An Analysis of Model Robustness across Concurrent Distribution Shifts

Anonymous authors
Paper under double-blind review

Abstract

Machine learning models, meticulously optimized for source data, often fail to predict target data when faced with distribution shifts (DSs). Previous benchmarking studies, though extensive, have mainly focused on simple DSs. Recognizing that DSs often occur in more complex forms in real-world scenarios, we broaden our study to include multiple concurrent shifts, such as unseen domain shifts combined with spurious correlations. We evaluate 26 algorithms that range from simple heuristic augmentations to zero-shot inference using foundation models, across 168 source-target pairs from eight datasets. Our analysis of over 56K models reveals that (i) concurrent DSs typically worsen performance compared to a single shift, with certain exceptions, (ii) if a model improves generalization for one distribution shift, it tends to be effective for others, (iii) heuristic data augmentations achieve the best overall performance on both synthetic and real-world datasets. Our project webpage: It will be available upon acceptance.

1 Introduction

Machine learning models deployed in real-world settings may face complex distribution shifts (DSs). These shifts can result in unreliable predictions. For instance, a self-driving car may be deployed in a place where there is a different driving etiquette, unfamiliar environment *and* weather condition. Thus, a systematic evaluation of such complex shifts is essential before deploying models in the real world.

Most research has focused on evaluating models under a single DS (**UniDS**) (Taori et al., 2020; Gulrajani & Lopez-Paz, 2020; Koh et al., 2021; Wiles et al., 2022; Miller et al., 2021; Wenzel et al., 2022). For instance, in the evaluation benchmark PACS (Li et al., 2017), the training data consists of photographs but the model is tested on sketches. Similarly, in the training data of Biased FFHQ (Lee et al., 2021), gender is spuriously correlated to age, but the model is tested on data where gender is anti-correlated to age. In a real-world scenario, we can encounter these two shifts *concurrently*, i.e., one where age and gender are spuriously correlated *and* where there is also a shift in the style of the image.

To address this, we introduce an evaluation framework that mirrors the complex DSs potentially found in real-world settings. We assume that the input is generated from a set of attributes, e.g., gender, age, image style, etc. Thus, we leverage existing multi-attribute datasets, such as CelebA (Karras et al., 2017), which includes 40 attribute labels such as gender, hair color, and smiling, to manipulate the type and number of DSs present in the dataset. We refer to this framework as **ConDS**. It can account for *the co-occurrence of different DSs across multiple attributes within paired source and target datasets.* An example of this protocol is illustrated in Figure 1. Consequently, evaluating existing methods against these shifts can provide insights unattainable with current benchmarks. Specifically, our proposed framework consists of the following components:

Distribution Shifts. We explore seven types of DSs, including spurious correlation (SC), low data drift (LDD), unseen data shift (UDS), and their combinations; that is, (SC, LDD), (SC, UDS), (LDD, UDS), and (SC, LDD, UDS). To offer a comprehensive analysis covering a wide range of cases, we select various attributes to induce DSs. This includes adjusting three attributes for each **UniDS** and creating six combinations from three attributes for each **ConDS**, leading to 33 distinct cases per dataset.

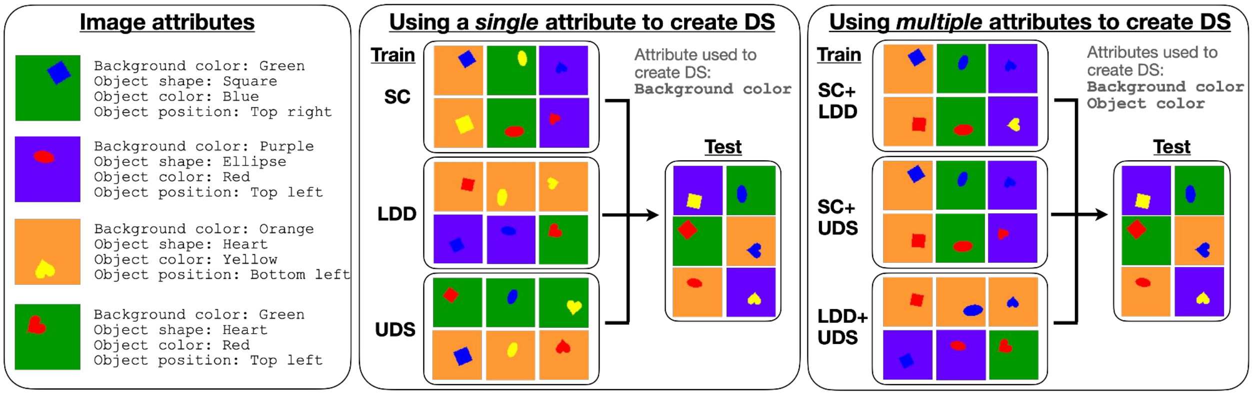


Figure 1: **Concurrent distribution shifts**. *Left*: We list some attributes of a few images from the **dSprites** dataset. In this dataset, the object shape is the label. *Center*: We show how a *single* attribute e.g., the background color, can be used to create different types of distribution shifts. Namely, spurious correlation (SC), where in this example, the background color is correlated with the object shape, low data drift (LDD), and unseen data shift (UDS). We assume that the test data consists of images where all attribute instances are equally likely to appear, i.e., each image is generated by randomly selecting each attribute instance with equal probability. *Right*: As the real world consists of more complex shifts, we also make use of *multiple* attributes to create combinations of distribution shifts. In the examples above, we use the background and object color to create combinations of distribution shifts. The first shift is created using the background attribute and the second shift, the shape attribute i.e., SC+UDS is created from a correlation between the background color and the shape (SC), ***and*** only using a subset of colors, for the shape attribute (UDS).

Algorithms. We evaluate 26 different methods from 7 popular approaches: architectural strategies, data augmentations, de-biasing, worst-case generalization, single domain generalization, out-of-distribution generalization, and even zero-shot inference using vision-language foundation models.

Experiments. We evaluate these algorithms on all proposed combinations of DSs. This results in over 56K experiments across 168 dataset pairs sourced from 3 synthetic and 5 real-world benchmarks. Further, we examine aspects such as the effectiveness of pre-training and the influence of prompts on foundation models for image classification.

As a result, we have made some intriguing findings: (i) Concurrent shifts are typically more challenging than single shifts; however, when a particularly difficult DS is combined with relatively easier ones, the harder shift tends to dominate, preventing the overall challenge from becoming more difficult. (ii) If a model improves generalization for one DS, it proves effective for others, even if it was originally designed to address a specific shift. (iii) Heuristic augmentation techniques outperform meticulously crafted generalization methods overall. (iv) Although vision-language foundation models (e.g., CLIP, LLaVA) perform well on simple datasets, even in the presence of DSs, their performance significantly deteriorates on more complex, real-world datasets.

2 Related Work

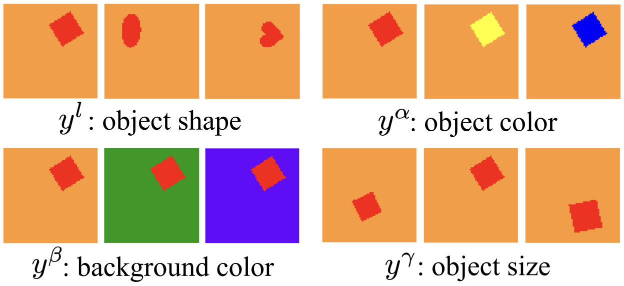
In this section, we review existing benchmarks for distribution shifts (DSs) and frameworks for evaluating model generalization.

Benchmark. Several types of benchmarks have been introduced to evaluate the generalization of models. One type of benchmark evaluates generalization on datasets collected from a different source than the training data (Li et al., 2017; Venkateswara et al., 2017; Peng et al., 2019; Koh et al., 2021). Another popular type applies transformations to the input (Hendrycks & Dietterich, 2019; Sagawa et al., 2019; Nam et al., 2020; Bahng et al., 2020; Jeon et al., 2022b). Such transformations range from analytical ones like rotation, or corruptions like 'brightness' or 'contrast', to learned ones like adversarial attacks. Images with the same transformations form a domain, while each transformation can be an attribute, creating spurious correlations by matching labels to transformations. The benchmarks most similar to ours involve changing the original train-test split of the dataset to induce different types of DSs (Kim et al., 2019; Koh et al., 2021; Jeon et al., 2022a; Atanov et al., 2022). In contrast to these methods, our paper focuses on creating controllable concurrent shifts by using attribute annotations in existing datasets.

Generalization Analysis. Although methods that enhance robustness against DSs have been extensively researched, significant variations across application domains mean that a method excelling in one dataset might not perform equally well in another. Consequently, recent efforts have been dedicated to comprehensively and fairly evaluating generalization methods. In their detailed experiments, Gulrajani & Lopez-Paz (2020) demonstrated that empirical risk minimization, when meticulously implemented and finely tuned, excels over domain generalization methods in robustness. Taori et al. (2020) found that generalizability to synthetic shifts does not guarantee robustness to natural shifts. Koh et al. (2021) introduced **Wilds**, a curated benchmark consisting of 10 datasets that encapsulate a diverse range of DSs encountered in real-world settings. Their findings indicate that current methods for generalization are inadequately equipped to address real-world DSs. Wiles et al. (2022) reported that both pre-training and augmentations significantly boost performance across many scenarios, although the most effective methods vary with different datasets and DSs. Additionally, Miller et al. (2021) and Wenzel et al. (2022) observed a positive correlation between out-of-distribution and in-distribution performance. However, these studies predominantly concentrate on single DSs, which do not fully capture real-world scenarios.

3 Problem Statement

In this section, we introduce the concept of distribution shift (DS) and its two categories: unique distribution shift (**UniDS**) and concurrent distribution shift (**ConDS**). To operationalize the concept of **ConDS**, annotations for multiple attributes are required. For clarity, we begin by introducing one of our experimental datasets, dSprites, in Figure 2.



y^l: object shape

y^α: object color

y^β: background color

y^γ: object size

Figure 2: **dSprites** samples. Even in simple synthetic data, multiple attributes can potentially lead to various distribution shifts. Visualizations for other datasets are included in Figure 5 in the supplementary A.1.

3.1 Distribution Shift

Consider an instance set $X = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{X}, i = 1, \ldots, N\}$ for a classification problem, where $\mathcal{X}$ denotes the input space. Each instance $\mathbf{x}$ can be represented by a finite set of attributes $Y(\mathbf{x}) = \{y^k | y^k \in \mathcal{Y}, k = 1, \ldots, K\}$ where $\mathcal{Y}$ denotes attribute space and K varies with $\mathbf{x}$. Within this framework, one attribute y^l from $Y(\mathbf{x})$ can be designated as the label. Let p, p_S, and p_T be the true, source, and target distributions, respectively. We denote the source and target datasets by $\mathcal{D}_S \sim p_S$ and $\mathcal{D}_T \sim p_T$, respectively, where each dataset consists of samples from its corresponding distribution. A machine learning model $\hat{f}$ is designed to minimize the empirical risk,

$$\hat{f} := \arg\min_f \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i^l, f(\mathbf{x}_i)), \tag{1}$$

where $(\mathbf{x}_i, y_i^l) \in \mathcal{D}_S$ for $i = 1, ..., N$. The ideal objective is to find the model f^* that maximizes its performance on true distribution $p(\mathbf{x}, y)$; that is,

$$f^* := \arg\min_{\hat{f}} \mathbb{E}_{p(\mathbf{x}, y^l)} \left[\mathcal{L}(y^l, \hat{f}(\mathbf{x})) \right]. \tag{2}$$

However, a significant shift in distribution from source to target data, denoted by $\mathcal{D}_S \not\approx \mathcal{D}_T$, can cause a substantial decrease in performance during inference. Assuming that the target distribution p_T aligns with p, we define DSs in the following subsections.

3.2 Unique Distribution Shifts

We revisit spurious correlation (SC), low data drift (LDD), and unseen data shift (UDS) as delineated in the experimental framework by Wiles et al. (2022), grouping these under the category of unique distribution shifts (**UniDS**), namely **UniDS** := {SC, LDD, UDS}. To create the **UniDS**, we select one attribute from the candidates (y^α, y^β, y^γ) and then sample the images for each DS accordingly. All other attributes display a uniform distribution for **UniDS**.

Test distribution p_T. All the attributes $y^k \in Y(\mathbf{x})$ are uniformly distributed, ensuring that each attribute is represented and independent from the others. E.g., for $p_T \approx p$, there is an equal number of samples across 81 combinations, ranging from (square, red, orange, small) to (heart, blue, purple, big).

Spurious correlation. Under p_S, the label y^l and a specific attribute y^α exhibit correlation, which does not hold under p_T. E.g., there are only (square, red), (ellipse, yellow), and (heart, blue) samples in $\mathcal{D}_S$, exhibiting a correlation between 'shape' and 'color'.

Low data drift. Attribute values show an uneven distribution under p_S but are more evenly distributed under p_T. Generalization for LDD is also referred to as worst-case generalization (Sagawa et al., 2019; Seo et al., 2022). E.g., There are significantly fewer red samples compared to the numerous yellow and blue samples. Even among the red samples, the distribution of shapes is uneven.

Unseen data shift. Some attribute values that are not present under p_S appear under p_T. E.g., There are no blue images in $\mathcal{D}_S$, indicating that blue images in $\mathcal{D}_T$ are new to the model.

3.3 Concurrent Distribution Shifts

Previous benchmarking studies have predominantly focused on **UniDS**. However, in real-world contexts, a dataset could contain concurrent DSs. Given that an instance $\mathbf{x}$ encompasses numerous attributes, DSs can exhibit more complex structures, with each attribute governed by distinct **UniDS**. For instance, in the real-world dataset iWildCam, UDS are observed due to discrepancies in the camera traps used across source and target data. Concurrently, LDD emerges from variable image counts across animal classes. Therefore, to effectively assess the efficacy of methods in real-world applications, it is crucial to conduct a comprehensive evaluation that addresses these scenarios.

To reflect the complexities observed in real-world datasets, we introduce the concept of "concurrent distribution shift (**ConDS**)". This novel framework extends the conventional **UniDS** by modeling the diverse shifts associated with different attributes of a dataset. In this study, we focus on combining various shifts attributed to different attributes, rather than evaluating a single type of shift across multiple attributes—a topic extensively covered in previous literature.

In our formal definition, a **ConDS** is conceptualized as the ensemble of all possible combinations of two or more distinct **UniDS** elements,

$$\textbf{ConDS} := \{S \subseteq \textbf{UniDS} \mid |S| \geq 2\}, \tag{3}$$

where each subset S represents a unique configuration of shifts, creating a richer and more representative model of the underlying complexities in real-world data. Each pair (y^l, y^k) within this framework, where k belongs to a predefined set $\{\alpha, \beta, \gamma\}$, is governed by specific **UniDS**.

For instance, using the attributes shown in Figure 2, (SC, UDS) can be established by sampling combinations such as (square, red), (ellipse, yellow), and (heart, blue), while the background color varies between (orange, green) and the object size maintains a uniform distribution. The concept of **ConDS** and its importance were also introduced in Koh et al. (2021). However, their discussion was limited to LDD and UDS. We broaden this to include SC, LDD, and UDS, aiming for a thorough understanding of their interconnections. Additionally, we investigate a broader array of algorithms to assess how existing methods perform on **ConDS**.

4 Generalization to Distribution Shifts

As discussed in Section 3.1, it is assumed that during training, we only have access to source data, while the target distribution remains unknown (Vapnik, 1991; Jeon et al., 2023). While some algorithms are specifically designed for scenarios where partial knowledge of the target distribution is available (Adeli et al., 2021; Alvi et al., 2018; Kim et al., 2019; Geirhos et al., 2018; Bahng et al., 2020; Ganin et al., 2016), the setup where no target information is known generally poses a broader challenge for model generalization. Consequently, our analysis focuses on this more general setup. We evaluate 26 algorithms suitable for this scenario, spanning a wide spectrum of methods, as shown in Table 1.

Table 1: **Generalization algorithms evaluated.** We outline the methods evaluated and the distribution shifts they are designed to address.

Generalization Algorithms		*SC*	*LDD*	*UDS*
Architecture	ResNet18, ResNet50, ResNet101 (He et al., 2016), ViT (Dosovitskiy et al., 2020), MLP (Vapnik, 1991).			
Heuristic augmentation	Imagenet (He et al., 2016), AugMix (Hendrycks et al., 2019), RandAug (Cubuk et al., 2020), AutoAug (Cubuk et al., 2019).			✓
De-biasing	UBNet (Jeon et al., 2022a), PnD (Li et al., 2023), OccamNets (Shrestha et al., 2022).	✓		
Worst-case generalization	GroupDRO (Sagawa et al., 2019), BPA (Seo et al., 2022).		✓	
Single domain generalization	ADA (Volpi et al., 2018), ME-ADA (Zhao et al., 2020), SagNet (Nam et al., 2021), L2D (Wang et al., 2021).			✓
Out-of-distribution generalization	IRM (Arjovsky et al., 2019), CausIRL (Chevalley et al., 2022).	✓	✓	✓
Zero-shot inference with foundation model	CLIP (Radford et al., 2021), InstructBLIP (Dai et al., 2024), LLaVA-1.5 (Liu et al., 2023), Phi-3.5-vision (Abdin et al., 2024).			

Augmentation for generalization. Although some augmentation techniques are not specifically designed for robustness, they are commonly used to enhance generalization against the unseen domain and are well-known for their effectiveness (Yan et al., 2020; Wiles et al., 2022; Cugu et al., 2022; Zheng et al., 2024). The operative idea is that augmentation expands the input space, thereby increasing the likelihood that the model will recognize new test samples.

Generalization algorithms. *De-biasing*: UBNet utilizes feature maps from lower to higher layers to enable the model to access a broader feature space. PnD removes spurious correlations by detecting them using GCE (Zhang & Sabuncu, 2018). OccamNets employs architectural inductive biases, demonstrating their effectiveness in mitigating spurious correlations. *Worst-case generalization*: GroupDRO and BPA modify the weight of gradient flow to ensure balance across different groups. *Single domain generalization*: ADA and ME-ADA generate additional inputs by incorporating adversarial noise. SagNet is designed to be agnostic to styles, emphasizing content instead. L2D utilizes learned augmentations with AdaIN from StyleGAN (Karras et al., 2019). *Out-of-distribution generalization*: IRM and CausIR eliminate variant features by regularizing discrepancies among samples from the source data.

Foundation model for image classification. Foundation models acquire robust representations from comprehensive datasets, facilitating their application in downstream tasks such as image classification. The mechanisms for aligning inputs to outputs ($\mathcal{X} \rightarrow \mathcal{Y}$) during zero-shot inference differ among models. CLIP utilizes the similarity between image embeddings and textual label embeddings to calculate confidence scores. Conversely, InstructBLIP, LLaVA-1.5, and Phi-3.5-vision employ label-eliciting prompts as queries within the framework of visual question answering (Chen et al., 2022). Additional details about the prompts used for image classification can be found in Table 10 of the Appendix B.4.

5 Experiments

In this section, we begin by presenting the datasets we evaluated on and experimental setup. Next, we evaluate 26 distinct algorithms across 168 (source, target) pairs spanning six datasets, addressing both **UniDS** and **ConDS**. Aggregate results are illustrated in Figure 3, while comprehensive results are detailed in the supplementary B.1. We further examine how the challenges vary among different DSs (Figure 4) and evaluate the effectiveness of pre-training in enhancing robustness (Table 3). Further analysis investigates how zero-shot inference performance depends on distribution shifts (DSs), as shown in Table 5. Finally, we analyze the outcomes, summarizing them into eight key takeaways.

5.1 Datasets

Controlled datasets: We assess algorithms using five evaluation datasets: **dSprites**, **Shapes3D**, **SmallNorb**, **CelebA**, and **DeepFashion**. From these, we select four attributes: one is designated as the label (y^l), and the other three as attributes (y^α, y^β, y^γ) to create DSs. Table 2 list the attribute instances for y^l and $\{y^\alpha, y^\beta, y^\gamma\}$ for the different controlled datasets. We divide the source data into training and validation sets, both sharing the same distribution, with the validation set used for hyperparameter tuning.

Uncontrolled real-world datasets: We use **iWildCam**, **fMoW**, and **Camelyon17** for evaluation. **iWildCam** data in the Wilds benchmark (Koh et al., 2021) exhibits LDD over the animal distributions, and UDS occurs across camera trap locations. Similarly, the **fMoW** dataset (Koh et al., 2021) exhibits UDS and LDD across time and regions in satellite images. **Camelyon17** is a tumor detection dataset with various unexpected DSs resulting from different hospitals. We further discuss additional insights from the categorization of synthetic (**Dsprites**, **Shapes3D**, **Smallnorb**) and real-world datasets (**CelebA**, **DeepFashion**, **iWildCam**, **fMoW**, **Camelyon17**) in supplementary A.3.

Table 2: **Attributes for controlled datasets.** (y^l, y^α, y^β, y^γ) for datasets are (object shape, object color, background color, object size), (object shape, object color, background color, object size), (object, azimuth, lightening, elevation), (gender, black hair, smiling, straight hair), and (dress, pattern, sleeve, fabric), respectively. Image samples from each dataset are displayed in Figure 5 in the supplementary A.1.

	dSprites	**Shapes3D**	**SmallNorb**	**CelebA**	**DeepFashion**
y^l	{square, ellipse, heart}	{cube, cylinder, sphere, capsule}	{animal, human, car, truck, airplane}	{male, female}	{skirt, others}
y^α	{red, yellow, blue}	{red, orange, yellow, green}	{0, 80, 160, 240, 320}	{black, others}	{floral, solid}
y^β	{orange, green, purple}	{red, orange, yellow, green}	{0, 1, 2, 3, 4}	{smiling, no smiling}	{long sleeve, sleeveless}
y^γ	{small, middle, big}	{tiny, small, middle, big}	{30, 40, 50, 60, 70}	{straight hair, others}	{chiffon, cotton}

5.2 Experimental Setup and Results

Controlled datasets: To comprehensively understand the generalizability of the algorithms on diverse DSs, we evaluate with three different seeds for five datasets changing attributes $\{y^\alpha, y^\beta, y^\gamma\}$. Specifically, with labels fixed to y^l, we manipulate attributes to simulate different DSs. For instance, we select one attribute from $\{y^\alpha, y^\beta, y^\gamma\}$ to create **UniDS** (3 settings). For **ConDS**, we use six combinations (either ${}_3C_2$ or ${}_3C_3$ settings) to establish designated DSs, resulting in a total of 165 (source, target) pairs. E.g., for the (SC, LDD) setting, we choose one attribute to define SC and another from the remaining two to create LDD, resulting in 6 (source, target) pairs. The detailed description for this is provided in supplementary A.1. For all SC, we include 1% of counterexamples, in accordance with the setup established in research on SC robustness (Jeon et al., 2022a; Nam et al., 2020). We set ResNet18 as a backbone for all the algorithms. More details about hyperparameters are provided in supplementary A.5. **Real-world datasets:** We adhered to the use of the train, validation, and test sets for **iWildCam**, **fMoW**, and **Camelyon17**. Real-world datasets do not exhibit a clear distribution shift like controlled datasets, but they inherently contain various naturally occurring distribution shifts that may go unnoticed.

Results. Figure 3 exhibits the aggregate result of algorithms across DSs for the controlled datasets. Figure 8 in supplementary and Table 3 present the outcomes when methods are trained using ImageNet pre-trained weights. In Table 3, some algorithms, particularly augmentation techniques, surpass large models in scenarios without SC, an advantage not observed when training from scratch (see Figure 3). Table 3 indicates that pre-training enhances performance for all the algorithms and DSs. In our experiments, while foundation models perform well on controlled datasets, their effectiveness is often limited when applied to real-world datasets.

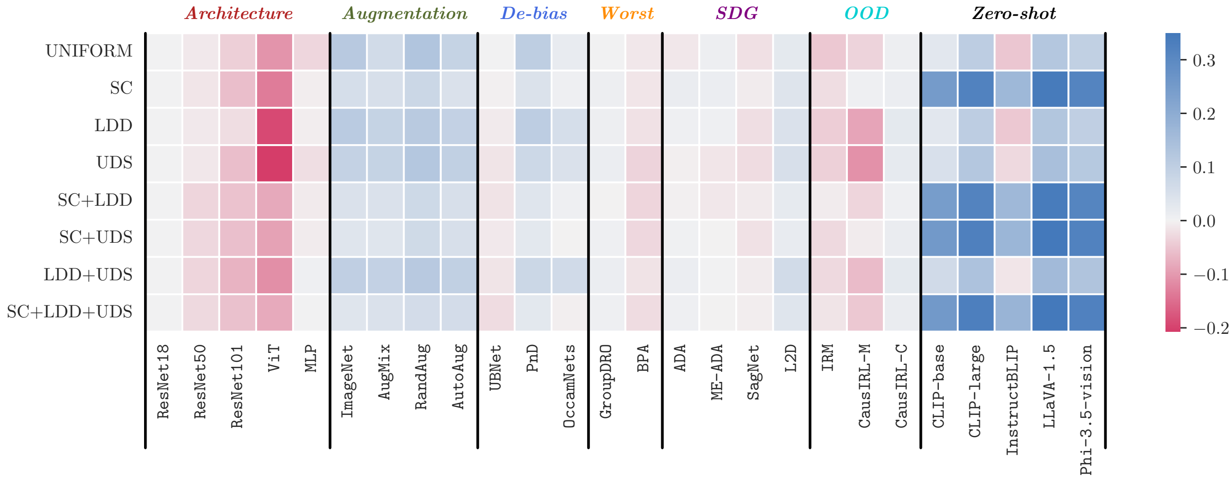


Figure 3: **Aggregate result on controlled datasets**. We plot the change in accuracy compared to the base model, ResNet18, averaged across all seeds and controlled datasets with varying attributes. Blue indicates improved performance, while red indicates a decline. Each row is independent of the others. The models used for *zero-shot inference* were only used for evaluation, thus, they have the same absolute performance for each row. However, as we show their accuracies relative to the ResNet18, the relative performance for each model is not the same for each row. We fine-tune all the algorithms and report their optimal results. Augmentation methods and zero-shot models perform well under the different types of shifts. We provide a breakdown of the accuracy for each algorithm and dataset in the supplementary B.1.

Comparison of distribution shifts. In Figure 4, we evaluate the challenges associated with varying numbers of DSs. To ensure a consistent comparison across these shifts, we standardize the sizes of the datasets. However, due to limitations in data availability that prevent standardization while adhering to specific DSs, our analysis primarily relies on the **Dsprites** and **CelebA**. As illustrated in Figure 4, although the overall performance of algorithms tends to degrade as the number of DSs increases, SC remains inherently challenging and shows almost no performance decline when combined with other DSs. Additionally, SC is significantly more challenging than other distribution shifts, even when considering multiple shifts combined, such as LDD+UDS.

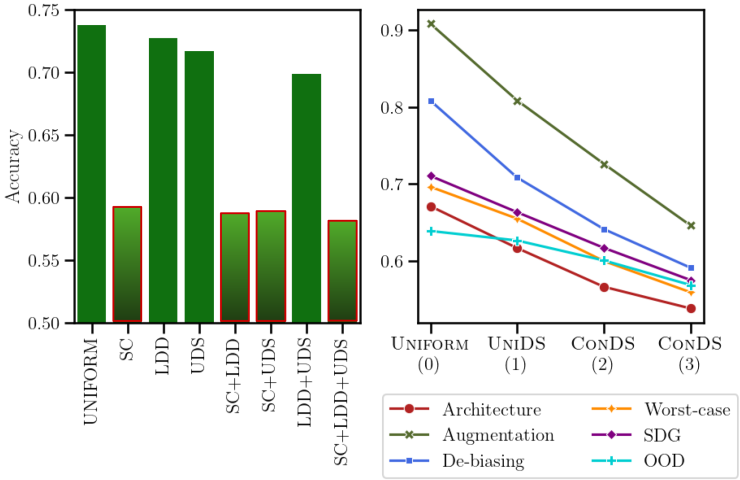


Figure 4: **Analysis of model robustness on distribution shifts.** *Left:* Average performance of all generalization methods under different combinations of DSs. *Right:* Comparing the different generalization methods under an increasing number of DSs.

Table 3: **Training from scratch *vs* Pre-training.** All the values reported for controlled datasets represent the average accuracy across all DSs and controlled datasets. We compare the performance of models trained from scratch with those pre-trained, bolding the better result between the two training strategies. The top average result in each column except *zero-shot inference* is indicated with a gray box, while the second-best is underlined. We used the original models for *zero-shot inference* for evaluations, thus, they are neither trained from scratch nor fine-tuned. **Trends on controlled datasets *vs* Real-world datasets.** We also show the results from evaluating on **iWildCam**, **fMoW**, and **Camelyon17**, as they are challenging real-world datasets with **ConDS**. The experimental results from our controlled datasets align with those from the real-world datasets, demonstrating that even on real-world datasets, *augmentation and pre-training significantly enhance robustness.*

		Controlled Dataset		**Real-world Dataset**	
		Scratch	*Pre-training*	*Scratch*	*Pre-training*
Architecture	ResNet18	62.43	**81.53**	53.91	**61.64**
	ResNet50	60.13	**82.64**	52.58	**67.76**
	ResNet101	56.98	**79.96**	47.01	**67.52**
	ViT	49.55	**78.53**	50.55	**70.94**
	avg.	*57.2 7*	*80.67*	51.01	66.97
Augmentation	*ImageNet*	69.25	**80.55**	48.82	**66.48**
	AugMix	68.99	**83.15**	51.99	**68.09**
	RandomAug	71.47	**85.28**	51.94	**68.29**
	AutoAug	69.65	**85.85**	52.15	**70.29**
	avg.	*69.84*	*83.71*	*51.22*	*68.29*
De-biasing	UBNet	61.08	**78.26**	41.01	**59.03**
	PnD	68.15	**77.67**	50.02	**60.34**
	OccamNets	64.98	**77.59**	52.57	**65.20**
	avg.	*64.74*	*77.84*	*47.87*	*61.52*
Worst-case generalization	GroupDRO	63.26	**81.23**	48.17	**64.40**
	BPA	59.99	**75.69**	29.44	**62.15**
	avg.	*61.63*	*78.46*	*38.81*	*63.28*
Single domain generalization	ADA	62.98	**80.20**	52.88	**69.71**
	ME-ADA	62.37	**78.33**	53.79	**66.41**
	SagNet	61.09	**81.56**	54.27	**66.75**
	L2D	66.42	**82.63**	52.67	**68.20**
	avg.	*63.22*	*80.68*	*53.40*	*67.77*
Out-of-distribution generalization	IRM	59.83	**80.27**	39.47	**57.90**
	CausIRL-M	57.55	**80.81**	48.90	**65.87**
	CausIRL-C	64.17	**82.02**	54.38	**68.06**
	avg.	*60.52*	*81.03*	*47.58*	*63.94*
Zero-shot inference	CLIP-base	78.92		25.76	
	CLIP-large	86.22		29.06	
	InstructBLIP	70.95		29.35	
	LLaVA-1.5	88.24		24.02	
	Phi-3.5-vision	85.63		29.17	
	avg.	*81.99*		*27.47*	

Further analysis on prompts of foundation model. We investigate various prompts to assess the effectiveness of foundation models for robust image classification. We assess the accuracy of the models in Table 5. For CLIP, we use widely adopted prompts, such as "a photo of a *label*" (Matsuura et al.). Given that vision-language models designed for visual question answering may depend on the query's context, the choice of prompts is crucial (Sahoo et al., 2024). We categorize prompts that are applicable to any dataset as 'General', and those specifically designed for each dataset as 'Tailored'. We provide several examples for the prompt in Table 4. In Matsuura et al. and Islam et al. (2023), diverse prompts are employed as queries

Table 4: **An example of zero-shot inference prompts.** The first two rows show the 'General' prompts and the next three rows the 'Tailored' prompts that we used with LLaVA-1.5. For a more comprehensive list of these prompts, please see Table 10 in the supplementary B.4.

Type	***Prompt***
General$_1$	*USER: <image>* *Classify the image into* $label_1, \dots,$ *or* $label_C$*. Please provide only the name of the label.*
General$_2$	*USER: <image>* *Choose a label that best describes the image. Here is the list of labels to choose from:* $label_1, \dots, label_C$*.* *Please provide only the name of the label.*
dSprites **Shapes3D** **SmallNorb**	*USER: <image>* *Classify the object in the image into* $label_1$*, ..., or* $label_C$*. Please provide only the name of the label.*
CelebA	*USER: <image>* *Classify the person in the image into* $label_1$ *or* $label_2$*. Please provide only the name of the label.*
DeepFashion	*USER: <image>* *Is a person wearing a dress or not? Please answer in yes or no.*

Table 5: **Comparison of the performance of visual language models** across different types of prompts. We report the best results obtained from this ablation study in Figure 3, 8 and Table 3. Note how sensitive the results are to the prompts used. There can be a drop in accuracy of 37% with a suboptimal prompt.

Model	**Prompt**	**dSprites**	**Shapes3D**	**SmallNorb**	**CelebA**	**DeepFashion**	**iWildCam**	**fMoW**	**Camelyon17**
CLIP-base	General	75.19	71.41	72.02	98.5	77.50	13.97	13.30	50.01
CLIP-large	General	90.86	85.39	86.00	97.62	71.25	13.70	23.46	50.01
InstructBLIP	General$_1$	50.25	50.08	26.59	**98.63**	61.25	1.32	11.61	50.57
	General$_2$	53.46	53.05	**36.22**	70.50	47.50	**1.86**	**18.17**	**68.03**
	Tailored	**58.89**	**74.77**	25.41	78.25	**86.25**	1.10	9.67	50.00
LLaVA-1.5	General$_1$	49.01	52.89	**90.05**	98.50	70.00	**4.64**	15.04	50.57
	General$_2$	72.72	**74.61**	87.81	**98.88**	**91.25**	1.60	**16.32**	**51.09**
	Tailored	**86.42**	71.02	88.16	**98.88**	88.75	1.85	14.24	50.00
Phi-3.5-vision	General$_1$	88.27	**72.73**	83.20	98.62	**82.50**	5.88	9.67	64.86
	General$_2$	88.27	72.50	**85.41**	**98.75**	78.75	4.06	**10.88**	57.09
	Tailored	**88.77**	72.58	83.30	98.62	53.75	**9.91**	9.17	**66.71**

for vision-language models, specifically tailored to the dataset and the labels targeted for classification. Our analysis builds on this approach.

5.3 Takeaways

Takeaway 1 – ConDS is, *on average*, more challenging than UniDS. While previous studies on model generalization have primarily focused on **UniDS**, we observe that most algorithms exhibit poorer performance in **ConDS**. Moreover, the more numerous the DSs, the greater the challenge they pose, as illustrated in Figure 4 (right).

Takeaway 2 – SC is the most challenging DS, followed by UDS and LDD. Figure 4 (left) breaks down the performance of generalization methods according to DS. We see that the presence of SC tends to dominate over other DSs in **ConDS**. Although **ConDS** presents more challenges than **UniDS** for LDD and UDS, there is almost no performance drop when moving from SC to SC+LDD or SC+UDS. Furthemore, for the DSs that include SC, the performance of most methods is inferior to that of foundation models even when it is pre-trained, as shown in Figure 8 in supplementary.

Takeaway 3 – Generalization tends to be consistent across DSs. If a method improves generalization for one DS, it tends to be effective for others. Namely, although models such as *de-biasing*, *worst-case generalization*, and *domain generalization* are designed to address a specific DS (as detailed in Section 4), their applicability is not confined to that particular shift.

5.3.1 Takeaways Common to both UniDS and ConDS

The following takeaways with **UniDS** have been previously identified by Sahoo et al. (2024); Wiles et al. (2021). We observe that these phenomena persist even in the **ConDS** scenario.

Takeaway 4 – Heuristic augmentations and pre-training are highly effective. As depicted in Figure 3, heuristic augmentations overall improve the model's robustness across all DSs. This remains consistent when pre-trained weights are utilized except for ImageNet augmentation (Figure 8 in the supplementary). All the augmentation methods even increase the performance of ResNet18 when used on Uniform, where there is no DS, establishing them as a more effective tool. Algorithms demonstrate improved performance with ImageNet pre-trained weights, as shown in Table 3. On LDD, UDS, and LDD+UDS, most methods even outperform foundation models, as depicted in Figure 8.

Takeaway 5 – Generalization algorithms provide a limited performance improvement. While some methods (L2D, PnD, OccamNets) perform well, most methods exhibit limitations. Existing generalization methods, though competitive in their original experimental setups and benchmarks, demonstrate reduced effectiveness in our more standardized and fair setup.

Takeaway 6 – Foundation pre-trained models are effective, but their performance can vary. Although foundation models demonstrate impressive performance on controlled datasets, they show limited performance compared to other baselines with ImageNet-pretrained weights, as shown in Figure 8. Notably, they exhibit the lowest accuracy on real-world datasets (Table 3). This observation underscores that the performance in image classification significantly depends on the datasets used to train these foundation models, highlighting potential limitations in their applicability. E.g., for specialized datasets like **Camelyon17**, large models with zero-shot inference can barely make accurate predictions (Table 10). Furthermore, we find that their performances are significantly influenced by the prompt we input (Please refer Table 5 and Table 10 in the supplementary B.4 for more details).

6 Concluding Remark

Contribution. In this paper, we introduce a novel evaluation framework to understand the robustness of models against various distribution shifts, including **UniDS** and **ConDS**. Using this protocol, we can create distribution shifts in any multi-attribute-annotated dataset, allowing for a more comprehensive understanding of robustness. Through our extensive evaluation involving 56K experiments, we find that **ConDS** present greater challenges compared to conventional **UniDS**, with spurious correlations being more problematic than low data drift and unseen domain shifts. Our results indicate that heuristic augmentations and pre-training are effective tools for enhancing generalization, while more complex models offer limited benefits. Additionally, while large models are viable for image classification, their performance is effective only in specific scenarios and requires careful application.

Limitation and future work. While we believe our work makes a promising step towards understanding how models behave under complex scenarios, there is still a lot more that can be done in this direction, we briefly discuss some of them. Our evaluation framework allows us to create controlled distribution shifts and assumes a uniform test distribution. Our framework also uses annotated, thus, interpretable attributes to create shifts. Using learned attributes would also be an interesting future direction. Furthermore, our study covers a limited range of attributes, particularly in controlled real-world datasets such as **CelebA** and **DeepFashion**, due to insufficient samples for other attributes. Future research could explore the use of advanced controllable generative models (e.g., diffusion models) to address this limitation and cover a broader range of conditions.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M Pohl. Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2513–2523, 2021.

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Andrei Atanov, Andrei Filatov, Teresa Yeo, Ajay Sohmshetty, and Amir Zamir. Task discovery: Finding the tasks that neural networks generalize on. *Advances in Neural Information Processing Systems*, 35: 15702–15717, 2022.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pp. 528–539. PMLR, 2020.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Mathieu Chevalley, Charlotte Bunne, Andreas Krause, and Stefan Bauer. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646*, 2022.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Ilke Cugu, Massimiliano Mancini, Yanbei Chen, and Zeynep Akata. Attention consistency on visual corruptions for single-source domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4165–4174, 2022.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

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

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

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

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

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Ashhadul Islam, Md Rafiul Biswas, Wajdi Zaghouani, Samir Brahim Belhaouari, and Zubair Shah. Pushing boundaries: Exploring zero shot object classification with large multimodal models. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 1–5. IEEE, 2023.

Myeongho Jeon, Daekyung Kim, Woochul Lee, Myungjoo Kang, and Joonseok Lee. A conservative approach for unbiased learning on unknown biases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16752–16760, 2022a.

Myeongho Jeon, Hyoje Lee, Yedarm Seong, and Myungjoo Kang. Learning without prejudices: continual unbiased learning via benign and malignant forgetting. In *The Eleventh International Conference on Learning Representations*, 2022b.

Myeongho Jeon, Myungjoo Kang, and Joonseok Lee. A unified framework for robustness on diverse sampling errors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1464–1472, 2023.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9012–9020, 2019.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Jiaxuan Li, Duc Minh Vo, and Hideki Nakayama. Partition-and-debias: Agnostic biases mitigation via a mixture of biases-specific experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4924–4934, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

Misaki Matsuura, Young Kyun Jung, and Ser Nam Lim. Visual-llm zero-shot classification.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pp. 7721–7735. PMLR, 2021.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8690–8699, 2021.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16742–16751, 2022.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. Occamnets: Mitigating dataset bias by favoring simpler hypotheses. In *European Conference on Computer Vision*, pp. 702–721. Springer, 2022.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

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

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 834–843, 2021.

Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, 35:7181–7198, 2022.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

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

Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Guangtao Zheng, Mengdi Huai, and Aidong Zhang. Advst: Revisiting data augmentations for single domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21832–21840, 2024.

A Detailed Experimental Setup

A.1 Setup for Controlled Distribution Shifts

We utilize multi-attribute datasets such as **dSprites**, **Shapes3D**, **SmallNorb**, **CelebA**, and **DeepFashion** to develop **ConDS**. These datasets enable us to sample images annotated with various attributes, showcasing a range of DSs from SC to (SC, LDD, UDS). Additionally, we configure different attributes for each type of DS. For instance, in SC, pairs like (y^{α}, y^{l}) and (y^{l}, y^{β}) may exhibit spurious correlations. By covering all possible attribute combinations (with the attributes depicted in Figure 5), we ensure a more comprehensive evaluation of scenarios.

For **UniDS**, we choose one attribute from three options, yielding three settings. In **ConDS**, we address two combinations: (SC, LDD), (SC, UDS), and (LDD, UDS). For each combination, one attribute is selected for the first DS and another for the second, following a ${}_3C_2$ selection method. For the three-attribute combination in **ConDS**, namely (SC, LDD, UDS), one attribute is chosen for the first DS, another for the second, and the remaining for the last DS, in line with a ${}_3C_3$ approach. With five datasets in a controlled setup, this generates a total of 165 (source, target) pairs.

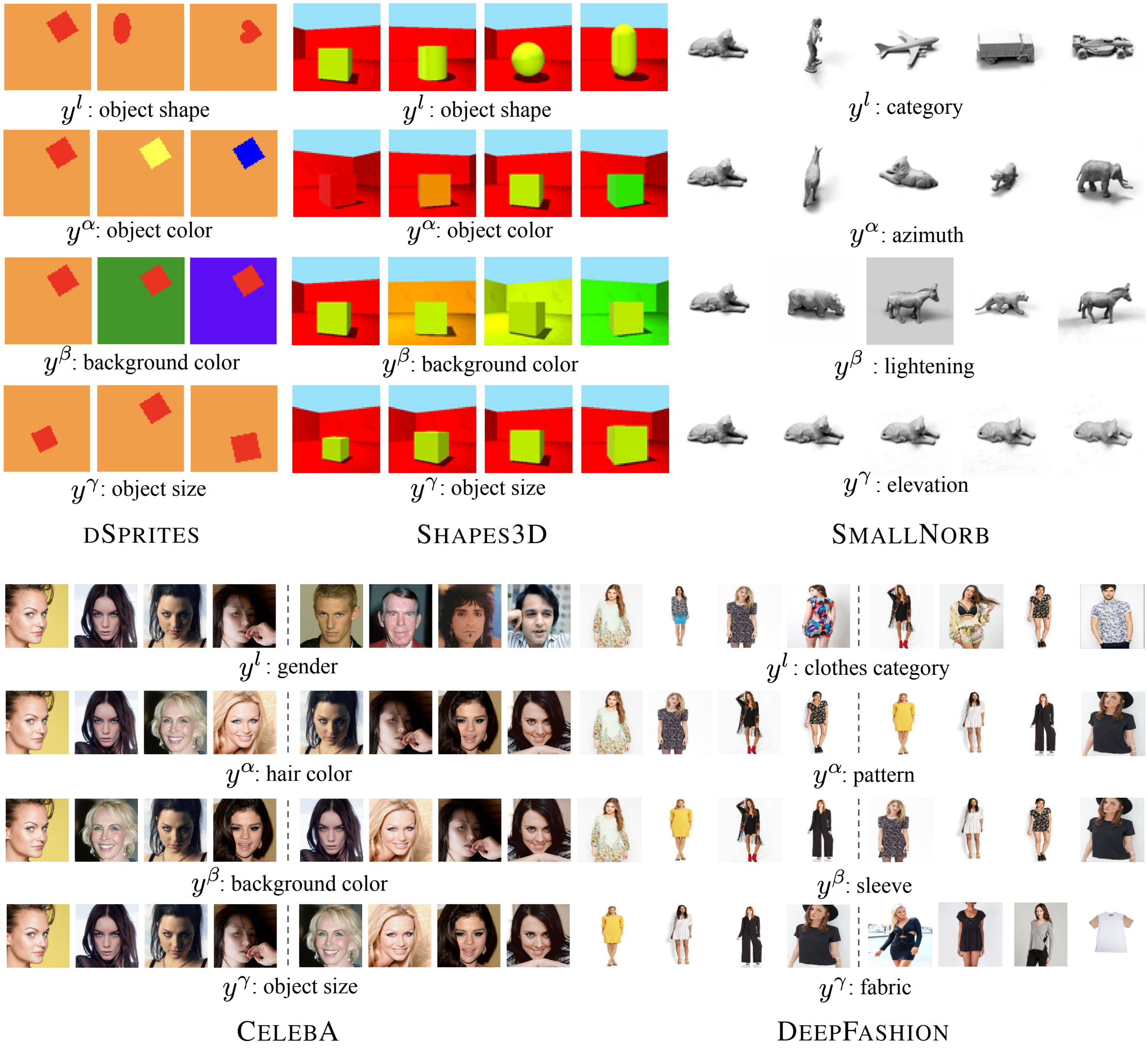


Figure 5: **Samples of controlled datasets.** We provide visualizations of some samples along with their attributes.

A.2 Dataset Configuration

We provide detailed counts for each split within the datasets. To create the DSs as outlined in Section 3, variations in dataset sizes across the DSs result from the limited availability of certain attribute combinations.

Table 6: **Dataset size.** Please note that we have included 1% counterexamples for SC in our input.

	dSprites	**Shapes3D**	**SmallNorb**	**CelebA**	**DeepFashion**
UNIFORM	1,091	65	127	227	97
SC	1,091	65	127	227	97
LDD	1,080	624	1,600	224	112
UDS	1,080	192	500	224	96
SC + LDD	1,091	158	324	227	99
SC + UDS	1,091	49	101	227	101
LDD + UDS	1,080	468	1,280	224	98
SC+ LDD + UDS	1,091	119	259	227	92
TEST	810	1,280	3,125	800	80

A.3 Synthetic and Real-world Datasets

We can precisely manage distribution shifts with synthetic datasets because they allow for complete control. Real-world datasets, on the other hand, offer the advantage of realism but may experience uncontrolled distribution shifts. For instance, suppose we create a gender classifier with a dataset that exhibits a spurious correlation by sampling images of older females and younger males. Despite our efforts, we cannot ensure uniformity across other attributes such as skin tone and gender. Given the advantages and disadvantages of each, we conducted experiments with a variety of synthetic datasets (**Dsprites**, **Shapes3D**, **Smallnorb**) and realistic datasets (**CelebA**, **DeepFashion**, **iWildCam**, **fMoW**, **Camelyon17**).

A.4 Computing Resource

All our experiments were performed using 8 NVIDIA H100 80GB HBM3 GPUs and Intel(R) Xeon(R) Gold 6448Y.

A.5 Implementation Details

To generate all the results reported in this script, we fine-tuned the hyperparameters. For the controlled datasets, we adopt early stopping where training stops early once the patience limit is reached. Validation accuracy is measured every 100 iterations and one patience is consumed if the best validation accuracy does not improve. The specific values are detailed in Table 7 and Table 8. We conducted a grid search with these parameters to optimize results for each algorithm across all DSs and datasets.

Table 7: **General Hyperparameters.**

Dataset	Learning Rate	Batch Size	Duration	Patience	Optimizer
Controlled	[1e-2, 1e-3, 1e-4]	128	10000 iterations	20	Adam
Real-world	[1e-3, 5e-4, 1e-4, 5e-5, 1e-5]	128	10000 iterations	20	Adam

Table 8: **Hyperparameters for specific methods.**

Method	Dataset	Hyperparameters
ViT	All	hidden dim: 256, MLP dim: 512, #heads: 8, #layers: 8
	Controlled	patch size: 4 (**dSprites**/ **Shapes3D**), 12 (**SmallNorb**), 16 (**CelebA**)
	Real-world	patch size: 16
MLP	All	#layers: 4, hidden dim: 256
ImageNet	**Controlled**	scale lower bound: 0.08
	Real-world	scale lower bound: 0.3
AugMix	**Controlled**	severity: 3, mixture_width: 3
	Real-world	severity: 7, mixture width: 5
RandAug	**Controlled**	num_ops: 3, magnitude: 5
	Real-world	num_ops: 2, magnitude: 9
AutoAug	All	CIFAR10 policy for **dSprites**, **Shapes3D**, and **SmallNorb**. Otherwise, ImageNet policy
UBNet	All	base model patience: 10, base model training epochs: 50
PnD	All	α_1: 0.2, α_2: 2, β: 4, q: 0.7, EMA$_\alpha$: 0.7, base model patience: 10, base model training epochs: 50
OccamNets	All	λ_{CS}: 0.1, λ_G: 1, γ_0: 3, γ: 1, $\tau_{acc,0}$: 0.5
	fMoW:	$\tau_{acc,0}$: 0.2
GroupDRO	All	λ: 1e-2
BPA	All	k: 8, m: 0.3, base model patience: 10, base model training epochs: 50
	iWildcam	k: 12
ADA	All	T_{max}: 15, k: 2, γ: 1.0
ME-ADA	All	T_{max}: 15, k: 2, γ: 1, η: 1.0
SagNet	All	λ_{adv}: 0.1
L2D	All	α_1: 1, α_2: 1, β: 0.1
IRM	All	annealing iterations: 500, λ: 1 (during annealing), 100 (after annealing)
CausIRL	**Controlled**	γ: 0.5 (**SmallNorb**), 0.3 (**dSprites**), 0.1 (**DeepFashion**, **CelebA**), 1 (**Shapes3D**)
	Real-world	γ: 1
CLIP-base	All	model: "openai/clip-vit-base-patch32"
CLIP-Large	All	model: "openai/clip-vit-large-patch14"
InstructBLIP	All	model: "Salesforce/instructblip-vicuna-7b",num_beam: 5, load_in_4bit: True top_p: 0.9, repetition_penalty: 1.5, length_penalty: 1.0, temperature: 1, max_new_tokens: 20
LLaVA1.5	All	model: "llava-hf/llava-1.5-7b-hf", max_new_tokens: 200
Phi-3.5-vision	All	model: "microsoft/Phi-3.5-vision-instruct", max_new_tokens: 200, temperature: 0.2

B Experimental Results

B.1 Comprehensive Results

Figure 6 displays the results for all algorithms trained from scratch across all DSs, while Figure 7 presents the outcomes for all algorithms trained using ImageNet pre-trained weights.

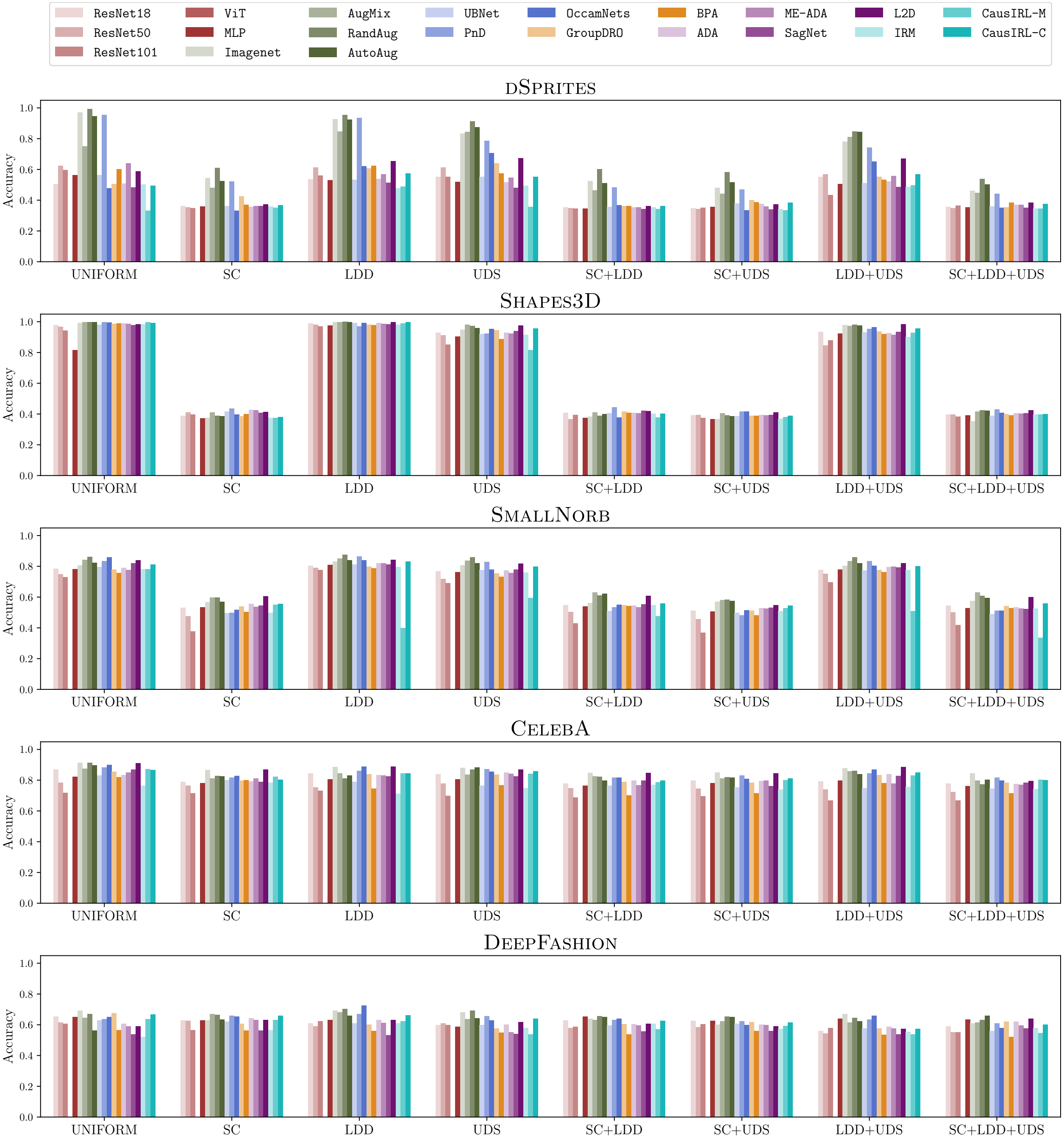


Figure 6: **Results for all algorithms from scratch.** We plot the top-1 accuracy.

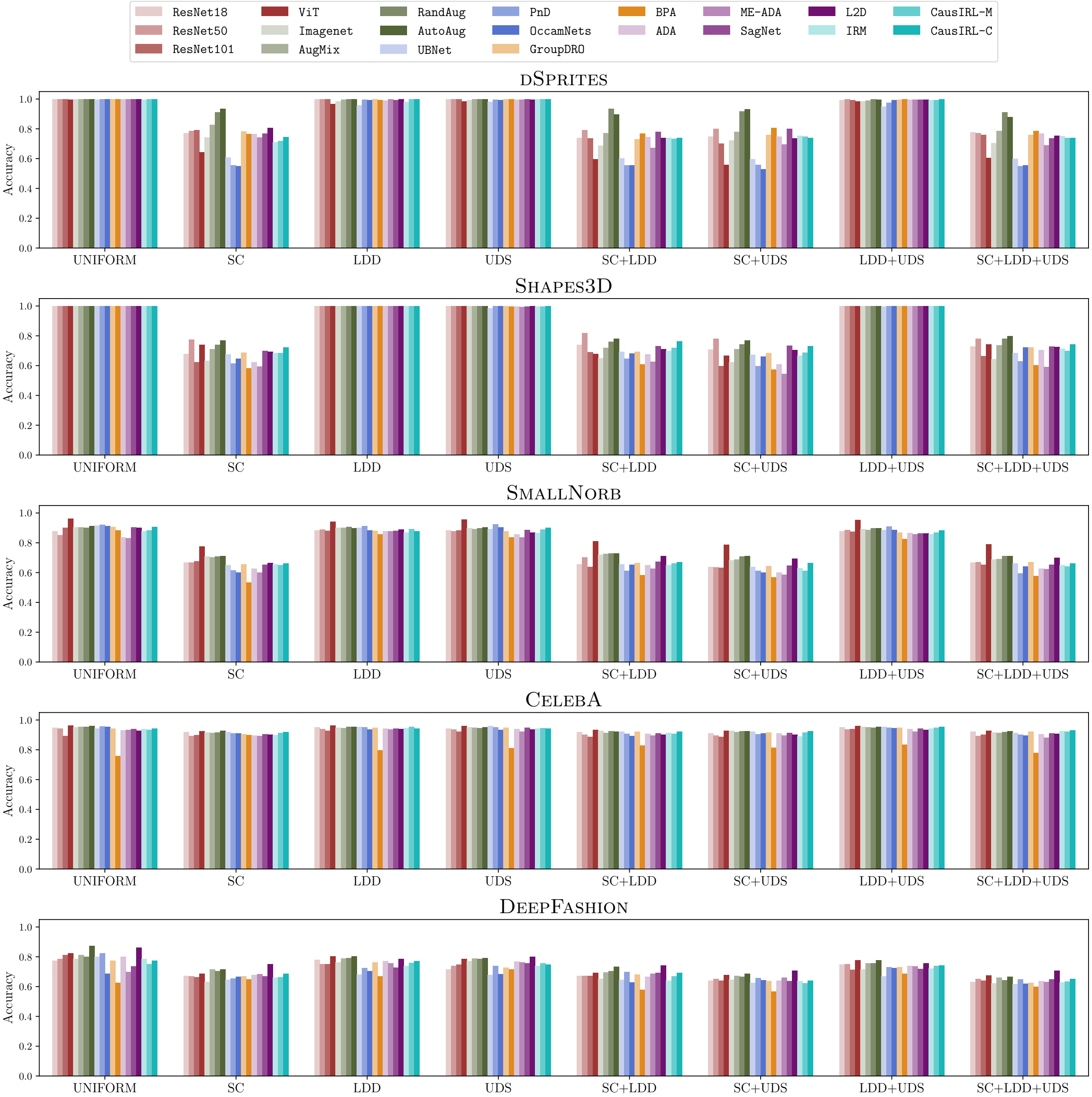


Figure 7: **Results for all algorithms with pre-trained weight.** We plot the top-1 accuracy.

B.2 Training from the scratch *vs* Pre-training

Figure 8 presents the aggregate results of pre-training across all algorithms and DSs.

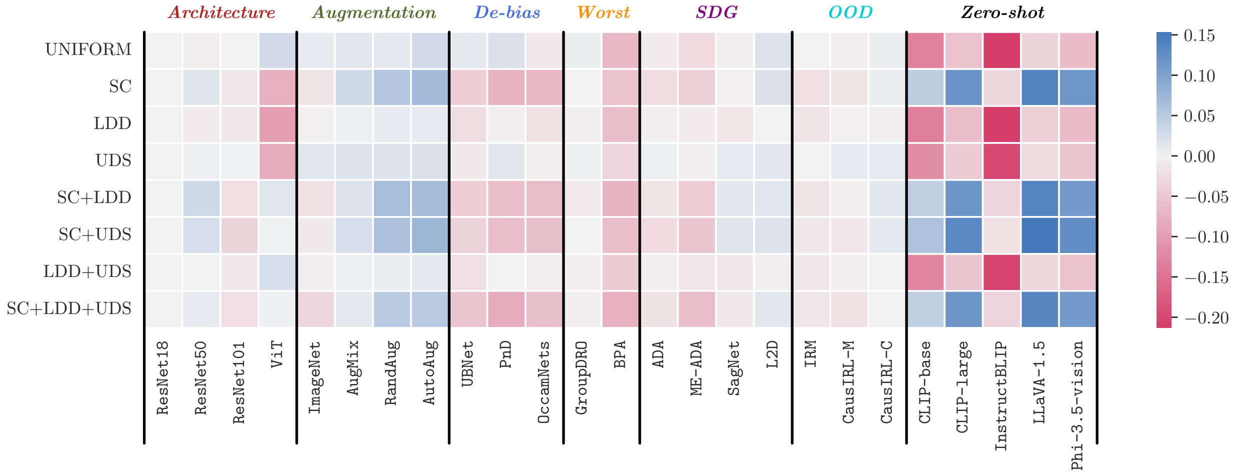


Figure 8: **Result with ImageNet pre-trained weight**. The setup is the same as Figure 3.

B.3 Results for Real-world datasets

Table 9 presents the detailed results for each dataset listed in Table 3.

Table 9: **Experimental results on real-world datasets.** The setting is exactly the same as Table 3.

		iWildCam		**fMoW**		**Camelyon17**	
		Scratch	*Pre-training*	*Scratch*	*Pre-training*	*Scratch*	*Pre-training*
Architecture	ResNet18	51.77	59.50	27.13	41.08	82.83	84.35
	ResNet50	53.35	66.75	28.55	49.75	75.85	86.78
	ResNet101	43.41	69.71	22.35	48.63	75.26	84.21
	ViT	51.68	69.82	23.20	51.74	76.78	91.27
	avg.	*50.05*	*66.45*	*25.31*	*47.80*	*77.68*	*86.65*
Augmentation	ImageNet	48.21	64.17	26.16	47.32	72.08	87.95
	AugMix	51.90	65.61	29.94	49.95	74.13	88.72
	RandomAug	47.49	65.14	31.12	49.48	77.21	90.24
	AutoAug	51.57	70.19	31.64	50.13	73.23	90.56
	avg.	*49.79*	*66.28*	*29.72*	*49.22*	*74.16*	*89.37*
De-biasing	UBNet	32.96	57.96	16.30	37.06	73.77	82.06
	PnD	45.25	61.03	21.01	40.29	83.81	79.69
	OccamNets	53.67	67.30	29.08	46.53	74.97	81.77
	avg.	*43.96*	*62.10*	*22.13*	*41.29*	*77.52*	*81.17*
Worst-case	GroupDRO	45.15	63.15	25.99	49.07	73.37	80.99
	BPA	2.62	56.06	8.71	44.32	76.99	86.08
	avg.	*23.89*	*59.61*	*17.35*	*46.70*	*75.18*	*83.54*
SDG	ADA	48.47	69.99	29.71	50.40	80.45	88.74
	ME-ADA	52.62	65.87	28.46	51.43	80.29	81.93
	SagNet	54.67	69.83	29.53	49.32	78.62	81.10
	L2D	47.78	64.09	23.90	45.65	86.32	94.85
	avg.	*50.89*	*67.45*	*27.90*	*49.20*	*81.42*	*86.66*
OOD	IRM	34.55	57.36	11.40	30.88	72.47	85.45
	CausIRL-M	52.28	68.34	22.32	49.44	72.10	79.83
	CausIRL-C	53.36	65.36	28.61	49.83	81.16	88.98
	avg.	*46.73*	*63.69*	*20.78*	*43.38*	*75.24*	*84.75*
Zero-shot	CLIP-base	13.97		13.30		50.01	
	CLIP-large	13.70		23.46		50.01	
	InstructBLIP	1.86		18.17		68.03	
	LLaVA-1.5	4.64		16.32		51.09	
	Phi-3.5-vision	9.91		10.88		66.71	
	avg.	*8.82*		*16.43*		*57.17*	

B.4 Further Analysis on Zero-shot Inference

To utilize foundation models for image classification, we employed various prompts to extract labels from the input images. Table 10 and Table 11 detail the prompts used in our prompt engineering approach.

Table 10: **Prompts used in Zero-shot inference.**

Foundation Model	**Prompt Type**	**Prompt**
CLIP	General	"a photo of a *label*"
LLaVA-1.5 InstructBLIP Phi-3.5-vision	General_1	Classify the image into $label_1, ...,$ or $label_C$. Please provide only the name of the label.
	General_2	Choose a label that best describes the image. Here is the list of labels to choose from: $label_1, ..., label_C$. Please provide only the name of the label.
	dSprites **Shapes3D** **SmallNorb**	Classify the object in the image into $label_1, ...,$ or $label_C$. Please provide only the name of the label.
	CelebA	Classify the person in the image into $male$ or $female$. Please provide only the name of the label.
	DeepFashion	Is a person wearing a dress or not? Please answer in yes or no.
	iWildcam	Classify the object or animal in the image. Here is the list of labels to choose from: $label_1, ..., label_C$. Please provide only the name of the label.
	fMoW	Classify the building or land-use in the image into $label_1, ..., label_C$. Please provide only the name of the label.
	Camelyon17 General_1	Please answer yes if the image contains any tumor tissue, and no otherwise. Please provide only the name of the label.
	Camelyon17 General_2	Please answer yes if the image contains any tumor tissue, and no otherwise. Please respond with a single word.
	Camelyon17 Tailored	Please analyze the image and determine if it contains any tumor tissue. Respond with 'Yes' if tumor tissue is present, or 'No' if it is not.

Table 11: **Prompt formats for vision-language models.**

	LLaVA-1.5, *InstructBLIP*	*Phi-3.5-vision*
Prompt Format	USER: < image >\n prompt \n ASSISTANT:	USER: < |image_1| >\n prompt