JOINT CONTRASTIVE-ADVERSARIAL REPRESENTATION LEARNING FOR UNSUPERVISED DOMAIN ADAPTATION

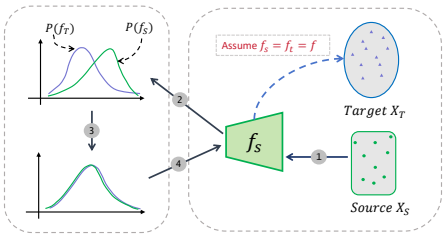
Anonymous authors
Paper under double-blind review

ABSTRACT

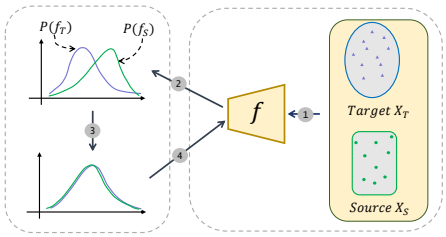
Effective unsupervised domain adaptation (UDA) requires a representation that captures transferable structure from both the labeled source domain and the unlabeled target domain. We introduce Joint Contrastive-Adversarial Representation Learning (JCAR), a two-stage framework that learns a common representation from the union of unlabeled source and target samples by combining contrastive learning with adversarial domain alignment, and then freezes the representation to fit a source-supervised predictor. We establish a non-asymptotic excess-risk bound for this procedure that couples representation and domain-discriminator estimation while separating downstream prediction, with explicit neural-network approximation errors. Under the stated assumptions and common Hölder smoothness $\beta \geq 1$, the bound identifies a sufficiently large unlabeled-sample regime in which the downstream rate is $\widetilde{\mathcal{O}}(m^{-\beta/(2\beta+d)})$, where m counts labeled source samples and logarithmic factors are suppressed. The labeled-sample exponent is governed by the latent dimension d rather than the ambient dimension p, which enters the pretraining cost. Experiments on Office-31, Office-Home, and DomainNet demonstrate clear transfer gains across diverse domain shifts, with particularly strong performance when representations are learned from scratch.

1 INTRODUCTION

Domain adaptation (DA) (Pan & Yang, 2010) aims to transfer a predictor from a labeled source domain to an unlabeled target domain. UDA methods use target samples in different ways: for example, statistical matching aligns feature distributions (Gretton et al., 2012), reweighting corrects distribution shift (Kato & Teshima, 2021), and adversarial methods train a feature extractor against a domain discriminator (Long et al., 2018). In many such methods, however, task-discriminative supervision is available only in the source domain, while target samples enter primarily through an alignment objective. This asymmetry motivates us to ask whether both domains can participate directly in unsupervised representation construction before the source-supervised predictor is fitted. Figure 1a contrasts the conventional source-supervised adaptation pipeline with the joint-domain representation-learning pipeline in Figure 1b.



(a) A conventional UDA representation-learning pipeline.



(b) The proposed joint-domain representation-learning pipeline.

To address this question, we introduce Joint Contrastive-Adversarial Representation Learning (JCAR). In Stage 1, contrastive learning (Chen et al., 2020) allows unlabeled source and target samples

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to jointly shape the representation, while an adversarial domain objective explicitly reduces the distinguishability of their feature distributions. In Stage 2, the learned feature extractor is frozen and a predictor is trained on labeled source samples. This design exploits target-domain structure without target labels and cleanly separates task-agnostic representation learning from downstream prediction.

Our non-asymptotic analysis jointly controls representation learning and domain discrimination while separating their cost from downstream prediction. It explicitly balances neural-network approximation and statistical errors, linking the coupled complexity of contrastive–adversarial pretraining to downstream label efficiency. Under the theorem's assumptions, sufficiently many unlabeled samples yield a downstream polynomial rate governed by the representation dimension d, while the ambient dimension p enters the pretraining cost.

Our contributions are summarized as follows:

- We formulate a two-stage UDA procedure in which unlabeled source and target samples jointly drive contrastive representation learning and explicit adversarial alignment. Freezing the learned extractor in the second stage decouples this task-agnostic representation objective from source-supervised prediction.
- We derive a coupled pretraining rate by jointly balancing the statistical and approximation costs of the feature extractor and domain discriminator. Under the assumptions of Theorem 3, common Hölder smoothness $\beta \geq 1$, and sufficiently many unlabeled samples, the excess-risk bound reduces to $\widetilde{\mathcal{O}}(m^{-\beta/(2\beta+d)})$, where m counts labeled source samples and the tilde suppresses logarithmic factors. This quantifies the distinct roles of unlabeled data in high-dimensional feature learning and labels in low-dimensional prediction.
- We evaluate JCAR under both random initialization and supervised pretraining on Office-31, Office-Home, and DomainNet. In the from-scratch setting, JCAR leads the ResNet group on four of six Office-31 transfers, and its ViT variant improves over ViT-B on four of six Office-31 and five of six Office-Home transfers; feature visualizations further illustrate the cross-domain alignment learned by the method.

1.1 Organization

Section 2 fixes the notation and basic concepts before JCAR is developed in Section 3. Its guarantees and empirical behavior are examined in Sections 4 and 5, respectively, while Section 6 closes with the main conclusions.

2 Preliminaries

2.1 Notations

Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and output spaces, respectively, where $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Y} \subset \mathbb{R}$. Without loss of generality, we constrain $\mathcal{X} = [0,1]^p$ in this work. Let $\mathcal{S}$ and $\mathcal{T}$ be subsets of $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{S}$ is the source domain and $\mathcal{T}$ is the target domain. Denote the source covariate–response pair by $(X^{\mathcal{S}}, Y^{\mathcal{S}}) \in \mathcal{X} \times \mathcal{Y}$ and its distribution by $P_S(X,Y)$. Similarly, $(X^{\mathcal{T}}, Y^{\mathcal{T}})$ has target distribution $P_T(X,Y)$. The corresponding covariate marginals are $P_S(X)$ and $P_T(X)$. In the domain-adaptation setting, we assume $P_S(X) \neq P_T(X)$ and $P_S(Y \mid X) = P_T(Y \mid X)$. Let $\mathcal{M}_f := \{f : \mathbb{R}^p \to \mathbb{R}^d\}$, $\mathcal{M}_h := \{h : \mathbb{R}^d \to \mathbb{R}\}$ and $\mathcal{M}_g := \{g : \mathbb{R}^d \to \mathbb{R}\}$ be the Hölder classes where $d < p$. Let $\mathcal{F} := \{f : \mathbb{R}^p \to \mathbb{R}^d\}$, $\mathcal{H} := \{h : \mathbb{R}^d \to \mathbb{R}\}$, and $\mathcal{G} := \{g : \mathbb{R}^d \to \mathbb{R}\}$ denote the hypothesis function classes. In this work, these functions are implemented as DNNs. Let $\|f\|_{L_q(\mu)}$ denote the L_q norm of f with respect to the probability measure μ. For $x \in \mathbb{R}^p$, we write $\|x\|_q := (\sum_{j=1}^p |x_j|^q)^{1/q}$ for $q \in [1, \infty)$. Let $a, b > 0$. We write $a \lesssim b$ if there exists a constant $c > 0$ such that $a \leq cb$. For any $y \in \mathbb{R}$, we denote $\lfloor y \rfloor := \max\{n \in \mathbb{Z} | n \leq y\}$, where $\mathbb{Z}$ is the integer set.

Definition 2.1 (Contrastive loss)**.** *Let $\{X_i\}_{i=1}^n$ be i.i.d. samples from $P_{\mathcal{X}}$ on $\mathcal{X}$. For each anchor X_i, let X_i^+ be an independently augmented view of the same underlying sample, and let X_i^- be a view associated with a different sample drawn independently from $P_{\mathcal{X}}$. Let $f : \mathbb{R}^p \to \mathbb{R}^d$ be a representation function, where $d < p$. For the theoretical analysis, we use the following pairwise*

logistic contrastive loss with one sampled negative:

$$\ell_{cl}(X, X^+, X^-, f) := \log\left(1 + \exp\{f(X)^\top \big(f(X^-) - f(X^+)\big)\}\right).$$

To simplify notation, we write $\ell_{cl}(X, f)$ when the sampling rule for (X^+, X^-) is clear. The experimental mini-batch objective extends this pairwise comparison by using other mini-batch samples as negatives (Appendix N), while the expression above enables a tractable statistical analysis of its core contrastive mechanism. The empirical risk of f is

$$\hat{\mathcal{R}}_{cl}(f) := \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\{f(X_i)^\top \big(f(X_i^-) - f(X_i^+)\big)\}\right).$$

Definition 2.2 (Domain discrimination loss)**.** *Let $\{(X_k, Z_k)\}_{k=1}^n$ be i.i.d. samples from a joint distribution $P_\mu(X, Z)$, where Z_k is a domain label taking values in $\{-1, 1\}$. Let $f : \mathbb{R}^p \to \mathbb{R}^d$ be a representation function, and $h : \mathbb{R}^d \to \mathbb{R}$ be a domain discriminator. We define the domain-classification utility as the negative logistic loss:*

$$\ell_d(X, Z, f, h) := -\log(1 + \exp(-Z \cdot h \circ f(X))).$$

The empirical risk of the pair (f, h) is defined by

$$\hat{\mathcal{R}}_d(f, h) := -\frac{1}{n}\sum_{k=1}^{n} \log(1 + \exp(-Z_k \cdot h \circ f(X_k))).$$

Under the minimax objective (3)*, the discriminator h maximizes $\hat{\mathcal{R}}_d$, equivalently minimizing the usual positive logistic loss, whereas the feature extractor f minimizes the same utility to make the two domains harder to distinguish. This sign convention therefore implements adversarial feature alignment.*

Definition 2.3 (Rademacher complexity)**.** *Let $X_{1:n} := \{X_i\}_{i=1}^n$ be n i.i.d. random variables on $\mathcal{X}$, and let $\mathcal{F}$ be a set of measurable functions defined on $\mathcal{X}$. Let $\zeta_{1:n} := \{\zeta_i\}_{i=1}^n$ be n i.i.d. Rademacher random variables, each uniformly taking values -1 or 1 with probability $1/2$. Then, the empirical Rademacher complexity of $\mathcal{F}$ is defined as*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_{\zeta_{1:n}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \zeta_i f(X_i)\right].$$

The Rademacher complexity of $\mathcal{F}$ is the expectation of $\hat{\mathfrak{R}}_n(\mathcal{F})$:

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{X_{1:n}}[\hat{\mathfrak{R}}_n(\mathcal{F})].$$

3 Problem Formulation

JCAR has two stages: joint-domain representation learning (Section 3.1) and source-supervised predictor fitting on frozen features (Section 3.2). The resulting predictor is then applied to the target domain.

3.1 Representation learning

Our central objective is to learn representations that not only capture rich information from both the source and target domains but also exhibit domain invariance. To this end, we propose an unsupervised representation learning framework that integrates the principles of contrastive learning and domain discrepancy minimization, and we accordingly formulate a contrastive domain-adaptive loss to jointly enforce semantic consistency and domain alignment. Unlike conventional domain-adaptation pipelines (Figure 1a), which use source supervision to shape the representation and then align it with the target domain, our approach uses source and target samples jointly during representation learning (Figure 1b). This directly incorporates target-domain structure into the learned representation.

We have a source domain dataset $D_{\mathcal{S}} := \{X_i^{\mathcal{S}}\}_{i=1}^{n_s}$ i.i.d. drawn from $P_S(X)$, and a target domain dataset $D_{\mathcal{T}} := \{X_j^{\mathcal{T}}\}_{j=1}^{n_t}$ i.i.d. drawn from $P_T(X)$. Let $n := n_s + n_t$, and denote $\alpha_n := \frac{n_s}{n}$.

Without loss of generality, we assume that $\alpha_n \to \alpha \in (0,1)$ as n tends to ∞. Accordingly, we introduce X as a mixture of source and target covariates, and let Z denote its pseudo label, with $Z = 1$ for $X = X^{\mathcal{S}} \sim P_S$ and $Z = -1$ for $X = X^{\mathcal{T}} \sim P_T$. Precisely, the mixture X can be formulated as

$$X = \frac{1+Z}{2} X^{\mathcal{S}} + \frac{1-Z}{2} X^{\mathcal{T}},$$

where Z is a Rademacher-type domain indicator with $P(Z=1) = \alpha$ and $P(Z=-1) = 1-\alpha$, and is independent of $X^{\mathcal{S}}$ and $X^{\mathcal{T}}$. In this case, we denote the joint distribution of (X, Z) by $P_\mu(X, Z)$, whose marginal distribution of X is $P_C(X) = \sum_{z\in\{-1,1\}} P(X \mid Z = z)P(Z = z) = \alpha P_S(X) + (1-\alpha)P_T(X)$. Conditionally on an anchor $X = x$, a positive view is sampled from an augmentation kernel $P_A(\cdot \mid x)$, whereas a negative X^- is associated with an independently drawn sample from P_C. Thus the joint sampling law used by Definition 2.1 is $P_C(dx)P_A(dx^+ \mid x)P_C(dx^-)$, rather than a product law that makes the positive view independent of its anchor.

Let $f : \mathbb{R}^p \to \mathbb{R}^d$ be a representation function, and let $h : \mathbb{R}^d \to \mathbb{R}$ be a domain discriminator. The contrastive domain-adaptive loss $\ell_{ad}(X, Z, f, h)$ is defined as

$$\ell_{ad}(X, Z, f, h) := \ell_{cl}(X, f) + \ell_d(X, Z, f, h), \tag{1}$$

where ℓ_{cl} is the contrastive loss (see Definition 2.1), and ℓ_d is the domain discrimination loss (see Definition 2.2). Let $D_J := \{(X_k, Z_k)\}_{k=1}^n$ be i.i.d. paired samples drawn from $P_\mu(X, Z)$. Let $D_C := \{X_1, \ldots, X_n\}$ denote the covariates and let $D_{\bar{C}} := \{(X_i, X_i^+, X_i^-)\}_{i=1}^n$. In this work, we assume by convention that $D_C = D_{\mathcal{S}} \bigcup D_{\mathcal{T}}$. The empirical risk associated with the contrastive domain-adaptive loss is then given by

$$\hat{\mathcal{R}}_{ad}(f, h) := \hat{\mathcal{R}}_{cl}(f) + \hat{\mathcal{R}}_d(f, h) = \frac{1}{n}\sum_{k=1}^{n} \ell_{ad}(X_k, Z_k, f, h). \tag{2}$$

Accordingly, the estimated representation function is defined as the solution to the following minimax problem:

$$\hat{f} \in \arg\min_{f\in\mathcal{F}} \max_{h\in\mathcal{H}} \hat{\mathcal{R}}_{ad}(f, h), \tag{3}$$

where $\mathcal{F}$ and $\mathcal{H}$ are the norm-constrained DNNs.

3.2 Downstream tasks

After obtaining the representation function $\hat{f}$ in (3), we proceed to learn a predictor for the downstream task. Owing to this domain-invariance property, the predictor can be effectively trained on the source domain, as established in Lemma 1. In this work, we focus on binary classification as a representative example; however, the framework can be readily extended to other supervised learning tasks.

Given m labeled source samples $\bar{D}_{\mathcal{S}} := \{(X_j^{\mathcal{S}}, Y_j^{\mathcal{S}})\}_{j=1}^m$ with $Y_j^{\mathcal{S}} \in \{-1, 1\}$, the predictor $\hat{g}$ in (4) is trained under the empirical risk minimization (ERM) framework by combining the representation $\hat{f}$ with the margin-logistic loss:

$$\hat{g} \in \arg\min_{g\in\mathcal{G}} \frac{1}{m}\sum_{j=1}^{m} \ell(g \circ \hat{f}(X_j^{\mathcal{S}}), Y_j^{\mathcal{S}}), \tag{4}$$

where $\mathcal{G}$ is a class of DNNs mapping $\mathbb{R}^d$ to $\mathbb{R}$. Here, $\ell(a, y) := \log(1 + \exp(-ya))$ is the margin-logistic loss for a score $a \in \mathbb{R}$ and a signed label $y \in \{-1, 1\}$. Since $|\partial\ell(a, y)/\partial a| = 1/(1 + \exp(ya)) \le 1$, the loss is 1-Lipschitz in its score argument, uniformly over y. The complete procedure is the two-stage algorithm in Appendix A: first estimate $\hat{f}$ using unlabeled source and target samples, then freeze $\hat{f}$ while fitting $\hat{g}$ on labeled source samples.

4 Error Analysis

We bound the excess risk of the two-stage estimator $(\hat{f}, \hat{g})$ through its representation and prediction errors. To connect source-supervised fitting with target prediction, we first consider the distributional conditions for transferring risk guarantees in the representation space.

Lemma 1. *Let f denote the common representation function for the source and target domains, and $\mathcal{M}_g$ be the Hölder class. Let $U^{\mathcal{S}} := f(X^{\mathcal{S}})$ and $U^{\mathcal{T}} := f(X^{\mathcal{T}})$. Assume that $P_S(U^{\mathcal{S}}) = P_T(U^{\mathcal{T}})$ and $P_S(Y^{\mathcal{S}} \mid U^{\mathcal{S}}) = P_T(Y^{\mathcal{T}} \mid U^{\mathcal{T}})$. Denote*

$$g^{**} \in \arg\min_{g\in\mathcal{M}_g} \mathbb{E}_{P_T(U^{\mathcal{T}},Y^{\mathcal{T}})}[\ell(g(U^{\mathcal{T}}), Y^{\mathcal{T}})], \quad g^* \in \arg\min_{g\in\mathcal{M}_g} \mathbb{E}_{P_S(U^{\mathcal{S}},Y^{\mathcal{S}})}[\ell(g(U^{\mathcal{S}}), Y^{\mathcal{S}})],$$

then $g^{**} = g^*$.

When the source and target feature marginals and conditional label distributions coincide, their prediction risks agree for every predictor in the common representation space. This risk identity transfers source excess-risk guarantees to the target domain and provides the distributional basis for source-supervised downstream fitting.

We next introduce the notation used in the analysis. As defined in Section 3.1, $P_\mu(X, Z)$ is the joint distribution of the mixture covariate $X \sim P_C$ and its domain label Z. Recalling the definitions of contrastive domain-adaptive loss ℓ_{ad} in (1) and its empirical risk $\hat{\mathcal{R}}_{ad}(f, h)$ in (2), we define the population-level contrastive domain-adaptive risk $\mathcal{R}_{ad}(f, h)$ as

$$\mathcal{R}_{ad}(f,h) := \mathbb{E}_{P_\mu,P_A}[\ell_{ad}(X,Z,f,h)] = \mathbb{E}_{P_C\times P_A}[\ell_{cl}(X,f)] + \mathbb{E}_{P_\mu}[\ell_d(X,Z,f,h)].$$

Let $\mathcal{M}_f$ and $\mathcal{M}_h$ be the Hölder classes. We denote $\mathcal{R}_{cl}(f) := \mathbb{E}_{P_C\times P_A}[\ell_{cl}(X, f)]$ and $\mathcal{R}_d(f, h) := \mathbb{E}_{P_\mu}[\ell_d(X, Z, f, h)]$, and define the representation function f^* at population level:

$$f^* \in \arg\min_{f\in\mathcal{M}_f} \max_{h\in\mathcal{M}_h} \mathcal{R}_{ad}(f,h).$$

To simplify the notations, we define

$$\mathcal{R}_{\mathcal{M}_h}(f) := \sup_{h\in\mathcal{M}_h} \mathcal{R}_{ad}(f,h), \quad \mathcal{R}_{\mathcal{H}}(f) := \sup_{h\in\mathcal{H}} \mathcal{R}_{ad}(f,h), \quad \hat{\mathcal{R}}_{\mathcal{H}}(f) := \sup_{h\in\mathcal{H}} \hat{\mathcal{R}}_{ad}(f,h). \tag{5}$$

Consequently, we have

$$f^* \in \arg\min_{f\in\mathcal{M}_f} \mathcal{R}_{\mathcal{M}_h}(f), \quad \hat{f} \in \arg\min_{f\in\mathcal{F}} \hat{\mathcal{R}}_{\mathcal{H}}(f).$$

By Lemma 1, we now consider the excess risk on the source domain between $\mathcal{R}_{\hat{f}}(\hat{g})$ and $\mathcal{R}_{f^*}(g^*)$, i.e.,

$$\mathcal{E}_{\hat{f}}(\hat{g}) := \mathcal{R}_{\hat{f}}(\hat{g}) - \mathcal{R}_{f^*}(g^*) = \mathbb{E}_{P_S(X,Y)}[\ell(\hat{g}\circ\hat{f}(X), Y)] - \mathbb{E}_{P_S(X,Y)}[\ell(g^*\circ f^*(X), Y)].$$

To establish the error bound of $\mathcal{E}_{\hat{f}}(\hat{g})$, we require the following regularity assumptions:

Assumption 4.1 (Smoothness and bounded representations)**.** *The ideal classes are the Hölder balls of Definition C.1, with fixed radii B_f, B_h, B_g and smoothness $\beta_f, \beta_h, \beta_g > 0$. The class $\mathcal{M}_f$ has d outputs; $\mathcal{M}_h, \mathcal{M}_g$ are scalar-valued on $[0,1]^d$. Inputs and augmented views lie in $[0,1]^p$. Both f^* and all $f \in \mathcal{F}$ take values in $[0,1]^d$, so $\|f(x)\|_2 \le R := \sqrt{d}$, independently of norm budgets. Write $\beta_j = \bar{\beta}_j + \tilde{r}_j$ with $\bar{\beta}_j = \lceil\beta_j\rceil - 1$, $\tilde{r}_j \in (0,1]$, and put $\rho_h = \min\{\beta_h, 1\}$.*

The range restriction specifies the theoretical network class. Clipping gives bounded ReLU approximants with constant-factor norm overhead and one extra hidden layer (Lemma 11); this does not assert that experimental backbones implement clipping.

Assumption 4.2 (Calibration)**.** *For the selected estimator $\hat{f}$, almost surely,*

$$\mathbb{E}_{P_C}\|\hat{f}(X) - f^*(X)\|_2^{\tilde{r}_g} \le C_{\mathrm{cal}}\{\mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{M}_h}(f^*)\},$$

where C_{cal} is independent of sample sizes and network budgets and $\tilde{r}_g$ is as in Assumption 4.1. The comparison uses a fixed representation-coordinate convention.

Remark 2. Assumption 4.1 specifies Hölder regularity, a standard framework for nonparametric approximation (Györfi et al., 2002; Wasserman, 2006). The ideal functions need not belong to the chosen neural-network classes. Our analysis retains the resulting approximation errors and balances them against statistical complexity, making network capacity an explicit part of the convergence guarantee.

Assumption 4.2 converts excess pretraining risk into representation-error control for the selected estimator, linking the contrastive–adversarial objective to the downstream prediction bound. More generally, a calibration condition can be expressed as $\mathbb{E}_X G(f(X)-f^*(X)) \lesssim \mathcal{R}_{\mathcal{M}_h}(f)-\mathcal{R}_{\mathcal{M}_h}(f^*)$, where $G : \mathbb{R}^d \longrightarrow \mathbb{R}^+$. This inequality asserts that the estimation error is controlled by the excess risk. In our setting, we take $G(t) = \|t\|_2^{\tilde{r}_g}$, $t \in \mathbb{R}^d$. Calibration conditions are widely used in the theoretical analysis of quantile regression (Belloni & Chernozhukov, 2011; Madrid Padilla & Chatterjee, 2022) and classification (Bartlett et al., 2006). For example, in quantile regression one often considers $G(t) = \min\{|t|, |t|^2\}, t \in \mathbb{R}$, while in classification, surrogate losses such as the logistic loss correspond to $G(t) = |t|^2$, and the hinge loss corresponds to $G(t) = |t|$, $t \in \mathbb{R}$.

4.1 Non-asymptotic error bounds

In this section, we present the non-asymptotic error bounds for the excess risk. We define the hypothesis spaces $\mathcal{F}$, $\mathcal{H}$, and $\mathcal{G}$ with the following characteristics:

(a). Let $\mathcal{F} := \mathcal{NN}(\mathcal{W}_f, \mathcal{D}_f, \mathcal{K}_f)$ be the norm-constrained DNNs with width $\mathcal{W}_f \gtrsim d\mathcal{K}_f^{(2p+\beta_f)/(2p+2)}$, depth $\mathcal{D}_f \geq 2\lceil \log_2(p + \bar{\beta}_f)\rceil + 3$, and norm constraint $\mathcal{K}_f \asymp n^{\beta_h(p+1)/(2\beta_f\beta_h+2\beta_f(d+1)+2\beta_h(p+1))}$,

(b). Let $\mathcal{H} := \mathcal{NN}(\mathcal{W}_h, \mathcal{D}_h, \mathcal{K}_h)$ be the norm-constrained DNNs with width $\mathcal{W}_h \gtrsim \mathcal{K}_h^{(2d+\beta_h)/(2d+2)}$, depth $\mathcal{D}_h \geq 2\lceil \log_2(d + \bar{\beta}_h)\rceil + 2$, and norm constraint $\mathcal{K}_h \asymp n^{\beta_f(d+1)/(2\beta_f\beta_h+2\beta_f(d+1)+2\beta_h(p+1))}$

(c). Let $\mathcal{G} := \mathcal{NN}(\mathcal{W}_g, \mathcal{D}_g, \mathcal{S}_g)$ be ReLU DNNs with width $\mathcal{W}_g \geq m^{d/4(2\beta_g+d)} \log(m)$, depth $\mathcal{D}_g \geq m^{d/4(2\beta_g+d)} \log(m)$, and size $\mathcal{S}_g \geq m^{3d/4(2\beta_g+d)} \log^3(m)$.

Theorem 3. *Under Assumptions 4.1–4.2 and network choices (a)–(c), the expected source and target excess risks coincide and satisfy*

$$\mathbb{E}_{D_{\bar{C}}, D_J, \bar{D}_{\mathcal{S}}}\left[\mathcal{E}_{\hat{f}}(\hat{g})\right] \leq \mathcal{O}\left(n^{-\frac{\beta_f\beta_h}{2\beta_f\beta_h+2(p+1)\beta_h+2(d+1)\beta_f}} \log(n)\right) + \mathcal{O}\left(m^{-\frac{\beta_g}{2\beta_g+d}}\right). \tag{6}$$

Moreover, suppose $\beta_f = \beta_h = \beta_g = \beta$ and let $n \geq m^{2(\beta+p+d+2)/(2\beta+d)}$, then

$$\mathbb{E}_{D_{\bar{C}}, D_J, \bar{D}_{\mathcal{S}}}\left[\mathcal{E}_{\hat{f}}(\hat{g})\right] \leq \mathcal{O}\left(m^{-\frac{\beta}{2\beta+d}}\right).$$

Remark 4. Theorem 3 separates the joint cost of learning representations and domain alignment from the labeled-sample cost of downstream prediction. The coupled pretraining rate reflects the combined smoothness and capacity of the feature extractor and discriminator: domain discrimination acts on learned features, so its statistical complexity depends on both components. This quantifies their interaction within pretraining, while the frozen representation enables a separate analysis of predictor fitting.

The resulting bound assigns the ambient-dimensional learning cost to the unlabeled stage and the low-dimensional prediction cost to the labeled stage. In the theorem's large-unlabeled-sample regime with common smoothness $\beta \geq 1$, the bound reduces to $\widetilde{\mathcal{O}}(m^{-\beta/(2\beta+d)})$, where the tilde suppresses logarithmic factors. Thus sufficiently accurate pretraining preserves a downstream label-sample exponent governed by d rather than p, explicitly linking unlabeled representation learning to low-dimensional prediction efficiency.

Remark 5. The analysis explicitly controls neural approximation errors, allowing the ideal representation, discriminator, and predictor to lie outside the chosen network classes. Increasing norm budgets improves the approximation bounds but enlarges the statistical bound; the coupled rate results from balancing these effects jointly for the feature extractor and discriminator. Under the maximum-row-sum product constraint of Definition C.3, the pretraining complexity bound has no explicit width penalty at fixed norm budgets and depths. This provides a quantitative basis for coordinating approximation capacity and statistical complexity in contrastive–adversarial learning.

Discussion. JCAR integrates joint-domain contrastive feature learning with explicit adversarial alignment, then freezes the representation for source-supervised prediction. Whereas Shen et al. (2022) explain transfer through augmentation-graph connectivity for spectral contrastive learning and a linear head, our theory quantifies the estimation and neural-approximation costs of the contrastive–adversarial objective. Its distinctive contribution is to connect the coupled complexity of feature learning and domain discrimination to downstream label efficiency, with separate roles for unlabeled and labeled sample sizes. Detailed comparisons are provided in Appendix A.

5 Numerical Experiments

To verify the effectiveness of our proposed method, we conduct a series of experiments in this section. Section 5.1 describes the datasets, model architectures and baselines. Section 5.2 reports from-scratch comparisons with ERM and established domain-adaptation baselines in Tables 1–2, together with qualitative t-SNE visualizations in Figure 2. Pretrained comparisons are reported separately in Appendix O.

5.1 Experimental setup

We evaluate JCAR on three domain-adaptation benchmarks: Office-31 (Saenko et al., 2010), Office-Home (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019), using ResNet (He et al., 2016) and ViT (Dosovitskiy et al., 2020) backbones. Baselines include DANN (Ganin & Lempitsky, 2015) and CDAN (Long et al., 2018). Appendix N provides details of the data augmentation, model architectures, and baselines.

Table 1: Top-1 accuracy (%) of the proposed method and baselines trained from scratch on Office-31.

Methods	A → W	D → W	W → D	A → D	D → A	W → A
ERM	10.1	28.9	31.3	9.2	6.1	6.0
CDAN	12.1	29.1	28.3	7.8	5.9	4.7
DANN	16.4	63.4	67.5	10.2	7.2	9.6
Ours (ResNet)	**16.7**	44.6	51.4	**12.1**	**10.5**	**13.0**
ViT-B	14.8	37.5	47.8	12.5	5.9	9.9
Ours (ViT)	**17.9**	**39.2**	42.9	11.0	**7.1**	**11.2**

Table 2: Top-1 accuracy (%) of the proposed method and baselines trained from scratch on Office-Home.

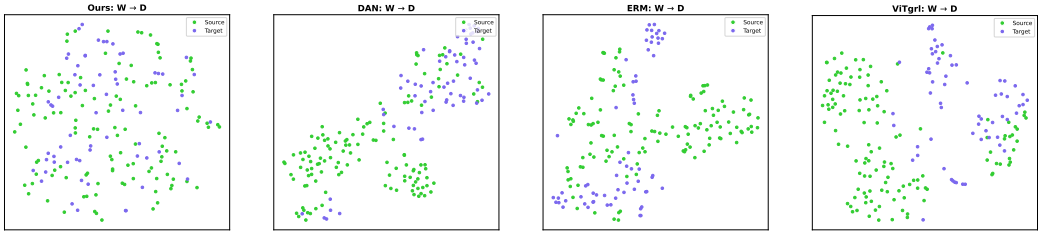
Methods	Ar → Rw	Pr → Rw	Cl → Pr	Ar → Cl	Pr → Ar	Cl → Rw
ERM	10.0	16.4	10.9	4.4	7.1	9.5
CDAN	9.4	16.0	12.5	3.9	5.0	8.4
DANN	9.6	16.4	10.3	4.9	4.4	10.2
Ours (ResNet)	7.6	10.4	10.0	4.3	5.0	7.2
ViT-B	9.2	14.3	3.1	5.8	7.0	9.1
Ours (ViT)	**9.4**	**14.5**	**8.2**	4.5	**7.8**	**10.5**

5.2 Evaluation

Tables 1 and 2 compare randomly initialized models, with ResNet-based and ViT-based methods grouped separately. JCAR delivers clear gains in several challenging from-scratch transfers. On Office-31, our ResNet variant obtains the best result in its architecture group on four of six transfers, including gains of 1.9, 3.3, and 3.4 percentage points over the strongest competing ResNet result on A→D, D→A, and W→A, respectively. Our ViT variant improves over ViT-B on four of six Office-31 transfers. On Office-Home, it improves over ViT-B on five of six transfers, with a 5.1-point gain on Cl→Pr. These results demonstrate the value of joint-domain contrastive-adversarial pretraining, especially for the ViT backbone. The ResNet variant is less effective on Office-Home and on the

D→W and W→D Office-31 transfers, showing that the gains depend on the backbone and domain pair. Tables 5–9 in the Appendix provide the complementary pretrained evaluation.

Figure 2 shows t-SNE (Maaten & Hinton, 2008) projections generated with scikit-learn (Pedregosa et al., 2011); green and purple denote source and target samples. For the displayed Office-31 transfer, JCAR produces substantial source–target overlap together with a broadly distributed feature geometry, providing qualitative evidence that the learned representation combines domain alignment with non-collapsed feature structure. Since t-SNE is a nonlinear projection, we use it as supporting visualization rather than a standalone quantitative measure; additional transfers are shown in Appendix O.



(a) Our method (b) DAN (c) ERM (d) ViT

Figure 2: Two-dimensional t-SNE projections for the Office-31 W→D task. Source and target samples are shown in green and purple, respectively. Compared with DAN (2b), ERM (2c), and ViT (2d), JCAR (2a) exhibits substantial cross-domain overlap while maintaining a broadly distributed feature geometry.

6 Conclusion

We introduced JCAR, a two-stage UDA framework combining joint-domain contrastive learning with adversarial alignment, followed by source-supervised prediction on frozen features. Our non-asymptotic analysis balances neural approximation and statistical errors, captures the coupled cost of feature learning and domain discrimination, and separates it from downstream prediction. Under the stated assumptions, sufficiently accurate unlabeled pretraining yields a downstream polynomial rate governed by the latent dimension rather than the ambient dimension. This establishes a quantitative link between contrastive–adversarial pretraining and label-sample efficiency. The experiments substantiate the practical value of this design, with strong from-scratch improvements across multiple Office-31 and Office-Home transfers, particularly for the ViT backbone. The current results also identify productive directions for extension: domain-specific augmentation design, stronger compatibility with supervised initialization, and component-wise ablations of contrastive learning, adversarial alignment, and stage-wise freezing.

AI use statement

Generative AI tools were used to assist with transcribing this manuscript into the ICLR 2027 LaTeX template, adapting figure and table formatting, and diagnosing compilation issues. They were not used to generate the paper's scientific claims, theoretical results, experimental data, or conclusions. All AI-assisted changes have been reviewed by the authors before submission. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 224–232. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arora17a.html.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Alexandre Belloni and Victor Chernozhukov. l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, pp. 82–130, 2011.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2724–2732, 2018.

Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2985–2994, 2019.

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

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pp. 1081–1090. PMLR, 2019.

Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Muhammad Ghifary, W Bastiaan Kleijn, and Mengjie Zhang. Domain adaptive neural networks for object recognition. In *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings 13*, pp. 898–904. Springer, 2014.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 513–520, 2011.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*, 9(2):473–504, 2020.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. *Advances in neural information processing systems*, 25, 2012.

László Györfi, Michael Kohler, Adam Krzyzak, Harro Walk, et al. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

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

Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649, 2016. URL `http://arxiv.org/abs/1612.02649`.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

Junguang Jiang, Baixu Chen, Bo Fu, and Mingsheng Long. Transfer-learning-library. `https://github.com/thuml/Transfer-Learning-Library`, 2020.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023a.

Yuling Jiao, Yang Wang, and Yunfei Yang. Approximation bounds for norm constrained neural networks with applications to regression and gans. *Applied and Computational Harmonic Analysis*, 65:249–278, 2023b.

Masahiro Kato and Takeshi Teshima. Non-negative bregman divergence minimization for deep direct density ratio estimation. In *International Conference on Machine Learning*, pp. 5320–5333. PMLR, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *International conference on machine learning*, pp. 4013–4022. PMLR, 2019.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and Philip S. Yu. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1076–1089, 2014. doi: 10.1109/TKDE.2013.111.

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

Mingsheng Long, Jianmin Wang, Yue Cao, Jiaguang Sun, and Philip S Yu. Deep learning of transferable representation for scalable domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2027–2040, 2016a.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016b.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pp. 2208–2217. PMLR, 2017.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Oscar Hernan Madrid Padilla and Sabyasachi Chatterjee. Risk bounds for quantile trend filtering. *Biometrika*, 109(3):751–768, 2022.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pp. 3–17. Springer, 2016.

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

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2022.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer vision–ECCV 2010: 11th European conference on computer vision, Heraklion, Crete, Greece, September 5-11, 2010, proceedings, part iV 11*, pp. 213–226. Springer, 2010.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pp. 2263–2291, 2013.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International conference on machine learning*, pp. 19847–19878. PMLR, 2022.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 961–968, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553497. URL https://doi.org/10.1145/1553374.1553497.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul Von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4068–4076, 2015.

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

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Xuezhi Wang and Jeff Schneider. Flexible transfer learning under support and model shift. *Advances in Neural Information Processing Systems*, 27, 2014.

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

Zhiyu Yao, Yunbo Wang, Mingsheng Long, and Jianmin Wang. Unsupervised transfer learning for spatiotemporal predictive networks. In *International Conference on Machine Learning*, pp. 10778–10788. PMLR, 2020.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Yao-Liang Yu and Csaba Szepesvári. Analysis of kernel mean matching under covariate shift. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1147–1154, 2012.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pp. 819–827. Pmlr, 2013.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pp. 7523–7532. PMLR, 2019.

APPENDIX

Implementation and methodological comparisons appear first in Appendix A, followed by broader context and technical definitions in Appendices B and C. Appendices D–L develop the proofs, Appendix M collects auxiliary results, and Appendices N–O supply the experimental details and additional results.

A THE PROPOSED ALGORITHM

Algorithm 1 Domain adaptation with common representation

Require: Unlabeled joint dataset from source and target domains: $D_C = \{X_k\}_{k=1}^n$, the augmentation dataset: $D_{\bar{C}} = \{(X_k, X_k^+, X_k^-)\}_{k=1}^n$, and domain tags $D_Z = \{Z_k\}_{k=1}^n$. Labeled source dataset $\bar{D}_{\mathcal{S}} = \{(X_j^{\mathcal{S}}, Y_j^{\mathcal{S}})\}_{j=1}^m$.

1: **Obtain representation function $\hat{f}$:** Learn the representation function by minimizing the contrastive domain-adaptive loss, i.e.,

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \max_{h \in \mathcal{H}} \hat{\mathcal{R}}_{ad}(f, h).$$

2: Freeze the learned representation $\hat{f}$; its parameters are not updated during downstream predictor training.

3: **Obtain predictor $\hat{g}$ for downstream task:** Learn the predictor through ERM using labeled source-domain data $\bar{D}_{\mathcal{S}}$, i.e.,

$$\hat{g} = \arg\min_{g \in \mathcal{G}} \frac{1}{m} \sum_{j=1}^{m} \ell(g \circ \hat{f}(X_j^{\mathcal{S}}), Y_j^{\mathcal{S}}).$$

4: **return** $\hat{g}$ and the frozen representation $\hat{f}$.

A.1 NOVELTY AND KEY DISTINCTIONS FROM PRIOR FRAMEWORKS

Subsection A.1.1 places the learning objectives side by side. Subsections A.1.2 and A.1.3 then contrast JCAR with conventional UDA and the framework of Shen et al. (2022), respectively.

A.1.1 UNIFIED LOSS FUNCTION FORMULATIONS

Let $G_f(\cdot;\theta_f)$ denote the feature extractor, $G_y(\cdot;\theta_y)$ the downstream predictor, and $G_d(\cdot;\theta_d)$ the domain discriminator; G_f produces D-dimensional representations. Let $\{(x_i^s, y_i^s)\}_{i=1}^n$ be the labeled source data and $\{x_i^t\}_{i=1}^N$ the unlabeled target data. We use $\mathcal{L}_y$, $\mathcal{L}_d$, and $\mathcal{L}_{\text{infoNCE}}$ for the supervised, domain, and contrastive losses, respectively. This subsection presents the multi-negative mini-batch objective used in the experiments. It generalizes the pairwise logistic contrastive mechanism in Definition 2.1, whose one-negative form enables the non-asymptotic analysis developed in Section 4.

- Traditional Adversarial UDA (e.g., DANN) [Single-Stage Coupled Objective]:

$$\mathcal{L}_{\text{Trad}}(\theta_f, \theta_y, \theta_d) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_y \left(G_y(G_f(\mathbf{x}_i^s; \theta_f); \theta_y), y_i \right)}_{\text{Supervised Prediction Loss}} - \lambda \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_d^i(G_f; \theta_d) + \frac{1}{N} \sum_{i=n+1}^{N+n} \mathcal{L}_d^i(G_f; \theta_d) \right)}_{\text{Domain Discrimination Loss}}$$

- Shen et al. (2022) [Two-Stage Decoupled Objective]:
  Stage 1 (Representation Pre-training):

$$\mathcal{L}_{\text{Shen}}(\theta_f) = \frac{1}{n+N} \sum_{i=1}^{n+N} \mathcal{L}_{\text{infoNCE}} \left(G_f(\mathbf{x}_i; \theta_f), G_f(\mathbf{x}_i^+; \theta_f), G_f(\mathbf{x}_i^-; \theta_f) \right)$$

  Stage 2 (Downstream Task Training):

$$\mathcal{L}_{\text{Downstream}}(\theta_y) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_y \left(G_y \left(\hat{G}_f(\mathbf{x}_i^s); \theta_y \right), y_i \right) \quad \text{where } \hat{G}_f \text{ is frozen.}$$

- JCAR [Two-Stage Decoupled Objective with Joint Contrastive Alignment]:
  Stage 1 (Contrastive Domain-Adaptive Pre-training):

$$\mathcal{L}_{\text{JCAR}}(\theta_f, \theta_d) = \underbrace{\frac{1}{n+N} \sum_{i=1}^{n+N} \mathcal{L}_{\text{infoNCE}} \left(G_f(\mathbf{x}_i; \theta_f), G_f(\mathbf{x}_i^+; \theta_f), G_f(\mathbf{x}_i^-; \theta_f) \right)}_{\text{Unsupervised Joint Contrastive Loss}}$$
$$- \lambda \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_d^i(G_f; \theta_d) + \frac{1}{N} \sum_{i=n+1}^{N+n} \mathcal{L}_d^i(G_f; \theta_d) \right)}_{\text{Domain Discrimination Loss}}$$

  Stage 2 (Downstream Task Training):

$$\mathcal{L}_{\text{Downstream}}(\theta_y) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_y \left(G_y \left(\hat{G}_f(\mathbf{x}_i^s); \theta_y \right), y_i \right) \quad \text{where } \hat{G}_f \text{ is frozen.}$$

A.1.2 Distinctions from Traditional UDA Methods

The formulations above highlight two differences between JCAR and standard single-stage adversarial UDA:

(A). Source-Supervised vs. Joint-Domain Contrastive Feature Learning.
In standard DANN-style training, task-discriminative supervision for G_f is supplied by labeled source examples, while unlabeled target examples contribute through the domain objective. Consequently, source labels remain the primary driver of feature construction even though target data participates in alignment.
JCAR instead applies the same self-supervised contrastive objective to the combined unlabeled dataset ($D_S \cup D_T$), so both domains contribute symmetrically and directly to feature construction. The adversarial term complements this instance-level objective by aligning the two feature distributions, yielding a representation shaped jointly by within-sample consistency and cross-domain agreement.

(B). Separating Representation Learning from Downstream Fitting.
Standard adversarial UDA jointly optimizes $\mathcal{L}_y$ and $\mathcal{L}_d$ in one minimax loop. JCAR separates joint-domain representation pretraining from fitting the downstream label predictor, isolating the task-agnostic representation objective from source-label supervision. JCAR completely decouples these two objectives:
Stage 1 (Task-agnostic representation learning):
We learn the representation using $\mathcal{L}_{\text{infoNCE}}$ and $\mathcal{L}_d$ without task labels. The representation is therefore constructed from the geometry of both domains and their domain-level alignment, independently of the source classification objective.
Stage 2 (Flexible task learning):
After pretraining, $\hat{G}_f$ is frozen and the downstream predictor G_y is trained on top of it. This decoupling prevents downstream fitting from overwriting the learned common representation and permits the representation to be reused across downstream tasks.

A.1.3 Distinctions from Shen et al. (2022)

Both JCAR and Shen et al. (2022) use contrastive pretraining followed by downstream classification. JCAR incorporates an explicit adversarial alignment objective and analyzes its joint statistical and approximation costs with feature learning:

(A). Representation Learning and Transfer Mechanisms. Under their augmentation-connectivity model, Shen et al. (2022) show that contrastive representations can support target prediction while retaining domain information. JCAR instead includes domain distinguishability in the pretraining objective, making feature alignment an explicit optimization target alongside contrastive learning.

(B). Mathematical Formulation and Explicit Alignment.

The contrastive objective in Shen et al. (2022) has no explicit adversarial alignment term. JCAR couples instance-level contrastive learning with domain discrimination through a minimax objective; its analysis therefore controls the composite class of discriminators acting on learned features.

(C). Implementation of Joint Pretraining.

At the implementation level, JCAR combines symmetric source–target feature processing with a progressive gradient-reversal schedule:

Symmetric Feature Processing: The same feature extractor and contrastive objective are applied to source and target mini-batch examples, allowing both domains to contribute uniformly to the representation rather than assigning target samples only an alignment role.

Progressive Gradient Reversal (Warm-Start GRL): We use a warm-start schedule that gradually increases the adaptation factor $\lambda(t)$:

$$\lambda(t) = \frac{2 \cdot (\text{hi} - \text{lo})}{1 + e^{-\alpha \cdot t/T}} - (\text{hi} - \text{lo}) + \text{lo}$$

This schedule introduces the adversarial signal gradually while contrastive features are being formed, avoiding an abrupt full-strength alignment objective at the start of training.

(D). Generalization and Theoretical Guarantees. Whereas Shen et al. (2022) derive target-error guarantees for spectral contrastive learning through augmentation-graph connectivity, our analysis explicitly tracks neural-network estimation and approximation errors for a contrastive–adversarial objective. Theorem 3 jointly balances feature-extractor and discriminator complexity, while treating downstream predictor fitting separately. Under its assumptions, common smoothness $\beta \geq 1$, and sufficiently many unlabeled samples, the bound reduces to $\widetilde{\mathcal{O}}(m^{-\beta/(2\beta+d)})$. This quantifies how unlabeled data support high-dimensional pretraining while labeled data govern low-dimensional prediction.

Together, joint-domain contrastive learning, explicit adversarial alignment, and frozen-feature prediction define JCAR's methodological contribution. The non-asymptotic analysis connects this unlabeled representation-learning stage to downstream statistical efficiency.

B Related Work

Distribution shifts between training and test data can degrade predictive performance (Quiñonero-Candela et al., 2022). Domain adaptation addresses such shifts to improve performance on downstream domains. Huang et al. (2006); Sugiyama et al. (2007) used reweighting techniques that estimate density ratios between domains through nonparametric methods such as kernel estimators. These methods, however, can be sensitive to kernel selection and difficult to apply to high-dimensional unstructured data (Sugiyama et al., 2007; 2008; Yu & Szepesvári, 2012; Cao et al., 2019). Furthermore, Maximum mean discrepancy (MMD) (Gretton et al., 2012; Sejdinovic et al., 2013) is widely used to align distributions. Long et al. (2015; 2016b;a) studied feature transferability in deep convolutional networks and used MMD in a reproducing kernel Hilbert space to align deeper-layer representations across unlabeled source and target data. Long et al. (2017; 2018) extended this approach to joint and conditional distribution matching, respectively. Motivated by spectral analyses of feature representations, Chen et al. (2019) penalized dominant singular values to strengthen less dominant directions and improve feature discriminability.

Domain-adversarial methods provide another approach to distribution alignment (Tzeng et al., 2017; Long et al., 2018). Deep domain-adaptation methods embed adaptation modules in neural networks to learn transferable representations while matching feature distributions across domains (Long et al., 2014; Tzeng et al., 2015; Ganin & Lempitsky, 2015; Ganin et al., 2016; Long et al., 2016b; Tzeng et al., 2017). Cao et al. (2018) relaxed the shared-label-space assumption by allowing the target label space to be a subset of the source label space; their method suppresses outlier source classes and aligns distributions within the shared label space. Adversarial alignment can nevertheless impair transfer when suppressing domain-specific variation distorts useful feature structure. Liu et al. (2019) instead generated examples that are adversarial to both the category classifier and the domain discriminator without altering the feature representations. When data distributions embody

complex multimodal structures, adversarial adaptation methods may fail to capture these structures for a discriminative alignment of distributions, potentially leading to model mismatch. This risk cannot be tackled by aligning distributions of features and classes via separate domain discriminators as Hoffman et al. (2016); Chen et al. (2017), because there is no guarantee that two distributions are sufficiently similar (Song et al., 2009; Arora et al., 2017) even when the discriminator is fully confused.

DNNs have achieved significant success, and they are widely believed to be able to learn more transferable features (Zhao et al., 2019) since they disentangle explanatory factors of variations. Ben-David et al. (2010); Pan et al. (2010); Zhang et al. (2013); Wang & Schneider (2014) studied feature representations for transfer between source and target domains. Several recent efforts have focused on learning domain-invariant features in the context of neural networks (Ajakan et al., 2014; Ghifary et al., 2014; Tzeng et al., 2014). Yosinski et al. (2014) showed that learned deep representations capture factors shared across populations and can transfer from the original tasks to related tasks. DNNs can also learn more transferable representations than traditional hand-crafted features (Bengio et al., 2013). DNNs can reduce domain discrepancy (Glorot et al., 2011), but residual dataset shift can increase target-domain risk (Mansour et al., 2009; Ben-David et al., 2010). Yao et al. (2020) noticed that DNNs may suffer from the problem of long-tail data distribution in the target domain. Furthermore, not all knowledge of the pretrained models can be directly applied to the target task due to the discrepancy of various domains.

C Definitions

Definition C.1. *The set $\mathfrak{V}^\beta([0,1]^p_d, B)$ represents the class of Hölder-continuous functions mapping $[0,1]^p$ to $\mathbb{R}^d$, where $p, d \in \mathbb{N}^+$, $\beta > 0$, with $\beta = \bar{\beta} + \tilde{r}$, $\bar{\beta} \in \mathbb{N}_0$, $\tilde{r} \in (0,1]$, and $B > 0$ is a constant. Specifically, this class is defined as*

$$\mathfrak{V}^\beta([0,1]^p_d, B) := \Big\{ f = \{f_1, \ldots, f_d\}; f_i : [0,1]^p \to \mathbb{R}, \\ \max_{\|s\|_1 \le \bar{\beta}} \|\partial^s f_i\|_\infty \le B, \max_{\|s\|_1 = \bar{\beta}} \sup_{x \ne y} \frac{|\partial^s f_i(x) - \partial^s f_i(y)|}{\|x - y\|_2^{\tilde{r}}} \le B,\ i = 1, \ldots, d \Big\},$$

where $s = (s_1, \ldots, s_p) \in \mathbb{N}^p$ and $\|s\|_1 = \sum_{j=1}^p s_j$, $\partial^s f_i(x) = \partial^{s_1} \cdots \partial^{s_p} f_i(x_1, \ldots, x_p)$ for $i = 1, \ldots, d$. For convenience, we use $\mathfrak{V}^\beta([0,1]^p, B)$ to denote the function class for which $d = 1$.

Definition C.2 (ReLU DNNs)**.** *Let $\mathcal{G} = \mathcal{NN}(\mathcal{W}, \mathcal{D}, \mathcal{S}, \mathcal{B})$ be the class of DNNs with parameters $\mathcal{W}, \mathcal{D}, \mathcal{S}, \mathcal{B}$ and ReLU activation function. Any $g \in \mathcal{G}$ can be parameterized as*

$$\begin{aligned} g_0(x) &= x, \\ g_{i+1}(x) &= \varrho(W_i g_i(x) + b_i), \quad i = 0, \ldots, \mathcal{D} - 1, \\ g(x) &= W_{\mathcal{D}} g_{\mathcal{D}}(x) + b_{\mathcal{D}}. \end{aligned}$$

Here, $\varrho(x) := \max\{x, 0\}$ is applied element-wise. For $i = 0, \ldots, \mathcal{D} - 1$, $W_i \in \mathbb{R}^{d_{i+1} \times d_i}$ and $b_i \in \mathbb{R}^{d_{i+1}}$ parameterize hidden layer $i + 1$; $W_{\mathcal{D}}$ and $b_{\mathcal{D}}$ parameterize the final affine output map. The symbol $\mathcal{D}$ denotes the number of hidden layers, $\mathcal{W}$ the maximum hidden-layer width, $\mathcal{S}$ the total number of trainable scalar parameters, and $\mathcal{B} > 0$ a uniform output bound, $\|g\|_\infty \le \mathcal{B}$.

Definition C.3 (Norm-constrained DNNs)**.** *Denote $\mathcal{NN}(\mathcal{W}, \mathcal{D}, \mathcal{K})$ as the norm-constrained DNNs with width $\mathcal{W}$, depth $\mathcal{D}$ and weight constraint $\mathcal{K}$. Specifically, for any $f \in \mathcal{NN}(\mathcal{W}, \mathcal{D}, \mathcal{K})$, the network can be parameterized as*

$$\begin{aligned} f_0 &= x, \\ f_{i+1}(x) &= \varrho(W_i f_i(x) + b_i), \\ f(x) &= W_{\mathcal{D}} f_{\mathcal{D}}(x). \end{aligned}$$

This definition is similar to the DNN class in Definition C.2, but imposes the following norm constraint on the weights of f:

$$\|W_{\mathcal{D}}\| \Pi_{i=0}^{\mathcal{D}-1} \max\{\|(W_i, b_i)\|, 1\} \le \mathcal{K},$$

where $\|W\| := \sup_{\|x\|_\infty \le 1} \|Wx\|_\infty$ and $\mathcal{K} > 0$.

Definition C.4 (Pseudo Dimension). *Let $\mathcal{H}$ be a set of real-valued functions defined on $\mathcal{X}$. The set $A = \{x_i\}_{i=1}^n$ is pseudo-shattered by $\mathcal{H}$ if there exists a fixed threshold vector $w \in \mathbb{R}^n$ such that, for every $b \in \{\pm 1\}^n$, there exists $h \in \mathcal{H}$ satisfying $b_i(h(x_i) - w_i) > 0$ for all $i = 1, \ldots, n$. The pseudo dimension is defined by*

$$Pdim(\mathcal{H}) := \sup\{|A| : A \subseteq \mathcal{X} \text{ is finite and pseudo-shattered by } \mathcal{H}\},$$

where $|\cdot|$ denotes cardinality; the supremum may be infinite.

Definition C.5 (ϵ-covering and Covering number). *Let $(\mathcal{F}, q)$ be a metric space, where the metric $q : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^+$. For any $\epsilon > 0$, we define $\mathcal{A} := \{f_i \in \mathcal{F}; i = 1, \ldots, |\mathcal{A}|\}$ as an ϵ-cover of $\mathcal{F}$, if for each $f \in \mathcal{F}$, there exists a $f_i \in \mathcal{A}$ such that $q(f, f_i) \leq \epsilon$. The ϵ-covering number of $\mathcal{F}$ is the minimal cardinality of an ϵ-cover:*

$$\mathcal{C}(\epsilon, \mathcal{F}, q) = \min\{|\mathcal{A}| : \mathcal{A} \text{ is an } \epsilon\text{-cover of } \mathcal{F}\}.$$

Definition C.6 (Uniform covering number). *Given $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, denote $\mathcal{F}_x = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$ as a subset of $\mathbb{R}^n$. Let $\mathcal{C}(\epsilon, \mathcal{F}_x, \|\cdot\|_\infty)$ be the covering number of $\mathcal{F}_x$ with radius $\epsilon > 0$ and metric $\|\cdot\|_\infty$. Then, the uniform covering number $\mathcal{C}_n(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ is defined as*

$$\mathcal{C}_n(\epsilon, \mathcal{F}, \|\cdot\|_\infty) := \max\{\mathcal{C}(\epsilon, \mathcal{F}_x, \|\cdot\|_\infty) : x \in \mathcal{X}^n\}.$$

D Proof sketch of Theorem 3

To bound $\mathcal{E}_{\hat{f}}(\hat{g})$, we first decompose it into three terms as follows:

$$\mathcal{R}_{\hat{f}}(\hat{g}) - \mathcal{R}_{f^*}(g^*) \leq \underbrace{2\sup_{g\in\mathcal{G}} |\hat{\mathcal{R}}_{\hat{f}}(g) - \mathcal{R}_{\hat{f}}(g)|}_{\text{(I)}} + \underbrace{\inf_{g\in\mathcal{G}} |\mathcal{R}_{\hat{f}}(g) - \mathcal{R}_{\hat{f}}(g^*)|}_{\text{(II)}} + \underbrace{\mathcal{R}_{\hat{f}}(g^*) - \mathcal{R}_{f^*}(g^*)}_{\text{(III)}}.$$

Appendix E derives this decomposition; the following lemmas bound its three terms.

Bound (I). Term (I) is the predictor's statistical error for the learned representation $\hat{f}$. Lemma 6 bounds it by symmetrization and contraction.

Lemma 6. *Suppose that Assumption 4.1 holds. Let $\mathcal{G}$ be ReLU DNNs $\mathcal{NN}(\mathcal{W}_g, \mathcal{D}_g, \mathcal{S}_g)$ bounded by a positive constant $B_\mathcal{G}$. If $m \geq Pdim(\mathcal{G})$, then*

$$\mathbb{E}_{\bar{D}_\mathcal{S}}\left[\sup_{g\in\mathcal{G}} |\hat{\mathcal{R}}_{\hat{f}}(g) - \mathcal{R}_{\hat{f}}(g)|\right] \leq B_\mathcal{G}\sqrt{\frac{\mathcal{S}_g\mathcal{D}_g \log(\mathcal{S}_g)\log(2meB_\mathcal{G})}{m}}.$$

This bound separates the effect of the labeled sample size m from the complexity of the predictor class $\mathcal{G}$ and decreases as m grows.

Bound (II). Term (II) represents the approximation error of g^* when using the DNN class $\mathcal{G}$ as the hypothesis class. More precisely, we have

$$\begin{aligned}
|\mathcal{R}_{\hat{f}}(g) - \mathcal{R}_{\hat{f}}(g^*)| &= |\mathbb{E}_{P_T(X,Y)}[\ell(g \circ \hat{f}(X), Y) - \ell(g^* \circ \hat{f}(X), Y)]| \\
&\leq \mathbb{E}_{P_T(X,Y)}[|\ell(g \circ \hat{f}(X), Y) - \ell(g^* \circ \hat{f}(X), Y)|] \\
&\leq c_\ell \mathbb{E}_{P_T}[|g(\hat{f}(X)) - g^*(\hat{f}(X))|] \\
&\leq c_\ell \|g - g^*\|_\infty,
\end{aligned}$$

where the second inequality follows from the c_ℓ-Lipschitz continuity of $\ell(\cdot, \cdot)$. By taking infimum on both sides of the above inequality Taking the infimum and applying the approximation result of Jiao et al. (2023a), stated in Lemma 10, bounds term (II).

Bound (III). Term (III) transfers the downstream representation error to the excess pretraining risk,

which we then decompose into statistical and approximation terms. By the loss's Lipschitz property and the compact-domain Hölder modulus C_g of g^*, Assumption 4.2 gives

$$\mathcal{R}_{\hat{f}}(g^*) - \mathcal{R}_{f^*}(g^*) \leq \frac{C_g C_{\text{cal}}}{\alpha}\{\mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{M}_h}(f^*)\}.$$

Here α is the source mixture proportion; the constants are fixed independently of the sample sizes. Thus, it suffices to bound the excess pretraining risk $\mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{M}_h}(f^*)$, which decomposes into the following three terms:

$$\begin{aligned}
&\mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{M}_h}(f^*) \\
&\leq \underbrace{2\sup_{f\in\mathcal{F}} |\mathcal{R}_{\mathcal{H}}(f) - \hat{\mathcal{R}}_{\mathcal{H}}(f)|}_{\text{statistical error: } \mathcal{E}_{sta}} + \underbrace{2\sup_{f\in\mathcal{F}} |\mathcal{R}_{\mathcal{H}}(f) - \mathcal{R}_{\mathcal{M}_h}(f)|}_{\text{approximation error: } \mathcal{E}_h} \\
&\quad + \underbrace{\inf_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{M}_h}(f) - \mathcal{R}_{\mathcal{M}_h}(f^*)}_{\text{approximation error: } \mathcal{E}_f}.
\end{aligned}$$

Details regarding the error transformation and decomposition can be found in Appendix I.

For the statistical error $\mathcal{E}_{sta}$, it can be bounded by the sum of the two parts below:

$$\sup_{f\in\mathcal{F}} |\mathcal{R}_{\mathcal{H}}(f) - \hat{\mathcal{R}}_{\mathcal{H}}(f)| \leq \sup_{f\in\mathcal{F}} |\hat{\mathcal{R}}_{cl}(f) - \mathcal{R}_{cl}(f)| + \sup_{f\in\mathcal{F},h\in\mathcal{H}} |\hat{\mathcal{R}}_d(f,h) - \mathcal{R}_d(f,h)|.$$

Appendix L derives this decomposition, and Lemmas 7–8 bound its two terms.

Lemma 7. *For independent sampled triples and the bounded representation class specified in Section 4.1,*

$$\mathbb{E}\sup_{f\in\mathcal{F}} |\mathcal{R}_{cl}(f) - \hat{\mathcal{R}}_{cl}(f)| \lesssim d\mathcal{K}_f \sqrt{\frac{\mathcal{D}_f + 2 + \log(p+1)}{n}}.$$

Lemma 8. *For independent domain-labeled anchors and the norm-constrained classes in Section 4.1,*

$$\mathbb{E}\sup_{f\in\mathcal{F},h\in\mathcal{H}} |\mathcal{R}_d(f,h) - \hat{\mathcal{R}}_d(f,h)| \lesssim \frac{\mathcal{K}_f\mathcal{K}_h d\sqrt{2(\mathcal{D}_f + 2 + \log(p+1))}}{\sqrt{n}}.$$

Lemmas 7 and 8 control the contrastive term over $\mathcal{F}$ and the domain term over the composite class $\mathcal{H}\circ\mathcal{F}$, respectively. Both use the joint unlabeled sample size n. The domain term depends on both network classes, which couples feature-extractor and discriminator complexity in the pretraining bound.

For the approximation errors $\mathcal{E}_h$ and $\mathcal{E}_f$, Appendix L shows that

$$\mathcal{E}_h \lesssim \sup_{h^*\in\mathcal{M}_h} \inf_{h\in\mathcal{H}} \|h - h^*\|_\infty, \qquad \mathcal{E}_f \lesssim \inf_{f\in\mathcal{F}} \|f - f^*\|_\infty.$$

Consequently, Lemma 11, based on the approximation results of Jiao et al. (2023b), bounds these two errors. Therefore, we can now bound the excess risk $\mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{M}_h}(f^*)$.

Lemma 9. *Under Assumptions 4.1 and independent representation sampling, and the choices (a)–(b),*

$$\mathbb{E}\{\mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{M}_h}(f^*)\} \lesssim \frac{d\mathcal{K}_f}{\sqrt{n}} + \mathcal{K}_f\mathcal{K}_h\sqrt{\frac{\mathcal{D}_h + 2 + \log(d+1)}{n}} + \mathcal{K}_f^{-\frac{\beta_f}{p+1}} + \mathcal{K}_h^{-\frac{\beta_h}{d+1}}.$$

This representation-stage result isolates the statistical and approximation costs of contrastive–adversarial pretraining before the downstream calibration step.

Lemma 9 combines the joint statistical complexity of $\mathcal{F}$ and $\mathcal{H}$ with their approximation errors. At fixed network budgets and depths, the statistical terms scale as $n^{-1/2}$; balancing these terms with approximation yields the coupled pretraining rate. Thus the bound quantifies the role of the joint unlabeled sample size without assuming that the ideal functions are exactly realizable by the chosen networks. By combining the results from **Bound (I)**–**Bound (III)**, we obtain the desired result in Theorem 3.

E Error Decompositions

Given the learned representation function $\hat{f}$, the excess risk for any $g \in \mathcal{G}$ can be decomposed as follows:

$$\begin{aligned}
\mathcal{R}_{\hat{f}}(\hat{g}) - \mathcal{R}_{f^*}(g^*) &= \mathcal{R}_{\hat{f}}(\hat{g}) - \hat{\mathcal{R}}_{\hat{f}}(\hat{g}) + \hat{\mathcal{R}}_{\hat{f}}(\hat{g}) - \hat{\mathcal{R}}_{\hat{f}}(g) + \hat{\mathcal{R}}_{\hat{f}}(g) - \mathcal{R}_{\hat{f}}(g) \\
&\quad + \mathcal{R}_{\hat{f}}(g) - \mathcal{R}_{\hat{f}}(g^*) + \mathcal{R}_{\hat{f}}(g^*) - \mathcal{R}_{f^*}(g^*) \\
&\le \mathcal{R}_{\hat{f}}(\hat{g}) - \hat{\mathcal{R}}_{\hat{f}}(\hat{g}) + \hat{\mathcal{R}}_{\hat{f}}(g) - \mathcal{R}_{\hat{f}}(g) + \mathcal{R}_{\hat{f}}(g) - \mathcal{R}_{\hat{f}}(g^*) \\
&\quad + \mathcal{R}_{\hat{f}}(g^*) - \mathcal{R}_{f^*}(g^*) \\
&\le \underbrace{2 \sup_{g \in \mathcal{G}} |\hat{\mathcal{R}}_{\hat{f}}(g) - \mathcal{R}_{\hat{f}}(g)|}_{\text{(I)}} + \underbrace{\inf_{g \in \mathcal{G}} |\mathcal{R}_{\hat{f}}(g) - \mathcal{R}_{\hat{f}}(g^*)|}_{\text{(II)}} + \underbrace{\mathcal{R}_{\hat{f}}(g^*) - \mathcal{R}_{f^*}(g^*)}_{\text{(III)}}.
\end{aligned}$$

Here, the first inequality follows from the definition of $\hat{g}$ in (4).

F Proof of Lemma 1

Proof. Let f be the common representation function. Write $U^{\mathcal{S}} := f(X^{\mathcal{S}})$ and $U^{\mathcal{T}} := f(X^{\mathcal{T}})$ for the source and target representations. Let $P_S(U^{\mathcal{S}}, Y^{\mathcal{S}})$ and $P_T(U^{\mathcal{T}}, Y^{\mathcal{T}})$ denote their joint distributions, with marginals $P_S(U^{\mathcal{S}})$ and $P_T(U^{\mathcal{T}})$. Then, we have

$$\begin{aligned}
\mathbb{E}_{P_T(U^{\mathcal{T}}, Y^{\mathcal{T}})}[\ell(g(U^{\mathcal{T}}), Y^{\mathcal{T}})] &= \int \ell(g(U^{\mathcal{T}}), Y^{\mathcal{T}}) P_T(U^{\mathcal{T}}, Y^{\mathcal{T}}) dU^{\mathcal{T}} dY^{\mathcal{T}} \\
&= \int \ell(g(U^{\mathcal{T}}), Y^{\mathcal{T}}) P_T(Y^{\mathcal{T}} | U^{\mathcal{T}}) P_T(U^{\mathcal{T}}) dU^{\mathcal{T}} dY^{\mathcal{T}} \\
&= \int \ell(g(U^{\mathcal{S}}), Y^{\mathcal{S}}) P_S(Y^{\mathcal{S}} \mid U^{\mathcal{S}}) P_S(U^{\mathcal{S}})\, dU^{\mathcal{S}} dY^{\mathcal{S}} \\
&= \mathbb{E}_{P_S(U^{\mathcal{S}}, Y^{\mathcal{S}})}[\ell(g(U^{\mathcal{S}}), Y^{\mathcal{S}})],
\end{aligned}$$

where the third equality follows from $P_S(U^{\mathcal{S}}) = P_T(U^{\mathcal{T}})$ and $P_S(Y^{\mathcal{S}}|U^{\mathcal{S}}) = P_T(Y^{\mathcal{T}}|U^{\mathcal{T}})$. Then, we have $g^{**} = g^*$. □

G Proof of Theorem 3

Proof. We set the network architectures of $\mathcal{F}, \mathcal{H}, \mathcal{G}$ as follows:

(a). Let $\mathcal{F} := \mathcal{NN}(\mathcal{W}_f, \mathcal{D}_f, \mathcal{K}_f)$ be the norm-constrained DNNs with width $\mathcal{W}_f \gtrsim d\mathcal{K}_f^{(2p+\beta_f)/(2p+2)}$, depth $\mathcal{D}_f \ge 2\lceil \log_2(p + \bar{\beta}_f) \rceil + 3$, and norm constraint $\mathcal{K}_f \asymp n^{\beta_h(p+1)/(2\beta_f\beta_h + 2\beta_f(d+1) + 2\beta_h(p+1))}$,

(b). Let $\mathcal{H} := \mathcal{NN}(\mathcal{W}_h, \mathcal{D}_h, \mathcal{K}_h)$ be the norm-constrained DNNs with width $\mathcal{W}_h \gtrsim \mathcal{K}_h^{(2d+\beta_h)/(2d+2)}$, depth $\mathcal{D}_h \ge 2\lceil \log_2(d + \bar{\beta}_h) \rceil + 2$, and norm constraint $\mathcal{K}_h \asymp n^{\beta_f(d+1)/(2\beta_f\beta_h + 2\beta_f(d+1) + 2\beta_h(p+1))}$

(c). Let $\mathcal{G} := \mathcal{NN}(\mathcal{W}_g, \mathcal{D}_g, \mathcal{S}_g)$ be ReLU DNNs with width $\mathcal{W}_g \ge m^{d/4(2\beta_g+d)} \log(m)$, depth $\mathcal{D}_g \ge m^{d/4(2\beta_g+d)} \log(m)$, and size $\mathcal{S}_g \ge m^{3d/4(2\beta_g+d)} \log^3(m)$.

By combining Lemmas 6, 9, and 10, we have

$$\begin{aligned}
\mathbb{E}\mathcal{E}_{\hat{f}}(\hat{g}) \lesssim B_{\mathcal{G}} \sqrt{\frac{\mathcal{S}_g \mathcal{D}_g \log(\mathcal{S}_g) \log(2meB_{\mathcal{G}})}{m}} + 19 B_g (\lfloor \beta_g \rfloor + 1)^2 d^{\lfloor \beta_g \rfloor + (\beta_g \vee 1)/2} (MN)^{-2\beta_g/d} \\
+ \frac{d\mathcal{K}_f}{\sqrt{n}} + \mathcal{K}_f \mathcal{K}_h \sqrt{\frac{\mathcal{D}_h + 2 + \log(d+1)}{n}} + \mathcal{K}_f^{-\frac{\beta_f}{p+1}} + \mathcal{K}_h^{-\frac{\beta_h}{d+1}},
\end{aligned}$$

we then obtain the results in Theorem 3 under the (a)–(c) network architectures. □

H PROOF OF LEMMA 6.

Proof. Let $\bar{D}_S := \{(X_i^{\mathcal{S}}, Y_i^{\mathcal{S}})\}_{i=1}^m$ and $\bar{D}_T := \{(X_j^{\mathcal{T}}, Y_j^{\mathcal{T}})\}_{j=1}^m$ be the labeled datasets from the source and target domains, respectively. Let $\bar{D}_{S'} := \{(X_i^{\mathcal{S}'}, Y_i^{\mathcal{S}'})\}_{i=1}^m$ and $\bar{D}_{T'} := \{(X_j^{\mathcal{T}'}, Y_j^{\mathcal{T}'})\}_{j=1}^m$ be independent ghost samples of $\bar{D}_S$ and $\bar{D}_T$. Let $\epsilon := \{\epsilon_j\}_{j=1}^m$ be independent Rademacher variables, each taking values in $\{-1, 1\}$ with equal probability. Taking the expectation of $\sup_{g\in\mathcal{G}}\{\mathcal{R}_{\hat{f}}(g) - \hat{\mathcal{R}}_{\hat{f}}(g)\}$ with respect to $\bar{D}_T$ yields we have

$$
\begin{aligned}
&\mathbb{E}_{\bar{D}_T}\left[\sup_{g\in\mathcal{G}} \mathcal{R}_{\hat{f}}(g) - \hat{\mathcal{R}}_{\hat{f}}(g)\right] \\
&= \mathbb{E}_{\bar{D}_T}\left[\sup_{g\in\mathcal{G}} \left(\mathbb{E}_{P_T}[\ell(g\circ\hat{f}(X^{\mathcal{T}}), Y^{\mathcal{T}})] - \frac{1}{m}\sum_{j=1}^m \ell(g\circ\hat{f}(X_j^{\mathcal{T}}), Y_j^{\mathcal{T}})\right)\right] \\
&= \mathbb{E}_{\bar{D}_T}\left[\sup_{g\in\mathcal{G}} \left(\mathbb{E}_{\bar{D}_{T'}}[\frac{1}{m}\sum_{j=1}^m \ell(g\circ\hat{f}(X_j^{\mathcal{T}'}), Y_j^{\mathcal{T}'})] - \frac{1}{m}\sum_{j=1}^m \ell(g\circ\hat{f}(X_j^{\mathcal{T}}), Y_j^{\mathcal{T}})\right)\right] \\
&= \mathbb{E}_{\bar{D}_T}\left[\sup_{g\in\mathcal{G}} \mathbb{E}_{\bar{D}_{T'}}[\frac{1}{m}\sum_{j=1}^m \ell(g\circ\hat{f}(X_j^{\mathcal{T}'}), Y_j^{\mathcal{T}'}) - \frac{1}{m}\sum_{j=1}^m \ell(g\circ\hat{f}(X_j^{\mathcal{T}}), Y_j^{\mathcal{T}})]\,\middle|\,\bar{D}_T\right] \\
&\le \mathbb{E}_{\bar{D}_T,\bar{D}_{T'}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{j=1}^m \left(\ell(g\circ\hat{f}(X_j^{\mathcal{T}'}), Y_j^{\mathcal{T}'}) - \ell(g\circ\hat{f}(X_j^{\mathcal{T}}), Y_j^{\mathcal{T}})\right)\,\middle|\,\bar{D}_T\right] \quad \text{(Jensen's inequality)} \\
&= 2\mathbb{E}_{\epsilon,\bar{D}_T}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{j=1}^m \epsilon_j \ell(g\circ\hat{f}(X_j^{\mathcal{T}}), Y_j^{\mathcal{T}})\right] \quad \text{(symmetrization)} \\
&\le 2c_\ell\mathbb{E}_{\epsilon,\bar{D}_T}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{j=1}^m \epsilon_j g\circ\hat{f}(X_j^{\mathcal{T}})\right] \\
&:= 2c_\ell\mathfrak{R}_{\bar{D}_T}(\mathcal{G}\circ\hat{f}).
\end{aligned}
$$

The last inequality follows from Talagrand inequality Ledoux & Talagrand (2013, Theorem 4.12), since ℓ is 1-Lipschitz. Let $(U^{\mathcal{T}}, Y^{\mathcal{T}}) = (\hat{f}(X^{\mathcal{T}}), Y^{\mathcal{T}})$ and $\bar{D}_{U^{\mathcal{T}}} = \{(U_j^{\mathcal{T}}, Y_j^{\mathcal{T}})\}_{j=1}^m$ with distribution $P_{\bar{D}_{U^{\mathcal{T}}}}(U^{\mathcal{T}}, Y^{\mathcal{T}})$. Since $\hat{f}$ is fixed,

$$
\mathfrak{R}_{\bar{D}_T}(\mathcal{G}\circ\hat{f}) = \mathbb{E}_{\epsilon,\bar{D}_T}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{j=1}^m \epsilon_j g\circ\hat{f}(X_j)\right] \le \mathbb{E}_{\epsilon,\bar{D}_{U^{\mathcal{T}}}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{j=1}^m \epsilon_j g(U_j)\right] := \mathfrak{R}_{\bar{D}_{U^{\mathcal{T}}}}(\mathcal{G}).
$$

Because the dataset $\bar{D}_T$ is unavailable, we cannot calculate the risk $\mathfrak{R}_{\bar{D}_{U^{\mathcal{T}}}}(\mathcal{G})$ in practice. However, by Lemma 1 and its proof, $P_{\bar{D}_T}(U^{\mathcal{T}}, Y^{\mathcal{T}}) = P_{\bar{D}_S}(U^{\mathcal{S}}, Y^{\mathcal{S}})$, and consequently $\mathfrak{R}_{\bar{D}_{U^{\mathcal{T}}}}(\mathcal{G}) = \mathfrak{R}_{\bar{D}_{U^{\mathcal{S}}}}(\mathcal{G})$.

Let $\mathcal{C}_m(\delta, \mathcal{G}, \|\cdot\|_\infty)$ denote the uniform covering number of $\mathcal{G}$, where $\|g\|_\infty := \max\{|g(U_j)| : j = 1, \ldots, m\}$. Since $\|g\|_\infty \le B_{\mathcal{G}}$ for $B_{\mathcal{G}} > 0$, Dudley's entropy integral (Anthony & Bartlett, 2009) gives

$$
\mathfrak{R}_{\bar{D}_{U^{\mathcal{S}}}}(\mathcal{G}) \le \mathbb{E}\inf_{0<\delta<B_{\mathcal{G}}}\left(4\delta + \frac{12}{\sqrt{m}}\int_\delta^{B_{\mathcal{G}}} \sqrt{\log(2\mathcal{C}_m(t, \mathcal{G}, \|\cdot\|_\infty))}\,dt\right). \tag{7}
$$

Let $Pdim(\mathcal{G})$ denote the pseudo dimension of $\mathcal{G}$. For $1 \le Pdim(\mathcal{G}) < m$ and $0 < \delta \le B_{\mathcal{G}}$, applying Anthony & Bartlett (2009, Theorem 12.2) to the shifted class $\{g + B_{\mathcal{G}} : g \in \mathcal{G}\}$, whose range lies in $[0, 2B_{\mathcal{G}}]$, gives

$$
\mathcal{C}_m(\delta, \mathcal{G}, \|\cdot\|_\infty) \le \left(\frac{2eB_{\mathcal{G}}m}{\delta Pdim(\mathcal{G})}\right)^{Pdim(\mathcal{G})}. \tag{8}
$$

Moreover, the ReLU network pseudodimension bound of Bartlett et al. (2019) gives, for $\mathcal{S}_g \ge 2$ and $\mathcal{D}_g \ge 1$, a universal constant $C > 0$ such that

$$
Pdim(\mathcal{G}) \le C \cdot \mathcal{S}_g\mathcal{D}_g \log(\mathcal{S}_g).
$$

Combining these bounds proves Lemma 6:

$$\mathfrak{R}_{\bar{D}_{U^{\mathcal{S}}}}(\mathcal{G}) \leq B_{\mathcal{G}} \sqrt{\frac{\mathcal{S}_g \mathcal{D}_g \log(\mathcal{S}_g) \log(2meB_{\mathcal{G}})}{m}}.$$

□

I ERROR TRANSFORM AND DECOMPOSITION

Proof. First, we bound $\mathcal{R}_{\hat{f}}(g^*) - \mathcal{R}_{f^*}(g^*)$ by $\mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{M}_h}(f^*)$:

$$\begin{aligned}
\mathcal{R}_{\hat{f}}(g^*) - \mathcal{R}_{f^*}(g^*) &= \mathbb{E}_{P_S}[\ell(g^* \circ \hat{f}(X), Y) - \ell(g^* \circ f^*(X), Y)] \\
&\leq c_\ell \mathbb{E}_{P_S}[|g^* \circ \hat{f}(X) - g^* \circ f^*(X)|] \\
&\leq c_\ell C_g \mathbb{E}_{P_S}[\|\hat{f}(X) - f^*(X)\|_2^{\tilde{r}_g}] \\
&\leq \frac{C_g}{\alpha} c_\ell \mathbb{E}_{P_C(X)}[\|\hat{f}(X) - f^*(X)\|_2^{\tilde{r}_g}] \\
&\leq \frac{C_g C_{\text{cal}}}{\alpha} c_\ell \big(\mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{M}_h}(f^*)\big),
\end{aligned}$$

The first inequality uses the c_ℓ-Lipschitz continuity of ℓ, the second uses the Hölder modulus of g^* with constant C_g, the third follows from $P_C(X) = \alpha P_S(X) + (1-\alpha) P_T(X) \geq \alpha P_S(X)$, and the last follows from Assumption 4.2. We therefore obtain the following decomposition:

$$\begin{aligned}
&\mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{M}_h}(f^*) \\
&= \mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{H}}(\hat{f}) + \mathcal{R}_{\mathcal{H}}(\hat{f}) - \hat{\mathcal{R}}_{\mathcal{H}}(\hat{f}) + \hat{\mathcal{R}}_{\mathcal{H}}(\hat{f}) - \hat{\mathcal{R}}_{\mathcal{H}}(f) \\
&+ \hat{\mathcal{R}}_{\mathcal{H}}(f) - \mathcal{R}_{\mathcal{H}}(f) + \mathcal{R}_{\mathcal{H}}(f) - \mathcal{R}_{\mathcal{M}_h}(f) + \mathcal{R}_{\mathcal{M}_h}(f) - \mathcal{R}_{\mathcal{M}_h}(f^*) \\
&\leq \mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{H}}(\hat{f}) + \mathcal{R}_{\mathcal{H}}(\hat{f}) - \hat{\mathcal{R}}_{\mathcal{H}}(\hat{f}) \\
&+ \hat{\mathcal{R}}_{\mathcal{H}}(f) - \mathcal{R}_{\mathcal{H}}(f) + \mathcal{R}_{\mathcal{H}}(f) - \mathcal{R}_{\mathcal{M}_h}(f) + \mathcal{R}_{\mathcal{M}_h}(f) - \mathcal{R}_{\mathcal{M}_h}(f^*) \\
&\leq \underbrace{2 \sup_{f \in \mathcal{F}} |\mathcal{R}_{\mathcal{H}}(f) - \hat{\mathcal{R}}_{\mathcal{H}}(f)|}_{\text{statistical error:}\mathcal{E}_{sta}} + \underbrace{2 \sup_{f \in \mathcal{F}} |\mathcal{R}_{\mathcal{H}}(f) - \mathcal{R}_{\mathcal{M}_h}(f)|}_{\text{approximation error:}\mathcal{E}_h} \\
&\quad + \underbrace{\inf_{f \in \mathcal{F}} \mathcal{R}_{\mathcal{M}_h}(f) - \mathcal{R}_{\mathcal{M}_h}(f^*)}_{\text{approximation error:}\mathcal{E}_f}.
\end{aligned}$$

□

J PROOF OF LEMMA 7

Proof. Considering the expectation of $\sup_{f\in\mathcal{F}}(\mathcal{R}_{cl}(f)-\hat{\mathcal{R}}_{cl}(f))$, we obtain

$$\begin{aligned}
&\mathbb{E}_{D_{\bar{C}}}[\sup_{f\in\mathcal{F}}(\mathcal{R}_{cl}(f)-\hat{\mathcal{R}}_{cl}(f))] = \mathbb{E}_{D_{\bar{C}}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{P_C}\left[\ell_{cl}(X,f)\right]-\frac{1}{n}\sum_{i=1}^{n}\ell_{cl}(X_i,f)\right)\right]\\
&=\mathbb{E}_{D_{\bar{C}}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{D_{\bar{C}'}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell_{cl}(X_i',f)\right]-\frac{1}{n}\sum_{i=1}^{n}\ell_{cl}(X_i,f)\right)\right]\\
&=\mathbb{E}_{D_{\bar{C}}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{D_{\bar{C}'}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell_{cl}(X_i',f)-\frac{1}{n}\sum_{i=1}^{n}\ell_{cl}(X_i,f)\right)\right]\middle| D_{\bar{C}}\right]\\
&=\mathbb{E}_{D_{\bar{C}}}\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{D_{\bar{C}'}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\ell_{cl}(X_i',f)-\ell_{cl}(X_i,f)\right)\right]\middle| D_{\bar{C}}\right]\\
&\le\mathbb{E}_{D_{\bar{C}},D_{\bar{C}'}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\left(\ell_{cl}(X_i',f)-\ell_{cl}(X_i,f)\right)\middle| D_{\bar{C}}\right]\quad\text{(Jensen's inequality)}\\
&=\mathbb{E}_{\epsilon,D_{\bar{C}},D_{\bar{C}'}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(\ell_{cl}(X_i',f)-\ell_{cl}(X_i,f)\right)\middle| D_{\bar{C}}\right]\quad\text{(Symmetrization)}\\
&=2\mathbb{E}_{\epsilon,D_{\bar{C}}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell_{cl}(X_i,f)\right]\\
&=2\mathfrak{R}_{\bar{C}}(\ell_{cl}\circ\mathcal{F}),
\end{aligned}$$

where $D_{\bar{C}}=\{(X_i,X_i^+,X_i^-)\}_{i=1}^n$, $D_{\bar{C}'}:=\{(X_i',X_i'^+,X_i'^-)\}_{i=1}^n$ is the ghost sample of $D_{\bar{C}}$, and ϵ_i are independent Rademacher variables, each taking values in $\{-1,1\}$ with equal probability. Since $f:\mathbb{R}^p\to\mathbb{R}^d$ and ℓ_{cl} is Lipschitz continuous,

$$\mathfrak{R}_{\bar{C}}(\ell_{cl}\circ\mathcal{F})\le\ell_{cl}\mathfrak{R}_{\bar{C}}(\mathcal{F})\le d\mathcal{K}_f\sqrt{\frac{\mathcal{D}_f+2+\log(p+1)}{n}}.$$

The first inequality uses the Lipschitz continuity of ℓ_{cl}, and the second follows from Lemma 13.

□

K PROOF OF LEMMA 8

Proof. We consider the expectation of $\sup_{f\in\mathcal{F},h\in\mathcal{H}}\left(\mathcal{R}_d(f,h)-\hat{\mathcal{R}}_d(f,h)\right)$, i.e.,

$$\begin{aligned}
&\mathbb{E}_{D_J}[\sup_{f\in\mathcal{F},h\in\mathcal{H}}(\mathcal{R}_d(f,h)-\hat{\mathcal{R}}_d(f,h))]\\
&=\mathbb{E}_{D_J}\left[\sup_{f\in\mathcal{F},h\in\mathcal{H}}\left(\mathbb{E}_{P_\mu}[\ell_d(X,Z,f,h)]-\frac{1}{n}\sum_{k=1}^{n}\ell_d(X_k,Z_k,f,h)\right)\right]\\
&=\mathbb{E}_{D_J}\left[\sup_{f\in\mathcal{F},h\in\mathcal{H}}\left(\mathbb{E}_{D_{J'}}[\frac{1}{n}\sum_{k=1}^{n}\ell_d(X_k',Z_k',f,h)]-\frac{1}{n}\sum_{k=1}^{n}\ell_d(X_k,Z_k,f,h)\right)\middle| D_J\right]\\
&\le\mathbb{E}_{D_J,D_{J'}}\left[\sup_{f\in\mathcal{F},h\in\mathcal{H}}\frac{1}{n}\sum_{k=1}^{n}\left(\ell_d(X_k',Z_k',f,h)-\ell_d(X_k,Z_k,f,h)\right)\middle| D_J\right]\quad\text{(Jensen's inequality)}\\
&=\mathbb{E}_{\epsilon,D_J,D_{J'}}\left[\sup_{f\in\mathcal{F},h\in\mathcal{H}}\frac{1}{n}\sum_{k=1}^{n}\epsilon_k\left(\ell_d(X_k',Z_k',f,h)-\ell_d(X_k,Z_k,f,h)\right)\middle| D_J\right]\quad\text{(Symmetrization)}\\
&=2\mathbb{E}_{\epsilon,D_J}\left[\sup_{f\in\mathcal{F},h\in\mathcal{H}}\frac{1}{n}\sum_{k=1}^{n}\epsilon_k\ell_d(X_k,Z_k,f,h)\right]\\
&=2\mathfrak{R}_C(\ell_d\circ(\mathcal{H}\circ\mathcal{F}))\\
&\le 2\mathfrak{R}_C(\mathcal{H}\circ\mathcal{F}).
\end{aligned}$$

The last inequality follows from Talagrand inequality (Ledoux & Talagrand, 2013, Theorem 4.12), since ℓ_d is 1-Lipschitz.

Let $\mathcal{C}(\delta, \mathcal{H} \circ \mathcal{F}, \|\cdot\|_2)$ denote the covering number of $\mathcal{H} \circ \mathcal{F}$ with respect to the empirical norm $\|h \circ f\|_2 := \sqrt{n^{-1}\sum_{k=1}^{n}(h \circ f(X_k))^2}$. Since $\|h\|_2 \le B_{\mathcal{H}}$, the chaining argument of Shalev-Shwartz & Ben-David (2014, Lemma 27.4) and Dudley's entropy integral (Dudley, 1967) give

$$\mathfrak{R}_C(\mathcal{H} \circ \mathcal{F}) \le \mathbb{E} \inf_{0 \le \delta \le B_{\mathcal{H}}} \left(4\delta + \frac{12}{\sqrt{n}} \int_{\delta}^{B_{\mathcal{H}}} \sqrt{\log(2\mathcal{C}(t, \mathcal{H} \circ \mathcal{F}, \|\cdot\|_2))}\, dt \right). \tag{9}$$

We bound the entropy integral in (9) by constructing a cover of $\mathcal{H} \circ \mathcal{F}$ from covers of $\mathcal{H}$ and $\mathcal{F}$. Recall that $D_C = \{X_i\}_{i=1}^n$. Let $\mathcal{C}_{\mathcal{F}}(\delta_1, \mathcal{F}, \|\cdot\|_2)$ be a δ_1-cover of $\mathcal{F}$ on D_C. For each $f \in \mathcal{C}_{\mathcal{F}}$, let $\mathcal{C}_{\mathcal{H}_f}(\delta_2, \mathcal{H}, \|\cdot\|_2)$ be a δ_2-cover of $\mathcal{H}$ on the feature inputs generated by f. Then $\mathcal{C}_{\mathcal{H}\circ\mathcal{F}} := \bigcup_{f \in \mathcal{C}_{\mathcal{F}}} \mathcal{C}_{\mathcal{H}_f}$ is a $(L(\mathcal{H})\delta_1 + \delta_2)$-cover of $\mathcal{H} \circ \mathcal{F}$ on D_C. Indeed, for arbitrary $f \in \mathcal{F}$ and $h \in \mathcal{H}$, choose $f' \in \mathcal{C}_{\mathcal{F}}$ within δ_1 of f and $h' \in \mathcal{C}_{\mathcal{H}_{f'}}$ within δ_2 of h on the inputs $\{f'(X_i) : X_i \in D_C\}$. Writing $\mathbf{X} = (X_1, \ldots, X_n)$, the triangle inequality gives

$$\begin{aligned} \|h \circ f(\mathbf{X}) - h' \circ f'(\mathbf{X})\|_2 &\le \|h \circ f(\mathbf{X}) - h \circ f'(\mathbf{X})\|_2 + \|h \circ f'(\mathbf{X}) - h' \circ f'(\mathbf{X})\|_2 \\ &\le L(\mathcal{H}) \cdot \delta_1 + \delta_2, \end{aligned}$$

The second inequality uses the uniform Lipschitz property of $\mathcal{H}$. Let $\mathbf{U} := \{U_i = f(X_i) : i = 1, \ldots, n\}$. Then $|\mathcal{C}_{\mathcal{H}\circ\mathcal{F}}| \le |\mathcal{C}_{\mathcal{F}}| \max_{f \in \mathcal{C}_{\mathcal{F}}} |\mathcal{C}_{\mathcal{H}_f}| \le |\mathcal{C}_{\mathcal{F}}| \mathcal{C}_n(\delta_2, \mathcal{H}, \|\cdot\|_2)$. Using this covering-number bound with $\delta_1 = \delta/(2L(\mathcal{H}))$ and $\delta_2 = \delta/2$, together with the subadditivity of the square root, gives

$$\begin{aligned} \int_{\delta}^{B_{\mathcal{H}}} \sqrt{\log \mathcal{C}(t, \mathcal{H} \circ \mathcal{F}, \|\cdot\|_2)}\, dt \le &\int_{\delta}^{B_{\mathcal{H}}} \sqrt{\log \mathcal{C}(t/2L(\mathcal{H}), \mathcal{F}, \|\cdot\|_2)}\, dt \\ &+ \int_{\delta}^{B_{\mathcal{H}}} \sqrt{\log \mathcal{C}_n(t/2, \mathcal{H}, \|\cdot\|_2)}\, dt. \end{aligned}$$

The Sudakov minoration theorem for Gaussian processes (Wainwright, 2019, Theorem 5.30), together with the relation between packing and covering numbers at the same scale, gives

$$\log \mathcal{C}_n(\delta, \mathcal{H}, \|\cdot\|_2) \le 4\left(\frac{\sqrt{n}\mathfrak{G}_{\mathbf{U}}(\mathcal{H})}{\delta} \right)^2, \quad \forall \delta > 0,$$

where $\mathfrak{G}_{\mathbf{U}}(\mathcal{H})$ is the Gaussian process indexed by $\mathcal{H}$ on $\mathbf{U}$, and $h(U) \in [-B_{\mathcal{H}}, B_{\mathcal{H}}]$ for $U \in \mathbf{U}$. Consequently,

$$\begin{aligned} \mathcal{R}_C(\mathcal{H} \circ \mathcal{F}) \le \frac{12}{\sqrt{n}} \mathbb{E} \inf_{0 \le \delta \le B_{\mathcal{H}}} \Bigg(4\delta + &\int_{\delta}^{B_{\mathcal{H}}} \sqrt{\log \mathcal{C}(t/2L(\mathcal{H}), \mathcal{F}, \|\cdot\|_2)}\, dt \\ &+ \int_{\delta}^{B_{\mathcal{H}}} \sqrt{\log \mathcal{C}_n(t/2, \mathcal{H}, \|\cdot\|_2)}\, dt \Bigg). \end{aligned}$$

By the relation between the covering number and Gaussian process, we have

$$\log \mathcal{C}_n(t/2, \mathcal{H}, \|\cdot\|_2) \le 4\left(\frac{2\sqrt{n}\mathfrak{G}_{\mathbf{U}}(\mathcal{H})}{t} \right)^2 \le 4\left(\frac{4\ln(n)\sqrt{n}\mathfrak{R}_{\mathbf{U}}(\mathcal{H})}{t} \right)^2, \tag{10}$$

where the second inequality holds by Bartlett & Mendelson (2002, Lemma 4), and $\mathfrak{R}_X(\mathcal{H})$ is the empirical Rademacher complexity. Similarly, we have

$$\log \mathcal{C}(t/(2L(\mathcal{H})), \mathcal{F}, \|\cdot\|_2) \le 4\left(\frac{2L(\mathcal{H})\sqrt{n}\mathfrak{G}_{\mathbf{X}}(\mathcal{F})}{t} \right)^2 \le 4\left(\frac{4L(\mathcal{H})\ln(n)\sqrt{n}\mathfrak{R}_{\mathbf{X}}(\mathcal{F})}{t} \right)^2. \tag{11}$$

Combining Lemmas 13-14 with (9), (10), and (11), we obtain

$$\mathcal{R}_C(\mathcal{H} \circ \mathcal{F}) \le \frac{\mathcal{K}_f \mathcal{K}_h d\sqrt{2(\mathcal{D}_f + 2 + \log(p+1))}}{\sqrt{n}} + \frac{\mathcal{K}_h \sqrt{2(\mathcal{D}_h + 2 + \log(d+1))}}{\sqrt{n}}.$$

□

L Proof of Lemma 9

Proof. We bound the three terms in the decomposition of $\mathcal{R}_{\mathcal{M}_h}(\hat{f}) - \mathcal{R}_{\mathcal{M}_h}(f^*)$ from Appendix I.
For the statistical error $\mathcal{E}_{sta}$, we have

$$\begin{aligned}
\sup_{f\in\mathcal{F}} |\mathcal{R}_{\mathcal{H}}(f) - \hat{\mathcal{R}}_{\mathcal{H}}(f)| &= \sup_{f\in\mathcal{F}} |\sup_{h\in\mathcal{H}} \hat{\mathcal{R}}_{ad}(f,h) - \sup_{h\in\mathcal{H}} \mathcal{R}_{ad}(f,h)| \\
&\le \sup_{f\in\mathcal{F},h\in\mathcal{H}} |\hat{\mathcal{R}}_{ad}(f,h) - \mathcal{R}_{ad}(f,h)| \\
&= \sup_{f\in\mathcal{F},h\in\mathcal{H}} |\hat{\mathcal{R}}_{cl}(f) - \mathcal{R}_{cl}(f) + \hat{\mathcal{R}}_{d}(f,h) - \mathcal{R}_{d}(f,h)| \\
&\le \sup_{f\in\mathcal{F}} |\hat{\mathcal{R}}_{cl}(f) - \mathcal{R}_{cl}(f)| + \sup_{f\in\mathcal{F},h\in\mathcal{H}} |\hat{\mathcal{R}}_{d}(f,h) - \mathcal{R}_{d}(f,h)|.
\end{aligned}$$

Using Lemmas 7 and 8, we bound the statistical error as follows:

$$\mathbb{E}\sup_{f\in\mathcal{F}} |\mathcal{R}_{\mathcal{H}}(f) - \hat{\mathcal{R}}_{\mathcal{H}}(f)| \lesssim d\mathcal{K}_f\sqrt{\frac{\mathcal{D}_f + 2 + \log(p+1)}{n}} + \mathcal{K}_f\mathcal{K}_h\sqrt{\frac{\mathcal{D}_f + \mathcal{D}_h + 2 + \log(p+1)}{n}}.$$

For the approximation error $\mathcal{E}_h$, we have

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} |\mathcal{R}_{\mathcal{H}}(f) - \mathcal{R}_{\mathcal{M}_h}(f)| \\
&= \sup_{f\in\mathcal{F}} |\sup_{h_2\in\mathcal{M}_h} \mathcal{R}_{ad}(f,h_2) - \sup_{h_1\in\mathcal{H}} \mathcal{R}_{ad}(f,h_1)| \\
&= \sup_{f\in\mathcal{F}} |\sup_{h_2\in\mathcal{M}_h} \inf_{h_1\in\mathcal{H}} \{\mathcal{R}_{ad}(f,h_2) - \mathcal{R}_{ad}(f,h_1)\}| \\
&= \sup_{f\in\mathcal{F}} |\sup_{h_2\in\mathcal{M}_h} \inf_{h_1\in\mathcal{H}} \{\mathcal{R}_{cl}(f) + \mathcal{R}_{d}(f,h_2) - (\mathcal{R}_{cl}(f) + \mathcal{R}_{d}(f,h_1))\}| \\
&\le \sup_{f\in\mathcal{F}} |\sup_{h_2\in\mathcal{M}_h} \inf_{h_1\in\mathcal{H}} \{\mathcal{R}_{d}(f,h_2) - \mathcal{R}_{d}(f,h_1)\}| \\
&= \sup_{f\in\mathcal{F}} |\sup_{h_2\in\mathcal{M}_h} \inf_{h_1\in\mathcal{H}} \mathbb{E}_{P_\mu}[\ell_d(f,h_2) - \ell_d(f,h_1)]| \\
&\lesssim \sup_{f\in\mathcal{F}} \sup_{h_2\in\mathcal{M}_h} \inf_{h_1\in\mathcal{H}} \mathbb{E}_{P_C}[|h_2\circ f(X) - h_1\circ f(X)|] \quad \text{(Lipschitz continuity)} \\
&\le \sup_{h_2\in\mathcal{M}_h} \inf_{h_1\in\mathcal{H}} \mathbb{E}_{P_{C(U)}}[|h_2(U) - h_1(U)|] \\
&\le \sup_{h_2\in\mathcal{M}_h} \inf_{h_1\in\mathcal{H}} \|h_2 - h_1\|_\infty.
\end{aligned}$$

Lemma 11 bounds this approximation error:

$$\sup_{h_2\in\mathcal{M}_h} \inf_{h_1\in\mathcal{H}} \|h_2 - h_1\|_\infty \le \mathcal{K}_h^{-\beta_h/(d+1)}.$$

For the approximation error of $\mathcal{E}_f$, we have

$$\begin{aligned}
&\inf_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{M}_h}(f) - \mathcal{R}_{\mathcal{M}_h}(f^*) \\
&\le \inf_{f\in\mathcal{F}} \sup_{h\in\mathcal{M}_h} \big(\mathcal{R}_{ad}(f,h) - \mathcal{R}_{ad}(f^*,h)\big) \\
&\le \inf_{f\in\mathcal{F}} \sup_{h\in\mathcal{M}_h} |\mathcal{R}_{cl}(f) + \mathcal{R}_{d}(f,h) - (\mathcal{R}_{cl}(f^*) + \mathcal{R}_{d}(f^*,h))| \\
&\le \inf_{f\in\mathcal{F}} \sup_{h\in\mathcal{M}_h} \Big[|\mathcal{R}_{cl}(f) - \mathcal{R}_{cl}(f^*)| + |\mathcal{R}_{d}(f,h) - \mathcal{R}_{d}(f^*,h)|\Big] \\
&= \inf_{f\in\mathcal{F}} \sup_{h\in\mathcal{M}_h} \Big[|\mathbb{E}_{P_C\times P_A}[\ell_{cl}(f) - \ell_{cl}(f^*)]| + |\mathbb{E}_{P_\mu}[\ell_d(f,h) - \ell_d(f^*,h)]|\Big] \\
&\le c \inf_{f\in\mathcal{F}} \mathbb{E}_{P_C}[\|f(X) - f^*(X)\|] \quad \text{(Lipschitz continuity of } \ell_{cl}, \ell_d \text{ and } h) \\
&\lesssim \inf_{f\in\mathcal{F}} \|f - f^*\|_\infty.
\end{aligned}$$

Applying Lemma 11 gives

$$\inf_{f\in\mathcal{F}} \sup_{h\in\mathcal{M}_h} \|f - f^*\|_\infty \lesssim \mathcal{K}_f^{-\beta_f/(p+1)}.$$

□

M Auxiliary Results

Lemma 10 (Corollary 3.1 of Jiao et al. (2023a)). *For any $N, M \in \mathbb{N}^+$, there exists a function $g \in \mathcal{G}$ with width $\mathcal{W} = 38(\lfloor\beta_g\rfloor+1)^2 3^d d^{\lfloor\beta_g\rfloor+1} N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta_g\rfloor+1)^2 M\lceil\log_2(8M)\rceil + 2d$, such that*

$$|g(x) - g^*(x)| \le 19B_g(\lfloor\beta_g\rfloor+1)^2 d^{\lfloor\beta_g\rfloor+(\beta_g\vee 1)/2}(MN)^{-2\beta_g/d},$$

for all $x \in [0,1]^d$.

Lemma 11 (Corollary 3.2 of Jiao et al. (2023b)). *Let $d \in \mathbb{N}$ and $\beta_f = r + \alpha > 0$, where $r \in \mathbb{N}_0, \alpha \in (0,1]$. There exists $c > 0$ such that for any $\mathcal{K} \ge 1$, any $\mathcal{W} \ge c\mathcal{K}^{(2d+\beta_f)/(2d+2)}$ and $\mathcal{D} \ge 2\lceil\log_2(d+r)\rceil + 2$,*

$$\mathcal{E}(\mathcal{H}^\alpha, \mathcal{NN}(\mathcal{W},\mathcal{D},\mathcal{K})) \lesssim \mathcal{K}^{-\beta_f/(d+1)}.$$

Lemma 12 (Corollary 4 of Maurer (2016)). *Let $\mathcal{V}$ be a class of maps from a set $\mathcal{Z}$ to $\mathbb{R}^q$, and let $\ell : \mathbb{R}^q \to \mathbb{R}$ be L-Lipschitz in the Euclidean norm. For fixed $z_1, \dots, z_n$,*

$$\mathbb{E}_\epsilon \sup_{v\in\mathcal{V}} \sum_{i=1}^n \epsilon_i \ell(v(z_i)) \le \sqrt{2}L\, \mathbb{E}_{\epsilon_{it}} \sup_{v\in\mathcal{V}} \sum_{i=1}^n \sum_{t=1}^q \epsilon_{it} v_t(z_i).$$

For the contrastive application, use $q = 3d$ and $v_f(z) = (f(x), f(x^+), f(x^-))$;

Lemma 13 (Coordinatewise extension of the scalar row-sum-norm bound). *Let $z_1, \dots, z_n \in [-B, B]^p$, $B_0 = \max\{B, 1\}$, and let $\mathcal{F}$ be a d-output class of networks with the norm constraint in Definition C.3. Then*

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^d \epsilon_{it} f_t(z_i) \le dB_0\mathcal{K}_f \sqrt{\frac{2(\mathcal{D}_f + 2 + \log(p+1))}{n}}.$$

Proof. Let $\mathcal{F}_t = \{f_t : f \in \mathcal{F}\}$. Selecting one output row does not increase the maximum row-sum norm, so each $\mathcal{F}_t$ is contained in the corresponding scalar network class. Thus

$$\mathbb{E}_\epsilon \sup_f \frac{1}{n} \sum_t \sum_i \epsilon_{it} f_t(z_i) \le \sum_t \mathbb{E}_\epsilon \sup_{g\in\mathcal{F}_t} \frac{1}{n} \sum_i \epsilon_{it} g(z_i),$$

by Lemma 14, Golowich et al. (2020) and Jiao et al. (2023b), we obtain the stated bound. □

Lemma 14 (Theorem 3.2 of Golowich et al. (2020)). *For any $x_1, \dots, x_n \in [-B, B]^p$ with $1 \le B$, let $\mathcal{F}$ be the class of norm-constrained DNNs $\mathcal{NN}(\mathcal{W},\mathcal{D},\mathcal{K})$, then*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \le \frac{1}{n}\mathcal{K}\sqrt{2(\mathcal{D}+2+\log(p+1))} \max_{1\le j\le p+1} \sqrt{\sum_{i=1}^n x_{i,j}^2} \le \frac{B\mathcal{K}\sqrt{2(\mathcal{D}+2+\log(p+1))}}{\sqrt{n}}.$$

N Experiment Details

Datasets and Processes. We use Office-31 (Saenko et al., 2010), Office-Home (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019). For the reported protocol, each domain is randomly partitioned into an 80% training split and a 20% held-out test split. Representation pretraining uses images from the source and target training splits without target class labels, and downstream predictor fitting uses labels only from the source training split. The target test split is reserved for final

evaluation. We apply the same split-based protocol to every compared method, enabling controlled comparisons across learning objectives and architectures.

Table 3: The dataset information for domain adaptation experiments.

Data	Domain	Class	Total sample
Office-31	3	31	4,625
Office-Home	4	65	15,588
DomainNet	6	345	586,753

We then combine the training data from both the source and target domains and apply a series of augmentations for contrastive learning. The augmentation procedures follow the same setup as (Chen et al., 2020), including:

- Randomly cropping the image to a size of 224×224 pixels.
- Randomly applying a horizontal flip to the image with a probability of 0.5.
- Randomly applying color jitter to the image with a probability of 0.8.
- Converting the image to grayscale with a probability of 0.2.
- Applying a custom Gaussian blur with a specified kernel size. All of the above augmentations are applied to two randomized copies of the original image x to generate a positive pair consisting of two augmented views, denoted by $x^+_{(1)}$ and $x^+_{(2)}$, respectively. Other images within the same training mini-batch serve as negative samples.

Model Architectures. We use ResNet (He et al., 2016) and ViT (Dosovitskiy et al., 2020) backbones. The domain discriminator and downstream predictor are fully connected networks with ReLU activations, and a shallow perceptron connects the backbone to the discriminator. During Stage 1, a mini-batch contains unlabeled source and target training examples. Their features are used by both the contrastive loss and the domain discriminator. The discriminator minimizes its domain-classification loss, while a gradient-reversal layer sends the reversed domain gradient to the feature extractor; the contrastive gradient is propagated normally. During Stage 2, the feature extractor is frozen and only the downstream source-supervised predictor is fitted. DANN also uses labeled source and unlabeled target examples, but couples source classification with adversarial alignment; our Stage 1 replaces source classification with joint-domain self-supervised contrastive learning.

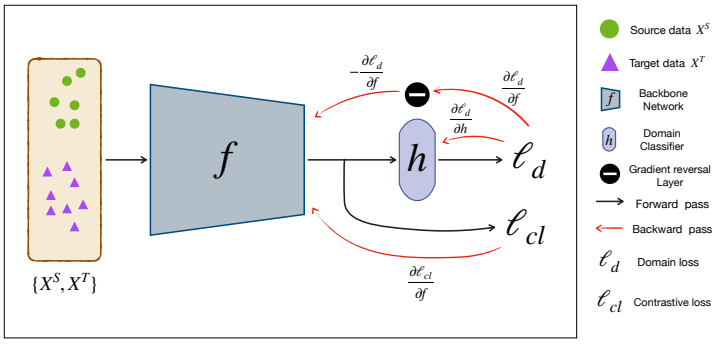


Figure 3: The framework for representation learning.

Hyperparameter Settings. Table 4 summarizes the hyperparameter settings used in our experiments. For each dataset, we train the backbone network for 800 epochs using stochastic gradient descent (SGD) as the optimizer. The initial learning rates are set to 0.01, 0.1, and $0.1(1.0)$ for the backbone, bottleneck, and discriminator, respectively. The value in parentheses indicates that the learning rate

for the discriminator is set to 1.0 when training without pretrained weights. The batch size is set to 16 to accommodate hardware constraints and is reduced to 8 when training without pretrained weights. We set the feature dimension to 512 for Office-Home and DomainNet, which is larger than that used for Office-31.

Table 4: Hyperparameter settings

Data	Epoch	Batch Size	Learning rate	Feature dimension
Office-31	800	16(8)	0.01, 0.1, 0.1(1.0)	256
Office-Home	800	16(8)	0.01, 0.1, 0.1(1.0)	512
DomainNet	800	16(8)	0.01, 0.1, 0.1(1.0)	512

Baselines. We use ERM, DANN (Ganin & Lempitsky, 2015), and CDAN (Long et al., 2018) as reference baselines, using their implementations in the Transfer Learning Library (Jiang et al., 2020). Together, these methods distinguish the contribution of source-only empirical risk minimization from adversarial and conditional domain adaptation. ViT-B (Dosovitskiy et al., 2020) provides an architecture-matched reference for the ViT experiments.

O ADDITIONAL EXPERIMENT RESULTS

Tables 5 and 6 report pretrained results on Office-31; Tables 7 and 8 report Office-Home; and Table 9 reports DomainNet. These experiments initialize the backbone from supervised pretraining on a large labeled dataset such as ImageNet (Krizhevsky et al., 2012), testing JCAR in a complementary regime to the from-scratch evaluation. On Office-31, our ViT variant remains within three percentage points of ViT-B on five of six single-source transfers. ViT-B is nevertheless higher across the pretrained single-source tables, indicating that JCAR's strongest empirical advantage currently arises when the representation is learned from scratch and that integrating joint contrastive pretraining with supervised initialization is an important direction for further improvement.

Table 5: The Top-1 accuracy (%) comparison of proposed method with others on the Office-31 dataset using pretrained model weights.

METHODS	A → W	D → W	W → D	A → D	D → A	W → A
ERM	74.8	95.0	93.9	69.7	43.2	55.4
CDAN	61.6	95.0	97.0	66.7	53.3	52.8
DANN	71.7	95.0	97.0	70.7	52.8	53.5
VIT-B	95.0	99.2	99.9	93.9	81.8	81.4
OURS (VIT)	92.0	98.9	99.5	82.0	79.1	79.8

Table 6: The Top-1 accuracy (%) comparison of proposed method with others on the Office-31 dataset using pretrained model weights.

Methods	{D,W} → A	{A, D } → W	{A, W}→ D
ERM	56.5	91.8	93.9
CDAN	56.7	92.5	89.9
DANN	56.5	81.1	90.9
ViT-B	81.4	99.2	100.0
Ours (ViT)	78.5	98.4	98.9

Table 7: The Top-1 accuracy (%) comparison of proposed method with others on the Office-Home dataset using pretrained model weights.

Methods	Ar → Rw	Pr → Rw	Cl → Pr	Ar → Cl	Pr → Ar	Cl → Rw
ERM	66.5	63.5	56.6	39.5	41.2	55.3
CDAN	64.3	65.4	54.3	36.4	41.6	50.2
DANN	65.0	67.3	57.4	39.1	41.2	51.7
ViT-B	86.8	87.8	82.8	62.3	78.2	85.6
Ours (ViT)	84.5	82.9	76.2	58.5	67.0	78.9

Table 8: The Top-1 accuracy (%) comparison of proposed method with others on the Office-Home dataset using pretrained model weights.

Methods	{Cl,Pr, Rw} ↓ Ar	{Ar, Rw, Pr } ↓ Cl	{Ar, Cl, Pr } ↓ Rw	{Ar, Cl, Rw} ↓ Pr
ERM	63.3	46.8	77.7	76.1
CDAN	67.4	51.5	77.2	73.8
DANN	62.7	48.3	75.5	73.3
ViT-B	83.7	64.4	90.0	89.6
Ours (ViT)	78.6	71.6	88.5	88.0

Table 9: The Top-1 accuracy (%) comparison of proposed method with others on the DomainNet dataset using pretrained model weights.

Methods	r → c	r → p	p → c	c → s	s → p	r → s
ERM	36.5	40.7	29.4	32.1	27.5	26.2
CDAN	51.4	50.1	41.1	40.1	41.0	40.7
DANN	48.9	48.8	35.9	41.1	40.2	39.8
ViT-B	47.4	53.5	37.8	44.7	47.1	34.8
Ours (ViT)	44.2	47.8	30.9	31.2	33.7	34.7

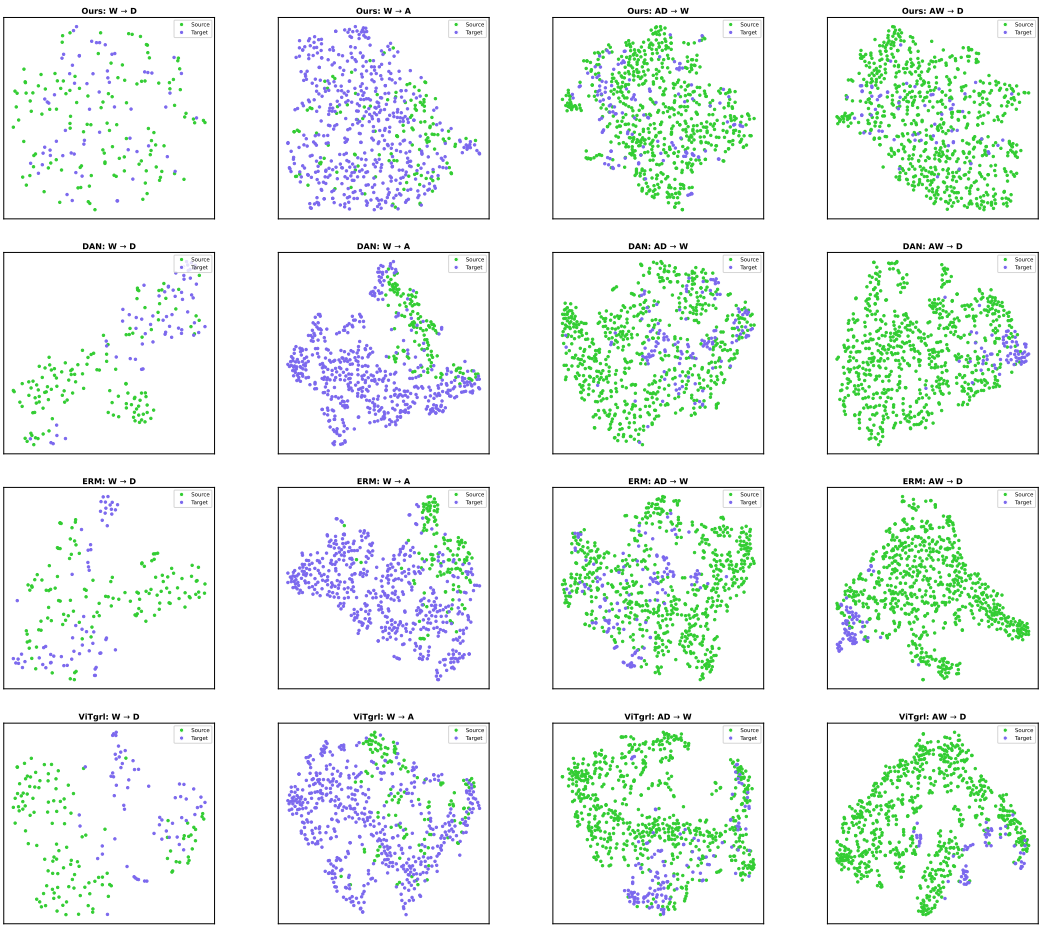


Figure 4: The t-SNE plots for the representation of the source and target domains in Office-31, the top row is our method, and other rows are the baselines. The first two columns correspond to single-source domain scenarios, and the remaining two correspond to multi-source domains.

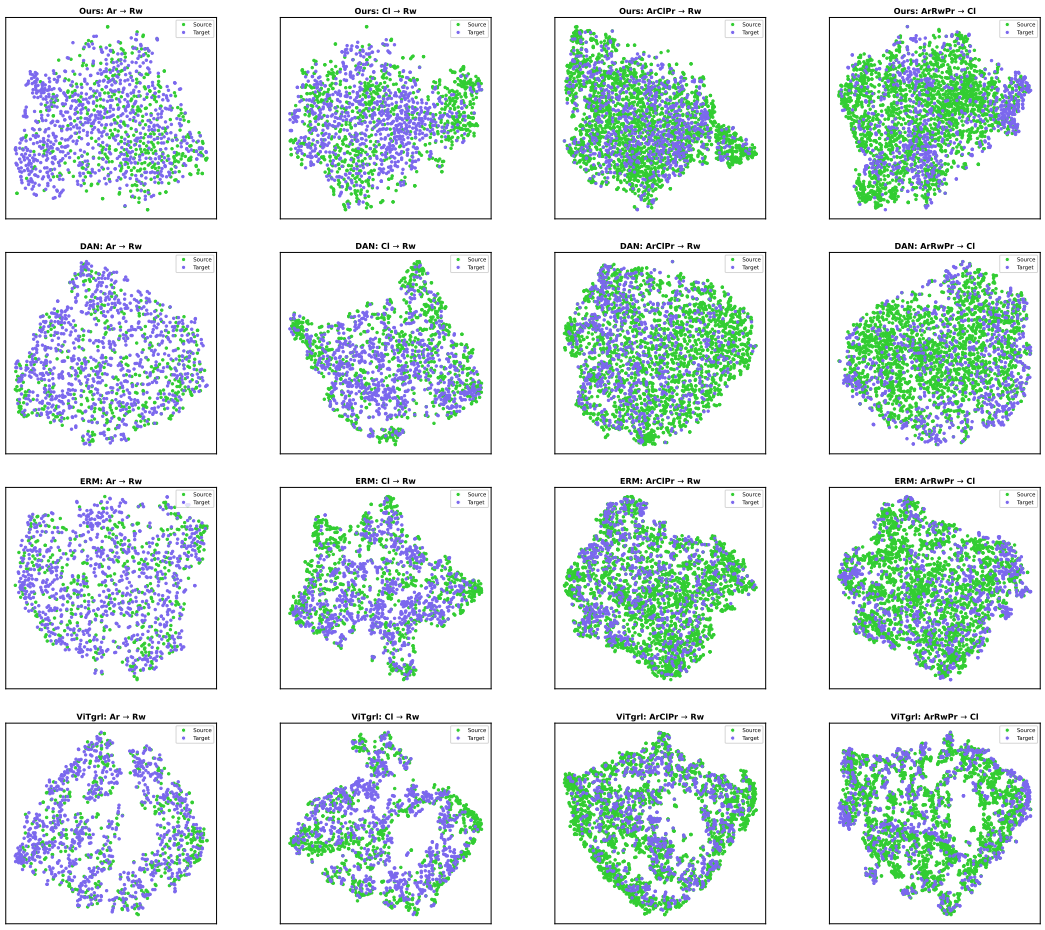


Figure 5: The t-SNE plots for the representation of the source and target domains in Office-Home, the top row is our method, and other rows are the baselines. The first two columns correspond to single-source domain scenarios, and the remaining two correspond to multi-source domains.