Techniques for Domain Adaptation and Generalization: An Empirical Study

Muhammad Sabeer Asad Muhammad Bin Tariq Raahim A Samad Poonawala

Abstract

Domain Adaptation and Domain Generalization (DA/DG) are open problems in machine learning that seek to address distribution shifts between training and inference data. This work presents an empirical investigation of several DA/DG techniques, and their robustness to domain shifts. We also evaluate the efficacy of these techniques by studying the computational resources required to achieve said performance. We also explore a cutting edge approach to DA/DG using CLIP and prompt learning as a control knob to adapt to new domains. Our results mostly agree with established literature and provide insights into the trade-offs between adaptation and generalization. All our code can be found here.

1. Introduction

The promise of machine learning relies heavily on the i.i.d. assumption, that training and test data originate from identical distributions. In real-world scenarios, this assumption is routinely violated, resulting in degraded model performance under distribution shift. Domain adaptation (DA) addresses this challenge by transferring knowledge from labeled source domains to related but distinct target domains, often with limited or no target labels. Domain generalization (DG) extends this goal by learning representations from multiple source domains that generalize to entirely unseen target domains. Both paradigms share a common theoretical foundation (Ben-David et al., 2010), emphasizing that cross-domain generalization depends on minimizing both source error and inter-domain divergence.

This work presents a systematic empirical investigation across the DA and DG landscape. We implement and compare major unsupervised adaptation paradigms, including statistical alignment (DAN), adversarial alignment (DANN, CDAN), and self-training, under consistent experimental protocols using the PACS benchmark. For domain generalization, we explore invariant representation learning (IRM), robust optimization (GroupDRO), and loss-flatness regularization (SAM), benchmarking each against the standard Empirical Risk Minimization (ERM) baseline.

Finally, we extend our analysis to large vision–language models through prompt learning with CLIP and CoOp, examining how context optimization facilitates both adaptation and generalization in foundation models. Across all settings, we evaluate the alignment–discrimination trade-off, robustness to class and label imbalance, and the impact of domain shift nature, offering a unified perspective on the mechanisms driving cross-domain robustness.

2. Methodology

Our experimental investigation focuses on unsupervised domain adaptation across stylistically diverse domains. This section describes the experimental infrastructure, datasets, evaluation protocols, and training procedures.

2.1. Unsupervised Domain Adaptation

2.1.1. Experimental Setup

We employ the PACS dataset from HuggingFace[1], containing 9,991 images across 7 object categories in four visually distinct domains: Art Painting, Cartoon, Photo, and Sketch. Following standard domain generalization protocols adapted for domain adaptation, we use three source domains (Art Painting, Cartoon, Photo) concatenated as labeled training data, and Sketch as the unlabeled target domain.

We split each domain 80-20 train-test with stratified sampling. All experiments employ ResNet-50 (He et al., 2016) pretrained on ImageNet as the feature extractor (2048-dim output) with a fully-connected classifier. Images are resized to 224×224 and normalized using ImageNet statistics. We use Adam optimizer (Kingma & Ba, 2014) with learning rate 10^{-4}.

We report target domain test accuracy (primary metric), source domain test accuracy (detect negative transfer), macro-averaged F1 scores (class-balanced evaluation), and proxy A-distance computed by training a linear classifier to distinguish source from target features: $\hat{d}_{\mathcal{A}} = 2(1 - 2\epsilon)$ where ϵ is held-out classifier error.

[1] https://huggingface.co/datasets/flwrlabs/pacs

2.1.2. Source-Only Baseline

The source-only baseline trains exclusively on labeled source data via empirical risk minimization for 5 epochs, establishing the no-adaptation reference point. This quantifies raw domain shift magnitude without any adaptation (Ben-David et al., 2010).

2.1.3. Deep Adaptation Network (DAN)

DAN (Long et al., 2015) minimizes Maximum Mean Discrepancy (MMD) between source and target feature distributions. We employ multi-kernel MMD with five Gaussian kernels, setting bandwidths adaptively using median heuristic. The adaptation weight follows progressive schedule $\lambda_p = \frac{2}{1+\exp(-10p)} - 1$ where $p \in [0, 1]$ represents training progress. We train for 10 epochs with base MMD weight 1.0.

2.1.4. Domain-Adversarial Neural Network (DANN)

DANN (Ganin & Lempitsky, 2015) learns domain-invariant features through adversarial training. A domain discriminator classifies whether features come from source or target, while the feature extractor produces features that fool the discriminator via gradient reversal layer (GRL) (Ganin et al., 2016). Our discriminator has three fully-connected layers $[2048 \to 1024 \to 1024 \to 1]$ with ReLU and 50% dropout. We use progressive schedule λ_p for adversarial weight and GRL strength $\alpha = \lambda_p$. Training runs for 10 epochs.

2.1.5. Conditional Domain Adversarial Network (CDAN)

CDAN (Long et al., 2018) conditions domain alignment on predicted class labels to enable class-specific adaptation. The discriminator receives concatenated features and class predictions $[h; p] \in \mathbb{R}^{2048+7}$, allowing class-aware alignment that preserves discriminative structure while reducing within-class domain gaps. We train for 10 epochs.

2.1.6. Self-Training with Pseudo-Labeling

We implement iterative self-training (Lee, 2013) as an alternative to explicit alignment. First, train on labeled source data (5 epochs). Second, generate predictions on unlabeled target samples and filter using confidence threshold $\tau = 0.7$. Third, fine-tune on high-confidence pseudo-labeled target data for 10 epochs with learning rate 10^{-4}, mixing source samples with weight 0.6 to prevent catastrophic forgetting. We freeze the backbone during fine-tuning to stabilize training.

2.1.7. Concept Shift Scenarios

To evaluate robustness under assumption violations, we introduce two concept shift scenarios. The *rare class scenario* undersamples class 3 to 20% of its original frequency in the target test set. The *label shift scenario* completely removes classes $\{2, 3, 5\}$ from the target test set. Both evaluate previously trained models without retraining, assessing zero-shot robustness.

2.2. Invariant and Robust Domain Generalization

2.2.1. Dataset and Setup

We use the PACS dataset with four domains: *Art Painting, Cartoon, Photo, Sketch*. The first three serve as source domains for training and validation, while the unseen *Sketch* domain is used exclusively for testing. The backbone is ResNet-50 pre-trained on ImageNet, optimized using Adam with a learning rate of 1×10^{-4} for 5 epochs per method.

2.2.2. Algorithms

ERM: Standard empirical risk minimization with cross-entropy across all combined source domains.

IRM: Enforces invariant representations by penalizing the variance of optimal classifiers across domains. Regularization strength $\lambda \in \{1, 10, 100\}$.

GroupDRO: Minimizes the worst-domain loss by up-weighting domains with higher losses. GroupDRO Weight $\eta_q = 0.1$

SAM: Minimizes the maximum loss under small perturbations of weights, promoting flatter minima and better generalization. Maximum perturbation $\rho = 0.05$

2.3. Domain Adaptation and Generalization via Prompt Learning

We employed OpenAI's vision-language model CLIP (Contrastive Language–Image Pre-training) (Radford et al., 2021) for all experiments, using the ViT-B/32 variant and the PACS dataset from HuggingFace. Four experiments were conducted to evaluate prompt learning for domain adaptation (DA) and generalization (DG): (1) CLIP Zero-Shot vs Linear Probing, (2) Prompt Learning with CoOp, (3) Gradient Alignment across Domains, and (4) Open-Set Evaluation.

2.3.1. CLIP Zero-Shot vs Fine-Tuned on Domains

We first evaluated CLIP's zero-shot performance on PACS as a baseline. Using the text encoder, we generated:

- Simple Prompts (SP): "a photo of a {class_name}"
- Domain-Specific Prompts (DSP): "a {domain_name} of a {class_name}"

Zero-shot accuracy was measured across all four PACS domains. For comparison, we trained a linear classifier on frozen CLIP image features for each domain (linear probing).

2.3.2. Prompt Learning

For domain adaptation, we applied Context Optimization (CoOp) (Zhou et al., 2022) using a unified set of 16 learnable prompt (context) vectors optimized jointly across all source domains. Prompts were initialized randomly and trained with SGD (lr=0.002) for 10 epochs. For domain generalization, pseudo-labels from the target domain were incorporated using an unsupervised loss on target data.

2.3.3. Gradient Analysis across Domains

To study gradient alignment under prompt learning, we computed cosine similarity of gradients between two source domains (Art Painting and Photo) and between a source (Photo) and the pseudo-labeled target (Sketch). Measurements were taken every 5 epochs to track how alignment evolved during training.

2.3.4. Open-Set Evaluation

Finally, we simulated an open-set scenario by training prompts on 80% of PACS classes and evaluating on the remaining unseen classes in the target domain to assess generalization to novel categories.

3. Results

3.1. Unsupervised Domain Adaptation

3.1.1. Baseline and Method Comparison

Table 1 shows the source-only baseline, quantifying a 35.04% performance gap due to substantial style shift. Table 2 compares all methods, revealing striking results: DANN achieves the best target accuracy, followed by CDAN and DAN. Surprisingly, pseudo-labeling shows modest improvement, dramatically underperforming alignment methods.

Domain	Accuracy (%)	Drop (%)
Source Domains Test	95.47	—
Target (Sketch) Test	60.43	35.04

Table 1. Source-only baseline on PACS multi-source to Sketch transfer.

Figure 1 visualizes per-class accuracy across all 7 classes. All methods show relatively consistent performance across classes, with classes 5 and 6 achieving near-perfect accuracy (dark green). DANN and DAN show the most consistent green coloring, while source-only and pseudo-labeling exhibit more red regions on challenging classes (0, 1, 2). Class 2 appears particularly difficult across all methods.

Method	Source (%)	Target (%)	Gain (%)
Source-Only	95.47	60.43	0.00
DAN	90.02	66.16	+5.73
DANN	92.42	**74.81**	**+14.38**
CDAN	88.95	70.48	+10.05
Pseudo-Label	90.27	62.47	+2.04

Table 2. Performance comparison across methods on PACS.

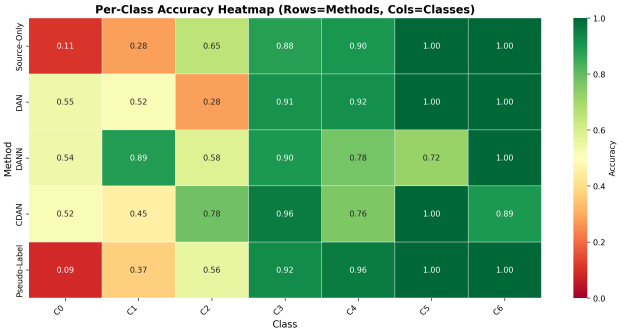


Figure 1. Per-class accuracy heatmap showing DANN achieves most consistent strong performance across all 7 classes, with fewer failure modes than other methods.

3.1.2. Domain Divergence Analysis

Table 3 presents proxy A-distance measurements. Remarkably, all methods achieve nearly identical A-distance, indicating that domain alignment quality does not explain performance differences.

Method	Classifier Acc	A-distance
Source-Only	0.9980	1.9960
DAN	0.9980	1.9960
DANN	0.9860	1.9720
CDAN	0.9960	1.9920
Pseudo-Label	0.9980	1.9960

Table 3. Proxy A-distance measurements showing near-identical alignment across all methods.

3.1.3. Concept Shift Robustness

Tables 4 and 5 show performance under assumption violations. Under rare class scenario, DANN maintains its advantage, with all methods showing graceful degradation. Pseudo-labeling achieves highest rare-class F1 (0.8378) despite lower overall accuracy, suggesting its confidence filtering successfully identifies and adapts to minority samples.

Method	Accuracy (%)	F1-Avg
Source-Only	56.01	0.6137
DAN	62.81	0.6660
DANN	**72.79**	**0.7375**
CDAN	67.15	0.6602
Pseudo-Label	57.31	0.6405

Table 4. Performance under rare class scenario (class 3 undersampled to 20%); F1-avg is over all classes.

Method	Accuracy (%)	F1-Avg
Source-Only	50.40	0.2494
DAN	70.16	0.3856
DANN	**76.09**	**0.4219**
CDAN	61.26	0.3624
Pseudo-Label	54.74	0.2805

Table 5. Performance under label shift (classes 2, 3, 5 removed from target); F1-avg is over all classes.

Under label shift, DANN again dominates, with DAN second. Interestingly, performance increases compared to standard setting for alignment methods, likely because removing challenging classes simplifies the task. The removed classes appear in Figure 2 as completely dark red columns, with models unable to predict absent categories. F1-macro scores drop substantially for all methods, reflecting the difficulty of the reduced label space.

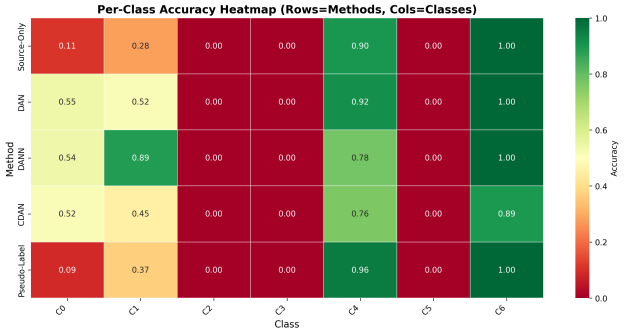


Figure 2. Per-class accuracy heatmap under label shift showing dark red columns for removed classes (2, 3, 5) as models cannot predict absent categories.

Figure 3 (left) shows F1 scores for the three rarest classes (5, 6, 3) in the PACS target test set. Source-only and pseudo-labeling achieve higher F1 on these classes, while alignment-based methods (DANN, DAN) yield more balanced but slightly lower scores—consistent with prior observations that excessive marginal alignment can hurt minority categories under label-prior shift (Zhao et al., 2019). As shown in Figure 3 (right), pseudo-labeling lies above the diagonal (rare classes benefit), while CDAN lies below (rare classes hurt). This pattern suggests that pseudo-labeling's confidence filtering preserves rare-class performance, even if it fails to exploit target structure broadly.

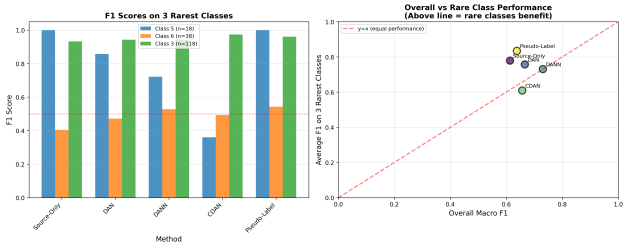


Figure 3. F1 scores on naturally rarest classes (left) and overall vs rare-class F1 comparison (right). Pseudo-labeling excels on rare classes despite lower overall accuracy.

3.2. Invariant and Robust Domain Generalization

3.2.1. Quantitative Comparison

Method	**Val. Avg**	**Target Acc.**
ERM	97.39%	67.19%
IRM ($\lambda = 10$)	92.32%	**75.41%**
GroupDRO	94.65%	72.82%
SAM	**98.27%**	73.43%

Table 6. Average validation and target-domain accuracies. IRM achieves the best target-domain generalization.

Method	**Art Painting**	**Cartoon**	**Photo**
ERM	94.63	96.16	99.40
IRM ($\lambda = 10$)	91.22	86.35	99.40
GDRO	93.41	90.83	99.70
SAM	96.83	97.87	99.40

Table 7. Per-source-domain validation accuracies on PACS dataset.

All the domain generalization methods performed better on an unseen target domain than the baseline ERM, and suffered a drop in source-domain validation accuracy. These results agree with established literature. While IRM achieved the best target domain performance, we note that SAM improved over ERM in both target *and source* domains. This is also expected, and we elaborate on the reasons in the Discussion section.

3.2.2. SAM vs ERM Flatness Analysis

The flatness of the region in the loss landscape where a model converges can be extrapolated as a measure of the model's generalizability. SAM is the method that tries to achieve generalization through this hypothesis. Figure 4

shows the how the loss function changes with small perturbations around the region where ERM-trained and SAM-trained model converged. ΔL for SAM is always smaller than ΔL for ERM, except for the discontinuous region in the middle – this only further shows how undulating and sensitive to perturbations the ERM converged loss is. SAM optimizer is therefore successful in finding a flatter loss region to converge the model to

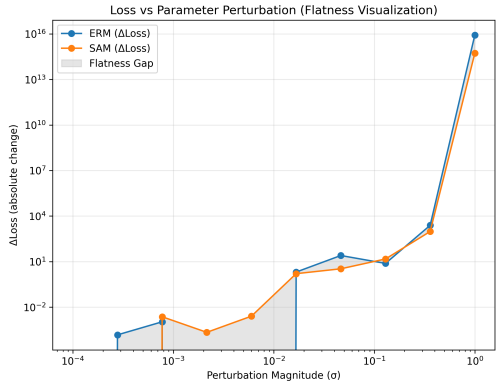


Figure 4. Comparison of sharpness measures between ERM and SAM.

Figures 5 explicitly visualizes the landscape, and it is apparent that with the SAM optimizer, the model converged to a flatter loss compared to ERM.

3.2.3. GroupDRO Domain Weight Evolution

Under the **Group Distributionally Robust Optimization** paradigm, generalization is approached by inducing domain invariance via making the model more robust against the domain it performs worse on. We can see in Figure 6 that the model learns to weight the domains differently across epochs. This adaptive reweighting during training allows the model to more robustly learn domain-invariant features that are expected to exist in an unseen target domain as well. Thus, the target domain performance improves compared to ERM.

3.2.4. IRM Penalty Weight Effect

Invariant Risk Minimzation method tries to induce domain invariance by minimizing the risk of a model given a domain-invariant task head. A domain-invariant task head is a common task head that minimizes the risk across all domains. This makes IRM a double-minimization problem that is intractible. The approximation used by (Arjovsky et al., 2020) makes convergence extremely sensitive to stationary points in the loss landscape.

Figure 7 shows how accuracy changes with different values of λ as the weight of the invariance penalty. Across multiple experiments, we found that $\lambda = 10$ amortized over

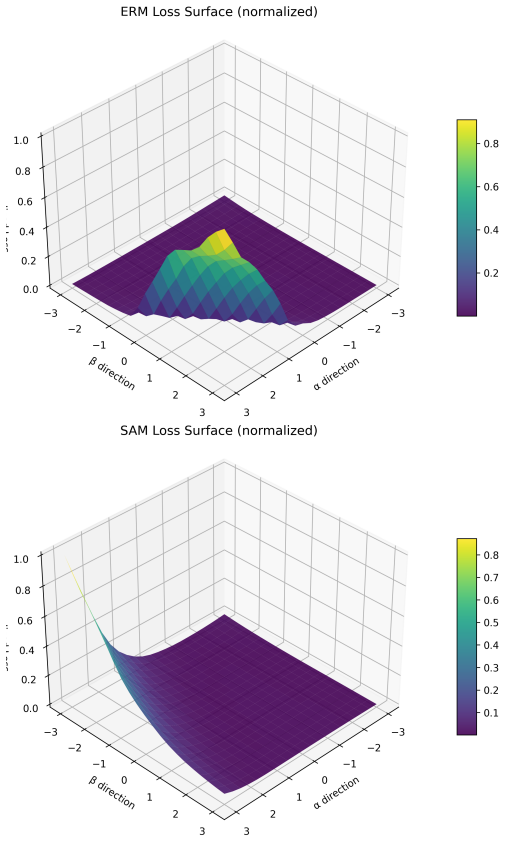


Figure 5. 3D visualization confirming flatter minima in SAM's loss landscape.

the `batch_size` being processed gave the best accuracy. While higher values of λ did reduce the IRM penalty, the model accuracy suffered. This suggests that for too high a weight, training collapse. However, for $\lambda = 10$, figure 8 shows how the penalty is appropriately minimized while our model retains high accuracy on target domain (in fact, the highest), so training did not collapse.

3.2.5. Feature Representation Visualization

Figure 9 shows t-SNE projections of learned features from each method (final logits just before classification), colored by domain and markers shaped by class. ERM exhibits clear domain clustering, indicating that it learned domain-dependent features. IRM and GroupDRO exhibit an attempt to induce cross-domain alignment, which increases generalizability. However, we also see cross-class mixing which may hurt classification performance. SAM also tries to align domains, though to a lesser degree than IRM and gDRO, but maintains tighter class separability. This may explain its improved performance both on source and target domains (over ERM)

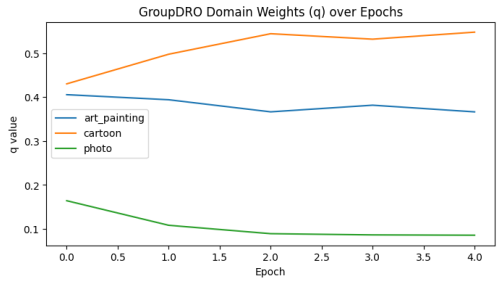


Figure 6. Evolution of domain weights in GroupDRO. Harder domains receive increasing importance over epochs.

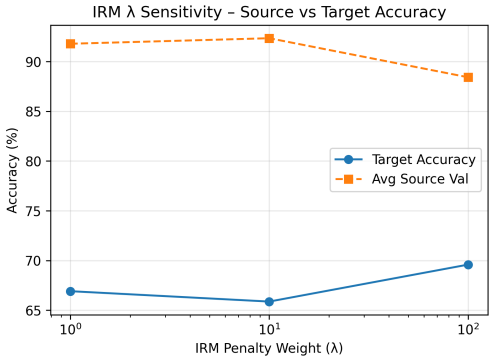


Figure 7. Effect of IRM penalty weight λ on validation and target accuracies. Moderate λ values yield optimal generalization.

3.3. Domain Adaptation and Generalization via Prompt Learning

3.3.1. CLIP Zero-Shot vs Fine-Tuned on Domains

Domains	Accuracy (%)
Source	99.44
Target	84.22

Table 8. Accuraccies of the linear classifier trained on frozen CLIP image features for each PACS domain.

Table 8 and 9 summarize our findings comparing CLIP's zero-shot capabilities with linear probing. The results indicate that while CLIP performs admirably in a zero-shot setting (consistently achieving more than 95% accuracy on all domains except sketch), fine-tuning it through linear probing yields a significant boost in accuracy across almost all domains in the PACS dataset. However, do note that this improvement is deceptive as the linear classifier is rather overfitting to the training data (99.44% accuracy on source domain) and achieves only 84.22% accuracy on the target domain (sketch). Due to this, the generalization capabilities of CLIP in a zero-shot setting are much more reliable and robust when compared to linear probing.

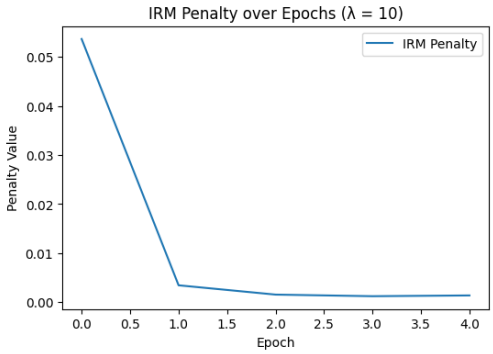


Figure 8. How the invariance penalty changes across epochs with a weight $\lambda = 10$

Domains	SP Accuracy (%)	DSP Accuracy (%)
Photo	99.82	99.82
Art Painting	96.29	95.36
Cartoon	98.17	97.91
Sketch	85.19	85.42

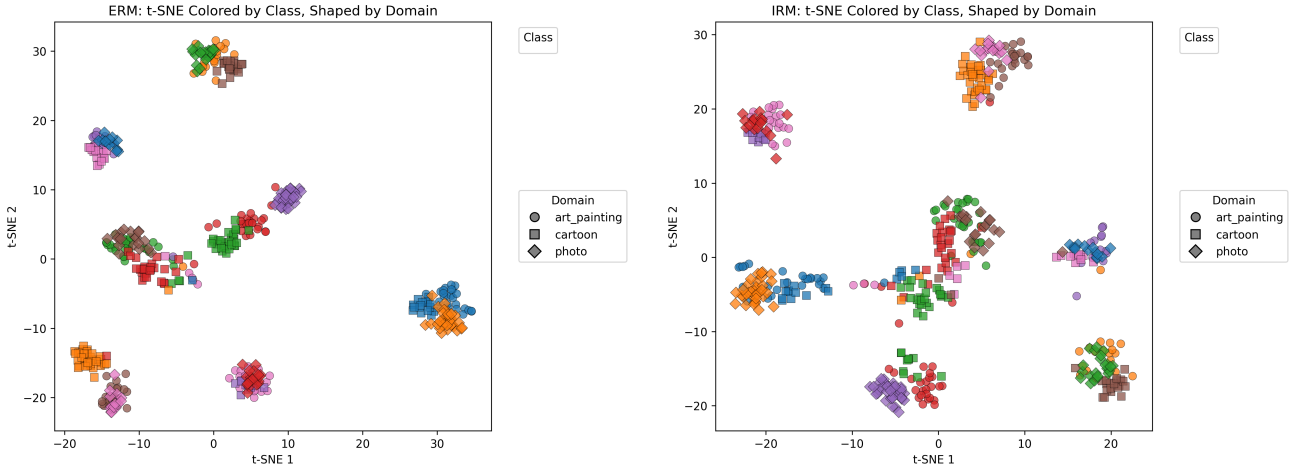
Table 9. CLIP Zero-Shot Accuracies using Simple Prompts (SP) and Domain-Specific Prompts (DSP) across PACS domains.

3.3.2. Prompt Learning

We observed that prompt learning using CoOp enhanced the target performance from 84.22% via linear probing to 85.19% on the sketch domain. Although modest, this improvement agrees with findings from (Zhou et al., 2022). This was done by optimizing a unified context across all source domains. For our domain generalization setup, incorporating pseudo-labels from the target domain during prompt learning led to even better generalization on the sketch domain, achieving an accuracy of **90%**. This is a significant boost over both zero-shot and linear probing methods, highlighting the effectiveness of prompt learning combined with pseudo-labeling for DA/DG.

Domains	Accuracy (%)
Source	99.34
Target ('sketch')	85.19

Table 10. Accuracies after Prompt Learning using Context Optimization (CoOp) for DA on PACS.



(a) ERM (b) IRM

(c) GroupDRO (d) SAM

Figure 9. t-SNE visualizations of learned representations across methods. Colors indicate domains; shapes indicate classes.

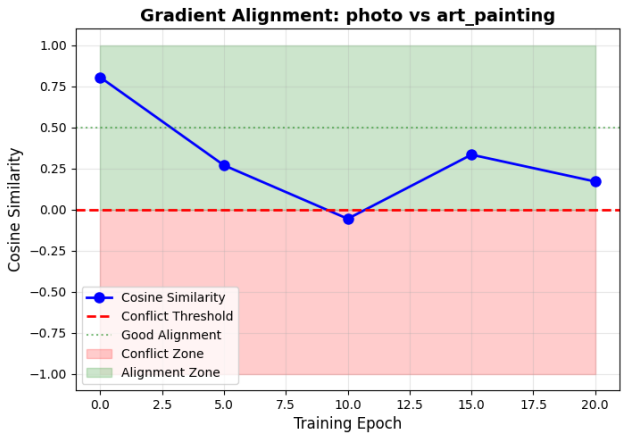


Figure 10. Cosine Similarity of Gradients between two source domains (Art Painting and Photo) during Prompt Learning using CoOp.

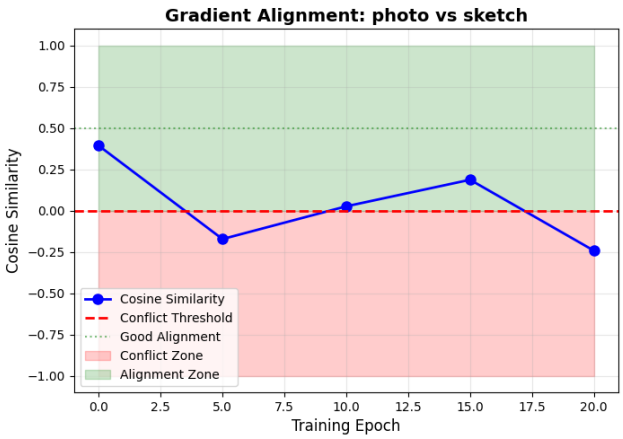


Figure 11. Cosine Similarity of Gradients between the target domain (Sketch) and the source domain (Photo) during Prompt Learning using CoOp.

3.3.3. Gradient Analysis Across Domains

Figures 10 and 11 illustrate the cosine similarity of gradients between two source domains (Art Painting and Photo) and between a source domain (Photo) and the pseudo-labeled target domain (Sketch) during prompt learning using CoOp. For the two source domains, a relatively higher cosine similarity is observed during the training epochs, suggesting that the gradients are aligned across these domains due to their similar styles. However, do note that with more epochs, the cosine similarity decreases and is converging towards 0-0.2, indicating that the model is learning domain-specific features as well.

In constrast, the gradient conflict between the source domain (Photo) and the target domain (Sketch) is much more pronounced. The angle between the gradient vectors (Figure 17) remains close to 90 degrees throughout the training epochs. This indicates that the gradients from these domains are pointing in nearly orthogonal directions, highlighting the significant domain shift. Also, the magnitude of the gradients (Figure 11) in this conflict is much higher than that of the source domains (Figure 14), suggesting that the model is struggling to reconcile the differences between these domains during training.

3.3.4. Open-Set Evaluation

Table 11 summarizes the results of our open-set evaluation. For learned prompts on 80% of the classes in PACS, the model achieved an average prediction confidence of 0.92 on seen classes in the target domain (Sketch) along with an accuracy of 80.74%. However, for unseen classes, the average prediction confidence dropped significantly to 0.66 depicting that the model becomes less certain when predicting classes it has not encountered during training. This shows that prompt learning in our case, while effective for DA, is less robust in open-set scenarios where novel classes are present.

Category	Accuracy (%)	Confidence
SEEN Classes ('sketch')	80.74	0.92
UNSEEN Classes ('sketch')	-	0.66

Table 11. Open Set Generalization Accuracies and average confidence after Prompt Learning using Context Optimization (CoOp) on PACS.

4. Discussion

4.1. Unsupervised Domain Adaptation

4.1.1. Why Adversarial Alignment Succeeds on PACS

Our results (Table 2) show that DANN's adversarial alignment clearly outperforms all alternatives on PACS, in contrast to semantic-shift benchmarks where pseudo-labeling dominates (Saito et al., 2017). This highlights that **adaptation method effectiveness depends on the nature of domain shift**.

PACS primarily exhibits style-based variation—objects remain semantically consistent across domains but differ in visual rendering (photographic, cartoon, sketch). Such shifts favor distribution-matching approaches: adversarial objectives can suppress stylistic artifacts while preserving object-level discriminative structure. DANN's superior performance over DAN, despite similar A-distances (Table 3), supports this view and aligns with theory showing adversarial training learns richer invariances than statistical moment matching (Ganin et al., 2016).

CDAN's intermediate performance indicates that class-conditional alignment helps preserve class structure but adds unnecessary complexity under purely stylistic shifts. Its conditioning mechanism can amplify noise when early predictions are unreliable, explaining its lower source accuracy. Similar concerns about degraded discriminability under weak supervision are discussed by Xiao et al. (2021), while Xiao et al. (2023) propose more structured alignment strategies beyond naive conditioning.

4.1.2. The Failure of Pseudo-Labeling

Pseudo-labeling performs poorly on PACS despite its success on semantic-shift tasks (Saito et al., 2017). The source-only model's weak 60.43% target accuracy yields unreliable pseudo-labels—confidence filtering either discards too much data or reinforces noise, stalling refinement. Moreover, adapting from photographic to sketch domains requires re-learning low-level style statistics, which simple self-training cannot capture. Adversarial alignment, by contrast, forces feature extractors to explicitly learn domain-invariant structure.

t-SNE embeddings (Figure 12) visualize this contrast: DANN and DAN show substantial source-target mixing, while pseudo-labeling and source-only remain disjoint. CDAN shows partial mixing, consistent with its intermediate accuracy. DANN forms tight within-class clusters across domains, explaining its superior target generalization.

Under rare-class conditions (Figure 13), the same pattern holds. DANN maintains alignment even for sparse categories, while pseudo-labeling continues to separate do-

mains. Its relatively high rare-class F1 (0.8378) suggests confidence filtering does help on minority samples, though without overall adaptation gains.

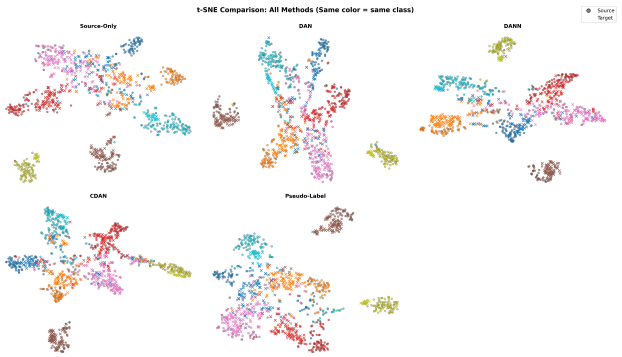


Figure 12. t-SNE visualization showing DANN achieves superior source-target mixing within class clusters, while pseudo-labeling fails to align domains.

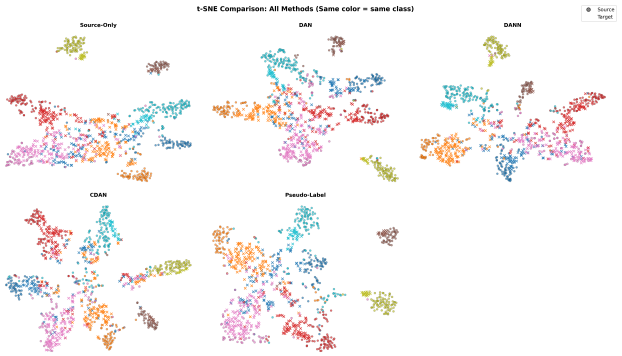


Figure 13. t-SNE visualization under rare-class scenario showing DANN maintains superior mixing while pseudo-labeling continues failing to align domains.

4.1.3. The Alignment–Discrimination Trade-Off Revisited

All alignment methods reduce source accuracy (Table 2), confirming the alignment–discrimination trade-off. On PACS, however, this trade-off is beneficial—target improvements outweigh modest source losses. In contrast, semantic-shift datasets often suffer negative transfer when alignment removes class-discriminative semantics (Zhao et al., 2019). This suggests that while alignment methods successfully reduce marginal distribution divergence similarly, DANN's adversarial mechanism learns more discriminative domain-invariant representations than DAN's statistical matching or CDAN's class-conditional approach.

In style-shift settings, alignment removes surface features (texture, shading) while retaining shape-based cues, which remain discriminative across domains. This explains DANN's robustness and CDAN's weaker balance. The nearly identical A-distances (Table 3) further support that *alignment degree* alone does not determine success—what matters is aligning the *right* features. Adversarial objectives selectively align stylistic factors while preserving discriminative ones (Johansson et al., 2019).

4.2. Invariant and Robust Domain Generalization

In our experiments, we see in Table 6 that overall best performance on unseen target domain is achieved with **Invariant Risk Minimization** (75.41% accuracy). This reinforces the hypothesis made by Arjovsky et al. (2020) that

> To learn invariances across environments, find a data representation such that the optimal classifier on top of that representation matches for all environments.

SAM was a close second with 73.43% accuracy on target domain and the highest average validation accuracy on source domain. This strongly makes a case for better generalizability of convergence to flat loss landscapes (**?**)

Improvements made on worst-case source domain performance does not seem to correlate with better target performance. In fact, IRM had the lowest worst-case performance, yet achieved the best target performance. Group DRO, which is *designed* to induce invariance by weighting against the worst-performing source domain, also did **not** achieve a higher accuracy than neither IRM nor SAM. This discrepancy may stem from implementation or training choices (e.g. η_q, no. of epochs, `batch_size` etc.)

We must note that IRM's high performance in our experimentation is subject to well-reported inconsistencies due to difficulties in training IRM because of its sensitivity to local minima in the loss landscape. This accuracy was achieved by finely tuning the penalty weight λ, because if λ is too small, IRM behaves like ERM as invariance is not enforced, and if it is too large, training collapses to mapping features to a constant classifier. This achieves a low penalty but with a severe cost to accuracy and performance. Our results for IRM illustrate this.

The transferable high performance of **Sharpness Aware Minimization** on source and target domains strongly suggests a stark correlation between flatness and domain invariance. We visualize loss landscape flatness in our results section, and also provided an anecdotal measure ΔL (change in loss) to be compared with ERM. Not only do the visualizations show that a SAM-trained model converged to a relatively broad flat region in the loss landscape, but the comparitive plot shows that ΔL is always less than ERM, suggesting that SAM forced the model to converge to a flatter minima than ERM. While we cannot be sure whether

the flat minima that SAM converged our model to is common across all domains, the improved target performance as well as improved source performance coincide with the theory that spawned SAM: that converging to a minima that is flat (i.e. less sensitive to perturbations) induces generalization by making the model robust to perturbations it may encounter in spurious correlations or a completely unseen domain (**?**). This does however come at the cost of higher computational power required, as SAM convergence makes two backprop passes per training step (one for default weights and one for perturbed weights). So the trade-off is significant.

In our experiments, the domains we attempted to generalize across differed in stylistic features (photo, painting, cartoon, sketch), but not in semantic meaning i.e. all domains had the same classes – closed group domain generalization. Therefore, forcing models to learn domain invariant features is a more suitable approach for our dataset, and that is what the 3 methods achieve. On the contrary, if there was semantic differences such as uncommon class labels across domains or class information entangled with domain-specific features, we would prioritize discriminability to maximize the model's classification performance.

4.3. Domain Adaptation and Generalization via Prompt Learning

Our experiments with CLIP and prompt tuning demonstrate that prompts are effective tools for adapting large vision-language models like CLIP and ALIGN (Zhou et al., 2022; Radford et al., 2021), to new domains. The fact that prompt learning with CoOp performs better than the zero-shot and linear probing baselines indicates that learning a shared context across source domains (given the assumption that all our domains share the same label space) helps the model adapt better to the target domain. Training a linear classifier on frozen CLIP features leads to the linear probe learning domain-dependent features, which fails to generalize well on unseen domains. In contrast, prompt learning may encourage the model to learn domain-invariant features by optimizing a unified context across all source domains. In either case, the model performance is impacted when the source domain and target domain have significant domain shifts (e.g. Photo to Sketch).

This leads us to an adaptation-generalization trade-off similar to the alignment-discrimination trade-off observed in traditional DA methods. Prompt learning effectively adapts to the target domain by learning domain-specific contexts, but this may come at the cost of generalization to new domains (Zhou et al., 2022). We explored thus by incorporating pseudo-labels from the target domain during prompt learning, which led to significant improvements in generalization performance on the target domain. However, as literature on this topic discusses, CoOp requires training on specific data distributions that harbors an inherent risk of learning spurious correlations that do not generalize well to unseen domains (Zhou et al., 2022; Radford et al., 2021).

Naive prompt approaches suffer across domain shifts due to gradient conflicts, as shown in Figures 10 and 11. This conflict hinders effective learning of domain-invariant features, as the model may be forced to learn domain specific features to minimize loss on both domains. Aligning gradients across domains could mitigate recurring conflicts observed during prompt learning, potentially improving adaptation and generalization performance. However, the fact that we are training a domain invariant feature extractor (using CLIP's frozen image encoder) limits the extent to which we can fully resolve these conflicts as aligning gradients may lead to overfitting to source domains.

Finally, our open-set evaluation reveals that while prompt learning enhances DA performance, it struggles with unseen domains and classes. This may correlate with the findings from (Zhou et al., 2022) that CoOp may result in spurious correlations given the nature of the training data. While this technique improves closed-set performance, it is poor on open-set scenarios.

5. Conclusion

This work presents a systematic empirical investigation of unsupervised domain adaptation on the PACS benchmark, revealing that DANN substantially outperforms alternative approaches including statistical matching (DAN), CDAN, and self-training (pseudo-labeling). DANN achieves 74.81% target accuracy (+14.38% over source-only baseline), demonstrating that explicit distribution matching can be highly effective when domain shifts are primarily style-based rather than semantic.

Through controlled experiments on PACS, we observe that enforcing invariance (IRM) and promoting flatness (SAM) both improve domain generalization beyond ERM and GroupDRO. SAM strikes a balance between simplicity and generalization, achieving flatter minima that correlate with better robustness to distribution shifts. Future work could explore combining SAM's flatness regularization with IRM's invariance constraints for stable and robust generalization.

Lastly, we demonstrate that prompt learning with CLIP effectively adapts to new domains, with pseudo-labeling further enhancing generalization. Gradient analyses reveal conflicts between source and target domains during prompt optimization, suggesting avenues for improved adaptation strategies. However, open-set evaluations indicate limitations in handling unseen classes, highlighting the need for robust prompt designs that generalize beyond closed-set scenarios.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2020. URL https://arxiv.org/abs/1907.02893.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, pp. 1180–1189. PMLR, 2015.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2030–2096, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Johansson, F. D., Sontag, D., and Ranganath, R. Support and invariance in domain adaptation. In *International Conference on Learning Representations*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL https://arxiv.org/abs/1412.6980.

Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, pp. 97–105. PMLR, 2015.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. *Advances in Neural Information Processing Systems*, 31, 2018. URL https://arxiv.org/abs/1705.10667.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Saito, K., Ushiku, Y., and Harada, T. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Xiao, T. et al. Simultaneously improve transferability and discriminability. *PMC (open access)*, 2021.

Xiao, Z. et al. Spa: A graph spectral alignment perspective for domain adaptation. In *NeurIPS (poster track)*, 2023.

Zhao, H., Des Combes, R. T., Zhang, K., and Gordon, G. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 7523–7532. PMLR, 2019.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2022. URL https://arxiv.org/abs/2109.01134.

Appendix

A. Additional Experimental Results

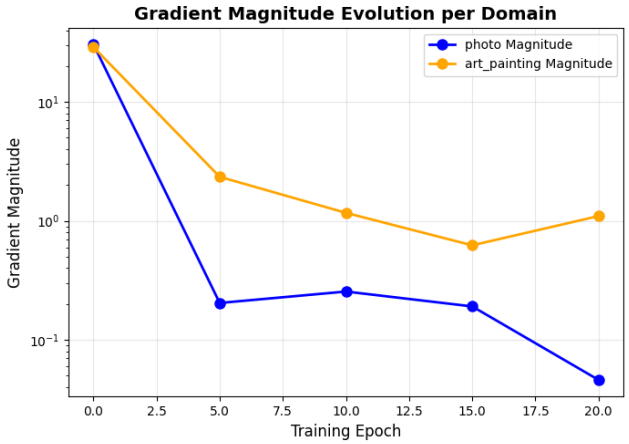


Figure 14. Evolution of gradient magnitudes for two source domains (Art Painting and Photo) during Prompt Learning using CoOp over 20 epochs.

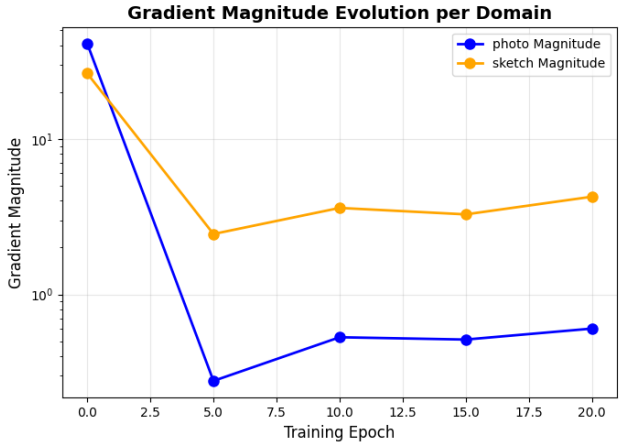


Figure 15. Evolution of gradient magnitudes for the target domain (Sketch) and source domain (Photo) during Prompt Learning using CoOp over 20 epochs.

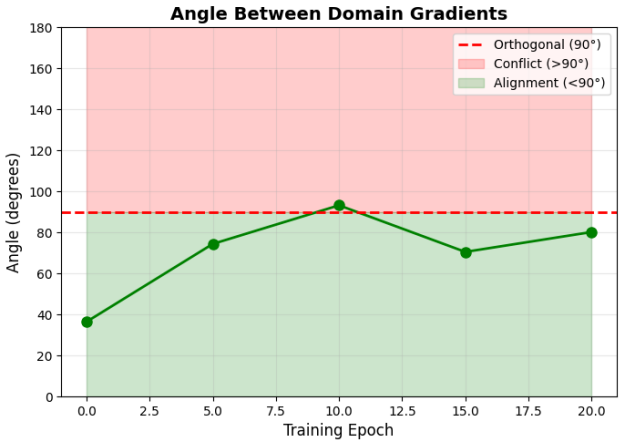


Figure 16. Evolution of angles between gradient vectors of the mentioned source domains during the CoOp Prompt Learning scheme over 20 epochs.

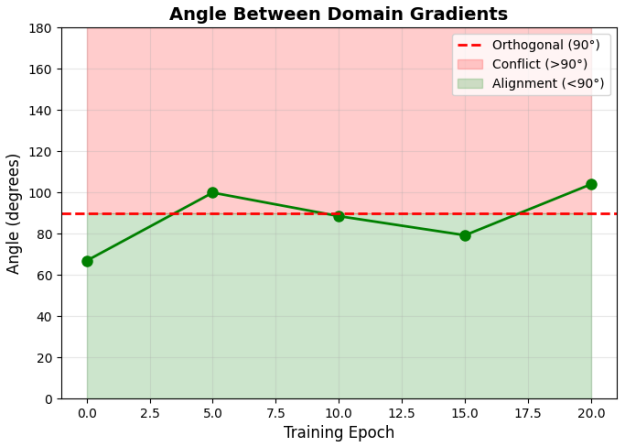


Figure 17. Evolution of angles between gradient vectors of the target domain (Sketch) and source domain (Photo) during the CoOp Prompt Learning scheme over 20 epochs.