Do Objectives Matter OOD? Understanding the Impact of Self-Supervised Objectives on Robustness of Vision Transformers

Eric Tang
Department of Computer Science
Stanford University
Stanford, CA 94305
eatang@stanford.edu

Abstract

In this project, we investigate the impact of self supervised pretraining objectives on OOD robustness on downstream tasks. The two dominant paradigms for self supervised pretraining of vision transformer models for image and video representation learning are contrastive learning (i.e. MoCo, DINO), and masked image modelling (i.e. BEiT, MAE, SimMIM). We conduct experiments on iWildCam-Wilds, and find that although with linear probing, ViTs pretrained with contrastive learning objectives outperform masked image modelling objectives in terms of both effective and absolute robustness, fine tuning largely closes the gap in terms of effective robustness between the two methods. Additionally, although empirical risk minimization (ERM) methods dominate the iWildCam-Wilds leaderboard, we experiment with using invariant risk minimization (IRM) as an additional method for increasing the effective robustness of ViT models, but find that it is unable to match the performance of a model trained with ERM. Finally, we provide an analysis of the self attention mechanism comparing across different pretraining objectives, fine tuning vs linear probing, and in distribution vs out of distribution test data from iWildCam-Wilds. Code at: github.com/erictang000/wilds.

1 Introduction

Self-supervised learning has become the default method for pretraining large machine learning models across domains, both in natural language processing, with the emergence of the Transformer for machine translation (Vaswani et al. [2017]), and follow-up works including BERT and GPT for more general language modelling (Devlin et al. [2018], Brown et al. [2020]), as well as recently in the field of computer vision, with work like SimCLR and MAE (He et al. [2021], Chen et al. [2020b]) showing strong results across various image recognition tasks by utilizing large scale unlabeled image datasets and self supervised pretraining tasks. With the emergence and stronger potential performance of the Vision Transformer (Dosovitskiy et al. [2020]), which requires more pretraining and data due in part to the lack of inductive biases provided by the self-attention mechanism relative to convolutional networks like ResNets (He et al. [2015]) that previously dominated Computer Vision benchmarks, picking a self-supervised pretraining objective for learning strong general representations from is more important than ever.

Self supervision has previously shown to be a strong method of increasing the robustness of machine learning models to distribution shifts via pretext tasks and self supervised pretraining on downstream task domains before pretraining (Hendrycks et al. [2019], Gururangan et al. [2020], Sun et al. [2019]). However, prior work in the application of self-supervision for out of distribution robustness in computer vision has largely focused on using self-supervision tasks like predicting the rotation of an

37th Conference on Neural Information Processing Systems (NeurIPS 2023), Vancouver, Canada.

image as an auxiliary task that is co-trained with a supervised classification task. Current pretraining techniques on the other hand, typically consist of large scale self supervised learning from images prior to any supervised learning for downstream classification tasks.

The two currently dominant paradigms for self supervised pretraining of ViT models can largely be grouped into two sets (Park et al. [2023], Shekhar et al. [2023]) - Contrastive Learning (CL) objectives (SimCLR, DINO, MoCo (Chen et al. [2020b], Caron et al. [2021], Chen et al. [2021])), and Masked Image Modelling (MIM) objectives (MAE, BeiT, SimMIM (He et al. [2021], Bao et al. [2022], Xie et al. [2022])). CL objectives involve putting images into an embedding space such that they are close to positive samples, which often include different views of the same image using data augmentations to prevent representation collapse, while being far away in the embedding space from negative samples. Masked Image Modelling objectives consist of masking out patches of images, and training a ViT model to reconstruct the masked tokens from the remaining patches. Intuitively, the CL objective tends to learn higher level global features that are invariant across image views, while the MIM objective tends to learn lower level pixel level features due to the nature of the reconstruction task.

Given the prior work on the differences in representations learned by these two pretraining objectives, as well as work showing that pretraining can help increase effective robustness of vision models on the iWildCam-Wilds dataset (Miller et al. [2021]), we investigate whether the pretraining objective of ViT models impacts their robustness to distribution shift. On extensive experiments on iWildCam-Wilds, we find that the models that perform the strongest in terms of absolute out of distribution robustness are linearly probed CL models, and that linearly probed MIM models show lower absolute and effective robustness. However, we find that fine tuning both CL and MIM models on the downstream task increases effective robustness for both pretraining objectives, closing the gap between the two, but also decreases the absolute robustness of CL models. We additionally experiment with replacing the standard ERM training with the IRM objective proposed by Arjovsky et al. [2020], however, it shows lower absolute robustness without significant gains in effective robustness. Finally, in order to investigate the nature of the representations learned by the two objectives, and how they might contribute to learning robust representations, we visualize and analyze the self attention layers from a CL and MIM Model (DINO and MAE), showing that fine tuning decreases the average attention distance, resulting in more precise representations that are potentially more robust to distribution shifts, and lead to higher effective robustness.

2 Related Work

2.1 Contrastive Learning vs Masked Image Modelling

Recent work from Shekhar et al. [2023] and Park et al. [2023] both investigate the differences in representations learned by CL and MIM. Both works find that CL produces representations that lead to stronger linear probing performance, since relevant class discriminative information is available in the final CLS token, and that fine tuning MIM models leads to stronger performance than CL, through a reorganization of class information to the final layer. The works also confirm the idea that CL learns higher frequency global shape features, while MIM models learn lower frequency texture based features. Park et al. [2023] also measure the robustness of both methods to artificially injected noise, finding that CL is more robust to the noise due to the more texture based features learned by MIM. However, neither of these works compare the robustness of different SSL pretrained ViT models to natural distribution shifts, which we aim to do in this work.

2.2 Self Supervised Pretraining for Robustness

Hendrycks et al. [2019] and Sun et al. [2019] demonstrated the benefits on robustness of using pretext tasks like rotation for self supervised training that accompanied traditional supervised methods. Gururangan et al. [2020] showed that continuing to perform self supervised pretraining prior to supervised fine tuning on downstream tasks shows benefits in model performance and absolute robustness. Bordes et al. [2022] find that self supervised methods may be more robust to adversarial perturbations than fully supervised methods. Additional work has also been done on adding adversarial training into self supervised pretraining recipes (Chen et al. [2020a]), however, we do not consider adversarial settings in this work. Our work specifically focuses on self supervision as a standalone task, and specifically on the comparison between the CL and MIM objectives

3 Setup

3.1 Contrastive Learning Methods

Following Park et al. [2023] and Shekhar et al. [2023], we study DINO (Caron et al. [2021]) and MoCo-v3 (Chen et al. [2021]) as the methods for SSL CL on ViTs. DINO uses a self distillation technique, where two views of the same image are passed through a student and teacher encoder, and a cross entropy loss is used to enforce their similarity. The gradients are then propagated through the student, and the teacher's weights are updated as an exponential moving average of the student's. Similarly, MoCo uses a momentum encoder that is updated by taking an average of the updates on the query encoder. MoCo takes the standard approach of comparing a query key to a dictionary for computing a contrastive loss, but extends the size of the dictionary during the contrastive learning process by using previously encoded minibatches.

3.2 Masked Image Modelling Methods

For MIM, we consider MAE (He et al. [2021]), BeiT (Bao et al. [2022]), and SimMIM (Xie et al. [2022]). MAE uses an encoder decoder architecture, where the encoder takes in a subset of the patches from an image, and generates representations for them. These representations are then added to a set of mask tokens, and fed to the decoder, which outputs a reconstruction of the image. The model is trained end to end using a pixel wise reconstruction loss. The encoder is then used for downstream image classification tasks. BeiT poses the task of reconstruction via visual tokens, where image patches are first processed by a tokenizer, then an encoder takes in a masked representation and attempts to match the corrupted representation with the visual tokens. SimMIM is largely similar to MAE, but uses an l_1 loss rather than an MSE loss over the reconstruction target.

3.3 iWildCam Wilds Dataset

The iWildCam dataset (Koh et al. [2020]) consists of images captured of 182 different animal species across a number of different camera traps. The distribution shift observed in the dataset is in the shift from different camera deployments, and the task consists of a standard multi-class image classification of animal species across different domains. The dataset is split into a training set with 129809 images, an in-distribution and out of distribution validation set with 7314 and 14961 images respectively, and an in-distribution and out of distribution test set with 8154 and 42791 images.

3.4 IRM

Invariant Risk Minimization is a learning algorithm proposed by Arjovsky et al. [2020], as an alterative to the standard approach of empirical risk minimization in machine learning, in which the goal is to learn representations such that the optimal classifier on top of that representation matches for all environments.

4 Approach

We carry out a controlled study of Masked Image Modelling and Contrastive Learning as self supervised pretraining objectives for vision transformer models. As Fang et al. [2022] find that the additional robustness of CLIP is in large part due to additional data rather than the captions or self supervised pretraining methods, we ensure that models that we evaluate are pretrained solely on the ImageNet-1k data, with no extra sources of supervision or image data that could impact robustness. Models are then either linear probed or fine tuned on the downstream task, with the default hyperparameters from the Wilds dataset being selected for tuning and evaluation. We do not perform a hyperparameter sweep using the validation set, since our objective is solely to compare the performance of various pretrained models, with all other things held constant, and we do not necessarily care about the strength of the methods relative to other models and methods.

5 Empirical Results

Holding all other things constant, we fine tune and/or linear probe the following models using ERM:

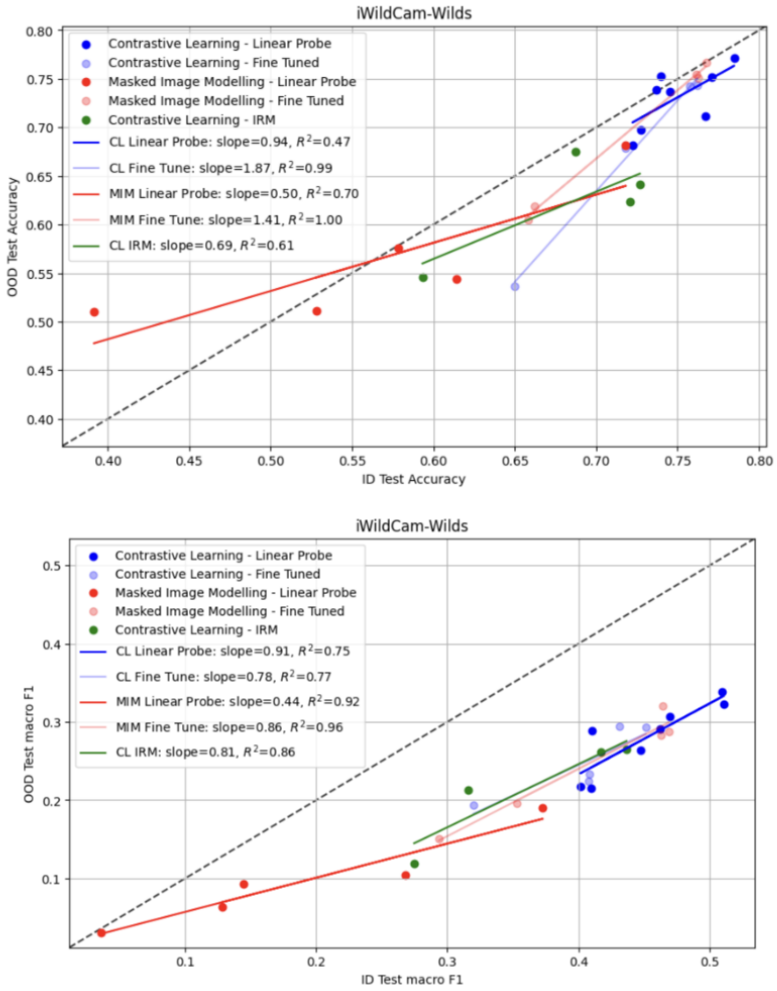


Figure 1: ID vs OOD Accuracy and F1 for CL and MIM models on iWildCam-Wilds. CL models with Linear Probing show the strongest absolute robustness, with significantly higher effective robustness than MIM models with Linear Probing. However, fine tuning closes the gap, increasing effective robustness for both methods, decreasing accuracy and F1 for CL, and increasing accuracy and F1 for MIM. Additionally, we show results for CL with IRM, and find that it performs similarly in terms of effective robustness to using a standard ERM method.

- Contrastive Learning:
  - DINO: ViT-B/16, ViT-S/16, ViT-S/8, ViT-B/8, XCiT-S/16, XCiT-S/8, XCiT-M/16, XCiT-M/8
  - MoCo-v3: ViT-B/16
- Masked Image Modelling
  - MAE: ViT-B/16, ViT-L/16
  - BeiT-v2: ViT-B/16, ViT-L/16
  - SimMIM: ViT-B/16

Additionally, we fine tune a subset of the DINO models using the IRM algorithm, the results of which can be found in 1. We use a batch size of 16, an input image resolution of (224, 224), a standard cross entropy loss, a learning rate of 3e-5, no weight decay, and an Adam optimizer for 12 epochs.

Inspired by Miller et al. [2021], in Figure 1, we plot the accuracy and F1 scores of our SSL pretrained models on the in distribution and out of distribution test sets of iWildCam-Wilds. We do not use the

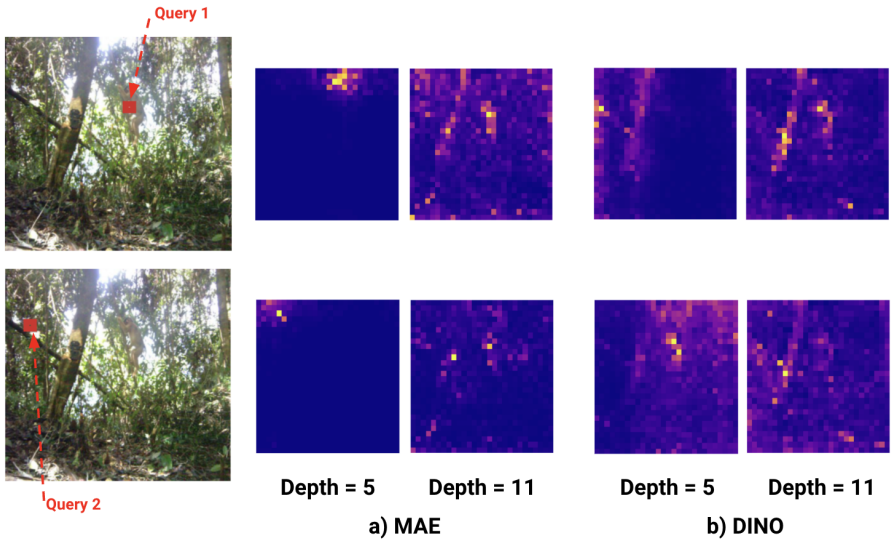


Figure 2: MIM tends to capture more local features, especially in early transformer layers, compared to CL, which more consistently captures more global features. Results shown using ViT-B/16 fine tuned for both MAE and DINO.

probit transform for Figure 1, but doing so does not change any of the observed trends. From Figure 1, we can see that CL models with linear probing show the best absolute robustness, with high ID and OOD test accuracy. CL linear probed models show a stronger effective robustness than MIM linear probed models, with a slope of 0.94 for CL compared to a slope of 0.5 for MIM considering accuracy and a slope of 0.91 vs 0.44 for F1. However, this gap is somewhat closed when both methods are fine tuned, with MIM showing a slope of 0.86 compared to a slope of 0.78 for CL when consdering F1, and a slope of 1.41 compared to 1.87 when considering accuracy. Noticeably, the slopes for both CL and MIM are above 1 when considering accuracy, indicating stronger effective robustness than when linear probing.

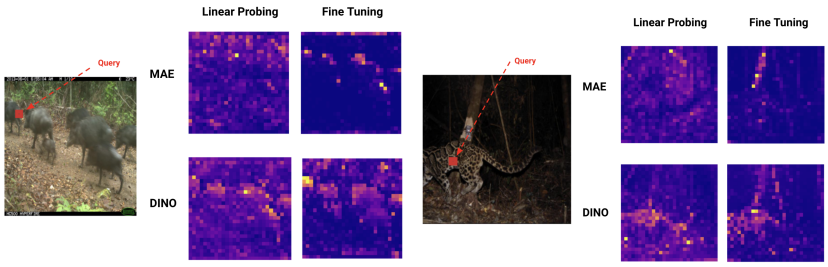
Additionally, we observe that fine tuning CL pretrained models with IRM rather than ERM does not seem to result in an increase in either absolute or effective robustness, which is consistent with prior results on iWildCam-Wilds that show that ERM methods seem to be able to perform the best on the OOD test set.

From our experiments, we conclude that CL models with linear probing are more robust than MIM models with linear probing, but that fine tuning for models trained with either pretraining objective leads to an increase in effective robustness, although potentially at the cost of absolute robustness.

5.1 Analyzing Self-Attention

In order to better understand the mechanisms behind the differences in learned representations between MIM and CL pretrained models, we visualize the attention maps on examples from the ID test set and OOD test set, which can be seen in Figure 2 and Figure 3. We see confirmation of prior results that MIM tends to capture more local features, particularly in earlier layers, while CL tends to capture more global features. From Figure 3 we can see that fine tuning of both DINO and MAE models results in less noisy self attention maps for examples both from the ID and OOD test sets. This potentially corresponds with the increase that we observed in the effective robustness of the models when fine tuned.

In addition to visualizing self attention maps, we also measure the average distance between the query token and the key tokens across both fine tuned and linearly probed MAE and DINO models. This roughly corresponds to the size of the receptive field in convolutional models. From Figure 4, we can see that fine tuning leads to a decrease in the average attention distance, especially for the MAE model, which we suggest corresponds to the increase in effective robustness observed in Figure 1 when fine tuning MIM models. The drop in attention distance for the fine tuned DINO model in



(a) Test ID (b) Test OOD

Figure 3: Fine tuning decreases the noise of the self attention operation for both MAE and DINO, both ID and OOD, corresponding with an increase in effective robustness. However, for CL models, this also leads to a decrease in the accuracy and F1 of models, suggesting that some useful global features from SSL pretraining may be lost during fine tuning. Results shown use ViT-B/16.

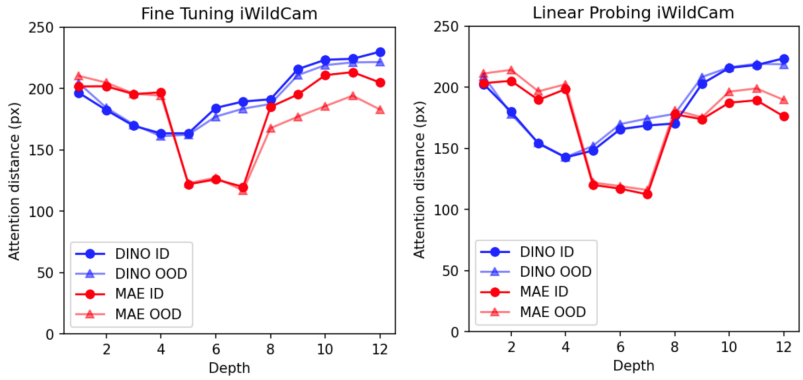


Figure 4: Fine Tuning leads to lower average distance attentions for MAE OOD, which also corresponds to an increase in both absolute and effective robustness. Results shown use ViT-B/16

Figure 4 corresponds with a decrease in absolute robustness and OOD accuracy, suggesting that fine tuning could potentially cause CL trained models to lose some global features learned during pretraining.

6 Discussion and Conclusions

In this work, we investigated the impact of SSL objectives on OOD robustness using iWildCam-Wilds. We found that with linear probing, ViTs pretrained with contrastive learning objectives outperform masked image modelling objectives in terms of both effective and absolute robustness. Although fine tuning largely closes the gap in terms of effective robustness between the two methods, it comes at the cost of absolute robustness of CL models. In order to investigate the effect of fine tuning on both MIM and CL pretrained models, we visualize and analyze the self attention layers from a CL and MIM Model (DINO and MAE), showing that fine tuning decreases the average attention distance, resulting in more precise representations that are potentially more robust to distribution shifts, and lead to higher effective robustness. However, this potentially harms CL accuracy more than MIM accuracy due to the global nature of the features learned by CL methods.

Future work would involve replicating these results across additional distribution shifts and with stronger models trained with CL and MIM. In addition, many models are now combining the two objectives in order to get the best of both worlds - seeing whether these models show stronger robustness OOD and learn strong representations would also be an interesting line of work to explore.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022.

Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about, 2022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL `https://arxiv.org/abs/2005.14165`.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. *CoRR*, abs/2003.12862, 2020a. URL `https://arxiv.org/abs/2003.12862`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020b. URL `https://arxiv.org/abs/2002.05709`.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

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. *CoRR*, abs/2010.11929, 2020. URL `https://arxiv.org/abs/2010.11929`.

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), 2022.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964, 2020. URL `https://arxiv.org/abs/2004.10964`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. URL `https://arxiv.org/abs/2111.06377`.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *CoRR*, abs/1906.12340, 2019. URL `http://arxiv.org/abs/1906.12340`.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. *CoRR*, abs/2012.07421, 2020. URL https://arxiv.org/abs/2012.07421.

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization, 2021.

Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn?, 2023.

Shashank Shekhar, Florian Bordes, Pascal Vincent, and Ari Morcos. Objectives matter: Understanding the impact of self-supervised objectives on vision transformer representations, 2023.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training for out-of-distribution generalization. *CoRR*, abs/1909.13231, 2019. URL http://arxiv.org/abs/1909.13231.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling, 2022.