Exploring the Design of Adaptation Protocols for Improved Generalization and Machine Learning Safety

Puja Trivedi [1] **Danai Koutra** [1] **Jayaraman J. Thiagarajan** [2]

Abstract

While directly fine-tuning large-scale, pretrained models on task-specific data is well-known to induce strong in-distribution task performance, recent works have demonstrated that different adaptation protocols, such as linear probing before fine-tuning, can improve OOD generalization. However, the design space of such adaptation protocols remains under-explored and the evaluation of such protocols has primarily focused on distribution shifts. Therefore, in this work, we evaluate common adaptation protocols across distributions shifts and machine learning safety metrics (e.g., anomaly detection, calibration). We find that protocols induce disparate trade-offs that were not apparent from prior evaluation. Finally, we demonstrate that appropriate pairing of data augmentation and protocol can substantially mitigate this trade-off.

1. Introduction

Through larger datasets (Yalniz et al., 2019), better architectures (Zhai et al., 2022; Chen et al., 2021; Steiner et al., 2021; Tolstikhin et al., 2021), and novel self-supervised learning (SSL) frameworks (He et al., 2020; Chen et al., 2020; Grill et al., 2020; Caron et al., 2020), the quality of large-scale, pretrained models has drastically and rapidly improved; resulting in more robust (Hendrycks et al., 2019b; Liu et al., 2021), transferable (Ericsson et al., 2021) and semantically consistent (Caron et al., 2021) representations. While directly fine-tuning (`FT`) such models on task-specific data is known to improve in-distribution task performance (Neyshabur et al., 2020; Zhuang et al., 2019; Chen et al., 2020), recent work finds `FT` does not effectively leverage the expressiveness of large-scale, pretrained representations and fails to match the out-of-distribution (OOD) performance of other adaptation protocols, such as linear probing (`LP`) prior to `FT` (Kumar et al., 2022). Concurrently, Kirichenko et al. (2022) find that simply retraining the last (classifier) layer with a small amount of "re-weighting" or minority group data, can safeguard against spurious correlations. Crucially, both works suggest that well-designed adaptation protocols can improve both in-distribution task performance and robustness.

However, practical deployment requires that models are not only robust to such shifts, but that they also perform well with respect to safety metrics, such as anomaly detection and calibration error (Hendrycks et al., 2021a). Yet, recently proposed protocols focus only on a particular aspect of generalization behavior, potentially to the detriment of others. For example, while the `LP + FT` protocol improves out-of-distribution accuracy, its performance lags behind simple `FT` on several other metrics (see Fig. 1). Understanding and mitigating this trade-off is critical as all aspects are important to high-impact, low data tasks, such as healthcare.

Diversity-promoting data augmentation, such as RandAug (Cubuk et al., 2020), and CutMix (Yun et al., 2019), are becoming the *de facto* approach to improve model generalization; however, when not designed carefully, such sophisticated augmentations can adversely impact safety metric performance (Chun et al., 2020). In practice, it is unknown what characteristics of augmentations are beneficial to adaptation and where augmentations should be incorporated into adaptation protocols to maximize their benefits. Indeed, as shown in Fig. 1, naïvely incorporating such augmentations into adaptation protocols can lead to poorer performance than both `LP + FT` and simple fine-tuning. Therefore, in this paper, we holistically evaluate the behavior of adaptation protocols under distribution shifts and with respect to various safety metrics, and investigate how augmentations can be effectively leveraged to improve adaptation generalization behavior.

Proposed Work. We first make an important finding that the state-of-the-art adaptation protocol, `LP + FT`, improves

[1]Univesity of Michigan, Ann Arbor [2]Lawrence Livermore National Laboratory. Correspondence to: Puja Trivedi <pujat@umich.edu>.

Presented at ICML Workshop on Principles of Distribution Shift 2022,

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC.and was supported by the LLNL-LDRD Program under Project No. 22-ERD-006.

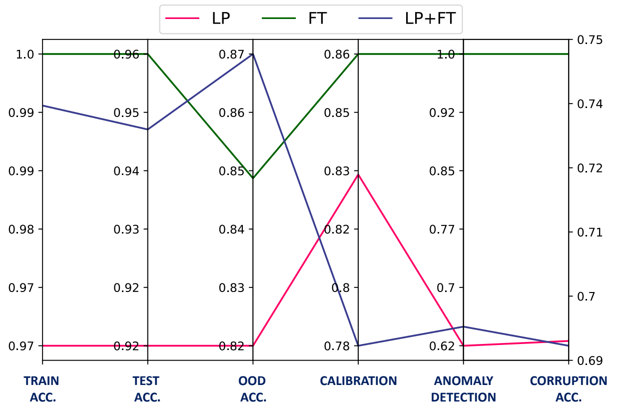

Figure 1: **Adaptation Protocols induce Trade-Offs.** While recently proposed protocol, `LP + FT`, can improve OOD accuracy, there are trade-offs with respect to other metrics.

out-of-distribution accuracy but lacks in the safety metrics relative to simple `FT`. This emphasizes the need for a holistic evaluation when understanding the behavior of task-specific adaptation strategies. We then use this evaluation to explore how augmentations influence the generalization behavior of different adaptation protocols. Given insights from this study, we hypothesize that hardness-promoting augmentations are needed during `LP`, while diversity-promoting augmentations can be used during `FT`, for effectively implementing `LP + FT` in practice. We verify this by employing virtual-adversarial training (Miyato et al., 2017) during `LP` and demonstrate substantial improvements across all measures.

- **Holistic Evaluation.** We evaluate all protocols with respect to in-distribution accuracy, out-of-distribution accuracy, calibration error, anomaly detection performance and robustness to corruptions.
- **The Effect of Augmentations in Adaptation.** We show that incorporating augmentations at different stages of adaptation protocols can lead to disparate generalization performance.
- **Hardness-Based Augmentations Improve Performance.** By using virtual-adversarial training during `LP` to promote a more amenable initialization for subsequent `FT`, we are able to improve generalization behavior across all metrics.

2. Background

In this section, we briefly discuss recently proposed adaptation protocols, relevant augmentation strategies and ML safety metrics.

Adaptation Protocols. In addition to directly `FT` on downstream task data or simply training a new classifier through `LP`, additional adaptation protocols have recently proposed to improve the robustness of adapted models. Kumar et al. (2022) argue that large-scale, pretrained models have expressive enough features to perform well on both in-distribution (ID) and OOD data. However, directly `FT` significantly distorts pretrained features toward ID data, harming OOD performance. To mitigate this distortion, they propose `LP` prior to `FT` and find this improves OOD performance. Concurrently, Kirichenko et al. (2022) find that retraining the last-layer of a model on minority group or re-weighting data can significantly improve robustness to spurious correlations. Like Kumar et al. (2022), they argue that the model has learned expressive features. However, these features are being poorly utilized, e.g., the classifier relies upon spurious instead of core features. Here, we focus on the `LP +FT` protocol as it is effective, inexpensive and does not require re-weighting data. We use adaptation, instead of transfer, to emphasize that we desire strong performance across robustness and safety measures in addition to strong downstream task performance.

Data Augmentation. Instead of building larger models or obtaining more data, data augmentation has been shown to be highly effective at improving the robustness, generalization and OOD generalization of models. We focus on popular, effective strategies AugMix (Hendrycks et al., 2020), AutoAug (Cubuk et al., 2018), CutMix (Yun et al., 2019), CutOut (Devries & Taylor, 2017), MixUp (Zhang et al., 2018) and RandAug (Cubuk et al., 2020).

Machine Learning Safety. Safe deployment of ML models requires that models are robust and reliable. While there are several aspect of model safety including robustness to distribution shift and adversarial samples (Hendrycks et al., 2021a), we focus on how well models are able to classify corrupted images (Hendrycks & Dietterich, 2019), how well calibrated uncertainty estimates are (Guo et al., 2017) and how well anomalous samples can be detected (Hendrycks & Gimpel, 2017; Hendrycks et al., 2019a). Evaluating additional aspect of ML safety is left to future work.

3. Designing Adaptation Protocols

In this section, we investigate the behavior of three adaptation protocols, with respect to both OOD generalization and ML safety metrics. We then investigate how incorporating popular diversity-promoting augmentations into these protocols impacts performance. Finally, we find that hardness-promoting augmentation can improve performance across all metrics. We first introduce the experimental setup.

Experimental Set-up. A ResNet-50 MoCoV2 (He et al., 2020) model pretrained on ImageNet-1K is used as the base-feature extractor to ensure high quality, expressive representations. CIFAR-10 is the in-distribution adaption dataset, while STL10 is the out-of-distribution dataset for which strong performance is also desired. Mean corruption accuracy (mCA) on CIFAR-10-C (Hendrycks & Dietterich, 2019), RMS calibration error, and AUROC when detecting anomalous inputs are the considered safety metrics (Hendrycks & Gimpel, 2017). mCA is the model's accuracy over 15 different corruptions and 5 different severities. Calibration error is measured as: $\sqrt{\mathbb{E}_C\left[(\mathbb{P}(Y=\hat{Y} \mid C=c)-c)^2\right]}$, where C is confidence and $\hat{Y}$ is the model's prediction. Blobs, Gaussian, LSUN, Places69, Rademacher, Textures, SVHN are the out-of-distribution datasets whose samples should be classified as anomalous. Our evaluation protocol closely follows Hendrycks et al. For the LP protocol, we only train the classifier for 200 epochs with LR=30. For FT, the entire model is trained for 20 epochs with LR=1e-5. For LP + FT, the model's classifier is initialized with the solution found by LP, and then it is fine-tuned for 20 epochs. A grid-search was conducted to determine the LR for LP and FT. When using augmented protocols, the same LRs are used. Note, all results were obtained by averaging over 3 seeds [2].

Need for Holistic Evaluation. As discussed in Sec. 2, Kumar et al. (2022) propose LP prior to FT as a means of mitigating feature distortion and improving OOD accuracy. In Table. 1, we indeed see that LP + FT has better OOD accuracy than both LP and FT. However, LP + FT's performance lags behind FT's on robustness to corruptions, calibration, and anomaly detection. This indicates that while mitigating feature distortion is important to ensure that FT does not over-fit to the in-distribution task, additional distortion may in fact be necessary for improving safety performance. Therefore, we ask how to modify the existing LP + FT protocol such that pretrained features are distorted in a way that is amenable to both improved OOD accuracy and safety.

Table 1: **Protocol Performance vs. Safety.** Best performance is shown in bold. Second best is underlined. FT outperforms LP + FT on all metrics but OOD Accuracy.

Protocol	mCA	RMS	AUROC	ID Acc.	OOD Acc.
LP	<u>0.6909</u>	<u>0.1697</u>	0.6206	0.9150	0.8194
FT	**0.7468**	**0.1366**	**1.0**	**0.9557**	0.8492
LP + FT	0.69	0.2166	<u>0.6454</u>	<u>0.9452</u>	**0.8714**

Role of Augmentations. Data augmentation is well-known to be effective in improving the robustness and generalization of end-to-end training (Hendrycks et al., 2020; Chun et al., 2020). However, relatively less work has focused on the role data augmentation plays when adapting high-quality pretrained representations to downstream tasks. Here, we evaluate 7 different diversity promoting augmentation strategies AugMix (Hendrycks et al., 2020), AutoAug (Cubuk et al., 2018), RandCrop+RandFlip, CutMix (Yun et al., 2019), CutOut (Devries & Taylor, 2017), MixUp (Zhang et al., 2018), and RandAug (Cubuk et al., 2020)) as they are applied at different points of adaptation protocols: LP +aug, FT+aug, during LP and not during FT, *i.e.*, (LP+aug) + FT, and vice-versa LP + (FT+aug). We begin by determining how these augmented protocols effect ID vs OOD performance and make the following observations from Fig. 2

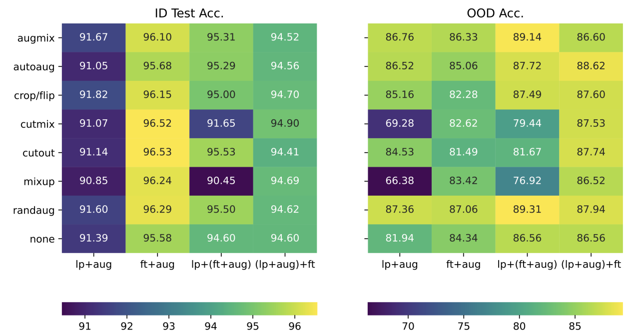

Figure 2: **Adding Augmentations to Protocols.** Most augmentation improve ID and OOD performance. However, naively adding MixUp and CutMix can be harmful.

Soft-Cross Entropy Loss Distorts Features. CutMix and MixUp interpolate between samples and labels during training. Therefore, these strategies require models to be trained with the soft cross entropy loss. Noticeably, under the LP + (FT+aug) protocol, with CutMix or MixUp, models have worse ID and OOD performance than LP + FT, without augmentations. Similarly, LP +aug and FT +aug protocols also produce poorer OOD performance than their augmentation-free counter-parts.

Using Augmentations improves OOD and ID Acc. Across all protocols and all augmentations (except for MixUp and CutMix), we see that OOD accuracy and often ID accuracy are improved relative to analogous augmentation-free protocols. Indeed, we particularly observe that, incorporating augmentations with LP + (FT + aug), can substantially improve OOD performance with comparable ID performance. For example, RandAug and AugMix both achieve $\geq 89\%$ OOD accuracy, in comparison to plain LP + FT's 87.14%.

Fine-tuning without augmentations can recover from poor LP solutions. While the OOD accuracy for CutMix and MixUp deteriorates across almost all protocols, (LP + aug) + FT is a notable exception. Here, we see that even if the linear-probing solution is poorly initialized, FT without any augmentations is still able to recover strong ID and OOD performance. Indeed, even with MixUp/CutMix LP initialization, (LP +aug) + FT still outperforms plain LP + FT's OOD accuracy.

[2]Code to be released at: https://anonymous.4open.science/r/classifier_playground-4EC6

In summary, these observations suggest that pixel-level augmentations are more effective at mitigating feature distortion to improve OOD accuracy and FT after LP is robust to poor LP initializations. Furthermore, we find that when taking into account the aforementioned holistic evaluation, incorporating augmentations can also improve safety performance. For brevity, we report safety performance for RandAug and Augmix, the two best performing augmentation strategies, as well as CutMix, a poorly performing strategy.

Table 2: **Protocol with Augmentation Performance vs. Safety.** Results shown for AugMix, RandAug and CutMix due to space constraints. Best performance in bold. Second best are underlined.

Protocol	mCA	RMS	AUROC	ID Acc.	OOD Acc.
LP	0.6909	0.1697	0.6206	0.9150	0.8194
LP+augmix	0.7264	0.1312	0.6477	0.9172	0.8673
LP+cutmix	0.6891	0.1333	0.5397	0.9121	0.6943
LP+randaug	0.7126	0.1259	0.6357	0.9167	<u>0.8728</u>
FT	0.7468	0.1367	**1.0000**	0.9558	0.8493
FT+augmix	**0.8139**	**0.0890**	**1.0000**	**0.9610**	0.8647
FT+cutmix	0.7669	0.1345	**1.0000**	**0.9652**	0.8217
FT+randaug	<u>0.7871</u>	**0.0824**	**1.0000**	**0.9629**	0.8682
LP + FT	0.6900	0.2166	0.6455	0.9452	<u>0.8714</u>
LP + (FT+augmix)	<u>0.7829</u>	<u>0.1089</u>	0.8074	<u>0.9528</u>	**0.8912**
LP + (FT+cutmix)	0.6663	0.1477	0.2677	0.9117	0.7923
LP + (FT+randaug)	0.7714	0.1190	<u>0.8305</u>	<u>0.9553</u>	**0.8934**
(LP+augmix) + FT	0.7136	0.1856	0.5533	0.9452	0.8669
(LP+cutmix) + FT	0.6926	0.1724	0.8062	0.9470	<u>0.8750</u>
(LP+randaug) + FT	0.7126	0.1259	0.6357	0.9167	<u>0.8728</u>

Augmented Protocols Improve Safety. Across all protocols, we see that RandAug and AugMix improve the safety performance in comparison to not using any augmentation. CutMix, which significantly harmed OOD accuracy under most protocols, occasionally provides some improved calibration or corruption performance. Notably, FT , which already had strong performance without augmentations, provides further improvement, while LP + (FT+aug) has the second best performance across safety metrics and the best OOD accuracy.

Hardness Promoting Augmentations. Our results in Table. 2 suggest that modifying the LP step prior to FT may be a viable strategy for simultaneously improving both OOD and safety performance. In particular, we hypothesize that hardness-promoting augmentations should be used during LP and diversity-promoting augmentations should be used during FT. Hardness-promoting augmentations will ensure that a smooth and robust classifier is learnt during LP, which should improve OOD performance during the subsequent fine-tuning step. While we leave theoretical analysis of our hypothesis to future work, we empirically verify it by using virtual adversarial training (VAT) (Miyato et al., 2017) during LP.

In a nutshell, VAT enforces local distribution smoothness by minimizing the KL-divergence between the predictions of perturbed pairs of examples, where the samples are adversarially perturbed such that outputs differ after perturbation. By training on such hard samples, classifiers become more robust and locally smooth. Moreover, because we are only applying VAT to the penultimate layer's representation (during LP), this step remains relatively inexpensive.

Table 3: **Benefits of Hardness Promoting Augmentations.** Incorporating VAT into the LP step leads to further improvements over previously identified high performing protocols.

Protocol	mCA	RMS	AUROC	ID Acc.	OOD Acc.
FT+augmix	**0.8139**	**0.0890**	**1.0000**	**0.9610**	0.8647
FT+randaug	0.7871	**0.0824**	**1.0000**	**0.9629**	0.8682
LP + (FT+augmix)	0.7829	0.1089	0.8074	<u>0.9528</u>	0.8912
LP + (FT+randaug)	0.7714	0.1190	0.8305	<u>0.9553</u>	0.8934
(LP+vat) + FT	0.7442	0.1645	0.871	**0.9611**	0.8909
(LP+vat) + (FT+augmix)	**0.8135**	**0.0817**	0.9253	**0.9638**	<u>0.9132</u>
(LP+vat) + (FT+randaug)	<u>0.8006</u>	<u>0.0900</u>	<u>0.9467</u>	**0.9655**	**0.9219**

As shown in Table. 3, we find that training on such examples leads to significant improvements across all-metrics relative to the best performing augmented protocols. Indeed, by incorporating VAT during LP, we are able to surpass the best OOD accuracy while performing comparably to the FT +aug protocol in terms of the safety metrics. Overall, the proposed (LP +vat) + (FT +aug) is a viable strategy for improving performance with respect to both distributional shifts and safety measures.

4. Conclusion and Future Directions

In this work, we explored how modifications to common adaptation protocols influence generalization under distribution shifts and performance with respect to various safety metrics. We make the somewhat surprising finding that while LP + FT does achieve impressive OOD accuracy, simple FT outperforms on the safety scores. We then find that diversity-inducing pixel-based augmentations can circumvent this challenge to an extent. However, to jointly achieve the benefits of plain FT and the LP + FT, hardness-inducing augmentation strategies, such as VAT, to generate challenging input perturbations are critical. Indeed, doing so allows models to match the performance of FT, while surpassing the best OOD accuracy by other protocols. There are several interesting directions for future work:

Expanded Evaluation. We would like to expand our analysis to include adversarial robustness and larger datasets. Moreover, we are also interested in using representational analysis tools such as prediction depth (Baldock et al., 2021) or CKA (Kornblith et al., 2019) to better understand how augmentation strategies and protocols change the representations and subsequent generalization behavior.

Theoretical Analysis. Extending the feature distortion analysis (Kumar et al., 2022) to analytically explain the benefits of hardness-promoting augmentations is also an interesting future direction.

References

Baldock, R. J. N., Maennel, H., and Neyshabur, B. Deep learning through the lens of example difficulty. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *NeurIPS*, 2021.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Chen, X., Hsieh, C.-J., and Gong, B. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

Chun, S., Oh, S. J., Yun, S., Han, D., Choe, J., and Yoo, Y. An empirical evaluation on robustness and uncertainty of regularization methods. *CoRR*, abs/2003.03879, 2020.

Cubuk, E. D., Zoph, B., Mané, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation policies from data. *CoRR*, 2018.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Devries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017.

Ericsson, L., Gouk, H., and Hospedales, T. M. How well do self-supervised models transfer? In *CVPR*, 2021.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent - A new approach to self-supervised learning. In *NeurIPS*, 2020.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In Precup, D. and Teh, Y. W. (eds.), *Proc. of the Int. Conf. on Machine Learning, (ICML)*, 2017.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL `https://doi.org/10.1109/CVPR42600.2020.00975`.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations, ICLR*, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Hendrycks, D., Mazeika, M., and Dietterich, T. G. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019a.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. In *In Proc. of NeurIPS*, 2019b.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ML safety. *CoRR*, abs/2109.13916, 2021a.

Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Li, B., Song, D., and Steinhardt, J. Pixmix: Dreamlike pictures comprehensively improve safety measures. *CoRR*, abs/2112.05135, 2021b.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *CoRR*, abs/2204.02937, 2022.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. E. Similarity of neural network representations revisited. In *Proceedings of the Int. Conf. on Machine Learning, ICML*, 2019.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2022.

Liu, H., HaoChen, J. Z., Gaidon, A., and Ma, T. Self-supervised learning is more robust to dataset imbalance. *CoRR*, abs/2110.05025, 2021.

Miyato, T., Maeda, S., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Neyshabur, B., Sedghi, H., and Zhang, C. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, 2020.

Steiner, A., Kolesnikov, A., , Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., and Dosovitskiy, A. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546, 2019.

Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., and Choe, J. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. of Int. Conf. on Computer Vision, ICCV*, 2019.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. *CVPR*, 2022.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *CoRR*, abs/1911.02685, 2019.