WHEN TO SWITCH: FROM ADVERSARIAL TRAINING TO ERM FOR OUT-OF-DISTRIBUTION GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A key challenge in out-of-distribution (OOD) learning is that models trained with empirical risk minimization (ERM) may learn spurious features. It is well known that adversarial training (AT) can suppress such learning, but OOD generalization may deteriorate in later training stages. In the present paper, we propose an adaptive two-stage AT-to-ERM strategy which uses early AT to encourage invariant feature learning, followed by ERM for empirical fitting based on the learned representation. The key is to determine when to switch. We introduce class conditional cross domain variance (CDV) to quantify cross domain discrepancies in the within-class variances of normalized features. We switch from AT to ERM when CDV shifts from decreasing to increasing. In the invariant spurious feature model, we prove that CDV is proportional to the fourth power of the normalized spurious feature coefficient, and gradient flow analysis shows that, beyond a finite time, the decay rate of CDV is at most a prescribed threshold. Experiments on five benchmarks with convolutional and Transformer backbones show that our strategy improves mean OOD accuracy by $0.93 - 2.70$ percentage points over AT and by $1.17 - 1.60$ percentage points over ERM, while outperforming MAT, Fish, and IRM. The strategy is also competitive when integrated into the strong ERM++ baseline, and CDV guided switching yields better average OOD performance than CORAL or MMD guided switching.

1 INTRODUCTION

Deep networks trained with empirical risk minimization (ERM) learn highly predictive features that generalize surprisingly well to in-distribution data (Chen et al., 2023). At the same time, they also learn spurious features, such as image backgrounds or textures, whose correlations with labels do not hold in out-of-distribution data, leading to substantial degradation in OOD performance. To address this issue, methods that learn features with stable predictive relationships to labels under distribution shifts have been widely studied in multi-source domain generalization (DG) (Muandet et al., 2013; Peters et al., 2016; Ramé et al., 2022; Shi et al., 2022; Ganin et al., 2017; Sun & Saenko, 2016; Sun et al., 2016; Ye et al., 2021; Bai et al., 2024; Arjovsky et al., 2019; Krueger et al., 2021; Kamath et al., 2021). However, DomainBed and subsequent evaluations found that many DG methods did not consistently outperform carefully implemented ERM baselines (Gulrajani & Lopez-Paz, 2021; Wiles et al., 2022; Izmailov et al., 2022; Kirichenko et al., 2023; Rosenfeld et al., 2022). These observations suggest that reducing reliance on spurious features while maintaining the strength of ERM in empirical fitting is a promising direction for improving OOD generalization.

Adversarial training (AT), which minimizes worst-case risk within local perturbation neighborhoods, can reduce models' reliance on spurious features (Madry et al., 2018; Wang et al., 2021; Wu et al., 2020). Prior work has used AT to improve OOD generalization by viewing distribution shifts as larger scale perturbations (Volpi et al., 2018; Yi et al., 2021). Wang et al. (2022) argue that sample-level adversarial perturbations may not fully capture domain-level distribution shifts and thus yield only limited improvements in OOD performance. However, a potential problem is that they ignore the fact that models trained with AT may favor different features as training progresses. For example, Wei et al. (2025) found that, in the IID setting, models trained with AT initially learn more features shared across classes but increasingly rely on class-specific features as training con-

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.

tinues and robust overfitting emerges. This raises the question of whether the benefits of AT for OOD generalization also depend on the training stage.

To this end, we conduct a two-stage training experiment in which we first use AT to encourage invariant feature learning and then continue training with ERM for empirical fitting based on the learned representation. We compare the final OOD accuracy obtained by switching at different AT checkpoints while keeping the total number of training epochs fixed. As shown in Figure 1, switching at an appropriate checkpoint yields higher OOD accuracy than AT and ERM, whereas switching too early or too late can reduce these gains. These results suggest that the benefits of AT for OOD generalization are stage-dependent. The key is therefore to determine *when to switch from AT to ERM*. In this paper, we introduce class conditional cross domain variance (CDV) to quantify cross domain discrepancies in the within-class variances of normalized features. We monitor CDV during AT and switch to ERM when CDV shifts from decreasing to increasing, This yields an adaptive two-stage AT-to-ERM strategy. Our contributions are threefold.

- Motivated by AT's potential to encourage invariant feature learning and ERM's strength in empirical fitting, we conduct a two-stage training experiment in which we first use AT to learn a representation and then continue training with ERM. The results suggest that benefits of AT for OOD generalization are stage-dependent and that the key is to determine when to switch from AT to ERM. To this end, we introduce CDV, a lightweight measure of cross-domain discrepancies in class-conditional variance, to guide the switch, forming an adaptive two-stage AT-to-ERM strategy.
- In the invariant spurious feature model, we prove that CDV is proportional to the fourth power of the normalized spurious-feature coefficient. Gradient-flow analysis further shows that, within finite time, the decay rate of CDV does not exceed a prescribed positive threshold. These results provide theoretical justification for using CDV as a lightweight metric of feature stability across domains and establish a principled basis for determining when to switch from AT to ERM.
- Extensive experiments on the DomainBed benchmark compare our adaptive AT-to-ERM strategy with existing baselines, evaluate CDV as a switching criterion and the performance of our strategy when integrated into the strong ERM++ baseline, and conduct ablations on CDV-guided checkpoint selection. These results validate the proposed strategy and demonstrate both the practical utility of CDV for guiding AT-to-ERM switching and the extensibility of our two-stage training framework to stronger baselines.

2 Related Work

Cross-domain feature statistics and model selection. Statistical discrepancies between feature distributions have been used to characterize feature invariance and guide representation learning across domains. Maximum mean discrepancy (MMD) measures distributional discrepancies through distances between kernel mean embeddings (Gretton et al., 2012). CORAL reduces differences between feature covariance matrices to align second-order statistics (Sun et al., 2016). These discrepancies have been incorporated into alignment losses to encourage domain-invariant feature learning, as in DAN and Deep CORAL (Long et al., 2015; Sun & Saenko, 2016). Ye et al. (2021) quantify cross-domain variation in class-conditional feature distributions and use it to guide final model selection. In contrast, we introduce CDV, a lightweight measure of cross-domain discrepancies in class-conditional variance, which is computed from normalized source-domain features to monitor invariant feature learning during AT and guide the switch to ERM. Final model selection follows DomainBed's training-domain validation protocol (Gulrajani & Lopez-Paz, 2021).

Adversarial training and training schedules. There is work that explores different strategies to retain the benefits of AT and use the learned representations. Rice et al. (2020) mitigate robust overfitting and improve adversarial robustness through validation-based early stopping. Yi et al. (2021) investigate adversarial pretraining followed by standard fine-tuning, using the pretrained model as an initialization to improve downstream OOD generalization. Beyond AT, Teterwak et al. (2025) use Auto-LR in ERM++ to determine the training length and learning rate schedule based on source-domain validation performance. In this work, we use CDV to guide the two-stage training schedule

and select an AT checkpoint. We then continue training from this checkpoint with ERM for empirical fitting, building on the representation learned during AT. We also incorporate this CDV guided training scheme into ERM++.

3 Stage Dependence of AT for OOD Generalization

3.1 Preliminaries

Out-of-distribution (OOD) generalization. Let $x \in \mathcal{X}$ denote the input and $y \in \mathcal{Y}$ the corresponding label. We consider predictors of the form $f = w \circ \phi(x)$, where $\phi : \mathcal{X} \to \mathcal{Z}$ is a feature extractor and $w : \mathcal{Z} \to \mathcal{Y}$ is a classifier. We consider m training domains $\mathcal{E} = \{E_1, \ldots, E_m\}$, where each domain $E_e = \{(x_i^e, y_i^e)\}_{i=1}^{n_e}$ contains n_e i.i.d. samples drawn from a distribution $\mathcal{P}_e$, and a test domain E_{m+1} with data following the distribution of $\mathcal{P}_{\text{te}}$, where $\mathcal{P}_{\text{te}} \neq \mathcal{P}_e$, $e = 1, \ldots, m$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ denote the loss function. The goal of OOD generalization is to learn a predictor f that minimizes the risk on the unseen test domain:

$$\min_f \ \mathbb{E}_{(x,y)\sim\mathcal{P}_{\text{te}}}\big[\ell(f(x), y)\big]. \tag{1}$$

Empirical Risk Minimization (ERM). In the multi domain setting, ERM minimizes the average empirical risk across training domains:

$$\min_f \frac{1}{|\mathcal{E}|} \sum_{E_e \in \mathcal{E}} \frac{1}{n_e} \sum_{i=1}^{n_e} \ell\big(f(x_i^e), y_i^e\big), \tag{2}$$

where $R_{ERM}^{E_e}(f) := \frac{1}{n_e} \sum_{i=1}^{n_e} \ell\big(f(x_i^e), y_i^e\big)$ is the empirical risk on domain E_e.

Adversarial Training (AT). AT optimizes the worst-case loss within a bounded perturbation set. For a single distribution $\mathcal{P}(x, y)$, its objective is

$$\min_f \ \mathbb{E}_{(x,y)\sim\mathcal{P}(x,y)}\Big[\max_{\delta\in\mathcal{B}} \ell\big(f(x+\delta), y\big)\Big],$$

where δ denotes an adversarial perturbation and $\mathcal{B} = \{\delta : \|\delta\|_\infty \leq \epsilon\}$ denotes the ℓ_∞-norm ball of radius ϵ. In this work, we approximate the inner maximization using the fast gradient sign method (FGSM) (Goodfellow et al., 2015), which generates the adversarial example

$$x^{\text{adv}} = x + \epsilon \ \text{sgn}(\nabla_x \ell(f(x), y)) \,.$$

In the multi domain setting, AT minimizes the average adversarial empirical risk across training domains:

$$\min_f \frac{1}{|\mathcal{E}|} \sum_{E_e \in \mathcal{E}} R_{\text{AT}}^{E_e}(f), \tag{3}$$

where $R_{\text{AT}}^{E_e}(f) := \frac{1}{n_e} \sum_{i=1}^{n_e} \max_{\delta\in\mathcal{B}} \ell\big(f(x_i^e + \delta), y_i^e\big)$ is the adversarial empirical risk on domain E_e.

3.2 AT-to-ERM Switching Experiments

We conducted AT-to-ERM switching experiments to verify that the contribution of AT to OOD generalization depends on the training stage.

We first train a model with AT and save checkpoints $\{c_t\}_{t\in\mathcal{T}}$, where $\mathcal{T} \subseteq \{1, \ldots, T\}$ is the set of checkpoint epochs. From each checkpoint c_t, we continue training with ERM for the remaining $T - t$ epochs and denote the obtaining OOD test accuracy by $\text{Acc}_{\text{AT}\to\text{ERM}}(t)$. As a baseline, we train a model with ERM for T epochs and denote its OOD test accuracy by Acc_{ERM}. The OOD accuracy difference from this baseline is

$$\Delta\text{Acc}_{\text{OOD}}(t) := \text{Acc}_{\text{AT}\to\text{ERM}}(t) - \text{Acc}_{\text{ERM}},$$

where $\Delta\text{Acc}_{\text{OOD}}(t) > 0$ indicates that switching from AT to ERM at checkpoint c_t yields higher OOD accuracy than the ERM baseline.

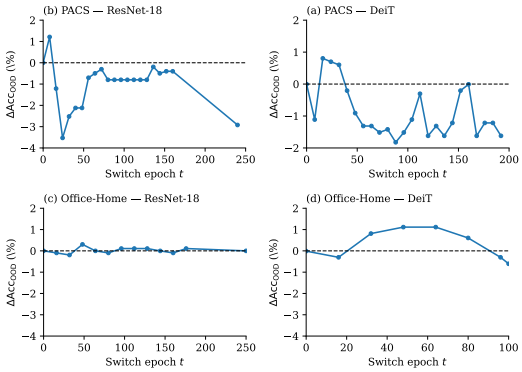


Figure 1: OOD test accuracy differences between models switched from AT to ERM at different checkpoints t and the ERM baseline on PACS and OfficeHome with ResNet-18 and DeiT.

Figure 1 shows that switching from AT to ERM at appropriate checkpoints improves OOD accuracy over ERM baseline, whereas switching at other checkpoints yields little improvement or lower accuracy. Extending the AT phase does not consistently improve final OOD accuracy. These results suggest that the benefits of AT for subsequent ERM training depend on when the switch occurs.

4 CDV Guided AT-to-ERM Training

In this section, we propose a CDV guided adaptive two-stage AT-to-ERM strategy that combines the strengths of AT and ERM to improve OOD generalization.

4.1 Class Conditional Cross Domain Variance

We introduce class conditional cross domain variance (CDV), which measures differences in the within-class total variance of normalized features across source domains.

For each source domain $E_e \in \mathcal{E}$, let $\Phi(x_i^e) \in \mathbb{R}^d$ denote the feature representation of sample x_i^e. We first normalize these representations as $\hat{\Phi}_i^e := \Phi(x_i^e)/(\|\Phi(x_i^e)\|_2 + \xi)$, where $\xi > 0$ is a small constant for numerical stability.

Let $\mathcal{I}_{e,y} := \{i : y_i^e = y\}$ denote the set of sample indices for class $y \in \mathcal{Y}$ in domain E_e, and let $n_{e,y} := |\mathcal{I}_{e,y}|$ be the number of such samples. When $n_{e,y} \geq 2$, we compute the class conditional mean $\boldsymbol{\mu}_{e,y} := \frac{1}{n_{e,y}} \sum_{i \in \mathcal{I}_{e,y}} \hat{\Phi}_i^e$ and define the within-class total variance as

$$v_{e,y} := \frac{1}{n_{e,y} - 1} \sum_{i \in \mathcal{I}_{e,y}} \left\| \hat{\Phi}_i^e - \boldsymbol{\mu}_{e,y} \right\|_2^2, \tag{4}$$

where $v_{e,y}$ equals the trace of the class conditional sample covariance matrix. For $n_{e,y} < 2$, we set $v_{e,y} = 0$.

We average the total variances across source domains for each class y as $\bar{v}_y := \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} v_{e,y}$, and define the cross-domain variation in these total variances by

$$\mathrm{CDV}_y := \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \|v_{e,y} - \bar{v}_y\|_2^2 .$$

Finally, the overall CDV is obtained by averaging over classes:

$$\mathrm{CDV} := \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \mathrm{CDV}_y. \tag{5}$$

A smaller CDV indicates more similar within-class total variances across source domains. We monitor CDV during AT as a proxy for invariant feature learning, and use it to guide checkpoint selection and the switch to ERM.

4.2 ADAPTIVE TWO-STAGE TRAINING

We first train with the AT objective in Equation 3 to encourage invariant feature learning and then continue with the ERM objective in Equation 2 for empirical fitting based on the learned representation. The switch from AT to ERM is guided by CDV, as described below.

Let $\mathcal{D}_{\text{val}}$ denote a held-out validation set drawn from the source domains. At each AT epoch t, we first compute CDV_t on $\mathcal{D}_{\text{val}}$ and apply an exponential moving average to reduce epoch-to-epoch fluctuations, obtaining the smoothed signal $\widetilde{\text{CDV}}_t$:

$$\widetilde{\text{CDV}}_t = \alpha\widetilde{\text{CDV}}_{t-1} + (1-\alpha)\text{CDV}_t, \qquad t \geq 2,$$

where $\alpha \in (0,1)$ is the smoothing coefficient and $\widetilde{\text{CDV}}_1 = \text{CDV}_1$.

We then use $\widetilde{\text{CDV}}_t$ together with a source-validation accuracy constraint, which helps preserve predictive performance on the source domains, to retain an AT checkpoint for subsequent ERM training. Let b_t denote the smoothed CDV value of the most recently retained checkpoint before epoch t. We measure the relative deterioration in $\widetilde{\text{CDV}}$ as

$$r_t := \frac{\widetilde{\text{CDV}}_t - b_t}{|b_t| + \xi},$$

where $\xi > 0$ is a small constant for numerical stability. When r_t exceeds a threshold $\delta > 0$ for a prescribed number of consecutive evaluations satisfying the accuracy constraint, we terminate the AT phase and restore the retained checkpoint. Training then continues with ERM for empirical fitting based on the representation learned during AT. The complete procedure is given in Algorithm 1.

Algorithm 1 CDV-Guided Adaptive AT-to-ERM Training

Input: Source training splits $\mathcal{D}_{\text{tr}} = \{D_e^{\text{tr}}\}_{e=1}^m$; held-out source splits $\mathcal{D}_{\text{val}} = \{D_e^{\text{val}}\}_{e=1}^m$; initial parameters θ_0; update rules $\mathcal{A}_{\text{AT}}$ and $\mathcal{A}_{\text{ERM}}$. **Training parameters:** Epoch budget T; warm-up threshold $T_{\min}$; learning rate η; per-domain batch size B. **Monitoring parameters:** EMA decay α; accuracy ratio κ; deterioration threshold δ; patience P; improvement tolerances $\epsilon_{\min}$ and ρ; numerical stabilizer $\epsilon_{\text{num}} > 0$.
Output: Trained model f_θ.

$\theta \leftarrow \theta_0$; $b^\star \leftarrow +\infty$; $\theta^\star \leftarrow \perp$; $\tau^\star \leftarrow 0$; $A_{\max} \leftarrow 0$; $c \leftarrow 0$; $s \leftarrow T$
Stage 1: AT with source-domain monitoring
for $t = 1, \ldots, T$ **do**
  $\theta \leftarrow \text{TrainEpoch}(\theta, \mathcal{D}_{\text{tr}}, \mathcal{A}_{\text{AT}}; \eta, B)$
  $C_t \leftarrow \text{CDV}(f_\theta, \mathcal{D}_{\text{val}})$; $A_t \leftarrow \text{Acc}(f_\theta, \mathcal{D}_{\text{val}})$
  $\widetilde{C}_t \leftarrow C_t$ if $t = 1$, otherwise $\alpha\widetilde{C}_{t-1} + (1-\alpha)C_t$
  $A_{\max} \leftarrow \max\{A_{\max}, A_t\}$
  if $t \geq T_{\min}$ **and** $A_t \geq \kappa A_{\max}$ **then**
    $\epsilon_t \leftarrow \max\{\epsilon_{\min}, \rho|\widetilde{C}_t|\}$
    if $\tau^\star = 0$ **or** $\widetilde{C}_t < b^\star - \epsilon_t$ **then**
      $(b^\star, \theta^\star, \tau^\star, c) \leftarrow (\widetilde{C}_t, \text{Copy}(\theta), t, 0)$
    else
      $r_t \leftarrow (\widetilde{C}_t - b^\star)/(|b^\star| + \epsilon_{\text{num}})$; $c \leftarrow c + 1$ if $r_t \geq \delta$, otherwise 0
    end if
    if $\tau^\star > 0$ **and** $c \geq P$ **then**
      $s \leftarrow t$; $\theta \leftarrow \text{Copy}(\theta^\star)$
      break
    end if
  end if
end for
Stage 2: ERM within the remaining epoch budget
for $t = s + 1, \ldots, T$ **do**
  $\theta \leftarrow \text{TrainEpoch}(\theta, \mathcal{D}_{\text{tr}}, \mathcal{A}_{\text{ERM}}; \eta, B)$
end for
return f_θ

5 THEORETICAL ANALYSIS

Following prior work Nagarajan et al. (2021); Wang et al. (2022), we consider the invariant spurious feature model, establish the relationship between CDV and the normalized spurious feature coefficient, and use gradient-flow analysis to prove the existence of a finite switching time.

Notations. We consider a binary classification problem with label $y \in \{+1, -1\}$. Let x_{inv} and x_{sp} denote the invariant features and the spurious features respectively. For simplicity, we assume there exists an identity mapping $\Phi : \mathcal{X}_{\text{inv}} \times \mathcal{X}_{\text{sp}} \to \mathcal{X}$ such that each distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ is induced by a distribution over $\mathcal{X}_{\text{inv}} \times \mathcal{X}_{\text{sp}}$ (i.e., $x = (x_{\text{inv}}, x_{\text{sp}})$). We further assume $x_{\text{sp}} \in \{+\beta, -\beta\}$ for some $\beta > 0$.

A simple OOD task. Assume two training domains (e.g., grassland vs. desert backgrounds) and one test environment. The key distribution shift is a *label–spurious correlation flip* from training to testing. Concretely, the invariant feature is perfectly predictive, i.e., $x_{\text{inv}} = y$. For each source domain $e \in \{e_1, e_2\}$, the spurious feature is generated conditionally on y as

$$x_{\text{sp}}^e = \begin{cases} +\beta y, & \text{with probability } p_e, \\ -\beta y, & \text{with probability } 1 - p_e, \end{cases}$$

where p_e controls the strength of the label–spurious correlation in domain e. We assume $p_{e_1} \neq p_{e_2}$ so that the two source domains exhibit different spurious correlations. Since the two source domains are equally weighted, the corresponding pooled training distribution satisfies $p = \frac{p_{e_1} + p_{e_2}}{2}$, and hence

$$x_{\text{sp}} = \begin{cases} +\beta y, & \text{with probability } p, \\ -\beta y, & \text{with probability } 1 - p, \end{cases}$$

where $p \in [0.5, 1)$ controls the strength of the spurious correlation in this distribution. When $p = 0.5$, the spurious feature is independent of the label in the pooled training distribution.

We first establish the relationship between CDV and the normalized spurious feature coefficient in the following lemma, and then prove that the CDV decay rate falls below any prescribed positive threshold within finite time.

Lemma 5.1 (CDV and the normalized spurious feature coefficient)**.** *Under the binary OOD model defined above, consider the linear predictor $h_t(x) = w_{\text{inv}}(t)x_{\text{inv}} + w_{\text{sp}}(t)x_{\text{sp}}$, with the two component representation $\Phi_t(x) = \big(w_{\text{inv}}(t)x_{\text{inv}}, w_{\text{sp}}(t)x_{\text{sp}}\big)^\top$. For $\xi > 0$, define the normalized coefficients $\widetilde{w}_j(t) = \frac{w_j(t)}{\sqrt{w_{\text{inv}}^2(t) + \beta^2 w_{\text{sp}}^2(t) + \xi}}, j \in \{\text{inv}, \text{sp}\}$, such that $\widehat{\Phi}_t(x) = \big(\widetilde{w}_{\text{inv}}(t)x_{\text{inv}}, \widetilde{w}_{\text{sp}}(t)x_{\text{sp}}\big)^\top$. Then the population CDV of $\widehat{\Phi}_t$ satisfies*

$$\text{CDV}(t) = 4\beta^4(2p-1)^2\big(p_{e_1} - p_{e_2}\big)^2 \widetilde{w}_{\text{sp}}^4(t).$$

Remark 5.2. For fixed source distributions with $p > 1/2$ and $p_{e_1} \neq p_{e_2}$, the proportionality constant is positive and independent of t. Thus, CDV is strictly increasing in the magnitude of the normalized spurious feature coefficient.

Theorem 5.3 (Existence of a finite switching time based on the CDV decay rate)**.** *Under the binary OOD model of Lemma 5.1, consider the linear predictor $h_t(x) = w_{\text{inv}}(t)x_{\text{inv}} + w_{\text{sp}}(t)x_{\text{sp}}$, trained by AT with exponential loss and an ℓ_∞ perturbation radius ϵ. Suppose that $0 < \epsilon < \min\{1, \beta\}, p > \frac{\beta+\epsilon}{2\beta}$, and $w_{\text{inv}}(0) = 0, w_{\text{sp}}(0) > w_{\text{sp}}^*$, where $w_{\text{sp}}^* = \frac{1}{2\beta}\ln\Big(\frac{p}{1-p}\frac{\beta-\epsilon}{\beta+\epsilon}\Big)$. Then, for any prescribed threshold $\gamma > 0$, the switching time $t^* := \inf\big\{t \geq 0 : -\frac{d}{dt}\text{CDV}(t) \leq \gamma\big\}$ is finite and satisfies*

$$0 \leq t^* \leq \frac{\text{CDV}(0)}{\gamma} < \infty.$$

Remark 5.4. Theorem 5.3 establishes that, along the AT gradient flow trajectory, the CDV decay rate must fall below any prescribed positive threshold within finite time. This follows from the nonnegativity of CDV, which implies that the decay rate cannot remain uniformly positive indefinitely. Therefore, the CDV decay rate criterion defines a finite AT phase and provides a theoretical basis for switching from AT to ERM.

Proof. The proofs of Lemma 5.1 and Theorem 5.3 are provided in Appendices A.1 and A.2, respectively. □

6 EXPERIMENTS

In this section, we evaluate the proposed adaptive AT-to-ERM strategy, the effectiveness of CDV as a switching criterion, and the AT-to-ERM++ variant that integrates our strategy into the ERM++ baseline. We additionally analyze CDV dynamics during training across different algorithms, with results reported in Appendix A.4.

6.1 EXPERIMENTAL SETUP

We compare our adaptive AT-to-ERM strategy with ERM, AT (Madry et al., 2018), MAT (Wang et al., 2022), Fish (Shi et al., 2021), CORAL (Sun & Saenko, 2016), and IRM (Arjovsky et al., 2019). ERM and AT serve as the primary single-stage baselines. The remaining baselines cover domain-wise adversarial perturbation design (MAT), invariant prediction (IRM), cross-domain gradient alignment (Fish), and second-order feature alignment (CORAL), which also achieves strong performance in the original DomainBed study (Gulrajani & Lopez-Paz, 2021). Experiments are conducted on five multi-source DG datasets: OfficeHome (Venkateswara et al., 2017), DomainNet (Peng et al., 2019), PACS (Li et al., 2017), VLCS (Fang et al., 2013), and TerraIncognita (Beery et al., 2018). We use four backbones: ResNet-18 (He et al., 2016) and ResNet-50 (He et al., 2016), DeiT (Touvron et al., 2021) and ViT-B/16 (Oquab et al., 2024).

Implementation details. We follow the DomainBed training procedure. For ResNet-based and DeiT backbones, we use the default hyperparameters from DomainBed (Gulrajani & Lopez-Paz, 2021), e.g., batch size of 32 (per-domain), learning rate of 5e-5, ResNet dropout value of 0, and weight decay of 0. For ViT-B/16, we follow the training configuration in Teterwak et al. (2025). Final model selection follows DomainBed's training-domain validation protocol. We compute CDV on the same held-out source-domain validation set, which contains no target-domain data. The CDV-guided switching hyperparameters are provided in Appendix A.3.

6.2 EXPERIMENTAL RESULTS

Table 1: OOD accuracy (%, mean $\pm$ standard deviation) with ResNet-18 and DeiT under hyperparameter search. Avg. is the mean over the three datasets; the best and second-best results are shown in **bold** and underlined, respectively.

Backbone	Method	PACS	OfficeHome	VLCS	Avg.
ResNet-18	ERM	80.8 ± 0.5	59.6 ± 0.5	74.5 ± 0.7	71.63
	AT	80.8 ± 0.1	59.7 ± 0.3	74.8 ± 0.2	71.77
	MAT	80.5 ± 1.0	59.1 ± 0.2	74.2 ± 0.2	71.27
	Fish	81.5 ± 0.3	60.4 ± 0.2	$\underline{75.7} \pm 0.1$	72.53
	CORAL	$\mathbf{81.7} \pm 0.8$	$\mathbf{63.2} \pm 0.1$	$\underline{75.7} \pm 1.0$	**73.53**
	IRM	80.0 ± 0.8	59.5 ± 0.4	75.1 ± 0.6	71.53
	AT-to-ERM (ours)	$\underline{81.6} \pm 0.1$	$\underline{60.6} \pm 0.1$	$\mathbf{77.2} \pm 0.0$	$\underline{73.13}$
DeiT	ERM	84.3 ± 0.1	72.4 ± 0.1	79.1 ± 0.4	78.60
	AT	85.7 ± 0.3	71.6 ± 0.2	79.4 ± 0.2	78.90
	MAT	85.4 ± 0.1	70.8 ± 0.7	79.6 ± 0.2	78.60
	Fish	85.3 ± 0.5	72.1 ± 0.3	78.6 ± 0.3	78.67
	CORAL	85.5 ± 0.2	72.4 ± 0.0	$\underline{79.8} \pm 0.5$	$\underline{79.23}$
	IRM	$\underline{85.9} \pm 0.3$	$\underline{72.5} \pm 0.4$	79.3 ± 0.3	$\underline{79.23}$
	AT-to-ERM (ours)	$\mathbf{86.5} \pm 0.2$	$\mathbf{73.0} \pm 0.2$	$\mathbf{80.0} \pm 0.2$	**79.83**

Comparison with Baseline Methods As shown in Table 1, AT-to-ERM outperforms ERM, AT, MAT, Fish, and IRM in mean OOD accuracy with both ResNet-18 and DeiT. With DeiT, it achieves the highest accuracy on all datasets; with ResNet-18, it ranks second in mean OOD accuracy behind CORAL. With ResNet-50 and ViT-B/16 (Table 2), AT-to-ERM achieves the highest mean OOD accuracy. Compared with CORAL, our method improves mean OOD accuracy by 0.84 and 1.33 percentage points with ResNet-50 and ViT-B/16, respectively.

Table 2: OOD accuracy (%) with ResNet-50 and ViT-B/16. TerraInc. denotes TerraIncognita. Within each backbone, the best and second-best results are **bold** and underlined, respectively.

Method	PACS	OfficeHome	VLCS	DomainNet	TerraInc.	Avg.
ResNet-50						
ERM	86.3 ± 0.3	68.6 ± 0.5	77.6 ± 0.7	43.8 ± 0.2	$\underline{47.4} \pm 1.5$	64.74
AT	85.0 ± 0.2	68.1 ± 0.8	$\underline{77.8} \pm 0.3$	43.6 ± 0.0	43.6 ± 0.9	63.62
Fish	86.4 ± 0.6	$\underline{71.8} \pm 0.2$	77.3 ± 0.6	43.1 ± 0.1	43.1 ± 0.7	64.34
CORAL	$\underline{87.1} \pm 0.3$	71.6 ± 0.5	77.2 ± 0.4	$\mathbf{44.6} \pm 0.1$	47.0 ± 0.6	$\underline{65.50}$
IRM	86.1 ± 0.6	67.9 ± 0.5	77.6 ± 0.7	43.8 ± 0.2	46.8 ± 1.2	64.44
AT-to-ERM (ours)	$\mathbf{87.8} \pm 0.1$	$\mathbf{72.2} \pm 0.2$	$\mathbf{78.6} \pm 0.2$	$\underline{43.9} \pm 0.0$	$\mathbf{49.2} \pm 0.4$	**66.34**
ViT-B/16						
ERM	93.7 ± 0.4	82.9 ± 0.1	81.5 ± 0.2	–	–	86.03
AT	94.6 ± 0.4	81.9 ± 0.3	80.9 ± 0.2	–	–	85.80
CORAL	94.3 ± 0.2	82.7 ± 0.0	80.6 ± 0.3	–	–	85.87
AT-to-ERM (ours)	$\mathbf{95.6 \pm 0.0}$	$\mathbf{83.2 \pm 0.0}$	$\mathbf{82.8 \pm 0.1}$	–	–	**87.20**

Table 3: OOD accuracy (%) of ERM++ and AT-to-ERM++. AT-to-ERM++ integrates our CDV-guided training strategy into ERM++.

Method	PACS	OfficeHome	VLCS	TerraIncognita	Avg.
ERM++	89.82	74.57	78.37	52.92	73.92
AT-to-ERM++ (ours)	**90.40**	**74.96**	**78.66**	**53.75**	**74.44**

Integration with ERM++ As shown in Table 3, AT-to-ERM++ improves OOD accuracy over ERM++ on all four datasets, with the average increasing from 73.92% to 74.44%. These results suggest that our strategy also benefits the stronger ERM++ baseline.

Table 4: Ablation of the AT-to-ERM switching time. Earlier and Later are switching before and after the CDV-selected checkpoint, respectively; the best result is shown in **bold**.

Backbone	Switching time	PACS	OfficeHome	VLCS	Avg.
ResNet-18	Earlier	77.7	59.9	76.0	71.20
	Later	81.3	60.0	75.2	72.17
	CDV-selected	**82.1**	**60.4**	**77.2**	**73.23**
DeiT	Earlier	85.9	73.3	79.9	79.70
	Later	85.3	73.6	**80.6**	79.83
	CDV-selected	**87.3**	**73.8**	80.2	**80.43**

Table 5: OOD accuracy (%) for AT checkpoint selection with ResNet-18. ERM is initialized from either the deterioration-trigger or CDV-selected checkpoint; the best result is **bold**.

ERM initialization	PACS	OfficeHome	VLCS	Avg.
Deterioration-trigger checkpoint	81.2 ± 0.1	60.5 ± 0.2	75.1 ± 0.7	72.27
CDV-selected checkpoint	$\mathbf{81.9} \pm 0.2$	$\mathbf{60.8} \pm 0.2$	$\mathbf{75.7} \pm 0.8$	**72.80**

Ablation of CDV-Guided Switching Table 4 evaluates the timing of the AT-to-ERM transition. CDV-guided switching achieves the highest mean OOD accuracy with both ResNet-18 (73.23%) and DeiT (80.43%). It performs best on all three datasets with ResNet-18 and on PACS and OfficeHome

with DeiT; later switching performs better only on VLCS with DeiT. Table 5 evaluates checkpoint restoration. With ResNet-18, restoring the CDV-selected checkpoint improves OOD accuracy by 0.7, 0.3, and 0.6 percentage points on PACS, OfficeHome, and VLCS, respectively, increasing the mean from 72.27% to 72.80%. These results support restoring the CDV-selected checkpoint rather than switching directly from the deterioration-trigger checkpoint.

Table 6: Comparison of OOD accuracy (%) using MMD, CORAL, and CDV as switching criteria for AT-to-ERM training with ResNet-50. All three criteria are used only to guide switching and are not added as training losses. The best result is shown in **bold**.

Criterion	PACS	OfficeHome	VLCS	TerraIncognita	Avg.
MMD	85.2	71.8	**78.7**	43.8	69.88
CORAL	85.2	72.4	78.3	**46.9**	70.70
CDV (ours)	**86.0**	**72.5**	78.5	46.5	**70.88**

Comparison of Switching Criteria Table 6 shows that using CDV as the switching criterion yields a mean OOD accuracy of 70.88% across the four datasets, compared with 69.88% for MMD and 70.70% for CORAL. These results show that our two-stage can accommodate different discrepancy measures, with CDV achieving the highest mean OOD accuracy.

Table 7: Representation collapse diagnostics on PACS using the ℓ_2-normalized source-validation features used for CDV computation. T_0 denotes the CDV-selected AT checkpoint, and Final denotes the checkpoint used to report the final OOD test accuracy; the best result is shown in **bold**.

Model / checkpoint	Eff. rank ↑	C-rank ↑	Trace	Top-1 ↓	Sep. ↑	Deg.
AT	14.055	0.2283	0.5440	0.2279	1.0324	0/21
ERM	11.476	0.1942	0.6151	**0.1955**	1.2282	0/21
CORAL	13.618	**0.2749**	0.4595	0.2097	0.8647	0/21
AT-to-ERM (T_0)	**18.435**	0.2705	0.4892	0.2156	0.8672	0/21
AT-to-ERM (Final)	12.520	0.2282	0.6707	0.2049	**1.3231**	0/21

Analysis of Representation Collapse Table 7 shows that the CDV-selected checkpoint has the highest effective rank (Eff. rank 18.435) and a higher class-conditional normalized rank (C-rank) than AT and ERM, with no degenerate groups (Deg.) detected. After further ERM training, effective rank decreases while inter-class separation (Sep.) increases to 1.3231, exceeding AT, ERM, and CORAL. These results indicate that the low CDV at the selected checkpoint is not caused by representation collapse, supporting CDV as a reliable criterion for checkpoint selection.

7 Conclusion

We study adversarial training for OOD generalization from a training dynamics perspective and show that its benefits are stage-dependent. Motivated by this observation, we propose CDV to monitor invariant feature learning during AT and guide the switch from AT to ERM. The theoretical results establish CDV as a lightweight and principled metric of cross-domain feature stability in multi-source DG and support the existence of a finite switching point from AT to ERM. Experiments on five DomainBed benchmarks with convolutional and Transformer backbones further validate the effectiveness of CDV-guided switching and show that the resulting two-stage training strategy can be integrated into stronger baselines, such as ERM++, to better exploit the potential of AT for OOD generalization. Our study is currently limited to standard image classification benchmarks under the DomainBed protocol, and the theoretical analysis relies on a simplified invariant spurious feature model. Future work will investigate the proposed stage-wise training principle under broader distribution shifts, with larger-scale models, and within stronger training frameworks.

AI USE STATEMENT

Generative AI tools were used solely to aid and polish the writing of this paper, including language refinement and grammatical correction. We reviewed all AI-assisted edits and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We support reproducibility by providing the methodological, experimental, and theoretical details required to reproduce our results. The proposed CDV-guided AT-to-ERM procedure is described in the main text and Algorithm 1, while the experimental setup specifies the datasets, model backbones, training protocol, model-selection procedure, and CDV computation. The hyperparameters specific to CDV-guided switching are reported in Appendix A.3. For the theoretical results, the model assumptions and problem setting are stated in the main text, and complete proofs of the relationship between CDV and the normalized spurious-feature coefficient and of the finite switching time result are provided in Appendices A.1 and A.2, respectively.

REFERENCES

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. URL http://arxiv.org/abs/1907.02893.

Haoyue Bai, Yifei Ming, Julian Katz-Samuels, and Yixuan Li. HYPO: hyperspherical out-of-distribution generalization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=VXak3CZZGC.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pp. 472–489. Springer, 2018. doi: 10.1007/978-3-030-01270-0_28. URL https://doi.org/10.1007/978-3-030-01270-0_28.

Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d73d5645ddbb9ada6c862116435574f6-Abstract-Conference.html.

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pp. 1657–1664. IEEE Computer Society, 2013. doi: 10.1109/ICCV.2013.208. URL https://doi.org/10.1109/ICCV.2013.208.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. *Domain-Adversarial Training of Neural Networks*, pp. 189–209. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58347-1. doi: 10.1007/978-3-319-58347-1_10. URL https://doi.org/10.1007/978-3-319-58347-1_10.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012. doi: 10.5555/2503308.2188410. URL https://dl.acm.org/doi/10.5555/2503308.2188410.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Pritish Kamath, Akilesh Tangella, Danica J. Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 4069–4077. PMLR, 2021. URL http://proceedings.mlr.press/v130/kamath21a.html.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 5543–5551. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.591. URL https://doi.org/10.1109/ICCV.2017.591.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 97–105. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/long15.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 10–18. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/muandet13.html.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=fSTD6NFIW_b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 1406–1415. IEEE, 2019. doi: 10.1109/ICCV.2019.00149. URL https://doi.org/10.1109/ICCV.2019.00149.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(5):947–1012, 2016. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/44682904.

Alexandre Ramé, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18347–18377. PMLR, 2022. URL https://proceedings.mlr.press/v162/rame22a.html.

Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8093–8104. PMLR, 2020. URL http://proceedings.mlr.press/v119/rice20a.html.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: ERM may already learn features sufficient for out-of-distribution generalization. *CoRR*, abs/2202.06856, 2022. URL https://arxiv.org/abs/2202.06856.

Yuge Shi, Jeffrey Seely, Philip H. S. Torr, N. Siddharth, Awni Y. Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *CoRR*, abs/2104.09937, 2021. URL https://arxiv.org/abs/2104.09937.

Yuge Shi, Jeffrey Seely, Philip H. S. Torr, Siddharth Narayanaswamy, Awni Y. Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=vDwBW49HmO.

Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou (eds.), *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, volume 9915 of *Lecture Notes in Computer Science*, pp. 443–450, 2016. doi: 10.1007/978-3-319-49409-8_35. URL https://doi.org/10.1007/978-3-319-49409-8_35.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In Dale Schuurmans and Michael P. Wellman (eds.), *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 2058–2065. AAAI Press, 2016. doi: 10.1609/AAAI.V30I1.10306. URL https://doi.org/10.1609/aaai.v30i1.10306.

Piotr Teterwak, Kuniaki Saito, Theodoros Tsiligkaridis, Kate Saenko, and Bryan A. Plummer. ERM++: an improved baseline for domain generalization. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025*, pp. 8525–8535. IEEE, 2025. doi: 10.1109/WACV61041.2025.00826. URL https://doi.org/10.1109/WACV61041.2025.00826.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 2021. URL http://proceedings.mlr.press/v139/touvron21a.html.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5385–5394. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.572. URL https://doi.org/10.1109/CVPR.2017.572.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5339–5349, 2018.

Qixun Wang, Yifei Wang, Hong Zhu, and Yisen Wang. Improving out-of-distribution generalization by adversarial training with structured priors. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. *CoRR*, abs/2112.08304, 2021. URL https://arxiv.org/abs/2112.08304.

Zeming Wei, Steven Y. Guo, and Yisen Wang. Identifying and understanding cross-class features in adversarial training. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=FvBYG5jA7k.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=Dl4LetuLdyK.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 23519–23531, 2021.

Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhiming Ma. Improved OOD generalization via adversarial training and pretraing. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11987–11997. PMLR, 2021. URL http://proceedings.mlr.press/v139/yi21a.html.

A Appendix

A.1 Proof of Lemma 5.1

Proof. Following the definition of CDV in Section 4.1, we first compute the class conditional variance of the normalized representation. For any source domain $e \in \{e_1, e_2\}$ and label $y \in \{-1, +1\}$, the invariant component $\widetilde{w}_{\text{inv}}(t)x_{\text{inv}} = \widetilde{w}_{\text{inv}}(t)y$ is deterministic conditional on (e, y). Therefore,

$$\text{Var}\left(\widetilde{w}_{\text{inv}}(t)x_{\text{inv}} \mid e, y\right) = 0. \tag{6}$$

For the spurious feature, the assumed conditional distribution gives

$$\mathbb{E}[x^e_{\text{sp}} \mid e, y] = \beta y(2p_e - 1), \qquad \mathbb{E}[(x^e_{\text{sp}})^2 \mid e, y] = \beta^2.$$

Since $y^2 = 1$, its conditional variance is

$$\begin{aligned}\text{Var}(x^e_{\text{sp}} \mid e, y) &= \mathbb{E}[(x^e_{\text{sp}})^2 \mid e, y] - \left(\mathbb{E}[x^e_{\text{sp}} \mid e, y]\right)^2 \\ &= \beta^2 - \beta^2(2p_e - 1)^2 \\ &= 4\beta^2 p_e(1 - p_e).\end{aligned}$$

It follows that

$$\text{Var}\big(\widetilde{w}_{\text{sp}}(t)x^e_{\text{sp}} \mid e, y\big) = 4\beta^2 p_e(1 - p_e)\widetilde{w}^2_{\text{sp}}(t). \tag{7}$$

Combining Equations equation 6 and equation 7 therefore gives

$$v_{e,y}(t) = 4\beta^2 p_e(1 - p_e)\widetilde{w}^2_{\text{sp}}(t).$$

For the two equally weighted source domains, let $\bar{v}_y(t) := \frac{v_{e_1,y}(t)+v_{e_2,y}(t)}{2}$. By the population definition of CDV, we obtain

$$\begin{aligned}\text{CDV}_y(t) &= \frac{1}{2}\sum_{e\in\{e_1,e_2\}} \big(v_{e,y}(t) - \bar{v}_y(t)\big)^2 \\ &= \frac{1}{4}\big(v_{e_1,y}(t) - v_{e_2,y}(t)\big)^2 \\ &= 4\beta^4\widetilde{w}^4_{\text{sp}}(t)\left[p_{e_1}(1 - p_{e_1}) - p_{e_2}(1 - p_{e_2})\right]^2. \end{aligned} \tag{8}$$

Using $p_{e_1} + p_{e_2} = 2p$, we further obtain

$$\begin{aligned}&p_{e_1}(1 - p_{e_1}) - p_{e_2}(1 - p_{e_2}) \\ &= (p_{e_1} - p_{e_2})\big[1 - (p_{e_1} + p_{e_2})\big] \\ &= (1 - 2p)(p_{e_1} - p_{e_2}).\end{aligned} \tag{9}$$

Substituting Equation (9) into Equation (equation 8) yields

$$\text{CDV}_y(t) = 4\beta^4(2p - 1)^2(p_{e_1} - p_{e_2})^2\widetilde{w}^4_{\text{sp}}(t).$$

Averaging over over $y \in \{-1, +1\}$ therefore gives

$$\text{CDV}(t) = 4\beta^4(2p - 1)^2(p_{e_1} - p_{e_2})^2\widetilde{w}^4_{\text{sp}}(t).$$

□

A.2 Proof of Theorem 5.3

Proof. Let $\mathcal{S}_{\text{min}}$ and $\mathcal{S}_{\text{maj}}$ denote the subsets of data points in $\mathcal{S}$ for which the spurious feature and label are negatively correlated and positively correlated, respectively, i.e.,

$$\mathcal{S}_{\text{min}} = \{(x, y) \in \mathcal{S} : x_{sp} \cdot y < 0\}, \qquad \mathcal{S}_{\text{maj}} = \{(x, y) \in \mathcal{S} : x_{sp} \cdot y > 0\}.$$

Let $\mathcal{D}_{inv}$ denote the distribution over (x_{inv}, y) induced by drawing (x, y) uniformly from $\mathcal{S}_{min}$. The corresponding marginal distribution of y is denoted as $\mathcal{D}_y$. By the assumption of the theorem,

this distribution $\mathcal{D}$ would be the same if x was drawn uniformly from $\mathcal{S}_{maj}$. In this setting, the adversarial risk minimized by the classifier f_w can be written as

$$\mathcal{L}(f(x), y) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\|\delta\|\leq\epsilon} \ell(f_w(x+\delta), y)\right], \text{ where } \ell(f, y) = \exp\{-yf\}. \tag{10}$$

We first compute the inner maximization

$$\max_{\|\delta\|\leq\epsilon} \exp\{-y\, f_w(x+\delta)\}$$

for a linear classifier $f_w(x) = w_{inv}x_{inv}+w_{sp}x_{sp}$. Let $\delta = (\delta_{inv},\, \delta_{sp})$ be an adversarial perturbation with $\|\delta\| \leq \epsilon$. Then

$$\max_{\|\delta\|\leq\epsilon} \exp\{-yf_w(x+\delta)\} = \max_{\|\delta\|\leq\epsilon} \exp\{-y[w_{inv}(x_{inv}+\delta_{inv}) + w_{sp}(x_{sp}+\delta_{sp})]\}.$$

The gradient sgn for each coordinate gives the optimal perturbation, i.e.,

$$\begin{aligned} x_{inv} + \delta_{inv} =& x_{inv} + \epsilon\text{sgn}(\nabla_{x_{inv}}\ell(f(x)y) \\ =& x_{inv} + \epsilon\text{sgn}\left(\frac{\partial(\exp\{-y(w_{inv}x_{inv}+w_{sp}x_{sp})\})}{\partial x_{inv}}\right) \\ =& x_{inv} + \epsilon\text{sgn}\left(\exp\{-y(w_{inv}x_{inv}+w_{sp}x_{sp})\}(-yw_{inv})\right) \end{aligned}$$

and

$$\begin{aligned} x_{sp} + \delta_{sp} =& x_{sp} + \epsilon\text{sgn}(\nabla_{x_{sp}}\ell(f(x)y) \\ =& x_{sp} + \epsilon\text{sgn}\left(\frac{\partial(\exp\{-y(w_{inv}x_{inv}+w_{sp}x_{sp})\})}{\partial x_{sp}}\right) \\ =& x_{sp} + \epsilon\text{sgn}\left(\exp\{-y(w_{inv}x_{inv}+w_{sp}x_{sp})\}(-yw_{sp})\right). \end{aligned}$$

Thus the optimal adversarial perturbation shifts each coordinate by

$$x_{inv} + \delta_{inv} = x_{inv} + -y\epsilon\text{sgn}(w_{inv}),\;\; x_{sp} + \delta_{sp} = x_{sp} + -y\epsilon\text{sgn}(w_{sp}).$$

We consider the case $w_{inv} > 0$, $w_{sp} > 0$, so the optimal perturbation is $\delta^* = (\delta^*_{inv}, \delta^*_{sp}) = (-y\epsilon, -y\epsilon)$. Substituting δ^* into (10) yields

$$\begin{aligned} \mathcal{L}(f(x), y) =& \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[e^{-f(x+\delta^*)y}\right] \\ =& \mathbb{E}_{(x_{inv},y)\sim\mathcal{D}_{inv}}\left[\mathbb{E}_{x_{sp}|x_{inv},y}\left(e^{f(x+\delta^*)y}\right)\right]. \end{aligned}$$

Under the conditional distribution of $x_{sp} \mid (x_{inv}, y)$, we have

$$\Pr(x_{sp} = +\beta y \mid x_{inv}, y) = p, \qquad \Pr(x_{sp} = -\beta y \mid x_{inv}, y) = 1-p.$$

Therefore,

$$\begin{aligned} & \mathcal{L}(f(x), y) \\ =\;& \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[e^{-f(x+\delta^*)y}\right] \\ =\;& \mathbb{E}_{(x_{inv},y)\sim\mathcal{D}_{inv}}\left[\mathbb{E}_{x_{sp}|x_{inv},y}\left(e^{-f(x+\delta^*)y}\right)\right] \\ =\;& \mathbb{E}_{(x_{inv},y)\sim\mathcal{D}_{inv}}\left[pe^{-yw_{inv}x_{inv}-w_{sp}\beta+\epsilon(w_{inv}+w_{sp})} + (1-p)e^{-yw_{inv}x_{inv}+w_{sp}\beta+\epsilon(w_{inv}+w_{sp})}\right] \\ =\;& \mathbb{E}_{(x_{inv},y)\sim\mathcal{D}_{inv}} e^{w_{inv}(-yx_{inv}+\epsilon)}\left[pe^{-w_{sp}\beta+\epsilon w_{sp}} + (1-p)e^{w_{sp}\beta+\epsilon w_{sp}}\right]. \end{aligned} \tag{11}$$

We prove that the CDV decay rate falls below a prescribed threshold within finite time. The proof proceeds in three steps.

Step 1: AT gradient flow dynamics. From Equation 11, the invariant and spurious coefficients satisfy

$$\dot{w}_{\text{inv}}(t) = (1-\epsilon)e^{-(1-\epsilon)w_{\text{inv}}(t)}\left[pe^{-(\beta-\epsilon)w_{\text{sp}}(t)} + (1-p)e^{(\beta+\epsilon)w_{\text{sp}}(t)}\right], \tag{12}$$

$$\dot{w}_{\text{sp}}(t) = e^{-(1-\epsilon)w_{\text{inv}}(t)}\left[p(\beta-\epsilon)e^{-(\beta-\epsilon)w_{\text{sp}}(t)} - (1-p)(\beta+\epsilon)e^{(\beta+\epsilon)w_{\text{sp}}(t)}\right]. \tag{13}$$

Since $0 < \epsilon < 1$, Equation 12 implies

$$\dot{w}_{\text{inv}}(t) > 0.$$

Together with $w_{\text{inv}}(0) = 0$, this gives $w_{\text{inv}}(t) \geq 0$ for all $t \geq 0$.

To analyze the spurious coefficient, define

$$g(v) := p(\beta - \epsilon)e^{-(\beta-\epsilon)v} - (1-p)(\beta+\epsilon)e^{(\beta+\epsilon)v}.$$

Then Equation 13 can be written as

$$\dot{w}_{\text{sp}}(t) = e^{-(1-\epsilon)w_{\text{inv}}(t)} g\big(w_{\text{sp}}(t)\big),$$

so the sign of $\dot{w}_{\text{sp}}(t)$ is determined by $g\big(w_{\text{sp}}(t)\big)$. Moreover,

$$g'(v) = -p(\beta-\epsilon)^2 e^{-(\beta-\epsilon)v} - (1-p)(\beta+\epsilon)^2 e^{(\beta+\epsilon)v} < 0,$$

and hence g is strictly decreasing.

Letting $g(v) = 0$ yields the unique equilibrium

$$w^*_{\text{sp}} = \frac{1}{2\beta} \ln\left(\frac{p}{1-p}\frac{\beta-\epsilon}{\beta+\epsilon}\right).$$

By assumption, $w_{\text{sp}}(0) > w^*_{\text{sp}} > 0$. Since g is strictly decreasing and $g(w^*_{\text{sp}}) = 0$, we have $g(w_{\text{sp}}(0)) < 0$.

Furthermore, $w_{\text{sp}} = w^*_{\text{sp}}$ is invariant under the gradient flow. Since the vector field is smooth, uniqueness of the solution prevents a trajectory initialized with $w_{\text{sp}}(0) > w^*_{\text{sp}}$ from crossing this equilibrium. Therefore,

$$w_{\text{sp}}(t) > w^*_{\text{sp}} > 0, \qquad \dot{w}_{\text{sp}}(t) < 0, \qquad t \geq 0.$$

Step 2: Monotonicity of CDV along the AT trajectory. Let $C(t) := \text{CDV}(t)$. By Lemma 5.1,

$$C(t) = C_{\mathcal{D}} \widetilde{w}^4_{\text{sp}}(t), \qquad C_{\mathcal{D}} = 4\beta^4(2p-1)^2(p_{e_1} - p_{e_2})^2, \tag{14}$$

where

$$\widetilde{w}_{\text{sp}}(t) = \frac{w_{\text{sp}}(t)}{R(t)+\xi}, \qquad R(t) = \sqrt{w^2_{\text{inv}}(t) + \beta^2 w^2_{\text{sp}}(t)}.$$

Differentiating $R(t)$ gives

$$\dot{R}(t) = \frac{w_{\text{inv}}(t)\dot{w}_{\text{inv}}(t) + \beta^2 w_{\text{sp}}(t)\dot{w}_{\text{sp}}(t)}{R(t)}.$$

Therefore,

$$\begin{aligned}\dot{\widetilde{w}}_{\text{sp}}(t) &= \frac{(R+\xi)\dot{w}_{\text{sp}} - w_{\text{sp}}\dot{R}}{(R+\xi)^2} \\ &= \frac{\big(w^2_{\text{inv}} + \xi R\big)\dot{w}_{\text{sp}} - w_{\text{inv}} w_{\text{sp}} \dot{w}_{\text{inv}}}{R(R+\xi)^2},\end{aligned} \tag{15}$$

where the dependence on t is omitted above for brevity.

From Step 1,

$$w_{\text{inv}}(t) \geq 0, \qquad w_{\text{sp}}(t) > 0, \qquad \dot{w}_{\text{inv}}(t) > 0, \qquad \dot{w}_{\text{sp}}(t) < 0.$$

Moreover, since $\xi > 0$ and $R(t) > 0$, $w^2_{\text{inv}} + \xi R > 0$. Then the first term in the numerator of Equation 15 is strictly negative, while the second is non-positive. Therefore,

$$\dot{\widetilde{w}}_{\text{sp}}(t) < 0.$$

Combining the conditions $p > (\beta+\epsilon)/(2\beta) > 1/2$ and $p_{e_1} \neq p_{e_2}$, we have $C_{\mathcal{D}} > 0$. Differentiating Equation 14 yields

$$\dot{C}(t) = 4C_{\mathcal{D}} \widetilde{w}^3_{\text{sp}}(t) \dot{\widetilde{w}}_{\text{sp}}(t) < 0.$$

Thus, $C(t)$ is strictly decreasing along the AT trajectory. Since the gradient flow vector field is smooth and $R(t) > 0$, $C(t)$ is continuously differentiable and $-\dot{C}(t)$ is continuous on $[0, \infty)$.

Step 3: Existence of a finite rate-threshold time. Fix any $\gamma > 0$ and define

$$t^* := \inf \left\{ t \geq 0 : -\dot{C}(t) \leq \gamma \right\}.$$

Let

$$T_\gamma := \frac{C(0)}{\gamma}.$$

We show that the threshold condition must be satisfied no later than T_γ.

Suppose, to the contrary, that

$$-\dot{C}(t) > \gamma \qquad \text{for all } t \in [0, T_\gamma].$$

Integrating over $[0, T_\gamma]$ gives

$$\begin{aligned} C(T_\gamma) &= C(0) + \int_0^{T_\gamma} \dot{C}(t)\, dt \\ &= C(0) - \int_0^{T_\gamma} [-\dot{C}(t)]\, dt \\ &< C(0) - \gamma T_\gamma \\ &= 0. \end{aligned}$$

This contradicts the nonnegativity of CDV.

Therefore, there exists some $t \in [0, T_\gamma]$ such that $-\dot{C}(t) \leq \gamma$. Then the set defining t^* is nonempty, and

$$0 \leq t^* \leq T_\gamma = \frac{C(0)}{\gamma} < \infty.$$

By continuity of $-\dot{C}(t)$, its infimum is attained, so t^* is the first finite time at which the CDV decay-rate criterion is satisfied.

□

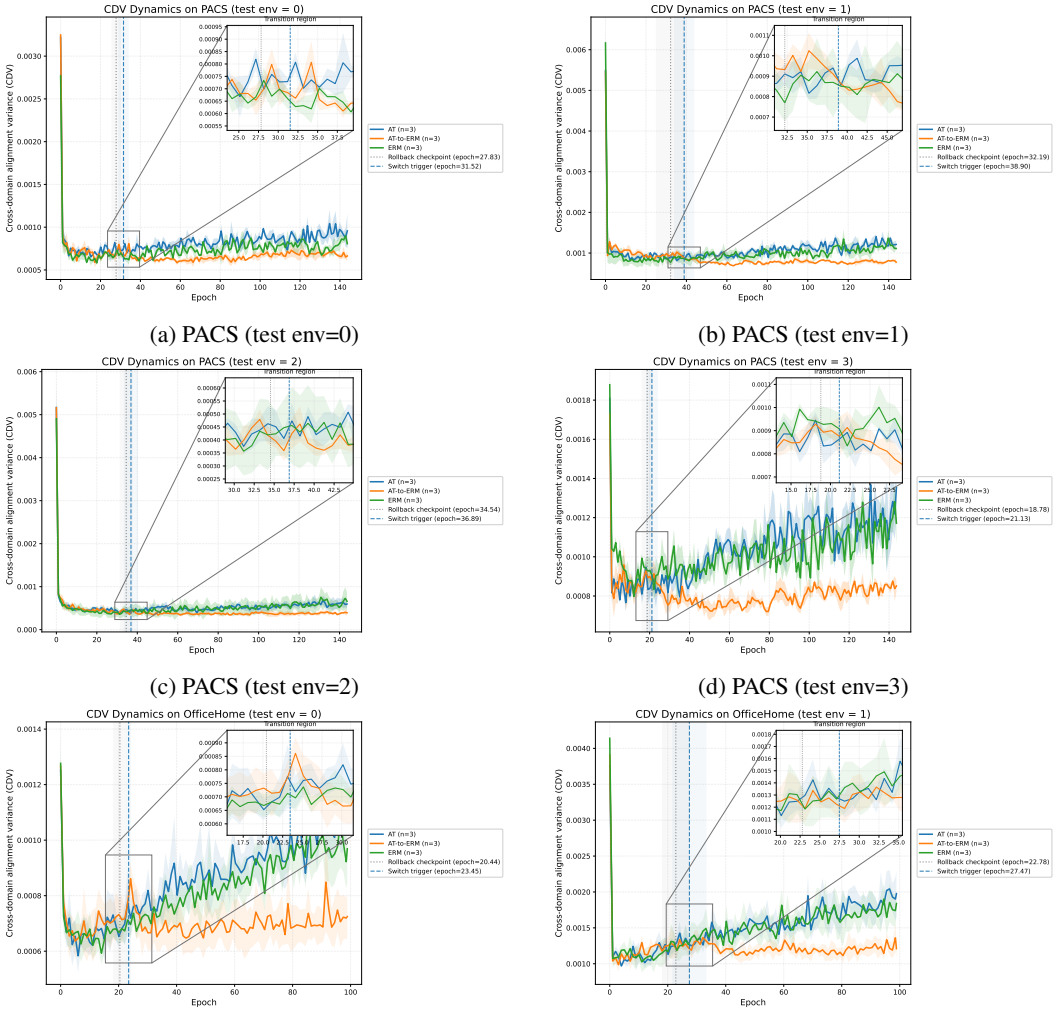
A.3 CDV-Guided Switching Hyperparameters

Table 8: Hyperparameters for the proposed CDV-guided AT-to-ERM switching rule. All other training configurations follow the corresponding backbone-specific settings.

Parameter	Description	Value
α	EMA coefficient for CDV smoothing	0.85
κ	Validation-accuracy ratio threshold	0.95
ρ	Relative tolerance for updating $b^\star$	0.01
δ	Relative-deterioration threshold	0.01
P	Patience for triggering the switch	2
$T_{\min}$	Minimum epoch before switching	5

A.4 CDV Dynamics across Training Methods

Results. As shown in Figures 2 and 3, CDV decreases rapidly during the early stage of training, gradually levels off, and then begins to increase. This trend is consistent with our theoretical result that CDV is proportional to the fourth power of the normalized spurious-feature coefficient. In particular, the early decrease in CDV reflects a reduction in the normalized spurious contribution, whereas the subsequent increase indicates a growing contribution from spurious features. After the switch from AT to ERM, AT-to-ERM shows a slower increase in CDV than either AT or ERM. This suggests that the proposed two-stage strategy better preserves the representation learned during the early AT phase and limits the subsequent increase in spurious feature contribution. Thus, these results support CDV as a reliable indicator for monitoring invariant representation learning during training.



(a) PACS (test env=0)

(b) PACS (test env=1)

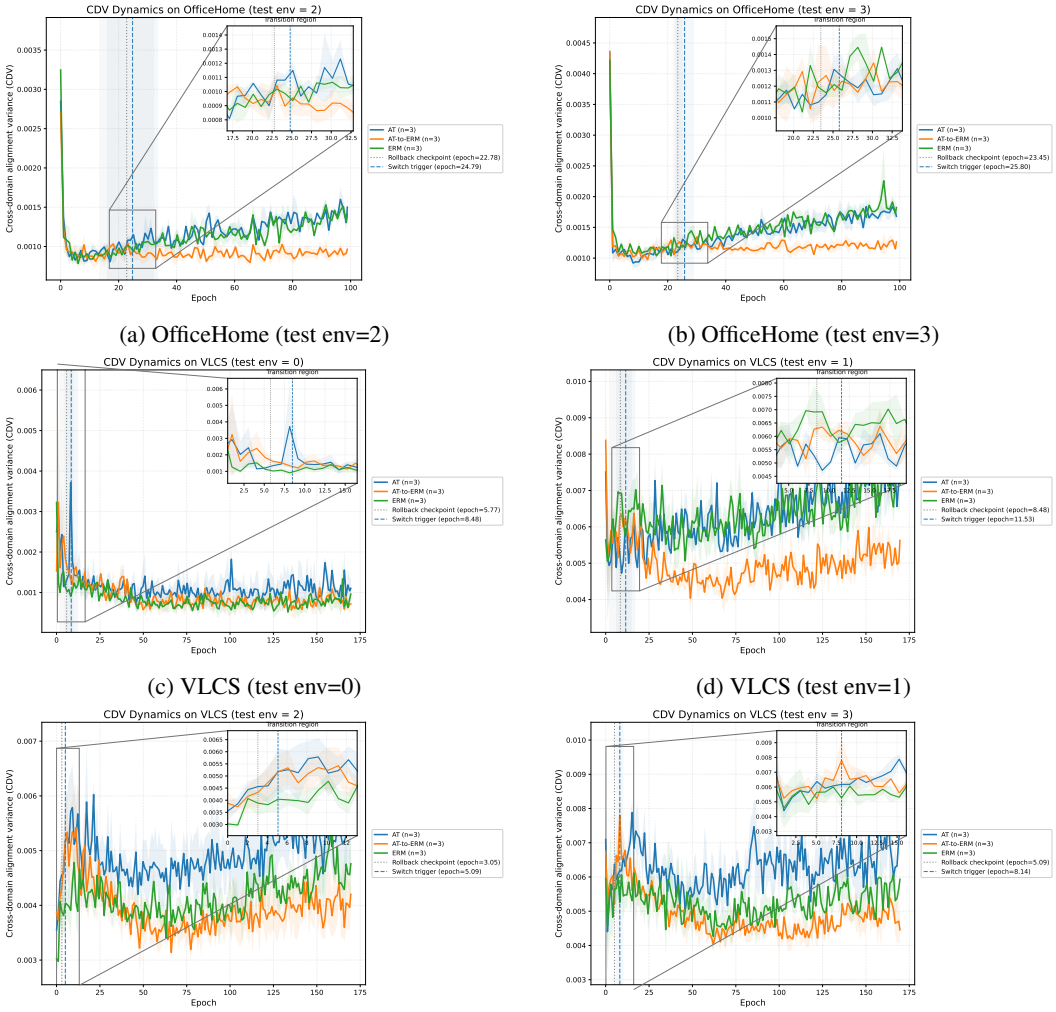
(c) PACS (test env=2)

(d) PACS (test env=3)

(e) OfficeHome (test env=0)

(f) OfficeHome (test env=1)

Figure 2: CDV dynamics during training across different methods, Part I. Results are shown for PACS and two test environments of OfficeHome. Shaded regions denote standard errors over three runs, and vertical lines indicate the rollback checkpoint and the AT-to-ERM switch trigger.



(a) OfficeHome (test env=2)

(b) OfficeHome (test env=3)

(c) VLCS (test env=0)

(d) VLCS (test env=1)

(e) VLCS (test env=2)

(f) VLCS (test env=3)

Figure 3: CDV dynamics during training across different methods, Part II. Results are shown for the remaining OfficeHome test environments and all VLCS test environments. Shaded regions denote standard errors over three runs, and vertical lines indicate the rollback checkpoint and the AT-to-ERM switch trigger.