Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Role of Backbone Capacity in Domain-Adversarial Neural Networks

Candelaria Elizabeth Sansores Pérez

csansores@ucaribe.edu.mx

Universidad del Caribe

Antonio Neme

National Autonomous University of Mexico

Mirbella Gallareta Negrón

Universidad del Caribe

Research Article

Keywords: Domain Adaptation, Adversarial Neural Networks, Negative Transfer, Feature Flattening

Posted Date: April 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8682145/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The Role of Backbone Capacity in Domain-Adversarial Neural Networks

Candelaria Elizabeth Sansores-Pérez[1*], Antonio Neme[2†] and Mirbella Gallareta Negrón[1†]

[1*]Ciencias Básicas e Ingenierías, Universidad del Caribe, Esq. Tabachines, Cancún, 77528, Qroo, México.
[2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Ciudad Universitaria, Mexico City, 04510, México.

*Corresponding author(s). E-mail(s): csansores@ucaribe.edu.mx;
Contributing authors: antonio.neme@iimas.unam.mx;
mgallareta@ucaribe.edu.mx;
†These authors contributed equally to this work.

Abstract

Domain-Adversarial Neural Networks (DANN) are often treated as a general solution to covariate shift, yet their effectiveness is highly context dependent. Through a systematic empirical study on the MNIST↔USPS benchmarks, we examine how backbone capacity, adversarial hyperparameters, and shift direction jointly shape adaptation outcomes. We show that when representational capacity is limited, adversarial adaptation fails to stabilize regardless of hyperparameter refinement, revealing an architectural bottleneck. In contrast, with high-capacity backbones under mild shifts, explicit adversarial alignment is frequently redundant and can induce *negative transfer* by degrading class-discriminative structure. We analyze this failure mode as *feature flattening*, where statistical alignment is achieved at the expense of semantic separability. By reversing the shift direction to a data-scarce regime, we identify conditions under which DANN becomes beneficial, acting as a *conditional regularizer* that induces *feature compaction* and improves target performance. Rather than advocating for or against adversarial adaptation, our results provide evidence-based guidance on when DANN is unnecessary, harmful, or essential, highlighting backbone capacity as a key determinant of successful domain adaptation.

Keywords: Domain Adaptation, Adversarial Neural Networks, Negative Transfer, Feature Flattening

1 Introduction

Deep neural networks (DNNs) have become the default approach for pattern recognition in many settings, from medical imaging to environmental monitoring [1]. However, its strong empirical performance is usually based on a delicate assumption: the distribution of data observed at the time of training resembles that found in the deployment. In practical scenarios, this assumption is frequently violated as a result of variations in data acquisition conditions, sensor properties, imaging techniques, or shifts in the underlying population. Such distribution mismatch—commonly described as covariate shift—can substantially degrade performance and complicate the reliable deployment of DNN-based systems [2, 3].

Unsupervised Domain Adaptation (UDA) seeks to mitigate this problem by learning representations that transfer to a target domain without requiring target labels [4]. Among UDA methods, Domain-Adversarial Neural Networks (DANN) [5, 6] are widely used: they frame adaptation as a minimax game in which a feature extractor is trained to support the main task while simultaneously confusing a domain discriminator, typically implemented through a gradient reversal mechanism. This simple and general recipe has made DANN a common baseline between benchmarks and variants [7].

However, the practical reliability of DANN is less clear than its popularity suggests. Recent studies have raised concerns about whether adversarial alignment remains consistently beneficial—particularly when coupled with modern high-capacity backbones [8, 9]. Across the literature, reported gains are often discussed in terms of adversarial strength and hyperparameter choices; however, how these design decisions interact with backbone capacity and with the severity or direction of domain shift remains insufficiently characterized. As a result, practitioners face an unresolved question: when does DANN genuinely improve generalization, and when is it redundant or even harmful?

This study was motivated by a concrete deployment scenario in our university's physicochemical laboratory: helping environmental science students by automating phytoplankton genus classification. A convolutional model trained on public datasets—largely based on Scanning Electron Microscopy images—did not generalize to our target setting, where Optical Microscopy images are used. We initially adopted DANN to bridge this modality gap, but the results remained disappointing even after extensive hyperparameter exploration. Because systematic experimentation with the entire phytoplankton dataset was computationally costly, we turned to simpler and well-controlled benchmarks (MNIST [10] and USPS [11]) to examine the underlying cause of this adversarial adaptation failure more directly. The guiding question became: *under what conditions does adversarial domain adaptation actually improve generalization, and when does it fail despite optimization efforts?* In particular, we focus on the role of backbone capacity in shaping DANN's behavior across different shift regimes.

To study these interactions with finer granularity, we decouple adversarial regulation into two components: λ, which weights the domain loss, and α, which scales the gradient reversal signal. This separation allows us to probe how task discrimination

and domain invariance trade off as adversarial pressure varies, and how that trade-off changes with the representational capacity of the feature extractor.

Our experiments reveal a clear dichotomy. With relatively simple architectures (e.g., standard CNNs), we observe an *architectural bottleneck*: even broad hyperparameter sweeps do not consistently yield stable adaptation. In this regime, limited representational capacity constrains the model's ability to satisfy both objectives (class discrimination and domain confusion), so optimization alone cannot reliably rescue performance [12, 13].

In contrast, introducing a high-capacity backbone such as ResNet18 [14] changes the adaptation setting. Under mild shift conditions (MNIST→USPS), the backbone often reduces the domain gap sufficiently on its own, making explicit adversarial alignment unnecessary. More importantly, when DANN is applied to pre-trained models, it can lead to *negative transfer* [15, 16]. We refer to this failure mode as *feature flattening*: adversarial alignment can emphasize marginal distribution matching while eroding the class-discriminative structure inherited from pre-training.

This naturally leads to the question of whether adversarial adaptation is still necessary when deeper backbones already exhibit substantial robustness under mild shifts. By reversing the adaptation direction (USPS→MNIST), we create a data-scarce regime in which the backbone's inductive bias alone is not sufficient. In this setting, DANN becomes beneficial and, in many runs, essential: adversarial training acts as a conditional regularizer that promotes *feature compaction*, sharpening class structure and closing a generalization gap that the backbone does not bridge by itself.

In this article, we provide a systematic re-evaluation of DANN through the lens of backbone capacity, with the following contributions:

1. *Limits of optimization under insufficient capacity:* We show empirically that hyperparameter refinement cannot compensate for inadequate architectural capacity. Simple CNNs fail to achieve stable adaptation across broad ranges of λ and α, indicating a fundamental architectural bottleneck.
2. *Redundancy of adversarial adaptation in mild shift scenarios:* We present evidence that under mild covariate shift, modern high-capacity backbones (e.g., ResNet18) can achieve strong target performance without adversarial alignment, making DANN unnecessary and in some cases detrimental.
3. *Analysis of negative transfer mechanisms:* We analyze *feature flattening* as a mechanism for negative transfer, where aggressive alignment collapses class-discriminative structure in pre-trained representations.
4. *Identification of an optimal regime for adversarial adaptation:* We delineate conditions—high-capacity backbones combined with data scarcity—under which DANN acts as an effective regularizer, improving semantic structure and target performance via *feature compaction*.

Rather than arguing for or against DANN as a general tool, we aim to provide evidence-based guidance about when adversarial alignment is likely to help, when it is redundant, and when it risks damaging useful structure. Overall, our results challenge a "one-size-fits-all" use of domain-adversarial training and highlight backbone capacity as a key determinant of adaptation outcomes.

The remainder of this paper is organized as follows. Section 2 reviews related work on domain adaptation, backbone capacity, and negative transfer. Section 3 details our methodology, including the decoupled DANN hyperparameters, the architectures considered, and the experimental protocols. Section 4 analyzes the experimental results, combining quantitative performance evaluation with qualitative feature-space visualizations via t-SNE [17]. Section 5 discusses the implications of these findings and summarizes the conditions under which adversarial training is redundant versus beneficial. Finally, Section 6 concludes and outlines future work, including extending these insights to phytoplankton classification under severe modality shift.

2 Related work

Domain adaptation (DA) addresses a central obstacle for deploying machine learning models in real world settings: performance can drop sharply when the data distribution at training time (source) differs from the one observed at deployment (target). A common formalization is the covariate-shift setting, in which the input distribution changes while the conditional label distribution is assumed to remain stable [2, 3]. In many real applications, target labels are unavailable; accordingly, UDA seeks representations that transfer across domains using labeled source data and unlabeled target samples [4]. From a theoretical standpoint, classic analyses relate the target risk to the source risk plus a measure of domain discrepancy, offering guidance on when adaptation is expected to be feasible [18].

Discrepancy-based alignment. Initial domain adaptation methods typically focused on explicitly aligning feature distributions through the use of statistical distance measures. Maximum Mean Discrepancy (MMD) and its deep variants align source and target by minimizing distribution discrepancy in a reproducing kernel Hilbert space [19, 20]. CORrelation ALignment (CORAL) follows a related principle by aligning second-order statistics, typically by minimizing the distance between the source and target covariance matrices of deep features [21]. Related discrepancy criteria extend beyond mean matching by also accounting for covariance information; for example, Maximum Mean and Covariance Discrepancy (MMCD) combines mean and covariance matching to capture richer distributional differences between domains [22]. While discrepancy-based methods can be effective, their behavior becomes less predictable as representations grow high-dimensional and class structure becomes harder to preserve under purely marginal alignment.

Adversarial domain adaptation. Adversarial learning presents a different pathway to alignment, motivated by the minimax formulation underlying Generative Adversarial Networks (GANs) [23]. DANN [5] is a representative and widely used approach: via the Gradient Reversal Layer (GRL), it trains a feature extractor that supports source-task discrimination while simultaneously reducing domain predictability. A number of variants refine this idea in different ways. Adversarial Discriminative Domain Adaptation (ADDA) [7] decouples source and target encoders; Multi-Adversarial Domain Adaptation (MADA) [24] introduces class-conditional discriminators to encourage finer-grained alignment; and recent work provides an in-depth analysis of adversarial discriminative adaptation [25]. Conditional alignment

strategies aim to better preserve class structure during adaptation [26]. For a broad synthesis of adversarial approaches and their design choices, we refer to the recent survey [27]. Collectively, these methods have been reported to perform well on standard digit and object recognition benchmarks, contributing to adversarial adaptation becoming a common baseline in UDA.

Optimization fragility and negative transfer. Nevertheless, the minimax formulation can be difficult to optimize consistently. The resulting adversarial game often exhibits sensitivity to hyper-parameters, architectural choices, and training dynamics [12]. Stability improvements developed for GAN training (e.g., Wasserstein distance) [28] have influenced the field, but they do not eliminate the core tension in adversarial DA: pushing domains together can inadvertently damage the class-discriminative structure needed for prediction. This failure mode is closely related to *negative transfer* [15], where adaptation performs worse than a source-only baseline. Prior work has explicitly discussed the transferability–discriminability trade-off in adversarial DA [16], and recent overviews also emphasize negative transfer as a recurring practical concern [27]. Several strategies attempt to mitigate this tension by dynamically reweighting the alignment and classification objectives [13]; nevertheless, identifying when adversarial pressure helps versus harms remains an open problem. In this paper, we analyze this trade-off through the lens of *feature flattening*, highlighting how aggressive alignment can collapse useful geometry in representation space.

Backbone capacity and architectural effects. The architecture and capacity of the feature extractor (backbone) are critical factors in domain adaptation, yet they have not been extensively studied in a systematic manner. Capacity, governed by factors such as depth, width, and architectural design, has a direct impact on representation expressiveness and stability. Results on over-parameterization and generalization suggest that larger models can learn features that are more robust or transferable under certain regimes [29, 30]. In transfer learning, empirical evidence shows that representation quality depends strongly on backbone choice and pre-training, yet the interaction between backbone capacity and adversarial adaptation mechanisms remains underexplored [31, 32]. These findings motivate a careful re-examination of whether adversarial alignment is beneficial once backbone capacity becomes sufficiently strong.

Pre-training and modern feature extractors. Large-scale pre-training and high-capacity architectures (e.g., ResNets and Vision Transformers [33]) are increasingly reported to provide robust abstractions that reduce the marginal benefit of explicit alignment, particularly under mild shifts [8, 9]. This naturally raises two practical questions that connect directly to our study: (i) do the gains commonly associated with DANN on classic settings persist when the backbone is already strong, and (ii) can adversarial alignment become redundant—or even harmful—when it interferes with pre-trained structure? In a related line of work, recent studies show that aggressive fine-tuning may distort representations and worsen out-of-distribution behavior [34], which is conceptually aligned with the representation degradation that we observe under strong adversarial pressure.

Positioning our work. Building on these perspectives, we focus on how DANN's behavior changes across backbone capacity and shift regimes. While prior studies have

documented optimization fragility or the benefits of strong backbones in isolation, we provide a controlled investigation of their interaction. Specifically, we examine whether adversarial adaptation (i) fails under insufficient capacity, (ii) becomes redundant or harmful under mild shift with strong backbones, and (iii) re-emerges as beneficial regularization in data-scarce settings where capacity is available but labeled source evidence is limited. By decoupling adversarial hyperparameters and analyzing feature-space geometry, we characterize the mechanisms underlying these outcomes and delineate boundary conditions for effective adversarial domain adaptation.

3 Methodology

We designed a controlled evaluation protocol to isolate the main factors that constrain adversarial adaptation. Our analysis focuses on the interaction among three core factors: backbone capacity, adversarial hyperparameters, and the magnitude/direction of the domain shift.

3.1 Datasets and preprocessing

To simulate different covariate-shift regimes, we use the MNIST↔USPS benchmark as our primary testbed. The phytoplankton dataset, while practically relevant, is large enough that exhaustive hyperparameter exploration becomes computationally impractical. By contrast, MNIST and USPS share the same label space and similar low-level structure (grayscale digits, 10 classes) while still exhibiting a well-studied covariate shift, which helps isolate the effect of backbone capacity without confounds such as strong semantic mismatch or severe class imbalance. Their moderate scale further enables a systematic sweep of the adversarial control parameters λ and α under consistent training conditions. Moreover, the relative simplicity of the task reduces the likelihood that observed failures are driven by undertraining, making architectural and algorithmic limitations easier to diagnose.

1. **MNIST** [10]**:** 60,000 training and 10,000 test grayscale images at 28×28.
2. **USPS**[11]**:** 7,291 training and 2,007 test grayscale images at 16×16.

All input samples were resized to a resolution of 28×28 pixels and normalized to the range $[0, 1]$. During training, we apply dynamic augmentation (affine transforms, Gaussian blur, and additive noise) to introduce controlled variability.

Data splits. For each random seed, we create stratified splits in two stages. We reserve 20% of the data as a test set. From the remaining 80%, we allocate 10% for validation, yielding 72%/8%/20% train/validation/test splits per seed. Mean±std values reported throughout therefore reflect both training stochasticity and seed-dependent stratified partitioning.

3.2 Model architectures

We evaluated two backbone families to test the capacity hypothesis:

1. **Simple CNN (low capacity):** Two convolutional blocks (32 and 64 filters, 3×3), each followed by Batch Normalization [35] and ReLU [36]. The feature extractor

terminates in Global Average Pooling (GAP) [37] to reduce parameters, followed by a linear classification head.

2. **ResNet18** [14] **(high capacity):** A deep residual backbone evaluated in two modes:
   - **From scratch:** Random initialization to isolate architectural capacity without prior knowledge.
   - **Pre-trained:** Initialized from ImageNet [38] weights and fine-tuned with early stages frozen (stages 1–2), unfreezing from stage 3 onward (trainable layers begin at `stage3_unit1_conv1`) to assess the impact of pre-training.

3.3 Adversarial decoupling framework

Our DANN implementation consists of a feature extractor G_f, a label classifier G_y, and a domain discriminator G_d. Given inputs $\mathbf{x}$ with class labels y (source only) and domain labels $d \in \{\text{source}, \text{target}\}$, the optimization objective is:

$$L = L_y(G_y(G_f(\mathbf{x})), y) + \lambda L_d(G_d(G_f(\mathbf{x})), d) \tag{1}$$

where λ controls the contribution of the domain-discrimination loss.

Unlike the canonical design, we decouple adversarial regulation into two independent control signals. In addition to λ, we introduce a Gradient Reversal Layer (GRL) scaling factor α:

$$R(\mathbf{x}) = \mathbf{x}, \qquad \frac{\partial R}{\partial \mathbf{x}} = -\alpha \mathbf{I} \tag{2}$$

This decoupling allows us to study how the importance of the auxiliary domain task (λ) interacts with the intensity of adversarial pressure (α).

Logistic schedule for adversarial pressure. When using the logistic schedule, we increase the GRL scaling smoothly as a function of normalized training progress $p \in [0, 1]$:

$$\alpha(p) = \alpha_{\max} \left(\frac{2}{1 + \exp(-\gamma p)} - 1 \right) \tag{3}$$

with $\gamma = 3$ and $\alpha_{\max} = 1.0$ in our implementation.

Domain discriminator design. To improve stability, G_d is implemented as a two-layer MLP (128 and 64 units) rather than a single linear layer. This design helps avoid overly simplistic domain separation and encourages more transferable features.

3.4 Experimental strategy

We conduct experiments under three scenarios, designed to disentangle optimization limits, backbone capacity, and data availability:

1. **Scenario 1: optimization limits (low-capacity regime).** Using the Simple CNN, we analyze hyperparameter sensitivity via a two-stage sweep. First, we vary the domain-loss weight $\lambda \in \{0.1, 0.3, 0.5\}$ while keeping GRL scaling fixed ($\alpha = 1.0$) to probe whether increasing the importance of the domain objective can compensate for limited capacity. Second, fixing the best-performing value from stage one ($\lambda =$

0.3), we vary adversarial pressure using $\alpha \in \{0.1, 0.3, 0.5, \text{logistic}\}$. In addition, we report two reference settings to contextualize these sweeps: the canonical DANN configuration $(\lambda, \alpha) = (1.0, \text{logistic})$ and a high-pressure fixed configuration $(\lambda, \alpha) = (1.0, 1.0)$.

2. **Scenario 2: capacity hypothesis (high-capacity, mild shift).** We deploy ResNet18 in MNIST→USPS to probe redundancy and potential negative transfer when the backbone is already robust. We compare baseline and DANN under identical training settings to assess whether adversarial alignment yields gains or induces negative transfer in this mild-shift regime.
3. **Scenario 3: boundary conditions (inverse shift, data-scarce source).** We reverse the direction to USPS→MNIST to create a data-scarce, small-to-large shift and identify regimes where DANN acts as beneficial regularization. This setting stresses generalization under limited labeled source data, while target labels are used only for evaluation.

Optimization uses Adam [39] with batch size 64. To accommodate non-monotonic and unstable adversarial dynamics, we train for up to 200 epochs to increase the likelihood of capturing peak performance rather than terminating after transient drops.

Oracle model selection. In standard UDA, labeled target data are unavailable and cannot be used for model selection. Here, we deliberately adopt an oracle model-selection protocol: we monitor labeled target validation accuracy (`val_acc_target`) with a patience of 20 epochs and retain the best-performing checkpoint. This is analysis-oriented, providing an estimate of best achievable performance under oracle selection rather than reflecting deployment conditions. Target labels are used exclusively for checkpoint selection and reporting; they never contribute to training gradients. Moreover, λ and α are fixed *a priori* and are not tuned on labeled target data (i.e., no target-labeled hyperparameter optimization is performed). We emphasize that this selection protocol is not available in deployment settings.

Reproducibility across random seeds and splits. Each configuration is run with three random seeds. Each seed defines a stratified train/validation/test split; mean±std therefore reflects both training stochasticity and split variability. Representation metrics are computed per seed on the seed-specific test split using the feature extractor output immediately before the classification head (Simple CNN: 64-d embedding after GAP; ResNet18: backbone embedding before the final classifier).

3.5 Evaluation metrics

Beyond standard accuracy, we define metrics to quantify adaptation trade-offs and diagnose failure modes such as negative transfer.

Performance metrics. We report deviations from the non-adapted baseline: **Target gain (Δ_T):**

$$\Delta_T = \text{Acc}_T^{\text{DANN}} - \text{Acc}_T^{\text{Baseline}} \tag{4}$$

Source penalty (Δ_S):

$$\Delta_S = \text{Acc}_S^{\text{Baseline}} - \text{Acc}_S^{\text{DANN}} \tag{5}$$

Net benefit (Ω):

$$\Omega = \Delta_T - \Delta_S \tag{6}$$

Accuracy-based domain gap. To quantify cross-domain consistency in terms of performance (distinct from distributional discrepancy metrics), we define: **Initial domain gap (G_{init}):**

$$G_{\text{init}} = \text{Acc}_S^{\text{Baseline}} - \text{Acc}_T^{\text{Baseline}} \tag{7}$$

Residual domain gap (G_{res}):

$$G_{\text{res}} = \text{Acc}_S^{\text{DANN}} - \text{Acc}_T^{\text{DANN}} \tag{8}$$

A reduction from G_{init} to G_{res} indicates improved cross-domain consistency, although it may arise from target gains or from source degradation; therefore, we interpret these gaps jointly with Δ_S and Ω.

Representation metrics. To quantify domain-level statistical divergence and class structure in the learned feature space, we compute:

Maximum Mean Discrepancy (MMD). We compute empirical (squared) MMD using an RBF kernel with fixed bandwidth $\gamma = 1.0$ [19, 40]. Throughout the paper, we refer to this quantity simply as MMD. Lower MMD indicates closer agreement between source and target feature distributions in the induced feature space.

LDA separability proxy. We measure class separability via a trace-ratio proxy based on between-class and within-class scatter:

$$\text{LDA}_{\text{trace}} = \frac{\text{Tr}(S_b)}{\text{Tr}(S_w) + \epsilon} \tag{9}$$

where S_b and S_w are computed from class means and within-class deviations, and we use $\epsilon = 10^{-10}$ for numerical stability. Higher values indicate more compact and better-separated class structure. We use this metric to support patterns such as *feature flattening*, where reduced MMD (apparent alignment) coincides with degraded separability.

t-SNE visualization protocol. t-SNE [17] is used exclusively for qualitative visualization (Section 4.4) with `init=pca`, `perplexity=30`, and `learning_rate=auto` (random seed fixed for reproducibility). For each configuration, we run three seeds; to avoid cherry-picking, the t-SNE figure is generated from the seed whose *target test accuracy* is closest to the mean across the three seeds, yielding a representative visualization. Quantitative comparisons rely on mean±std MMD and LDA across seeds.

4 Experimental results

4.1 Scenario 1: optimization limits (low-capacity regime)

To establish a reference point for domain adaptation, the performance of the model trained solely on the source domain (MNIST) and evaluated on the target (USPS) was initially assessed. Table 1 (row 1) summarizes this baseline, establishing the starting

point for our optimization. Table 1 is designed to characterize the low-capacity regime by jointly reporting transfer outcomes (Acc_T, Δ_S, Ω) and representation diagnostics (MMD, LDA_T), rather than providing a full accuracy breakdown. This distinction clarifies that alignment metrics are interpreted as evidence about feature-space behavior, while transfer effectiveness is assessed through the Δ_S–Ω trade-off.

In this standard MNIST→USPS direction, the non-adapted Simple CNN achieved a high source accuracy of 0.9712. However, generalization was poor, with target accuracy (Acc_T) dropping to 0.6828. This discrepancy reveals a substantial *initial domain gap* of $G_{\text{init}} = 0.2884$, highlighting the severity of the covariate shift and establishing the potential margin for improvement via adversarial adaptation.

Table 1 Impact of hyperparameters on Simple CNN (MNIST → USPS). All values are reported as mean ± std over three random seeds, using oracle checkpoint selection per seed. Compared to the *source-only baseline* (no adaptation), the dynamic α scheduling (best config) is associated with a lower *source penalty* (Δ_S), suggesting improved stability in source-task retention, and yields the least negative *net benefit* (Ω) among the tested Simple CNN DANN variants.

Configuration	λ	α	Acc_T ($\mu \pm \sigma$)	Δ_S ($\mu \pm \sigma$)	Ω ($\mu \pm \sigma$)	MMD ($\mu \pm \sigma$)	LDA_T ($\mu \pm \sigma$)
Baseline (No Adaptation)	-	-	0.6828 ± 0.0326	0.0000 ± 0.0000	0.0000 ± 0.0000	0.2094 ± 0.0182	0.7469 ± 0.0096
DANN (max, fixed)	1.0	1.0	0.6484 ± 0.0228	0.0660 ± 0.0262	-0.1004 ± 0.0034	0.1308 ± 0.0315	0.9885 ± 0.0645
λ **Exploration**	0.1	1.0	0.6366 ± 0.0292	0.0606 ± 0.0065	-0.1069 ± 0.0243	0.2234 ± 0.0583	0.8044 ± 0.0386
(Fixed $\alpha = 1.0$)	0.3	1.0	0.6384 ± 0.0532	0.0624 ± 0.0189	-0.1069 ± 0.0676	0.2268 ± 0.0648	0.9270 ± 0.0750
	0.5	1.0	0.6321 ± 0.0218	0.0684 ± 0.0247	-0.1191 ± 0.0080	0.2109 ± 0.0576	0.9064 ± 0.0716
α **Exploration**	0.3	0.1	0.6590 ± 0.0032	0.0490 ± 0.0081	-0.0728 ± 0.0099	0.2185 ± 0.0594	0.7487 ± 0.0526
(Fixed $\lambda = 0.3$)	0.3	0.3	0.6403 ± 0.0238	0.0545 ± 0.0119	-0.0970 ± 0.0357	0.2181 ± 0.0614	0.7851 ± 0.0457
(Best Config)	**0.3**	**Logistic**	$\mathbf{0.6742 \pm 0.0677}$	$\mathbf{0.0392 \pm 0.0104}$	$\mathbf{-0.0478 \pm 0.0753}$	$\mathbf{0.1824 \pm 0.0691}$	$\mathbf{0.8560 \pm 0.0802}$
Canonical DANN	1.0	Logistic	0.6172 ± 0.0255	0.0996 ± 0.0404	-0.1652 ± 0.0528	0.2360 ± 0.0680	0.8647 ± 0.0544

4.1.1 Hyperparameter sensitivity analysis

We started by evaluating the initial configuration of DANN with maximum adversarial pressure ($\lambda = 1.0, \alpha = 1.0$) to establish a baseline for hyperparameter tuning. Table 1 (row 2) reveals the failure of this setup. The model suffered from clear negative transfer, with source accuracy falling sharply compared to the baseline (0.9052 vs 0.9712), resulting in a *source penalty* of $\Delta_S = 0.0660$. More concerningly, the target accuracy (0.6484) also failed to surpass the non-adapted baseline (0.6828), leading to a negative *net benefit* of $\Omega = -0.1004$. This severe degradation confirms that applying maximum adversarial pressure to a simple CNN is counterproductive, prioritizing alignment over task discrimination. Although this configuration reduces MMD relative to the baseline (from 0.2094 to 0.1308), the *residual domain gap* remains high $G_{\text{res}} = 0.2568$, indicating that statistical alignment alone is insufficient and can be achieved at the expense of discriminability. This interpretation is consistent with the validation dynamics: `val_acc_domain` remained above chance (0.7047 ± 0.1553), indicating that the discriminator was not consistently confused, while `val_acc_target` exhibited erratic oscillations and abrupt drops, suggesting destructive interference under strong adversarial pressure.

We then tested different values for the domain loss weight λ while fixing the adversarial pressure at $\alpha = 1.0$. The objective was to identify a configuration that minimizes

negative transfer. As summarized in Table 1, varying λ failed to overcome the architectural limitations of the Simple CNN. Target accuracy plateaued in the 0.63–0.64 interval, failing to match the 0.6828 baseline. Lower λ values (0.1 and 0.3) proved less detrimental, keeping the *source penalty* near 0.06 and yielding the least negative *net benefit* among the fixed-α runs (-0.1069). Notably, none of these settings produced a consistent reduction in MMD relative to the baseline; for instance, $\lambda = 0.5$ yields an MMD of 0.2109, essentially unchanged from the baseline (0.2094). More broadly, the lack of target improvement despite occasional changes in MMD reinforces that reducing distributional discrepancy is not sufficient for improving target performance in this low-capacity regime. MMD reduction without target improvement is consistent with alignment that does not preserve class structure. The validation dynamics confirmed persistent instability; `val_acc_domain` remained consistently high with substantial variance, indicating that the domain classifier was rarely fooled. We identified $\lambda = 0.3$ as the most favorable static value in this regime, offering the best trade-off between source retention and target performance among the fixed-α configurations.

Following this, we evaluated the impact of the gradient reversal coefficient α while fixing $\lambda = 0.3$. As detailed in Table 1, dynamic scheduling via a logistic function proved critical for maximizing performance with the Simple CNN. This configuration achieved a target accuracy of 0.6742, significantly outperforming fixed values of α and nearly matching the non-adapted baseline (0.6828). Consequently, it yielded the least negative *net benefit* (-0.0478) and the lowest *source penalty* (0.0392), alongside improved alignment (MMD 0.1824) and stronger class separability (highest target LDA: 0.8560). The dynamic scheduler proved beneficial because it mitigates destructive interference: while constant pressure forces the extractor to prioritize invariance prematurely, the logistic schedule allows discriminative features to form before gradually increasing adversarial pressure. Although validation curves indicate that instability persists, the scheduler enabled higher performance peaks (max 0.7527) and better stability than fixed setups. Ultimately, we adopted $\lambda = 0.3$ and $\alpha =$ `logistic` as the optimal setup for this architecture.

4.1.2 Comparative analysis: canonical DANN and optimization limits

In the final stage of this phase, we evaluated the impact of advanced learning rate scheduling and compared our optimized configuration against the canonical DANN formulation. First, we explored a cosine decay scheduler applied to the optimal hyperparameters ($\lambda = 0.3, \alpha = \textit{logistic}$). Surprisingly, this approach actually worsened performance compared to the fixed learning rate. Target accuracy dropped to 0.6371 (vs 0.6742), and the *net benefit* worsened to -0.1046. Validation metrics indicated that the decay schedule failed to mitigate training volatility, likely interfering with the delicate balance required for adversarial convergence. Due to this performance degradation, this configuration is excluded from Table 1 to focus on adversarial hyperparameters. Notably, cosine decay did not improve alignment or separability metrics: compared to the fixed-learning-rate best configuration, it increased MMD (0.2140 vs. 0.1824) and slightly reduced target LDA (0.8272 vs. 0.8560), consistent with the observed drop in target accuracy.

Finally, we tested the standard DANN setup exactly as proposed by Ganin et al. (2016) with $\lambda = 1.0$ and dynamic α (*logistic*). As presented in the final row of Table 1, this setup yielded the poorest performance of the study. The high adversarial pressure resulted in a target accuracy of only 0.6172 and a severe *source penalty* of 0.0996, leading to the most negative *net benefit* recorded (-0.1652). Validation curves confirmed extreme volatility, with `val_acc_domain` fluctuating widely, suggesting that $\lambda = 1.0$ is too aggressive for the simple CNN architecture. These results strongly suggest that granular tuning ($\lambda = 0.3$) is superior to the canonical formulation for capacity-constrained backbones.

In summary, despite extensive tuning identifying $\lambda = 0.3$ and dynamic scheduling as the optimal regime, the simple CNN architecture failed to consistently outperform the non-adapted baseline. The persistent instability observed suggests fundamental limitations in this architecture's capacity to support the adversarial optimization required by DANN. The model failed to reach a balance between domain confusion and task discrimination. This confirms that tuning cannot fix a lack of architectural capacity. This motivates the subsequent investigation using a high-capacity backbone (ResNet18).

4.2 Scenario 2: the capacity hypothesis (high-capacity regime).

In this phase, we investigated whether a high-capacity backbone renders adversarial adaptation redundant or detrimental in the mild shift scenario.

We began with the ResNet18 model trained from scratch. As detailed in Table 2 (ResNet18 *Scratch*, Baseline), the baseline performance was remarkably strong. Target accuracy on USPS hit 0.9586, leaving a minimal domain gap of only 0.0363. Even without explicit adaptation, we found an exceptionally low MMD (0.0238 ± 0.0007) alongside a high target LDA (2.3916 ± 0.2731). While conventionally a low MMD suggests distributional alignment, our subsequent visual analysis (Section 4.4) will show that this metric can be misleading in this context. The high accuracy and discriminability were not achieved by unifying the domains, but rather by the backbone's capacity to **segregate** them into distinct, effective sub-manifolds within the same feature space. This distinction is crucial, the baseline solved the task through separation, not invariant alignment.

Table 2 The capacity hypothesis (MNIST → USPS). Using ResNet18 trained from scratch and ImageNet initialization under a mild domain shift. All values are reported as mean ± std over three random seeds, using oracle checkpoint selection per seed. Positive Ω indicates a net gain over the source-only baseline; negative values indicate negative transfer.

Configuration	$\mathbf{Acc}_S$ ($\mu \pm \sigma$)	$\mathbf{Acc}_T$ ($\mu \pm \sigma$)	Δ_S ($\mu \pm \sigma$)	Δ_T ($\mu \pm \sigma$)	Ω ($\mu \pm \sigma$)	MMD ($\mu \pm \sigma$)	$\mathbf{LDA}_T$ ($\mu \pm \sigma$)
ResNet18 *(Scratch)*							
Baseline (No Adaptation)	0.9949 ± 0.0005	0.9586 ± 0.0074	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0238 ± 0.0007	2.3916 ± 0.2731
DANN	0.9900 ± 0.0003	0.9710 ± 0.0057	0.0050 ± 0.0003	0.0124 ± 0.0057	0.0074 ± 0.0058	0.0191 ± 0.0006	3.3155 ± 0.2333
ResNet18 *(ImageNet)*							
Baseline (No Adaptation)	0.9938 ± 0.0003	0.8993 ± 0.0148	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0009 ± 0.0001	2.7379 ± 0.2317
DANN	0.9588 ± 0.0500	0.8385 ± 0.1214	0.0350 ± 0.0500	-0.0608 ± 0.1214	-0.0957 ± 0.1714	0.0026 ± 0.0027	1.8923 ± 0.5326

Subsequently, we integrated DANN with this backbone using the same hyperparameters identified in Scenario 1 ($\lambda = 0.3$, $\alpha =$ *logistic*) for consistency across

backbones. This setup remained effective (Table 2, ResNet18 *Scratch*, DANN). Target accuracy increased to 0.9710 ± 0.0057 relative to the baseline (0.9586 ± 0.0074), with a minimal *source penalty* ($\Delta_S = 0.0050 \pm 0.0003$), yielding a positive *net benefit* ($\Omega = +0.0074 \pm 0.0058$). The MMD also decreased (0.0191 ± 0.0006 vs. 0.0238 ± 0.0007), although, as discussed in Section 4.4, low MMD should not be interpreted as evidence of unified manifolds. Importantly, in this scratch setting the adversarial objective did not compromise class separability, as reflected by the increased target LDA (3.3155 ± 0.2333). Overall, this suggests that with sufficient capacity and training from scratch, adversarial adaptation can act as a mild regularizer rather than inducing feature degradation.

We then assessed the impact of prior knowledge by fine-tuning an ImageNet-pretrained ResNet18, with early stages frozen (stages 1–2), unfreezing from stage 3 onward. As reported in Table 2 (ResNet18 *ImageNet*, Baseline), this configuration achieved a target accuracy of 0.8993 ± 0.0148. However, performance was lower than the model trained from scratch (0.9586 ± 0.0074), suggesting that ImageNet features, while robust, may introduce an inductive bias that is less optimal for digit recognition than features learned *from scratch*. A notable observation is the extremely low MMD (0.0009 ± 0.0001), the lowest in this study, indicating very strong distributional similarity under this metric even without adversarial adaptation. Importantly, target discriminability remained relatively high ($\text{LDA}_T = 2.7379 \pm 0.2317$), implying that class structure was initially preserved. Overall, this supports the view that under mild shift and sufficient data, explicit adaptation may offer limited additional benefit; however, as shown next, imposing further adversarial pressure on top of an already tightly constrained representation can be associated with a degradation of class separability.

Integrating DANN into this pre-trained setup, however, produced a clear pattern of negative transfer (Table 2, ResNet18 *ImageNet*, DANN). Target accuracy dropped to 0.8385 ± 0.1214, below the non-adapted baseline (0.8993 ± 0.0148), yielding a negative *net benefit* of $\Omega = -0.0957 \pm 0.1714$. Validation traces were also consistent with unstable optimization, with erratic `val_acc_target` and lack of stable convergence. Notably, this degradation occurred despite very low MMD (0.0026 ± 0.0027), illustrating that strong statistical overlap can coexist with poorer task structure. In parallel, target LDA decreased substantially relative to the baseline (1.8923 ± 0.5326 vs. 2.7379 ± 0.2317), consistent with reduced class separability. Together, these observations support the interpretation that adversarial pressure can erode discriminative structure when applied on top of features that are already strongly constrained by pre-training, aligning with the *feature flattening* behavior discussed in this work.

4.3 Scenario 3: boundary conditions (inverse shift)

In the final phase of our experiment, we reversed the domain direction to create a data-scarce scenario (USPS as source), aiming to identify the boundary conditions where adaptation becomes essential.

Inverting the task to USPS $\rightarrow$ MNIST exposed a much harsher initial domain gap. Although Simple CNN performed well on the source (USPS accuracy: 0.9496), it collapsed on the target (MNIST: 0.4360), producing a massive initial gap of $G_{\text{init}} = 0.5136$. Data scarcity likely amplifies this disparity, as the limited USPS

dataset prevents the model from learning features that generalize well to the larger, more complex MNIST domain. Table 3 summarizes the Simple CNN baseline and the two DANN variants discussed below. For this low-capacity setting under scarcity, we focus on performance trade-offs (Acc_S, Acc_T, Δ_S, Δ_T, and Ω), as the model does not produce sufficiently stable representations for reliable feature-space diagnostics.

Table 3 Simple CNN under inverse shift (USPS → MNIST). Performance trade-offs in a data-scarce regime. All values are reported as mean ± std over three random seeds.

Configuration	$\textbf{Acc}_S$ $(\mu \pm \sigma)$	$\textbf{Acc}_T$ $(\mu \pm \sigma)$	Δ_S $(\mu \pm \sigma)$	Δ_T $(\mu \pm \sigma)$	Ω $(\mu \pm \sigma)$
Baseline (No Adaptation)	0.9496 ± 0.0081	0.4360 ± 0.0629	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
DANN (Optimized)	0.8473 ± 0.0258	0.4744 ± 0.0292	0.1023 ± 0.0258	0.0384 ± 0.0292	-0.0639 ± 0.0528
DANN (Canonical)	0.8597 ± 0.0514	0.5073 ± 0.0266	0.0900 ± 0.0514	0.0714 ± 0.0266	-0.0186 ± 0.0519

4.3.1 Low capacity under scarcity (Simple CNN)

To evaluate robustness under this data scarcity, we first applied the optimized DANN configuration ($\lambda = 0.3, \alpha =$ `logistic`) to the Simple CNN. The model reached 0.4744 on the target, edging out the baseline (0.4360) for a gain of +0.0384, but this marginal boost came with a substantial degradation in source performance. Source accuracy dropped to 0.8473, corresponding to a *source penalty* of 0.1023 and an overall negative *net benefit* of −0.0639. Validation curves showed extreme volatility, consistent with unstable adversarial dynamics in this low-capacity setting. Overall, these findings suggest that the Simple CNN lacks sufficient capacity to learn transferable features when the source domain is small, yielding an unfavorable trade-off in which adaptation erodes source performance without securing reliable target gains.

Under the canonical formulation ($\lambda = 1.0$), the model exhibited a more aggressive adaptation behavior. Target accuracy reached 0.5073, surpassing both the baseline and the optimized DANN. This yielded a higher target improvement (Δ_T) of +0.0714. Crucially, a significant *source penalty* of 0.0900 offset these gains, leaving a negligible *net benefit* of −0.0186. Although the higher adversarial pressure showed a slightly lower discrepancy under MMD (MMD decreased to 0.2638), the validation metrics revealed extreme volatility. The `val_acc_domain` remained high on average (0.7493 ± 0.1863), indicating that the discriminator was rarely fooled consistently. This suggests that in data-scarce scenarios, high adversarial pressure can force partial adaptation gains, but the Simple CNN architecture lacks the capacity to maintain stability under such a strong signal.

4.3.2 High capacity as a regularizer (ResNet18)

Moving to high-capacity architectures, the ResNet18 backbone demonstrated remarkable robustness even in this challenging direction. As shown in Table 4 (ResNet18 *Scratch*, Baseline), target accuracy reached 0.8160 ± 0.0712, a substantial improvement over the Simple CNN baseline (0.4360). The MMD remained low (0.0195 ± 0.0016) and the target LDA indicated preserved separability ($\text{LDA}_T = 1.4798 \pm 0.0429$), suggesting that higher capacity mitigates part of the covariate shift. However, unlike the mild

shift scenario where the gap was negligible, a substantial performance gap remains relative to source accuracy (0.9858 ± 0.0011). This residual gap ($G_{\text{res}} \approx 0.17$) indicates that, despite the backbone's capacity, data scarcity still limits target generalization.

Table 4 Domain inversion (USPS → MNIST). Using ResNet18 with scratch and ImageNet initialization in a data-scarce regime. All values are reported as mean ± std over three random seeds, using oracle checkpoint selection per seed. Positive Ω indicates a net gain over the source-only baseline; negative values indicate negative transfer.

Configuration	Acc_S ($\mu \pm \sigma$)	Acc_T ($\mu \pm \sigma$)	Δ_S ($\mu \pm \sigma$)	Δ_T ($\mu \pm \sigma$)	Ω ($\mu \pm \sigma$)	MMD ($\mu \pm \sigma$)	LDA_T ($\mu \pm \sigma$)
ResNet18 *(Scratch)*							
Baseline (No Adaptation)	0.9858 ± 0.0011	0.8160 ± 0.0712	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0195 ± 0.0016	1.4798 ± 0.0429
DANN	0.9884 ± 0.0012	0.9488 ± 0.0189	-0.0025 ± 0.0012	0.1328 ± 0.0189	0.1353 ± 0.0194	0.0201 ± 0.0033	1.9090 ± 0.2934
ResNet18 *(ImageNet)*							
Baseline (No Adaptation)	0.9805 ± 0.0046	0.7511 ± 0.0779	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0012 ± 0.0005	1.1880 ± 0.2180
DANN	0.9794 ± 0.0038	0.9102 ± 0.0107	0.0011 ± 0.0038	0.1591 ± 0.0107	0.1580 ± 0.0136	0.0008 ± 0.0000	1.7675 ± 0.1459

The ResNet18 backbone responded strongly to adversarial adaptation in this inverse-shift setting. Target accuracy increased to 0.9488 ± 0.0189 (Table 4, ResNet18 *Scratch*, DANN), clearly improving over the baseline (0.8160 ± 0.0712). Notably, this gain did not incur a source penalty; instead, Δ_S was slightly negative (-0.0025 ± 0.0012), yielding a strongly positive *net benefit* ($\Omega = +0.1353 \pm 0.0194$). In representation terms, target separability also improved (LDA_T: 1.9090 ± 0.2934 vs. 1.4798 ± 0.0429), consistent with tighter class structure under adaptation. We interpret this as *feature compaction*: under data scarcity, adversarial training can regularize the decision space by consolidating discriminative structure rather than flattening it. This outcome highlights a boundary condition for DANN: when the target generalization gap is driven by limited source diversity, adversarial pressure can act constructively as a regularizer, unlike in mild-shift regimes where it may be redundant or even harmful.

4.3.3 Resolving the pre-training paradox

Finally, the pre-trained baseline revealed a notable paradox when compared to its from-scratch counterpart. As shown in Table 4 (ResNet18 *ImageNet*, Baseline), target accuracy (0.7511 ± 0.0779) exceeded the Simple CNN baseline yet still lagged behind the ResNet18 trained from scratch (0.8160 ± 0.0712). Although MMD was extremely low (0.0012 ± 0.0005), indicating very strong distributional similarity under this metric, this did not translate into improved task performance. Consistent with this mismatch, target separability was lower in the pre-trained baseline ($\text{LDA}_T = 1.1880 \pm 0.2180$) than in the scratch baseline ($\text{LDA}_T = 1.4798 \pm 0.0429$), suggesting that ImageNet initialization introduces an inductive bias that is less suitable for digit recognition under inverse shift than features learned from scratch.

Coupling DANN with the pre-trained backbone produced a substantial turnaround (Table 4, ResNet18 *ImageNet*, DANN). Target accuracy increased to 0.9102 ± 0.0107, yielding a strongly positive *net benefit* ($\Omega = +0.1580 \pm 0.0136$) with negligible source penalty ($\Delta_S = 0.0011 \pm 0.0038$). This improvement coincided with a marked increase in class separability (LDA_T: 1.7675 ± 0.1459 vs. 1.1880 ± 0.2180), indicating that adversarial training reshaped the representation to better preserve class structure in the target domain, even when MMD was already very low. This contrasts with the

negative transfer observed under mild shift (Scenario 2; Section 4.2), supporting the conclusion that DANN is most beneficial in challenging adaptation regimes where a substantial target generalization gap remains to be closed.

4.4 Visual analysis of the feature space

t-SNE is used here as a qualitative visualization tool; quantitative interpretations of alignment and separability are supported by MMD and LDA. For each configuration, we visualize the run whose *target test accuracy* is closest to the mean across the three seeds to provide a representative projection.

4.4.1 The simple CNN feature space

The t-SNE visualizations in Fig. 1 illustrate the impact of the optimized DANN on the simple CNN architecture.

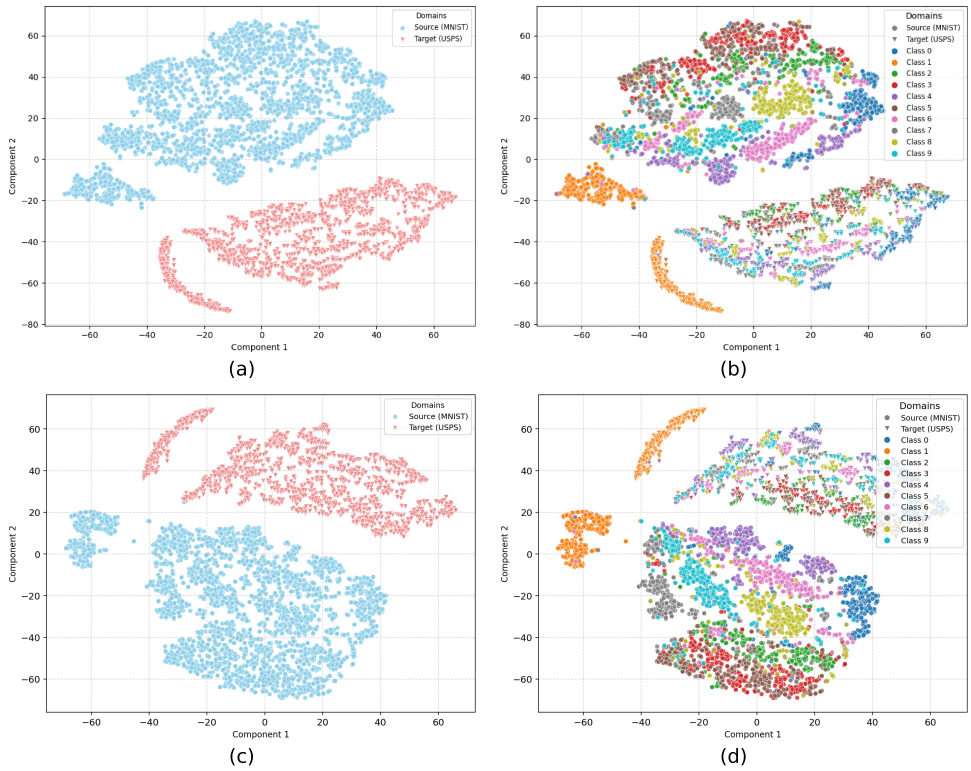


Fig. 1 Limited reorganization under adversarial training (Simple CNN). t-SNE visualization of feature-extractor embeddings for MNIST→USPS. Top: baseline model, showing pronounced domain separation (a) and only partially separated class structure (b). Bottom: optimized DANN ($\lambda = 0.3$, $\alpha = \textit{logistic}$), where domains remain largely separated (c) and the class geometry shows no clear qualitative reorganization (d). These visual patterns are consistent with the low-capacity regime, where adversarial training does not reliably reshape the feature space.

Fig. 1a visualizes the feature embeddings generated by the baseline model. The source (MNIST) and target (USPS) domains appear on opposite sides of the space, forming two distinct macro-clusters with a clear separation between them. This separation is consistent with the large initial performance gap ($G_{\text{init}} = 0.2884$), suggesting that the learned features do not generalize invariantly across domains. Fig. 1b suggests that some class-wise structure is present, but clusters are not uniformly compact or sharply separated—particularly in the target domain—consistent with the limited representational capacity of the Simple CNN.

After adversarial training (Fig. 1c–d), the global geometry remains largely dominated by domain separation rather than domain overlap. The source and target samples still occupy distinct regions in the embedding, and we do not observe a consistent qualitative reorganization of class structure relative to the baseline. In this low-capacity regime, the quantitative metrics indicate partial pressure toward alignment (e.g., reduced MMD), yet the visual evidence suggests that adversarial training does not reliably reshape the representation into a shared, class-structured manifold. This is consistent with the optimization limits observed in Scenario 1 (see Section 4.1): distributional matching signals may change without yielding a stable improvement in target generalization.

4.4.2 Domain segregation under high capacity

Fig. 2 provides a qualitative comparison of the ResNet18 (scratch) feature space with and without adversarial training, highlighting how domain structure can remain segregated even under strong performance.

The top row challenges an interpretation based solely on performance and MMD. Although the scratch baseline attains high target accuracy (0.9586) with a low MMD (0.0238), Fig. 2a shows that source and target samples remain largely separated in the projected space. When colored by class (Fig. 2b), the embedding exhibits compact and well-separated class clusters, yet many digits appear to occupy domain-specific sub-regions rather than overlapping across domains. This pattern is consistent with a *domain segregation* regime, where high accuracy is achieved without enforcing full invariance between MNIST and USPS.

The bottom row suggests a more nuanced effect of adversarial training. In Fig. 2c, domains appear slightly closer in several regions, which is consistent with the modest reduction in MMD (0.0191). However, the embedding does not indicate full mixing into a single invariant manifold. When colored by class (Fig. 2d), class structure remains compact and clearly separated, and some digits exhibit increased proximity between source and target sub-clusters. Importantly, this qualitative pattern aligns with the improved target accuracy (0.9710) and higher LDA_T reported in Table 2, suggesting that adaptation can refine representations without compromising class separability, while only partially reducing domain segregation, in this scratch-trained regime.

In summary, the visual comparison supports the interpretation that the scratch baseline operates in a segregated regime, achieving strong performance with limited cross-domain overlap. Adversarial training appears to reduce domain separation locally while preserving class structure, consistent with the observed improvement in Acc_T and LDA_T. These results suggest that, in this setting, the adversarial signal can

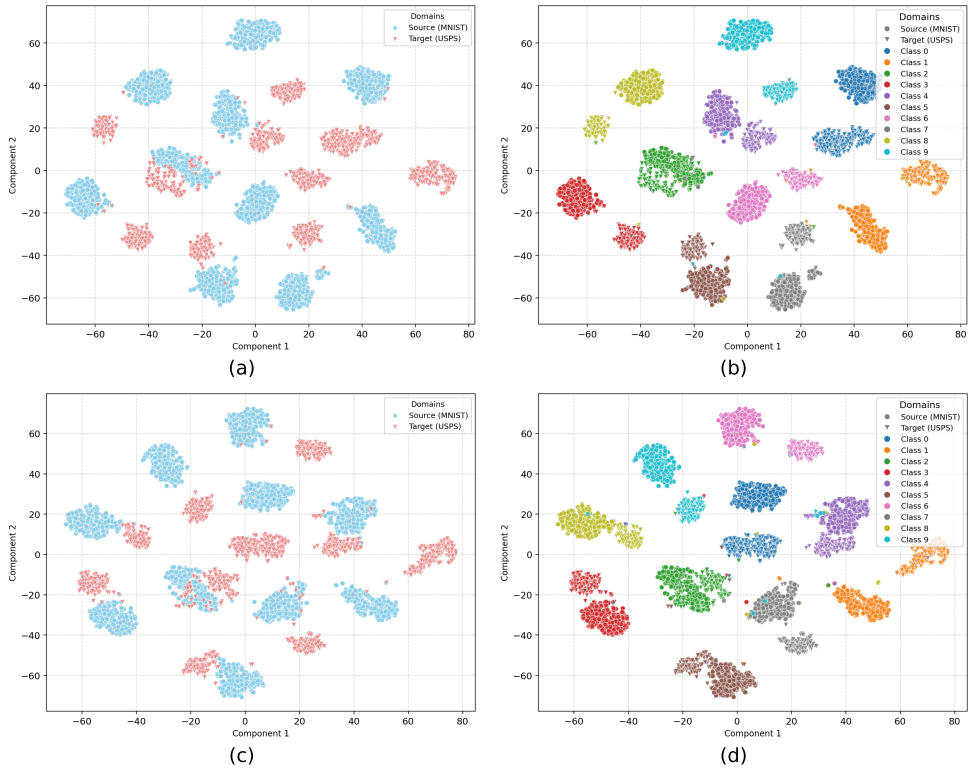


Fig. 2 Segregation vs. partial alignment (ResNet18 Scratch). t-SNE visualization. Top (Baseline): source and target representations form largely separated sub-clusters, indicating domain segregation despite high class separability (a, b). Bottom (DANN): adversarial training reduces domain separation locally, but does not yield full invariant alignment; class structure remains compact and well preserved (c, d). Quantitative comparisons rely on Acc_T, MMD and LDA (Table 2).

act constructively as a mild regularizer rather than inducing the feature flattening observed in lower-capacity regimes.

4.4.3 Negative transfer via feature flattening

Fig. 3 uses t-SNE as a qualitative visualization to contextualize the negative transfer observed when applying DANN to the ImageNet-pretrained ResNet18 backbone in the MNIST→USPS setting.

The top row depicts the baseline feature space. In Fig. 3a, source and target points form a single global cloud with multiple sub-clusters (roughly two per class), reflecting the class–domain decomposition discussed earlier. Fig. 3b suggests heterogeneous class-wise behavior: some digits show close proximity between source and target clusters (e.g., '4'), whereas others remain clearly separated (e.g., '9'). Importantly, most

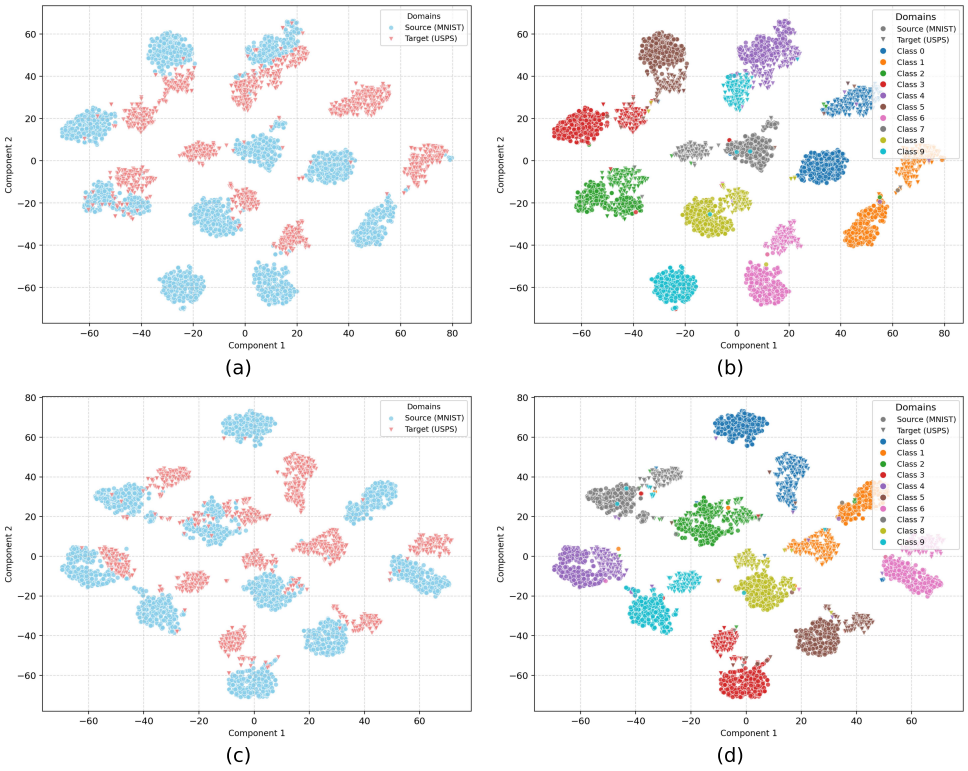


Fig. 3 Feature flattening cues (pre-trained ResNet18). t-SNE visualization of the MNIST→USPS setting. Top (Baseline): class clusters appear relatively compact (b), despite limited domain mixing (a). Bottom (DANN): while some local domain proximity is encouraged (c), several class clusters become more diffuse (d), suggesting increased intra-class variance.

baseline class clusters appear visually compact, which is consistent with the high target discriminability quantified by the LDA_T (2.7379) and the strong baseline target accuracy (0.8993) reported in Table 2.

The bottom row illustrates the effect of adversarial training in this pre-trained regime. Visually, DANN encourages local domain proximity for certain previously disjoint class–domain pairs (Fig. 3c), including the digit '9' (Fig. 3d). However, this apparent class-wise alignment is accompanied by a broader loss of compactness: several class clusters become more diffuse and expanded relative to the baseline. While t-SNE is only qualitative, this pattern is consistent with an increase in intra-class variance and a reduction in class margins. Quantitatively, this degradation is captured by the drop in LDA_T from 2.7379 to 1.8923, alongside the decline in target accuracy from 0.8993 to 0.8385, yielding a negative net benefit ($\Omega = -0.0957$). Taken together, these results support the *feature flattening* interpretation: in a backbone whose representations are already strongly structured by pre-training, additional adversarial pressure can

improve class-wise domain proximity at the cost of diluting discriminative structure, ultimately producing negative transfer.

4.4.4 Feature compaction and semantic alignment

The t-SNE visualizations in Fig. 4 offer qualitative cues that are consistent with DANN's improved performance in the challenging USPS→MNIST scenario, suggesting a pattern of feature compaction and increased source–target class proximity.

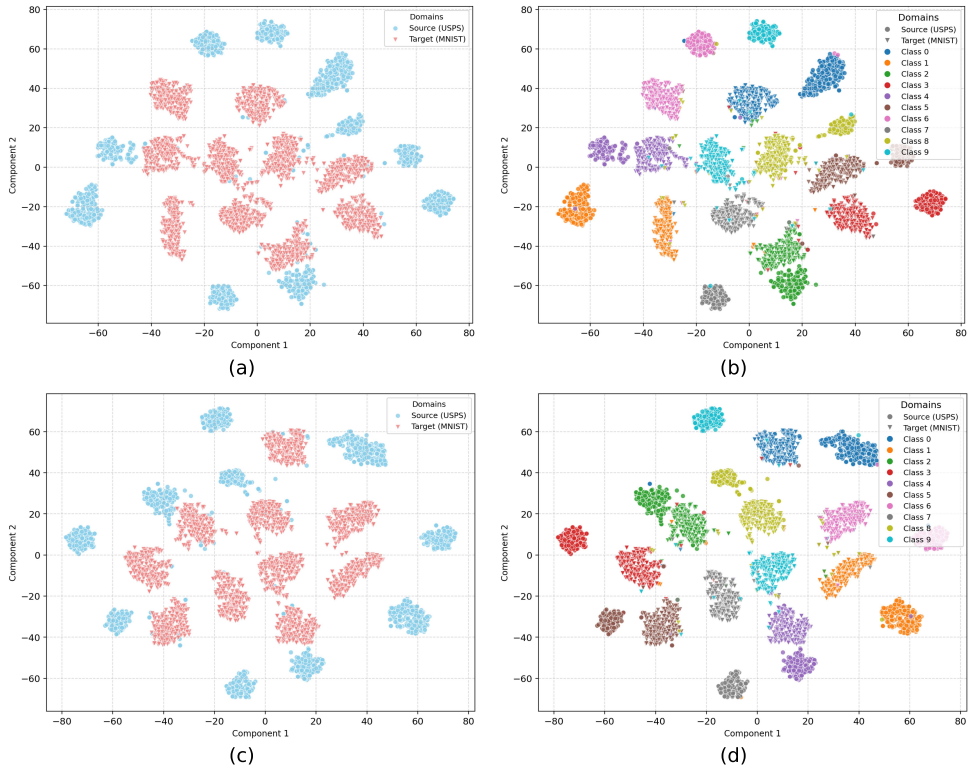


Fig. 4 Feature refinement (USPS → MNIST). t-SNE visualization of domain inversion. Top (Baseline): target clusters appear dispersed and "loose" with numerous stray points (b), indicating weak structure. Bottom (DANN): in contrast to feature flattening, DANN is associated with *feature compaction*. Target clusters tighten visibly, and source "satellites" are pulled into closer proximity (d), consistent with DANN's regularizing effect in data-scarce settings.

The top row shows the baseline feature space. In Fig. 4a, source (USPS) and target (MNIST) points form two largely separated regions, with USPS distributed around the periphery and MNIST concentrated near the center. Fig. 4b further suggests that several target class clusters are comparatively *loose* and exhibit noticeable dispersion, with scattered points drifting away from their main groupings. This qualitative

pattern is consistent with the baseline target LDA of 1.4798 and the corresponding target accuracy of 0.8160, indicating that target representations are less compact and therefore harder to classify reliably.

The bottom row illustrates how DANN reshapes this geometry. Compared to the baseline, Fig. 4d shows tighter, more compact target clusters, alongside increased proximity between the USPS "satellites" and their MNIST counterparts. Importantly, this visual tightening is supported by the quantitative rise in target LDA to 1.9090, indicating improved class separability in the target feature space. In parallel, the closer source–target positioning is consistent with improved alignment, reflected in the accompanying discrepancy metrics. Together, these changes help explain the substantial target accuracy gain to 0.9488, suggesting that in this data-scarce regime DANN acts as a constructive regularizer: it discourages diffuse, ambiguous target representations and promotes more stable, discriminative structure.

Taken together, these visual patterns—interpreted cautiously and in conjunction with the quantitative MMD and LDA trends—suggest that adversarial alignment may either dilute or refine representations, depending on whether the adversarial objective conflicts with or supports class-discriminative structure.

5 Discussion

In this work, we sought to delineate the *operational limits* of DANN beyond the binary question of whether it "works." Rather than treating adversarial alignment as an off-the-shelf remedy for covariate shift, we isolated the conditions under which it (i) improves generalization, (ii) becomes redundant, or (iii) induces negative transfer. Across MNIST↔USPS, our evidence indicates that DANN's effect is strongly situational and primarily governed by the interaction between feature-extractor capacity, prior knowledge (e.g., pre-training), and the magnitude/direction of the shift.

Architectural constraints vs. hyperparameter refinement. A consistent empirical finding is that hyperparameter tuning alone is unlikely to offset insufficient representational capacity. The persistent failure of the Simple CNN across the explored λ settings and α strategies suggests that when a backbone cannot learn stable class-discriminative structure in the first place, the adversarial signal may act less as a regularizer and more as a source of optimization interference. Conversely, the strong performance of the ResNet18 *source-only* baseline under mild shift (MNIST→USPS) challenges the assumption that explicit adaptation is universally beneficial: with adequate capacity, robust class structure can emerge that is already tolerant to mild covariate differences, making adversarial alignment unnecessary and sometimes harmful.

The pre-training paradox and *feature flattening*. The negative transfer observed with pre-trained ResNet18 highlights a critical failure mode: statistical alignment and predictive utility can diverge. In this regime, extremely low MMD coexisted with degraded target performance, while LDA dropped sharply, indicating a loss of class separability. We interpret this pattern as *feature flattening*: in aggressively minimizing domain distinguishability, the adversarial objective can collapse the rich discriminative geometry inherited from ImageNet, erasing precisely the structure that enables

transfer. Taken together, these patterns—interpreted cautiously and supported by the quantitative MMD and LDA trends—suggest that for pre-trained backbones, the standard adversarial objective may undermine task discrimination by over-penalizing domain-specific but class-informative variations.

Importantly, these outcomes (*net benefit* < 0) are unlikely to be explained by checkpoint timing. Under our oracle model-selection protocol, the selected model is the checkpoint achieving the best observed target validation accuracy, so the failure to surpass the "source only" baseline cannot be attributed to an unfortunate stopping epoch. Instead, the evidence supports the interpretation that the adversarial objective itself can actively erode class separability in this setting, consistent with the *feature flattening* mechanism. This feature flattening phenomenon shares similarities with observations that fine-tuning can distort pre-trained representations [34], suggesting that aggressive adaptation pressures may generally degrade robust features.

From a theoretical perspective [18], successful adaptation requires both minimizing source risk and reducing domain divergence without collapsing discriminative structure.

The optimal regime: adaptation as essential regularization. Our domain inversion experiments (USPS→MNIST) identify the regime in which DANN is most clearly advantageous: high-capacity backbones operating under data scarcity, where the baseline does not fully close the generalization gap. In this intermediate zone, adversarial training behaves as an *essential regularizer* rather than a generic aligner. By suppressing source-specific feature patterns that overfit the source data and produce unstable or outlier representations when applied to the target domain, DANN can refine decision boundaries and improve target performance when capacity is available but source-labeled data are scarce.

Redefining alignment: semantic proximity over geometrical indistinguishability. Finally, our results motivate a practical reframing of what successful adaptation should optimize. The best-performing configurations did not correspond to the lowest possible MMD or to complete geometrical overlap. Instead, they approached what we term *semantic proximity*: source and target clusters become compatible with shared decision boundaries while preserving internal class-discriminative structure. This suggests that the theoretical ideal of total invariance can be practically suboptimal, and that functional alignment—reducing domain-level statistical divergence (e.g., MMD) without collapsing class-discriminative geometry—should be the primary objective.

6 Conclusions and future work

This study delineates the practical boundaries of adversarial domain adaptation. Across MNIST↔USPS, we showed that while hyperparameter refinement is important for stability, it cannot compensate for insufficient representational capacity: when the backbone fails to learn robust class-discriminative structure, adversarial training does not reliably recover adaptation. These results position DANN not as a default remedy for covariate shift, but as a conditional mechanism whose utility depends on the interaction between backbone capacity, prior knowledge (e.g., pre-training), and shift characteristics.

Our experiments further indicate that under mild covariate shift with high-capacity backbones, explicit adaptation can be redundant and may even be detrimental. In particular, for pre-trained models we observed a negative-transfer regime in which strong statistical alignment coexisted with degraded target performance, consistent with a *feature flattening* mechanism where adversarial pressure erodes class-discriminative geometry. Conversely, in the inverse-shift setting characterized by source data scarcity (USPS→MNIST), adversarial training acted as a beneficial regularizer—provided sufficient backbone capacity—and yielded consistent target gains. Taken together, these outcomes support a functional view of alignment: reducing domain-level statistical divergence without collapsing class geometry is more predictive of success than pursuing total invariance.

Importantly, these conclusions are strengthened by our evaluation protocol. Under our oracle model-selection protocol, the selected model is the checkpoint with the highest target validation accuracy; therefore, reported performance drops reflect each configuration's best attainable target performance under this protocol rather than an unfortunate stopping epoch. This makes a checkpoint-timing explanation unlikely and supports the interpretation that the observed limitations arise from the interaction between backbone capacity and adversarial pressure.

Future work will extend this analysis in three directions. First, we will evaluate non-convolutional feature extractors, particularly Vision Transformers [33], to test whether global attention alters susceptibility to negative transfer and feature flattening. Second, we will validate the observed dynamics on higher-variance natural image benchmarks (e.g., CIFAR-10 [41] and STL-10 [42]) to assess robustness under greater intra-class variability. Third, we will explore adaptive strategies that modulate adversarial pressure based on online indicators of class structure (e.g., LDA trends), with the goal of preserving class geometry while still reducing domain divergence.

Finally, we will return to the application that motivated this study: adapting phytoplankton classification from Scanning Electron Microscopy to Optical Microscopy. The severity of this modality shift likely places the problem outside the regime where adversarial alignment is reliably beneficial, helping to explain the limited gains observed in practice. Building on the regimes identified here, we will revisit phytoplankton classification using a more principled selection of adversarial pressure and adaptation strategy, with the goal of improving transfer while reducing the risk of negative transfer.

Declarations

Funding
There was no funding for this work.

References

[1] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015) https://doi.org/10.1038/nature14539

[2] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset Shift in Machine Learning. MIT Press, Cambridge, MA (2009)

[3] Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010) https://doi.org/10.1109/TKDE.2009.191

[4] Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312**, 135–153 (2018) https://doi.org/10.1016/j.neucom.2018.05.083

[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016)

[6] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. ICML'15, pp. 1180–1189. JMLR.org, Lille, France (2015)

[7] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2962–2971 (2017). https://doi.org/10.1109/CVPR.2017.316

[8] Ozyurt, Y., Feuerriegel, S., Zhang, C.: Contrastive Learning for Unsupervised Domain Adaptation of Time Series (2023). https://arxiv.org/abs/2206.06243

[9] Kim, D., Wang, K., Sclaroff, S., Saenko, K.: A broad study of pre-training for domain generalization and adaptation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, pp. 621–638. Springer, Cham (2022)

[10] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998) https://doi.org/10.1109/5.726791

[11] Hull, J.J.: A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence **16**(5), 550–554 (1994) https://doi.org/10.1109/34.291440

[12] Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3723–3732 (2018). https://doi.org/10.1109/CVPR.2018.00392

[13] You, K., Long, M., Cao, Z., Wang, J., Jordan, M.I.: Universal domain adaptation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2715–2724 (2019). https://doi.org/10.1109/CVPR.2019.00283

[14] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[15] Wang, Z., Dai, Z., Póczos, B., Carbonell, J.G.: Characterizing and avoiding negative transfer. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11285–11294 (2018)

[16] Chen, X., Wang, S., Long, M., Wang, J.: Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 1081–1090. PMLR, Long Beach, California, USA (2019). https://proceedings.mlr.press/v97/chen19i.html

[17] Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)

[18] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine Learning **79**(1), 151–175 (2010) https://doi.org/10.1007/s10994-009-5152-4

[19] Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 97–105. PMLR, Lille, France (2015). https://proceedings.mlr.press/v37/long15.html

[20] Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17, pp. 2208–2217. JMLR.org, Sydney, NSW, Australia (2017)

[21] Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Hua, G., Jégou, H. (eds.) Computer Vision – ECCV 2016 Workshops, pp. 443–450. Springer, Cham (2016)

[22] Zhang, W., Zhang, X., Lan, L., Luo, Z.: Maximum mean and covariance discrepancy for unsupervised domain adaptation. Neural Processing Letters **51**(1), 347–366 (2020) https://doi.org/10.1007/s11063-019-10090-0

[23] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. NIPS'14, pp. 2672–2680. MIT Press, Cambridge, MA, USA (2014)

[24] Pei, Z., Cao, Z., Long, M., Wang, J.: Multi-adversarial domain adaptation. In:

Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. AAAI'18. AAAI Press, New Orleans, Louisiana, USA (2018)

[25] Choi, E., Rodriguez, J., Young, E.: An In-Depth Analysis of Adversarial Discriminative Domain Adaptation for Digit Classification (2025). https://arxiv.org/abs/2412.19391

[26] Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS'18, pp. 1647–1657. Curran Associates Inc., Red Hook, NY, USA (2018)

[27] HassanPour Zonoozi, M., Seydi, V.: A survey on adversarial domain adaptation. Neural Processing Letters **55**(3), 2429–2469 (2023) https://doi.org/10.1007/s11063-022-10977-5

[28] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 214–223. PMLR, Sydney, Australia (2017). https://proceedings.mlr.press/v70/arjovsky17a.html

[29] Allen-Zhu, Z., Li, Y., Song, Z.: A convergence theory for deep learning via over-parameterization. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 242–252. PMLR, Long Beach, CA, USA (2019). https://proceedings.mlr.press/v97/allen-zhu19a.html

[30] Neyshabur, B., Salakhutdinov, R., Srebro, N.: Path-sgd: path-normalized optimization in deep neural networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2. NIPS'15, pp. 2422–2430. MIT Press, Cambridge, MA, USA (2015)

[31] Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2656–2666 (2019). https://doi.org/10.1109/CVPR.2019.00277

[32] Nguyen, C.V., Hassner, T., Seeger, M., Archambeau, C.: Leep: a new measure to evaluate transferability of learned representations. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20. JMLR.org, Vienna, Austria (2020)

[33] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021). https://openreview.net/forum?id=YicbFdNTTy

[34] Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. ArXiv **abs/2202.10054** (2022)

[35] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 448–456. PMLR, Lille, France (2015). https://proceedings.mlr.press/v37/ioffe15.html

[36] Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. ICML'10, pp. 807–814. Omnipress, Madison, WI, USA (2010)

[37] Lin, M., Chen, Q., Yan, S.: Network in network. CoRR **abs/1312.4400** (2013)

[38] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848

[39] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015). http://arxiv.org/abs/1412.6980

[40] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.J.: A kernel two-sample test. Journal of Machine Learning Research **13**(25), 723–773 (2012)

[41] Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario (2009). https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[42] Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Gordon, G., Dunson, D., Dudík, M. (eds.) Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 15, pp. 215–223. PMLR, Fort Lauderdale, FL, USA (2011). https://proceedings.mlr.press/v15/coates11a.html