Hint-Aug: Drawing Hints from Foundation Vision Transformers towards Boosted Few-shot Parameter-Efficient Tuning

Zhongzhi Yu[1], Shang Wu[2], Yonggan Fu[1], Shunyao Zhang[2], Yingyan (Celine) Lin[1]
[1]Georgia Institute of Technology [2]Rice University
{zyu401, yfu314, celine.lin}@gatech.edu {sw99,sz74}@rice.edu

Abstract

Despite the growing demand for tuning foundation vision transformers (FViTs) on downstream tasks, fully unleashing FViTs' potential under data-limited scenarios (e.g., few-shot tuning) remains a challenge due to FViTs' data-hungry nature. Common data augmentation techniques fall short in this context due to the limited features contained in the few-shot tuning data. To tackle this challenge, we first identify an opportunity for FViTs in few-shot tuning: pretrained FViTs themselves have already learned highly representative features from large-scale pretraining data, which are fully preserved during widely used parameter-efficient tuning. We thus hypothesize that leveraging those learned features to augment the tuning data can boost the effectiveness of few-shot FViT tuning. To this end, we propose a framework called ***Hint****-based Data* ***Aug****mentation (****Hint-Aug****), which aims to boost FViT in few-shot tuning by augmenting the over-fitted parts of tuning samples with the learned features of pretrained FViTs. Specifically, Hint-Aug integrates two key enablers: (1) an* ***A****ttentive* ***O****ver-fitting* ***D****etector (****AOD****) to detect over-confident patches of foundation ViTs for potentially alleviating their over-fitting on the few-shot tuning data and (2) a* ***C****onfusion-based* ***F****eature* ***I****nfusion (****CFI****) module to infuse easy-to-confuse features from the pretrained FViTs with the over-confident patches detected by the above AOD in order to enhance the feature diversity during tuning. Extensive experiments and ablation studies on five datasets and three parameter-efficient tuning techniques consistently validate Hint-Aug's effectiveness:* $0.04\% \sim 32.91\%$ *higher accuracy over the state-of-the-art (SOTA) data augmentation method under various low-shot settings. For example, on the Pet dataset, Hint-Aug achieves a 2.22% higher accuracy with 50% less training data over SOTA data augmentation methods.*

1. Introduction

Foundation vision transformers (FViTs) [16, 41, 54, 55, 64] with billions of floating point operations (FLOPs) and parameters have recently demonstrated significant potential in various downstream tasks [40, 41]. The success of FViTs has ushered in a new paradigm in deep learning: pretraining-then-tuning [16, 40, 67], which first pretrains an FViT on a large-scale dataset, then uses recently developed parameter-efficient tuning methods (e.g., visual prompt tuning (VPT) [34], visual prompting [2], LoRA [33], and Adapter [72]) to tune pretrained FViTs on downstream tasks with limited tuning data. However, although it is highly desirable, effectively tuning pretrained FViTs for real-world applications, especially under few-shot tuning scenarios, remains a particularly challenging task. The reason is that although parameter-efficient tuning methods are dedicatedly designed for FViTs and can alleviate the overfitting issue by reducing the number of trainable parameters [2, 34, 72], the data-hungry nature of FViTs [16, 54] is not mitigated and thus the achievable accuracy under data-limited scenarios (e.g., few-shot tuning scenarios) are still limited. Therefore, how to effectively tune pretrained FViTs on various downstream tasks with few-shot tuning is still an open question.

To enhance the effectiveness of parameter-efficient FViT tuning under few-shot settings, one promising direction is to leverage data augmentation techniques to increase the data diversity and thus the feature diversity of the models when being tuned on few-shot data, boosting the achievable accuracy [12, 31, 68, 71]. Nevertheless, it has been shown that existing data augmentation techniques fall short in boosting the model accuracy under few-shot tuning scenarios. This is because most of the existing data augmentation techniques are random-based (e.g., RandAugment [13], AutoAugment [12], color jitter, mixup [71], and cutmix [68]), which only randomly permute existing features in the training data and thus cannot generate new and meaningful features [63]. As illustrated in Fig. 1, we observe that neither the widely-used random-based data augmentation techniques (i.e., a dedicated combination of techniques including RandAugment [13], color jitter, and random erasing [74] as in [72]) nor training without data augmentation can consistently achieve a satisfactory accuracy across dif-

[†]Our code is available at https://github.com/GATECH-EIC/Hint-Aug

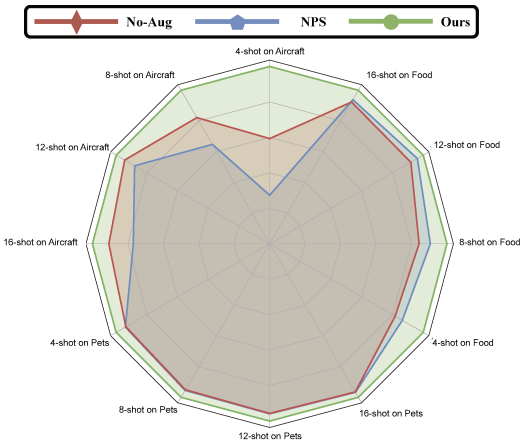

Figure 1. The normalized achieved accuracies when few-shot tuning the ViT-base model [16] on various datasets and numbers of tuning shots using (1) vanilla training without augmentation (i.e., No-Aug), (2) the SOTA parameter-efficient tuning technique [72] (i.e., NPS), and (3) our proposed Hint-Aug.

ferent datasets under few-shot tuning. Specifically, when being applied to fine-grained classification tasks, e.g., the Aircraft dataset [45], these random-based data augmentation techniques actually hurt the achievable accuracy. The reason is that random-based data augmentation techniques can easily create out-of-manifold samples [68, 71], especially on commonly used fine-grained datasets. Such out-of-manifold samples can largely degrade the achievable accuracy given the limited number of training samples under few-shot tuning scenarios [27]. Therefore, it is crucial to develop data augmentation techniques that can adaptively augment the given training samples with diverse, but still within-manifold, features to boost the effectiveness of tuning FViTs on various downstream tasks.

This work sets out to close the increasing gap between the growing demand for effective few-shot FViT tuning and the unsatisfactory achievable accuracy by existing techniques. In particular, we identify that in few-shot parameter-efficient tuning, the pretrained FViTs' weights are fixed during tuning. Meanwhile, existing works have shown that (1) pretrained transformer models have already learned complex but generalizable features [16, 41, 54] and (2) gradient-based methods can extract the learned features from pretrained models and then add them to the input images [44, 46]. Therefore, we hypothesize that FViTs' few-shot tuning accuracies can be non-trivially improved by leveraging the learned features in the pretrained FViTs. Specifically, we make the following contributions:

- We propose a framework called **Hint**-based Data **Aug**mentation (**Hint-Aug**), which is dedicated to boosting the achievable accuracy of FViTs under few-shot tuning scenarios by leveraging the learned features of pretrained FViTs to guide the data augmentation strategy used for the training dataset in an input-adaptive manner.

- Our Hint-Aug framework integrates two enablers: (1) an **A**ttentive **O**ver-fitting **D**etector (AOD) to identify the over-fitting samples and patches in the given training dataset by making use of the attention maps of pretrained FViTs and (2) a **C**onfusion-based **F**eature **I**nfusion (CFI) module to adaptively infuse pretrained FViTs' learned features into the training data to better tuning those models on downstream tasks, alleviating the commonly recognized challenge of having limited features under few-shot tuning.

- Extensive experiments and ablation studies on five datasets and three parameter-efficient tuning techniques consistently validate the effectiveness of our proposed Hint-Aug framework, which achieves a $0.04\% \sim 32.91\%$ higher accuracy over state-of-the-art (SOTA) data augmentation methods [72] across different datasets and few-shot settings. For example, on the Pets dataset, Hint-Aug achieves a 2.22% higher accuracy with 50% less training data compared with the SOTA augmentation method.

2. Related Works

2.1. FViTs

Inspired by the recent success of vision transformers (ViTs), one of the most notable directions in ViTs is to scale up ViTs' model size to build FViTs, aiming to replicate the success of large-scale neural language processing foundation models [15, 19, 50] in the field of computer vision [16, 40, 67]. Existing efforts in developing FViTs mainly fall into two categories: (1) exploring how to scale up ViTs' architectures to construct powerful FViTs [40, 69, 75]; (2) developing self-supervised pretraining techniques to train FViTs so that their learned representations can be more effectively generalized to downstream tasks [4, 7, 17, 30, 36].

Unlike conventional convolutional neural networks (CNNs), FViTs extensively use the self-attention mechanism to extract global features, resulting in improved task accuracy with larger models (e.g., over 10G FLOPs). Specifically, in ViTs, a series of N input image patches $X = [x_1, \cdots, x_N]^\top \in \mathbb{R}^{N \times D}$, where D is the embedding dimension, is sequentially processed by ViT blocks. In each block, the input is first converted into queries $Q \in \mathbb{R}^{N \times d}$, keys $K \in \mathbb{R}^{N \times d}$ and values $V \in \mathbb{R}^{N \times d}$ (d denotes the hidden dimension) via linear projection, followed by the computation of the self-attention, which is calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \qquad (1)$$

The outputs are then fed into a feed-forward network to extract information in the channel dimension.

2.2. Parameter-efficient Tuning

Motivated by the impressive pretraining performance of FViTs on large-scale datasets, there has been a growing interest in applying FViTs to real-world applications. The common solution follows the pretraining-then-tuning paradigm, which tunes pretrained FViTs on various downstream tasks based on the corresponding applications' needs. However, with conventional weight tuning, each task would need to store an additional set of model weights, which can lead to cumbersome and prohibitive storage overhead. To this end, various parameter-efficient tuning methods have been proposed [2,33,53,72]. In parameter-efficient tuning, a set of tiny learnable modules are added to the pretrained FViTs, while the weights of the backbone FViTs remain unchanged during tuning [32–34]. This approach offers two benefits: (1) it allows FViTs to be tuned on new downstream tasks with negligible additional parameters, and (2) the pretrained FViTs can be easily retrieved at any time by simply removing the added parameter-efficient tuning modules.

Among recent parameter-efficient tuning techniques, LoRA [33] proposes to learn a set of low-rank weights and apply them on top of the backbone weights, and VPT [34] proposes to use the idea of prompt tuning, inserting a set of task-specific prompts as additional tokens. More recently, NPS [72] proposes to search for the optimal combination of parameter-efficient tuning techniques and their corresponding hyperparameters through neural architecture search.

2.3. Few-shot Tuning

Few-shot tuning aims to tune pretrained models on new tasks with limited samples per class [18, 21, 28, 39, 42]. It has gained increasing attention in recent years [59] as high-quality data is scarce in many real-world applications [3]. Recently, a few pioneering works that target few-shot tuning for ViTs propose to customize meta-learning tasks and learning objectives under the guidance of self-attention modules [8, 10, 38, 61, 65]. In this paper, we aim to enhance FViTs' few-shot tuning accuracy from an orthogonal direction, i.e., adaptively augmenting the few-shot tuning samples to compensate for their lack of diverse features.

2.4. Data Augmentation

Data augmentation aims to enhance data diversity and thus the feature diversity of the models [9, 11, 24, 31, 49, 60, 68, 71, 74]. An effective data augmentation strategy should properly enhance data diversity, while simultaneously avoiding the generation of out-of-manifold data caused by excessive augmentation intensity [57, 71]. Although various data augmentation techniques have been proposed, how to effectively augment the data under few-shot tuning settings is still an open question. The limited data diversity in few-shot data calls for techniques that can generate novel but meaningful features [62,63]. To this end, most existing few-shot data augmentation techniques adopt generative models to generate in-domain data, which, however, further increase the memory and storage overhead of tuning FViTs [23, 29, 37, 43].

One potential way to alleviate the aforementioned challenges is to use adversarial techniques to generate samples with beneficial features [20, 51, 59]. However, the majority of these works focus on improving adversarial robustness instead of the clean accuracy [20, 26, 51, 59, 70, 73]. In contrast, our work explores the opportunities of leveraging adversarial training to generate beneficial features that can boost the clean accuracy during few-shot parameter-efficient tuning.

3. The Proposed Hint-Aug Framework

3.1. Hint-Aug: Motivation

We first identify that the characteristics of parameter-efficient tuning together with pretrained FViTs provide a unique opportunity for FViTs' parameter-efficient tuning. Based on this observation, we then propose our Hint-Aug framework, which utilizes these characteristics to enhance the tuning effectiveness. We describe each of the characteristics in detail below:

Characteristics of parameter-efficient tuning: As mentioned in Sec. 2.1 and Sec. 2.2, the weights of pretrained FViTs are fixed during tuning. Therefore, the tuned FViTs behave the same as their pretrained counterpart after the added tuning modules (e.g., those adopted in Adapter [32], VPT [34], and LoRA [33]) are removed [72]. This motivates us to consider whether we can make use of this characteristic to improve the achievable few-shot tuning accuracy by leveraging the pretrained FViTs.

Characteristics of pretrained FViTs: Existing works have shown that pretrained FViTs have two promising characteristics regarding their learned features: (1) pretrained FViTs can identify complex but meaningful features [16, 41], even on unseen datasets without tuning [6, 30, 36]; (2) the learned features in FViTs can be reversely projected to the input image space using gradient-based methods [22, 44, 46].

Given the aforementioned characteristics of both parameter-efficient tuning and pretrained FViTs, we hypothesize that these characteristics provide a unique opportunity to effectively leverage the pretrained FViTs to augment the few-shot tuning data. To validate our hypothesis, we aim to explore proper ways to leverage the learned features in pretrained FViTs to boost the effectiveness of few-shot tuning. Specifically, given the two commonly recognized major challenges of few-shot tuning, which are overfitting [1, 58] and the lack of data diversity in the tuning data [62, 63], we set out to answer the following questions:

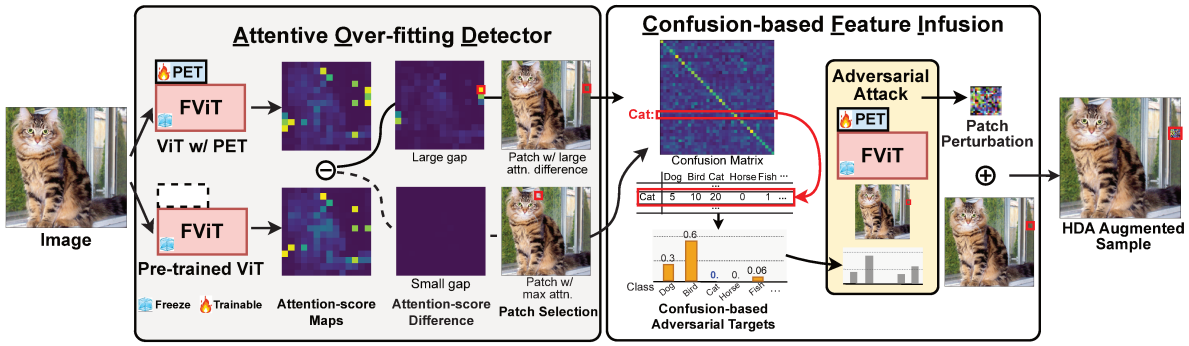

Figure 2. An overview of our proposed Hint-Aug framework, which consists of two enablers: (1) an AOD to detect whether the current sample is prone to over-fitting and which patch is prone to over-fitting, and (2) a CFI to infuse easy-to-confuse features to the over-fitted patches detected by the aformentioned AOD to increase the feature diversity of the tuning data and thus alleviate the over-fitting issue. In the figure, PET represents the parameter-efficient tuning module (e.g., Adapter [32], VPT [34], and LoRA [33]) added on top of the pretrained FViTs.

Q1 - *Can pretrained FViTs detect the potential over-fitting issue during few-shot tuning?* and **Q2** - *Can we leverage pretrained FViTs to generate meaningful features for enhancing the data diversity?* Our proposed Hint-Aug framework provides an effective solution to these two questions.

3.2. Hint-Aug: Overview

We first give an overview of our proposed Hint-Aug framework, which is dedicatedly designed for few-shot parameter-efficient tuning of FViTs by leveraging the characteristic of parameter-efficient tuning that pretrained FViTs' weights are not updated during tuning, allowing the features learned in pretrained FViTs to be utilized to augment the tuning data. As shown in Fig. 2, Hint-Aug adopts a two-stage detect-then-augment pipeline. In particular, to answer the above **Q1**, Hint-Aug uses AOD to detect (1) whether the tuned FViT is over-fitted on this image and (2) which patch in the image is more prone to be over-fitted; To address **Q2**, Hint-Aug further augments the patch detected from AOD by infusing the easy-to-confuse features with our proposed CFI module. We introduce our AOD and CFI modules in Sec. 3.3 and Sec. 3.4, respectively.

3.3. Enabler 1: Attentive Over-fitting Detector

Over-fitting is a well-known issue in few-shot tuning scenarios [37, 62], and becomes even more severe due to the combination of larger model size and limited data size during few-shot FViT tuning. Therefore, our AOD aims to explore whether we can detect the underlying over-fitting issue for each tuning sample on-the-fly during parameter-efficient tuning of FViTs.

Inspired by the various visualizations showing FViTs' attention distributions in previous works [22, 30, 52, 66], we hypothesize that the evolution of attention distributions during the tuning process contains hidden traces for identifying the existence of over-fitting. To validate this hypothesis, we utilize an attention-score map to quantify the impact of each input image patch on the FViT's attention distribution. Specifically, an attention-score map is constructed with the attention-score corresponding to each patch of the input image and we define the attention-score as follows: given the attention distribution $[a_1^{(l,h,i)}, \cdots, a_N^{(l,h,i)}]$ for the i-th patch of the h-th head in the l-th layer, the attention-score $s_j^{(l,k)}$ of the j-th patch for the k-th query patch is defined as:

$$s_j^{(l,k)} = \sum_h a_j^{(l,h,k)} \tag{2}$$

For the sake of simplicity, we omit the superscript l and k in the following text.

By visualizing the attention-score at different stages of tuning, as shown in Fig. 3, we can draw two observations: (1) the attention-score map of the pretrained FViT itself (see Fig. 3(a)) shares a high correlation with that of the half-tuned FViT model (see Fig. 3(b)), and a relatively higher tuning accuracy (e.g., 64.37%) suggests that the over-fitting issue is not severe at the corresponding tuning stage; (2) the attention-score map at the end of tuning (see Fig. 3(c)) focuses more on certain patches (marked in red) that are not focused on by the pretrained FViT, and a lower tuning accuracy (e.g., 61.55%) indicates the existence of the over-fitting issue. Additionally, we observe that patches with a newly attracted higher attention-score (marked in red) do not contain human-readable information for identification. For example, some patches only consist of a black background that does not contribute to identifying the target class, e.g., a cheese plate. This finding suggests that these patches could be the reason for over-fitting.

Based on the observations above, we propose an AOD module to use the attention-score map difference between the pretrained FViT and the corresponding one being tuned

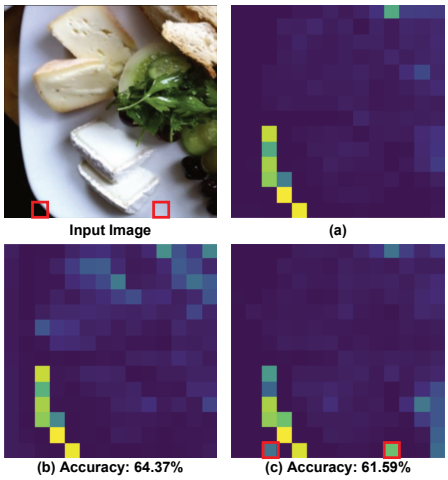

Figure 3. Visualization of the attention-score map from the (a) pretrained foundation model, called ViT-base, (b) parameter-efficient tuned ViT-base model with 20% of the total tuning epochs, achieving an accuracy of 64.37%, and (c) parameter-efficient tuned ViT-base model with an accuracy of 61.55%.

to identify both the existence of over-fitting and which patch contributes most to the over-fitting issue. Specifically, given the attention-score maps $S^P = [s_1^P, \cdots, s_N^P]$ generated from the pretrained FViT (denoted as P) and $S^T = [s_1^T, \cdots, s_N^T]$ generated from the FViT model to be tuned (denoted as T), we define the over-fitting indicator as:

$$I = \begin{cases} 0, & \sum_i \|s_i^P - s_i^T\| < \lambda \sum_i \|s_i^P\| \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

where λ is a hyperparameter to control the sensitivity of over-fitting detection.

When over-fitting occurs (i.e., $I = 1$), we select the patch that significantly changes the attention-score map as the target patch to be augmented in order to alleviate the over-fitting issue. Thus, we select the patch p to augment, where p is defined by:

$$p = \arg\max_i (\|s_i^P - s_i^T\|) \tag{4}$$

Otherwise, when there is no detected over-fitting issue, we select the patch p with the highest attention-score as the target patch to be augmented from all patches in the corresponding image.

3.4. Enabler 2: Confusion-based Feature Infusion

With the selected over-fitted patch detected by the AOD above, the remaining question is how to augment the selected patch with meaningful features to (1) alleviate the over-fitting issue and (2) increase the diversity of tuning data with meaningful features. Therefore, we propose CFI that uses adversarial attack-based methods to extract the learned features from the pretrained FViT model and infuse them into the selected patch with the aim of improving the feature diversity in a meaningful way, thus alleviating the over-fitting issue.

However, achieving a meaningful feature extraction and infusion that can help boost the few-shot tuning accuracy is non-trivial. Naively augmenting samples with commonly used attack objectives (e.g., perturbing the image to reduce the value of the model's output logit on the correct class) can easily lead to out-of-manifold samples, as shown in our alation study in Sec. 4.3.2. To overcome this, the CFI module incorporates injected features to steer the model prediction towards a synthetic target label. This target label is determined by utilizing a confusion matrix, which quantifies the degree to which the model is prone to confusion between pairs of classes.

Specifically, we construct a confusion matrix $C \in \mathbb{R}_{\geq 0}^{M \times M}$ in CFI, where M is the total number of classes. As shown in a recent study on open set detection [56], a pre-softmax model output has a better ability to preserve a model's uncertainty of samples. We thus define C as follows:

$$C_{i,j} = \sum_{X: y(X)=j} \Big(f_i(X) - \min_{i'} f_{i'}(X) \Big) \tag{5}$$

where i and j are coordinates in C that represent two classes; y and $f \in \mathbb{R}^M$ are the ground truth label and pre-softmax output given the input image X. The generated confusion matrix C helps to identify the class-wise similarity learned by the model and distinguish the class pairs that are easy to be confused by the model.

To infuse the easy-to-confuse features to the patch, given input X with label y, we propose to design the attack label $\tilde{f}(X) \in \mathbb{R}_{\geq 0}^M$ where the i-th element is computed as:

$$\tilde{f}_i(X) = \begin{cases} \frac{C_{i,y}}{\sum_j C_{j,y} - C_{y,y}}, & i \neq y \\ 0, & i = y \end{cases} \tag{6}$$

The loss function is defined as

$$\mathcal{L}_{tar} = \text{CrossEntropy}(\text{softmax}(f), \text{softmax}(\tilde{f})) \tag{7}$$

By optimizing the patch to minimize the above loss, the generated features are further shifted towards the direction where the model considers an easy-to-confuse class from the current class. This shift allows the model to learn to differentiate between the current class and the easy-to-confuse class, effectively extending the decision boundary of the current class.

4. Experimental Results

4.1. Experimental Setup

Datasets, few-shot settings, models, and parameter-efficient tuning techniques. Datasets and few-shot

settings. We adopt five commonly-used datasets for few-shot tuning, including Food [5], Pet [48], Cars [35], Flowers [34], and Aircraft [45], and benchmark our Hint-Aug under 1/2/4/8/12/16-shot scenarios to provide a thorough evaluation of its achieved accuracy across different few-shot tuning scenarios. Models. We conduct our experiment on a widely used FViT model, i.e., ViT-Base [16]. Adopted parameter-efficient tuning methods. We consider three most widely used parameter-efficient tuning methods including Adapter [32], LoRA [33], and VPT [34].

Baselines. We benchmark our proposed Hint-Aug against two baselines, including the SOTA data augmentation technique for parameter-efficient FViT tuning introduced in [72] (denoted as NPS) and the vanilla tuning without augmentation (denoted as No-Aug). It is worth noting that, given the unique challenge of limited data diversity in the few-shot tuning scenarios, even the SOTA data augmentation technique, i.e., the aforementioned NPS [72], can lead to an inferior accuracy than that of the vanilla tuning without augmentation (as shown in Fig. 1). Thus, it is necessary to include No-Aug as one of the baselines.

Tuning settings. In our experiments, we set $l = 5$ and adopt the center patch in each image as the query patch (i.e., $k = 90$), following [22]. We follow the widely adopted few-shot tuning settings in [72]. Specifically, we tune the model for 100 epochs using a batch size of 256, a learning rate of 0.01, and an SGD optimizer starting from the ImageNet [14] pretrained ViT-Base [16]. Following NPS [72], we also use data augmentation techniques including color-jitter with a factor of 0.4 and RandAugment [13] with a magnitude of 9 and a standard deviation equal to 0.5. We set λ in Eq. 3 as 0.1 and use FGSM [25] to generate the adversarial samples with attack radius $\epsilon = 0.001$. Additionally, we run all experiments in the paper three times and report the average accuracy, following NPS [72].

4.2. Benchmark on Few-shot Image Classification

We first benchmark our proposed method on five commonly used few-shot image classification datasets [5,34,35, 45,48] with different parameter-efficient tuning techniques and few-shot settings. As shown in Fig. 4, although the SOTA augmentation baseline NPS [72] suffers from considerable accuracy degradation compared with the vanilla tuning method No-Aug on fine-grained image classification dataset (e.g., a 5.55% accuracy drop on [45]), our proposed Hint-Aug achieves $0.25\% \sim 6.10\%$, $0.10\% \sim 32.91\%$, and $0.04\% \sim 6.17\%$ higher accuracies across different shot selections over baselines when using Adapter [32], VPT [34], and LoRA [33] tuning, respectively.

In particular, we draw the following two exciting observations: (1) the features generated by Hint-Aug can compensate for the lack of sufficient tuning data and improve accuracy under more stringent few-shot settings. Specifically, Hint-Aug boosts the accuracy of 8-shot tuning by $2.45\% \sim 4.96\%$ and surpasses the 12-shot tuning with NPS [72] by a $0.73\% \sim 2.22\%$ higher accuracy when tuning Adapter and LoRA on the Food and Pets datasets; (2) Hint-Aug's ability to extract features from the pretrained FViTs and infuse them into the tuning data can considerably boost accuracy in extreme few-shot scenarios (e.g., 1-shot tuning). For example, on the Pets dataset, tuning VPT with Hint-Aug under a 1-shot setting leads to a 32.91% higher accuracy than that of NPS [72].

Table 1. Ablation study on each enabler's contribution to the final accuracy.

AOD	CFI	Food	Pets	Cars
		66.25	86.97	40.83
✓		68.53	88.01	42.17
	✓	70.52	89.07	43.55
✓	✓	71.04	89.42	44.80

4.3. Ablation Studies

4.3.1 Accuracy Improvement Breakdown

Setup. To better understand the contribution of each enabler of Hint-Aug, including AOD and CFI, to the final accuracy, we conduct an ablation study where we run 8-shot tuning with Adapter [32] on three datasets, namely Food [5], Pets [48], and Cars [35]. We implement this accuracy improvement breakdown experiment as follows: (1) when using AOD only, we adopt the data augmentation method in [72] to augment the selected patch; (2) when using CFI only, we generate the samples with $\mathcal{L}_{tar}$ loss and randomly select a patch to augment in each image.

Observations. As shown in Tab. 1, when augmenting a selected patch, we can observe that (1) enabling either AOD or CFI can lead to an accuracy improvement of $1.04\% \sim 2.28\%$ and $2.10\% \sim 4.27\%$ over the baseline (e.g., neither AOD nor CFI enabled), respectively. This indicates that both key challenges (i.e., the over-fitting issue and lack of feature diversity as analyzed in Sec. 3.1) indeed hurt the achievable accuracy of few-shot tuning and our proposed enablers can effectively alleviate the challenge in over-fitting; (2) Combining both AOD and CFI can marry the merit of both, thus further boosting the achievable accuracy by $0.35\% \sim 2.63\%$ over that of enabling only one of AOD or CFI.

4.3.2 Ablation on Adversarial Objectives

Setup. We conduct ablation studies to validate the choice of loss functions for generating the adversarial sample for feature infusion. As mentioned in Sec. 3.4 and Sec. 2.3, different loss functions can have different impacts on the tuning accuracy and an improper loss function can lead to inferior clean accuracy. In Tab. 2, we validate the objective function we selected with other potential candidates when tuning on

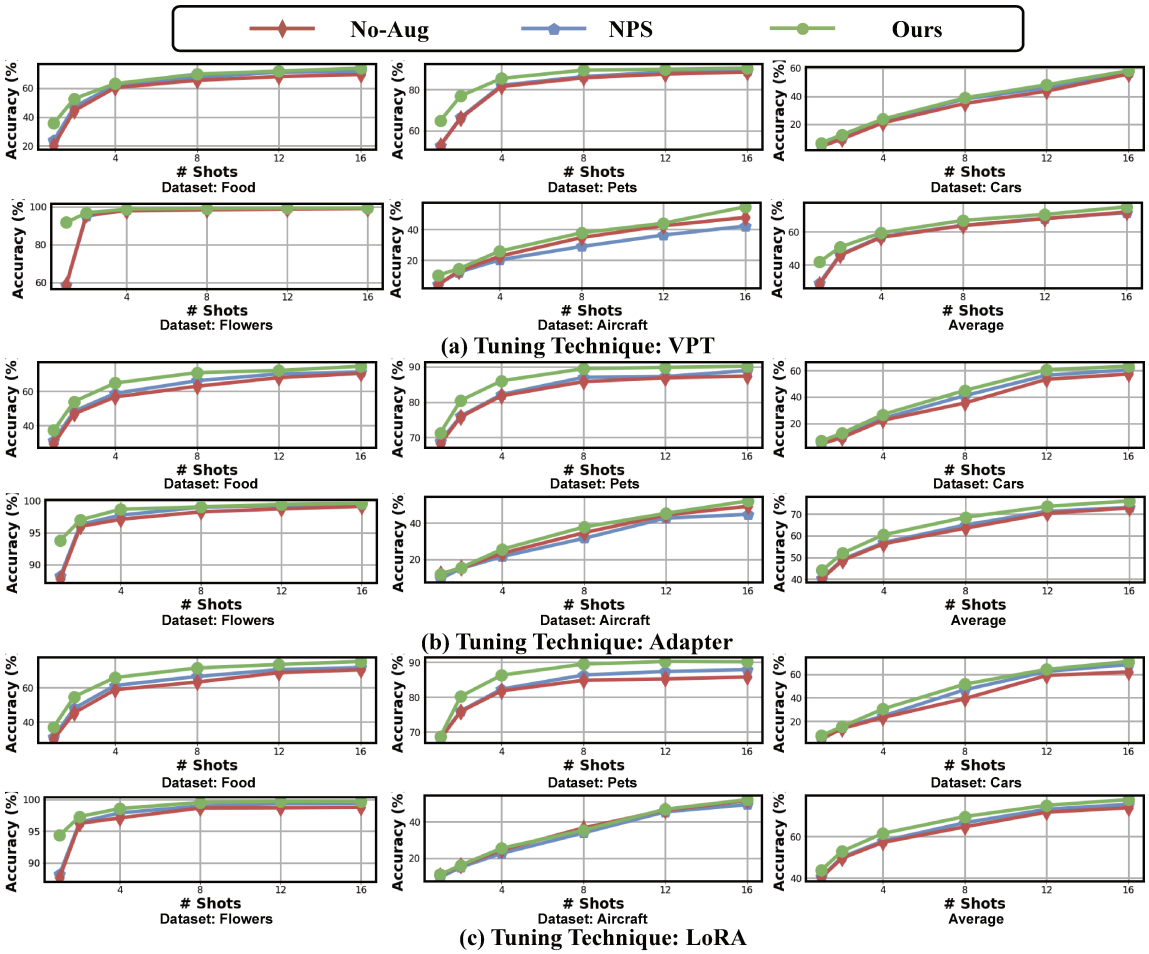

Figure 4. Benchmark Hint-Aug on Food [5], Pets [48], Cars [35], Flowers [47], and Aircraft [45] with parameter-efficient tuning methods (a) VPT, (b) Adapter, and (c) LoRA on 1/2/4/8/12/16-shot setting.

Table 2. Ablation study on different adversarial objectives.

Target	Full	Untarget	Random	Proposed
4-shot	57.49	59.21	62.35	64.92
8-shot	66.04	67.36	69.14	71.04
16-shot	70.78	71.58	73.85	74.90

the Food dataset [5] using Adapter [32], where "Full" indicates generating adversarial samples with the whole image, instead of the selected patch, "Untarget" means using the conventional attack target that minimizes the value of the model's output logit on the correct class by augmenting the selected patch, and "Random" means augmenting the selected patch to mislead the output of the augmented image toward another randomly selected class.

Observations. As shown in Tab. 2, "Full" leads to the worst achieved accuracy which is $0.80\% \sim 1.72\%$ lower than the second worst object "Untarget". "Untarget" also leads to a $3.32\% \sim 5.71\%$ lower accuracy than our proposed method. These two observations suggest that (1) attacking the image as a whole cannot effectively help with FViT tuning, and (2) naively using the "Untarget" attack can easily lead to out-of-manifold data. Furthermore, the $1.78\% \sim 3.14\%$ accuracy improvement of "Random" over "Untarget" suggests that despite the simple method of selecting the direction to add features, adding features from other classes can help with tuning. However, the lack of a more precise augmentation direction still limits the achievable accuracy when using the "Random" adversarial objective, leading to a $1.05\% \sim 2.57\%$ lower accuracy than the adversarial objective adopted in Hint-Aug.

4.3.3 Sensitivity to Augmentation Intensity

According to recent studies [52, 54], augmentation intensity is a crucial factor in FViT tuning. Thus, we investigate the impact of the adversarial attack radius ϵ on the achievable accuracy of Hint-Aug. When tuning with Adapter [32] on Food [5] under an 8-shot setting, Hint-Aug achieves relatively stable achieved accuracy under the drastic change in attack radius. Specifically, as shown in Tab. 3, increasing or decreasing ϵ by 5 times only leads to a $0.03\% \sim 0.21\%$

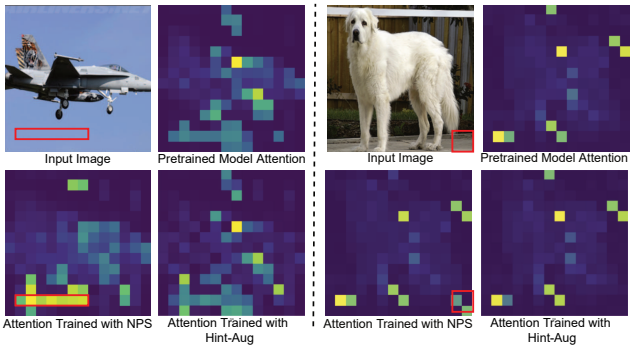

Figure 5. Visualization of attention score maps of images trained with different augmentation techniques.

Table 3. Impact of adversarial attack radius on the achievable accuracy of the Hint-Aug framework.

ϵ	0.01	0.005	0.001	0.0002	0.0001
Acc. (%)	70.01	70.83	71.04	71.01	70.15

Table 4. Ablation on the number of selected patches to augment.

# patches	1	2	3	8	32	All
Average Acc.	65.42	65.44	65.35	65.08	64.59	63.72

accuracy change, while changing ϵ by 10 times leads to a $0.89\% \sim 1.03\%$ accuracy change compared with a radius of 0.001 that we select in Hint-Aug, proving the robustness of Hint-Aug in different selections of hyperparameters. It is worth noting that changing ϵ by 10 times is a non-trivial change. As suggested in [22], changing ϵ by 8 times leads to an accuracy change larger than 27.94% when attacking DeiT-Tiny on ImageNet.

4.3.4 Number of Patches to Augment

Motivated by the promising accuracy-data efficiency trade-off achieved by Hint-Aug, an interesting question arises whether augmenting more than one patch for each image can further push forward the accuracy-efficiency trade-off. To answer this question, we conduct an ablation study on Hint-Aug with different numbers of augmented patches and report the average achieved accuracy when tuning with an 8-shot VPT [34] across five datasets. Notably, augmenting all patches (i.e., column "All" in Tab. 4) is equivalent to augmenting the whole image without considering the patch information. Our experiments show that augmenting one to three patches in each image leads to similar average accuracy (less than 0.1% accuracy change). However, when augmenting more patches in the image, the average accuracy drops by $0.34\% \sim 1.70\%$ when augmenting more than 8 patches in each image. We suspect this is because only a few patches are prone to over-fitting in each image, as suggested in Fig. 3. Augmenting too many patches may ruin the attention-score map instead, leading to reduced accuracy.

4.4. Visualization of Attention Score Maps

To verify Hint-Aug's effectiveness in alleviating the over-fitting issue, we visualize the attention score maps of the pretrained FViT, FViT tuned by NPS [72], and FViT tuned with our proposed Hint-Aug. As shown in Fig. 5, we can observe that (1) after tuning with our proposed Hint-Aug, the over-fitted patches (marked in red) that are commonly observed in the attention score maps tuned by NPS [72] are successfully eliminated, and (2) the attention score map obtained from Hint-Aug features similar locations of high-attention score patches to those obtained from the pretrained FViT, indicating that Hint-Aug effectively alleviates the over-fitting issue.

4.5. Visualization of the Confusion Matrix

We visualize the confusion matrix using a 4-shot Adapter [32] tuning setting on Pets [48] to interpret the discovered class-wise similarity. We calculate the averaged confusion matrix value of the Cats and Dogs meta-group and visualize them in Tab. 5. We observe that the FViT is much more confused in distinguishing between different classes within the Cat or Dog meta-group than distinguishing between the Cat and Dog meta-groups. This suggests that despite the simplicity of our strategy that uses the pre-softmax output, the generated confusion matrix can effectively identify the class pairs with easy-to-confuse features and thus provide correct guidance for CFI.

Table 5. The averaged confusion matrix value of the Cats and Dogs meta-group.

	Cats	Dogs
Cats	4.94	3.96
Dogs	3.96	5.72

5. Conclusion

In this paper, we propose a framework called Hint-Aug, which is dedicated to boosting the few-shot parameter-efficient tuning accuracy of FViTs. Specifically, Hint-Aug features two enablers called AOD and CFI, aiming to alleviate the over-fitting issue and the lack of diverse data in few-shot tuning, respectively. Extensive experiments and ablation studies validate that Hint-Aug achieves a $0.04\% \sim 32.91\%$ higher accuracy over SOTA data augmentation methods, opening up a new perspective towards more effectively tuning pretrained FViTs on downstream tasks in a realistic low-data scheme.

Acknowledgement

The work was supported by the National Science Foundation (NSF) through the NSF CCF program (Award number: 2211815) and supported in part by CoCoSys, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340, 2017. 3

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274, 1(3):4, 2022. 1, 3

[3] Ms Aayushi Bansal, Dr Rewa Sharma, and Dr Mamta Kathuria. A systematic review on data scarcity problem in deep learning: solution and applications. ACM Computing Surveys (CSUR), 54(10s):1–29, 2022. 3

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021. 2

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In European conference on computer vision, pages 446–461. Springer, 2014. 6, 7

[6] Shuhao Cao, Peng Xu, and David A Clifton. How to understand masked autoencoders. arXiv preprint arXiv:2202.03670, 2022. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294, 2021. 2

[8] Haoxing Chen, Huaxiong Li, Yaohui Li, and Chunlin Chen. Sparse spatial transformers for few-shot learning. arXiv preprint arXiv:2109.12932, 2021. 3

[9] Jie-Neng Chen, Shuyang Sun, Ju He, Philip Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. arXiv preprint arXiv:2111.09833, 2021. 3

[10] Yuzhong Chen, Zhenxiang Xiao, Lin Zhao, Lu Zhang, Haixing Dai, David Weizhong Liu, Zihao Wu, Changhe Li, Tuo Zhang, Changying Li, et al. Mask-guided vision transformer (mg-vit) for few-shot learning. arXiv preprint arXiv:2205.09995, 2022. 3

[11] Ziyi Cheng, Xuhong Ren, Felix Juefei-Xu, Wanli Xue, Qing Guo, Lei Ma, and Jianjun Zhao. Deepmix: Online auto data augmentation for robust visual object tracking. In 2021 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2021. 3

[12] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501, 2018. 1

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 702–703, 2020. 1, 6

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 6

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 1, 2, 3, 6

[17] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. arXiv preprint arXiv:2205.09113, 2022. 2

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning, pages 1126–1135. PMLR, 2017. 3

[19] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4):681–694, 2020. 2

[20] Nic Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial examples are a natural consequence of test error in noise. arXiv preprint arXiv:1901.10513, 2019. 3

[21] Ahmed Frikha, Denis Krompaß, Hans-Georg Köpken, and Volker Tresp. Few-shot one-class classification via meta-learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 7448–7456, 2021. 3

[22] Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-fool: Are vision transformers always robust against adversarial perturbations? arXiv preprint arXiv:2203.08392, 2022. 3, 4, 6, 8

[23] Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. Low-shot learning via covariance-preserving adversarial augmentation networks. Advances in Neural Information Processing Systems, 31, 2018. 3

[24] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Keepaugment: A simple information-preserving data augmentation approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1055–1064, 2021. 3

[25] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014. 6

[26] Shupeng Gui, Haotao Wang, Haichuan Yang, Chen Yu, Zhangyang Wang, and Ji Liu. Model compression with adversarial robustness: A unified optimization framework. Advances in Neural Information Processing Systems, 32, 2019. 3

[27] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 3714–3722, 2019. 2

[28] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In European conference on computer vision, pages 124–141. Springer, 2020. 3

[29] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features.

In Proceedings of the IEEE international conference on computer vision, pages 3018–3027, 2017. 3

[30] 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). arXiv preprint arXiv:2111.06377, 2021. 2, 3, 4

[31] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781, 2019. 1, 3

[32] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning, pages 2790–2799. PMLR, 2019. 3, 4, 6, 7, 8

[33] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021. 1, 3, 4, 6

[34] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. arXiv preprint arXiv:2203.12119, 2022. 1, 3, 4, 6, 8

[35] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In Proceedings of the IEEE international conference on computer vision workshops, pages 554–561, 2013. 6, 7

[36] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. arXiv preprint arXiv:2106.09785, 2021. 2, 3

[37] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13470–13479, 2020. 3, 4

[38] Lu Liu, William Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. A universal representation transformer layer for few-shot image classification. arXiv preprint arXiv:2006.11702, 2020. 3

[39] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. arXiv preprint arXiv:1805.10002, 2018. 3

[40] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. arXiv preprint arXiv:2111.09883, 2021. 1, 2

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030, 2021. 1, 2, 3

[42] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In International conference on machine learning, pages 97–105. PMLR, 2015. 3

[43] Qinxuan Luo, Lingfeng Wang, Jingguo Lv, Shiming Xiang, and Chunhong Pan. Few-shot learning via feature hallucination with variational inference. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3963–3972, 2021. 3

[44] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017. 2, 3

[45] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151, 2013. 2, 6, 7

[46] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. Google Research Blog, 2015, 2015. 2, 3

[47] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 1447–1454. IEEE, 2006. 7

[48] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In 2012 IEEE conference on computer vision and pattern recognition, pages 3498–3505. IEEE, 2012. 6, 7, 8

[49] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. arXiv preprint arXiv:2110.07858, 2021. 3

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140):1–67, 2020. 2

[51] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In European Conference on Computer Vision, pages 53–69. Springer, 2020. 3

[52] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. arXiv preprint arXiv:2106.10270, 2021. 4, 7

[53] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. arXiv preprint arXiv:2206.06522, 2022. 3

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR, 2021. 1, 2, 7

[55] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. arXiv preprint arXiv:2204.07118, 2022. 1

[56] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. arXiv preprint arXiv:2110.06207, 2021. 5

[57] Shashanka Venkataramanan, Yannis Avrithis, Ewa Kijak, and Laurent Amsaleg. Alignmix: Improving representation by interpolating aligned features. arXiv preprint arXiv:2103.15375, 2021. 3

[58] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. Advances in neural information processing systems, 29, 2016. 3

[59] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. Advances in neural information processing systems, 31, 2018. 3

[60] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. Advances in Neural Information Processing Systems, 34, 2021. 3

[61] Xixi Wang, Xiao Wang, Bo Jiang, and Bin Luo. Few-shot learning meets transformer: Unified query-support transformers for few-shot classification. arXiv preprint arXiv:2208.12398, 2022. 3

[62] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur), 53(3):1–34, 2020. 3, 4

[63] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7278–7286, 2018. 1, 3

[64] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22–31, 2021. 1

[65] Chengming Xu, Siqian Yang, Yabiao Wang, Zhanxiong Wang, Yanwei Fu, and Xiangyang Xue. Exploring efficient few-shot adaptation for vision transformers. arXiv preprint arXiv:2301.02419, 2023. 3

[66] Zhongzhi Yu, Yonggan Fu, Sicheng Li, Chaojian Li, and Yingyan Lin. Mia-former: Efficient and robust vision transformers via multi-grained input-adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 8962–8970, 2022. 4

[67] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432, 2021. 1, 2

[68] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6023–6032, 2019. 1, 2, 3

[69] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12104–12113, 2022. 2

[70] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. Advances in Neural Information Processing Systems, 32, 2019. 3

[71] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017. 1, 2, 3

[72] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. arXiv preprint arXiv:2206.04673, 2022. 1, 2, 3, 6, 8

[73] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. Advances in Neural Information Processing Systems, 33:14435–14447, 2020. 3

[74] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 13001–13008, 2020. 1, 3

[75] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:2103.11886, 2021. 2