Robust Machine Learning for Domain Adapation and Domain Generalization

Qasim Ayub [1] Muhammad Usman [1] Usman Ahad [1]

Abstract

Popular machine learning architectures such as deep neural networks achieve remarkable accuracy when tested in the same domain, but degrade sharply when exposed to distribution shifts across domains. This paper presents a comprehensive study of domain adaptation (DA) and domain generalization (DG) methods, aiming to understand when and why alignment, invariance, or robustness-based strategies succeed or fail. We evaluate four DA methods; DAN, DANN, CDAN, and self-training under different conditions to analyse their strengths and tradeoffs. For DG, we investigate ERM, IRM, Group DRO, and SAM to assess their ability to learn representations that generalize to unseen domains. Finally, we extend the study to vision–language models by exploring CLIP prompt tuning (CoOp, CoCoOp) and gradient-alignment techniques, examining their stability. Across multiple benchmark datasets (e.g., PACS, Office-Home, DomainNet), our results highlight that no single method dominates: alignment-based DA can overfit under semantic shift, while flatness-aware and gradient-aligned approaches yield more consistent generalization. Repository here.

1. Introduction

This project studies Domain Adaptation (DA) and Domain Generalization (DG) through a sequence of hypothesis-driven experiments designed to (i) quantify how different adaptation strategies perform under different conditions and what are their trade offs, (ii) evaluate methods that explicitly seek invariant solutions across multiple sources i.e IRM and SAM, and (iii) investigate whether prompt tuning and gradient alignment provide a stable control knob for adaptation when using large pretrained vision–language models.

In Section 2 we perform unsupervised DA from a labeled source to an unlabeled target, comparing a source-only ERM baseline with four alignment techniques, DAN, DANN, class-aware CDAN, and with self-training via pseudo-labels. We complement numeric metrics like accuracies with visualizations like t-SNE embeddings to analyse how alignment affects class separability and rare-class performance. Task 1 also includes controlled *concept-shift* experiments by removing certain class labels from target domain to test alignment methods under extreme conditions.

Section 3 is focused on domain generalization. We treat several source domains and evaluate on a held-out unseen domain. We implement and compare ERM, IRM (with stability diagnostics to detect trivial solutions), Group DRO (to optimize worst-case domain performance), and SAM-based sharpness-aware training. Our analysis focuses on (i) target-domain accuracy, (ii) worst-source-domain behaviour, (iii) optimization stability, and (iv) loss-landscape flatness as explanation for improved OOD robustness.

Section 4 explores CLIP and prompt-based adaptation. We first measure CLIP zero-shot and linear-probe baselines across domains, then implement prompt-learning schemes (CoOp / CoCoOp) for source supervision plus unsupervised target constraints (consistency or pseudo-labeling).

Across all tasks we adhere to rigorous experimental practices: using a consistent backbone for fair comparisons (pretrained ResNet/ViT or CLIP variants as specified), fixing random seeds where possible, and reporting per-domain and worst-group metrics.

2. Domain Adaptation

2.1. Methodology

This section focuses on domain adaptation experiments conducted using four methods: Deep Alignment Network (DAN), Domain Adversarial Neural Network (DANN), Conditional Domain Adversarial Network (CDAN), and Domain Adaptation using Pseudo-Labels. All experiments are con-

[1] Department of Computer Science, LUMS, Lahore, Pakistan. Correspondence to: Qasim Ayub <27100168@lums.edu.pk>, Muhammad Usman <27100046@lums.edu.pk>, Usman Ahad <27100041@lums.edu.pk>.

ducted using the OfficeHome dataset, normalized using ImageNet mean and standard deviation, with **Real World** domain as the source, and the **Art** domain as the target.

A ResNet50 backbone is used across all experiments to ensure consistency and enable a comparative analysis of the different domain adaptation techniques. Each model is trained for 10 epochs using the Adam optimizer with a learning rate of 0.001. The baseline model is trained solely on the source domain, without any feedback from the target domain, to provide a fair comparison against the adaptation methods.

After the initial experiments, we simulate a concept shift scenario by introducing two types of modifications: (1) a *label shift*, achieved by removing several randomly selected classes from the target domain, and (2) a *rare-class* scenario, created by ensuring one class is under-represented in target domain. All domain adaptation methods are subsequently evaluated under these shifted conditions.

2.1.1. Deep Alignment Network (DAN)

In this approach, features are extracted after layers 2, 3, and 4 of ResNet50. Forward passes are performed separately for the source and target datasets to obtain corresponding feature representations. Adaptive pooling is applied to reduce feature dimensionality to a 2D space, to reduce computation. Alignment is achieved using the Maximum Mean Discrepancy (MMD) loss with a Gaussian kernel, which is combined with the cross-entropy loss to train the model.

2.1.2. Domain Adversarial Neural Network (DANN)

For DANN, features are extracted immediately before the classification head and passed to a discriminator. Forward passes are performed on both source and target datasets. The discriminator is trained using binary cross-entropy loss, while the feature extractor is trained with a gradient reversal layer, to incorporate adversarial feedback, and standard cross entropy loss. The discriminator architecture is shown in Table 1.

Table 1. Architecture of the discriminator network.

Layer	Details
Linear (Input → 1024)	Fully connected layer
ReLU	Activation
Dropout (0.5)	Regularization
Linear (1024 → 1024)	Fully connected layer
ReLU	Activation
Dropout (0.5)	Regularization
Linear (1024 → 1)	Output layer

2.1.3. Conditional Domain Adversarial Network (CDAN)

This method extends DANN by embedding class information into the feature representation. Specifically, feature vectors are combined with class predictions via an outer product before being passed to the discriminator, allowing class discriminator to be trained on class information as well.

2.1.4. Domain Adaptation using Pseudo-Labels

In this method, the fine-tuned ResNet50 model from the baseline experiment is used to generate pseudo-labels for the target domain. Target samples with prediction confidence above an initial threshold of 90% are selected for fine-tuning. The threshold is increased by 1% after each iteration, and pseudo-labels are re-generated. The model is fine-tuned for five epochs in this way and finally it is evaluated on target domain.

2.2. Results

The baseline model achieved a source accuracy of 91% and a target accuracy of 23%. After applying concept shift, target accuracy decreased slightly to 22%.

2.2.1. Deep Alignment Network

Applying DAN reduced the source accuracy to 41% and the target accuracy to 18%. Under concept shift, these dropped further to 19% and 7%, for source and target respectively.

2.2.2. Domain Adversarial Neural Network

DANN achieved 82% source accuracy and 23% target accuracy, which increased to 83% and 26% under concept shift conditions.

2.2.3. Conditional Domain Adversarial Network

CDAN achieved 86% source accuracy and 27% target accuracy under normal conditions. After applying concept shift the accuracies fropped to 83% and 25% for source and target respectively.

2.2.4. Domain Adaptation with Pseudo-Labels

This approach resulted in the lowest performance, with the target accuracy decreasing to 10% on the standard dataset. Interestingly, under concept shift, the target accuracy increased slightly to 15%.

Table 2. Source and target domain accuracies before and after applying concept shift.

Method	Src (%)	Tgt (%)	Src (Shift) (%)	Tgt (Shift) (%)
Base	91	23	91	22
DAN	41	18	19	7
DANN	82	23	83	26
CDAN	86	27	83	25
Pseudo-Labelled	22	12	20	13

Table 3. Difference between source and target accuracies.

Method	∇Acc (Normal) (%)	∇Acc (Concept Shift) (%)
Base	-68	-69
DAN	-23	-12
DANN	-59	-57
CDAN	-59	-58
Pseudo-Labelled	-10	-13

2.3. Analysis

Training on the source domain (*Real World*) yielded a source accuracy of 91%, indicating substantial feature learning within that domain. However, testing the same model on the target domain resulted in a significant accuracy drop to 23%, suggesting that the model captured non-causal, domain-specific features. This degradation can be due to CNNs' reliance on local semantic cues rather than global structure. Consequently, the model struggled to generalize to the *Art* domain, which differs from the source in color, style, and object shape, as shown in Figure 1. The model overfitted on the source domain producing tight clusters while struggling on test domain as seen by the feature representations in Figure 2 which are very scattered in target domain. Introducing a concept shift further reduced target accuracy by 1%, as some relatively high-performing classes were removed from the target set. Table 4 lists the classes removed and their per-class accuracies in the base model.

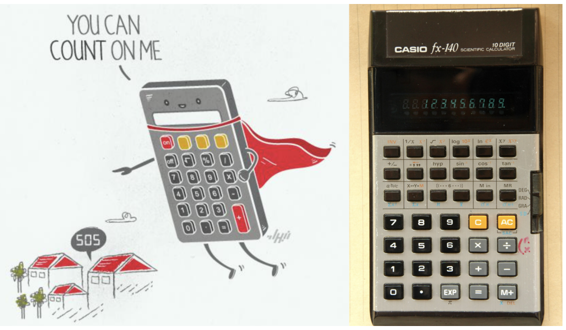

Figure 1. Left: Calculator from the target dataset (Art). Right: Calculator from the source dataset (Real World).

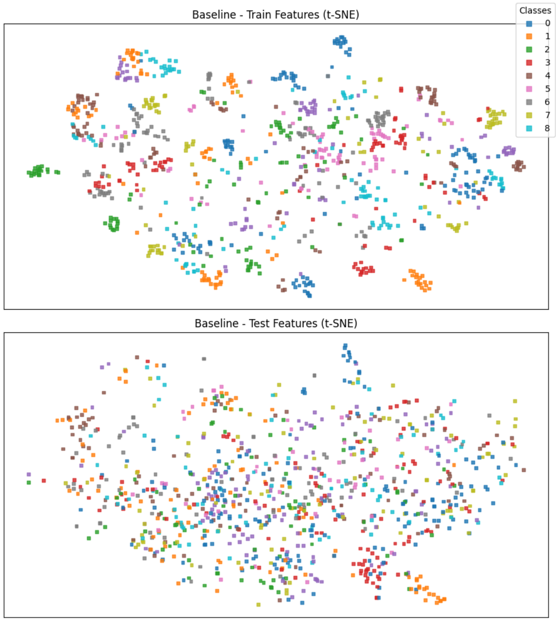

Figure 2. Baseline TSNE Feature Map

Class Label	Target Accuracy Before Concept Shift (%)
3	15
14	17.5
28	28
31	35
35	22

Table 4. Per-class accuracies of removed classes in the concept shift scenario

2.3.1. Deep Alignment Network

DAN is computationally demanding, as Gaussian-kernel MMD alignment in high-dimensional feature space requires substantial computation. Although feature alignment reduced the discrepancy between domains (as visualized in Figure 3), it also caused a significant decline in source accuracy. Since, there were major semantic differences in source and target domains aligning their feature spaces reduced source accuracy as well as large shifts in domain can confuse the model (Zhang et al., 2021). Under concept shift, the method proved even less effective, as the underlying data distributions diverged further. This method might yield results if trained for more epochs but under these standard conditions this was not as efficient as some other approaches.

2.3.2. Domain Adversarial Neural Network

DANN employs adversarial training to encourage domain invariance. Although its target accuracy matched the baseline, the reduction in ∇acc indicates improved balance between

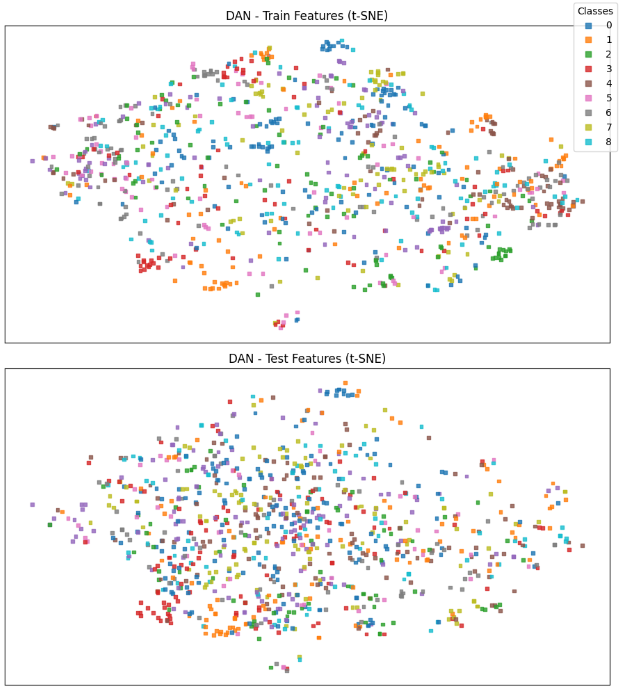

Figure 3. t-SNE visualization of DAN feature space alignment.

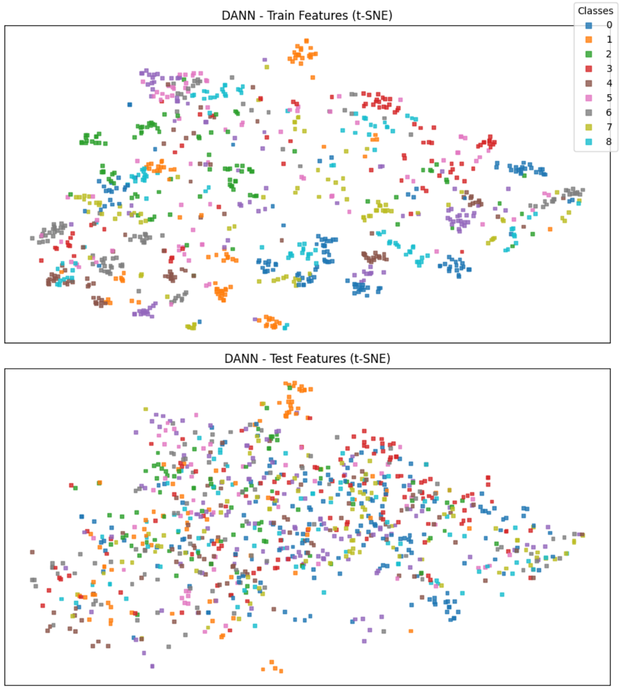

Figure 4. t-SNE visualization of DANN feature space.

domains. As shown in Figure 4, the model began forming coherent clusters for certain classes (e.g., "Backpack (Label 1)"), suggesting partial domain alignment. However, further training is required for stronger convergence.

2.3.3. Conditional Domain Adversarial Network

CDAN achieved the best overall performance among the adaptation methods, maintaining high source accuracy while slightly improving target performance. By incorporating class-conditional feedback, it provides more informative gradient signals that reflect both class-discriminative and domain-invariant characteristics (Long et al., 2018). This leads to better alignment in the source and target as compared to other methods. This method resulted in a drop in accuracy under concept shift since certain class labels were removed and this method relied on class labels heavily to align domains.

2.3.4. Domain Adaptation with Pseudo-Labels

Although pseudo-labeling is widely regarded as an effective self-supervised strategy, it underperformed in this setup due to substantial domain differences. The model frequently generated incorrect pseudo-labels with high confidence, reinforcing incorrect labels. As a result, both source and target accuracies dropped. Similar to the DAN, the feature space is scattered, for both source and target, since any information learnt about source is lost when the model is fine-tuned using incorrect labels. Incorporating an auxiliary function to predict pseudo labels could mitigate this issue in future work.

2.3.5. Overall Analysis

While domain adaptation methods conceptually aim to bridge domain gaps, practical implementation reveals considerable training overhead with negligible accuracy improvement. Each technique required extensive computation, yet none substantially surpassed the baseline. This may stem from limited training epochs, restricted dataset size, and CNNs' inherent semantic bias.

Overall, despite some clustering improvements in adapted models, the baseline achieved the most stable performance-to-effort ratio, emphasizing that in real-world settings, architecture selection and data diversity are as critical as adaptation strategy design.

3. Domain Generalization via Invariant and Robust Learning

3.1. Methodology

Recent literature has focused on the use of specialized generalization-focused learning strategies to improve performance on Domain Generalization (DG) tasks, where the target domain data is inaccessible during training. Modern DG approaches often build upon, or combine, methods such as Invariant Risk Minimisation (IRM), Group Distributionally Robust Optimization (GroupDRO), and Sharpness Aware Minimisation (SAM) (Arjovsky et al., 2019; Foret et al., 2020; Sagawa et al., 2019), as seen in DGSAM and curriculum-enhanced GroupDRO (Song et al., 2025; Barbalau, 2024).

We assess the efficacy of these IRM, GroupDRO, and SAM by applying them to the PACS dataset (Li et al., 2017) containing 9,991 images from seven classes: Dog, Elephant, Giraffe, Guitar, Horse, House, and Person. The four visually distinct domains - Photo, Art, Cartoon, Sketch - provide a lightweight yet challenging benchmark for DG. We train using the Photo, Art, and Cartoon domains as source domains and evaluate with the entire Sketch set as our target domain.

To assess performance in a moderately-sized network, we apply the learning strategies to ResNet-18 pre-trained on ImageNet weights. We crop images to 224-pixel resolution, normalize using ImageNet values, and train for 9 epochs.

We train two Empirical Risk Minmization (ERM) classifiers as baselines for comparison. The first is trained using the Adam optimizer with batch size of 64 and learning rate of 0.001. The second baseline uses Stochastic Gradient Descent (SGD) with 3 warm-up epochs, batch size of 512, weight decay of 0.0004, momentum of 0.9, and cosine learning rate schedule with initial learning rate of 0.08. Although ERM with Adam is a suboptimal choice for DG (Naganuma et al., 2022), we include it for context because the IRM and GroupDRO setups discussed below use Adam internally. SGD is expected to set a challenging benchmark (Lin et al., 2021) and is thus used as the primary baseline.

We use the DomainBed benchmark suite implementations (Gulrajani & Lopez-Paz, 2020) of IRM and GroupDRO with batch size of 64 and learning rate of 0.001, following from ERM-Adam. For IRM, we use penalty term of $\lambda = 1$. When performing penalty term ablation analysis, we anneal the penalty term to by 1.0 for the first 3-of-9 epochs following the warm-up schedule of ERM-SGD. For GroupDRO, we use an update rate for group weights of $\eta = 0.01$.

We implement SAM following the original paper (Sagawa et al., 2019), applying it with $\rho = 0.05$ and the SGD optimizer following the same training setup as ERM-SGD.

3.2. Results

Our results in Table 5 show that ERM-Adam is the least optimal learning strategy for Domain Generalization, achieving 86.56% accuracy across unseen source samples and 36.93% accuracy on the target domain. This 49.63% difference demonstrates poor generalization ability despite strong in-domain performance and suggests that the model over-fitted to source idiosyncrasies. This undesirable performance is likely the result of Adam's adaptive updates converging quickly to sharp minima found in the training set leading to low robustness to data perturbations in the target domain (Naganuma et al., 2022). Figure 5 confirms that ERM-Adam finds the sharpest minima relative to the other models. Further, the model may rely on spurious correlations from the source domain, such as background textures i.e. sky and grass or source-specific styling, a phenomenon which has been observed when using PACS for DG testing (Wang et al., 2022a).

	ERM-Adam	ERM-SGD	IRM*	GroupDRO	SAM
D_1: Art	82.52%	93.2%	-1.94%	-10.68%	-0.97%
D_2: Cartoon	82.20%	91.53%	$+3.39\%$	0.0%	$+1.69\%$
D_3: Photo	97.62%	96.43%	-1.19%	-2.38%	$+1.19\%$
All Source	86.56%	93.44%	$+0.33\%$	-4.26%	$+0.66\%$
D_T: Sketch	36.93%	66.51%	-3.14%	-0.79%	$+9.29\%$

Table 5. Accuracy across unseen source samples and target domain. Final 3 columns are with reference to ERM-SGD. *IRM with $\lambda = 1$

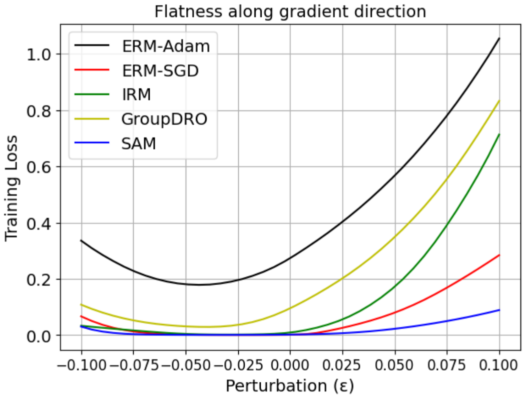

Figure 5. Loss landscape visualization: effect of weight perturbation along gradient direction on training loss

The ERM-SGD results in Table 5 demonstrate that strong DG performance can be achieved without using generalization-specific learning strategies while also retaining high in-domain accuracy. ERM-SGD achieved 66.51% accuracy on the target domain and 93.44% across source domains. The 29.58% increase in OOD accuracy over ERM-Adam suggests that SGD converges at more robust, flatter loss minima which are less prone to data perturbations despite the opportunity to rely on spurious correlations present in the dataset. When comparing ERM-SGD and ERM-Adam's loss landscapes in Figure 5, we see that the SGD loss is 0.18-0.42 lower than Adam's across the perturbation range and it is also stable across perturbations from -0.075 to 0.0125, while Adam has an unstable, sharp minima. Because these results are achieved using hyperparameter tuning which maximizes in-domain performance, they encourage the exploration of optimizing ERM methods

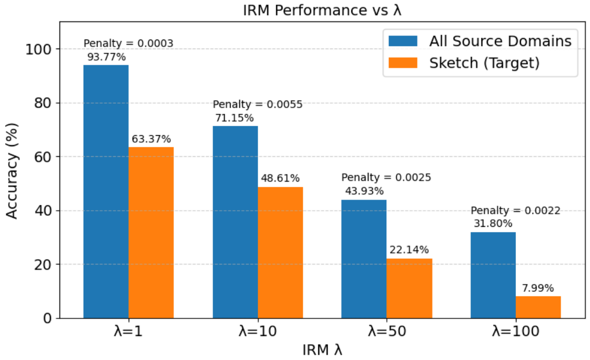

Figure 6. Effect of IRM penalty weight vartion on unseen source and target accuracy. Final IRM penalty term shown.

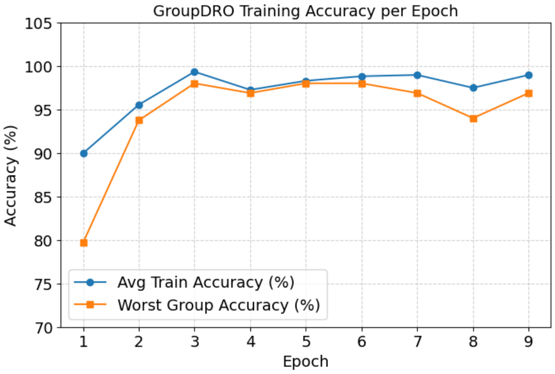

Figure 7. GroupDRO Average and Worst Group Training Losses

for generalization tasks which may further increase performance (Lin et al., 2021).

IRM with $\lambda = 1$ exhibits 3.14% lower target domain accuracy and comparable $+0.33\%$ in-domain accuracy relative to ERM-SGD, our primary baseline (Table 5). The final penalty weight during training was an extremely low 0.0003. This indicates that the model was successful in minimizing the regularization term. However, the high in-domain and lower target-domain performance suggests that the model relies on spurious correlations in the source domains, as discussed above, and is thus not robust to perturbations in the target domain. This is reinforced in Figure 5, where the loss landscape for IRM is sharper than ERM-SGD, although more stable than ERM-Adam at low perturbations. Thus, this model is an example of the high instability of IRM because the intended capturing of invariant features across source domains is not successful. The issue of instability is also seen in Figure 6, which shows the tendency of IRM to collapse to trivial, non-discriminative representations based on the choice of penalty weight λ. We note that, despite always achieving near-zero penalty time, the model's in-domain and target domain performance reduces significantly as λ grows. For example the drop in target and in-domain accuracies by 41.23% and 49.83% respectively from $\lambda = 1$ to $\lambda = 50$ show that enforcing invariance more strongly by increasing λ does not garner expected results. Interestingly, IRM yields a 3.39% increase in worst-domain (Cartoon) accuracy over ERM-SGD, suggesting that the invariance penalty is effective in improving performance on the most challenging source domain.

Our GroupDRO model attempts to increase generalization ability by maximizing performance on the worst-performing source domain, essentially implementing min-max optimization. The increase in worst group accuracy during training from 79.69% to 96.88% is seen in Figure 7. Table 5 shows that this strategy does not have the intended effect, causing 4.26% lower in-domain performance and 0.79% lower target domain performance comapred to our baseline. Notably, this model also achieves significantly low accuracy on the Art domain (source) - 10.68% lower than the baseline. These results show that worst-domain optimization does not effectively enforce the learning of domain-invariant features, instead they promote the possibility that the model learned spurious correlations in the data. This is also supported by the sharp loss landscape seen in Figure 5. Additionally, the low Art accuracy shows that worst-group optimization can come at the cost of how well the model fits to other source domains, which can be highly undesirable: the Art domain is comparable to Sketch. However, if Art is not the lowest-performing group, the model is restricted from learning highly transferable features. It is also notable that GroupDRO may be especially ineffective when using a small number of source domains, as in our example, due to higher chances of capturing idiosyncrasies when doing so (Wang et al., 2022b).

Lastly, the results for SAM in Table 7 highlight optimality for DG compared to all other learning strategies. SAM yields a 9.29% increase in target domain accuracy while retaining very high in-domain accuracy, 0.66% greater than our baseline ERM-SGD. This relatively high OOD performance can be attributed to an effective learning strategy which directly addresses the DG problem by attempting to find a flat loss minima in a stable manner. As seen in Figure 5, SAM outperforms other models in this sub-task, converging at a minima with loss comparable to ERM-SGD, but one which is significantly more robust to perturbations due to it's smoother curvature. For example, at $\epsilon = 0.1$, SAM records 0.088 while ERM-SGD has a training loss of 0.284. This high flatness region allows SAM to be robust to perturbations in the unseen target domain, leading to high generalization ability. Additionally, by searching for a flatter loss minima, SAM may have avoided learning spurious correlations in the training data, effectively enforcing some level of invariance. The higher in-domain performance of SAM suggests that enforcing generalization may also have a positive impact on non-DG tasks. The 1.69% and 1.19% increases in Cartoon and Photo domains point to the effectiveness of SAM in reduced over-fitting to training data, highlighting potential application beyond DG tasks. It must be remembered that the benefits of SAM come at the cost of increased compute, as SAM computes the gradient of the gradient at each step. This may be infeasible for some applications, so we encourage the prospect of a more efficient SAM-inspired solution.

3.3. Discussion

Our experiments, albeit limited by model size and training effort, produce informative insights into the differing models produced based on which learning strategy is used. We see that early generalization-specific strategies such as IRM and GroupDRO are highly unstable and limited in increasing generalization ability by enforcing invariance or maximizing worst-case performance.

In contrast, SAM demonstrates that converging at a flat minima explicitly improves DG performance while also reducing the risk of learning spurious correlations in data. The SAM performance shows that learning discriminative features was more effective for performance on PACS than strictly enforcing invariance. This may be because the PACS data is highly varied in style, and simply enforcing invariance encourages collapsed solutions or underfitting to the training data.

We note that there is still an 18.30% difference in accuracies between source and target domains in our best results (SAM). Additionally, a common challenge for the deployment of all of the DG solutions discussed is increased hyperparameters and instability with imperfect choices of these. There is a need for future literature that continues to address the source-target domain gap, training stability, and compute requirements. As each of the strategies above have their own merits and demerits, it may be beneficial to combine two or more strategies.

4. Fine Tuning Prompts via CLIP

4.1. Methodology

This study investigates domain generalization and adaptation for the CLIP model on the PACS dataset. The dataset's domains include Photo, Art Painting, and Cartoon as source domains, with Sketch serving as the unseen target domain for evaluating generalization in this experiment. Our methodology is structured across four distinct phases.

4.1.1. Phase 1: Zero-Shot vs. Linear Probing

We first establish baseline performance using two settings.

1. **Zero-Shot Evaluation:** The standard CLIP model is evaluated on all domains using a fixed prompt “a photo of a $\{$class$\}$”, “a painting of a $\{$class$\}$”, “a cartoon of a $\{$class$\}$”, and “a sketch of a $\{$class$\}$” for all classes.
2. **Linear Head Fine-Tuning:** The CLIP image encoder is frozen, and a linear classifier is trained on its output features. The classifier is trained for 10 epochs on the combined source domains (Photo, Art, Cartoon) to assess the quality of the frozen features and the effect of source-domain fine-tuning.

4.1.2. Phase 2: Prompt Tuning for Domain Adaptation

To explore lightweight adaptation, we implement a Context Optimization (CoOp) prompt learner. The prompt consists of learnable context vectors followed by frozen, class-specific label tokens. The context vectors are initialized to represent the phrase “a photo of a $\{$class$\}$”. To compare the effect of prompt length, we evaluated two configurations: a short prompt with 4 context tokens (`ctx_len=4`) and a longer one with 16 tokens (`ctx_len=16`). For both configurations, the model was trained for 10 epochs with the CLIP vision and text backbones fully frozen, updating only the prompt's context vectors.

Optimization is performed using Adam with a learning rate of 1×10^{-3}. The training objective is a combination of a standard supervised cross-entropy loss on labeled source data and an unsupervised consistency regularization term on unlabeled target (‘Sketch‘) data, weighted by $\lambda_u = 0.5$.

4.1.3. Phase 3: Gradient Conflict Analysis in Multi-Domain Learning

To analyze inter-domain gradient dynamics, we conduct a multi-domain prompt learning experiment using only the Art and Cartoon domains. We initialize a prompt that has already been fine-tuned on multiple source domains (from the previous phase) and further train it for 10 epochs using PCGrad (Projected Conflicting Gradients) to manage gradient conflicts. Throughout this phase, the CLIP backbone remains frozen. We monitor the training dynamics by calculating the cosine similarity between the gradients originating from the Art domain and the Cartoon domain at each epoch. A positive similarity shows gradient alignment, while a negative value is a sign of conflict.

4.1.4. Phase 4: Open-Set Generalization Analysis

Finally, we evaluate the impact of prompt tuning on CLIP's open-set recognition capabilities. From the 7 classes in the PACS dataset, we leave two classes (`dog`, `house`) unseen. A new prompt is trained for 10 epochs on the remaining five seen classes (`['elephant','giraffe','guitar','horse',` 'person']) using the combined source domains.

Evaluation is performed using several metrics:

- **Seen-Class Accuracy:** Classification accuracy on the held-in classes for each domain.
- **Unseen-Class Accuracy:** Zero-shot performance of the standard manual CLIP prompt on the held-out classes to measure baseline recognition.
- **Open-Set Detection:** The ability to distinguish between in-distribution (seen classes) and out-of-distribution (unseen classes) images is measured by AUROC and FPR@95%. We use two scoring functions for this task:
  1. **MSP-inv:** $1 -$ max softmax probability.
  2. **Cosine:** The negative of the maximum cosine similarity between an image embedding and the learned text embeddings.
- **Prompt Embedding Similarity:** We compute the mean diagonal cosine similarity between the learned prompt embeddings and the manual prompt embeddings for the same classes to quantify semantic drift.

4.2. Results

4.2.1. Zero-Shot and Linear Head Performance

Both the zero-shot CLIP model and the fine-tuned linear head achieved high accuracy on the source domains, with near-perfect scores on Photo (99.70%), Art Painting (95.56%) and Cartoon (97.65%). However, both models experienced a significant performance drop of approximately 14-15% on the unseen Sketch target domain, highlighting a clear domain shift. The linear head trained smoothly over 10 epochs, with the training loss decreasing from 1.75 to 0.35 and accuracy on the source domains improving from 81.4% to 97.7%. Detailed results are summarized in Table 6.

Table 6. Accuracy (%) of Zero-Shot CLIP and a fine-tuned Linear Head on the PACS dataset. Sketch is the unseen target domain.

Domain	CLIP Zero-Shot (%)	Linear Head (%)
Art Painting	95.56	95.26
Cartoon	97.65	98.46
Photo	99.70	99.70
Sketch (Target)	**85.16**	**84.50**

4.2.2. Prompt Learning Performance

The CoOp-style prompt learning demonstrated stable training dynamics for both context lengths, with losses decreasing steadily. When comparing the two configurations, we observed minimal impact on the target domain, as both the 4-token and 16-token prompts achieved an identical accuracy of 85.18% on the Sketch domain. Both versions performed slightly below the zero-shot CLIP baseline of 85.24%.

Interestingly, performance on the source domains diverged. The shorter prompt (ctx_len=4) slightly improved its source accuracy from 95.97% to 96.39% after training. In contrast, the longer prompt (ctx_len=16) experienced a slight degradation in source accuracy, dropping from 96.39% to 95.66%. Across both experiments, the learned context tokens were found to be semantically noisy, with their nearest neighbors in the vocabulary corresponding to tokens like "igers," "chelse," and various emojis. The results are summarized in Table 7.

4.2.3. Gradient Conflict Dynamics

The cosine similarity between gradients from the Art and Cartoon domains showed a distinct and volatile pattern during training, as illustrated in Figure 8. At initialization (epoch 0), the gradients were highly aligned with a similarity of approximately 0.89. As training began, the similarity dropped sharply into a state of conflict, hitting a low of approximately -0.07 at epoch 2. With the help of PCGrad, the gradients entered a period of partial realignment, recovering into positive territory. However, this alignment remained unstable, with significant fluctuations, peaking near 0.5 at epoch 6 before dropping sharply again to a negative similarity of approximately -0.05 in the final epoch. This volatility suggests that while PCGrad mitigated conflict, it did not fully resolve it. The two-source model achieved a final accuracy of 79.36%.

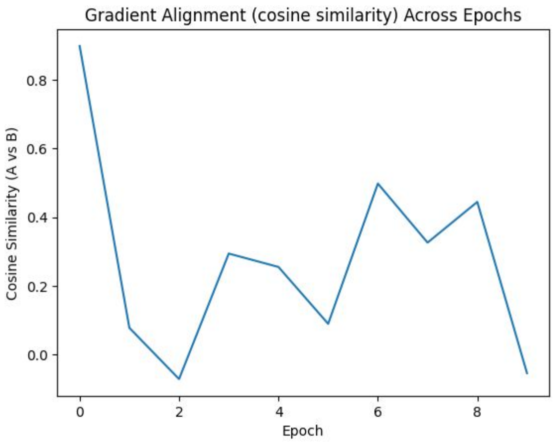

Figure 8. Cosine similarity between Art and Cartoon gradients over 10 epochs. Positive values indicate alignment, while negative values denote conflict. The stabilization after epoch 3 reflects PCGrad's role in mitigating interference.

Table 7. Comparison of Zero-Shot CLIP and CoOp models with different context lengths. Source accuracy is averaged over Art, Cartoon, and Photo.

Model	Source Accuracy (%)	Target (Sketch) Accuracy (%)
CLIP Zero-Shot (manual prompt)	97.86	85.24
CoOp (learned, ctx_len=4)	96.39	85.18
CoOp (learned, ctx_len=16)	95.66	85.18

4.2.4. Open-Set Generalization after Prompt Tuning

The prompt tuned on a closed set of five classes achieved high accuracy on those seen classes, though it consistently performed slightly below the manual zero-shot baseline across all domains, as shown in Table 8. For instance, on the `Photo` domain, the learned prompt achieved 99.58% accuracy, while on the more challenging `Sketch` domain, it reached 87.10%.

The model's open-set detection performance varied drastically depending on the scoring method. The Maximum Softmax Probability (MSP-inv) score proved to be a reasonable detector, achieving an AUROC of 0.8911, as seen in the ROC curve in Figure 10. The histogram in Figure 9 visually confirms this, showing a clear, albeit overlapping, separation between the score distributions for seen (in-distribution) and unseen (out-of-distribution) samples.

In stark contrast, cosine-based scoring failed completely, yielding an AUROC of 0.1867—worse than random chance. The histogram illustrates this failure, revealing a heavy overlap between the cosine similarity distributions for seen and unseen images.

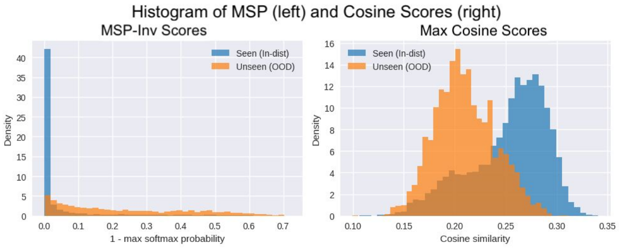

Figure 9. Histogram of MSP (left) and max-cosine scores (right) for seen vs unseen-class images under the learned prompt. MSP separates the two distributions reasonably well; cosine does not.

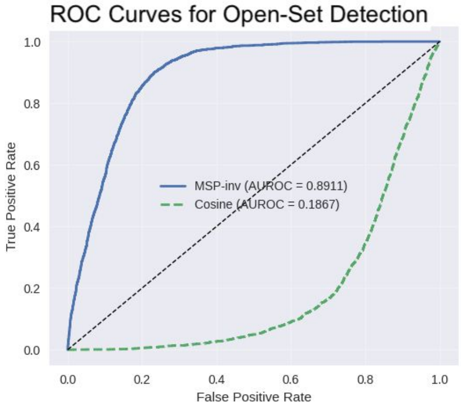

Figure 10. ROC curves for open-set detection. MSP-inv achieves AUROC 0.8822 while cosine-based detection fails (AUROC 0.2304).

Table 8. Seen-class accuracy across domains for the learned prompt (closed-set) and manual (zero-shot) baseline for comparison.

Domain	Manual Prompt - Zero-Shot (%)	Learned Prompt - Closed Set (%)
Art Painting	97.67	99.85
Cartoon	99.52	100.00
Photo	99.92	100.00
Sketch	86.29	99.88

4.3. Discussion

4.3.1. Domain Shift and the Adaptation-Generality Trade-Off

Our initial experiments confirm CLIP's strong zero-shot performance, especially on domains like Photo that align well with its pre-training distribution. However, the large accuracy drop of 15% on the highly stylized Sketch domain shows a clear domain shift. This occurs because the minimal texture and edge-based nature of sketches differ sharply from the rich visual data CLIP was trained on and optimized for. This leads to the adaptation-generality trade-off; fine-tuning a linear head on the source domains did not improve, and in fact slightly worsened, performance on the unseen target domain (84.50% vs. 85.16%). This suggests that training even a small component on the source domains can cause overfitting to source-specific styles, reinforcing that CLIP's zero-shot capability serves as a powerful and hard-to-surpass baseline.

4.3.2. Prompt Tuning as a Lightweight Adaptation Mechanism

Prompt tuning, such as the Context Optimization (CoOp) method we implemented, presents a parameter-efficient alternative to full fine-tuning (Zhou et al., 2022b). Our results show it can be a stable adaptation mechanism, but it does not resolve the generalizability challenge, as the learned prompts failed to outperform the zero-shot accuracy on the target domain. The key challenges are overfitting and brittleness. Our comparison between a short (4-token) and a long (16-token) prompt provides direct evidence of these limitations; increasing the prompt's parameter count did not improve generalization and slightly degraded source-domain performance. This combined with the semantically noisy nature of the learned tokens aligns with findings that CoOp often discovers syntactic shortcuts rather than meaningful semantic phrases (Zhou et al., 2022b). These static, class-agnostic prompts lack the ability to adapt dynamically to each image, a limitation addressed by subsequent methods like CoCoOp, which conditions prompts on image features (Zhou et al., 2022a).

4.3.3. Resolving Domain Conflict via Gradient Alignment

Our gradient analysis provides direct evidence of domain conflict during multi-domain learning. An initial high alignment (cosine similarity $\approx$ 0.89) was followed by a sharp drop into conflict as the model learned domain-specific stylistic features. To help solve this, PCGrad was used, which projects conflicting gradients to find a compromise update direction (Yu et al., 2020). While this approach forced the gradients back into a positive, more stable alignment, the continued volatility in similarity suggests that conflict was not fully resolved. By encouraging the model to learn domain-invariant features over superficial styles, gradient alignment can lead to more stable optimization, preventing the model from oscillating between domains. As an exploratory alternative, one could devise adaptive reweighting schemes or regularize updates to be orthogonal to learned ”style” vectors to achieve a similar goal.

4.3.4. The Impact of Prompt Tuning on Open-Set Recognition

Our final experiment reveals a critical risk of closed-set fine-tuning: the degradation of open-set capabilities. While the tuned prompt excels at classifying its seen classes, it slightly underperforms the robust zero-shot baseline, hinting at a loss of generality. More critically, it compromises the model's ability to identify unseen classes, a task where the manual prompt excels, confirming CLIP's well-structured embedding space for open-set recognition.

The stark failure of cosine similarity for open-set detection (AUROC of 0.1867) is particularly revealing. This failure occurs because prompt tuning reshapes the embedding space to maximize closed-set discriminability, causing ”semantic drift.” The learned prompt embeddings move away from their original general-purpose anchors to new, task-specific positions, destroying the geometric separation that cosine similarity relies on to detect outliers. In contrast, Maximum Softmax Probability (MSP) remains a viable detector (AUROC $\approx$ 0.89) because it relies on the *relative confidence* of the classifier, not the embedding geometry. This shows a critical trade-off: optimizing a model for one specific job can erase the broad knowledge it learned from its initial training, making it less flexible for new, unseen tasks.

5. Conclusion

Our results demonstrate that while some Domain Learning strategies are limited, other methods such as SAM and prompt-tuning for LLMs showcase more promising results with higher stability guarantees. These should be tested at scale with larger datasets so that real-world deployment can be imitated. There is also a need for literature that explores novel or hybrid strategies and improves the balance of invariance, robustness, and discriminability in order to maximize Domain Learning performance.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. v3, revised 27 Mar 2020.

Barbalau, A. Curriculum-enhanced groupdro: Challenging the norm of avoiding curriculum learning in subpopulation shift setups. *arXiv preprint arXiv:2411.15272*, 2024. https://arxiv.org/abs/2411.15272.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. v3, revised 29 Apr 2021.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. URL https://arxiv.org/abs/2007.01434.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization, 2017. URL https://arxiv.org/abs/1710.03077.

Lin, S.-B., Wang, Y., and Zhou, D.-X. Generalization performance of empirical risk minimization on over-parameterized deep relu nets. *arXiv preprint arXiv:2111.14039*, 2021. URL https://arxiv.org/abs/2111.14039.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. *Neural Information Processing Systems (NeurIPS)*, 2018.

Naganuma, H., Ahuja, K., Takagi, S., Motokawa, T., Yokota, R., Ishikawa, K., and Mitliagkas, I. Empirical study on optimizer selection for out-of-distribution generalization. *arXiv preprint arXiv:2211.08583*, 2022. https://arxiv.org/abs/2211.08583.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. v2, revised 2 Apr 2020.

Song, Y., Hwang, Y., Lee, J., Lee, H., and Lim, D.-Y. Dgsam: Domain generalization via individual sharpness-aware minimization. *arXiv preprint arXiv:2503.23430*, 2025. https://arxiv.org/abs/2503.23430.

Wang, Q., Wang, Y., Zhu, H., and Wang, Y. Improving out-of-distribution generalization by adversarial training with structured priors. *arXiv preprint arXiv:2210.06807*, 2022a. URL https://arxiv.org/abs/2210.06807.

Wang, R., Yi, M., Chen, Z., and Zhu, S.-C. Out-of-distribution generalization with causal invariant transformations, 2022b. URL https://arxiv.org/abs/2203.11528.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

Zhang, W., Deng, L., Zhang, L., and Wu, D. A survey on negative transfer. *IEEE Transactions on Neural Networks and Learning Systems*, -(-):1–24, 2021.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022b.