributing to the predicted class.

4.1. Limitations

Despite the solid performance of the CREDA framework on unsupervised domain adaptation tasks, several limitations must be acknowledged, which in turn open relevant opportunities for future research. Firstly, its validation has been confined to moderate-scale architectures like ResNet-18; consequently, its scalability and representational capacity on deeper networks, such as ResNet-50 or Transformer-based models, are yet to be explored [70]. Secondly, a single hyperparameter tuning strategy was employed for all tasks, lacking specific optimization per domain pair. Incorporating automated search schemes for adaptation could potentially improve performance and generalization, though this would increase computational cost [71]. Thirdly, the combination of the classification loss and the Rényi divergence-based regularization relies on a static weighting coefficient. Exploring an adaptive normalization method for the loss functions could lead to more stable training by balancing gradient scales. Moreover, given that the regularizer depends on kernel-based estimations, the model's performance is sensitive to the kernel bandwidth, a parameter that was not dynamically optimized in this work. Finally, while CREDA was conceived for the standard unsupervised adaptation setting, its extension to more demanding scenarios, such as few-shot or source-free adaptation, has not yet been investigated [72]. Overcoming these shortcomings would enhance the robustness of the proposed framework and expand its applicability to more complex transfer learning problems.

5. Conclusions

This work introduced a novel domain adaptation framework, termed Conditional Rényi α-Entropy Domain Adaptation (CREDA), a novel deep learning-based strategy integrating kernel-based conditional alignment from a matrix-based formulation of Rényi's quadratic entropy. CREDA is structured around three key components. First, a deep feature extractor is used to learn domain-invariant representations by leveraging labeled source data and unlabeled target data. Second, an entropy-weighted strategy attenuates the influence of low-confidence pseudo-labels, thereby enhancing robustness in ambiguous regions. Third, a class-conditional alignment loss, expressed as a Rényi divergence, is introduced to promote semantic consistency across domains within the latent representation space. In contrast to supervised or semi-supervised approaches, the proposed method does not require labels in the target domain, making it particularly suitable for scenarios where annotation is costly or unavailable. Besides, our class-wise alignment is formulated in a non-parametric and differentiable manner by leveraging kernel-based information potentials, enabling the preservation of semantic structure across domains.

Experimental results across diverse visual adaptation scenarios demonstrate that CREDA consistently outperforms conventional methods such as DANN, ADDA, and CDAN+E in terms of predictive accuracy, representational quality, and interpretability. Notably, CREDA maintains class separability even under complex distribution shifts and when the predicted labels in the target domain exhibit low confidence. The integration of UMAP- and GradCAM++-based visualizations offers valuable insights into the learned representations, reinforcing its applicability in real-world settings where traceability and semantic coherence are critical. From an implementation standpoint, CREDA does not require modifications to the classification loss function. Its confidence-aware weighting scheme and class-conditional regularization enhance robustness to pseudo-label noise and class imbalance. Moreover, its modular architecture facilitates seamless integration into existing deep learning pipelines.

As future work, we will pursue two major extensions. First, we aim to extend CREDA to multi-source and continual domain adaptation settings, where domain shifts occur either simultaneously or sequentially. Attention-based class-conditioned alignment across multiple source domains has been shown to mitigate negative transfer and effectively address class imbalance [73]. Second, we plan to incorporate class-conditional kernel alignment and attention-guided feature disentanglement to improve both interpretability and discriminative alignment, particularly in contexts characterized by subtle inter-class distinctions or limited labeled data. Recent advances in attention-aware class-conditioned alignment suggest that these mechanisms yield robust feature representations and highlight relevant discriminative regions in multi-source adaptation [74]. Together, these extensions aim to broaden the applicability of CREDA to dynamic and heterogeneous environments, while preserving its core strengths of semantic coherence, interpretability, and transparency.

Author Contributions: Conceptualization, D.A.P.-R., A.M.Á.-M. and C.G.C.-D.; data curation, D.A.P.-R.; methodology, D.A.P.-R., A.M.Á.-M. and C.G.C.-D.; project administration, A.M.Á.-M.; supervision, A.M.Á.-M. and C.G.C.-D.; resources, D.A.P.-R. and A.M.Á.-M. All authors have read and agreed to the published version of the manuscript.

Funding: Under grants provived by the project: "Prototipo funcional de lengua electrónica para la identificación de sabores en cacao fino de origen colombiano”, funded by Minciencias-82729-ICETEX 2022-0740 and Casa Luker. Also, A.M. Alvarez thanks to the project: "Aprendizaje de máquina cuántico utilizando espines electrónicos", Hermes-62836, funded by Universidad Nacional de Colombia and Universidad Nacional de Caldas.

Data Availability Statement: The publicly available dataset analyzed in this study and our Python codes can be found at https://github.com/Daprosero/Domain_Adaptation (accessed on 4 July 2025).

References

1. Lu, X.; Yao, X.; Jiang, Q.; Shen, Y.; Xu, F.; Zhu, Q. Remaining useful life prediction model of cross-domain rolling bearing via dynamic hybrid domain adaptation and attention contrastive learning. *Computers in Industry* **2025**, *164*, 104172.
2. Wu, H.; Shi, C.; Yue, S.; Zhu, F.; Jin, Z. Domain Adaptation Network Based on Multi-Level Feature Alignment Constraints for Cross Scene Hyperspectral Image Classification. *Knowledge-Based Systems* **2025**, p. 113972.
3. Huang, X.Y.; Chen, S.Y.; Wei, C.S. Enhancing Low-Density EEG-Based Brain-Computer Interfacing With Similarity-Keeping Knowledge Distillation. *IEEE Transactions on Emerging Topics in Computational Intelligence* **2023**, *8*, 1156–1166.
4. Jiang, J.; Zhao, S.; Zhu, J.; Tang, W.; Xu, Z.; Yang, J.; Liu, G.; Xing, T.; Xu, P.; Yao, H. Multi-source domain adaptation for panoramic semantic segmentation. *Information Fusion* **2025**, *117*, 102909.
5. Imtiaz, M.N.; Khan, N. Towards Practical Emotion Recognition: An Unsupervised Source-Free Approach for EEG Domain Adaptation. *arXiv preprint arXiv:2504.03707* **2025**.
6. Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; Yu, P.S. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering* **2022**, *35*, 8052–8072.
7. Galappaththige, C.J.; Baliah, S.; Gunawardhana, M.; Khan, M.H. Towards generalizing to unseen domains with few labels. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23691–23700.
8. Zhu, H.; Bai, J.; Li, N.; Li, X.; Liu, D.; Buckeridge, D.L.; Li, Y. FedWeight: mitigating covariate shift of federated learning on electronic health records data through patients re-weighting. *npj Digital Medicine* **2025**, *8*, 1–19.
9. Li, L.; Zhang, X.; Liang, J.; Chen, T. Addressing Domain Shift via Imbalance-Aware Domain Adaptation in Embryo Development Assessment. *arXiv preprint arXiv:2501.04958* **2025**.
10. Yuksel, G.; Kamps, J. Interpretability Analysis of Domain Adapted Dense Retrievers. *arXiv preprint arXiv:2501.14459* **2025**.
11. Adachi, K.; Yamaguchi, S.; Kumagai, A.; Hamagami, T. Test-time Adaptation for Regression by Subspace Alignment. *arXiv preprint arXiv:2410.03263* **2024**.
12. Zhang, G.; Zhou, T.; Cai, Y. CORAL-based Domain Adaptation Algorithm for Improving the Applicability of Machine Learning Models in Detecting Motor Bearing Failures. *Journal of Computational Methods in Engineering Applications* **2023**, pp. 1–17.
13. Wang, J.; Feng, W.; Chen, Y.; Yu, H.; Huang, M.; Yu, P.S. Visual domain adaptation with manifold embedded distribution alignment. In Proceedings of the Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 402–410.
14. Yun, K.; Satou, H. GAMA++: Disentangled Geometric Alignment with Adaptive Contrastive Perturbation for Reliable Domain Transfer. *arXiv preprint arXiv:2505.15241* **2025**.
15. Sanodiya, R.K.; Yao, L. A subspace based transfer joint matching with Laplacian regularization for visual domain adaptation. *Sensors* **2020**, *20*, 4367.
16. Wei, F.; Xu, X.; Jia, T.; Zhang, D.; Wu, X. A multi-source transfer joint matching method for inter-subject motor imagery decoding. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2023**, *31*, 1258–1267.
17. Battu, R.S.; Agathos, K.; Monsalve, J.M.L.; Worden, K.; Papatheou, E. Combining transfer learning and numerical modelling to deal with the lack of training data in data-based SHM. *Journal of Sound and Vibration* **2025**, *595*, 118710.
18. Yano, M.O.; Figueiredo, E.; da Silva, S.; Cury, A. Foundations and applicability of transfer learning for structural health monitoring of bridges. *Mechanical Systems and Signal Processing* **2023**, *204*, 110766.
19. Liang, S.; Li, L.; Zu, W.; Feng, W.; Hang, W. Adaptive deep feature representation learning for cross-subject EEG decoding. *BMC bioinformatics* **2024**, *25*, 393.
20. Chen, G.; Xiang, D.; Liu, T.; Xu, F.; Fang, K. Deep discriminative domain adaptation network considering sampling frequency for cross-domain mechanical fault diagnosis. *Expert Systems with Applications* **2025**, *280*, 127296.
21. Wei, G.; Lan, C.; Zeng, W.; Chen, Z. Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 16643–16653.
22. Zhang, Y.; Wang, X.; Liang, J.; Zhang, Z.; Wang, L.; Jin, R.; Tan, T. Free lunch for domain adversarial training: Environment label smoothing. *arXiv preprint arXiv:2302.00194* **2023**.

23. Lu, M.; Huang, Z.; Zhao, Y.; Tian, Z.; Liu, Y.; Li, D. DaMSTF: Domain adversarial learning enhanced meta self-training for domain adaptation. *arXiv preprint arXiv:2308.02753* **2023**.
24. Wu, Y.; Spathis, D.; Jia, H.; Perez-Pozuelo, I.; Gonzales, T.I.; Brage, S.; Wareham, N.; Mascolo, C. Udama: Unsupervised domain adaptation through multi-discriminator adversarial training with noisy labels improves cardio-fitness prediction. In Proceedings of the Machine Learning for Healthcare Conference. PMLR, 2023, pp. 863–883.
25. Mehra, A.; Kailkhura, B.; Chen, P.Y.; Hamm, J. Understanding the limits of unsupervised domain adaptation via data poisoning. *Advances in Neural Information Processing Systems* **2021**, *34*, 17347–17359.
26. Zhu, Y.; Zhuang, F.; Wang, J.; Chen, J.; Shi, Z.; Wu, W.; He, Q. Multi-representation adaptation network for cross-domain image classification. *Neural Networks* **2019**, *119*, 214–221.
27. Madadi, Y.; Seydi, V.; Sun, J.; Chaum, E.; Yousefi, S. Stacking Ensemble Learning in Deep Domain Adaptation for Ophthalmic Image Classification. In Proceedings of the Ophthalmic Medical Image Analysis: 8th International Workshop, OMIA 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 8. Springer, 2021, pp. 168–178.
28. Zhu, Y.; Zhuang, F.; Wang, J.; Ke, G.; Chen, J.; Bian, J.; Xiong, H.; He, Q. Deep subdomain adaptation network for image classification. *IEEE transactions on neural networks and learning systems* **2020**, *32*, 1713–1722.
29. Li, X.; Chen, H.; Li, S.; Wei, D.; Zou, X.; Si, L.; Shao, H. Multi-kernel weighted joint domain adaptation network for cross-condition fault diagnosis of rolling bearings. *Reliability Engineering & System Safety* **2025**, *261*, 111109.
30. Chen, L.; Chen, H.; Wei, Z.; Jin, X.; Tan, X.; Jin, Y.; Chen, E. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 7181–7190.
31. Lu, W.; Luu, R.K.; Buehler, M.J. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials* **2025**, *11*, 84.
32. Xiao, R.; Liu, Z.; Wu, B. Teacher-student competition for unsupervised domain adaptation. In Proceedings of the 2020 25th international conference on pattern recognition (ICPR). IEEE, 2021, pp. 8291–8298.
33. Choi, E.; Rodriguez, J.; Young, E. An In-Depth Analysis of Adversarial Discriminative Domain Adaptation for Digit Classification. *arXiv preprint arXiv:2412.19391* **2024**.
34. Kumar, A.; Raghunathan, A.; Jones, R.; Ma, T.; Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054* **2022**.
35. Liu, Y.; Wong, W.; Liu, C.; Luo, X.; Xu, Y.; Wang, J. Mutual Learning for SAM Adaptation: A Dual Collaborative Network Framework for Source-Free Domain Transfer. In Proceedings of the Proceedings of the 42nd International Conference on Machine Learning (ICML), 2025. Poster presentation.
36. Gao, Y.; Baucom, B.; Rose, K.; Gordon, K.; Wang, H.; Stankovic, J.A. E-ADDA: Unsupervised Adversarial Domain Adaptation Enhanced by a New Mahalanobis Distance Loss for Smart Computing. In Proceedings of the 2023 IEEE International Conference on Smart Computing (SMARTCOMP). IEEE, 2023, pp. 172–179.
37. Dan, J.; Jin, T.; Chi, H.; Dong, S.; Xie, H.; Cao, K.; Yang, X. Trust-aware conditional adversarial domain adaptation with feature norm alignment. *Neural Networks* **2023**, *168*, 518–530.
38. Wang, H.; Naidu, R.; Michael, J.; Kundu, S.S. Ss-cam: Smoothed score-cam for sharper visual feature localization. *arXiv preprint arXiv:2006.14255* **2020**.
39. Mirkes, E.M.; Bac, J.; Fouché, A.; Stasenko, S.V.; Zinovyev, A.; Gorban, A.N. Domain adaptation principal component analysis: base linear method for learning with out-of-distribution data. *Entropy* **2022**, *25*, 33.
40. Jeon, H.; Park, J.; Shin, S.; Seo, J. Stop Misusing t-SNE and UMAP for Visual Analytics. *arXiv preprint arXiv:2506.08725* **2025**.
41. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**.
42. Huang, H.; Wang, Y.; Rudin, C.; Browne, E.P. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Communications biology* **2022**, *5*, 719.
43. Wei, G.; Lan, C.; Zeng, W.; Zhang, Z.; Chen, Z. Toalign: Task-oriented alignment for unsupervised domain adaptation. *Advances in Neural Information Processing Systems* **2021**, *34*, 13834–13846.
44. Langbein, S.H.; Koenen, N.; Wright, M.N. Gradient-based Explanations for Deep Learning Survival Models. *arXiv preprint arXiv:2502.04970* **2025**.
45. Santos, R.; Pedrosa, J.; Mendonça, A.M.; Campilho, A. Grad-CAM: The impact of large receptive fields and other caveats. *Computer Vision and Image Understanding* **2025**, p. 104383.

46. Singh, A.K.; Chaudhuri, D.; Singh, M.P.; Chattopadhyay, S. Integrative CAM: Adaptive Layer Fusion for Comprehensive Interpretation of CNNs. *arXiv preprint arXiv:2412.01354* **2024**.
47. Ahmad, J.; Rehman, M.I.U.; ul Islam, M.S.; Rashid, A.; Khalid, M.Z.; Rashid, A. LAYER-WISE RELEVANCE PROPAGATION IN LARGE-SCALE NEURAL NETWORKS FOR MEDICAL DIAGNOSIS.
48. Ding, R.; Liu, J.; Hua, K.; Wang, X.; Zhang, X.; Shao, M.; Chen, Y.; Chen, J. Leveraging data mining, active learning, and domain adaptation for efficient discovery of advanced oxygen evolution electrocatalysts. *Science Advances* **2025**, *11*, eadr9038.
49. Murphy, K.P. *Probabilistic machine learning: an introduction*; MIT press, 2022.
50. Scholkopf, B.; Smola, A.J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*; MIT press, 2018.
51. Wilson, A.; Adams, R. Gaussian process kernels for pattern discovery and extrapolation. In Proceedings of the International conference on machine learning. PMLR, 2013, pp. 1067–1075.
52. Principe, J.C. *Information theoretic learning: Renyi's entropy and kernel perspectives*; Springer Science & Business Media, 2010.
53. Bishop, C.M.; Nasrabadi, N.M. *Pattern recognition and machine learning*; Vol. 4, Springer, 2006.
54. Silverman, B.W. *Density estimation for statistics and data analysis*; Routledge, 2018.
55. Xu, J.W.; Paiva, A.R.; Park, I.; Principe, J.C. A reproducing kernel Hilbert space framework for information-theoretic learning. *IEEE Transactions on Signal Processing* **2008**, *56*, 5891–5902.
56. Bromiley, P. Products and convolutions of Gaussian probability density functions. *Tina-Vision Memo* **2003**, *3*, 1.
57. Giraldo, L.G.S.; Rao, M.; Principe, J.C. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory* **2014**, *61*, 535–548.
58. Giraldo, L.G.S.; Principe, J.C. Information theoretic learning with infinitely divisible kernels. *arXiv preprint arXiv:1301.3551* **2013**.
59. Hatefi, E.; Karshenas, H.; Adibi, P. Probabilistic similarity preservation for distribution discrepancy reduction in domain adaptation. *Engineering Applications of Artificial Intelligence* **2025**, *158*, 111426.
60. Sankaranarayanan, S.; Balaji, Y.; Castillo, C.D.; Chellappa, R. Generate to adapt: Aligning domains using generative adversarial networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8503–8512.
61. Cheng, J.; Liu, L.; Liu, B.; Zhou, K.; Da, Q.; Yang, Y. Foreground object structure transfer for unsupervised domain adaptation. *International Journal of Intelligent Systems* **2022**, *37*, 8968–8987.
62. Odusami, M.; Maskeliūnas, R.; Damaševičius, R.; Krilavičius, T. Analysis of Features of Alzheimer's Disease: Detection of Early Stage from Functional Brain Changes in Magnetic Resonance Images Using a Finetuned ResNet18 Network. *Diagnostics* **2021**, *11*. https://doi.org/10.3390/diagnostics11061071.
63. Jin, Y.; Song, X.; Yang, Y.; Hei, X.; Feng, N.; Yang, X. An improved multi-channel and multi-scale domain adversarial neural network for fault diagnosis of the rolling bearing. *Control Engineering Practice* **2025**, *154*, 106120.
64. Li, B.; Liu, H.; Ma, N.; Zhu, S. Cross working conditions manufacturing process monitoring using deep convolutional adversarial discriminative domain adaptation network. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* **2025**, p. 09544054251324677.
65. Deng, M.; Zhou, D.; Ao, J.; Xu, X.; Li, Z. Bearing fault diagnosis of variable working conditions based on conditional domain adversarial-joint maximum mean discrepancy. *The International Journal of Advanced Manufacturing Technology* **2025**, pp. 1–18.
66. Feng, Y.; Liu, P.; Du, Y.; Jiang, Z. Cross working condition bearing fault diagnosis based on the combination of multimodal network and entropy conditional domain adversarial network. *Journal of Vibration and Control* **2024**, *30*, 5375–5386.
67. Qiao, D.; Ma, X.; Fan, J. Federated t-sne and umap for distributed data visualization. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 20014–20023.
68. Raveenthini, M.; Lavanya, R.; Benitez, R. Grad-CAM based explanations for multiocular disease detection using Xception net. *Image and Vision Computing* **2025**, p. 105419.
69. Long, M.; Cao, Z.; Wang, J.; Jordan, M.I. Conditional adversarial domain adaptation. *Advances in neural information processing systems* **2018**, *31*.
70. Wang, X.; Zhuo, J.; Zhang, M.; Wang, S.; Fang, Y. Rethinking Effectiveness of Unsupervised Domain Adaptation Models: a Smoothness Perspective. In Proceedings of the ACCV, 2022. shows diminishing gains of UDA methods with ResNet-50 vs stronger backbones [**?**].

71. Saito, K.; Kim, D.; Teterwak, P.; Sclaroff, S.; Darrell, T.; Saenko, K. Tune it the Right Way: Unsupervised Validation of Domain Adaptation via Soft Neighborhood Density. *arXiv preprint arXiv:2108.10860* **2021**. highlights the importance of hyperparameter tuning in UDA without target labels [**?**].
72. Chen, Q.; Zhuang, X. Incorporating Pre-training Data Matters in Unsupervised Domain Adaptation. *IEEE Transactions on Medical Imaging* **2023**. addresses pre-training and mentions source-free and few-shot scenarios [**?**].
73. Deng, Z.; Zhou, K.; Yang, Y.; Xiang, T. Domain Attention Consistency for MultiSource Domain Adaptation. In Proceedings of the International Conference on Computer Vision (ICCV), 2021.
74. Belal, A.; et al. Attention-based Class-Conditioned Alignment for Multi-Source Domain Adaptation of Object Detectors. *arXiv preprint arXiv:2403.09918* **2024**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.