Advanced Topics in Machine Learning - PA2

Ahmad Ashraf [1] **Sarim Malik** [2]

Abstract

This assignment explores Domain Adaptation (DA) and Domain Generalization (DG) techniques to improve model robustness under distribution shifts. We implement and evaluate state-of-the-art DA methods, including Deep Adaptation Network, Domain-Adversarial Neural Network, and Conditional Domain-Adversarial Network, which aim to minimize domain divergence between labeled source and unlabeled target domains. Building upon these insights, we extend our analysis to DG methods such as Invariant Risk Minimization, Sharpness-Aware Minimization, and robust optimization via Group DRO to enhance performance on unseen domains. Finally, we investigate CLIP-based prompt tuning using Context Optimization (CoOp) as a modern DA approach, leveraging vision–language representations for improved transferability.

Our codebase with complete experiments can be found here: https://github.com/ATML-AshrafxSarim/PA2 and backup: https://github.com/s-malix21/atml-fall2025.

1. Introduction

Domain shift is one of the most central challenges in deploying machine learning systems, where models are faced with data distributions at test time that differ significantly from their training distributions. A fundamental problem is that models trained using standard empirical risk minimization (ERM) can tend to learn spurious correlations, rather than pure predictive patterns, therefore when these models are deployed across different domains (e.g ML for autonomous vehicles but in varying weather conditions), they fail to maintain their training performance; underscoring a really major challenge of cross-domain robustness which persists in real world ML applications. This research aims to empirically analyze how domain adaptation and domain generalization techniques can minimize distribution shift challenges. The key research questions that we aim to tackle through this study are as follows:

- How can learning algorithms effectively adapt to new domains without labeled data; what factors influence the balance between domain alignment and task-specific performance?
- What training strategies enable models to generalize robustly across diverse source domains and maintain reliable performance on completely unseen target domains?
- How can multimodal vision-language models be adapted or guided to perform better in open-set environments, also what does this reveal about their capacity for generalization?

The overarching goal of this study is to delve deeper into how different algorithms and training strategies influence model robustness to domain shift. Also we aim to discover principles for building generalizable ML systems. Our experiments reveal that while techniques such as adaptation and invariant learning can mitigate domain gaps, prompt learning with CLIP introduces new trade-offs (in open-set recognition setting), highlighting the importance of calibration and confidence estimation.

2. Methodology

2.1. Unsupervised Domain Adaptation

2.1.1. Dataset Selection

The PACS dataset was chosen for the unsupervised domain adaptation task. It contains seven classes and approximately 9,991 images. The dataset includes four domains: art painting, cartoon, photo (real), and sketch, providing a diverse range of visual styles and features for testing domain adaptation methods. For all experiments, the source domain was set to art painting and the target domain to cartoon ($A \rightarrow C$). These domains were kept consistent across all experiments to ensure fair and accurate evaluation. For both domains, the data is split 80/20 into training and testing sets to enable balanced and reproducible evaluation.

2.1.2. Model Architecture and Hyperparameters

We utilize a ResNet-18 model implemented in *PyTorch* as the feature extractor and pretrained backbone, initialized with ImageNet-pretrained weights. For all experiments except pseudo-label training, the backbone learning rate is set to 0.001, while the classifier head uses a higher rate of 0.01 (due to its random initialization). Each domain adaptation method is fine-tuned for 20 epochs, a duration selected to allow minimax-based architectures to stabilize and yield comparable results. The feature extractor is connected to a linear classifier in all experiments, except where a specific classifier head configuration is required by the adaptation method. Finally, all experiments use stochastic gradient descent (SGD) with a momentum of 0.9 and Nesterov acceleration enabled.

2.1.3. UDA Techniques and Setup

To evaluate the effectiveness of our approach, we compare performance against several established domain adaptation baselines implemented using the *DALIB* library suite. These methods were selected to represent the different range of adaptation techniques. Each technique is built upon the same ResNet-18 backbone described earlier, ensuring fair and consistent comparison across experiments. The training setups for each approach are detailed below.

Source-only Baseline. The baseline model consists of our backbone followed by a linear classifier trained exclusively on the source domain data using cross-entropy loss. No adaptation losses or domain alignment modules are used. The total loss function for the baseline model is simply:

$$L_{\text{total}} = L_{\text{CE}}$$

where L_{CE} denotes the supervised cross-entropy loss computed on the source domain.

Domain Adaptation Network (DAN). DAN minimizes the domain shift by aligning the source and target feature distributions through the Multiple Kernel Maximum Mean Discrepancy (MK-MMD) loss ($L_{\text{MK-MMD}}$). The MK-MMD measures the distance between domain distributions in a reproducing kernel Hilbert space (RKHS) and is combined with the classification loss to jointly optimize for both class discrimination and domain alignment:

$$L_{\text{total}} = L_{\text{CE}} + L_{\text{MK-MMD}}$$

In our setup, we employ a series of Gaussian kernels with multiple bandwidths to capture distribution discrepancies at different scales. The bandwidth values are logarithmically spaced between 2^{-8} and 2^{8} with a step size of 0.5, ensuring broad coverage across kernel scales. The use of multiple kernels stabilizes training and avoids overfitting to a specific bandwidth range.

Following the original DAN paper, we introduce a bottleneck layer between the feature extractor and classifier head to constrain feature dimensionality and improve adaptation stability. This modification deviates slightly from the baseline but is essential for effective MK-MMD computation and convergence.

Domain Adversarial Neural Network (DANN). DANN introduces adversarial training to promote domain invariance. A binary domain discriminator is trained to distinguish whether a feature originates from the source or target domain, while the feature extractor simultaneously learns to confuse this discriminator through gradient reversal. This adversarial setup results in a minimax objective, combining the supervised and adversarial domain losses:

$$L_{\text{total}} = L_{\text{CE}} + L_{\text{dom}}$$

where L_{dom} is the binary cross-entropy loss from the domain discriminator.

The discriminator takes features from the backbone as input and passes them through a single hidden layer of size 1024, followed by a binary output for domain prediction. The learning rate for the discriminator utilized for training is set to 0.001.

Conditional Domain Adversarial Network (CDAN). CDAN extends DANN by conditioning the domain discriminator on both the extracted features and the classifier predictions. This conditioning captures the interaction between feature representations and class information, allowing the model to align conditional (class-specific) distributions across domains rather than marginal ones.

The input to the discriminator is formed by the outer product between the feature vector and the class prediction probabilities, encoding the class-conditional structure. The discriminator architecture mirrors that of DANN except for its input, with a hidden layer size of 1024. The total loss for CDAN+E is defined as:

$$L_{\text{total}} = L_{\text{CE}} + \mathbb{E}_{x \sim \mathcal{D}_t}[(1 + H(p(y|x)))\, L_{\text{dom}}]$$

where $H(p(y|x))$ denotes the prediction entropy, used to weight the domain loss such that high-confidence samples contribute more. This entropy conditioning improves stability by reducing the influence of uncertain target samples.

An inverse learning rate scheduler is also employed to balance the learning dynamics between the feature extractor and the discriminator, following the configuration proposed in the original paper. The learning rate employed for the discriminator is 0.0001 in order to ensure training stability.

Pseudo-label Training. Pseudo-label training is utilized as a self-training approach to domain adaptation, where the source-trained baseline model is used to generate labels for unlabeled target-domain samples. Target samples with prediction confidence above a predefined threshold are assigned pseudo-labels and used for further fine-tuning. In our experiments, the confidence threshold is set to 0.9, and fine-tuning is performed with a reduced learning rate of 2e-5 to minimize the impact of potential label noise and ensure stable adaptation.

2.1.4. t-SNE Visualizations

We use t-SNE visualizations to analyze feature representations after training and to assess whether domain invariance is achieved between the source and target domains. Ideally, the plots should show class clusters with minimal separation between domains. To quantify this, we also compute silhouette scores, which measure how well features from different classes are separated and whether domain gaps remain. Higher silhouette scores indicate better domain alignment and clearer class separation.

2.1.5. Concept Shift

To systematically evaluate the robustness of different domain adaptation methods, we design experiments that simulate label shift and rare-class scenarios. We induce these shifts by selectively removing certain classes and oversampling others within the target domain, thereby altering the target label distribution $P_t(Y)$ relative to the source $P_s(Y)$. This setup creates situations where the label distributions no longer align, challenging the adaptability of each method. Using these modified datasets, we train and evaluate **DAN**, **CDAN**, and **DANN**, assessing their performance in terms of overall and per-class accuracy. These experiments allow us to examine the limits of domain invariance and identify the conditions under which each method succeeds or fails.

2.2. Domain Generalization via Invariant & Robust Learning

2.2.1. Model Architecture, Setup, Datasets

For our primary experiments we have utilized the PACS dataset, which comprises 4 visual domains (art painting, cartoon, photo, sketch) with 9,991 total images across 7 object classes, designed to evaluate cross-domain generalization under style variations. We further simulate a domain generalization setup by utilizing the leave-one-out protocol on domains available, with 3 source domains art painting, cartoon, photo for training and sketch domain held out as unseen target for evaluation only.

We incorporated ResNet-18 backbone pretrained on ImageNet-1K with final classification layer adapted for 7-class PACS classification task. The training configuration for this backbone model remains consistent across all of our experiments to ensure fairness in the results, such that we use a 10 epoch training regime with Adam optimizer and learning rate set at 1e-4 and weight decay regularization also set at 1e-4. We also utilize the standard ImageNet normalization with center cropping to 224×224 resolution, to maintain consistency with pretrained weights.

2.2.2. ERM Baseline and Invariant Risk Minimization; Core Methods

Our baseline follows the empirical risk minimization (ERM) framework, where all source domain samples are pooled into a single training set. The model is optimized by minimizing the average cross-entropy loss

$$\min_{\theta} \frac{1}{|D_{source}|} \sum_{(x,y)\in D_{source}} \ell(f_\theta(x), y),$$

using gradient clipping for stability. Generalization performance is assessed through target domain accuracy

$$A_{target} = \frac{1}{|D_{target}|} \sum_{(x,y)\in D_{target}} \mathbf{1}[f(x) = y],$$

evaluated on an unseen domain without training exposure. To quantify the effect of domain shift, we compute the generalization gap

$$\Delta = A_{source} - A_{target}, \quad \text{where } A_{source} = \frac{1}{|S|} \sum_{s\in S} A_s.$$

Invariant Risk Minimization (IRM) extends this baseline by encouraging feature representations that lead to an optimal classifier invariant across domains. The IRM objective jointly minimizes the empirical loss and a gradient penalty,

$$\min_{\Phi,w} \sum_e \mathcal{L}_e(\Phi, w) + \lambda \sum_e \left\| \nabla_{w|w=1.0} \mathcal{L}_e(\Phi, w) \right\|^2,$$

where Φ denotes the feature extractor and w the classifier head. The penalty term

$$R_{IRM} = \sum_e \left\| \nabla_{w|w=1.0} \mathcal{L}_e \right\|^2$$

enforces gradient invariance across environments, thus it guides the model toward stable representations wrt causality. Monitoring of the IRM penalty alongside target accuracy helps identify trivial solutions where invariance holds but discriminative power is lost.

2.2.3. GROUP DRO AND SHARPNESS-AWARE MINIMIZATION; CORE METHODS

Group Distributionally Robust Optimization (Group DRO) aims to improve robustness across domains by minimizing the worst-case loss over source environments,

$$\min_{\theta} \max_{e \in \mathcal{E}} \mathcal{L}_e(\theta),$$

focusing optimization on the most challenging domain. Performance is evaluated using the worst-domain accuracy

$$A_{worst} = \min_{s \in S} A_s,$$

and training stability is promoted through exponential reweighting of domains,

$$w_e \leftarrow w_e \cdot \exp(\eta \cdot \mathcal{L}_e),$$

followed by normalization to upweight harder domains adaptively. Domain-level variance

$$\sigma^2_{domains} = \frac{1}{|S|} \sum_{s \in S} (A_s - A_{source})^2$$

is used to measure balance and fairness among source domains. Early stopping based on worst-case performance further prevents overfitting in deep architectures.

Sharpness-Aware Minimization seeks parameters that generalize better by finding flatter minima through a two-step optimization process. The objective is

$$\min_{\theta} \max_{\|\epsilon\| \leq \rho} \mathcal{L}(\theta + \epsilon),$$

where ρ controls the neighborhood radius. The first step locates the worst-case perturbation direction, and the second updates parameters using the perturbed gradients. Varying ρ allows analysis of the trade-off between sharpness and performance. We also viusalize loss sensitivity curves and normalized sharpness measures in order to compare the flatness of ERM and SAM solutions.

2.2.4. FEATURE REPRESENTATION ANALYSIS

For feature representation analysis, we utilize t-SNE visualizations, which do dimensionality reduction of learned feature representations to visualize domain clustering and class separability patterns. Silhouette scores are utilized for computing domain separation metrics, measuring feature space organization w.r.t domain identity versus semantic class structure.

For spurious correlation quantification, we domain silhouette scores as a proxy for spurious feature learning, and compare representation quality across methods. For invariance assessment, we utilize feature level analysis of cross-domain alignment, and target domain positioning relative to source distributions

2.2.5. EXPERIMENTAL DESIGN AND ABLATIONS

We also utilize a series of ablation studies and additional compariteve experiments for testing the affect of key parameters such as IRM penalty weights and SAM perturbation radii to characterize method robustness. We also conduct a head-to-head comparison of all methods using identical architectures, training procedures and evaluation protocols. For evaluating statistical significance we also test multiple runs on some experiment with different random seeds to ensure robustness of results and reliability.

2.3. Prompt Tuning with CLIP

2.3.1. DATASET SELECTION

The PACS dataset was chosen for the unsupervised domain adaptation task. It contains seven classes and approximately 9,991 images. The dataset includes four domains: art painting, cartoon, photo (real), and sketch, providing a diverse range of visual styles and features for testing domain adaptation methods. For all domains, the data is again split 80/20 into training and testing sets to enable balanced and reproducible evaluation.

2.3.2. ZERO-SHOT VS. LINEAR PROBING ON CLIP

We evaluate CLIP out-of-the-box on all PACS domains to assess its zero-shot generalization performance. This is compared with linear probing, where a linear classifier is attached to CLIP and only the classifier head is fine-tuned on the source domain. For linear probing, we use the Adam optimizer with a learning rate of 0.001 and a weight decay of 1e-4. This setup allows us to quantify the generalization performance of CLIP and any improvements obtained from fine-tuning the classifier head.

It must be noted that the prompts used for each domain are customized to reflect the domain-specific visual style, as shown in Table 1.

Table 1. Prompt Templates Used for Each PACS Domain in CLIP Evaluation

Domain	**Prompt Template**
Photo	"a photo of a {class}"
Sketch	"a sketch of a {class}"
Cartoon	"a cartoon of a {class}"
Art Painting	"an art painting of a {class}"

2.3.3. PROMPT TUNING (CONTEXT OPTIMIZATION)

We simulate a domain adaptation approach with CLIP using Context Optimization (CoOp), which injects a series of learnable context vectors into the prompt template. While CLIP's text and image encoders remain frozen, the context

vectors are optimized per epoch to improve performance, avoiding manual prompt engineering.

For our experiments, we reproduce the CoOp setup from the official GitHub repository and follow the hyperparameters reported in the original paper. The prompts are trained for approximately 10 epochs with a learning rate of 0.002 using SGD with momentum 0.9. Each trainable prompt consists of 16 context vectors, with the class token always appended at the end. CoOp is applied in two settings: (i) using only the source domain (*cartoon*), and (ii) including the target domain (*sketch*), where the target data is unlabeled and pseudo-label training is applied. This The combined training loss consists of the cross-entropy loss on the source domain (L_{CE}) and the pseudo-label loss (*Lpseudo*) on the target domain, if utilizing the target domain for CoOp.

2.3.4. Gradient Conflict and Alignment

We analyze gradient alignment during prompt training under the same conditions as in 2.3.3, using two source domains (*art painting* and *sketch*) and omitting the target domain. Specifically, we compute the gradient of the loss with respect to the prompt parameters for both source domains and calculate the cosine similarity between these gradient vectors. A cosine similarity greater than 0 indicates that the gradients point in the same direction, 0 indicates they are orthogonal, and less than 0 indicates opposing directions. Opposing gradients reveal a conflict, which can be harmful to performance on both source domains.

To mitigate this, we conduct an experiment using the baseline training setup, and a separate experiment applying GradCos, a method that re-aligns conflicting gradients to improve training stability. We then evaluate and compare the accuracies to determine whether GradCos improves performance.

2.3.5. Open-Set and Generalization Analysis

To evaluate the effect of prompt tuning on open-set performance, we design an experiment in which prompts are tuned on a closed set of classes (80% of the PACS classes) and tested on both seen and unseen classes.

We evaluate open-set performance using seen accuracy, unseen accuracy, and AUROC based on Maximum Softmax Probability (MSP). A higher AUROC reflects stronger separation between seen and unseen classes, which is preferred for open-set detection.

3. Results

3.1. Unsupervised Domain Adaptation

3.1.1. Domain Adaptation Techniques

Table 2 summarizes the results of various domain adaptation (DA) methods for the $(A \rightarrow C)$ transfer scenario. As expected, the source-only baseline achieves the highest accuracy on the source domain, but its target accuracy is very poor, since it receives no signal to guide learning on the target data.

Among the adaptation methods, CDAN achieves the best target performance with 77.61% accuracy, benefiting from its conditional alignment that leverages class-specific information while maintaining source performance close to the baseline. DANN also improves target accuracy compared to the baseline, but it significantly reduces source performance, likely due to the difficulty of balancing source-target alignment through the domain discriminator. Additionally, while DAN achieves comparable target accuracy, it is highly sensitive to the choice of kernel numbers and bandwidths, and the need for a bottleneck layer adds further training complexity.

Interestingly, pseudo-label training, a simpler self-training approach, also yields notable improvements on the target domain. However, self-training is sensitive and prone to confirmation bias: using overly confident pseudo-labels can reinforce spurious correlations and misguide the model. While pseudo-label training does not match the performance of state-of-the-art methods, these results highlight that self-training can be effective when the confidence threshold and learning rate are carefully tuned.

3.1.2. T-SNE Visualizations and Silhouette Scores

Table 3. Silhouette Scores for Source-Only Baseline and Domain Adaptation Models

Model	Silhouette Score
Source-Only Baseline	0.1612
DAN	0.3080
DANN	0.2626
CDAN	**0.3441**

After training, we extract features from the source and target domains using the same data loaders for each method. These features are projected into 2D using t-SNE for visualization, and silhouette scores are computed to quantify class separation and domain alignment. As shown in Table 3, CDAN achieves the highest silhouette score (0.3441), indicating the most well-separated and compact class clusters with minimal domain gaps. CDAN performs best because

Table 2. Performance of Domain Adaptation Methods on Source and Target Domains (% Accuracy). The **best target accuracy** is highlighted in bold.

Method	Source Accuracy (%)	Target Accuracy (%)	Average Accuracy (%)
Source Only Baseline	91.95	53.94	72.95
DAN	86.59	74.20	80.40
DANN	81.71	75.69	78.70
CDAN	87.32	**77.61**	82.47
Pseudo-label	84.39	60.77	72.58

its conditional adversarial alignment explicitly incorporates class information when matching source and target feature distributions, ensuring that features from different classes do not collapse together. DAN and DANN also improve over the source-only baseline, with DAN slightly outperforming DANN in cluster separation. These results align with the accuracy trends in Table 2, confirming that stronger domain alignment correlates with higher target performance. The t-SNE visualizations for all methods can be found in the Appendix A.1.1.

3.1.3. Concept Shift

Compared to the original accuracies (refer to Table 2), all models suffer from substantial drops in overall performance when exposed to severe label shift. This occurs because the removal of certain classes and oversampling of others heavily skews the target label representation, breaking the assumption of similar label distributions between the source and target domains.

Moreover, the imbalance introduced by missing and oversampled classes distorts the domain-level alignment. Since DANN and CDAN perform global rather than class-aware alignment, the source domain, which still contains all seven classes. is forced to align with an incomplete target label space. This global misalignment pulls the features of shared classes toward the domain mean influenced by missing-class source features, leading to cross-class confusion among visually similar or low-margin categories. Our results show that missing classes surprisingly still have better performance, while oversampled classes suffer from collapse. The per-class accuracies are shown in Appendix A.1.2. Overall, these results show that under label shifts, these domain adaptation techniques are not robust in terms of their performance.

Table 4. Performance of Domain Adaptation Methods under Severe Label Shift

Method	Overall Accuracy (%)
DANN	69.51
DAN	65.46
CDAN	67.38

3.2. Domain Generalization via Invariant & Robust Learning

3.2.1. ERM Baseline Performance and Spurious Correlation Analysis

The ERM baseline achieved strong source domain performance (91.95%, 95.74%, 98.20% on art painting, cartoon, photo respectively) but poor target generalization (61.59% on sketch), yielding a 33.71% generalization gap that evidences overfitting to source-specific features (Table 5).

Our t-SNE analysis revealed the underlying failure mechanism: ERM features exhibited strong domain clustering (domain silhouette: 0.2847) over semantic clustering (class silhouette: 0.1653), with poor target-source alignment (distance: 12.4387). This pattern reflects spurious correlation learning.ERM exploited photorealistic cues (color gradients, textures, lighting) present across all source domains but absent in sketches (Gulrajani & Lopez-Paz, 2021). The model learned predictive shortcuts that worked within the photorealistic distribution but failed catastrophically when these domain-specific features disappeared, demonstrating the need for invariant learning approaches(Gulrajani & Lopez-Paz, 2021).

Table 5. Comparison of ERM and IRM performance on the PACS dataset using a leave-one-domain-out setup (sketch as unseen target). IRM shows a smaller generalization gap and higher target accuracy.

Metric	ERM (%)	IRM (%)
Art Painting (Source)	91.95	90.73
Cartoon (Source)	95.74	91.90
Photo (Source)	98.20	98.20
Average Source	**95.30**	93.61
Target (Sketch)	61.59	**63.96**
Generalization Gap	33.71	**29.65**
Final Penalty (R_{IRM})	–	0.0005
Target Improvement	+2.37	
Source Trade-off	-1.69	

3.2.2. Invariant Risk Minimization (IRM)

IRM achieved 63.96% target accuracy with $\lambda = 1.0$ using variance-based penalty formulation, representing a 2.37% improvement over ERM baseline while maintaining reasonable source performance (93.61% average) (Table 5). Our ablation study revealed critical sensitivity to penalty weight selection—moderate values ($\lambda \in [0.1, 1.0]$) balanced invariance enforcement with discriminative performance, while high penalties ($\lambda = 100$) triggered model collapse to 23.43% source accuracy, indicating trivial constant predictions (Table 16) (Arjovsky et al., 2019).

Our t-SNE analysis confirmed IRM's effectiveness in learning invariant representations: domain silhouette scores decreased from 0.0471 (ERM) to 0.0183 (IRM), while target-source feature distance improved from 47.61 to 24.18, demonstrating reduced spurious correlation reliance and better cross-domain alignment (Table 6). However, IRM exhibited the expected source-target trade-off, sacrificing 1.69% average source performance to achieve invariant features that generalized better to the sketch domain's distribution.

Table 6. t-SNE based quantitative analysis comparing ERM and IRM feature representations on the PACS dataset.

Metric	ERM	IRM	Better
Domain Silhouette (↓)	0.0471	**0.0183**	IRM
Class Silhouette (↑)	**0.3097**	0.2980	ERM
Target–Source Dist. (↓)	47.61	**24.18**	IRM

3.2.3. Group DRO (Distributionally Robust Optimization)

Group DRO reached a target accuracy of 65.59%, showing a 4.00% gain compared to ERM. This improvement came from its adaptive domain weighting, which focused more on the most difficult source domain during training (Table 7). The model identified *art painting* as the hardest domain (final weight: 0.3747), as seen in the weight progression. However, this focus on the hardest domain reduced overall source performance, with average source accuracy dropping by 3.55% and the worst-case source accuracy by 4.88%.

Interestingly, Group DRO showed a kind of train-test mismatch: even though its worst-case source performance dropped during training, it still achieved better generalization on the target domain (Sagawa et al., 2020). This may indicate that focusing on the hardest source domain (art painting) helped the model learn features that transferred more effectively to the sketch target, even though sketch was not the weakest domain during training (Sagawa et al., 2020; Zhou et al., 2021).

The min–max optimization objective, $\min_\theta \max_{e \in \mathcal{E}} \mathcal{L}_e(\theta)$, reduced spurious correlations by discouraging reliance on domain-specific cues and encouraging the learning of more stable features that generalize across distribution (Zhou et al., 2021). The training dynamics for Group DRO were also visualized in (Figure 8) and the visualziations corroborate the theory well.

Table 7. Summary of Group DRO results compared to ERM on the PACS dataset.

Metric	ERM (%)	Group DRO (%)
Worst Source Domain	91.95	87.07
Best Source Domain	98.20	96.70
Gap (Best - Worst)	6.25	9.63
Average Source Accuracy	95.30	91.75
Target (Sketch) Accuracy	61.59	**65.59**
Worst-Case Change		–4.88
Domain Gap Change		–3.38
Average Source Change		–3.55
Target Improvement		+4.00

3.2.4. Sharpness-Aware Minimization (SAM) and Loss Landscape Analysis

Table 8. Comparison between SAM and ERM on PACS dataset. Sharpness-Aware Minimization (SAM) maintains comparable source performance while achieving significantly better target domain generalization and flatter loss landscapes

Metric	ERM (%)	SAM (%)	delta
art_painting	91.95	92.93	+0.98
cartoon	95.74	94.46	-1.28
photo	98.20	98.20	+0.00
Average Source	95.30	95.20	-0.10
Target (Sketch)	61.59	**67.70**	+6.11
Max Loss Increase	0.1860	0.0835	↓55.1%
Normalized Sharpness	-7.67	0.99	↓112.9%

SAM achieved the best target performance at 67.70% accuracy; there is a 6.11% improvement over ERM while maintaining source performance (Table 8). The two-step gradient procedure optimizing $\max_{||\epsilon|| \leq \rho} \mathcal{L}(\theta + \epsilon)$ successfully found flatter minima that generalize better across domains Foret et al. (2021).

Loss landscape analysis confirmed SAM's flatness advantage: 55.1% reduction in maximum loss increase and 112.9% improvement in normalized sharpness compared to ERM (Figure 1). This flatness explains the generalization improvementForet et al. (2021); since distribution shifts act like parameter perturbations, so flatter minima remain stable across domain boundaries.

Our ablation study showed sensitivity to perturbation radius

ρ (Table 15). Moderate values ($\rho \in [0.1, 0.2]$) worked best, achieving up to 69.18% target accuracy, while extreme settings failed. $\rho = 0.01$ provided insufficient flatness and $\rho = 1.0$ caused training collapse (52.80% source accuracy). SAM operates through optimization geometry rather than loss constraints, offering a complementary approach to IRM's invariance penalties and Group DRO's worst-case objectives Foret et al. (2021).

3.2.5. Comparative Analysis and Method Integration Discussion

SAM proved to be the most effective approach , reaching 67.70% target accuracy with almost no drop in source performance, showing that flat-minima optimization can improve domain generalization without reducing discriminative ability (Table 14). IRM and Group DRO achieved moderate gains, but both required trading some source accuracy for improved invariance, highlighting the inherent balance between generalization and discrimination (Gulrajani & Lopez-Paz, 2021).

SAM's strength lies in its optimization geometry. It seeks flatter minima through parameter perturbations, acting as an implicit regularizer against domain-specific overfitting Foret et al. (2021). In contrast, IRM enforces invariance through explicit loss constraints, and Group DRO focuses on reweighting domains to guard against worst-case shifts, which can sometimes limit flexibility.

ERM's baseline performance (61.59%) reflects how well-tuned models with proper regularization can still perform robustly. Yet, the consistent improvements from domain-aware methods confirm the value of explicitly accounting for robustness in domain generalization.

Combining methods offers a promising direction: SAM x IRM could help stabilize IRM's training and reduce sharp minima, while SAM x DRO might strengthen worst-case robustness without harming average accuracy (Sagawa et al., 2020). Since these methods operate at different levels; geometry (SAM), invariance (IRM), and domain weighting (DRO; their integration could enable more comprehensive cross-domain generalization.

3.3. Prompt Tuning on CLIP

3.3.1. Zero-Shot CLIP vs. Linear Probing on CLIP

Table 9 shows CLIP performance in zero-shot and linear probing settings. Zero-shot CLIP already achieves strong performance across all four PACS domains. When linear probing on the source domains *art painting, photo, cartoon*, small improvements are observed across all domains, suggesting that CLIP's broad pretraining provides a strong baseline.

Table 9. CLIP performance on PACS in zero-shot and linear probing settings. LP1: linear probing on Art Painting, Photo, and Cartoon; LP2: linear probing on Sketch, Photo, and Art Painting.

Domain	**Zero-Shot (%)**	**LP1 (%)**	**LP2 (%)**
Art Painting	94.39	95.61	95.37
Cartoon	97.65	99.36	97.65
Photo	100.00	100.00	99.70
Sketch	83.72	85.62	90.46
Average Acc.	93.94	95.15	95.79

When linear probing on *art painting, photo, sketch* with the target domain being *cartoon*, the overall average accuracy increases to 95.79%. However, we notice a slight decrease in accuracy for the *photo* domain, indicating that performance gains depend on the choice of source domains. Including a source domain that differs significantly from others (e.g., Sketch) can reduce generality and slightly degrade performance on some domains, even if overall accuracy improves.

3.3.2. Context Optimization (CoOp)

Table 10. Evaluation of CoOp on Source and Target Domains (Cartoon → Sketch)

Setting	**Source Accuracy (%)**	**Target Accuracy (%)**
CoOp (Source Only)	99.36	81.81
CoOp (Source + Target)	98.93	83.33

Table 10 shows the results of prompt tuning using CoOp. We observe that CoOp improves source-domain performance compared to zero-shot CLIP on Cartoon (97.31%). For the target domain, incorporating pseudo-label adaptation further increases accuracy from 81.81% to 83.33%, although it remains comparable to zero-shot CLIP and LP techniques without using any target labels.

There are, however, several considerations when using CoOp. First, the number of context vectors has a significant impact on performance: more vectors increase model capacity but also raise the risk of overfitting, so careful tuning is required (e.g., testing 4 or 8 context vectors to balance generalization). Second, while CoOp is more computationally efficient than full fine-tuning or linear probing, it can be brittle: the learned context vectors may overly specialize on certain features or styles within the source domain, reducing generalization to other domains. It is highlighted with the extreme accuracy difference between the target and source domains.

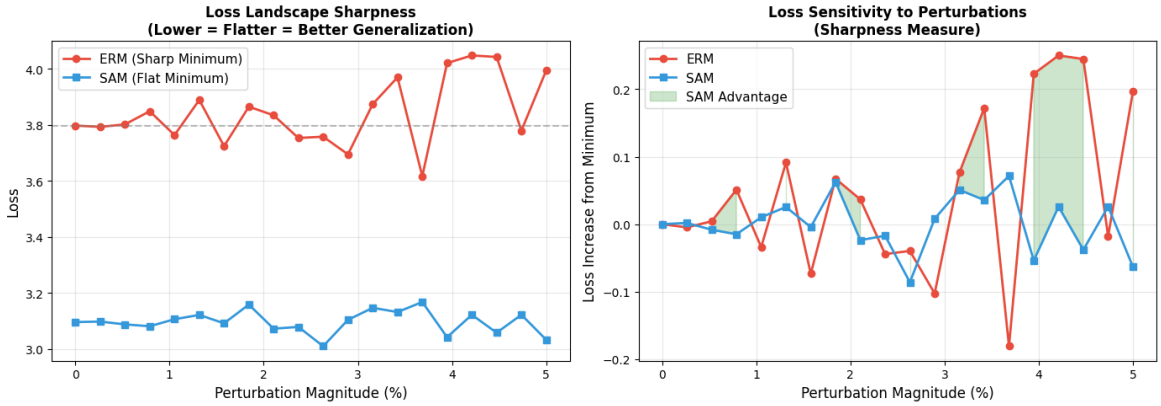

Figure 1. Comparison: SAM yields a flatter loss landscape than ERM, showing lower loss and reduced sensitivity to input perturbations.

Table 11. Evaluation of CoOp without and with GradCos on Two Source Domains (*art painting* and *sketch*)

Domain / Metric	**CoOp (w/o GradCos)**	**CoOp (with GradCos)**
Source 1 (Art Painting)	97.32	96.83
Source 2 (Sketch)	91.09	91.98
Average Accuracy	94.21	**94.41**

Table 12. Open-Set Evaluation of Zero-Shot CLIP and CoOp (Prompt Tuning) on Seen and Unseen Classes

Metric	**Zero-Shot CLIP**	**CoOp (Prompt Tuning)**
Seen Accuracy (%)	99.33	100.00
Unseen Accuracy (%)	100.00	86.73
Avg MSP (Seen)	0.983	0.998
Avg MSP (Unseen)	0.988	0.785

3.3.3. Gradient Conflict and Alignment on CoOp

Table 11 summarizes the evaluation of CoOp with pseudo-label adaptation on two source domains (Art Painting and Sketch). We observe that GradCos slightly improves the average accuracy (from 94.21% to 94.41%) by reducing conflicts in gradient directions between the two source domains, as indicated by the increased mean cosine similarity during training. While the source 1 accuracy decreases marginally, the improvement on source 2 compensates, demonstrating the effectiveness of GradCos in balancing multi-domain learning. It should also be noted that the improvement in average accuracy is marginal, likely because the two source domains (*art painting* and *sketch*) are somewhat similar. Larger improvements might be observed with more diverse source domains, such as *photo* and *sketch*, where greater stylistic differences could lead to stronger gradient conflicts and hence more benefit from GradCos.

3.3.4. Open Set Adaptation Analysis

Table 12 shows the open-set evaluation of Zero-Shot CLIP and CoOp on seen and unseen classes. Zero-shot CLIP achieves very high accuracy on both seen (99.33%) and unseen (100.00%) classes due to its broad pretraining.

With CoOp, seen class accuracy reaches 100.00% due to prompts specifically tuned for those classes. However, the accuracy on unseen classes drops to 86.73%, and the average MSP for unseen classes decreases to 0.785. This indicates that the model becomes overconfident on seen classes while less confident on unseen classes, making it more difficult to reliably detect unknown samples using MSP. The drop in unseen MSP highlights that prompt tuning can negatively impact the model's practical ability to identify truly unknown classes.

4. Discussion

4.1. Unsupervised Domain Adaptation

Overall, the domain adaptation methods demonstrate strong effectiveness in bridging the gap between source and target domains, as seen in Table 2. Conditional networks such as CDAN consistently outperform others, confirming that integrating class information into the alignment process enhances transferability. However, each approach has its own pitfalls: adversarial methods like DANN can struggle with training instability due to the minimax optimization, while kernel-based approaches such as DAN depend heavily on hyperparameter tuning, particularly in kernel bandwidths and bottleneck configurations.

Under concept shift scenarios, where the label distribution between domains diverges significantly, these methods exhibit some unexpectedly unstable behavior. Surprisingly,

despite their strong performance under standard conditions, global alignment strategies such as DANN and CDAN fail when the target domain lacks certain classes, as the models continue aligning features across an incomplete label space. This counterintuitive effect causes distorted feature representations and cross-class confusion, particularly among oversampled or visually similar categories, even though the missing classes sometimes appear to retain high accuracy. While conditional alignment provides marginal resilience, these findings reveal that conventional domain adaptation techniques are far less robust to label or concept shift than anticipated, underscoring the surprising brittleness of adversarial alignment and the need for more adaptive, class-aware strategies.

4.2. Cross-Domain Robustness Through Invariant Representation Learning

Our study highlights key trade-offs between invariance learning, optimization geometry, and discriminative performance that shape how models achieve cross-domain robustness. ERM's large generalization gap underscores how strongly neural networks rely on spurious visual cues such as texture and lighting, rather than underlying causal features.

IRM successfully reduced these correlations through gradient-based invariance penalties but remained highly sensitive to hyperparameter tuning; too strong a penalty often caused optimization collapse and loss of expressivity. Group DRO, by reweighting domains adaptively, exposed an implicit hierarchy in the PACS dataset: the most challenging source domain led to better target transfer, suggesting that domain difficulty during training can align with higher-level feature abstraction.

SAM outperformed other methods by regularizing the optimization landscape rather than altering the loss. Its flat-minima solutions improved robustness without explicit constraints, supporting the view that geometry-aware optimization can enhance generalization.

Overall, these methods reveal complementary mechanisms.IRM through invariant losses, DRO through distributional balancing, and SAM through geometric regularization. Future research could explore integrating these ideas within unified architectures that adaptively combine invariance, robustness, and flatness to address both data-driven and architectural biases in domain generalization.

4.3. Prompt Tuning on CLIP

Overall, the results demonstrate that CLIP's zero-shot performance remains remarkably strong across all domains in the PACS dataset, as shown in Table 9. This reinforces CLIP's ability to generalize to new visual styles through large-scale pretraining on image–text pairs. However, despite this robustness, zero-shot performance can still benefit from domain adaptation, especially when target domains exhibit distinct style or distributional differences. Direct fine-tuning of CLIP is computationally expensive and risks catastrophic forgetting, making prompt tuning an attractive alternative due to its efficiency and minimal parameter overhead.

Prompt tuning methods such as CoOp (Table 10) adapt CLIP to specific source domains by learning a small number of context vectors. While effective in improving in-domain accuracy, these models tend to overfit, as the learned prompts may capture domain-specific biases rather than domain-invariant features. This brittleness becomes more pronounced in multi-domain adaptation setups, where gradient conflicts emerge between domains with distinct visual characteristics. As illustrated by the modest improvements in Table 11, techniques like GradCos aim to mitigate such conflicts by aligning gradient directions across domains. However, in scenarios with multiple diverse domains, these conflicts can cause instability in optimization, leaving the model in a partially unlearned state or suppressing gradients that encode useful discriminative information.

Finally, Table 12 highlights a key limitation of prompt tuning in open-set adaptation. While CoOp achieves perfect seen-class accuracy, its performance on unseen classes drops substantially (from 100.00% to 86.73%), and the average MSP on unseen samples decreases sharply to 0.785. This suggests that the tuned prompts cause the model to become overconfident on the seen label space while failing to recognize truly novel samples. In essence, prompt tuning narrows the feature space around seen class embeddings, reducing flexibility and impairing the model's ability to generalize under open-set or out-of-distribution conditions.

5. Conclusion

Our study reveals that domain robustness arises from distinct yet complementary mechanisms—alignment, invariance, and optimization geometry. In unsupervised adaptation, class-conditional methods like CDAN outperformed adversarial baselines but failed under concept shift, exposing the fragility of global alignment. In domain generalization, Sharpness-Aware Minimization (SAM) achieved the highest target accuracy (67.7%), showing that geometry-aware optimization can enhance robustness without explicit invariance constraints, while IRM and Group DRO traded source accuracy for stability. Prompt-based adaptation in vision-language models highlighted a similar balance between specialization and generalization: Context Optimization (CoOp) improved in-domain accuracy but reduced unseen class performance. Overall, our findings suggest that achieving cross-domain robustness requires integrating insights from alignment, invariance, and geometric optimization to

balance adaptability and generalization in future architectures.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2019.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. S. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1911.08731.

Zhou, K., Liu, Z., Qiao, Y., and Xiang, T. Examining and combating spurious features under distribution shift. *arXiv preprint arXiv:2106.07171*, 2021. URL https://arxiv.org/abs/2106.07171.

A. Appendix

A.1. Supplementary Material - Unsupervised Domain Adaptation

A.1.1. T-SNE Visualizations

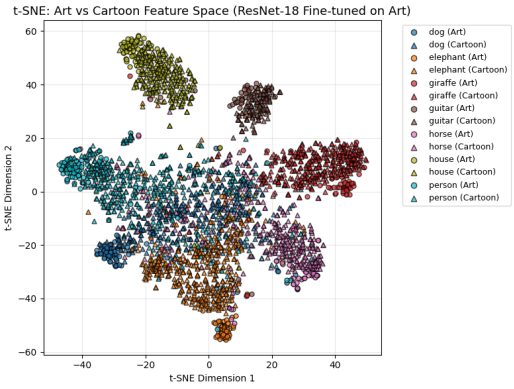

Figure 2. t-SNE visualization of feature embeddings for the **Source Only** model.

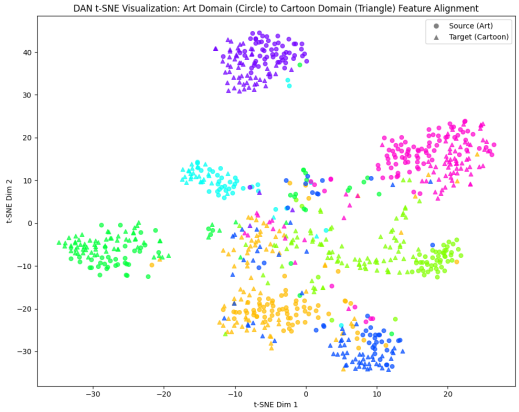

Figure 3. t-SNE visualization of feature embeddings for the **DAN** model.

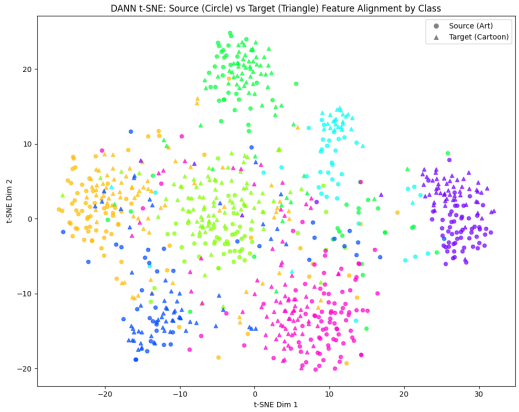

Figure 4. t-SNE visualization of feature embeddings for the **DANN** model.

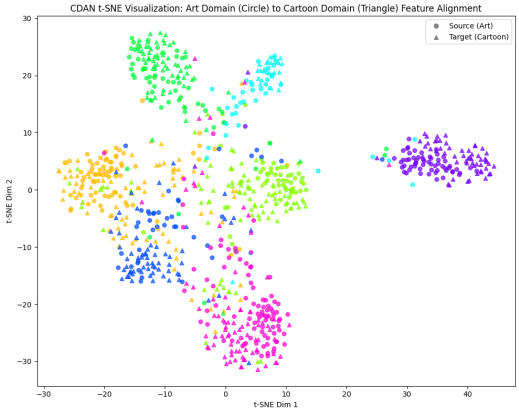

Figure 5. t-SNE visualization of feature embeddings for the **CDAN** model.

A.1.2. Concept Shift (Class Accuracies)

Table 13. Per-Class Accuracies (%) of Domain Adaptation Methods under Severe Label Shift

Class	DANN	DAN	CDAN
Dog (0)	65.38	44.87	64.10
Elephant (1)	26.37	20.88	18.68
Giraffe (2)	84.06	85.51	89.86
Guitar (3)	92.59	96.30	92.59
Horse (4)	55.38	63.08	63.08
House (5)	94.83	94.83	94.83
Person (6)	86.42	82.72	81.48

A.2. Supplementary Material - DG via Invariant & Robust Learning

Table 14. Summary of domain generalization performance across methods

Method	Avg Src	Worst Src	Target	T↑ vs ERM
ERM	95.30	91.95	61.59	–
IRM	93.61	90.73	63.96	+2.37
GDRO	93.87	88.54	65.59	+4.00
SAM	95.20	92.93	67.70	+6.11

Table 15. Ablation results for SAM with different sharpness radius (ρ) values. Moderate perturbation radii (0.1–0.2) yield the best trade-off between source performance and target generalization, while extreme ρ values cause instability and degradation.

ρ	Source Avg (%)	Target (%)
0.01	95.40	62.92
0.05	95.54	66.71
0.10	96.28	**67.75**
0.20	95.61	69.18
0.50	94.59	65.84
1.00	52.80	19.55

Table 16. **IRM Ablation Study across λ values.** Higher λ overly constrains invariance, degrading both source and target performance. Moderate λ (0.1–1.0) balances penalty and generalization.

λ	Avg Src (%)	Target (%)	art	cartoon	photo
0.1	93.49	64.47	91.22	91.04	98.20
1.0	93.24	62.81	88.78	92.75	98.20
10.0	79.21	54.21	75.37	74.84	87.43
100.0	23.43	23.72	22.44	22.39	25.45

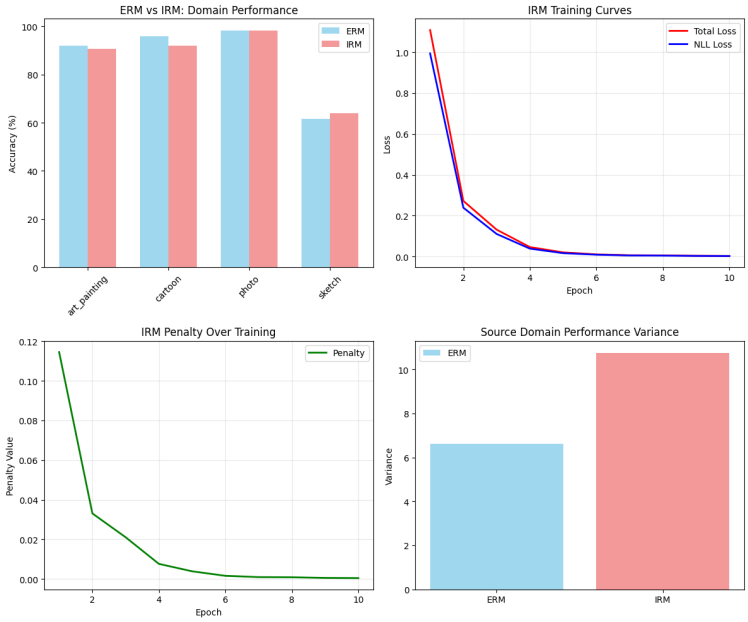

Figure 6. Comparison of ERM and IRM on domain generalization: ERM achieves higher source accuracy but larger generalization gap, while IRM reduces domain variance and improves target performance.

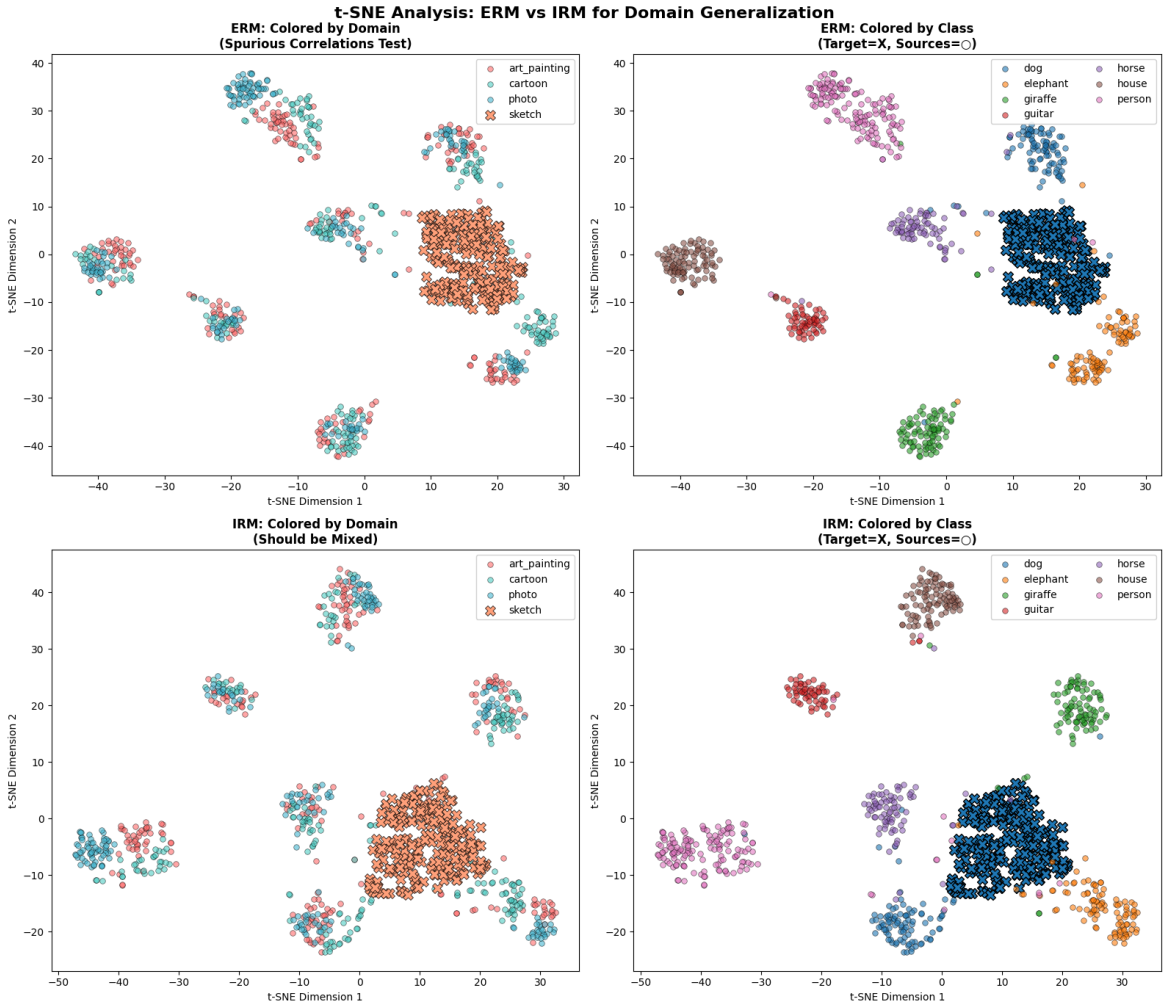

Figure 7. t-SNE visualizations compare ERM and IRM feature embeddings across PACS domains, highlighting domain clustering and class separation. IRM reduces domain-specific clustering and improves class-based grouping, indicating more invariant representations and better domain alignment.

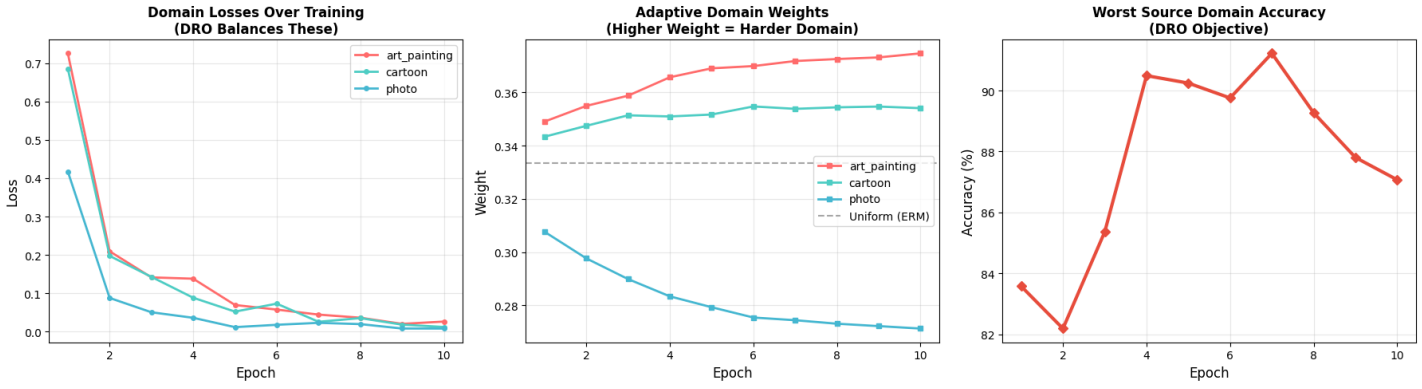

Figure 8. Group DRO training dynamics: the model adaptively increases weights for harder domains, balances domain losses, and improves worst-case source accuracy over epochs.