Robust Fine-Tuning of Vision-Language Models for Domain Generalization

Kevin Vogt-Lowell*
Artificial Intelligence Technology
MIT Lincoln Laboratory
kevin.vogt-lowell@ll.mit.edu

Noah Lee*
Homeland Sensors and Analytics
MIT Lincoln Laboratory
noah.lee@ll.mit.edu

Theodoros Tsiligkaridis
Artificial Intelligence Technology
MIT Lincoln Laboratory
ttsili@ll.mit.edu

Marc Vaillant
Homeland Sensors and Analytics
MIT Lincoln Lab
marc.vaillant@ll.mit.edu

***Abstract*—Transfer learning enables the sharing of common knowledge among models for a variety of downstream tasks, but traditional methods suffer in limited training data settings and produce narrow models incapable of effectively generalizing under distribution shifts. Foundation models have recently demonstrated impressive zero-shot inference capabilities and robustness under distribution shifts. However, zero-shot evaluation for these models has been predominantly confined to benchmarks with simple distribution shifts, limiting our understanding of their effectiveness under the more realistic shifts found in practice. Moreover, common fine-tuning methods for these models have yet to be evaluated against vision models in few-shot scenarios where training data is limited. To address these gaps, we present a new recipe for few-shot fine-tuning of the popular vision-language foundation model CLIP and evaluate its performance on challenging benchmark datasets with realistic distribution shifts from the WILDS collection. Our experimentation demonstrates that, while zero-shot CLIP fails to match performance of trained vision models on more complex benchmarks, few-shot CLIP fine-tuning outperforms its vision-only counterparts in terms of in-distribution and out-of-distribution accuracy at all levels of training data availability. This provides a strong incentive for adoption of foundation models within few-shot learning applications operating with real-world data.**

***Index Terms*—foundation model, vision-language model, CLIP, fine-tuning, distribution shift, out-of-distribution robustness**

I. Introduction

In transfer learning, the knowledge acquired by a model trained to solve one task is leveraged to solve other different yet related tasks. Traditionally, the approach to transfer learning has involved the fine-tuning of a uni-modal, shared feature extractor using a large, annotated, task-specific dataset [1]. While this method is simple to implement and works well if the training dataset is similar to the target dataset, the resulting models have limited flexibility. Additionally, these models perform poorly without access to a large amount of quality target labels, so successfully training models via transfer learning proves difficult in disciplines that often operate within limited data environments. In these situations, successful deployment of new systems can involve considerable amounts of time and resources spent on data collection.

However, over the last several years, artificial intelligence research has increasingly shifted toward the creation of larger, more flexible models capable of reuse in a variety of applications. In particular, a new class of models known as foundation models, a term first-popularized by the Stanford Institute for Human-Centered AI, has begun to transform the way AI systems are built and developed [2]–[4]. Foundation models are defined as large neural networks that are trained on broad, unlabelled datasets through self-supervision at scale. Importantly, the unlabelled datasets used to train these models constitute samples of many different data types (i.e. RGB, thermal, text, audio, etc.) originating from a variety of perceptual data sources. Through self-supervised training over such diverse source data, foundation models become powerful, robust, general-purpose engines capable of multi-modal information processing and adaptation to a wide variety of downstream tasks with far less source data [5].

Domain generalization refers to learning models that are capable of maintaining high levels of performance on unseen target domains, or data distributions. In most applications involving real-world data, there exist real-world distribution shifts associated with the target environments in which the models are deployed. This shift can lead to a significant disparity between the distribution of data used to fine-tune a model and the distribution in which that model is actually tested [6]. As a result, traditional transfer learning solutions also underperform when applied to domain generalization problems. Research has mostly attempted to address this issue through experimentation involving different pre-training and data augmentation techniques, producing some improvement with methods like empirical risk minimization [7]–[11]. Yet, vision-language foundation models have been shown to possess significant robustness to distribution shifts. Given this robustness and their reduced fine-tuning needs, can these models overcome the limitations of transfer learning faced by vision-only models in regard to limited data scenarios and generalization under challenging distribution shifts?

[*] These authors contributed equally.

In [12], Radford et al. explored generalization and zero-shot transfer using the vision-language foundation model, CLIP (Contrastive Language-Image Pre-training) [12]. By leveraging natural language supervision during visual representation learning, CLIP achieves impressive zero-shot performance across a variety of benchmarks, demonstrating significant generalization capabilities. Notably, the paper highlights that the accuracy of zero-shot CLIP matches that of a pre-trained ResNet50 on ImageNet. These models were also evaluated on additional benchmarks representing ImageNet distribution shifts, and CLIP produced consistent zero-shot accuracy across all distribution shifts while ResNet50 could not. However, the ImageNet distributions tested represent relatively simple examples when compared to distributions encountered in real-world applications, like those provided by the WILDS collection [13]. When deploying zero-shot CLIP on more complex benchmarks with realistic distribution shifts, we found that the model significantly under-performed compared to trained state-of-the-art (SoTA) vision-only models.

Many studies have been conducted to explore and improve fine-tuning strategies for large, pre-trained models like CLIP. While existing fine-tuning strategies can substantially improve performance on target distributions, they also significantly reduce the model's robustness to distribution shifts [14]. Wortsman et al. explored methods to maintain CLIP's robustness during fine-tuning [15], finding that the ensembling of weights from zero-shot and fine-tuned CLIP models improved both accuracy under distribution shifts and accuracy on target distributions. Wortsman et al. further explored weight ensembling strategies in [16], improving robustness and inference accuracy by averaging the weights of CLIP models fine-tuned using different hyper-parameters. Xin Zhang et al. tried an entirely different approach for adapting CLIP to unseen target distributions [17], proposing a novel approach called Domain Prompt Learning which automatically generated tailored text prompts for CLIP by estimating domain-specific features from the target distribution. Yet these studies, like [12], only present results on distribution shifts derived from ImageNet, and they don't evaluate model performances in terms of gains over typical vision-only models. Yang Shu et al. [18] evaluate a fine-tuning strategy employing margin-based cross-entropy loss and beta moving average to improve generalization for CLIP, but add only DomainBed [19] as a benchmark alongside ImageNet. Furthermore, none of these studies evaluated their fine-tuning strategies in limited data environments.

In the field of remote sensing, deep learning models often operate within limited data environments. Smaller amounts of training data are available for these models due to a lack of raw data, increased difficulty of dataset annotation, and limitations of sensor characteristics. As a result, many studies on remote sensing image interpretation have explored few-shot learning methods in an effort to leverage the benefits of deep learning [20]. While these studies have explored data augmentation and prior-knowledge based transfer learning approaches, none have evaluated foundation models as potential few-shot learning solutions.

In this paper, we explore the use of vision-language foundation models as a modern solution to the limited training data and domain generalization issues inherent to previous transfer learning approaches. We provide the following contributions:

1) We reveal that the zero-shot classification performance of CLIP does not match that of a fine-tuned SoTA vision-only model under challenging realistic distribution shifts.
2) We demonstrate the superior performance of few-shot CLIP over a few-shot vision-only model in limited data environments containing realistic distribution shifts. We verify these results extend at various levels of training data availability by evaluating both in-distribution (ID) and out-of-distribution (OOD) robustness.
3) We present a fine-tuning strategy for CLIP that combines cross-entropy training and stochastic weight averaging to further improve out-of-distribution robustness.

II. Methodology

Our experiments compare results for both vision-only and vision-language models. Vision-language models are models where the visual output (i.e. image classification, object detection, segmentation) is conditioned on an additional text input. We adopt CLIP [12] as our model of choice to experiment with vision-language models.

A. Vision-Language Models

CLIP models are pre-trained with the task of matching text captions to images. The key advantages of CLIP as a research platform lie in its modularity, zero-shot capability, and simple formulation. Its encoders can be swapped out depending on desired accuracy versus model size trade-off, and the resulting performance scales predictably. Thus, CLIP is commonly used in systems as a foundation model: once pre-trained, CLIP can serve as the backbone of models fine-tuned for other common vision tasks. A key result of CLIP conditioning prediction on encoded text instead of predefined labels is the ability to transfer well to other tasks without fine-tuning (called zero-shot transfer). In [12], the authors show that zero-shot CLIP outperforms a trained ResNet50-based [21] logistic regression classifier on 16 lower-complexity datasets.

CLIP is implemented as two encoders, one for images and one for text. In (1), $I : Images \to U$ is the image encoder, $T : Text \to V$ is the text encoder. The encoder outputs are each projected to the embedding space W by $\mathcal{P}_I : U \to W$ and $\mathcal{P}_T : V \to W$ respectively, and classification is determined by the cosine similarity ($S_{cos}(a, b) := a \cdot b/(\|a\|_2 \|b\|_2)$) between each pair of text and image embeddings as shown in the below equation.

$$\arg\max_{y \in Text} S_{cos}(\mathcal{P}_I(I(x)), \mathcal{P}_T(T(y))) \tag{1}$$

As a training objective, CLIP uses an InfoNCE loss function with temperature scaling (τ), popularized by van den Oord et al. [22] and adapted for image-text learning by Zhang et al. [17]. For a batch size $|B|$, and a given image sample and embedding i_k, we compute the cross-entropy of the sample with the aligned text t_k and other unaligned text samples $t_{j\neq k}$ in the batch. The loss for a given text sample is computed in a similar yet symmetric manner. Losses of all image and text samples in the batch are averaged for back-propagation.

$$\mathcal{L}_{\text{image}} = -\mathop{\mathbb{E}}_{k\in|B|}\left[\log\frac{\exp(S_{cos}(i_k,t_k)/\tau)}{\sum_{t_j\in T}\exp(S_{cos}(i_k,t_j)/\tau)}\right] \quad (2)$$

$$\mathcal{L}_{\text{text}} = -\mathop{\mathbb{E}}_{k\in|B|}\left[\log\frac{\exp(S_{cos}(i_k,t_k)/\tau)}{\sum_{i_j\in I}\exp(S_{cos}(i_j,t_k)/\tau)}\right] \quad (3)$$

B. Vision-Only Models

For vision-only implementations of models, we extract the image encoder from CLIP. A randomly initialized trainable linear layer is added to the image encoder, with output dimension equal to the number of classes for the task we are fine-tuning for. Classification is performed using the softmax function, and training uses the standard cross-entropy objective. Following convention from [12], we differentiate between vision-only models depending on whether the image encoder's weights are frozen or not. If the encoder's weights are frozen and only the classification layer is left trainable, we call those 'linear probes'. Else, we label the models as 'fine-tuned'.

C. Weight Space Averaging

Recent work has demonstrated that averaging the weights of multiple models can approximate the generalization capabilities of deep ensembles [16], [23]–[25], as well as provide additional robustness to distribution shift [15]. To investigate this idea, we implement SWA from [25]. Our implementation runs for the final $n_{\text{SWA}} = 10$ epochs of training. At the end of each SWA epoch ($n_{\text{cur}} = 0, 1, \ldots, n_{\text{SWA}}$), the weights of the current model are aggregated with the moving average of model weights from past SWA epochs (4), which adds minimal overhead to the total training time.

$$w_{\text{SWA}} \leftarrow \frac{w_{\text{SWA}}\cdot n_{\text{cur}} + w_{\text{current}}}{n_{\text{cur}}+1} \quad (4)$$

During each epoch, the learning rate for each weight update per mini-batch follows a cosine annealing schedule (5), where $\eta_{\min}$ and $\eta_{\max}$ are the lower and upper bound learning rates and T_i and $T_{\max}$ are the current step and number of mini-batches in the epoch.

$$\eta_{i+1} = \eta_{\min} + \frac{1}{2}(\eta_{\max}-\eta_{\min})\left(1+\cos\left(\frac{T_i}{T_{\max}}\cdot\pi\right)\right) \quad (5)$$

D. Distributed Training

We also evaluate the effects of scaling the training of vision-language models. To successfully implement scaling for training CLIP, we make use of data parallelism [26] to split the training dataset across nodes. Communication collectives via NCCL [27] are used to synchronize gradients for models on different nodes. We also modify the training recipe of CLIP models by including the linear learning rate scaling rule and warmup periods suggested by Goyal et al. in [28]. For our distributed training environment, we run on the TX-GAIA system at the MIT Lincoln Laboratory Supercomputing Center [29]. The system allows the use of 2 Intel Xeon Gold 6248 CPUs (40 total cores) and 2 NVIDIA Volta V100 GPUs (32 GB RAM each) per node. We launch jobs using a SLURM workload manager [30], starting from 1 V100 and scaling up to 32 V100 GPUs.

III. Experimental Results

This section presents the key findings from our experimental results. We begin by providing a brief overview of the datasets used, as well as relevant details regarding our experimental setup and hyper-parameter testing. Then, we show zero-shot CLIP's failure to match the in-distribution (ID) and out-of-distribution (OOD) performance of a pre-trained Vision Transformer (ViT) in domain generalization tasks with challenging distribution shifts. Next, we demonstrate that CLIP outperforms the ViT at all levels of training data availability above zero-shot in these same domain generalization tasks, but particularly within few-shot scenarios. We show these results extend across both ID and OOD testing, verifying CLIP's improved robustness to challenging distribution shifts over vision-only models in limited data environments. Finally, we conclude with an evaluation of the performance improvements provided by SWA and the training time speed-ups gained from scaling training to multiple GPUs.

A. Datasets

To evaluate model robustness and performance on realistic distribution shifts, our chosen datasets came from Stanford University's WILDS collection, a set of domain generalization benchmarks representing challenging distribution shifts faced in the wild. [13] From WILDS, we selected the WILDS-FMoW (Functional Map of the World) [31] and the WILDS-iWildCam datasets [32], each containing geographic distribution shifts in satellite and wildlife imagery, respectively. Training and test distributions from the WILDS collection comprise disjoint sets of domains, allowing for adequate evaluation of model generalization to OOD test data.

B. Experimental Setup and Hyper-Parameters

For our experiments, we used a ViT-B/32 and a CLIP model using ViT-B/32 as its image backbone. To compare their performances at varying levels of data availability, the models were each trained on randomly sampled fractions of the training splits provided by FMoW and iWildCam. The exact fractions of available training data evaluated in these experiments can be seen in Table I. Models were then evaluated on the entirety of the ID and OOD test splits from each of the respective datasets.

In each of these varied data experiments, ViT-B/32 and CLIP models were separately fine-tuned twice: once as linear probes and once end-to-end (E2E). By testing different fine-tuning strategies for each model, we evaluate the optimal strategy for each. This approach resulted in four model permutations for comparison: ViT-B/32 linear probe, ViT-B/32

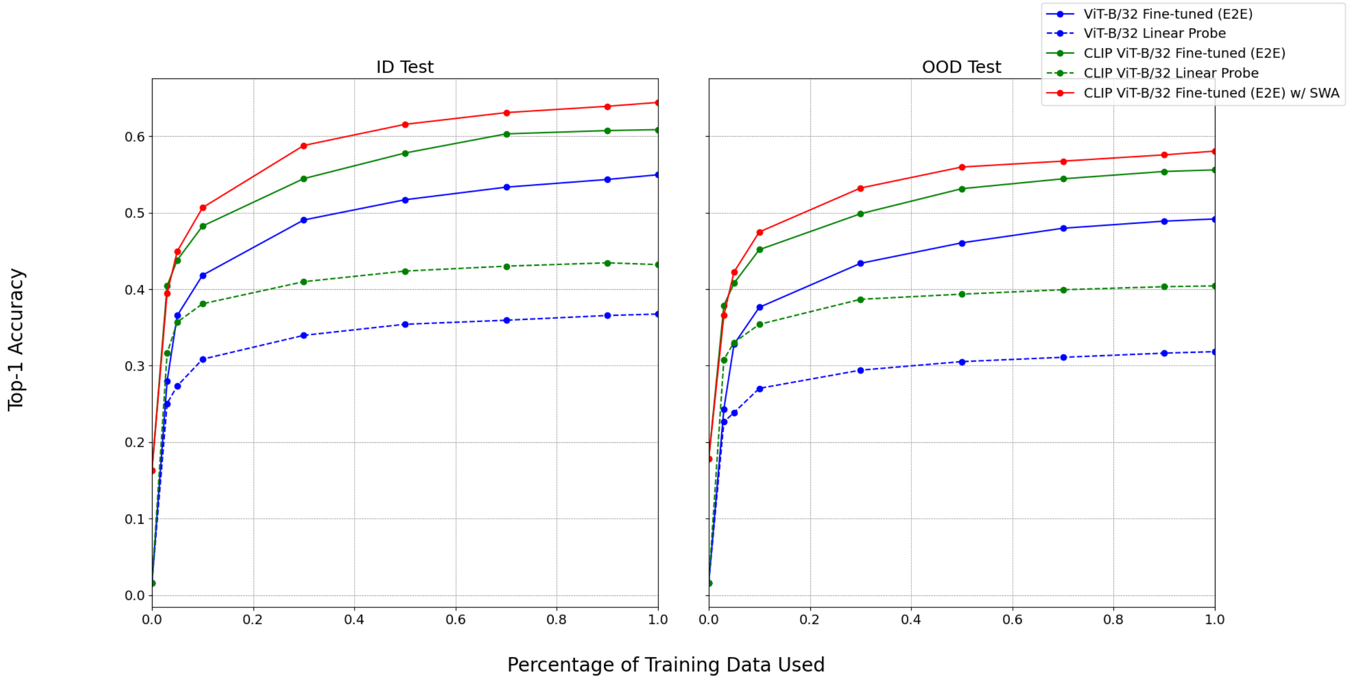

Fig. 1. Top-1 accuracies obtained on the FMoW in-distribution (ID Test) and out-of-distribution (OOD Test) test splits at different training data availabilities.

TABLE I
FMoW Top-1 Accuracies

	Training Data Used								
	0%	3%	5%	10%	30%	50%	70%	90%	100%
In-Distribution									
ViT-B/32 Linear Probe	1.6	25.1±0.3	27.4±0.1	31.8±0.1	34.0±0.2	35.4±0.4	35.9±0.1	36.6±0.1	36.7±0.1
CLIP ViT-B/32 Linear Probe	1.6	31.6±0.2	35.7±0.2	38.1±0.1	41.0±0.1	42.4±0.1	43.0±0.2	43.4±0.1	43.2±0.1
ViT-B/32 Fine-tuned (E2E)	1.6	28.0±2.5	36.5±0.7	41.8±0.3	49.0±0.5	51.7±0.2	53.3±0.1	54.3±0.1	55.0±0.5
CLIP ViT-B/32 Fine-tuned (E2E)	16.3	**40.4**±0.5	43.8±0.8	48.2±0.6	54.4±0.4	57.8±0.3	60.3±0.4	60.7±0.4	60.9±0.1
CLIP ViT-B/32 Fine-tuned (E2E) w/ SWA	16.3	39.5±1.2	**44.9**±0.5	**50.7**±0.4	**58.8**±0.5	**61.5**±0.1	**63.1**±0.3	**63.9**±0.4	**64.4**±0.1
Out-of-Distribution									
ViT-B/32 Linear Probe	1.6	22.7±0.3	23.9±0.1	27.0±0.1	29.4±0.1	30.5±0.1	31.1±0.1	31.6±0.1	31.8±0.1
CLIP ViT-B/32 Linear Probe	1.6	30.8±0.6	33.0±0.5	35.4±0.2	38.7±0.0	39.3±0.1	39.9±0.2	40.3±0.1	40.4±0.1
ViT-B/32 Fine-tuned (E2E)	1.6	24.3±2.2	32.8±0.6	37.6±0.2	43.4±0.4	46.1±0.1	48.0±0.2	48.9±0.5	49.2±0.4
CLIP ViT-B/32 Fine-tuned (E2E)	17.8	**37.9**±0.6	40.8±0.7	45.2±0.6	49.9±0.4	53.1±0.2	54.4±0.4	55.4±0.3	55.6±0.4
CLIP ViT-B/32 Fine-tuned (E2E) w/ SWA	17.8	36.6±0.8	**42.2**±0.9	**47.5**±0.1	**53.2**±0.4	**56.0**±0.2	**56.7**±0.1	**57.6**±0.5	**58.0**±0.1

fine-tuned, CLIP-ViT-B/32 linear probe, and CLIP-ViT-B/32 fine-tuned.

To account for the stochasticity inherent to the fine-tuning process, every model's top-1 accuracy (FMoW) or macro F1 score (iWildCam) was reported as the average of three separate, 20 epoch training runs. As part of model fine-tuning on each dataset, we included an exhaustive search over four learning rates of varying magnitudes: {1e-2, 1e-3, 1e-4, 1e-5}. All model permutations were fine-tuned using each of these learning rates in separate trials, and the best learning rate for each model selected based on the highest OOD performance for the given model configuration. The ID and OOD metrics reported in all figures and tables represent the averages yielded using each model permutation's best learning rate.

The magnitude of weight decay used during training on different datasets was chosen in a manner similar to the learning rate, examining {1e-1, 1e-2, 1e-3, 1e-4} as possible values. For both FMoW and iWildCam, a weight decay of 1e-4 was deemed appropriate. Batch size was held constant at a size of 128 across all experiments.

C. CLIP in Limited Data Scenarios

While zero-shot CLIP is capable of achieving comparable performance to vision-only linear probes on simpler distributions [12], our results highlight that the zero-shot performance

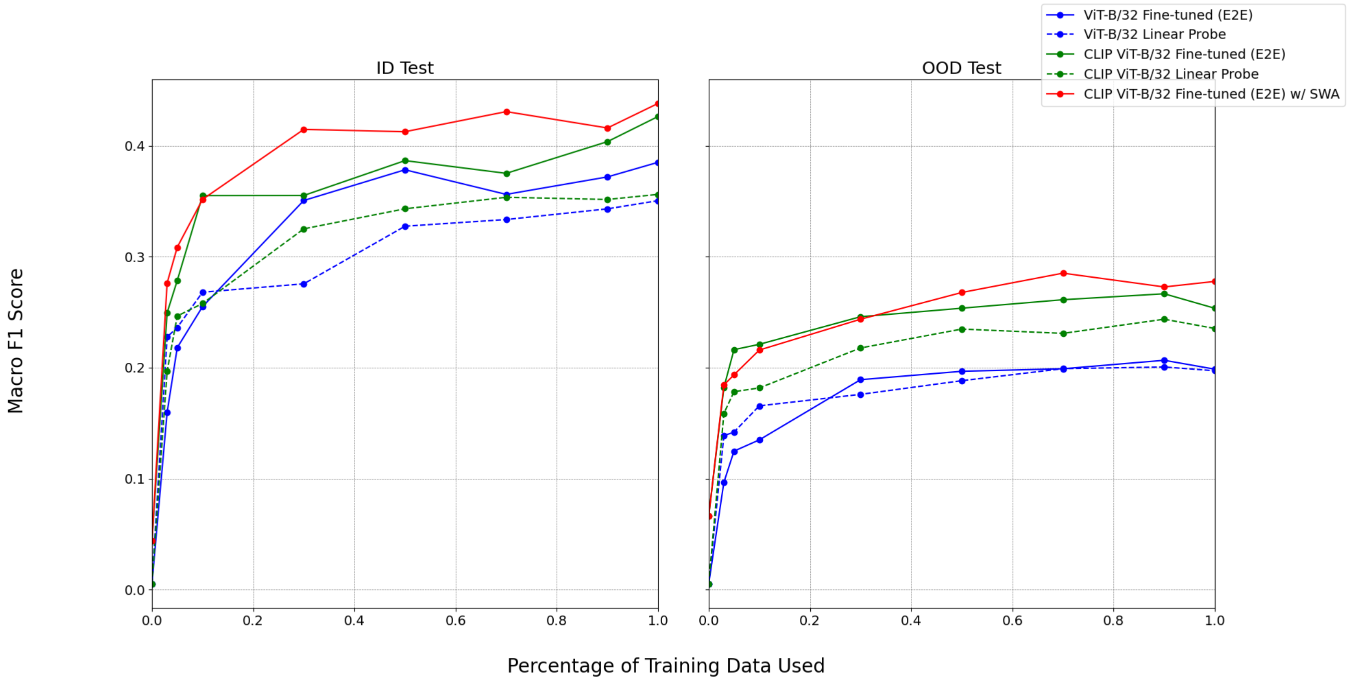

Fig. 2. Macro F1 scores obtained on the iWildCam in-distribution (ID Test) and out-of-distribution (OOD Test) test splits at different training data availabilities.

TABLE II
iWILDCAM MACRO F1 SCORES

	Training Data Used								
	0%	3%	5%	10%	30%	50%	70%	90%	100%
In-Distribution									
ViT-B/32 Linear Probe	0.5	22.8±0.7	23.6±0.6	26.8±1.5	27.6±0.2	32.8±0.7	33.4±0.8	34.3±0.6	35.1±0.6
CLIP ViT-B/32 Linear Probe	0.5	19.7±0.1	24.7±1.1	25.8±1.0	32.5±0.8	34.3±1.4	35.4±0.7	35.2±0.2	35.6±1.2
ViT-B/32 Fine-tuned (E2E)	0.5	16.0±0.6	21.8±1.7	25.5±0.1	35.1±1.4	37.8±1.7	35.6±0.8	37.2±0.4	38.5±0.6
CLIP ViT-B/32 Fine-tuned (E2E)	4.1	25.0±2.0	27.9±1.6	**35.5**±1.1	35.5±3.6	38.7±3.7	37.5±1.2	40.4±0.5	42.6±2.1
CLIP ViT-B/32 Fine-tuned (E2E) w/ SWA	4.1	**27.6**±0.9	**30.9**±1.1	35.2±0.9	**41.5**±0.9	**41.3**±1.4	**43.1**±0.9	**41.6**±0.2	**43.8**±0.5
Out-of-Distribution									
ViT-B/32 Linear Probe	0.5	13.9±0.2	14.2±0.3	16.6±0.6	17.6±0.3	18.8±0.3	19.9±0.3	20.1±0.7	19.7±0.2
CLIP ViT-B/32 Linear Probe	0.5	15.9±0.4	17.8±0.6	18.2±0.4	21.8±0.7	23.5±0.7	23.1±0.4	24.4±0.3	23.5±0.9
ViT-B/32 Fine-tuned (E2E)	0.5	9.7±0.3	12.5±1.6	13.5±1.0	18.9±0.6	19.7±0.7	19.9±0.3	20.7±0.6	19.9±0.7
CLIP ViT-B/32 Fine-tuned (E2E)	6.6	18.2±0.3	**21.6**±1.1	**22.1**±1.0	**24.6**±0.5	25.4±1.3	26.1±0.9	26.7±1.0	25.4±1.4
CLIP ViT-B/32 Fine-tuned (E2E) w/ SWA	6.6	**18.4**±0.8	19.4±0.7	21.6±0.6	24.4±0.4	**26.8**±0.7	**28.5**±0.7	**27.3**±0.8	**27.8**±0.8

of CLIP does not match that of a fine-tuned vision-only linear probe when applied to benchmarks from WILDS with challenging and realistic distribution shifts. As seen in Table I, zero-shot inference from CLIP scores a top-1 ID accuracy of 16.3% and OOD accuracy of 17.8% on FMoW, whereas the ViT-B/32 obtains an ID accuracy of 36.7% and OOD accuracy of 31.8% when trained as a linear probe using the full dataset. This trend holds for iWildCam as well, and can be seen in Table II. However, the zero-shot accuracies from CLIP still lie far above the accuracies expected from random guessing (1.6% and 0.5%, respectively), indicating that pre-trained vision-language models have some semantic understanding that can be leveraged.

The top-1 accuracies on FMoW for the linear probe and E2E fine-tuned ViT-B/32 and CLIP model permutations across multiple levels of training data availability above zero-shot can be seen in Fig. 1 and Table I. We can see that, at all levels of data availability, each of the CLIP model variations significantly outperforms its vision-only ViT-B/32 counterpart in terms of top-1 accuracy. This trend holds across testing on both ID and OOD samples. In particular, some of the biggest differences in performance between CLIP and ViT-B/32 can be seen in the few-shot scenarios, where data was limited. For example, the average difference between the top-1 accuracies for the E2E fine-tuned models was 6.1% ID and 6.7% OOD at training data availabilities above 5%, but was 9.9% ID and 10.8% OOD at training data availabilities at and below 5%. For all models, accuracy expectedly improves given increased amounts of available training data. The results for the same model permutations evaluated on iWildCam can be found in Figure 2 and Table II. Similar to the findings for FMoW, the CLIP model variants outperform the the ViT-B/32 at nearly all levels of training data availability when evaluated on both ID and OOD test samples, with the largest differences occurring when training data is most limited. Macro F1 scores tend to improve as models are given increased access to training data but some fluctuation occurs, likely related to the higher amount of rare class labels found in iWildCam compared to FMoW. In addition, the results from both datasets demonstrate that E2E fine-tuning of either model architecture yields higher accuracies for both ID and OOD testing than those obtained through linear probing alone. By incorporating SWA into the fine-tuning process, we further improved outcomes on both datasets at nearly all levels of data availability.

In terms of model robustness, the CLIP model variants were more robust than their ViT-B/32 counterparts. For FMoW, the average drop in top-1 accuracy between the ID and OOD test results was 4.4% for ViT-B/32 and 2.6% for CLIP when linearly probed, and 4.9% for ViT-B/32 and 4.3% for CLIP when fine-tuned E2E. For iWildCam, we observe the same trends regarding robustness under distribution shifts: F1 scores drop by 11.6% for ViT-B/32 and 9.4% for CLIP when linearly probed, and 14.0% for ViT-B/32 and 11.6% for CLIP when fine-tuned E2E. As expected from [14], the differences in robustness between the ViT and CLIP were most pronounced in few-shot environments, yet robustness for CLIP diminished when fine-tuned with larger amounts of training data. While SWA improved performance metrics for our best CLIP models, robustness across distribution shifts remained approximately the same.

D. Distributed Training

We investigate the effects of scaling training with SWA to multiple GPUs for Vision-Language models. Our experiments run for 30 epochs on the FMoW dataset. Training includes a 5 epoch linear warmup period, and the learning rate is scaled $\eta_1 \leftarrow k \cdot \eta_0$ by the number of GPUs used for training k. Note that scaling the number of GPUs is equivalent to scaling the effective batch size of training, since each GPU receives 128 samples per batch. We report the Top-1 ID and OOD accuracy, as well as the scale efficiency, which we define as the ratio between the time per epoch on one GPU T_1 scaled to k GPUs T_1/k and the actual time per epoch training on k GPUs T_k. As expected in III-D, the scale efficiency decreases as k increases due to inter-process communication overheads, necessary synchronization during training, and serial execution during evaluation. We also see in III-D that OOD accuracy does not begin to significantly degrade until 32 GPUs are used. This matches with the results of [28], where using a linear warmup, they observe no degradation in performance training on ImageNet until reaching batch sizes greater than 8192. Our results are further validated by 3, which shows that all training runs except for 32 GPUs converge to the same loss after 30 epochs, even though there is variation in the loss at the beginning of training. Overall, including the linear warmup helps make the training procedure robust to scaling.

TABLE III
DISTRIBUTED TRAINING SCALABILITY ON FMOW

GPUs	Top-1 Accuracy ID	OOD	% Scale Efficiency
1	**61.0** ±0.5	55.1 ±0.9	100.0 (367.21s)
2	60.4 ±0.3	**55.6** ±0.4	85.97 (213.53s)
4	58.9 ±0.4	54.2 ±0.6	81.75 (112.30s)
8	58.8 ±0.6	53.5 ±0.1	76.53 (59.98s)
16	59.6 ±0.7	54.2 ±0.5	78.98 (29.06s)
32	55.7 ±1.6	50.7 ±1.2	61.46 (18.67s)

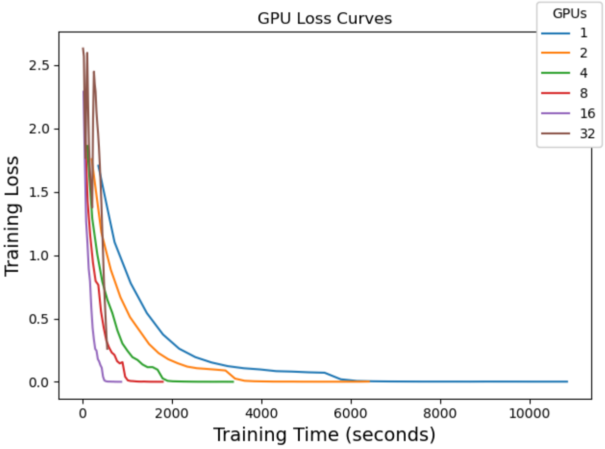

Fig. 3. Training loss curves for models trained on varying numbers of GPUs.

IV. CONCLUSION

This paper addresses the limitations of traditional transfer learning methods and the need for robust models capable of generalizing under distribution shifts. A new recipe for few-shot fine-tuning of the vision-language foundation model CLIP is proposed and its performance is evaluated on challenging benchmark datasets with realistic distribution shifts. The experimentation demonstrates that while CLIP's zero-shot inference capabilities do not match the performance of trained vision models on complex benchmarks, few-shot CLIP fine-tuning surpasses its vision-only counterparts in terms of in-distribution and out-of-distribution accuracy across different levels of training data availability. Adapting to limited training data, foundation models like CLIP offer enhanced performance and robustness in the face of distribution shifts, paving the way for more effective and practical machine learning applications. In future work, we will focus on improving domain generalization and novel class generalization performance of vision-language models in few-shot regimes by leveraging robust optimization and parameter-efficient methods.

Acknowledgment

The authors acknowledge the MIT Lincoln Laboratory Supercomputing Center for providing the high performance computing resources that have contributed to the research results reported within this paper.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

References

[1] K. Weiss, T. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, 05 2016.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[3] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *CoRR*, vol. abs/2102.12092, 2021.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020.

[5] R. Bommasani, D. A. Hudson, E. Adeli, R. B. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, and et al., "On the opportunities and risks of foundation models," *CoRR*, vol. abs/2108.07258, 2021.

[6] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *CoRR*, vol. abs/2103.02503, 2021.

[7] O. Wiles, S. Gowal, F. Stimberg, S.-A. Rebuffi, I. Ktena, K. D. Dvijotham, and A. T. Cemgil, "A fine-grained analysis on distribution shift," in *International Conference on Learning Representations*, 2022.

[8] T. Tsiligkaridis and A. Tsiligkaridis, "Diverse gaussian noise consistency regularization for robustness and uncertainty calibration," in *IJCNN*, 2023.

[9] R. Soklaski, M. Yee, and T. Tsiligkaridis, "Fourier-based augmentations for improved robustness and uncertainty calibration," in *NeurIPS 2021 Workshop on Distribution Shifts*, 2021.

[10] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *ICLR*, 2021.

[11] P. Teterwak, K. Saito, T. Tsiligkaridis, K. Saenko, and B. Plummer, "Erm++: An improved baseline for domain generalization," in *ICML 2023 Workshop on Spurious Correlations, Invariance and Stability*, 2023.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[13] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang, "WILDS: A benchmark of in-the-wild distribution shifts," *CoRR*, vol. abs/2012.07421, 2020.

[14] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," 2022.

[15] M. Wortsman, G. Ilharco, M. Li, J. W. Kim, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt, "Robust fine-tuning of zero-shot models," *CoRR*, vol. abs/2109.01903, 2021.

[16] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998, PMLR, 17–23 Jul 2022.

[17] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," *CoRR*, vol. abs/2010.00747, 2020.

[18] Y. Shu, X. Guo, J. Wu, X. Wang, J. Wang, and M. Long, "Clipood: Generalizing clip to out-of-distributions," 2023.

[19] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *CoRR*, vol. abs/2007.01434, 2020.

[20] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2387–2402, 2021.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[22] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[23] A. Ramé, M. Kirchmeyer, T. Rahier, A. Rakotomamonjy, P. Gallinari, and M. Cord, "Diverse weight averaging for out-of-distribution generalization," 2023.

[24] J. Cha, S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park, "Swad: Domain generalization by seeking flat minima," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[25] P. Izmailov, D. Podoprikhin, T. Garipov, D. P. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," in *Conference on Uncertainty in Artificial Intelligence*, 2018.

[26] S. Li, Y. Zhao, R. Varma, O. Salpekar, P. Noordhuis, T. Li, A. Paszke, J. Smith, B. Vaughan, P. Damania, and S. Chintala, "Pytorch distributed: Experiences on accelerating data parallel training," 2020.

[27] Nvidia, "Nvidia collective communication library."

[28] P. Goyal, P. Dollár, R. B. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch SGD: training imagenet in 1 hour," *CoRR*, vol. abs/1706.02677, 2017.

[29] A. Reuther, J. Kepner, C. Byun, S. Samsi, W. Arcand, D. Bestor, B. Bergeron, V. Gadepally, M. Houle, M. Hubbell, M. Jones, A. Klein, L. Milechin, J. Mullen, A. Prout, A. Rosa, C. Yee, and P. Michaleas, "Interactive supercomputing on 40,000 cores for machine learning and data analysis," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pp. 1–6, IEEE, 2018.

[30] A. B. Yoo, M. A. Jette, and M. Grondona, "Slurm: Simple linux utility for resource management," in *Job Scheduling Strategies for Parallel Processing* (D. Feitelson, L. Rudolph, and U. Schwiegelshohn, eds.), (Berlin, Heidelberg), pp. 44–60, Springer Berlin Heidelberg, 2003.

[31] G. A. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," *CoRR*, vol. abs/1711.07846, 2017.

[32] S. Beery, E. Cole, and A. Gjoka, "The iwildcam 2020 competition dataset," *CoRR*, vol. abs/2004.10340, 2020.