A Whac-A-Mole Dilemma: Shortcuts Come in Multiples Where Mitigating One Amplifies Others

[†]Zhiheng Li[2] [*]Ivan Evtimov[1] Albert Gordo[1] Caner Hazirbas[1] Tal Hassner[1]
Cristian Canton Ferrer[1] Chenliang Xu[2] [*]Mark Ibrahim[1]
[1]Meta AI [2]University of Rochester
{ivanevtimov,agordo,hazirbas,thassner,ccanton,marksibrahim}@meta.com
{zhiheng.li,chenliang.xu}@rochester.edu

Abstract

Machine learning models have been found to learn shortcuts—unintended decision rules that are unable to generalize—undermining models' reliability. Previous works address this problem under the tenuous assumption that only a single shortcut exists in the training data. Real-world images are rife with multiple visual cues from background to texture. Key to advancing the reliability of vision systems is understanding whether existing methods can overcome multiple shortcuts or struggle in a Whac-A-Mole game, i.e., where mitigating one shortcut amplifies reliance on others. To address this shortcoming, we propose two benchmarks: 1) UrbanCars, a dataset with precisely controlled spurious cues, and 2) ImageNet-W, an evaluation set based on ImageNet for watermark, a shortcut we discovered affects nearly every modern vision model. Along with texture and background, ImageNet-W allows us to study multiple shortcuts emerging from training on natural images. We find computer vision models, including large foundation models—regardless of training set, architecture, and supervision—struggle when multiple shortcuts are present. Even methods explicitly designed to combat shortcuts struggle in a Whac-A-Mole dilemma. To tackle this challenge, we propose Last Layer Ensemble, a simple-yet-effective method to mitigate multiple shortcuts without Whac-A-Mole behavior. Our results surface multi-shortcut mitigation as an overlooked challenge critical to advancing the reliability of vision systems. The datasets and code are released: https://github.com/facebookresearch/Whac-A-Mole.

1. Introduction

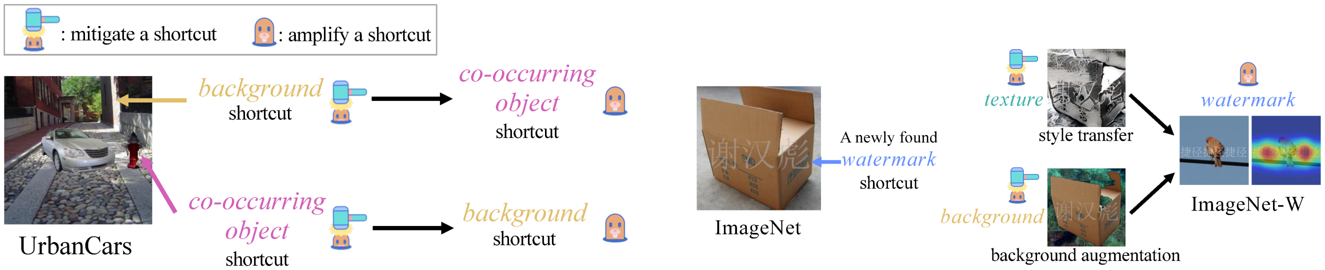
Machine learning often achieves good average performance by exploiting unintended cues in the data [26]. For instance, when backgrounds are spuriously correlated with objects, image classifiers learn background as a rule for object recognition [93]. This phenomenon—called "shortcut learning"—at best suggests average metrics overstate model performance and at worst renders predictions unreliable as models are prone to costly mistakes on out-of-distribution (OOD) data where the shortcut is absent. For example, COVID diagnosis models degraded significantly when spurious visual cues (*e.g.*, hospital tags) were removed [17].

Most existing works design and evaluate methods under the tenuous assumption that a *single shortcut* is present in the data [33,61,74]. For instance, Waterbirds [74], the most widely-used dataset, only benchmarks the mitigation of the background shortcut [7,15,59]. While this is a useful simplified setting, real-world images contain multiple visual cues; models learn multiple shortcuts. From ImageNet [18,82] to facial attribute classification [51] and COVID-19 chest radiographs [17], multiple shortcuts are pervasive. Whether existing methods can overcome multiple shortcuts or struggle in a *Whac-A-Mole* game—where mitigating one shortcut amplifies others—remains a critical open question.

We directly address this limitation by proposing two datasets to study *multi-shortcut* learning: **UrbanCars** and **ImageNet-W**. In UrbanCars (Fig. 1a), we precisely inject two spurious cues—background and co-occurring object. UrbanCars allows us to conduct controlled experiments probing multi-shortcut learning in standard training as well as shortcut mitigation methods, including those requiring shortcut labels. In ImageNet-W (IN-W) (Fig. 1b), we surface a new *watermark* shortcut in the popular ImageNet dataset (IN-1k). By adding a transparent watermark to IN-1k validation set images, ImageNet-W, as a new test set, reveals vision models ranging from ResNet-50 [31] to large foundation models [10] *universally rely on watermark as a spurious cue* for the "carton" class (*cf*. cardboard box in Fig. 1b). When a watermark is added, ImageNet top-1 accuracy drops by 10.7% on average across models. Some, such as ResNet-50, suffer a catastrophic 26.7% drop (from 76.1% on IN-1k to 49.4% on IN-W) (Sec. 2.2)). Along with texture [27,34] and background [93] benchmarks, ImageNet-W allows us to study *multiple shortcuts* emerging in natural images.

We find that across a range of supervised/self-supervised methods, network architectures, foundation models, and shortcut mitigation methods, vision models struggle when multiple shortcuts are present. Benchmarks on UrbanCars and multiple shortcuts in ImageNet (including ImageNet-W) reveal an overlooked challenge in the shortcut learning problem: *multi-shortcut mitigation resembles a Whac-A-Mole game, i.e., mitigating one shortcut amplifies reliance on others*. Even methods specifically designed to combat shortcuts

[†]Work done during the internship at Meta AI. [*]Equal Contribution.



(a) We construct UrbanCars, a new dataset with multiple shortcuts, facilitating the study of multi-shortcut learning under the *controlled setting*.

(b) We discover the new watermark shortcut emerged from a *natural image* dataset—ImageNet, and create ImageNet-W test set for ImageNet.

Figure 1. Our benchmark results on both datasets reveal the overlooked Whac-A-Mole dilemma in shortcut mitigation, *i.e.*, mitigating one shortcut amplifies the reliance on other shortcuts .

decrease reliance on one shortcut at the expense of amplifying others (Sec. 5). To tackle this open challenge, we propose Last Layer Ensemble (LLE) as the first endeavor to mitigate multiple shortcuts jointly without Whac-A-Mole behavior. LLE uses data augmentation based on only the knowledge of the shortcut type without using shortcut labels—making it scalable to large-scale datasets.

To summarize, our contributions are (1) We create UrbanCars, a dataset with precisely injected spurious cues, to better benchmark multi-shortcut mitigation. (2) We curate ImageNet-W—a new out-of-distribution (OOD) variant of ImageNet benchmarking a pervasive watermark shortcut we discovered— to form a more comprehensive multi-shortcut evaluation suite for ImageNet. (3) Through extensive benchmarks on UrbanCars and ImageNet shortcuts (including ImageNet-W), we uncover that mitigating multiple shortcuts is an overlooked and universal challenge, resembling a Whac-A-Mole game, *i.e.*, mitigating one shortcut amplifies reliance on others. (4) Finally, we propose Last Layer Ensemble as the first endeavor for multi-shortcut mitigation without the Whac-A-Mole behavior. We hope our contributions advance research into the overlooked challenge of mitigating multiple shortcuts.

2. New Datasets for Multi-Shortcut Mitigation

While most previous datasets [4,60,61,74] are based on the oversimplified single-shortcut setting, we introduce the UrbanCars dataset (Sec. 2.1) and the ImageNet-Watermark dataset (Sec. 2.2) to benchmark multi-shortcut mitigation.

2.1. UrbanCars Dataset

Overview We construct the UrbanCars dataset with multiple shortcuts: *background* (BG) and *co-occurring object* (CoObj). As shown in Fig. 2, each image in UrbanCars has a car at the center on a natural scene background with a co-occurring object on the right. The task is to classify the car's body type (*i.e.*, target) by overcoming two shortcuts in the training set, which correlate with the target label.

Formally, we denote the dataset as a set of N tuples, $\{(x_i, y_i, b_i, c_i)\}_{i=1}^N$, where each image x_i is annotated with three labels: target label y_i for the car body type, *background* label b_i, and *co-occurring object* label c_i. We use a shared label space for all three labels with two classes: `urban` and `country`, *i.e.*, $y_i, b_i, c_i \in \{\texttt{urban}, \texttt{country}\}$. Based on the combination of three labels, the dataset is partitioned into $2^3 = 8$ groups, *i.e.*, $\{\texttt{urban}, \texttt{country}\}$ car on the $\{\texttt{urban}, \texttt{country}\}$ BG with the $\{\texttt{urban}, \texttt{country}\}$ CoObj. We introduce the data distribution and construction below and include details in Appendix A.1.

Data Distribution The training set of UrbanCars has two spurious correlations of BG and CoObj shortcuts, whose strengths are quantified by $P(\mathbf{b} = \mathbf{y} \mid \mathbf{y})$ and $P(\mathbf{c} = \mathbf{y} \mid \mathbf{y})$, respectively. That is, the ratio of common BG (or CoObj) given a target class. We set both to 0.95 by following the correlation strength in [74]. We assume that two shortcuts are independently correlated with the target, *i.e.*, $P(\mathbf{b}, \mathbf{c} \mid \mathbf{y}) = P(\mathbf{b} \mid \mathbf{y}) P(\mathbf{c} \mid \mathbf{y})$. As shown in Fig. 2, most urban car images have the urban background (*e.g.*, alley) and urban co-occurring object (*e.g.*, fire plug), and vice versa for country car images. The frequency of each group in the training set is in Fig. 2. The validation and testing sets are balanced without spurious correlations, *i.e.*, ratios are 0.5.

Data Construction The UrbanCars dataset is created from several source datasets. The car objects and labels are from Stanford Cars [50], where the urban cars are formed by classes such as sedan and hatchback. The country cars are from classes such as truck and van. The backgrounds are from Places [99]. We use classes such as alley and crosswalk

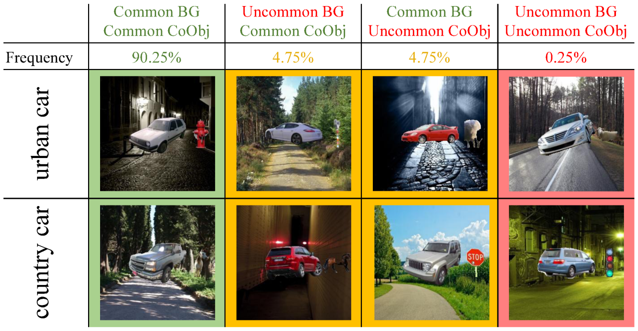


Figure 2. Unbalanced groups in UrbanCars's training set based on two shortcuts: *background* (BG) and *co-occurring object* (CoObj).

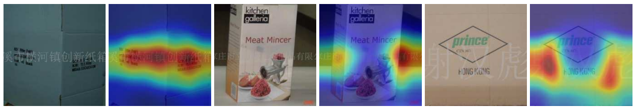


Figure 3. Many carton class images in the ImageNet training set contain the watermark. Saliency maps [78] of ResNet-50 [31] show that the watermark serves as the shortcut for the carton class.

to form the urban background. The country background images are from classes such as forest road and field road. Regarding co-occurring objects, we use LVIS [29] to obtain the urban ones (*e.g.*, fireplug and stop sign), and country ones (*e.g.*, cow and horse). After obtaining the source images, we paste the car and co-occurring object onto the background.

UrbanCars Metrics We first report the *In Distribution Accuracy* (**I.D. Acc**) on UrbanCars. It computes the weighted average over accuracy per group, where weights are proportional to the training set's correlation strength (*i.e.*, frequency in Fig. 2) by following "average accuracy" [74] to measure the performance when no group shift happens.

To measure robustness against the group shift, previous single-shortcut benchmarks [15,59,74] use worst-group accuracy [74], *i.e.*, the lowest accuracy among all groups. However, this metric does not capture multi-shortcut mitigation well since it only focuses on groups where both shortcut categories are uncommon (*cf*. the last column in Fig. 2).

To address this shortcoming, we introduce three new metrics: **BG Gap**, **CoObj Gap**, and **BG+CoObj Gap**. BG Gap is the accuracy drop from I.D. Acc to accuracy in groups where BG is uncommon but CoObj is common (*cf*. 1st yellow column in Fig. 2). Similarly, CoObj Gap computes the accuracy drop from I.D. Acc to groups where only CoObj is uncommon (*cf*. 2nd yellow column in Fig. 2). BG+CoObj Gap computes accuracy drop from I.D. Acc to groups where both BG and CoObj are uncommon (*cf*. red column in Fig. 2). The first two metrics measure the robustness against the group shift for each shortcut, and the last metric evaluates the model's robustness when both shortcuts are absent.

2.2. ImageNet-Watermark (ImageNet-W)

In addition to the precisely controlled spurious correlations in UrbanCars, we study naturally occurring shortcuts in the most popular computer vision benchmark: ImageNet [18]. While ImageNet lacks shortcut labels, we can evaluate models' reliance on texture [27] and background [93] shortcuts. We additionally discovered a pervasive watermark shortcut and contribute ImageNet-Watermark (ImageNet-W or IN-W), an evaluation set to expose models' watermark shortcut reliance. Along with texture and background, this forms a comprehensive suite to evaluate reliance on the multiple naturally occurring shortcuts in ImageNet.

Watermark Shortcut in ImageNet In the training set of the *carton* class, many images contain a watermark at the center written in Chinese characters and ImageNet-trained ResNet-50 [31] focuses on the watermark region to predict the carton class (Fig. 3). Since the watermark reads carton factory names or contact person's names of a carton factory, we conjecture that this watermark shortcut originates from the real-world spurious correlation of web images. In the validation set, none of the carton class images contain the watermark, so ResNet-50 underperforms on the carton class (48%) relative to overall accuracy (76%) across 1k classes.

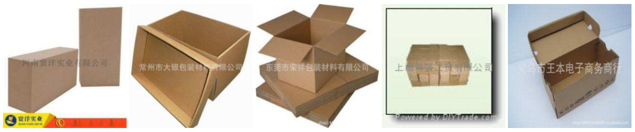


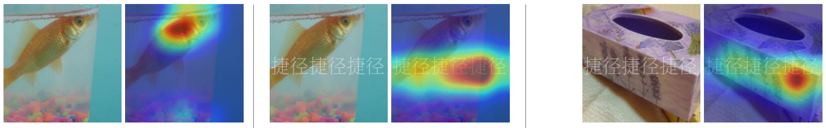
Figure 4. Carton images from LAION [75,76], a large-scale dataset with 400 million to 2 billion images used in CLIP [67] pretraining, also contain watermarks, enabling CLIP's reliance on the watermark shortcut in zero-shot transfer to ImageNet and ImageNet-W.

Data Construction To test the robustness against the watermark shortcut, we create ImageNet-Watermark (ImageNet-W or IN-W) dataset, a new out-of-distribution evaluation set of ImageNet. As shown in Tab. 1, we overlay a transparent watermark written in "捷径捷径捷径" at the center of all images from ImageNet validation set to mimic the watermark pattern in IN-1k, where "捷径" means "shortcut" in Chinese. We do this because we find that models use the watermark even when the content is not identical to the watermark in the training set of carton images, suggesting that it is watermark's presence rather than its content that serves as the shortcut. We evaluate watermark in other contents and languages in Appendix A.2.

ImageNet-W Metrics We mainly use two metrics to measure watermark shortcut reliance: (1) **IN-W Gap** is the accuracy on IN-W minus the accuracy on IN-1k validation set. A smaller accuracy drop indicates less reliance on the watermark shortcut across all 1000 classes. (2) **Carton Gap** is the carton class accuracy increase from IN-1k to IN-W. A smaller Carton Gap indicates less reliance on the watermark shortcut for predicting the carton class.

To demonstrate that the watermark shortcut is used for predicting carton, we use the following in Tab. 1: (1) $P(\hat{y} = \text{carton})$, the predicted probability of carton on all IN-1k validation set images, (2) $\Delta P(\hat{y} = \text{carton})$, the predicted probability increase from IN-1k to IN-W of all 1k classes, and (3) $\Delta P(\hat{y} = \text{carton} \mid y = \text{carton})$, the predicted probability increase from IN-1k to IN-W of the carton class.

Ubiquitous reliance on the watermark shortcut To study reliance on the watermark shortcut, we use ImageNet-W to benchmark a broad range of State-of-The-Art (SoTA) vision models, including standard supervised training, using different architectures [22,31,68], augmentations and regularizations [27,36,94,95]. We also benchmark foundation models [10] pretrained on larger datasets [28,67,75,76,81] with different pretraining supervision and transfer learning techniques [13,28,30,67,81,91]. In Tab. 1, we find a considerable IN-W Gap of up to -26.7 and -10.7 on average and



method	architecture	(pre)training data	IN-1k Acc ↑	$P(\hat{y} = \text{carton})$ (%)	IN-W Gap ↑	$\Delta P(\hat{y} = \text{carton})$ (%) ↓	Carton Gap ↓	$\Delta P(\hat{y} = \text{carton} \mid y = \text{carton})$ (%) ↓
Supervised	ResNet-50 [31]	IN-1k [18]	76.1	0.07	-26.7	+7.56	+40	+42.46
MoCov3 [13] (LP)	ResNet-50	IN-1k	74.6	0.08	-20.7	+2.94	+44	+44.37
Style Transfer [27]	ResNet-50	SIN [27]	60.1	0.10	-17.3	+4.91	+52	+50.06
Mixup [95]	ResNet-50	IN-1k	76.1	0.07	-18.6	+3.43	+38	+39.78
CutMix [94]	ResNet-50	IN-1k	78.5	0.09	-14.8	+1.92	+22	+29.61
Cutout [20,98]	ResNet-50	IN-1k	77.0	0.08	-18.0	+2.93	+32	+38.06
AugMix [36]	ResNet-50	IN-1k	77.5	0.09	-16.8	+2.61	+36	+34.44
Supervised	RG-32gf	IN-1k	80.8	0.09	-14.1	+3.74	+32	+33.43
SEER [28] (FT)	RG-32gf [68]	IG-1B [28]	83.3	0.09	-6.5	+0.56	+18	+24.26
Supervised	ViT-B/32 [22]	IN-1k	75.9	0.09	-8.7	+1.20	+34	+34.31
Uniform Soup [91] (FT)	ViT-B/32	WIT [67]	79.9	0.09	-7.9	+0.32	+24	+23.87
Greedy Soup [91] (FT)	ViT-B/32	WIT	81.0	0.09	-6.5	+0.35	+16	+23.87
Supervised	ViT-L/16	IN-1k	79.6	0.08	-6.2	+0.82	+34	+32.57
CLIP [67] (zero-shot)	ViT-L/14	WIT	76.5	0.06	-4.4	**+0.01**	+12	**+1.75**
CLIP (zero-shot)	ViT-L/14	LAION-400M [76]	72.7	0.05	-4.9	+0.03	+12	+13.76
MAE [30] (FT)	ViT-H/14	IN-1k	86.9	0.08	-3.5	+0.43	+30	+29.59
SWAG [81] (LP)	ViT-H/14	IG-3.6B [81]	85.7	0.09	-4.9	+0.19	**+8**	+12.80
SWAG (FT)	ViT-H/14	IG-3.6B	88.5	0.09	**-3.1**	+0.35	+18	+20.25
CLIP (zero-shot)	ViT-H/14	LAION-2B [75]	77.9	0.06	-3.6	+0.03	+16	+12.01
	average		78.6	0.08	-10.7	+1.74	+26.7	+27.96

Table 1. **Models rely on the watermark as a shortcut for the carton class.** LP and FT denote linear probing and fine-tuning on ImageNet-1k, respectively. Because models exhibit drops (*i.e.*, IN-W Gap) and an increase in accuracy and predicted probability of the carton class from IN-1k to IN-W, we conclude that various vision models suffer from the watermark shortcut (more results in Appendices E.1 and E.2).

a Carton Gap of up to +52 and +26.7 on average. While all models exhibit uniform ($1/1000 = 0.1\%$) predicted probabilities for carton class ($P(\hat{y} = \text{carton})$) on IN-1k, we observe a considerable increase in the predicted probability of carton on IN-W ($\Delta P(\hat{y} = \text{carton})$) and a significant predicted probability increase in carton class images ($\Delta P(\hat{y} = \text{carton} \mid y = \text{carton})$). Although compared to supervised ResNet-50, some models with larger architectures or extra training data can decrease reliance on the watermark shortcut, none of them fully close the performance gaps. Interestingly, CLIP with zero-shot transfer still suffers from the watermark shortcut with +12 to +16 Carton Gap, which could be explained by many carton images in the pretraining data (*e.g.*, LAION) also containing watermarks (*cf*. Fig. 4). To the best of our knowledge, this is the first real-world example of **the existence of shortcut in billion-scale datasets for foundation model pretraining**, which also confirms findings that data quality, not quantity [25,62], matters most to CLIP's robustness.

Multi-Shortcut Mitigation Metrics on ImageNet To measure the mitigation of multiple shortcuts, we evaluate models on multiple OOD variants of ImageNet. In this work, we study three shortcuts on ImageNet—background, texture, and watermark. The background shortcut is evaluated on ImageNet-9 (IN-9) [93], and we use **IN-9 Gap** (*i.e.*, BG-Gap in [93]) as the evaluation metric, which is the accuracy drop from Mixed-Same to Mixed-Rand in IN-9, where a lower accuracy drop implies less background shortcut reliance. The texture shortcut is evaluated on Stylized ImageNet (SIN) [27] and ImageNet-R (IN-R) [34], where we use **SIN Gap**, top-1 accuracy drop from IN-1k to SIN, and **IN-R Gap**, the top-1 accuracy drop from IN-200 (*i.e.*, a subset of IN-1k with 200 classes used in IN-R) to IN-R.

3. Benchmark Methods and Settings

On all datasets, we first evaluate standard training that minimizes the empirical risk on the training set (*i.e.*, **ERM** [85]) using ResNet-50 [31] as the network architecture, which serves as the baseline. On ImageNet, we additionally show ERM's results with other architectures, pretraining datasets, and supervision.

In addition to ERM, we comprehensively evaluate shortcut mitigation methods across four categories based on the level of shortcut information required (Tab. 2).

Category 1: Standard Augmentation and Regularization Methods in this category use general data augmentation or regularization without prior knowledge of the shortcut, which are commonly used to improve accuracy on IN-1k, *e.g.*, new training recipes [86,90]. Some works [11,65] show

Category	Summary	Shortcut Information	Methods
1	Standard Augmentation and Regularization	None	Mixup [95], Cutout [20,98], CutMix [94], AugMix [36], SD [64]
2	Targeted Augmentation for Mitigating Shortcuts	Types of shortcuts (w/o shortcut labels)	CF+F Aug [11], Style Transfer (TXT Aug) [27], BG Aug [73,93], WMK Aug
3	Using Shortcut Labels	Image-level ground-truth shortcut label	gDRO [74], DI [89], SUBG [39], DFR [46]
4	Inferring Pseudo Shortcut Labels	Image-level pseudo shortcut label	LfF [61], JTT [59], EIIL [15], DebiAN [54]

Table 2. Existing methods for multi-shortcut mitigation benchmark.

that they can also improve OOD robustness.

Category 2: Targeted Augmentation for Mitigating Shortcuts Other works use data augmentation that modifies shortcut cues. We evaluate CF+F Aug [11] on UrbanCars. On ImageNet, we benchmark texture augmentation (TXT Aug) via style transfer [27] and background augmentation (BG Aug) [73,93]. To counter the watermark shortcut, we design watermark augmentation (WTM Aug) that randomly overlays the watermark onto images (*cf*. Appendix B.1).

Category 3: Using Shortcut Labels In this category, methods use shortcut labels for mitigation, which are generally used to reweight [74] or resample training data [39, 46,74]. We only benchmark methods in this category on UrbanCars since ImageNet does not have shortcut labels.

Category 4: Inferring Pseudo Shortcut Labels Following the ideas of methods using shortcut labels, one line of works [15,54,59,61] estimates the pseudo shortcut labels when ground-truth labels are unavailable.

Benchmark Settings We introduce the experiment settings here (details in Appendix B.3). On UrbanCars, we use worst-group accuracy [74] on the validation set to select the early stopping epoch and report test set results. All methods except DFR [46] use end-to-end training on UrbanCars. On ImageNet, following the last layer re-training [46] setting, we only train the last classification layer upon a frozen feature extractor. On both datasets, we use ResNet-50 as the network architecture. On ImageNet, we also benchmark self-supervised and foundation models.

4. Our Approach

Motivation Our multi-shortcut benchmark results (Sec. 5) show that many existing methods suffer from the Whac-A-Mole problem, motivating us to design a method to mitigate multiple shortcuts simultaneously.

We focus on mitigating multiple *known* shortcuts—the number and types of shortcuts are given, but shortcut labels are not. The absence of shortcut labels makes it scalable to large datasets (*e.g*., ImageNet). Although mitigating unknown numbers and types of shortcuts seems more desirable, not only do our empirical results show their underperformance, but also it is theoretically impossible to mitigate shortcuts without any inductive biases [58].

We follow methods that use data augmentation to modify the shortcut cues (*i.e*., category 2). Formally, given a set of K shortcuts $\{\mathrm{s}_i\}_{i=1}^{K}$ for mitigation, we create a set of augmentations $\mathrm{S}_{\text{aug}} = \{\mathcal{A}_i\}_{i=1}^{K} \cup \{\mathcal{I}\}$, where the augmentation $\mathcal{A}_i$ (*e.g*., style transfer [27]) modifies the visual cue of the shortcut s_i (*e.g*., texture). $\mathcal{I}$ denotes the identity transformation, *i.e*., no augmentation applied.

Based on the augmentation set S_{aug}, a straightforward way is to minimize the empirical risk [85] over all augmented and original images. However, different augmentations can be incompatible, leading to suboptimal results. That is, augmentation $\mathcal{A}_i$ could be detrimental to mitigating a different shortcut s_j, where $i \neq j$. For example, mitigating the texture shortcut via style transfer [27] augmentation unexpectedly amplifies the saliency of the watermark (Fig. 1b), leading to worse watermark mitigation results (Tab. 1).

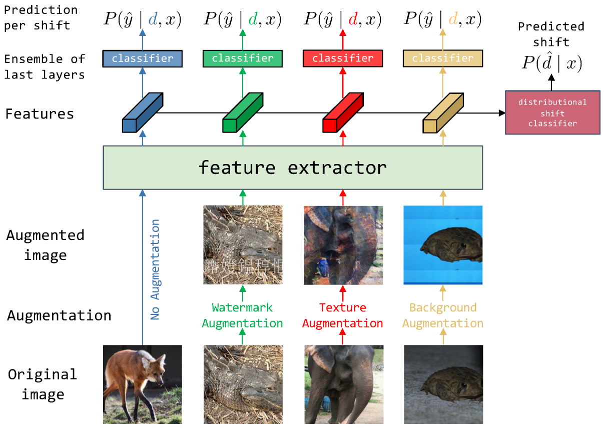


Figure 5. An overview of Last Layer Ensemble (LLE). LLE trains an ensemble of the last classification layers upon a feature extractor, where each last layer is trained with images in one augmentation type. The distributional shift classifier, supervised by the augmentation type, is trained to predict the distributional shift and dynamically aggregates the predictions per shift during testing.

Last Layer Ensemble To address this issue, we propose Last Layer Ensemble (LLE), a new method for mitigating multiple shortcuts simultaneously (Fig. 5). Since it is hard to use a single model to learn the invariance among incompatible augmentations, we instead train an ensemble [21] of classification layers (*i.e*., last layers) on top of a shared feature extractor so that each classification layer only trains on data from a single type of augmentation that simulates one type of distributional shift d. In this way, each last layer predicts the probability of the target $P(\hat{y} \mid d, x)$.

At the same time, we train a *distributional shift classifier*, another classification layer on top of the feature extractor, to predict the type of augmentation that simulates the distributional shift, *i.e*., $P(\hat{d} \mid x)$. During testing, LLE dynamically aggregates the logits from the ensemble of the last layers based on the predicted distributional shift. *E.g*., when the testing image contains the texture shift, the *distributional shift classifier* gives higher weights for the logits from the classifier trained with texture augmentation, alleviating the impact from other classification layers trained with incompatible augmentations. In addition, when the weights of the feature extractor are not frozen, we stop the gradient from the *distributional shift classifier* to the feature extractor, preventing the feature extractor from learning the shortcut information. Compared to standard ensemble approaches [21] that train multiple full networks and add significant inference overhead, our method uses minimal additional training parameters and has better computational efficiency.

	I.D. Acc	shortcut reliance BG Gap ↑	 CoObj Gap ↑	 BG+CoObj Gap ↑
ERM	97.6	-15.3	-11.2	-69.2
Mixup	98.3	-12.6	-9.3	-61.8
CutMix	96.6	-45.0 (×2.94 🐹)	-4.8	-86.5
Cutout	97.8	-15.8 (×1.03 🐹)	-10.4	-71.4
AugMix	98.2	-10.3	-12.1 (×1.08 🐹)	-70.2
SD	97.3	-15.0	-3.6	-36.1
CF+F Aug	96.8	-16.0 (×1.04 🐹)	**+0.4**	-19.4
LfF	97.2	-11.6	-18.4 (×1.64 🐹)	-63.2
JTT (E=1)	95.9	-8.1	-13.3 (×1.18 🐹)	-40.1
EIIL (E=1)	95.5	-4.2	-24.7 (×2.21 🐹)	-44.9
JTT (E=2)	94.6	-23.3 (×1.52 🐹)	-5.3	-52.1
EIIL (E=2)	95.5	-21.5 (×1.40 🐹)	-6.8	-49.6
DebiAN	98.0	-14.9	-10.5	-69.0
LLE (ours)	96.7	**-2.1**	-2.7	**-5.9**

Table 3. **Many methods not using shortcut labels (category 1,2,4) amplify shortcut on UrbanCars.** 🐹: increased reliance on a shortcut relative to ERM. ×2.94: 2.94 times larger than ERM.

5. Experiments

Based on UrbanCars and ImageNet-W datasets, we show results on multi-shortcut mitigation. We first study if standard supervised training (*i.e.*, ERM) relies on multiple shortcuts (Sec. 5.1). Next, we show the multi-shortcut setting is significantly challenging: mitigating one shortcut increases reliance on other shortcuts compared to ERM. We name this phenomenon *Whac-A-Mole*, which is observed in many SoTA methods, including mitigation methods (Sec. 5.2) and self-supervised/foundation models (Sec. 5.3). Finally, we show that our Last Layer Ensemble method can reduce reliance across multiple shortcuts more effectively (Sec. 5.4).

5.1. Standard training relies on multiple shortcuts

On both datasets, we find that standard training (*i.e.*, ERM [85]) relies on multiple shortcuts. On UrbanCars, Tab. 3 shows that ERM achieves near zero in-distributional error (97.6% I.D. Acc.). However, ERM's performance drops when group shift happens. When the background shortcut is absent, ERM's performance drops by 15.3% in BG Gap. Similarly, the accuracy drops by 11.2% in CoObj Gap when the CoObj shortcut is absent. When neither shortcut is present, models suffer catastrophic drops of 69.2% in BG+CoObj Gap. On ImageNet, Tab. 4 shows that ERM achieves good top-1 accuracy of 76.39% on IN-1k. However, it suffers considerable drops in accuracy when watermark, texture, or background cues are altered, *e.g.*, 30% Carton Gap for watermark, 56-69% for texture, and 5.19% for background, suggesting that standard training on natural images from ImageNet leads to reliance on multiple shortcuts.

5.2. Results: Mitigation Methods

Results: Standard Augmentation and Regularization (Category 1) We first show the results of methods using augmentation and regularization without using inductive biases of shortcuts. On UrbanCars (Tab. 3), we observed that CutMix and Cutout amplify the background shortcut with a larger BG Gap relative to ERM. AugMix increases the reliance on the CoObj shortcut with a larger CoObj Gap (*i.e.*, -12.2%) compared to ERM. Although Mixup and SD do not produce Whac-A-Mole results, they only yield marginal improvement or can only mitigate one shortcut well. On ImageNet, the results in Tab. 4 show that all approaches amplify at least one shortcut. For instance, AugMix achieves a worse Carton Gap to amplify the watermark shortcut compared to ERM. For CutMix, we again observe that it amplifies the BG shortcut on ImageNet. We show more results of CutMix and analyze its background shortcut reliance in Appendix G.

	IN-1k	shortcut reliance Watermark (WTM) IN-W Gap ↑	 Carton Gap ↓	 Texture (TXT) SIN Gap ↑	 IN-R Gap ↑	 Background (BG) IN-9 Gap ↑
ERM	76.39	-25.40	+30	-69.43	-56.22	-5.19
Mixup	76.17	-24.87	+34 (×1.13 🐹)	-68.18	-55.79	-5.60 (×1.08 🐹)
CutMix	75.90	-25.78 (×1.01 🐹)	+32 (×1.06 🐹)	-69.31	-56.36	-5.65 (×1.09 🐹)
Cutout	76.40	-25.11	+32 (×1.06 🐹)	-69.39	-55.93	-5.35 (×1.03 🐹)
AugMix	76.23	-23.41	+38 (×1.26 🐹)	-68.51	-54.91	-5.85 (×1.13 🐹)
SD	76.39	-26.03 (×1.02 🐹)	+30	-69.42	-56.36	-5.33 (×1.03 🐹)
WTM Aug	76.32	**-5.78**	+14	-69.31	-56.22	-5.34 (×1.03 🐹)
TXT 🔨 Aug	75.94	-25.93 (×1.02 🐹)	+36 (×1.20 🐹)	-63.99	**-53.24**	-5.66 (×1.09 🐹)
BG 🔨 Aug	76.03	-25.01	+36 (×1.20 🐹)	-68.41	-54.51	-4.67
LfF	76.35	-26.19 (×1.03 🐹)	+36 (×1.20 🐹)	-69.34	-56.02	-5.61 (×1.08 🐹)
JTT	76.33	-26.40 (×1.04 🐹)	+32 (×1.06 🐹)	-69.48	-56.30	-5.55 (×1.07 🐹)
EIIL	71.55	-33.48 (×1.31 🐹)	+24	-66.04	-61.35 (×1.09 🐹)	-6.42 (×1.24 🐹)
DebiAN	76.33	-26.40 (×1.04 🐹)	+36 (×1.20 🐹)	-69.37	-56.29	-5.53 (×1.07 🐹)
LLE (ours)	76.25	-6.18	**+10**	**-61.00**	-54.89	**-3.82**

Table 4. **Existing methods fail to combat multiple shortcuts by amplifying at least one shortcut relative to ERM on ImageNet.** All models use ResNet-50 with last layer re-training [46].

> **Takeaway**: Standard augmentation and regularization methods can mitigate some shortcuts (*e.g.*, texture) 🔨 but amplify others 🐹.

Results: Targeted Augmentation for Mitigating Shortcuts (Category 2) Further, we benchmark methods using data augmentation to mitigate a specific shortcut. Compared to methods in category 1, augmentations here use stronger inductive biases about the shortcut by modifying the shortcut visual cue. On UrbanCars, although CF+F Aug achieves good results for the CoObj shortcut, it amplifies the BG shortcut. On ImageNet, texture and background augmentation improve the reliance on the watermark shortcut, which can be explained by the retained or even increased saliency of the watermark in Fig. 1b and Appendix's Figs. 9 and 10.

> **Takeaway**: Augmentations tackling a specific type of shortcut 🔨 (*e.g.*, style transfer for texture shortcut) can amplify other shortcuts 🐹 (*e.g.*, watermark).

Results: Using Shortcut Labels (Category 3) Then, we show the results of methods using shortcut labels on UrbanCars in Tab. 5. Methods can mitigate multiple shortcuts when labels of both shortcuts are used (*cf*. first section in Tab. 5). However, when using labels of either shortcut, which

	shortcut label			shortcut reliance		
	Train	Val	I.D. Acc	BG Gap ↑	CoObj Gap ↑	BG+CoObj Gap ↑
ERM	✗	BG+CoObj	97.6	-15.3	-11.2	-69.2
gDRO	BG+CoObj	BG+CoObj	91.6	-10.9	-3.6	-16.4
DI	BG+CoObj	BG+CoObj	89.0	**-2.2**	-1.0	**+0.4**
SUBG	BG+CoObj	BG+CoObj	71.1	-4.7	**-0.3**	-6.3
DFR	BG+CoObj	BG+CoObj	89.7	-10.7	-6.9	-45.2
ERM	✗	BG	97.8	-14.6	-11.3	-68.5
gDRO	BG 🔨	BG	96.0	-4.2	-26.9 (×2.39 🐹)	-56.5
DI	BG 🔨	BG	94.7	+2.2	-27.0 (×2.40 🐹)	-25.2
SUBG	BG 🔨	BG	92.6	+1.3	-36.4 (×3.24 🐹)	-35.8
DFR	BG 🔨	BG	97.4	-9.8	-13.6 (×1.21 🐹)	-58.9
ERM	✗	CoObj	97.6	-15.4	-11.0	-68.8
gDRO	CoObj 🔨	CoObj	95.7	-31.4 (×2.03 🐹)	-0.5	-54.9
DI	CoObj 🔨	CoObj	94.2	-36.1 (×2.34 🐹)	+2.8	-35.8
SUBG	CoObj 🔨	CoObj	93.1	-60.2 (×3.90 🐹)	+2.5	-62.4
DFR	CoObj 🔨	CoObj	97.4	-19.1 (×1.24 🐹)	-8.6	-64.9

Table 5. **Methods using shortcut labels (category 3) amplify the unlabeled shortcut when mitigating the labeled shortcut on UrbanCars.** 🔨: mitigate a shortcut, *e.g.*, using shortcut labels.

is the typical situation for in-the-wild datasets where shortcut labels are incomplete, they exhibit a higher performance gap in the other shortcut relative to ERM. *E.g.*, when only using the CoObj labels, models achieve poorer BG Gap results.

> **Takeaway**: Methods using shortcut labels mitigate the labeled shortcut 🔨 but amplifies the unlabeled one 🐹.

Results: Inferring Pseudo Shortcut Labels (Category 4) The Whac-A-Mole problem of methods using shortcut labels motivates us to study whether the problem can be solved by inferring pseudo labels of multiple shortcuts. Here we analyze the results of LfF, JTT, EIIL, and DebiAN. Their key idea is based on ERM's training dynamics of learning different visual cues. LfF infers soft shortcut labels by assuming that the shortcut is learned earlier. Similarly, JTT and EIIL use an under-trained ERM trained with E epochs as the reference model to infer pseudo shortcut labels. We use E=1 and E=2 for JTT and EIIL. Instead of using a fixed reference model, DebiAN jointly trains the reference and mitigation models. The results in Tab. 3 show that LfF, JTT (E=1), and EIIL (E=1) still exhibit Whac-A-Mole results by achieving a larger CoObj Gap than ERM. On the other hand, JTT (E=2) and EIIL (E=2) also show the Whac-A-Mole results by achieving larger BG Gap than ERM. On ImageNet, we observe Whac-A-Mole results produced by LfF, JTT, EIIL, and DebiAN in Tab. 4.

To investigate the reason for their Whac-A-Mole results, we analyze the *training dynamics of ERM*. In Fig. 6, we plot the accuracy of three visual cues—object (*i.e.*, car body type), background, and co-occurring object on the validation set. The accuracy is computed based on ERM's {`urban`, `country`} predictions against labels of object, BG, and CoObj. We observe a Whac-A-Mole game in ERM's training. At epoch 1, ERM mainly predicts the background (82.6%), suggesting that the background shortcut is learned first. Thus, LfF, JTT (E=1), and EIIL (E=1) can infer the BG shortcut labels well to amplify the CoObj shortcut. As the training continues to epoch 2, the reliance on the BG shortcut decreases (82.6% to 71.2%), but the reliance on the CoObj shortcut is increased (60.6% to 71.8%). It renders JTT (E=2) and EIIL (E=2) better infer CoObj shortcut labels, which, in turn, amplifies the BG shortcut.

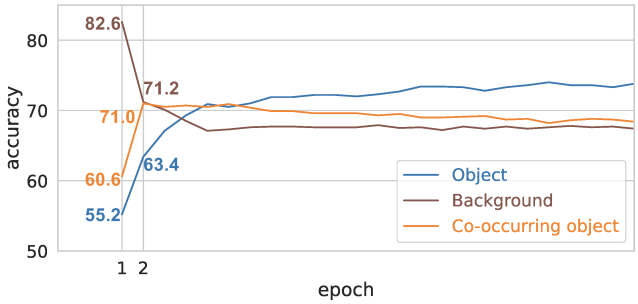


Figure 6. On UrbanCars, **ERM learns BG and CoObj shortcuts at different training epochs, making it difficult to infer pseudo labels (category 4) of multiple shortcuts from ERM.**

> **Takeaway**: Methods inferring pseudo shortcut labels still amplify shortcuts 🐹 because ERM learns different shortcuts *asynchronously* during training, making it hard to infer labels of all shortcuts 🔨 for mitigation.

5.3. Results: Self-Supervised & Foundation Models

On ImageNet, we further benchmark self-supervised pre-training methods, *i.e.*, MoCov3 [13], MAE [30], SEER [28]. We also benchmark foundation models that use extra training data, *i.e.*, Uniform Soup [91], Greedy Soup [91], CLIP [67], SEER [28], and SWAG [81]. The results in Tab. 6 show that many of them fail to mitigate multiple shortcuts jointly. Regarding self-supervised methods, MoCov3 achieves worse results on all three shortcuts, and MAE achieves a worse SIN Gap for the texture shortcut relative to ERM. Regarding foundation models, although SWAG with linear probing (LP) achieves a much better IN-R Gap (-19.79%), it also has a stronger reliance on the background in BG Gap compared to ERM. Similarly, SEER, Uniform Soup, and Greedy Soup mitigate the watermark shortcut but amplify the background shortcut. When using ViT-L, although CLIP with zero-shot transfer does not produce Whac-A-Mole results, they do not fully close the performance gap. Besides, they also show much lower IN-1k accuracy than other foundation models. We show results using other architectures in Appendix F.2.

> **Takeaway**: Self-supervised and foundation models can mitigate some shortcuts 🔨 but amplify others 🐹.

5.4. Results: Last Layer Ensemble (LLE)

We show that our Last Layer Ensemble (LLE) can better tackle multi-shortcut mitigation. LLE mitigates shortcuts via a set of data augmentations. Specifically, we augment background (BG) and co-occurring object (CoObj) by swapping

		shortcut reliance				
		Watermark		Texture		Background
	IN-1k	IN-W ↑ Gap	Carton ↓ Gap	SIN ↑ Gap	IN-R ↑ Gap	IN-9 ↑ Gap
arch: RG-32gf						
ERM	80.88	-14.15	+32	**-69.27**	-52.43	**-6.40**
SEER (FT,IG-1B)	83.35	**-6.50**	**+18**	-73.04 (×1.05)	**-50.42**	-7.14 (×1.11)
arch: ViT-B/32						
ERM	75.92	-8.71	+34	**-57.16**	-49.45	**-6.86**
Uniform Soup (FT,WIT)	79.96	-7.90	+24	-59.67 (×1.04)	**-27.51**	-7.78 (×1.13)
Greedy Soup (FT,WIT)	81.01	**-6.47**	**+16**	-59.61 (×1.04)	-30.01	-7.21 (×1.05)
arch: ViT-B/16						
ERM	81.07	-6.69	+26	-62.60	-50.36	-5.36
SWAG (LP,IG-3.6B)	81.89	-7.76 (×1.16)	+18	-67.33 (×1.08)	**-19.79**	-10.39 (×1.94)
SWAG (FT,IG-3.6B)	85.29	-5.43	+24	-66.99 (×1.07)	-29.55	-4.44
MoCov3 (LP)	76.65	-16.0 (×2.39)	+22	-63.36 (×1.01)	-56.86 (×1.12)	-7.80 (×1.45)
MAE (FT)	83.72	-4.60	+24	-65.20 (×1.04)	-47.10	-4.45
MAE+**LLE (ours)**	83.68	**-2.48**	**+6**	**-58.78**	-44.96	**-3.70**
arch: ViT-L/16 or 14						
ERM	79.65	-6.14	+34	-61.43	-53.17	-6.50
SWAG (LP,IG-3.6B)	85.13	-5.73	**+6**	-60.26	-10.17	-7.26 (×1.12)
SWAG (FT,IG-3.6B)	88.07	-3.16	+20	-63.45 (×1.03)	-12.29	-2.92
CLIP (zero-shot,WIT)	76.57	-4.47	+12	-61.27	**-6.26**	-3.68
CLIP (zero-shot,LAION)	72.77	-4.94	+12	-56.85	-8.43	-4.54
MAE (FT)	85.95	-4.36	+22	-62.48 (×1.02)	-36.46	-3.53
MAE+**LLE (ours)**	85.84	**-1.74**	+12	**-56.32**	-34.64	**-2.77**

Table 6. On ImageNet, many **self-supervised and foundation models amplify shortcuts**, whereas LLE mitigates multiple shortcuts jointly. (·): transfer learning (and extra data).

BG and CoObj across target classes on UrbanCars (details in Appendix B.5). On ImageNet, we use watermark augmentation (WMK Aug), style transfer [27] (TXT Aug), and background augmentation [73,93] (BG Aug) for watermark, texture, and background shortcuts, respectively.

The results on UrbanCars in Tab. 3 show that LLE beats all other methods in BG Gap and BG+CoObj Gap metrics and achieves second best CoObj Gap to CF+F Aug, a method amplifies the background shortcut. The results of ImageNet with ResNet-50 are in Tab. 4. LLE achieves the best multi-shortcut mitigation results in Carton Gap, SIN Gap, and IN-9 Gap. Regarding IN-W Gap and IN-R Gap, LLE achieves better results than ERM. *I.e.*, no Whac-A-Mole problems. On ImageNet, we further use MAE as the feature extractor, and the results on ImageNet are in Tab. 6. LLE achieves the best results in IN-W Gap, SIN Gap, and IN-9 Gap. LLE also achieves the best results in the remaining metrics comparing to methods not using extra pretraining data.

Ablation Study In Tab. 7, we show the ablation study of LLE: (1) w/o ensemble: training a single last layer. (2) AugMix (without ensemble): based on (1) and use JS divergence in AugMix to improve the invariance across augmentations. (3) w/o dist cls.: remove *domain shift classifier* and directly take the mean over the output of ensemble classifiers. Except for IN-R Gap, the full model achieves better results in all other metrics. Although the w/o ensemble achieves a better IN-R Gap, it suffers from reliance on other shortcuts.

6. Related Work

Group Shift Datasets Most previous works use single-shortcut datasets [4,33,44,48,56,60,61,74] to benchmark group shift robustness [74]. Although [8,79,97] use labels of multiple attributes [60] for evaluation, there lacks a sanity check on whether the selected attributes are learned as spurious shortcuts. [54,80] create MNIST-based [53] synthetic datasets with multiple shortcuts, where the shortcuts are unrealistic. In contrast, our UrbanCars dataset is more photo-realistic and contains commonly seen shortcuts. Besides, our ImageNet-W dataset better evaluates shortcut mitigation on the large-scale and real-world ImageNet dataset.

		Shortcut Reliance				
		Watermark		Texture		Background
	IN-1k	IN-W Gap ↑	Carton Gap ↓	SIN Gap ↑	IN-R Gap ↑	IN-9 Gap ↑
w/o ensemble	76.03	-6.71	+18	-66.81	**-52.55**	-5.08
AugMix	75.17	-7.27	+22	-66.33	-56.38	-5.38
w/o dist. cls.	75.82	-17.77	+36	-66.45	-53.58	-4.81
LLE (full model)	**76.25**	**-6.18**	**+10**	**-61.20**	-54.89	**-3.82**

Table 7. Ablation study of Last Layer Ensemble on ImageNet.

OOD Datasets of ImageNet While many models achieve great performance on ImageNet [18], they suffer under various distributional shifts, *e.g.*, corruption [35], sketches [87], rendition [34], texture [27], background [93], or unknown distributional shifts [37,69]. In this work, we construct ImageNet-W, where SoTA vision models rely on our newly discovered watermark shortcut.

Shortcut Mitigation and Improving OOD Robustness To address the shortcut learning problem [26], [39,74,89] use shortcut labels for mitigation. With only knowledge of the shortcut type, [5,88] use architectural inductive biases. [27,73,93] use augmentation and [42,46] re-trains the last layer for mitigation. Without knowledge of shortcut types, [3,15,54,59,61,79,84,96] infer pseudo shortcut labels, which is theoretically impossible [58], and we show that they struggle to mitigate multiple shortcuts. Other works suggest that self-supervised pretraining [30,45] and foundation models [10,28,28,41,67,91,92] improve OOD robustness. We show that many of them suffer from the Whac-A-Mole problem or struggle to close performance gaps.

7. Conclusion

We propose novel benchmarks to evaluate multi-shortcut mitigation. The results show that state-of-the-art models, ranging from shortcut mitigation methods to foundation models, fail to mitigate multiple shortcuts in a Whac-A-Mole game. To tackle this open challenge, we propose Last Layer Ensemble method to mitigate multiple shortcuts jointly. We leave to future work for shortcut mitigation without knowledge of shortcut types. Another promising future direction is to provide a theoretical analysis of the Whac-A-Mole phenomenon. Finally, we call for discarding the tenuous single-shortcut assumption and hope our work can inspire future research into the overlooked challenge of multi-shortcut mitigation.

Acknowledgment This work has been partially supported by the National Science Foundation (NSF) under Grant 1909912 and by the Center of Excellence in Data Science, an Empire State Development-designated Center of Excellence. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] Whack A Mole image is obtained from Flaticon.com.

[2] Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating Example Difficulty Using Variance of Gradients. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 27

[3] Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021. 8

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2, 8

[5] 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*, 2020. 8

[6] Yujia Bao and Regina Barzilay. Learning to Split for Automatic Bias Detection. *arXiv:2204.13749 [cs]*, 2022. 27

[7] Yujia Bao, Shiyu Chang, and Dr Regina Barzilay. Learning Stable Classifiers by Transferring Unstable Features. In *International Conference on Machine Learning*, 2022. 1

[8] Yujia Bao, Shiyu Chang, and Regina Barzilay. Predict then Interpolate: A Simple Algorithm to Learn Stable Classifiers. In *International Conference on Machine Learning*, 2021. 8

[9] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, 2019. 24, 27

[10] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021. 1, 3, 8

[11] Chun-Hao Chang, George Alexandru Adam, and Anna Goldenberg. Towards Robust Classification Model by Counterfactual and Invariant Data Generation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4, 5, 13, 17

[12] Hila Chefer, Idan Schwartz, and Lior Wolf. Optimizing Relevance Maps of Vision Transformers Improves Robustness. *Advances in Neural Information Processing Systems*, 2022. 19, 22

[13] Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 4, 7, 19

[14] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In *Advances in Neural Information Processing Systems*, 2021. 13

[15] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment Inference for Invariant Learning. In *International Conference on Machine Learning*, 2021. 1, 3, 4, 5, 8, 17, 26

[16] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 26

[17] Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 2021. 1

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1, 3, 4, 8, 19

[19] Greg d'Eon, Jason d'Eon, James R. Wright, and Kevin Leyton-Brown. The Spotlight: A General Method for Discovering Systematic Errors in Deep Learning Models. In *ACM Conference on Fairness, Accountability, and Transparency*, 2022. 27

[20] Terrance DeVries and Graham W Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*, 2017. 4, 19

[21] Thomas G. Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, 2000. 5

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. 3, 4, 19

[23] Elias Eulig, Piyapat Saranrittichai, Chaithanya Kumar Mummadi, Kilian Rambach, William Beluch, Xiahan Shi, and Volker Fischer. DiagViB-6: A Diagnostic Benchmark Suite for Vision Models in the Presence of Shortcut and Generalization Opportunities. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 27

[24] Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. Domino: Discovering Systematic Errors with Cross-Modal Embeddings. In *International Conference on Learning Representations*, 2022. 27

[25] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP). *International Conference on Machine Learning*, 2022. 4

[26] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. 1, 8

[27] 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. In *International Conference on Learning Representations*, 2019. 1, 3, 4, 5, 8, 17, 19, 21, 22

[28] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision Models Are More Robust And Fair When Pretrained On Uncurated Images Without Supervision. *arXiv preprint arXiv:2202.08360*, 2022. 3, 4, 7, 8, 19

[29] Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 14

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4, 7, 8, 16, 19

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3, 4, 19

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity Mappings in Deep Residual Networks. In *The European Conference on Computer Vision (ECCV)*, 2016. 19, 22

[33] Yue He, Zheyan Shen, and Peng Cui. Towards Non-I.I.D. image classification: A dataset and baselines. *Pattern Recognition*, 2021. 1, 8

[34] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 4, 8

[35] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*, 2019. 8

[36] 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. In *International Conference on Learning Representations*, 2020. 3, 4, 19

[37] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8, 24

[38] Mark Ibrahim, Quentin Garrido, Ari Morcos, and Diane Bouchacourt. The Robustness Limits of SoTA Vision Models to Natural Variation. *arXiv preprint arXiv:2210.13604*, 2022. 27

[39] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. *Conference on Causal Learning and Reasoning*, 2022. 4, 5, 8, 13, 26

[40] Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. ImageNet-X: Understanding Model Mistakes with Factor of Variation Annotations. In *International Conference on Learning Representations*, 2023. 27

[41] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. 8

[42] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On Feature Learning in the Presence of Spurious Correlations. In *Advances in Neural Information Processing Systems*, 2022. 8

[43] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling Model Failures as Directions in Latent Space. In *International Conference on Learning Representations*, 2023. 27

[44] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. BiaSwap: Removing Dataset Bias With Bias-Tailored Swapping Augmentation. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 8

[45] Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning Debiased Classifier with Biased Committee. In *Advances in Neural Information Processing Systems*, 2022. 8

[46] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. In *International Conference on Learning Representations*, 2023. 4, 5, 6, 8, 16, 22, 26

[47] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic Segmentation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 13

[48] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 8

[49] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan

Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big Transfer (BiT): General Visual Representation Learning. In *The European Conference on Computer Vision (ECCV)*, 2020. 19, 22

[50] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *The IEEE International Conference on Computer Vision Workshops*, 2013. 2, 13

[51] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. Explaining in Style: Training a GAN to explain a classifier in StyleSpace. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[52] Guillaume Leclerc, Hadi Salman, Andrew Ilyas, Sai Vemprala, Logan Engstrom, Vibhav Vineet, Kai Yuanqing Xiao, Pengchuan Zhang, Shibani Santurkar, Greg Yang, Ashish Kapoor, and Aleksander Madry. 3DB: A Framework for Debugging Computer Vision Models. In *Advances in Neural Information Processing Systems*, 2022. 27

[53] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 8

[54] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and Mitigate Unknown Biases with Debiasing Alternate Networks. In *The European Conference on Computer Vision (ECCV)*, 2022. 4, 5, 8, 17, 26

[55] Zhiheng Li and Chenliang Xu. Discover the Unknown Biased Attribute of an Image Classifier. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 27

[56] Weixin Liang and James Zou. MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts. In *International Conference on Learning Representations*, 2022. 8

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *The European Conference on Computer Vision (ECCV)*, 2014. 13

[58] Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. ZIN: When and How to Learn Invariance Without Environment Partition? In *Advances in Neural Information Processing Systems*, 2022. 5, 8, 25, 27

[59] Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information. *International Conference on Machine Learning*, 2021. 1, 3, 4, 5, 8, 17, 26

[60] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 8

[61] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: De-biasing Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 4, 5, 8, 17, 26

[62] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP. In *Advances in Neural Information Processing Systems*, 2022. 4

[63] Zoe Papakipos and Joanna Bitton. AugLy: Data Augmentations for Robustness. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 16

[64] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks. *Advances in Neural Information Processing Systems*, 2021. 4

[65] Francesco Pinto, Harry Yang, Ser-Nam Lim, Philip H. S. Torr, and Puneet K. Dokania. RegMixup: Mixup as a Regularizer Can Surprisingly Improve Accuracy and Out Distribution Robustness. In *Advances in Neural Information Processing Systems*, 2022. 4

[66] Xavier Soria Poma, Edgar Riba, and Angel Sappa. Dense Extreme Inception Network: Towards a Robust CNN Model for Edge Detection. In *The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 23

[67] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning*, 2021. 3, 4, 7, 8, 19

[68] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing Network Design Spaces. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 4, 19

[69] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, 2019. 8, 18, 24

[70] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 26

[71] Evgenia Rusak, Steffen Schneider, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. ImageNet-D: A new challenging robustness dataset inspired by domain adaptation. In *ICML 2022 Shift Happens Workshop*, 2022. 24

[72] Evgenia Rusak, Steffen Schneider, George Pachitariu, Luisa Eck, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. If your data distribution shifts, use self-learning. *Transactions on Machine Learning Research*, 2022. 24

[73] Chaitanya K. Ryali, David J. Schwab, and Ari S. Morcos. Characterizing and Improving the Robustness of Self-Supervised Learning through Background Augmentations, 2021. 4, 5, 8, 16, 22

[74] 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. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 4, 5, 8, 13, 16, 18, 26

[75] Christoph Schuhmann, Romain Beaumont, Cade W. Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta,

Clayton Mullis, Patrick Schramowski, Srivatsa R. Kundurthy, Katherine Crowson, Mitchell Wortsman, Richard Vencu, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3, 4, 19

[76] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *Advances in Neural Information Processing Systems Workshops*, 2021. 3, 4, 19

[77] Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun. Which Shortcut Cues Will DNNs Choose? A Study from the Parameter-Space Perspective. In *International Conference on Learning Representations*, 2022. 27

[78] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 3, 20

[79] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised Learning of Debiased Representations With Pseudo-Attributes. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[80] Robik Shrestha, Kushal Kafle, and Christopher Kanan. An Investigation of Critical Issues in Bias Mitigation Techniques. In *The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 8

[81] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting Weakly Supervised Pre-Training of Visual Perception Models. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4, 7, 19

[82] Sahil Singla and Soheil Feizi. Salient ImageNet: How to discover spurious features in Deep Learning? In *International Conference on Learning Representations*, 2022. 1

[83] Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, and Eric Horvitz. Understanding Failures of Deep Networks via Robust Feature Extraction. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 27

[84] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. In *Advances in Neural Information Processing Systems*, 2020. 8

[85] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 1999. 4, 5, 6

[86] Vasilis Vryniotis. How to Train State-Of-The-Art Models Using TorchVision's Latest Primitives. https://pytorch.org/blog/how-to-train-state-of-the-art-models-using-torchvision-latest-primitives, 2021. 4, 16

[87] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems*, 2019. 8, 23

[88] Haohan Wang, Zexue He, Zachary C. Lipton, and Eric P. Xing. Learning Robust Representations by Projecting Superficial Statistics Out. *International Conference on Learning Representations*, 2019. 8

[89] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4, 8, 26

[90] Ross Wightman, Hugo Touvron, and Hervé Jégou. ResNet strikes back: An improved training procedure in timm, 2021. 4

[91] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022. 3, 4, 7, 8, 19

[92] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[93] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or Signal: The Role of Image Backgrounds in Object Recognition. In *International Conference on Learning Representations*, 2021. 1, 3, 4, 5, 8, 16, 22

[94] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3, 4, 19, 25

[95] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*, 2018. 3, 4, 19

[96] Jianyu Zhang, David Lopez-Paz, and Léon Bottou. Rich Feature Construction for the Optimization-Generalization Dilemma. In *International Conference on Machine Learning*, 2022. 8

[97] Eric Zhao, De-An Huang, Hao Liu, Zhiding Yu, Anqi Liu, Olga Russakovsky, and Anima Anandkumar. Scaling Fair Learning to Hundreds of Intersectional Groups. In *Submitted to International Conference on Learning Representations*, 2022. 8

[98] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random Erasing Data Augmentation. *AAAI Conference on Artificial Intelligence*, 2020. 4, 19

[99] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2, 14