StarFT: Robust Fine-tuning of Zero-shot Models via Spuriosity Alignment

Younghyun Kim[1] , **Jongheon Jeong**[2] , **Sangkyung Kwak**[3] ,
Kyungmin Lee[4] , **Juho Lee**[4] and **Jinwoo Shin**[4]
[1]Samsung
[2]Korea University
[3]General Robotics
[4]KAIST
yh990220.kim@samsung.com, jonghj@korea.ac.kr,
sangkyung.kwak@generalrobotics.company, {kyungmnlee, juholee, jinwoos}@kaist.ac.kr

Abstract

Learning robust representations from data often requires scale, which has led to the success of recent zero-shot models such as CLIP. However, the obtained robustness can easily be deteriorated when these models are fine-tuned on other downstream tasks (*e.g.*, of smaller scales). Previous works often interpret this phenomenon in the context of domain shift, developing fine-tuning methods that aim to preserve the original domain as much as possible. However, in a different context, fine-tuned models with limited data are also prone to learning features that are spurious to humans, such as background or texture. In this paper, we propose StarFT (Spurious Textual Alignment Regularization), a novel framework for fine-tuning zero-shot models to enhance robustness by preventing them from *learning spuriosity*. We introduce a regularization that aligns the output distribution for spuriosity-injected labels with the original zero-shot model, ensuring that the model is not induced to extract irrelevant features further from these descriptions. We leverage recent language models to get such spuriosity-injected labels by generating alternative textual descriptions that highlight potentially confounding features. Extensive experiments validate the robust generalization of StarFT and its emerging properties: zero-shot group robustness and improved zero-shot classification. Notably, StarFT boosts both worst-group and average accuracy by 14.30% and 3.02%, respectively, in the Waterbirds group shift scenario, where other robust fine-tuning baselines show even degraded performance.

1 Introduction

Large-scale vision-language models [Radford *et al.*, 2021; Jia *et al.*, 2021; Zhai *et al.*, 2023] pre-trained on massive image-caption pairs are shown to have rich representations that generalize to a wide range of tasks, even without fine-tuning on task-specific data (*i.e.*, zero-shot generalization). These zero-shot models, such as CLIP [Radford *et al.*, 2021], have demonstrated impressive performance on diverse downstream tasks (defined by a set of textual prompts) without any fine-tuning on a specific target dataset. More intriguingly, zero-shot models are further reported to achieve unprecedented robustness across a range of benchmarks involving distribution shifts, which have been a major challenge in the literature [Taori *et al.*, 2020; Miller *et al.*, 2021]. This suggests that the ability to generalize on "out-of-distribution" (OOD) inputs may be an emergent property at scale of data.

Although existing zero-shot models provide reasonable performance on diverse tasks, one is often tempted to further *fine-tune* the models in practice when there are task-specific data available to improve their in-distribution (ID) performance. While such fine-tuning methods effectively enhance ID performance, they are also known to compromise the OOD robustness of the original zero-shot models [Bommasani *et al.*, 2022; Wortsman *et al.*, 2022]. As such, efforts have been recently made to understand the underlying causes of degradation in OOD robustness and mitigate the issue, which is referred to as *robust fine-tuning*. For example, [Goyal *et al.*, 2023] have shown that aligning fine-tuning objective with the pre-training stage can improve robustness, and several other works have introduced additional regularization terms [Mao *et al.*, 2022a; Nushi *et al.*, 2018].

However, in a broader context, the lack of model robustness is often attributed to learning *spurious features* [Geirhos *et al.*, 2020; Jaini *et al.*, 2024; Wichmann and Geirhos, 2023], *i.e.*, features that are not aligned with human decision-making but are present in the training data: *e.g.*, background [Xiao *et al.*, 2021], texture [Geirhos *et al.*, 2019], and resolution [Touvron *et al.*, 2019]. Existing robust fine-tuning approaches overlook this contributing factor that fine-tuned models depend on confounding decision rules. As it is likely that existing zero-shot models and their fine-tuned derivatives also possess certain types of spuriosity that may affect their robustness, further research has been demanded to explore and address them, particularly in the context of broader OOD robustness benchmarks.

Contribution. Motivated by this, we aim to interpret the degradation in OOD robustness of zero-shot models by focusing on the notion of *spuriosity*. We guide fine-tuned models to avoid constructing unnecessary decision rules so that

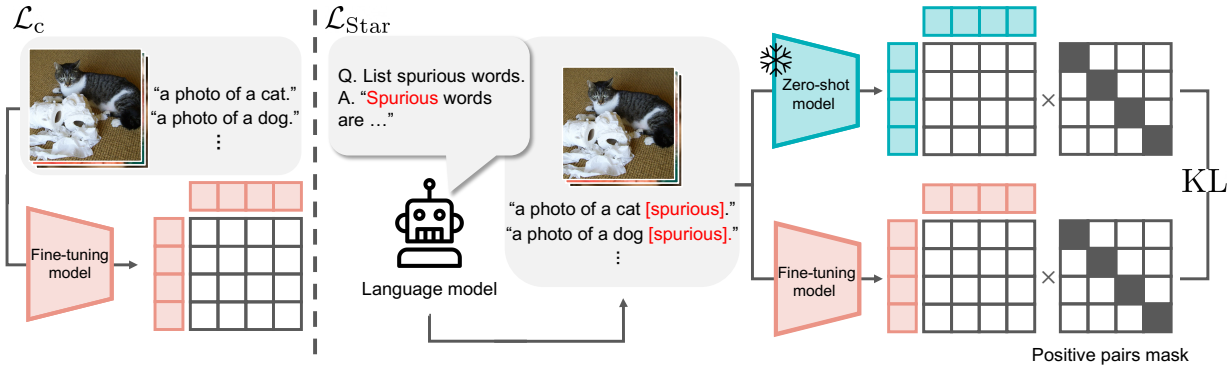


Figure 1: **Overview of StarFT.** Aside from the base contrastive objective $\mathcal{L}_{\text{c}}$, we propose a novel spuriosity textual alignment regularization $\mathcal{L}_{\text{Star}}$. We first extract spurious textual descriptions from language model, and corrupt the label textual descriptions. We then prevent fine-tuned models from learning spuriosity by minimizing the KL divergence of negative pairs' corrupted textual descriptions.

the models are more closely aligned with human decision making, which results in improved OOD robustness across diverse benchmarks. We mitigate the models' reliance on spuriosity by specifying irrelevant textual descriptions that they should not learn during fine-tuning. We identify the confounding features, such as background, with the aid of language models (LMs).[1] [2]

Specifically, we propose *Spurious Textual Alignment Regularization* (StarFT), a novel framework tackling spuriosity for robust fine-tuning of zero-shot models, *e.g.*, CLIP. We generate spurious textual descriptions by querying LMs using only general information, such as "image classification," without any other task-specific prompts, requiring no additional efforts to construct prompts. Then, we inject the obtained textual spuriosity into the label textual descriptions and regularize fine-tuned models to align the logits distribution of corrupted textual descriptions with zero-shot models. By directly providing spurious cues to the models and preventing models from extracting such cues during the fine-tuning process, we retrieve the obtained robustness of large-scale pre-trained zero-shot models.

We show that StarFT enhances various aspects of zero-shot models: OOD robustness, group shift robustness, zero-shot classification, and transfer learning. Although devising an advanced fine-tuning method for CLIP is a popular research topic recently, there exists no such "universally-good" method in the literature to the best of our knowledge.

Our contributions are:

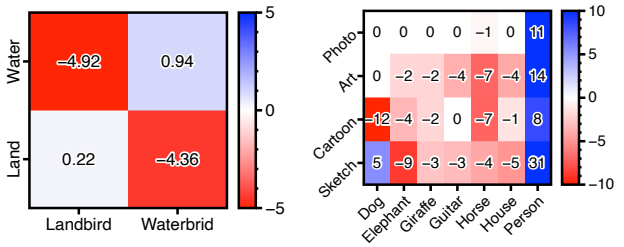
- From the observation that fine-tuned models learn spuriosity (Section 3.1), we propose a novel regularization loss that enforces fine-tuned models not to learn the spuriosity (Section 3.2).
- We demonstrate that spuriosity textual alignment has indeed improve OOD robustness, as supported by our experiments in domain shift benchmarks (Section 4.1).
- We also explore the application of spuriosity textual alignment, achieving zero-shot group robustness (Section 4.2); StarFT demonstrates group robustness in the Waterbirds dataset without having seen any Waterbirds data, where background bias is artificially injected.
- Finally, we show that our method enjoys a broader usage by applying it to zero-shot classification (Section 4.3) and transfer learning scenarios (Section 4.4).

2 Related Work

Out-of-distribution generalization. To deploy machine learning models for real-world applications, the generalization ability to unseen data distribution is crucial [Wiles *et al.*, 2022]. To remedy the performance degradation under distribution shifts, extensive efforts have been proposed to enhance the performance under benchmarks that focus on evaluating robustness [Torralba and Efros, 2011; Recht *et al.*, 2019; Hendrycks *et al.*, 2020; Shankar *et al.*, 2020; Hendrycks *et al.*, 2021a; Tramèr and Boneh, 2019; Paul and Chen, 2021; Mao *et al.*, 2022b; Wang *et al.*, 2021]. Prior works aim to increase OOD robustness by training with sophisticated data augmentations [Hendrycks *et al.*, 2020; Hendrycks *et al.*, 2021a], adversarial training [Tramèr and Boneh, 2019], using advanced network architecture [Paul and Chen, 2021; Mao *et al.*, 2022b] or extra information from test-time samples [Wang *et al.*, 2021]. Despite the tremendous efforts, there still exists a clear gap between the ID and OOD accuracy [Miller *et al.*, 2021]. Recent works on zero-shot models [Radford *et al.*, 2021; Jia *et al.*, 2021; Zhai *et al.*, 2023] have shown that the existing gap between ID and OOD performance can be notably reduced by scaling up the data curation, demonstrating large improvements in various robustness benchmarks. Our work is based upon these advances, by focusing on robust fine-tuning of recent zero-shot models.

Robust fine-tuning of zero-shot models. Motivated by the fact that fine-tuning zero-shot models is often at the cost of OOD generalization [Andreassen *et al.*, 2021; Kumar *et al.*, 2022; Wortsman *et al.*, 2022], various works have explored techniques to preserve the robustness of the zero-shot model while improving its ID accuracy [Li *et al.*, 2018; Wortsman *et al.*, 2022; Kumar *et al.*, 2022; Tian *et al.*, 2023;

[1] **Code**: https://github.com/alinlab/StarFT
[2] **Extended version**: https://arxiv.org/abs/2505.13232



(a) Waterbirds (b) PACS

Figure 2: **Subgroup accuracies in group shift benchmarks.** Differences of subgroup accuracies (%) between zero-shot and FLYP fine-tuned models.

Goyal *et al.*, 2023; Mao *et al.*, 2022a; Nam *et al.*, 2024; Oh *et al.*, 2024; Choi *et al.*, 2024]. Overall, they have studied post-hoc approaches [Wortsman *et al.*, 2022; Tian *et al.*, 2023], training schemes [Kumar *et al.*, 2022; Goyal *et al.*, 2023; Choi *et al.*, 2024], and regularization schemes [Mao *et al.*, 2022a; Nam *et al.*, 2024; Oh *et al.*, 2024] to better preserve the zero-shot model as prior knowledge. For instances, WiSE-FT [Wortsman *et al.*, 2022] uses a weight ensembling between the zero-shot and fine-tuned models, and [Tian *et al.*, 2023] have proposed a projection of the fine-tuned weights to be close to the zero-shot model; both in a post-hoc manner. FLYP [Goyal *et al.*, 2023] has shown that aligning fine-tuning objective with those of the pre-training stage can improve the OOD performance; AutoFT [Choi *et al.*, 2024] considers a bi-level optimization to search for a fine-tuning objective based on small OOD validation data; CAR-FT [Mao *et al.*, 2022a] regularizes context distributions induced by zero-shot and fine-tuned models, Lipsum-FT [Nam *et al.*, 2024] reduces energy gap using random texts, and CaRot [Oh *et al.*, 2024] adds self-distillation regularization.

3 Method

We propose StarFT, a robust fine-tuning method leveraging the spurious concept. Our motivation stems from the weakness of naïve fine-tuning objective in Section 3.1 and we explain the proposed method in Section 3.2. Our goal is to enhance generalization ability of fine-tuned models to OOD domains without compromising ID domain performance. Given a foundation model and a set of spurious concepts, StarFT regularizes models to avoid extracting spurious features other than label information, *i.e.*, "`[class]`," to preserve the robustness of the pre-trained foundation model.

Throughout the paper, we consider an open vocabulary image classification task, where the goal is to map an image $I \in \mathcal{I}$ to a label $y \in \mathcal{Y}$ using image-text aligned vision language models like CLIP [Radford *et al.*, 2021]. Given image-label pairs, we design label textual description $T \in \mathcal{T}_y$ of image I using templates such as "`a photo of a [class]`" with class names of each label y.

3.1 Motivation

Contrastive loss for fine-tuning. We start by considering a fine-tuning of the CLIP [Radford *et al.*, 2021] model on a labeled image dataset, specifically using contrastive loss as done in FLYP [Goyal *et al.*, 2023]. A typical practice for fine-tuning CLIP is to minimize cross-entropy loss; *viz.*, it initializes a new linear head upon the CLIP image embedding to define logits, discarding the textual labels of a given dataset. Therefore, the fine-tuned model is no longer suitable for open vocabulary tasks. In contrast, employing contrastive loss enables the fine-tuned model to continue processing language inputs as like CLIP. Formally, given a batch of N image-text pairs $\mathcal{B} = \{(I_i, T_i)\}_{i=1}^N$, let us denote $\boldsymbol{f}_\theta : \mathcal{I} \to \mathbb{R}^d$ an image encoder and $\boldsymbol{g}_\theta : \mathcal{T} \to \mathbb{R}^d$ a text encoder. Then the contrastive loss $\mathcal{L}_{\mathrm{C}}(\theta; \mathcal{B})$ is given as follows:

$$\mathcal{L}_{\mathrm{C}} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{e^{\mathbf{x}_i\cdot\mathbf{y}_i/\tau}}{\sum_{j=1}^{N}e^{\mathbf{x}_i\cdot\mathbf{y}_j/\tau}} + \log\frac{e^{\mathbf{x}_i\cdot\mathbf{y}_i/\tau}}{\sum_{j=1}^{N}e^{\mathbf{x}_j\cdot\mathbf{y}_i/\tau}}\right),$$

where $\mathbf{x}_i = \frac{\boldsymbol{f}_\theta(I_i)}{\|\boldsymbol{f}_\theta(I_i)\|_2}$ and $\mathbf{y}_i = \frac{\boldsymbol{g}_\theta(T_i)}{\|\boldsymbol{g}_\theta(T_i)\|_2}$ are ℓ_2-normalized image and text embeddings, and $\tau > 0$ is a temperature. When applied for fine-tuning, it first transforms the class labels of the given dataset into texts using some prompt template, *e.g.*, "`a photo of a [class]`," and optimizes the contrastive loss (3.1) updating both image and text encoders as well as the temperature τ.

However, it is uncertain whether the contrastive fine-tuning loss truly preserves the original ability of CLIP to associate text prompts to images, especially for prompts beyond the template-based prompts derived from the class labels (used during fine-tuning). To investigate this, we conduct the following group shift experiment asking whether the fine-tuned model retains its ability to handle diverse textual inputs.

Spuriosity in fine-tuned zero-shot models. We discover that although current robust fine-tuning method achieves higher accuracies in both ID and OOD, it also exhibits shortcut learning that learns spurious correlations from biased data [Sagawa *et al.*, 2020; Geirhos *et al.*, 2020] similar to the conventional fine-tuning. To see this, we examine how the subgroup accuracies of a CLIP model change through its fine-tuning (via FLYP) on ImageNet, as shown in Figure 2. We utilize group shift datasets such as Waterbirds [Sagawa *et al.*, 2020] and PACS [Li *et al.*, 2017], where the samples have extra labels indicating their subgroup domains. For example, in Waterbirds, we analyze an ImageNet fine-tuned CLIP using textual prompts such as "`a photo of a tench`," and subsequently test it under group shift scenarios with prompts like "`a photo of a waterbird in the mountain`." Here, FLYP tends to focus more on confounding features, such as background, rather than the core features, evidenced by its lower performance compared to CLIP on the minor subgroups (*i.e.*, landbirds in water background and waterbirds in land background).

Overall, the results demonstrate that even the most recent advanced robust fine-tuning methods still struggle to avoid spurious correlations during the fine-tuning process. This is an unfavorable behavior, especially since these fine-tuned models are often tested in more challenging conditions, such as real-world scenarios [Geirhos *et al.*, 2020].

3.2 StarFT: Fine-tuning with Spurious Textual Alignment Regularization

Motivated by the observation that fine-tuned zero-shot models show shortcut learning, we prevent the models from further constructing unnecessary decision rules during fine-tuning. To achieve this, our method, Spurious Textual Alignment Regularization fine-tuning (StarFT), introduces spuriosity suppressing regularization term that restricts model from shortcut learning as illustrated in Figure 1.

Spurious textual alignment regularization. To prevent the model from learning spurious features, *i.e.*, that leads to learning shortcut, we propose a novel fine-tuning objective named *spurious textual alignment regularization*, which uses the spurious descriptors. During fine-tuning with contrastive loss, we construct additional captions that include the spurious words and regularize the output of fine-tuning model with the zero-shot model. In particular, given a batch of N image-text pairs $\mathcal{B} = \{(I_i, T_i)\}_{i=1}^N$, we construct an additional batch $\mathcal{B}_\text{S} = \{(I_i, S_i)\}_{i=1}^N$ of image and spuriosity-augmented caption S_i, which is generated by attaching spurious keywords at each T_i. For instance, given the textual description of class name (*e.g.*, "`a photo of a [class]`"), we uniformly sample a spurious descriptor and augment to the textual description to generate spuriosity-augmented caption (*e.g.*, "`a photo of a [class] in the [descriptors]`").

Then, we compute regularization loss by measuring the KL divergence between the softmax outputs of fine-tuning and zero-shot models on $\mathcal{B}_S$. Specifically, we get the similarities between each image I_i and text S_i using fine-tuning model and zero-shot models. Then, we mask the logits (both fine-tuning and zero-shot) to exclude the pairs that are of same class labels and define q_i to be the softmax probability over the column of masked logits. q_i is given as follows:

$$[q_i]_j = \frac{e^{\mathbf{x}_i \cdot \mathbf{s}_j / \tau}}{\sum_{c(j') \neq c(i)} e^{\mathbf{x}_i \cdot \mathbf{s}_{j'} / \tau}}, \tag{1}$$

where $c(i)$ denotes the class label of I_i, τ is a temperature, and $\mathbf{x}_i = \frac{\boldsymbol{f}_\theta(I_i)}{\|\boldsymbol{f}_\theta(I_i)\|_2}$, $\mathbf{s}_j = \frac{\boldsymbol{g}_\theta(S_j)}{\|\boldsymbol{g}_\theta(S_j)\|_2}$ are image and text embeddings, respectively. By computing q_i, we represent the relative likelihood of image I_i that is related to the spurious descriptors S_j. Here, we mask out logits from the true class to address cases when zero-shot models perform poorly initially, so that distilling their confidence can be unfavorable: *e.g.*, as shown in Table 9 in Appendix. We define $\tilde{q}_i$ be the softmax over the masked logits of zero-shot models in a similar manner, and compute the spurious textual alignment regularization (Star) loss by KL divergence between q_i and $\tilde{q}_i$:

$$\mathcal{L}_\text{Star}(\theta; \mathcal{B}_S) = \frac{1}{N} \sum_{i=1}^{N} D_\text{KL}\left(\tilde{q}_i \,\|\, q_i\right). \tag{2}$$

Finally, we use linear combination of contrastive fine-tuning loss (*i.e.*, Eq. (3.1)) and Star regularization loss (*i.e.*, Eq. (2)) for fine-tuning, defining the objective of StarFT:

$$\mathcal{L}(\theta; \mathcal{B}, \mathcal{B}_S) = \mathcal{L}_\text{C}(\theta; \mathcal{B}) + \lambda_\text{Star} \mathcal{L}_\text{Star}(\theta; \mathcal{B}_S), \tag{3}$$

`[background]`	mountains	beach	desert
`[texture]`	rough	smooth	soft
`[resolution]`	blurred	bright	overexposed

Table 1: **Examples of spurious words.** Samples of the obtained spurious words for each spurious concepts. The list of concrete textual descriptions can be found in Appendix C.2.

where $\lambda_\text{Star} > 0$ is a hyperparamter.[3] This hyperparamter λ_Star is linearly decayed during the course of fine-tuning, which helps balancing the tradeoff between ID and OOD accuracy.

Obtaining spurious descriptors from LMs. As we leverage language supervision in contrastive learning objective, it is natural to use rich semantics of languages in aligning spuriosity. The advances of large language models have open possibilities of constructing some useful concept banks [Menon and Vondrick, 2023; Oikarinen *et al.*, 2023] or extracting task-specific spurious words [Adila *et al.*, 2024] for classifying images. Motivated by this, we aim to construct a new set of textual concepts that represents spuriosity, regardless of specific domains; see Table 1 for examples. We obtain textual descriptions of spurious concept that models often rely on while making decisions by querying language models. We prompt the LM with the input: "List possible spurious correleations while classifying natural images. Answer in a word." Given minimal task description, LMs provide spurious concepts like "background," "texture," and "resolution" that corresponds to our belief. Then, to directly corrupt the label textual description T, we ask LMs for more fine-grained spurious descriptors s. For example, for spurious concept "`[background],`" we get spurious descriptors such as "in the mountains" or "on the beach." We compose the spurious descriptors set $\mathcal{S}$ for each spurious concept, which is used to corrupt the textual description of an image.

Comparison with other methods. Recent works tackle the robust fine-tuning by introducing additional regularization term, such as CAR-FT [Mao *et al.*, 2022a] and CaRot [Nushi *et al.*, 2018]. CAR-FT optimizes the cross-entropy loss with additional dataset-specific context regularization to guide the fine-tuned model towards the zero-shot model. On the other hand, we regularize spuriosity rather than context, making it scalable to any other datasets. By directly addressing spuriosity, StarFT achieves better robustness than CAR-FT. Also, StarFT integrates with contrastive loss by injecting irrelevant information into the label textual descriptions. This direct adaptation eliminates the computational overhead associated with CAR-FT, which relies on averaged weights of context prompts across classes. CaRot, another recent baseline, optimizes the contrastive loss with additional regularization to maintain the image and textual logit distributions of zero-shot models. Additionally, CaRot updates zero-shot models using an exponential moving average (EMA). Due to this EMA update, CaRot fails on benchmarks where zero-shot models suffer. However, ours does not use EMA update to boost performance, and thus is not affected by whether zero-shot models

[3] In our experiments, we use $\lambda_\text{Star} = 0.5$ by default.

Method	ViT-B/16						ViT-L/14					
	IN	IN-R	IN-A	IN-S	IN-V2	Avg.	IN	IN-R	IN-A	IN-S	IN-V2	Avg.
Zeroshot	68.3	**77.7**	50.0	48.3	61.9	59.5	75.6	87.9	70.8	59.6	69.9	72.0
FT	81.3	71.3	44.5	49.1	71.7	59.1	84.7	75.4	55.7	54.4	75.3	65.2
FLYP	82.6	71.4	48.5	49.8	72.7	60.6	86.2	83.8	68.9	60.2	78.2	72.8
CAR-FT	81.9	75.6	50.0	51.5	72.8	62.5	86.3	84.2	66.6	60.0	76.8	71.9
CaRot	**83.1**	76.2	51.3	51.9	**74.3**	63.7	**87.0**	88.0	72.7	62.7	**79.3**	75.6
StarFT (Ours)	82.9	**77.7**	**53.7**	**52.5**	73.8	**64.4**	86.4	**88.7**	**73.8**	**63.2**	78.9	**76.1**

Table 2: **Evaluation on domain shifts.** We report Top-1 accuracies (%) on ImageNet (IN) and OOD datasets (IN-R, IN-A, IN-S, IN-V2), with their average values (Avg.) for two architectures (ViT-B/16 and ViT-L/14). StarFT outperforms the baselines under domain shifts scenarios on ImageNet. For example, on ViT-B/16, StarFT outperforms full fine-tuning by 5.3% on OOD and 1.6% on IN. We **bold** and underline the top two values in each column.

Method	Waterbirds		PACS		CIFAR-10.02	
	WG	Avg.	WG	Avg.	WG	Avg.
Zeroshot	25.9	87.1	87.4	93.0	47.0	87.0
FT	27.1	85.6	87.6	91.6	62.5	85.7
FLYP	21.5	87.2	86.0	92.5	61.0	88.2
CAR-FT	24.8	86.6	87.2	93.2	63.5	87.3
CaRot	27.1	89.5	88.5	94.2	63.0	89.3
StarFT (Ours)	**40.2**	**90.1**	**89.1**	**94.7**	**64.5**	**90.2**

Table 3: **Evaluation on group shifts.** We report worst-group (WG) and average (Avg.) accuracies (%) of ImageNet fine-tuned models on Waterbirds, PACS, and CIFAR-10.02 datasets. Notably, StarFT (Ours) consistently outperforms baselines including pre-trained CLIP without seeing any data from the group shift benchmarks. We **bold** and underline the top two values in each column.

Method	C-10	C-100	Cal101	STL10	Avg.
Zeroshot	90.8	68.2	89.6	98.3	86.7
FT	87.7	63.6	85.7	95.3	83.1
FLYP	90.0	64.2	87.4	98.5	85.0
CAR-FT	89.7	65.9	88.2	96.7	85.2
CaRot	91.1	66.7	89.0	98.7	86.5
StarFT (Ours)	**91.4**	**69.0**	**89.7**	**99.0**	**87.3**

Table 4: **Zero-shot evaluation.** We evaluate ImageNet fine-tuned models in different zero-shot scenarios on CIFAR-10 (C-10), CIFAR-100 (C-100), Caltech101 (Cal101), and STL10. Notably, without seen any data from the zero-shot benchmarks, ours consistently outperforms baselines including pre-trained CLIP. We **bold** and underline the top two values in each column.

perform poorly or not.

4 Experiments

First, we demonstrate the robustness of StarFT in diverse distribution shift scenarios (Section 4.1). Then, we present that the model fine-tuned with StarFT is robust to group shift benchmarks (Section 4.2) and outperforms on various object classification datasets (Section 4.3), even outperforming pre-trained zero-shot models. Lastly, we apply StarFT for standard transfer learning tasks (Section 4.4) demonstrating our method's efficacy in various scenarios.

Baselines. We compare StarFT with various fine-tuning methods for pre-trained zero-shot models; standard fine-tuning with classification loss (FT), and robust fine-tuning approaches such as FLYP [Goyal *et al.*, 2023], CAR-FT [Mao *et al.*, 2022a], and CaRot [Oh *et al.*, 2024].

Implementation details. Throughout experiments, we utilize CLIP [Radford *et al.*, 2021] ViT-B/16 and ViT-L/14 trained on the LAION dataset [Schuhmann *et al.*, 2021] and fine-tune the model using AdamW [Kingma and Ba, 2017] optimizer with a cosine learning rate scheduler. We train models with a batch size of 512 for ImageNet, while all other datasets use a batch size of 256. OOD datasets are only used for evaluation, where we select the best-performing model based on ID validation set accuracy. Across all datasets, we use the same text-templates as CLIP [Radford *et al.*, 2021] and WiSE-FT [Wortsman *et al.*, 2022]. Further implementation details are in Appendix B.

4.1 Evaluation on domain shifts

Datasets. To assess the performance of our approach across domain shifts, we train StarFT on ImageNet (IN) [Russakovsky *et al.*, 2015], which comprises over a million natural images of 1,000 classes. We then evaluate our fine-tuned models on 4 well-known ImageNet OOD benchmarks: ImageNet-R (IN-R) [Hendrycks *et al.*, 2021a], ImageNet-A (IN-A) [Hendrycks *et al.*, 2021b], ImageNet-Sketch (IN-S) [Wang *et al.*, 2019], and ImageNetV2 (IN-V2) [Recht *et al.*, 2019]. ImageNet-R contains visual renditions such as "cartoons" of ImageNet classes while ImageNet-Sketch contains sketches of ImageNet classes. ImageNet-A consists of naturally occurring samples that are misclassified by ResNet models. ImageNetV2 is a newly curated test set of ImageNet. Each of these variants reflects the domain shift of ImageNet.

Results. As shown in Table 2, StarFT shows the best OOD average accuracy while maintaining high ID accuracy for both small and large CLIP models. This highlights the effectiveness of our spurious alignment loss in enhancing domain robustness, providing empirical evidence that eliminating spuriosity during fine-tuning enhances model robustness against distribution shifts. On ImageNet, with ViT-B/16, our method surpasses full fine-tuning (FT) by 5.3% in OOD av-

Method	Caltech101	Cars	Flowers	ImageNet	iWILD	FMoW	Avg. Rank
Zeroshot	87.7	64.4	81.2	68.3	8.70	20.4	6.0
FT	97.0	84.2	92.4	81.3	45.2	**68.6**	3.8
FLYP	<u>97.1</u>	89.0	<u>97.1</u>	82.6	<u>48.5</u>	**68.6**	<u>2.2</u>
CAR-FT	96.1	84.3	94.5	81.9	45.8	<u>68.4</u>	3.7
CaRot	96.0	**89.8**	95.7	**83.1**	40.6	51.9	3.3
StarFT (Ours)	**97.2**	<u>89.5</u>	**97.5**	<u>82.8</u>	**50.1**	<u>68.4</u>	**1.7**

Table 5: **Transfer learning.** We evaluate our proposed approach on 6 different transfer learning datasets. We fine-tune with the downstream datasets and report the ID test accuracy. We **bold** and <u>underline</u> the top two values in each column.

erage accuracy and by 1.57% in ID accuracy. Similarly, with the larger ViT-L/14 architecture, StarFT achieves performance improvements on both ID and OOD datasets. In fact, ours shows the best OOD average accuracy even when compared to baselines with weight ensembling, *i.e.*, via WiSE-FT [Wortsman *et al.*, 2022], as shown in Table 8 in Appendix.

4.2 Evaluation on group shifts

Datasets. We further evaluate our ImageNet fine-tuned models on three different group shift benchmarks: Waterbirds [Sagawa *et al.*, 2020], PACS [Li *et al.*, 2017], and CIFAR-10.02 [Zhang and Ré, 2022]. As we are given the class labels from these datasets, we construct textual descriptions to perform zero-shot evaluations. Waterbirds classifies bird images into landbird and waterbird, with each class has two groups based on the background: bird on land background, and bird on water background. PACS classifies images into 7 categories and each image is from either arts, cartoons, photos or sketches, indicating different groups. CIFAR-10.02 classifies images into 10 categories. We follow [Zhang and Ré, 2022] to construct CIFAR-10.02, which combines the original CIFAR-10 [Krizhevsky, 2009] and CIFAR-10.2 [Lu *et al.*, 2020] from different data sources, defining each data source as a distinct group.

Results. To verify the effect of our spurious alignment loss in group shifts, which are closely related to spurious correlations, we leverage popular group shift benchmarks as a means of measuring spuriosity. As our objective is to enhance a wide range of robustness by targeting spuriosity, we do not additionally fine-tune with datasets in group shift benchmarks. Since all fine-tuned models are capable of zero-shot classification, we assess the spuriosity inherent in fine-tuned models using zero-shot performance on group shift benchmarks. We report both the worst group accuracy (WG) and the average accuracy (Avg.), as the worst group accuracy reflects the robustness of the model across different groups within the data. Notably, our spurious alignment loss indeed improves the group robustness of the fine-tuned model, resulting in the best WG accuracy among all baselines as depicted in Table 3. It is well known that improving worst group accuracy often comes at the cost of average accuracy [Sagawa *et al.*, 2020]. However, as opposed to this common drawback of eliminating spuriosity, StarFT improves both in the worst group and the average accuracy. This stands out in the Waterbirds benchmark, where ours narrows the gap between the worst group and the average accuracy to 49.88%, which is 61.19% in zero-shot models.

4.3 Zero-shot classification

We also conduct evaluation on zero-shot classification from the ImageNet fine-tuned models on 4 natural image benchmarks: CIFAR-10 [Krizhevsky, 2009], CIFAR-100 [Krizhevsky, 2009], Caltech101 [Li *et al.*, 2022], and STL10 [Coates *et al.*, 2011]. Overall, we observe that StarFT preserves the generalization ability of zero-shot models even after its fine-tuning. As shown in Table 4, StarFT outperforms the baselines including the base zero-shot models, across all datasets considered. In CIFAR-100 and Caltech101, although all fine-tuned baselines show deteriorated zero-shot performance compared to the zero-shot models, ours could maintain or even improve their accuracy.

4.4 Transfer learning

We compare the transferability of StarFT between various fine-tuning methods. We use 6 common object classification datasets: Caltech101 [Li *et al.*, 2022], StanfordCars [Krause *et al.*, 2013], Flowers102 [Nilsback and Zisserman, 2008], ImageNet [Russakovsky *et al.*, 2015], WILDS-iWILDCam [Beery *et al.*, 2020; Koh *et al.*, 2021], and WILDS-FMoW [Christie *et al.*, 2018; Koh *et al.*, 2021]. In our experiments, we use the same set of spurious descriptors across datasets. Table 5 shows the results. Compared to a strong baseline such as CaRot [Oh *et al.*, 2024], StarFT (Ours) shows slight degradation on Cars and ImageNet datasets. This is partly due to the CaRot's EMA update on zero-shot models which iteratively self-distills the knowledge of zero-shot models. However, this approach has drawbacks in scenarios where the zero-shot model performs poorly, such as in WILDS-iWILDCam and WILDS-FMoW. In these datasets, the zero-shot model struggles with ID accuracies of 8.70% and 18.7%, respectively, unlike other datasets. Consequently, CaRot performs significantly worse than other baselines, while StarFT shows consistent improvement across all 6 datasets with the best average rank in transfer learning since ours does not rely on EMA updates.

4.5 Ablation study

Corruption to label textual descriptions. We show the effect of each component in our proposed method in Table 6. We start from adding regularization to FLYP [Goyal *et al.*, 2023] which is the standard contrastive learning objective. Regularizing clean label textual descriptions without any spurious corruptions improves OOD accuracy but degrades the

$\mathcal{L}_{\text{Star}}$	Suffix	Mask	Decay	ImageNet					
				IN	IN-R	IN-S	IN-A	IN-V2	Avg.
–	–	–	–	82.6	71.4	48.5	49.8	72.7	60.6
✓	–	–	–	80.4	77.3	51.5	52.4	71.6	63.2
✓	Random	–	–	82.3	77.6	52.4	52.1	73.2	63.8
✓	Spurious	–	–	82.7	78.0	52.8	52.5	73.5	64.2
✓	Spurious	✓	–	82.3	78.2	54.2	52.6	73.3	64.6
✓	Spurious	–	✓	82.9	77.4	53.2	52.3	73.8	64.2
✓	Spurious	✓	✓	82.9	77.7	53.7	52.5	73.8	64.4

Table 6: **Ablation on different components.** We ablate on different parts of StarFT by starting with the baseline, where we regularize the clean label descriptions, and then add random and spurious textual descriptions to corrupt them. Next, we fix to the spurious suffix and examine the effects of other components: masking positive pairs and diminishing regularization ratio.

Concept	ImageNet					
	IN	IN-R	IN-S	IN-A	IN-V2	Avg.
`[background]`	82.9	77.7	53.7	52.5	73.8	64.4
`[texture]`	82.8	78.3	53.7	53.0	73.8	64.7
`[resolution]`	82.7	78.3	53.6	53.0	73.8	64.7
all	82.4	78.6	53.9	53.1	73.5	64.8

Table 7: **Ablation on different spurious concepts.** We ablate on different spurious concepts, including "background," "texture," "resolution," and "all." Adding spurious textual descriptions consistently improves both ID and OOD accuracies.

ID performance as it continues to interfere with learning to fit the in-domain dataset. Then we append the random textual tokens to label textual descriptions to see the effect of spurious descriptions in improvements. Results indicate that adding random tokens in the suffix improves both ID and OOD performance, where adding spurious textual descriptions further boosts the performance.

Component-wise analysis. We further test two additional components of StarFT in Table 6: masking and decaying regularization. For the masking, we observe an overall gain in OOD accuracies on ImageNet, and even larger gains on WILDS-iWILDCam (see Table 9 in Appendix); where there exist many over-confident "wrong" positive pairs (as zero-shot model suffers). Since positive pairs in the mini-batch have very higher confidences than those of negative pairs, it suppresses the model from learning useful knowledge that lies in spurious descriptions, thereby resulting in deterioration of both ID and OOD accuracies. Regarding the decaying regularization; we adopt the decaying λ_{Star} in StarFT primarily due to its effectiveness in preserving ID accuracy. Note that a key practical objective of robust fine-tuning is not only to improve OOD robustness, but also to ensure that ID performance is not compromised. As observed in Table 6, we find that the gain in ID performance (*e.g.*, $+0.6\%$ ImageNet accuracy) from the decaying strategy often outweighs the slight drop in OOD robustness (*e.g.*, -0.2% average OOD accuracy). For an additional ablation study, *e.g.*, on the effect of λ, and spuriosity concepts, see Appendix A.

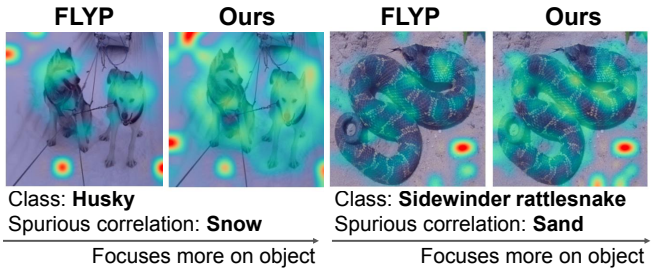


Figure 3: **Mitigation of spuriosity in ImageNet.** We display the GradCAM of fine-tuned models for comparison. Each class has the spurious correlations with background such as "snow" in "husky" and "sand" in "rattle snake." Rather than focusing on mostly background like FLYP, StarFT focuses on object itself to make decisions.

Choice of different spurious concept. By prompting language models, we attain the spurious concept descriptors of "background," "texture," and "resolution." We fine-tune StarFT using each of the spurious concepts and observe that adding spurious descriptions consistently improves both the ID and OOD accuracies. Combination of all concepts results in the slight improvement in OOD accuracy, however, at the cost of ID accuracy, meaning restricting spuriosity too much would sacrifice the ID accuracy. To study the efficacy of our methods in diverse downstream tasks, we fix our spurious concept to "background" throughout the experiments, which shows the best validation ID accuracy. However, we believe that further investigation on combinations of different spurious concepts would be promising research directions.

Spurious correlations in ImageNet. We observe that the StarFT reduces the spuriosity present in the zero-shot models while classifying ImageNet. To identify the spurious correlations in the zero-shot models, we adopt the setup of [Kim *et al.*, 2024] for each ImageNet classes via GradCAM [Selvaraju *et al.*, 2017]. We find that "snow" and "sand" are spuriously correlated to "husky" and "rattle snake," respectively, indicating a background bias in the zero-shot models. However, this reliance on the background is lower in StarFT as shown in Figure 3. Compared to ours, FLYP does not focus on object than background when classifying items.

5 Conclusion

It remains unclear which parts of the knowledge encoded in zero-shot models commit to their effective robustness, and how to preserve it during fine-tuning. We believe our approach of tackling *spuriosity* during fine-tuning suggests a novel view on understanding the robustness of zero-shot models. By devising a textual regularization that aims to prevent models from adopting spurious decision rules, aided by the textual interface of zero-shot models, we could improve the consistency of robustness across many benchmarks, where current methods have been inconsistent. We open the potential of studying the effect of spuriosity in zero-shot model's fine-tuning on wider notions of robustness. We further discuss on Limitations and Broader impacts in Appendix D.

Acknowledgements

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-II220184, 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics) and in part by Center for Applied Research in Artificial Intelligence(CARAI) grant funded by Defense Acquisition Program Administration(DAPA) and Agency for Defense Development(ADD) (UD230017TD). Jongheon Jeong acknowledges support from IITP grants funded by the Korea government (MSIT) (RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University); IITP-2025-RS-2024-00436857, Information Technology Research Center (ITRC); IITP-2025-RS-2025-02304828, Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents) and the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025).

Contribution Statement

The first two authors, Younghyun Kim and Jongheon Jeong, contributed equally to this work.

References

[Adila *et al.*, 2024] Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models. In *ICLR*, 2024.

[Andreassen *et al.*, 2021] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning, 2021.

[Beery *et al.*, 2020] Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 competition dataset, 2020.

[Bommasani *et al.*, 2022] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models, 2022.

[Choi *et al.*, 2024] Caroline Choi, Yoonho Lee, Annie Chen, Allan Zhou, Aditi Raghunathan, and Chelsea Finn. AutoFT: Learning an objective for robust fine-tuning, 2024.

[Christie *et al.*, 2018] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018.

[Coates *et al.*, 2011] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *ICML*, 2011.

[Geirhos *et al.*, 2019] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, et al. ImageNet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[Geirhos *et al.*, 2020] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[Goyal *et al.*, 2023] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *CVPR*, 2023.

[Grattafiori *et al.*, 2024] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Hendrycks *et al.*, 2020] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, et al. AugMix: A simple data processing method to improve robustness and uncertainty. In *ICLR*, 2020.

[Hendrycks *et al.*, 2021a] Dan Hendrycks, Steven Basart, Norman Mu, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021.

[Hendrycks *et al.*, 2021b] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

[Jaini *et al.*, 2024] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *ICLR*, 2024.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[Kim *et al.*, 2024] Younghyun Kim, Sangwoo Mo, Minkyu Kim, et al. Discovering and mitigating visual biases through keyword explanation. In *CVPR*, 2024.

[Kingma and Ba, 2017] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[Koh *et al.*, 2021] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[Krizhevsky, 2009] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

[Kumar *et al.*, 2022] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022.

[Li *et al.*, 2017] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

[Li *et al.*, 2018] Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, 2018.

[Li *et al.*, 2022] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022.

[Lu *et al.*, 2020] Shangyun Lu, Bradley Nott, Aaron Olson, et al. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

[Mao *et al.*, 2022a] Xiaofeng Mao, Yuefeng Chen, Xiaojun Jia, Rong Zhang, Hui Xue, and Zhao Li. Context-aware robust fine-tuning. *IJCV*, 2022.

[Mao *et al.*, 2022b] Xiaofeng Mao, Gege Qi, Yuefeng Chen, et al. Towards robust vision transformer. In *NeurIPS*, 2022.

[Menon and Vondrick, 2023] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023.

[Miller *et al.*, 2021] John Miller, Rohan Taori, Aditi Raghunathan, et al. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*, 2021.

[Nam *et al.*, 2024] Giung Nam, Byeongho Heo, and Juho Lee. Lipsum-FT: Robust fine-tuning of zero-shot models using random text guidance. In *ICLR*, 2024.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[Nushi *et al.*, 2018] Besmira Nushi, Ece Kamar, and Eric Horvitz. Towards accountable AI: Hybrid human-machine analyses for characterizing system failure. In *AAAI Conference on Human Computation and Crowdsourcing*, 2018.

[Oh *et al.*, 2024] Changdae Oh, Hyesu Lim, Mijoo Kim, Jaegul Choo, Alexander Hauptmann, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated robust fine-tuning of vision-language models, 2024.

[Oikarinen *et al.*, 2023] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *ICLR*, 2023.

[OpenAI, 2023] OpenAI. GPT-4 technical report, 2023.

[Paul and Chen, 2021] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *AAAI*, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Recht *et al.*, 2019] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, et al. ImageNet large scale visual recognition challenge. *IJCV*, 2015.

[Sagawa *et al.*, 2020] 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 *ICLR*, 2020.

[Schuhmann *et al.*, 2021] Christoph Schuhmann, Richard Vencu, Romain Beaumont, et al. LAION-400m: Open dataset of CLIP-filtered 400 million image-text pairs, 2021.

[Selvaraju *et al.*, 2017] 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 *ICCV*, 2017.

[Shankar *et al.*, 2020] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on ImageNet. In *ICML*, 2020.

[Taori *et al.*, 2020] Rohan Taori, Achal Dave, Vaishaal Shankar, et al. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020.

[Tian *et al.*, 2023] Junjiao Tian, Xiaoliang Dai, Chih-Yao Ma, Zecheng He, Yen-Cheng Liu, and Zsolt Kira. Trainable projected gradient method for robust fine-tuning. In *CVPR*, 2023.

[Torralba and Efros, 2011] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[Touvron *et al.*, 2019] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. In *NeurIPS*, 2019.

[Tramèr and Boneh, 2019] Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. In *NeurIPS*, 2019.

[Wang *et al.*, 2019] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

[Wang *et al.*, 2021] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

[Wichmann and Geirhos, 2023] Felix A. Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, pages 501–524, 2023.

[Wiles *et al.*, 2022] Olivia Wiles, Sven Gowal, Florian Stimberg, et al. A fine-grained analysis on distribution shift. In *ICML*, 2022.

[Wortsman *et al.*, 2022] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022.

[Xiao *et al.*, 2021] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*, 2021.

[Zhai *et al.*, 2023] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023.

[Zhang and Ré, 2022] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. In *NeurIPS*, 2022.