Anchor-based Robust Finetuning of Vision-Language Models

Jinwei Han[1], Zhiwen Lin[2], Zhongyisun Sun[2], Yingguo Gao[2], Ke Yan[2]
Shouhong Ding[2], Yuan Gao[3†], Gui-Song Xia[1†]
[1]School of Computer Science, Wuhan University [2]YouTu Lab, Tencent
[3]Electronic Information School, Wuhan University
{hanjinwei, guisong.xia}@whu.edu.cn, {sunzy12315, ethan.y.gao}@gmail.com
xavier.lin@foxmail.com, {yingguogao, kerwinyan, ericshding}@tencent.com

Abstract

We aim at finetuning a vision-language model without hurting its out-of-distribution (OOD) generalization. We address two types of OOD generalization, i.e., i) domain shift such as natural to sketch images, and ii) zero-shot capability to recognize the category that was not contained in the finetune data. Arguably, the diminished OOD generalization after finetuning stems from the excessively simplified finetuning target, which only provides the class information, such as "a photo of a [CLASS]". This is distinct from the process in that CLIP was pretrained, where there is abundant text supervision with rich semantic information. Therefore, we propose to compensate for the finetune process using auxiliary supervision with rich semantic information, which acts as anchors to preserve the OOD generalization. Specifically, two types of anchors are elaborated in our method, including i) text-compensated anchor which uses the images from the finetune set but enriches the text supervision from a pretrained captioner, ii) image-text-pair anchor which is retrieved from the dataset similar to pretraining data of CLIP according to the downstream task, associating with the original CLIP text with rich semantics. Those anchors are utilized as auxiliary semantic information to maintain the original feature space of CLIP, thereby preserving the OOD generalization capabilities. Comprehensive experiments demonstrate that our method achieves in-distribution performance akin to conventional finetuning while attaining new state-of-the-art results on domain shift and zero-shot learning benchmarks.

1. Introduction

Maintaining out-of-distribution (OOD) generalization is essential for a pretrained model to ensure applicability across
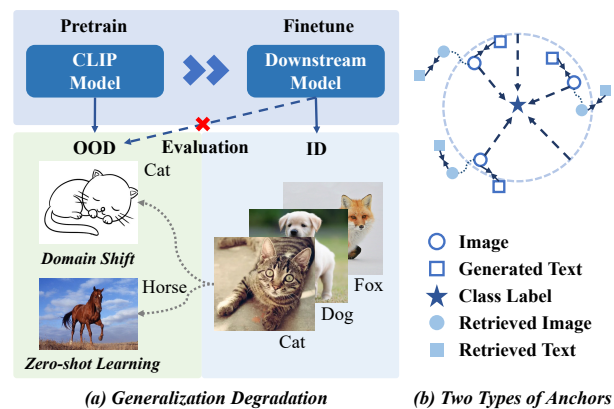

Figure 1. Motivation illustration. (a) The out-of-distribution generalization (*i.e.*, domain shift and zero-shot learning) capabilities of CLIP degrade significantly after finetuning on downstream tasks. (b) Images with generated texts and retrieved image-text pairs serve as two types of anchors to regularize the finetuning process of CLIP with auxiliary semantic information.

diverse circumstances (*e.g.*, domain shift and zero-shot learning), even after adaptation to downstream tasks. In recent years, contrastive vision-language pretrained models, such as CLIP [30] and ALIGN [17], have exhibited exceptional OOD generalization capabilities in the aforementioned situations. Although it is desirable to finetune these models using task-specific labeled data to improve performance on downstream tasks [8, 31], they often experience a significant degradation of OOD generalization [22, 39].

As illustrated in Fig. 1(a), OOD generalization encompasses domain shift, such as from natural to sketch images, and zero-shot capability to recognize categories not present in the finetuning data. Numerous methods have been proposed to preserve the OOD generalization of CLIP during the finetuning process. Prompt learning [1, 42, 43] optimizes a set of learnable vectors using a limited number of

† Corresponding authors.

labeled images while keeping the entire pretrained parameters of CLIP frozen. Although these approaches demonstrate potential in addressing zero-shot prediction, their performance on downstream tasks is insufficient for practical needs. Robust finetuning [12, 22, 36, 39] leverages all available data and employs a fully-tuned process to achieve high accuracy on in-distribution tasks without compromising performance under domain shift in OOD, while zero-shot capability remains neglected. In this study, we extend robust finetuning to a more challenging scenario, aiming to preserve OOD generalization capabilities in both domain shift and zero-shot learning. For instance, as shown in Fig. 1(a), our goal is to improve CLIP's capacity to recognize various types of animals while maintaining its original ability to classify animals from different domains (*e.g.*, sketch) and other categories of animals (*e.g.*, horse).

Our work is based on the observation that the feature space in CLIP, trained on large-scale image-text pairs, encompasses open-vocabulary semantic knowledge and thus demonstrates exceptional OOD generalization capabilities. The conventional supervised finetuning paradigm primarily minimizes a cross-entropy loss on an image classifier with class labels. FLYP [12] casts downstream class labels as text prompts (*e.g.*, "a photo of a [CLASS]") and optimizes the contrastive loss to align the image embeddings with the prompt embeddings, using only the class information.

We argue that the decline in OOD generalization stems from the semantic-scarce supervision containing only class labels during the finetuning process. Such an over-simplified finetuning target is distinct from the abundant text supervision employed in the pretraining of CLIP, leading to degraded OOD generalization. Specifically, image features originally possessing rich semantics tend to collapse into a single class center after finetuning. As a result, the feature space of the finetuned model shifts towards fitting the downstream dataset without the necessity of preserving rich semantics. Finally, as a consequence, the original feature space of CLIP deteriorates significantly, impairing its OOD generalization capabilities.

Based on the above analysis, we propose an Anchor-based Robust Finetuning (ARF) approach that regularizes the finetuning process of CLIP with auxiliary contrastive supervision. As illustrated in Fig. 1(b), our approach incorporates two types of anchors for maintaining the original feature space of CLIP. Specifically, one of them is the text-compensated anchor, which is derived from using the image from the finetune set while enriching the semantic text from a pretrained captioner, such as BLIP2 [24]. This is achieved by a carefully designed Text-Compensated Anchor Generation (TCAG) module. The other one is the image-text-pair anchor retrieved from a dataset similar to the pretraining data of CLIP according to the downstream task, whose text originally exhibits rich semantics. We implement this using

the Image-Text Anchor Retrieval (ITAR) module. These two types of image-text anchors are complementary to each other and are utilized as auxiliary supervision to regularize the finetuning process of CLIP, ensuring that the image features do not converge too close to the class prompt while retaining the original feature space of CLIP. Extensive experiments demonstrate that our ARF achieves in-distribution performance comparable to finetuning while attaining new state-of-the-art OOD generalization results on domain shift and zero-shot learning benchmarks.

Our main contributions are summarized as follows:

- We extend robust finetuning to a more challenging setting, aiming to preserve OOD generalization capabilities in both domain shift and zero-shot learning.
- We propose Anchor-based Robust Finetuning (ARF), using both text-compensated anchor and retrieved image-text-pair anchor to regularize the finetuning process.
- Extensive experiments reveal that our ARF attains new state-of-the-art performance on domain shift and zero-shot learning benchmarks while achieving in-distribution performance akin to finetuning.

2. Related Work

2.1. Vision-Language Contrastive Learning

Vision-language models [17, 30, 34] primarily focus on establishing a joint embedding space for cross-modal learning by aligning web-scale images and texts. Recent advancements in the contrastive vision-language pretraining paradigm, particularly CLIP [30], have demonstrated exceptional out-of-distribution (OOD) generalization capabilities across various downstream tasks (*e.g.*, domain shift and zero-shot learning). Numerous subsequent studies have been proposed to further enhance contrastive image-text pretraining, such as masked language/image modeling [9, 25], hard sample mining [38], and retrieval augmentation [16, 41]. BLIP [23] aims to improve image-text pretraining for unified vision-language understanding and generation, while BLIP2 [24] incorporates Large Language Models (LLMs) to further improve generalization. Although these methods have indeed enhanced vision-language pretraining, finetuning remains necessary to improve performance on downstream tasks. A recent study [8] demonstrates that CLIP attains superior or at least competitive performance on downstream tasks after finetuning, compared to conventional supervised pretraining models. ViFi-CLIP [31] implicitly models temporal cues and effectively finetunes image-level CLIP representations for videos. However, the OOD generalization capabilities significantly degrade after finetuning, impairing the applicability of CLIP across diverse circumstances.

2.2. Finetuning for Generalization

Maintaining the OOD generalization capabilities of CLIP during the finetuning process has been extensively investigated through various approaches. Prompt learning [19, 42, 43] incorporates a small number of learnable prompt vectors to finetune CLIP, using a limited set of labeled images while keeping the pretrained model weights fixed. Although these methods demonstrate potential in addressing zero-shot prediction, their performance on downstream tasks remains unsatisfactory. Robust finetuning [12, 22, 36, 39] leverages all available data and implements a fully-tuned process to achieve high accuracy on in-distribution tasks without sacrificing performance under domain shift in OOD. Wise-FT [39] ensembles the weights of finetuned and original models, yielding remarkable gains in domain shift. LP-FT [22] initially trains the classification layer and then finetunes the entire network. This two-stage process substantially mitigates distortion of pretrained features and improves generalization. FLYP [12] demonstrates that a straightforward method of mimicking contrastive pretraining consistently outperforms finetuning approaches, while TPGM [36] automatically learns the constraint imposed on each layer for finetuning regularization. Nonetheless, the zero-shot learning capability is neglected by robust finetuning, leading to a substantial degradation of OOD generalization. In light of these observations, we propose a more practical and challenging setting to finetune CLIP on downstream tasks without compromising both domain shift and zero-shot learning generalization capabilities.

3. Preliminaries

3.1. Summary of Notations

We summarize the notations used in the following sections. Specifically, we use the uppercase calligraphy font to denote a specific set, *e.g.*, a dataset $\mathcal{S}$, which consists of input images and their corresponding class information. The input images are denoted by $\mathcal{X}$, while the class information is either represented by the class labels $\mathcal{Y}$, or the text descriptions $\mathcal{T}$ encompassing the class information, *e.g.*, $\mathcal{S} = \{\mathcal{X}, \mathcal{Y}\}$ or $\mathcal{S} = \{\mathcal{X}, \mathcal{T}\}$. We denote the set of all the classes by $\mathcal{C}$. We use the lowercase font to denote a sample, *e.g.*, x is an image sample.

As for the scripts, we use superscript to indicate a specific data split, which can be a split of training (train), testing (test), in-distribution (id), domain shift (ds), zero-shot learning (zsl), etc. And the subscripts serve as indices. For example, x_i^{train} is the i-th image sample from the training dataset where $x_i^{\text{train}} \in \mathcal{X}^{\text{train}}$.

3.2. Contrastive Vision-Language Models

In this paper, we adapt the contrastive vision-language pretraining model CLIP [30] to downstream tasks, which com-

prises an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$. During the pretraining phase, image-text pairs are sampled from a web-scale training dataset $\mathcal{S}^{\text{train}} = \{(x_i^{\text{train}}, t_i^{\text{train}})\}$, where each text description t_i^{train} encompasses rich semantics with abundant class information. The categories $\mathcal{C}^{\text{train}}$ cover open-vocabulary semantic knowledge. Each image x_i^{train} and text t_i^{train} is mapped to an image embedding $f(x_i^{\text{train}})$ and a text embedding $g(t_i^{\text{train}})$, respectively. Subsequently, a contrastive loss function $\mathcal{L}_{CL}$ is utilized to align the image embedding $f(x_i^{\text{train}})$ with the corresponding text embedding $g(t_i^{\text{train}})$, which is formulated as,

$$
\begin{aligned}
\mathcal{L}_{CL} = &-\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(f(x_i^{\text{train}}) \cdot g(t_i^{\text{train}})/\tau)}{\sum_{j=1}^{B} \exp(f(x_j^{\text{train}}) \cdot g(t_j^{\text{train}})/\tau)} \\
&-\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(g(t_i^{\text{train}}) \cdot f(x_i^{\text{train}})/\tau)}{\sum_{j=1}^{B} \exp(g(t_j^{\text{train}}) \cdot f(x_j^{\text{train}})/\tau)},
\end{aligned}
\tag{1}
$$

where B represents the number of image-text pairs in the minibatch, and τ denotes the temperature parameter used to scale the pairwise similarities in the loss function.

We consider image recognition as evaluation, that is, we evaluate the performance on a test set $\mathcal{S}^{\text{test}} = \{(x_i^{\text{test}}, y_i^{\text{test}})\}$ with categories $\mathcal{C}^{\text{test}}$. CLIP utilizes predefined text prompts t_c^{test} as inputs for the text encoder, which describe each class $c \in \mathcal{C}^{\text{test}}$, such as,

$$
\text{text prompt: } a \; photo \; of \; a \; [\text{CLASS}], \tag{2}
$$

where "$[\text{CLASS}]$" represents the class name. The output text embeddings $g(t^{\text{test}})$ are employed as classification weights during evaluation. Given an image x_i^{test} in $\mathcal{S}^{\text{test}}$, it is fed into the image encoder of CLIP to obtain the corresponding image embedding $f(x_i^{\text{test}})$. The predicted label p_i^{test} is calculated as,

$$
p_i^{\text{test}} = \arg \max_{c \in C^{\text{test}}} (f(x_i^{\text{test}}) \cdot g(t_c^{\text{test}})), \tag{3}
$$

which implies that the class exhibiting the highest similarity between the image embedding $f(x_i^{\text{test}})$ and text embeddings $g(t_c^{\text{test}})$ of $\mathcal{C}^{\text{test}}$ classes is selected as the classification result.

We finetune the CLIP model by mimicking contrastive pretraining, as described in FLYP [12]. Specifically, class labels are formulated as text prompts in Eq. (2) and the contrastive loss function in Eq. (1) is utilized to align image embeddings with text prompt embeddings. We follow the same evaluation process as used in CLIP [30].

4. Method

4.1. Problem Setup

Given a pretrained CLIP model, we aim to finetune it on the in-distribution dataset $\mathcal{S}^{\text{id}} = \{(x_i^{\text{id}}, y_i^{\text{id}})\}$ sampled from the distribution P^{id} with classes $\mathcal{C}^{\text{id}}$, where each image x_i^{id} has a
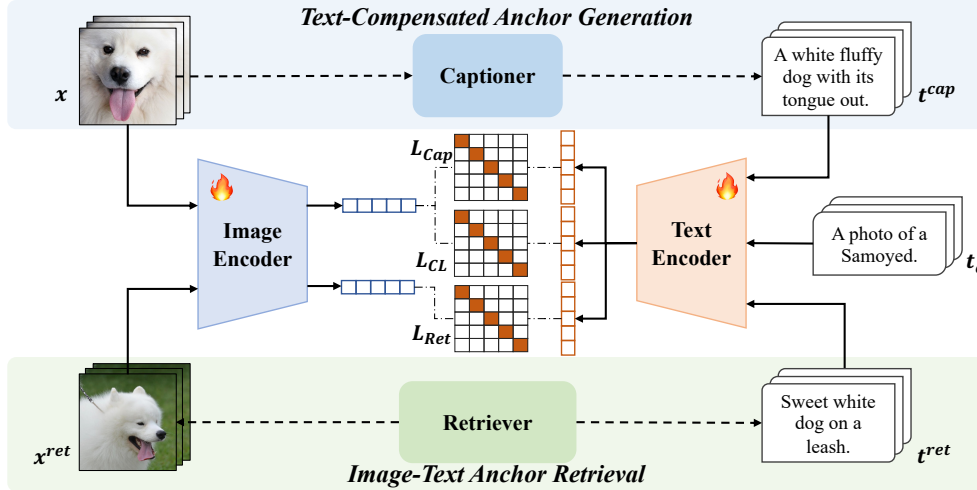
Figure 2. The pipeline of our proposed Anchor-based Robust Finetuning (ARF) comprises a Text-Compensated Anchor Generation (TCAG) module and an Image-Text Anchor Retrieval (ITAR) module. TCAG generates a caption for each image in the finetuning dataset utilizing a pretrained captioner as a text-compensated anchor with rich semantics. ITAR searches for image-text pairs from a candidate set similar to the data on which CLIP was pretrained, ensuring the presence of rich semantics in the image-text-pair anchor. We retrieve those samples related to our downstream tasks. A contrastive loss function, as used in CLIP, is employed for image-text alignment.

label $y_i^{\text{id}} \in \mathcal{Y}^{\text{id}}$. The adapted model should perform at least as well as conventional finetuning methods on the test set from the same distribution P^{id} and with the same categories $\mathcal{C}^{\text{id}}$ as the training data. Simultaneously, we also work towards preserving the out-of-distribution (OOD) generalization capabilities in both domain shift and zero-shot learning scenarios. In the domain shift situation, we evaluate the performance on the domain shift dataset $\mathcal{S}^{\text{ds}} = \{(x_i^{\text{ds}}, y_i^{\text{ds}})\}$. The test data are sampled from a different domain P^{ds} but share the same categories as the in-distribution data where we have $P(\mathcal{X}^{\text{id}}) \neq P(\mathcal{X}^{\text{ds}})$ while $P(\mathcal{Y}|\mathcal{X}^{\text{id}}) = P(\mathcal{Y}|\mathcal{X}^{\text{ds}})$. As for zero-shot learning, we have the downstream dataset as $\mathcal{S}^{\text{zsl}} = \{(x_i^{\text{zsl}}, y_i^{\text{zsl}})\}$, which composes of test image x_i^{zsl} from a different category $y_i^{\text{zsl}} \in \mathcal{Y}^{\text{zsl}}$ where $\mathcal{C}^{\text{zsl}} \cap \mathcal{C}^{\text{id}} = \emptyset$.

In other words, as illustrated in Fig. 1(a), consider finetuning on real cat images, if we use sketch cat images for testing, then it is a domain shift problem; while if we test on horse images whose category is not included in the finetune set, then it is a zero-shot learning situation. We aim to achieve high performance on the in-distribution testing set, while also preserving the OOD generalization for both domain shift and zero-shot learning scenarios.

4.2. Anchor-based Robust Finetuning

Overview. As shown in Fig. 2, our Anchor-based Robust Finetuning (ARF) approach finetunes both the image encoder and text encoder of CLIP using contrastive loss and incorporates two distinct modules to regularize the finetuning process. Specifically, *in the Text-Compensated Anchor Generation (TCAG) module*, we leverage a pretrained im-

age captioner to generate a caption as text-compensated information for each image and align them using contrastive loss. *In the Image-Text Anchor Retrieval (ITAR) module*, we search for image-text pairs from a dataset similar to CLIP's pretraining data, which are related to the downstream task. These samples, originally containing rich semantic information, serve as additional anchors during finetuning. These two types of anchors complement each other and preserve the original feature space of CLIP to ensure the OOD generalization after adaptation.

4.2.1 Text-Compensated Anchor Generation

It is essential to note that merely employing the contrastive training loss between images x and class prompts t_c in Eq. (1) can lead to overfitting the excessively simplified finetuning target, as class prompts t_c contain only class information shown in Eq. (2). This semantic-scarce supervision is distinct from the abundant text supervision exploited in the pretraining of CLIP and can result in the degradation of the original feature space.

Caption Generation. To mitigate the aforementioned issue, we propose to employ a pretrained image captioner, such as BLIP2 [24], to generate a text description t_i^{cap} (*i.e.*, caption) as compensated rich-semantic information for each image x_i, thereby preventing the overfitting on class prompts t_c. The generated captions t^{cap} encompass various descriptive words with more abundant semantics compared to class prompts t_c, similar to the texts utilized in CLIP's pretraining. The images x and corresponding captions t^{cap} form the text-compensated anchors for regularization.
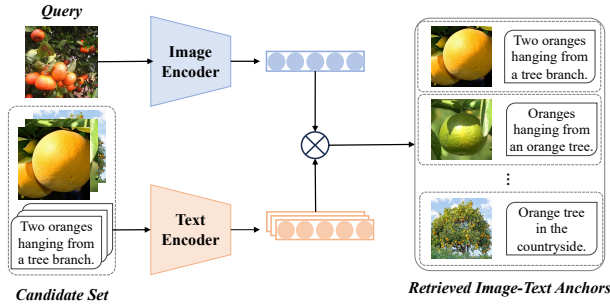
Figure 3. The pipeline of our Image-Text Anchor Retrieval (ITAR) module. We search for the most similar image-text pairs in the candidate set to obtain the rich semantic image-text anchors related to the downstream task for regularizing the finetuning process.

Image-Caption Contrastive Learning. We cast (x, t^{cap}) as text-compensated anchors and utilize a contrastive loss function $\mathcal{L}_{\text{Cap}}$ similar to Eq. (1) to align them within the feature space, thereby maintaining the semantic consistency between images and texts. The text-compensated anchors prevent overfitting by ensuring that the conventional finetuning process in Eqs. (1) and (2) does not pull the embedding of the image x_i too close to the text embedding of its corresponding class prompt t_c, regularizing the finetuning process of CLIP with auxiliary semantic supervision.

4.2.2 Image-Text Anchor Retrieval

Furthermore, we propose to search for image-text pairs according to the downstream task as auxiliary anchors. We construct a candidate set that is similar to the pretraining data of CLIP, originally containing rich semantic information. The pretrained CLIP model possesses an exceptional cross-modal retrieval capacity, which we employ to find image-text pairs with abundant semantics for regularization. These retrieved image-text-pair anchors are aligned using a contrastive loss function to preserve the original feature space during the finetuning process.

Candidate Set Construction. We construct the web-scale image-text dataset CC3M [33] as a candidate set $\mathcal{S}^{\text{can}} = \{(x_i^{\text{can}}, t_i^{\text{can}})\}$, which is employed to search for rich-semantic image-text pairs. The candidate set closely resembles the pretraining data of CLIP and encompasses abundant semantic information for maintaining the original feature space. We leverage the cross-modal retrieval capacity of CLIP to obtain image-text-pair anchors relevant to the downstream task. Specifically, we extract the embeddings of the images x in the downstream dataset as $f(x)$, as well as the embeddings of the texts t^{can} in the candidate set as $g(t^{\text{can}})$ for preparation. These embeddings are utilized for retrieval and can be precomputed offline.

Image-Text Pair Retrieval. In practice, only a small subset of the candidate set $\mathcal{S}^{\text{can}}$ is relevant to the downstream task and can be employed to preserve the original feature space of CLIP. We propose to search for image-text pairs from the candidate set $\mathcal{S}^{\text{can}}$ using KNN search.

Specifically, as illustrated in Fig. 3, we designate each image x_i as the query and find the most similar image-text pairs from the candidate set $\mathcal{S}^{\text{can}}$ by calculating the similarity between the image embedding $f(x_i)$ and the text embeddings $g(t^{\text{can}})$, which can be formulated as follows,

$$k = \arg\max(f(x_i) \cdot g(t^{\text{can}})), \qquad (4)$$

where k represents the index of retrieved image-text pair in the candidate set $\mathcal{S}^{\text{can}}$.

The retrieval process can be efficiently executed using existing libraries, such as Faiss [18]. Subsequently, we utilize the retrieved image-text pairs $\mathcal{S}^{ret} = \{(x_k^{ret}, t_k^{ret})\}$ with rich semantics as auxiliary anchors to regularize the finetuning process.

Image-Text Contrastive Learning. We denote (x^{ret}, t^{ret}) as retrieved image-text-pair anchors and employ a contrastive loss function $\mathcal{L}_{Ret}$, similar to Eq. (1), to preserve the original feature space of CLIP. The retrieved image-text pairs exhibit rich semantics related to the downstream task and serve as auxiliary supervision during finetuning.

These two types of image-text anchors with abundant semantic information are complementary to each other and are utilized as additional contrastive supervision to regularize the finetuning process of CLIP. The image encoder and text encoder of CLIP are finetuned together with the following loss function,

$$\mathcal{L} = \mathcal{L}_{CL} + \mathcal{L}_{Cap} + \mathcal{L}_{Ret}. \qquad (5)$$

5. Experiments

Overview. We assess the effectiveness of our proposed Anchor-based Robust Finetuning (ARF) approach by comparing it with several baselines and providing implementation details for reproducibility. The evaluation of maintaining out-of-distribution (OOD) generalization capabilities is divided into two sections. In Section 5.1, we present results for domain shift, which was the focus of the original robust finetuning methods. Subsequently, in Section 5.2, we show results for our extended scenario (*i.e.*, zero-shot learning). Furthermore, we conduct an ablation study in Section 5.3 to evaluate the efficacy of our approach and showcase qualitative examples of two types of anchors in Section 5.4.

Baselines. We compare our ARF with two conventional ways of finetuning pretrained models to downstream tasks using cross-entropy: linear probing (LP) and end-to-end fully finetuning (FT). Additionally, we investigate recent advancements of robust finetuning, such as LP-FT [22],

Methods	ImageNet							DomainNet					
	ID	Im-V2	Im-R	Im-A	Im-Sketch	ObjectNet	Avg. OOD	ID	Sketch	Painting	Infograph	Clipart	Avg. OOD
CLIP	68.3	61.9	77.7	50.0	48.3	54.2	58.4	84.8	65.7	68.5	50.2	72.1	64.1
LP	79.9	69.8	70.8	46.4	46.9	50.4	56.9	86.3	57.4	61.5	45.6	64.1	57.2
FT	81.3	71.2	66.1	37.8	46.1	51.6	54.6	89.5	61.8	65.6	49.0	71.7	62.1
LP-FT	81.7	72.1	73.5	47.6	50.3	54.4	59.6	89.5	63.6	67.4	50.7	73.4	63.8
FLYP	82.6	**73.0**	71.4	48.1	49.6	54.7	59.4	**89.8**	64.1	68.5	50.8	74.0	64.3
ARF	**82.7**	72.8	**75.6**	**50.3**	**51.8**	**55.8**	**61.3**	**89.8**	65.3	69.5	**51.1**	74.9	**65.2**

Table 1. Domain shift results (%) of state-of-the-art conventional finetuning and robust finetuning approaches on ImageNet and DomainNet benchmarks. The numbers represent the top-1 accuracy. We employ ImageNet and DomianNet-Real as the finetuning datasets, while the others serve as domain shift evaluation datasets. The best results are marked in **Black**.
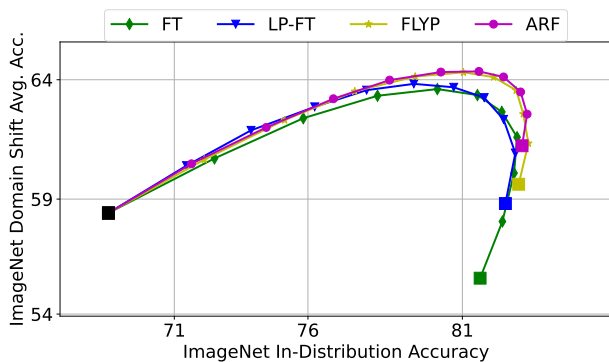

Figure 4. The ID and domain shift performance of our ARF compared with several baselines through linear interpolation of the finetuned model weights with the original model weights following Wise-FT [39]. The performance curves of ARF surpass (positioned in the upper right) those of the baselines on ImageNet, resulting in improved ID and domain shift accuracy.

which involves an initial linear probing followed by fully finetuning, and FLYP [12], wherein finetuning is conducted in a pretraining-like manner.

Implementation Details. We employ a batch size of 512 for finetuning on ImageNet [7] and DomainNet [29] with 10 epochs. We utilize a learning rate of 10^{-5} and a weight decay parameter of 0.1. ViT-B/16 [10] is utilized as the image encoder of CLIP for finetuning. Domain shift and zero-shot learning benchmarks are exclusively used for evaluation. We apply the same prompt templates as those employed in CLIP [30] and WiseFT [39] for training and inference.

5.1. Evaluation Under Domain Shift

Benchmarks. We evaluate domain shift performance on two widely employed benchmarks, namely, ImageNet [7] and DomainNet [29]. FMoW [20] and iWILDCam [3] are excluded from our evaluation due to the poor domain shift performance of the original CLIP. In the first benchmark, we finetune CLIP using ImageNet [7] for in-distribution

evaluation and assess its performance on five distinct variants of ImageNet with domain shift: ImageNet-V2 [32], ImageNet-Sketch [37], ImageNet-A [15], ImageNet-R [14] and ObjectNet [2]. We follow the training protocol outlined in FLYP [12] and employ the same ImageNet prompt templates used by CLIP during inference. For the second benchmark, we utilize the standard domain shift dataset DomainNet [29] for evaluation. We finetune CLIP using DomainNet-Real for in-distribution performance and assess its generalization ability to four domain shift splits: Clipart, Infograph, Painting, and Sketch.

Quantitative Results. We compare the performance of our ARF with several baselines on domain shift benchmarks, as detailed in Table 1. On the ImageNet dataset, our ARF exhibits a slight performance advantage over conventional finetuning methods and other robust finetuning methods in the in-distribution (ID) test dataset. The true strength of our ARF lies in its capacity to generalize to domain shift scenarios, achieving state-of-the-art performance with an average accuracy of 61.3% across five domain shift test datasets. Notably, our ARF performs significantly better than other finetuning methods, approaching or even surpassing the performance of CLIP on ImageNet-R, ImageNet-A, ImageNet-Sketch, and ObjectNet, which have large domain differences. In the case of DomainNet, we also observe that our ARF demonstrates state-of-the-art performance of 65.2% on the domain shift test dataset without sacrificing accuracy on in-distribution (ID) data. Our ARF outperforms CLIP on domain shift scenarios by 1.1%, whereas other finetuning methods either fail to surpass it or achieve only marginal improvements. These results indicate the effectiveness of the two types of anchors in our ARF. They effectively prevent the overfitting of the original feature space of CLIP to the downstream class prompts after finetuning, preserving the OOD generalization capability of CLIP for handling domain shift scenarios.

Weight Ensembling Curves. Wise-FT [39] demonstrates that a simple linear interpolation between the weights of the pretrained and finetuned models yields the optimal performance for both ID and domain shift. Therefore, we com-

	Zero-Shot Learning										
Methods	ImageNet	Caltech	Flowers	Food	SUN	DTD	Aircraft	Cars	Pets	EuroSAT	UCF	Avg. OOD
CLIP	68.3	89.3	70.4	89.2	65.2	46.0	27.1	65.6	88.9	54.1	69.8	66.6
FT	81.3	78.8	16.0	37.3	39.3	29.7	4.7	10.8	80.2	15.4	44.3	35.7
LP-FT	81.7	84.0	44.3	68.8	49.9	37.9	**15.8**	37.7	81.9	30.4	59.5	51.0
FLYP	82.6	87.6	36.8	62.8	52.0	36.9	8.7	31.1	77.6	34.3	58.6	48.6
ARF	**82.7**	**88.6**	**46.4**	**74.5**	**63.8**	**40.4**	13.9	**44.7**	**83.1**	35.8	**64.6**	**55.6**

Table 2. Zero-shot learning results (%) of state-of-the-art conventional finetuning and robust finetuning methods on numerous recognition tasks. The numbers represent the top-1 accuracy. We employ ImageNet as the finetuning dataset while the others serve as zero-shot learning evaluation datasets. The best results are marked in **Black**.

Method		ImageNet		Zero-Shot
TCAG	ITAR	ID	Domain Shift	Avg. Acc
baseline		82.6	59.4	48.6
✓		82.6	60.7(+1.3)	54.3(+5.7)
	✓	82.6	60.2(+0.8)	53.6(+5.0)
✓	✓	82.7	61.3(+1.9)	55.6(+7.0)

Table 3. Ablation study for Text-Compensated Anchor Generation (TCAG) module and Image-Text Anchor Retrieval (ITAR) module of our ARF. The baseline only conducts visual-language contrastive learning for finetuning like FLYP [12].

pare the performance of our ARF with the baselines by interpolating their model weights using 10 mixing coefficients ranging from 0 to 1. As depicted in Fig. 4, we can see that our ARF outperforms the baselines after finetuning on ImageNet, leading to enhanced ID and domain shift accuracy. Concretely, when comparing the coefficient that achieves the highest ID performance, our ARF with weight ensembling improves domain shift accuracy by 1.2% over the state-of-the-art method (*i.e.*, FLYP [12]). Comprehensive results are provided in the supplementary material.

5.2. Evaluation Under Zero-shot Learning

Benchmarks. We evaluate zero-shot learning performance using a diverse benchmark that encompasses a range of recognition tasks. To ensure a fair comparison, we employ the standard test split for inference, and the CLIP model is finetuned on ImageNet [7]. As for the evaluation of zero-shot learning, we utilize fine-grained object classification tasks such as OxfordPets [28], StanfordCars [21], Flowers102 [27] and Food101 [4]; as well as specific recognition tasks such as UCF101 [35] for action recognition, FGV-CAircraft [26] for aircraft classification, DTD [6] for texture classification, SUN397 [40] for scene recognition and EuroSAT [13] for satellite image classification. Additionally, we also evaluate our ARF on the general object classification dataset Caltech101 [11].

Quantitative Results. We present the results of our ARF and several baseline approaches on various zero-shot learning recognition tasks, as displayed in Table 2. It can be observed that previous finetuning methods exhibit substantial improvements in accuracy on the ImageNet test data after finetuning with ImageNet training data, compared to the original CLIP. However, their performance significantly deteriorates in zero-shot learning recognition on the categories that were not contained in the finetuning data. In contrast, our ARF optimally maintains the zero-shot recognition capability without compromising the performance on ImageNet. The experimental results demonstrate that our ARF effectively regularizes the finetuning process with auxiliary semantic supervision, preserving the OOD generalization capability of CLIP for handling zero-shot learning scenarios.

5.3. Ablation Study

Two Types of Anchors. To evaluate the effectiveness of our ARF, we conduct an ablation study to analyze the influence of the two types of anchors, as shown in Table 3. It can be observed that the Text-Compensated Anchor Generation (TCAG) module significantly improves domain shift and zero-shot learning accuracy by 1.3% and 5.7% over the baseline, respectively. These gains demonstrate that the rich semantic text descriptions generated by the pretrained captioner provide effective auxiliary supervision for the images to alleviate overfitting on class prompts. For retrieved image-text-pair anchors, we evaluate the influence of the Image-Text Anchor Retrieval (ITAR) module, which can improve domain shift and zero-shot learning accuracy over the baseline by 0.8% and 5.0% after finetuning on ImageNet, respectively. These results reveal that the rich semantic image-text pairs, retrieved from the candidate set similar to the pretraining data of CLIP according to the downstream task, are beneficial for regularizing the finetuning process. The two modules, working together, boost performance over the baseline by 1.9% on domain shift and 7.0% on zero-shot learning. The experimental results sug-

Task Image with Label	Generated Caption	Retrieved Image-Text Pair
Meerkat	A meerkat standing on a dirty road.	A pregnant meerkat resting in the sands.
Church	A blue dome on a church.	Blue-domed church of person in town.
Violin	A man playing a violin in a black shirt.	Studio shot of a classical violinist playing a violin on black background.

Figure 5. Visualization examples of captions generated by a pretrained captioner (*e.g.*, BLIP2 [24]) and retrieved image-text pairs relative to the downstream task from the CC3M dataset [33]. The downstream task images and generated captions serve as text-compensated anchors for regularization. The retrieved image-text pairs function as auxiliary anchors for maintaining the feature space.

| Captioner | ImageNet | | Zero-Shot |
	ID	Domain Shift	Avg. Acc
BLIP [23]	82.1	60.1	54.0
BLIP2 [24]	82.7	61.3	55.6
BLIP2 + Vicuna [5]	82.5	61.2	56.5

Table 4. Ablation study for the quality of generated captions. We evaluate the effectiveness of pretrained image captioners (*i.e.*, BLIP [23] and BLIP2 [24]) and further rewrite the text descriptions using Large Language Models (*e.g.*, Vicuna [5]) in our ARF.

gest that these two types of anchors complement each other and are beneficial to preserving the OOD generalization capabilities of CLIP.

The Quality of Captions. To assess the impact of caption quality in our ARF, we examine two pretrained image captioners (*i.e.*, BLIP [23] and BLIP2 [24]) and further rewrite the text descriptions using Large Language Models (*e.g.*, Vicuna [5]). As illustrated in Table 4, employing captions generated by BLIP2 results in a 1.2% improvement in domain shift performance and a 1.6% enhancement in zero-shot learning performance compared to using captions generated by BLIP. These gains demonstrate that more accurate text descriptions with rich semantics are beneficial for regularizing the finetuning process. Additionally, we utilize Vicuna [5] to rewrite the captions for increased diversity and richer semantic information. Since the text descriptions generated by BLIP2 are already sufficiently accurate, there is no improvement in ID and domain shift scenarios. However, the rich semantic knowledge accessed from Vicuna boosts zero-shot learning by 0.9%. These results indicate the effectiveness of auxiliary information from LLMs, which warrants further exploration.

5.4. Qualitative Examples of Anchors

In Fig. 5, we provide visualization examples to facilitate an understanding of how our Anchor-based Robust Finetuning (ARF) works. The text descriptions (*i.e.*, captions) generated by a pretrained captioner (*e.g.*, BLIP2 [24]) accurately describe the images, thus serving as effective anchors that supply rich information for maintaining the semantic consistency between images and texts. The image-text pairs, retrieved from the candidate set, bear a close resemblance to the pretraining data of CLIP and are related to the downstream task. This contributes auxiliary semantic knowledge to preserve the OOD generalization capabilities of CLIP.

6. Conclusion

In this study, we extend previous robust finetuning to a more challenging setting: preserving out-of-distribution (OOD) generalization capabilities in both domain shift and zero-shot learning during finetuning. We argue that the diminished OOD generalization results from the overly simplified finetuning target, which provides only class information. Consequently, we propose an Anchor-based Robust Finetuning (ARF) approach to regularize the finetuning process with auxiliary contrastive supervision. This approach incorporates a Text-Compensated Anchor Generation module and an Image-Text Anchor Retrieval module to generate image-text-pair anchors with rich semantic information and align these anchors with contrastive loss. Extensive experiments demonstrate the effectiveness of our approach.

Acknowledgement

References

[1] Shuanghao Bai, Min Zhang, Wanqi Zhou, Siteng Huang, Zhirong Luan, Donglin Wang, and Badong Chen. Prompt-based distribution alignment for unsupervised domain adaptation. *arXiv:2312.09553*, 2023. 1

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, 2019. 6

[3] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv:2004.10340*, 2020. 6

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV*, 2014. 7

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *https://vicuna.lmsys.org*, 2023. 8

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 7

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6, 7

[8] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. CLIP itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. *arXiv:2212.06138*, 2022. 1, 2

[9] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, 2023. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, 2004. 7

[12] 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. 2, 3, 6, 7

[13] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *Remote Sensing*, 2019. 7

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 6

[15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 6

[16] Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Retrieval-enhanced contrastive vision-text models. *arXiv:2306.07196*, 2023. 2

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2

[18] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 2021. 5

[19] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *CVPR*, 2023. 3

[20] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021. 6

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013. 7

[22] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pre-trained features and underperform out-of-distribution. In *ICLR*, 2022. 1, 2, 3, 5

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2, 8

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*, 2023. 2, 4, 8

[25] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. *CVPR*, 2023. 2

[26] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 7

[27] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 7

[28] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. 7

[29] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 6

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 6

[31] Hanoona Abdul Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned CLIP models are efficient video learners. *CVPR*, 2023. 1, 2

[32] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 6

[33] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 5, 8

[34] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022. 2

[35] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. 7

[36] 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. 2, 3

[37] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 6

[38] Haonan Wang, Minbin Huang, Runhui Huang, Lanqing Hong, Hang Xu, Tianyang Hu, Xiaodan Liang, and Zhenguo Li. Boosting visual-language models by exploiting hard samples. *arXiv:2305.05208*, 2023. 2

[39] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 1, 2, 3, 6

[40] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 7

[41] Chenwei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *CVPR*, 2023. 2

[42] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1, 3

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 1, 3