Flatness Improves Backbone Generalisation in Few-shot Classification

Rui Li[1] Martin Trapp[1] Marcus Klasson[1,2] Arno Solin[1,2]

[1]Aalto University [2]Finnish Center for Artificial Intelligence

Abstract

Deployment of deep neural networks in real-world settings typically requires adaptation to new tasks with few examples. Few-shot classification (FSC) provides a solution to this problem by leveraging pre-trained backbones for fast adaptation to new classes. However, approaches for multi-domain FSC typically result in complex pipelines aimed at information fusion and task-specific adaptation without consideration of the importance of backbone training. In this work, we introduce an effective strategy for backbone training and selection in multi-domain FSC by utilizing flatness-aware training and fine-tuning. Our work is theoretically grounded and empirically performs on par or better than state-of-the-art methods despite being simpler. Further, our results indicate that backbone training is crucial for good generalisation in FSC across different adaptation methods.

1. Introduction

Deep neural networks have shown remarkable successes when trained on large labelled data sets. However, in many real-world applications, access to labelled data is limited and, therefore, training a network with good generalisation behaviour is challenging. This has sparked research on methods to adapt pre-trained models to new data domains and concepts, *i.e.*, classes, even if only a few examples exist. Few-shot classification (FSC) methods [7, 53] have shown promising performance in these scenarios. Earlier works on FSC [18, 46, 50] focus on homogeneous/single-domain learning tasks (*e.g.*, [32, 50]), *i.e.*, training data and test data both come from the same domain. However, as later shown, these elaborated methods can often be surpassed by simple fine-tuning [7, 12, 47] when the distribution shift between training and test data is sufficiently small. Consequently, Meta-Dataset [49] was introduced as a heterogeneous, multi-domain benchmark to reflect more realistic settings in which models must be adapted to previously unseen data domains with potentially large distribution shifts.

Existing approaches to tackle the multi-domain FSC problem can roughly be grouped into three categories: *(i)* learn to fuse information of independent backbones to obtain generalisable features [15, 36], *(ii)* learn an auxiliary network that predicts parameters of task-specific layers added to the backbone [2, 3, 43], or *(iii)* directly learn the parameters of task-specific layers during adaptation [33, 34, 48]. Crucially, all of those approaches heavily depend on the generalisation behaviour of the backbone(s) to be transferable to new domains and concepts. However, investigating effective backbone training with good generalisation behaviour is an overlooked topic and still in its infancy.

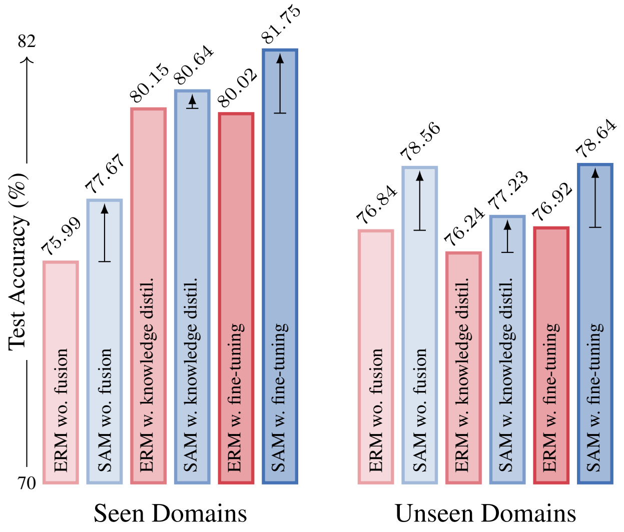


Figure 1. Average test accuracy on the Meta-Dataset benchmark for different backbone trainings using the adaptation by [34]. Across different information fusion methods, sharpness-aware minimisation (SAM) leads to better performance than empirical risk minimisation (ERM), showing flatness improves backbone generalisation.

Recently, the connection between model generalisation and flat optimum in the loss landscape has been studied empirically and theoretically [16, 27, 29] in the deep learning community. To this end, optimisers that seek flat minima have been proposed [19, 26] and have shown to improve generalisation in various deep learning applications [1, 8]. Moreover, recent findings in domain adaptation [6] indicate that flatness can improve generalisation in domain generalisation settings. Raising the question to what extend does flatness improve generalisation in FSC.

An additional challenge introduced in multi-domain settings is how to avoid domain conflicts when fusing information. A simple approach that has shown to be effective is to train one backbone per source domain rather than training one general backbone [15]. The information from each backbone can then be fused to obtain multi-domain feature representations that generalise to new tasks. However, as the number of training samples from each domain can vary (*e.g.*, in Meta-dataset), effective fusing of the information from different backbones can be challenging.

In this work, we introduce a theoretically justified and effective approach for backbone training and selection in multi-domain FSC. The approach is based on *(i)* seeking flat solutions during backbone training (*e.g.*, [19, 39]) to improve generalization, *(ii)* fusing information in the multi-domain setting using fine-tuning, *(iii)* and selecting the most compatible backbone for new tasks in cross-domain FSC settings. As shown in Fig. 1, combining these seemingly simple strategies results in a competitive approach compared to the state-of-the-art methods without changes of the adaptation method. Moreover, we observe that sharpness-aware training (SAM) of the backbone consistently improves generalisation in FSC over standard empirical risk-minimization (ERM). We present theoretical and empirical findings indicating that careful backbone training is crucial in FSC. Henceforth, we advocate for more careful treatments of the used backbones and a more competitive baseline.

Our contributions can be summarized as follows:

- We introduce an effective approach for multi-domain FSC which performs on par or better than state-of-the-art methods despite being simpler.
- We present theoretical results that flatness can improve generalisation in FSC, motivating our approach to backbone training and selection.
- We show empirical evidence that *(i)* flatness helps generalisation in FSC, *(ii)* fine-tuning is an effective information fusing method, and *(iii)* combining flatness and fine-tuning in the backbone training results in better performance compared to the state-of-the-art.

2. Background

We use calligraphic letters to denote sets (*e.g.*, the query set $\mathcal{Q}$), denote domains using fraktur font (*e.g.*, the target domain $\mathfrak{T}$), and use bold letters for vectors. Further, we denote the risk associated with a hypothesis/model $f_{\boldsymbol{\theta}}(\cdot)$ as $\mathcal{E}$ and use the hat-symbol $\hat{\mathcal{E}}$ whenever the risk is calculated w.r.t. an empirical distribution.

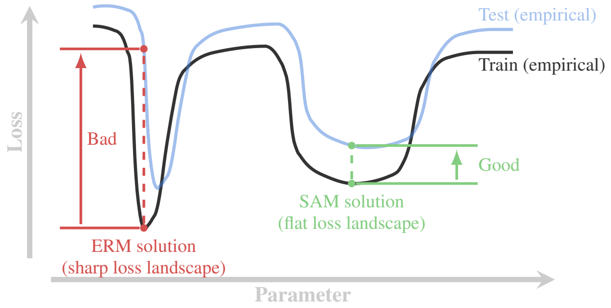


Figure 2. Illustration that solutions in flat areas on the training loss can result in better generalisation behaviour on the test loss.

2.1. Few-shot Classification

We assume to be given a training set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ of N input–output pairs, where $\mathbf{x}$ denotes the input and y its corresponding class label. In FSC, the goal is to learn a model $f_{\boldsymbol{\theta}}(\cdot)$ that can adapt to new classes or domains from few examples. At test time, we are given multiple tasks $t = 1, \ldots, T$ consisting of support $\mathcal{S}_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{S}_t|}$ and query sets $\mathcal{Q}_t = \{(\mathbf{x}_j, y_j)\}_{j=1}^{|\mathcal{Q}_t|}$ with $\mathcal{Q}_t \cap \mathcal{S}_\tau = \emptyset$ and $|\mathcal{S}_t| \ll N$. Support sets and query sets in FSC are similar to training data and test data respectively in supervised learning. Note the sets of classes at training and test time are disjoint and we might have a domain shift.

Let the model parameters be given as $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \boldsymbol{\psi}\}$, where we refer to $\boldsymbol{\phi}$ as the task-agnostic backbone and $\boldsymbol{\psi}$ as task-specific parameters. Task-agnostic parameters are learnt on the training set and task-specific parameters are learnt on the support set during evaluation. Specifically, learning a model in FSC can be divided into three phases: *(i)* during training, learning $\boldsymbol{\phi}$ on the training set, *(ii)* at test time, injecting task-specific layers parametrised by $\boldsymbol{\psi}$, and *(iii)* learning $\boldsymbol{\psi}$ during the adaptation on the support set of a sampled task while keeping $\boldsymbol{\phi}$ fixed. If the tasks are sampled from the same domain as the training set, we refer to the setting as in-domain and as cross-domain otherwise. If multiple training sets are available, *e.g.*, in multi-domain setting, we may have a backbone per data set or learn a general backbone.

2.2. Sharpness-aware Minimisation

Given a data set $\mathcal{D}$, the empirical risk minimisation (ERM) problem for a model $f_{\boldsymbol{\theta}}(\cdot)$ and a pointwise loss function $\ell(\cdot, \cdot)$ is defined as:

$$\arg\min_{\theta} \hat{\mathcal{E}}_{\text{ERM}}(\boldsymbol{\theta}; \mathcal{D}, \alpha) = \arg\min_{\theta} \mathcal{L}(\boldsymbol{\theta}) + \frac{\alpha}{2}\|\boldsymbol{\theta}\|^2, \quad (1)$$

where $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^N \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_n), y_n)$ and $\frac{\alpha}{2}\|\boldsymbol{\theta}\|^2$ is a regularization term. The goal of sharpness-aware minimisation (SAM) [19] is to reduce the generalisation error by additionally accounting for the loss geometry in the ERM objective.

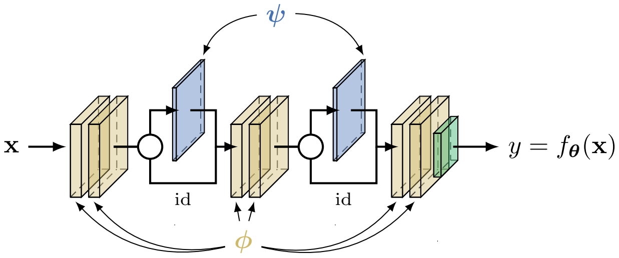


Figure 3. Decomposition of $f_{\boldsymbol{\theta}}(\cdot)$ with $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \boldsymbol{\psi}\}$ into task-agnostic layers parametrised by ϕ and task-specific layers parametrised by ψ. Additional gates ○ are used to switch between task-specific layers and the identity function. Note that this construction is only for theoretical purposes and does not imply any additional operations in practice.

In particular, SAM aims to simultaneously minimise the loss value and the loss sharpness by seeking parameters whose entire neighbourhood have uniformly low loss values under some $\boldsymbol{\epsilon}$ perturbation with $\|\boldsymbol{\epsilon}\| \leq \rho$, *i.e.*,

$$\hat{\mathcal{E}}_{\text{SAM}}(\boldsymbol{\theta}; \mathcal{D}, \rho, \alpha) = \max_{\|\boldsymbol{\epsilon}\| \leq \rho} \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}) + \frac{\alpha}{2}\|\boldsymbol{\theta}\|^2, \quad (2)$$

where $\rho > 0$ defines the radius of the neighbourhood of $\boldsymbol{\theta}$.

Fig. 2 provides an intuition that solutions to the SAM objective can result in improved generalisation behaviour of the trained model. Optimising the ERM objective can result in lower training loss compared to SAM. However, in the case of a distribution shift between the training and test sets, a solution in a flat loss landscape may yield better generalisation on the test loss. Based on this intuition, we study the use of SAM as a replacement of ERM in FSC.

3. Methods

We will now study the link between flatness and generalisation in the FSC in Sec. 3.1. Then in Sec. 3.2, we introduce our backbone training protocol: based on our theoretical results, we use a flatness-seeking objective for backbone training and introduce a backbone selection method to choose the best backbone for adaptation. For information fusion, we propose to use a fine-tuning strategy to fuse information in the multi-domain settings.

3.1. Flatness Leads to a Better Backbone for Adaptation

In FSC the backbone is trained using data from the source domain $\mathfrak{D}$ and later evaluated on data from the target domain $\mathfrak{T}$. We assume test tasks are sampled independently from sub-domains of the target domain $\mathfrak{T}_t \subset \mathfrak{T}$. Recall that the source and target (sub)-domains have disjoint sets of classes and a possible distribution shift.

Let $\ell(\cdot, \cdot)$ be a bounded loss function where $\ell(y_1, y_2) = 0$ iff $y_1 = y_2$. Given a model $f_{\boldsymbol{\theta}}(\cdot)$, we denote the *empirical* risk on the source domain as $\hat{\mathcal{E}}_{\text{ERM}}(\boldsymbol{\theta}; \mathcal{D})$ where $\mathcal{D} \in \mathfrak{D}$, the SAM loss on $\mathfrak{D}$ as $\hat{\mathcal{E}}_{\text{SAM}}(\boldsymbol{\theta}; \mathcal{D})$, and the risk on $\mathfrak{T}$ as:

$$\mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}) \triangleq \mathbb{E}_{\mathfrak{T}_t \sim \mathfrak{T}} \left[\mathbb{E}_{(\mathbf{x}, y) \sim \mathfrak{T}_t} \left[\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y) \right] \right]. \quad (3)$$

In the setting of domain generalisation, [6] showed the target domain loss can be bounded by the SAM loss on the source domain, the divergence between the source and the target, and a confidence bound that depends on the hyperparameter ρ of the SAM loss.

Theorem 3.1 ([6]). *First, let $\{\Theta_k \subset \mathbb{R}^d, k = 1, \ldots, K\}$, where d is dimension of Θ, be a finite cover of the parameter space Θ consisting of K closed balls with radius $\rho/2$ where $K \triangleq \lceil (\text{diam}(\Theta)/\rho)^d \rceil$. Denote the VC dimension of Θ and Θ_k as v and v_k, respectively. Then, for any $\boldsymbol{\theta} \in \Theta$, the following bound holds with probability at least $1 - \delta$:*

$$\mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}) \leq \hat{\mathcal{E}}_{SAM}(\boldsymbol{\theta}; \mathcal{D}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D}, \mathfrak{T}) + \max_k \sqrt{\frac{(v_k[\ln(N/v_k) + 1] + \ln(K/\delta))}{2N}}. \quad (4)$$

In Eq. (4), $\mathbf{Div}(\mathfrak{D}, \mathfrak{T})$ is the divergence between source and target domain. Building on their results, we show the expected generalisation gap on the target domain over test tasks in FSC can be upper bounded by the gap between the SAM and the ERM loss.

For this, let us assume a global labeling function and let $f_{\boldsymbol{\theta}}(\cdot)$ be decomposed as follows: *(i)* task-agnostic functions parametrised by $\boldsymbol{\phi}$, *(ii)* task-specific functions parametrised by $\boldsymbol{\psi}$, and *(iii)* gating functions that switch between task-specific functions and the identity function. Note that gating functions are only introduced to make the model contains task-agnostic and task-specific parameters during both training and testing so we can ensure theoretically correctness in the FSC setting. It does not add any extra operation in practice. During training on the source domain, all gates are set to choose the identity functions, while during adaptation on the target, all gates are set to select the task-specific layers. Fig. 3 illustrates our construction of the model functions in the FSC setting with the decomposition of $f_{\boldsymbol{\theta}}(\cdot)$ into task-agnostic and task-specific parts.

Theorem 3.2. *Let $\boldsymbol{\theta}^*_{SAM}$ denote the optimal solution of the SAM loss $\hat{\mathcal{E}}_{SAM}(\boldsymbol{\theta}; \mathcal{D})$,* i.e., *$\boldsymbol{\theta}^*_{SAM} \triangleq \arg\min_{\boldsymbol{\theta}} \hat{\mathcal{E}}_{SAM}(\boldsymbol{\theta}; \mathcal{D})$. Then, the gap between the loss $\min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta}; \mathfrak{T})$ and the loss of the optimal SAM solution on the training set, $\mathcal{E}(\boldsymbol{\theta}^*_{SAM}; \mathfrak{T})$,*

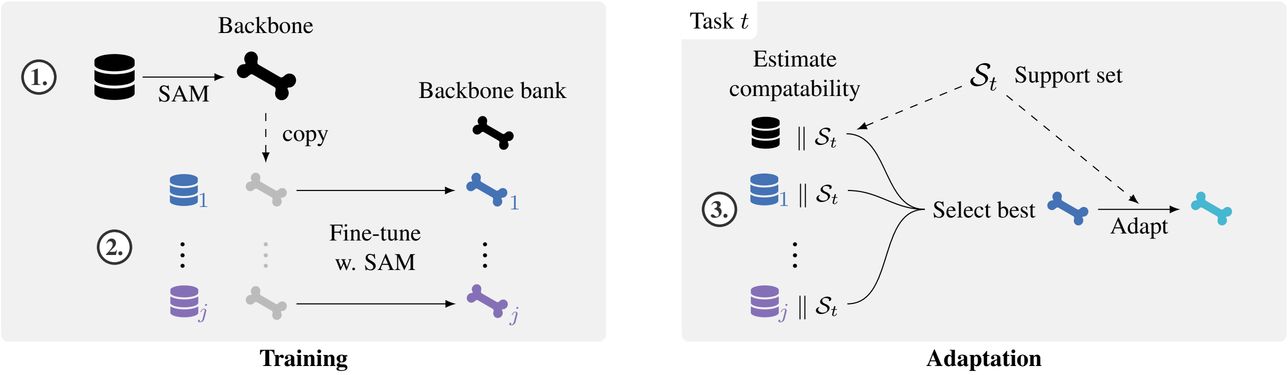


Figure 4. Our training protocol: ① SAM-based backbone training on a large and diverse data set (*e.g.*, ImageNet), ② SAM-based fine-tuning of on additional training data sets, ③ backbone selection and adaptation on the selected backbone →.

has the following bound with probability at least $1-\delta$*:*

$$\begin{aligned}
&\mathcal{E}(\boldsymbol{\theta}^*_{SAM};\mathfrak{T}) - \min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta};\mathfrak{T}) \\
&\quad \le \hat{\mathcal{E}}_{SAM}(\boldsymbol{\theta}^*_{SAM};\mathcal{D}) - \min_{\boldsymbol{\theta}} \hat{\mathcal{E}}_{ERM}(\boldsymbol{\theta};\mathcal{D}) \\
&\qquad + \mathbb{E}_{\mathfrak{T}_t\sim\mathfrak{T}}\left[\mathbf{Div}(\mathfrak{D},\mathfrak{T}_t)\right] + \sqrt{\frac{v\ln(N/v)+\ln(2/\delta)}{N}} \\
&\qquad + \max_{k}\sqrt{\frac{(v_k[\ln(N/v_k)+1]+\ln(K/\delta))}{2N}}, \qquad (5)
\end{aligned}$$

where N *is the number of training examples and* $\mathbf{Div}(\mathfrak{D},\mathfrak{T}_t)$ *is the divergence between domain* $\mathfrak{D}$ *and* $\mathfrak{T}_t$*. For further details and the proof, see App. A.*

Note that by our construction, $\boldsymbol{\theta}^*_{\text{SAM}}$ corresponds to the optimal solution under the SAM loss with gates switching to identity functions. Hence, Eq. (5) holds for any choice of ψ and its closeness will depend on the influence of ψ on the function value of $f_{\boldsymbol{\theta}}$. In practice, only a very small portion of $f_{\boldsymbol{\theta}}$ is task-specific, *e.g.*, for TSA [34] less than 1% of the parameters are task-specific, and we can consider the effect of ψ to be negligible.

Theorem 3.2 shows that the expected generalisation gap on the target domain can be bounded by: *(i)* the gap between the SAM and the ERM loss (in blue), *(ii)* the expected discrepancy between the source domain and the target domain (in green), *(iii)* and confidence bounds depending on ρ (in gray). Consequently, in the multi-domain FSC setting, *i.e.*, when we consider a collection of source domains $\{\mathfrak{D}_1,\ldots,\mathfrak{D}_D\}$, the bound in Theorem 3.2 suggests that the generalisation gap on the target depends on the selected source domain. Henceforth, we will suggest a backbone selection mechanism in Sec. 3.2 to minimise the expected generalisation gap on the target domain.

When the domain discrepancy between $\mathfrak{D}$ and $\mathfrak{T}$ is not large, the gap between SAM training loss and empirical training loss will play an important role at the bound. Given the complexity of loss landscapes in deep neural networks, it is sensible to assume that SAM with a proper ρ will find an optimal solution with similar training loss as ERM.

3.2. Backbone Training

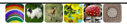
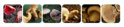
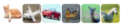
In this section, we introduce our proposed backbone training protocol for FSC. First, we use a flatness-seeking objective based on the SAM loss to train the backbone. Then, we propose to use a fine-tuning strategy to fuse information in the multi-domain settings. Finally, we introduce a backbone selection method that we use on unseen domains to choose the best backbone for adaptation. The training protocol and the respective steps are outlined in Fig. 4.

① Flatness Aware Training Objective Motivated by our theoretical result in Sec. 3.1, we propose to train the backbone with a flatness-seeking objective. More specifically, we propose to use the SAM objective [19] or variants thereof, *e.g.*, Bayesian-SAM (b-SAM) [39] or adaptive SAM [31]. This requires minor modification where we optimise the SAM objective rather than ERM when training the backbones, where any gradient-based optimiser can be applied to the SAM objective.

② Information Fusing using Fine-tuning One key challenge in multi-domain FSC is effective information fusion from different data sets. Simply training a single backbone on all domains will suffer from task conflicts [33]. To avoid this, we take inspiration from transfer learning where fine-tuning is a simple yet effective way to transfer knowledge between different data sets [42, 55]. Specifically, we propose to first train a base backbone on a diverse and extensive training data set (*e.g.*, ImageNet), then fine-tune the trained base backbone on smaller data sets. We experiment with standard fine-tuning and Low-Rank Adaptation (LoRA) [24].

③ Backbone Selection We propose using model selection scores on the backbone bank to determine which backbone is most suitable for feature extraction on unseen domain data.

Table 1. **Does flatness help generalisation? Yes.** Our performance comparison on the Meta-Dataset indicates that the SAM objective (SAM, b-SAM) results in better generalisation compared to ERM. Each trained backbone is combined with SUR or TSA for adaptation. Performance differences against ERM are indicated by ↑ and ↓. For visual comparison, we include example images from each data set.

			Adapted with SUR			*Adapted with TSA*		
		Example images	ERM	SAM	b-SAM	ERM	SAM	b-SAM
Seen during training	ILSVRC_2012		55.09±1.09	**56.72±1.11**	56.25±1.08	56.74±1.08	**58.99±1.08**	58.50±1.06
	OMNIGLOT		94.38±0.44	**94.69±0.44**	93.95±0.46	94.64±0.43	**94.87±0.42**	94.19±0.46
	AIRCRAFT		87.68±0.47	**89.51±0.43**	87.74±0.45	88.24±0.45	**89.91±0.42**	88.68±0.43
	CU_BIRDS		72.28±0.92	**74.18±0.83**	72.68±0.82	70.57±0.86	**74.27±0.81**	73.19±0.83
	DTD		72.08±0.73	72.99±0.80	**73.08±0.73**	61.33±0.71	**63.43±0.74**	62.73±0.77
	QUICKDRAW		83.36±0.56	**83.86±0.55**	83.75±0.56	83.64±0.58	**84.10±0.57**	84.00±0.58
	FUNGI		68.68±0.97	70.52±0.95	**70.67±0.95**	68.45±0.96	70.50±0.93	**70.54±0.92**
	VGG_FLOWER		87.11±0.52	**87.23±0.51**	86.42±0.54	84.28±0.58	**85.30±0.53**	83.49±0.62
Unseen	TRAFFIC_SIGN		45.33±1.02	**46.04±1.05**	44.05±1.15	80.73±0.97	**86.07±0.89**	80.87±0.92
	MSCOCO		**50.41±1.03**	50.28±1.08	48.54±1.03	56.07±1.05	**57.14±1.07**	55.87±1.02
	MNIST		94.71±0.42	94.16±0.40	**95.52±0.30**	96.59±0.35	96.84±0.34	**97.26±0.32**
	CIFAR10		68.76±0.76	66.93±0.86	**68.93±0.94**	79.62±0.73	**80.47±0.71**	80.04±0.71
	CIFAR100		59.32±1.05	59.31±1.07	**61.21±1.09**	71.21±0.96	**72.28±0.94**	71.95±0.93
	Average seen		77.58	**78.71** ↑**1.13**	78.07 ↑0.49	75.99	**77.67** ↑**1.68**	76.91 ↑0.92
	Average unseen		**63.71**	63.34 ↓0.37	63.65 ↓0.06	76.84	**78.56** ↑**1.72**	77.20 ↑0.36

As indicated in Theorem 3.2, the bound suggests that the generalisation gap on the target domain depends on the selected source domain. To narrow this gap in the cross-domain FSC setting, we use Pairwise Annotation Representation Comparison (PARC) [5] to select which backbone to use when adapting to unseen domains. During evaluation, we calculate the PARC scores for each backbone in the backbone bank on support set, and select the backbone with the highest score for current task. This only requires forward pass of trained backbones without the need for additional training of the backbones to compute the scores. See [5] for more details on PARC.

Our training protocol is adaptation- and backbone-agnostic, methodologically simple to allow easy integration into existing works, and theoretically motivated by Theorem 3.2.

4. Experiments

In this section, we first introduce the experimental setup and then study the following questions: *(i)* Does flatness help generalisation in FSC? *(ii)* How does fine-tuning compare against information fusion approaches? *(iii)* How does our proposed training protocol compare with the state-of-the-art methods?

4.1. Experimental Setup

We use the Meta-Dataset [49], a multi-domain FSC benchmark for in- and cross-domain generalisation, including data sets introduced by [43] for all evaluations. For this, we follow the standard varying-way varying-shot protocol. The Meta-Dataset contains the following training data sets: ImageNet [10], Omniglot [32], Aircraft [38], CU_Birds [51], VGG Flower [40], Quickdraw [28], Fungi [44], and Describable Textures [9]. For the cross-domain setting, we evaluated based on Traffic Signs [23], MSCOCO [35], MNIST [11], CIFAR-10 and CIFAR-100 [30].

To have a fair comparison with prior work, we use a ResNet-18 [21] as the backbone in main text experiments. In App. B, we also include results using a Vision Transformer [14] in Table 7 to verify that our proposed training procedure performs well for different network architectures. For the ResNet-18 backbone, we adopt the recent backbone-agnostic adaptation methods SUR [15] and TSA [34]. SUR combines features extracted from independently trained backbones and learns combination coefficients on the support set during adaptation. TSA adds task-specific feature adapters to the trained backbone and learns these on the support set. To ensure a fair comparison, we use the default hyperparameters provided for the adaptation. For unseen domains we select backbone based on PARC score for each task. To classify unseen classes during adaptation, we use a nearest-centroid classifier as typically adopted in FSC [46].

We compare our backbone training strategy against the following methods: *(i)* SUR [15] and URT [36] which learn to fuse information of independent backbones to obtain generalisable features; *(ii)* Simple CNAPS (S-CNAPS) [3] and Transductive CNAPS (T-CNAPS) [2] which learn an auxiliary network that predicts task-specific parameters of the backbone; *(iii)* URL [33] and TSA [34] which directly learn the parameters of task-specific layers during adaptation. We use a paired t-test ($p = 0.05$) to bold results with significant statistical difference. For more details, see App. C.

4.2. Does Flatness Help Generalisation in FSC?

To evaluate whether the flatness-seeking training leads to more generalisable backbones, we compare backbones trained with the SAM and ERM objectives, respectively.

Table 2. **Is fine-tuning effective? Yes.** Our performance comparison on the Meta-Dataset indicates that fine-tuning (LoRA, Vanilla) is an effective fusion strategy as it performs competitively against the other information fusion methods. We use TSA as the adaptation method, except for late fusion which is based on SUR. Surprisingly, fine-tuning outperforms knowledge distillation in the cross-domain (unseen) setting. Moreover, we observe that the performance on ImageNet deteriorates after knowledge distillation compared to no fusion.

	(N, C)	Late Fusion	No Fusion	Knowledge Distillation	Ours (LoRA)	Ours (Vanilla)
ILSVRC_2012	(11132759, 1000)	55.09±1.09	**56.74±1.08**	55.67±1.07	**56.74±1.08**	**56.74±1.08**
OMNIGLOT	(32460, 50)	94.38±0.44	94.64±0.43	**95.03±0.41**	93.48±0.49	94.00±0.46
AIRCRAFT	(10000, 100)	87.68±0.47	88.24±0.45	**89.95±0.45**	89.29±0.48	**89.94±0.43**
CU_BIRDS	(11788, 200)	72.28±0.92	70.57±0.86	**82.08±0.72**	80.84±0.74	81.46±0.68
DTD	(5640, 47)	72.08±0.73	61.33±0.71	**75.63±0.67**	74.80±0.76	74.35±0.74
QUICKDRAW	(50426266, 345)	83.36±0.56	**83.64±0.58**	82.33±0.62	80.96±0.65	83.08±0.60
FUNGI	(89760, 1394)	**68.68±0.97**	68.45±0.96	67.62±0.97	61.39±1.04	68.23±0.96
VGG_FLOWER	(8189, 102)	87.11±0.52	84.28±0.58	**92.90±0.43**	92.41±0.45	92.33±0.41
TRAFFIC_SIGN	(39209, 43)	45.33±1.02	80.73±0.97	**81.51±0.97**	80.73±0.97	80.73±0.97
MSCOCO	(860001, 80)	50.41±1.03	**56.07±1.05**	53.98±1.07	**56.07±1.05**	**56.07±1.05**
MNIST	(10000, 10)	94.71±0.42	96.59±0.35	96.65±0.37	**97.21±0.31**	97.00±0.34
CIFAR10	(10000, 10)	68.76±0.76	**79.62±0.73**	78.93±0.77	**79.62±0.73**	**79.62±0.73**
CIFAR100	(10000, 100)	59.32±1.05	**71.21±0.96**	70.11±1.00	**71.21±0.96**	**71.21±0.96**
Average seen		77.58	75.99	**80.15**	78.74	80.02
Average unseen		63.71	76.84	76.24	**76.97**	76.92

Table 3. Trace and top eigenvalues of the Hessian of the loss to measure the flatness of the trained backbones. Lower values mean flatter solutions. SAM in general results in flatter backbones compared with ERM.

	Trace of Hessian		*Top Eigenvalues of Hessian*	
	ERM	SAM	ERM	SAM
ILSVRC_2012	**8873.46**	13780.30	**237.13**	373.42
OMNIGLOT	199.15	**169.05**	9.39	**8.09**
AIRCRAFT	274.93	**170.70**	12.71	**11.53**
CU_BIRDS	**229.89**	341.53	**4.37**	10.60
DTD	134.64	**63.58**	7.60	**2.67**
QUICKDRAW	3508.08	**1785.32**	59.79	**37.50**
FUNGI	5359.20	**3998.80**	162.77	**149.88**
VGG_FLOWER	130.46	**89.39**	7.03	**6.66**

For the SAM objective, we use vanilla SAM [19] and b-SAM [39]. To assess the performance of varying adaptation methods, we employ both SUR and TSA. For SUR, the multi-domain features are fused during adaptation, while TSA needs backbone selection for the unseen domains. Hence, for TSA we employ our suggested backbone selection based on PARC [5]. We evaluate whether our backbone selection strategy chose the compatible backbone for unseen domain and report results in App. B.1. As shown in Table 6, PARC selects the most compatible backbone.

In Table 1, we observe that using the SAM objective during backbone training results in better generalisation on both seen and unseen domains in most cases. In particular, both SAM and b-SAM combined with TSA achieve better average performance on the seen and unseen domains compared to backbones trained with ERM. Note that for b-SAM, we are only using the posterior mean for computational reasons which might explain why it underperforms against SAM. For SUR, the information fusion combined with SAM might cause side effects that result in a slight drop in performance for the unseen domains. Further, the improvements on seen domains are larger than on the unseen domains in general. This performance gap is potentially caused by large domain shifts which would align with our theoretic findings, *c.f.* Theorem 3.2. Nevertheless, these results indicate that seeking flat minima during backbone training can improve generalisation in FSC.

Additionally, we measure the flatness of backbones trained with SAM and ERM using the trace and top eigenvalues of the Hessian of the loss [19]. [54] show that in the full-batch setting, SAM provably decreases the largest eigenvalue of Hessian, while in the stochastic setting (when batch size is 1), SAM provably decreases the trace of Hessian. Though in our experiment we use mini-batches where their theoretical results is inapplicable, we report trace and top eigenvalues of Hessian of trained backbones as they still measure the flatness in some degree. As shown in Table 3, SAM finds flatter solution in general compared with ERM.

4.3. Is Fine-tuning Enough for Information Fusion?

To evaluate how well fine-tuning performs for information fusion, we compare it against recent fusion methods. Our fine-tuning strategy on Meta-Dataset involves two steps: *(i)* train one backbone on the ImageNet data set, and *(ii)* fine-tune copies of the ImageNet-trained backbone on the remaining training data sets. We experiment with vanilla fine-tuning referred to as **Vanilla**, as well as fine-tuning using **LoRA** [24]. We compare against the following methods:

- **Late Fusion:** The multi-domain feature representation from SUR that fuses information from all backbones.
- **No Fusion:** Using single-domain backbones directly.
- **Knowledge Distillation:** The backbone from URL learned from distilling information from all backbones.

Note that all methods use TSA for adaptation, except for late fusion which is based on SUR. Moreover, all backbones are trained with the ERM objective.

In Table 2, we observe that our fine-tuning strategy performs competitively against other information fusion methods, especially on the unseen domains in Meta-Dataset. Compared to no fusion, both LoRA and vanilla fine-tuning yield better backbones on the seen domains, which means that the fine-tuning fuses information from ImageNet with the different domains successfully. On the unseen domains, our method and no fusion both select the ImageNet-backbone for the color datasets and Omniglot-backbone for MNIST, which is why their accuracies are similar.

When compared to the information fusion methods, our vanilla fine-tuning outperforms late fusion and performs competitively against knowledge distillation on both seen and unseen domains. While the smaller data sets benefit from using the universal backbone from knowledge distillation, we observe that our vanilla performs slightly better on the larger domains QuickDraw and Fungi. Furthermore, our method performs better than Knowledge Distillation on all unseen domains except Traffic Signs, which could be because the fine-tuning strategy mitigates the risk of task conflict when fusing the seen domains with ImageNet. These results demonstrate that our fine-tuning strategy is an effective alternative to previous information fusion methods.

4.4. How Does Our Approach Compare with SoTA?

We combine SAM-based backbone training with fine-tuning, which we denote as **SAM+FT**, and assess its performance by comparing it against SoTA on the Meta-Dataset. In addition, we combine SAM training with knowledge distillation based on URL [33], denoted as **SAM+KD**, to evaluate the effect of SAM on SoTA information fusion. We report the results with TSA adaptation for both approaches in Table 4.

We observe that our training protocol (SAM+FT) outperforms SoTA methods on most domains (10 out of 13) despite its simplicity. Furthermore, accounting for flatness in knowledge distillation (SAM+KD) results in mild improvements over TSA but is lacking behind our proposed approach. Our results show that flatness can improve generalisation in FSC across different information fusion strategies. To this end, we suggest that our simple yet effective training procedure be considered as a competitive baseline in FSC.

5. Related Work

Sharpness-aware minimisation The geometry of minima in neural network training has long been hypothesized to influence the generalisation behaviour of neural networks (*e.g.*, [22, 29]). Consequently, recent works studied theoretical links between flatness and generalisation of neural networks. Various algorithms accounting for the loss geometry have been proposed. For example, [29] showed a negative correlation between the sharpness of the loss landscape and the generalisation ability of the learner. Later, [13] related sharpness to the spectrum of the Hessian and [16] proposed a PAC-Bayes bound-based optimisation scheme to find flat minima. Recently, [52] introduced an augmented SAM loss which aims to further encourage flat minima and proposed a respective optimiser.

In the context of domain generalisation, [6] showed that flat minima can lead to a smaller generalisation gap on the target domain by leveraging results on generalisation in domain adaptation [4]. In the few-shot learning setting, [45] showed that flatness can help in overcoming catastrophic forgetting in the incremental learning setting and [17] recently proposed to account for flatness in the prompt tuning of large language models. However, to the best of our knowledge, sharpness-aware loss functions have not been leveraged or analysed in FSC settings.

Fine-tuning Transfer learning [42] involves utilizing and transferring knowledge learned from a set of source tasks to an unseen target task. For this, fine-tuning is a commonly adopted strategy where a pre-trained neural network, or a subset of its layers, is adapted to the target task. For example, [55] showed that fine-tuning a pre-trained network on a new data set can lead to better generalisation compared to training from scratch. For multi-domain FSC, avoiding task conflicts between different data sets during information fusing is an important problem. Motivated by the effectiveness of fine-tuning in transfer learning, fine-tuning has been adopted in the FSC setting.

In particular, fine-tuning has been shown to outperform elaborate methods in the adaptation stage (*e.g.*, [7, 12, 47]), in cases where the target domain is similar to the source domain (in-domain setting). Further, recent works showed that fine-tuning can be a successful adaptation strategy in both in-domain and cross-domain settings [20, 25, 37] and obtain competitive results compared with elaborate adaptation methods. However, fine-tuning the backbone before adaptation has received little to no attention.

Existing methods for Meta-Dataset FSC methods mainly focuses on two perspectives: *(i)* task-agnostic backbone training; *(ii)* adapting the task-agnostic backbone into a task-specific few-shot classifier.

For task-agnostic backbone training, FLUTE [48] trains a shared backbone jointly with domain-specific feature adapters on all training domains. It entangles adaptation with backbone training, limiting its applicability to other adaptation strategies. Later, [33] proposed to use knowledge distillation for information fusion, disentangling the backbone training from the adaptation. However, knowledge distillation is computationally expensive, and distilling knowledge from multiple domains simultaneously can suffer from task conflict.

Table 4. **A new baseline for FSC.** Our assessment on the Meta-Dataset shows that SAM-based training combined with fine-tuning (SAM+FT) outperforms SoTA methods in 10 out of 13 domains. Moreover, using SAM in conjunction with other information fusion methods, *e.g.*, knowledge distillation (SAM+KD), can improve generalisation performance. Henceforth, we advocate that our simple yet effective training procedure should be considered as a competitive baseline for both in-domain and cross-domain FSC.

	S-CNAPS	T-CNAPS	SUR	URT	URL	TSA	Ours (SAM+KD)	Ours (SAM+FT)
ILSVRC_2012	56.03±1.11	56.61±1.08	55.09±1.09	55.17±1.08	55.65±1.07	55.67±1.07	57.03±1.07	**59.01±1.08**
OMNIGLOT	91.45±0.62	92.91±0.50	94.38±0.44	94.42±0.46	94.76±0.41	**95.03±0.41**	**95.03±0.42**	94.46±0.43
AIRCRAFT	80.90±0.73	82.11±0.63	87.68±0.47	88.16±0.47	89.57±0.45	89.95±0.45	89.34±0.46	**92.63±0.35**
CU_BIRDS	75.10±0.86	77.35±0.77	72.28±0.92	79.04±0.77	81.51±0.69	82.08±0.72	82.58±0.70	**85.57±0.60**
DTD	68.90±0.72	68.76±0.73	72.08±0.73	73.05±0.67	74.66±0.65	75.63±0.67	**76.96±0.66**	75.35±0.72
QUICKDRAW	77.53±0.77	78.74±0.67	83.36±0.56	**83.51±0.56**	82.42±0.61	82.33±0.62	82.79±0.61	83.30±0.59
FUNGI	48.07±1.10	48.39±1.12	68.68±0.97	68.23±0.99	68.44±0.98	67.62±0.97	67.95±0.96	**70.13±0.91**
VGG_FLOWER	91.45±0.52	92.25±0.45	87.11±0.52	90.09±0.46	91.55±0.44	92.90±0.43	**93.41±0.43**	**93.59±0.39**
TRAFFIC_SIGN	58.33±1.03	56.83±1.13	45.33±1.02	47.11±1.02	60.15±1.18	81.51±0.97	82.83±0.93	**86.04±0.89**
MSCOCO	48.79±1.09	50.89±1.05	50.41±1.03	50.15±1.03	52.82±1.01	53.98±1.07	55.20±1.07	**57.13±1.07**
MNIST	93.81±0.42	95.13±0.30	94.71±0.42	88.89±0.48	94.84±0.42	96.65±0.37	96.71±0.37	**97.05±0.30**
CIFAR10	71.75±0.76	72.60±0.68	68.76±0.76	64.50±0.75	69.93±0.74	78.93±0.77	80.07±0.75	**80.60±0.71**
CIFAR100	61.62±1.05	62.35±1.02	59.32±1.05	55.85±1.07	61.58±1.09	70.11±1.00	71.34±0.98	**72.38±0.95**
Average seen	73.68	74.64	77.58	78.96	79.82	80.15	80.64	**81.75**
Average unseen	66.86	67.56	63.71	61.30	67.86	76.24	77.23	**78.64**

For adaptation strategies, [43] proposed CNAPS which learns an auxiliary network that predicts parameters of task-specific layers. Subsequently, [3] replaced the linear classifier in CNAPS with a nearest-centroid classifier using Mahalanobis distance inspired by ProtoNetwork [46], which was later extended with transductive learning [2]. However, learning an effective auxiliary network is difficult and its performance may suffer from distribution shift. Different approaches to adaptation are late fusion through linear combinations of features extracted from fixed backbones [15, 36], or learning task-specific layers through optimisation [34]. Nevertheless, little effort has been devoted to adaptation-agnostic backbone training.

6. Discussion and Conclusion

In this work, we have shown that flatness-seeking objectives, such as the SAM loss [19], can improve generalisation in few-shot classification (FSC). Combined with vanilla fine-tuning, minimising the SAM loss instead of the empirical risk results in a competitive baseline that outperforms current state-of-the-art methods in 10 out of 13 cases on the Meta-Dataset [49] benchmark (see Table 4).

In particular, we theoretically show that the generalisation gap on the target domain is upper bounded by the gap between the SAM and the ERM loss on the source domain and the difference between the domains. Motivated by this result, we proposed a backbone training protocol consisting of three steps: *(i)* SAM-based backbone training, *(ii)* information fusion using fine-tuning of the backbone(s), *(iii)* backbone selection in the multi-domain setting for unseen domains. We empirically showed that our approach is effective despite being methodologically simple, and that it can be combined with any adaptation method in FSC. Furthermore, we demonstrated that any information fusion method can potentially benefit from flat minima.

Limitations Our empirical findings are limited to the Meta-Dataset benchmark in which most data sets contain natural images or black-and-white drawings and written characters. As illustrated by the examples in Table 1, coloured training data sets can be considered similar (or even sub-sets) of ImageNet, and unseen domains are close to the in-domain data sets. Our additional results in Table 5 confirm this by highlighting the importance of the ImageNet backbone in the backbone selection. Moreover, our backbone selection strategy requires multiple forward passes, and the limitations of PARC [5] apply to our method. Lastly, SAM-based optimisation brings additional computational costs during backbone training.

Future directions Based on the improvement we have shown, investigating whether flatness helps generalisation in different model structures, *e.g.*, foundation models, and the wider few-shot learning context is a promising direction. Further, in consideration of the limitation of Meta-Dataset, it is important to investigate the performance of FSC methods in cross-domain settings. In particular, the generalisation gap on target domains with larger distribution shifts compared to the training data sets. Additional future directions include: leveraging the uncertainty estimates from b-SAM for more robust adaptation and backbone selection, improving the scalability of our approach, and assessing the performance in real-world downstream settings.

Code Publicly available under MIT license: https://github.com/AaltoML/FlatFSL

Acknowledgement AS and RL acknowledge funding from the Research Council of Finland (grant number 339730). MT acknowledges funding from the Research Council of Finland (grant number 347279). MK acknowledges funding from the Finnish Center for Artificial Intelligence (FCAI). We acknowledge CSC – IT Center for Science, Finland, for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through CSC. We acknowledge the computational resources provided by the Aalto Science-IT project.

References

[1] Dara Bahri, Hossein Mobahi, and Yi Tay. Sharpness-aware minimization improves language model generalization. In *60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7360–7371, 2022. 1

[2] Peyman Bateni, Jarred Barber, Jan-Willem Van de Meent, and Frank Wood. Enhancing few-shot image classification with unlabelled examples. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2796–2805. IEEE, 2022. 1, 5, 8

[3] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14481–14490. IEEE Computer Society, 2020. 1, 5, 8

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010. 7

[5] Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. Scalable diverse model selection for accessible transfer learning. *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 19301–19312, 2021. 5, 6, 8

[6] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 22405–22418, 2021. 1, 3, 7, 12

[7] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 7

[8] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[9] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613. IEEE Computer Society, 2014. 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE Computer Society, 2009. 5

[11] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 5

[12] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 7

[13] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 2017. 7

[14] 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. In *International Conference on Learning Representations (ICLR)*, 2020. 5

[15] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Selecting relevant features from a multi-domain representation for few-shot classification. In *European Conference on Computer Vision (ECCV)*, pages 769–786. Springer, 2020. 1, 2, 5, 8

[16] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017. 1, 7

[17] Shuo Fan, Liansheng Zhuang, and Aodi Li. Bayesian sharpness-aware prompt tuning for cross-domain few-shot learning. In *International Joint Conference on Neural Networks*, pages 1–7. IEEE, 2023. 7

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017. 1

[19] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International*

Conference on Learning Representations (ICLR), 2020. 1, 2, 4, 6, 8

[20] Yunhui Guo, Noel Codella, Leonid Karlinsky, James Codella, John Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision (ECCV)*, pages 124–141. Springer, 2020. 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE Computer Society, 2016. 5

[22] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997. 7

[23] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2013. 5

[24] Edward 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. In *International Conference on Learning Representations (ICLR)*, 2021. 4, 6

[25] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9068–9077. IEEE Computer Society, 2022. 7

[26] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pages 876–885. AUAI Press, 2018. 1

[27] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations (ICLR)*, 2019. 1

[28] Jonas Jongejan, Henry Rowley, Takashi Kawashima, Jongmin Kim, and Nick Fox-Gieg. The quick draw! a.i. experiment. https://quickdraw.withgoogle.com/, 2016. 5

[29] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2016. 1, 7

[30] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Toronto, ON, Canada, 2009. 5

[31] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 5905–5914. PMLR, 2021. 4

[32] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One-shot learning of simple visual concepts. In *Annual Meeting of the Cognitive Science Society*, 33, 2011. 1, 5

[33] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representation learning from multiple domains for few-shot classification. In *IEEE International Conference on Computer Vision*, pages 9526–9535. IEEE, 2021. 1, 4, 5, 7

[34] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain few-shot learning with task-specific adapters. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7161–7170. IEEE Computer Society, 2022. 1, 4, 5, 8

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 5

[36] Lu Liu, William L Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. A universal representation transformer layer for few-shot image classification. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 5, 8

[37] Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. A closer look at few-shot classification again. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 23103–23123. PMLR, 2023. 7

[38] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[39] Thomas Möllenhoff and Mohammad Emtiyaz Khan. Sam as an optimal relaxation of bayes. In *International Conference on Learning Representations*, 2022. 2, 4, 6

[40] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *60th Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 5

[41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li,

Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 16

[42] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 4, 7

[43] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*. Curran Associates, Inc., 2019. 1, 5, 8

[44] Brigit Schroeder and Yin Cui. Fgvcx fungi classification challenge. https://sites.google.com/view/fgvc5/competitions/fgvcx/fungi, 2018. 5

[45] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 6747–6761, 2021. 7

[46] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 30, 2017. 1, 5, 8

[47] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, pages 266–282. Springer, 2020. 1, 7

[48] Eleni Triantafillou, Hugo Larochelle, Richard Zemel, and Vincent Dumoulin. Learning a universal template for few-shot dataset generalization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 10424–10433. PMLR, 2021. 1, 7

[49] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 5, 8

[50] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, volume 29, pages 3630–3638. Curran Associates, Inc., 2016. 1

[51] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[52] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3769–3778, 2023. 7

[53] Yaqing Wang, Quanming Yao, James Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020. 1

[54] Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How sharpness-aware minimization minimizes sharpness? In *International Conference on Learning Representations (ICLR)*, 2022. 6

[55] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems 27 (NeurIPS)*, 2014. 4, 7

Appendices

A. Proofs

To prove Theorem 3.2, we first prove Lemmas A.1 and A.2. Then we use Lemmas A.1 and A.2 to prove Lemma A.3. After that, we use Lemma A.3 to prove Lemma A.4. At last, we use Lemma A.4 to prove Theorem 3.2.

For simplicity, in the bounded instance function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, m]$ where $\ell\left(y_1, y_2\right) = 0$ if and only if $y_1 = y_2$, we set $m = 1$ in the proof. We use proof techniques and results from [6].

A.1. Proof of Lemmas A.1 and A.2

We prove Lemmas A.1 and A.2 in this subsection.

Lemma A.1. *Define a functional error for two functions $f_1(\cdot)$ and $f_2(\cdot)$ on a domain $\mathfrak{D}$ as*

$$\mathcal{E}(f_1, f_2; \mathfrak{D}) \triangleq \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\mathfrak{D}}}[\ell(f_1(\mathbf{x}), f_2(\mathbf{x}))]. \tag{6}$$

For any domain $\mathfrak{D}$ and $\mathfrak{T}^{(i)}$, we have $|\mathcal{E}_{\mathfrak{D}}(f_1, f_2) - \mathcal{E}_{\mathfrak{T}}(f_1, f_2)| \leq \frac{1}{2}\mathbf{Div}(\mathfrak{D}, \mathfrak{T}^{(i)})$.

Proof. From the Fubini's theorem, we have

$$\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\mathfrak{D}}}\left[\ell\left(f_1(\mathbf{x}), f_2(\mathbf{x})\right)\right] = \int_0^{\infty} \mathbb{P}_{\mathfrak{D}}\left(\ell\left(f_1(\mathbf{x}), f_2(\mathbf{x})\right) > t\right) \mathrm{d}t. \tag{7}$$

Then,

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\mathfrak{D}}}\left[\ell\left(f_1(\mathbf{x}), f_2(\mathbf{x})\right)\right] - \mathbb{E}_{\mathbf{x}'\sim\mathbb{P}_{\mathfrak{T}^{(i)}}}\left[\ell\left(f_1(\mathbf{x}'), f_2(\mathbf{x}')\right)\right]\right| \\
&= \left|\int_0^{\infty} \mathbb{P}_{\mathfrak{D}}\left(\ell\left(f_1(\mathbf{x}), f_2(\mathbf{x})\right) > t\right) \mathrm{d}t - \int_0^{\infty} \mathbb{P}_{\mathfrak{T}^{(i)}}\left(\ell\left(f_1(\mathbf{x}'), f_2(\mathbf{x}')\right) > t\right) \mathrm{d}t\right| \\
&\leq \int_0^{\infty} \left|\mathbb{P}_{\mathfrak{D}}\left(\ell\left(f_1(\mathbf{x}), f_2(\mathbf{x})\right) > t\right) - \mathbb{P}_{\mathfrak{T}^{(i)}}\left(\ell\left(f_1(\mathbf{x}'), f_2(\mathbf{x}')\right) > t\right)\right| \mathrm{d}t \\
&\leq M \sup_{t\in[0,M]} \left|\mathbb{P}_{\mathfrak{D}}\left(\ell\left(f_1(\mathbf{x}), f_2(\mathbf{x})\right) > t\right) - \mathbb{P}_{\mathfrak{T}^{(i)}}\left(\ell\left(f_1(\mathbf{x}'), f_2(\mathbf{x}')\right) > t\right)\right| \\
&\leq M \sup_{f_1, f_2} \sup_{t\in[0,M]} \left|\mathbb{P}_{\mathfrak{D}}\left(\ell\left(f_1(\mathbf{x}), f_2(\mathbf{x})\right) > t\right) - \mathbb{P}_{\mathfrak{T}^{(i)}}\left(\ell\left(f_1(\mathbf{x}'), f_2(\mathbf{x}')\right) > t\right)\right| \\
&\leq M \sup_{A\in\mathcal{A}} \left|\mathbb{P}_{\mathfrak{D}}(A) - \mathbb{P}_{\mathfrak{T}^{(i)}}(A)\right| \\
&\triangleq \frac{1}{2}\mathbf{Div}(\mathfrak{D}, \mathfrak{T}^{(i)}),
\end{aligned} \tag{8}$$

where $\mathcal{A} = \{\mathbf{x}, \mathbf{x}' \mid \ell(f_1(\mathbf{x}), f_2(\mathbf{x})) > t, \ell(f_1(\mathbf{x}'), f_2(\mathbf{x}')) > t, \text{ for } t \in [0, M]\}$. □

Remark Because in FSC the source data and target data always have disjoint classes, we could assume there is a global labelling function $h(\cdot)$ for both source and target domain. Then in Lemma A.1, if we let $f_1(\mathbf{x})$ be the model $f_{\boldsymbol{\theta}}(\mathbf{x})$ and $f_2(\mathbf{x})$ be the global labelling function $h(\cdot)$, $\mathcal{E}(f_1, f_2; \mathfrak{D})$ becomes $\mathcal{E}(\boldsymbol{\theta}; \mathfrak{D})$ and $\mathcal{E}_{\mathfrak{T}^{(i)}}(f_1, f_2)$ becomes $\mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)})$, and we have

$$|\mathcal{E}(\boldsymbol{\theta}; \mathfrak{D}) - \mathcal{E}(\boldsymbol{\theta}; \mathfrak{T})| \leq \frac{1}{2}\mathbf{Div}(\mathfrak{D}, \mathfrak{T}). \tag{9}$$

Lemma A.2. *Let $\boldsymbol{\theta}_k \in \arg\max_{\Theta_k \cap \Theta} \mathcal{E}(\boldsymbol{\theta}; \mathfrak{D})$ be a local maximum in the k-th ball Θ_k. For any $\boldsymbol{\theta} \in \Theta$, the following bound holds with probability at least $1 - \delta$:*

$$\mathcal{E}(\boldsymbol{\theta}; \mathfrak{D}) - \hat{\mathcal{E}}_{SAM}(\boldsymbol{\theta}; \mathcal{D}) \leq \max_k \sqrt{\frac{(v_k[\ln(N/v_k) + 1] + \ln(K/\delta))}{2N}} \tag{10}$$

where N is the number of data points in the training set.

Proof. We first prove the following inequality holds

$$\mathcal{E}(\boldsymbol{\theta};\mathfrak{D})-\hat{\mathcal{E}}_{\text{SAM}}(\boldsymbol{\theta};\mathcal{D})\leq\max_{k}\left[\mathcal{E}\left(\boldsymbol{\theta}_{k'};\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_{k'};\mathcal{D}\right)\right], \tag{11}$$

then we prove for $\varepsilon_k\triangleq\sqrt{\frac{(v_k[\ln(N/v_k)+1]+\ln(K/\delta))}{2N}}$, $\varepsilon\triangleq\max_k\varepsilon_k$, we have $\mathbb{P}\left(\max_k\left[\mathcal{E}\left(\boldsymbol{\theta}_k;\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_k;\mathcal{D}\right)\right]>\varepsilon\right)\leq\delta$.

Since for any $\boldsymbol{\theta}$ there exists k' such that $\boldsymbol{\theta}\in\Theta_{k'}$, we have

$$\begin{aligned}\mathcal{E}(\boldsymbol{\theta};\mathfrak{D})-\hat{\mathcal{E}}_{\text{SAM}}(\boldsymbol{\theta};\mathcal{D})&=\mathcal{E}(\boldsymbol{\theta};\mathfrak{D})-\max_{\|\boldsymbol{\epsilon}\|\leq\rho}\hat{\mathcal{E}}_{\mathcal{D}}(\boldsymbol{\theta}+\boldsymbol{\epsilon})\\&\leq\mathcal{E}(\boldsymbol{\theta};\mathfrak{D})-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_{k'};\mathcal{D}\right)\\&=\mathcal{E}(\boldsymbol{\theta};\mathfrak{D})-\mathcal{E}\left(\boldsymbol{\theta}_{k'};\mathfrak{D}\right)+\mathcal{E}\left(\boldsymbol{\theta}_{k'};\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_{k'};\mathcal{D}\right)\\&\leq\mathcal{E}\left(\boldsymbol{\theta}_{k'};\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_{k'};\mathcal{D}\right)\\&\leq\max_{k}\left[\mathcal{E}\left(\boldsymbol{\theta}_{k'};\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_{k'};\mathcal{D}\right)\right],\end{aligned} \tag{12}$$

where the second inequality holds because $\boldsymbol{\theta}_k$ is the local maximum in Θ_k.

We now prove $\mathbb{P}\left(\max_k\left[\mathcal{E}\left(\boldsymbol{\theta}_k;\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_k;\mathcal{D}\right)\right]>\varepsilon\right)\leq\delta$. To do so, we first show the following inequality holds for the local maximum of N covers:

$$\begin{aligned}\mathbb{P}\left(\max_{k}\left[\mathcal{E}\left(\boldsymbol{\theta}_k;\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_k;\mathcal{D}\right)\right]>\varepsilon\right)&\leq\sum_{k=1}^{K}\mathbb{P}\left(\mathcal{E}(\boldsymbol{\theta}_k;\mathfrak{D})-\hat{\mathcal{E}}(\boldsymbol{\theta}_k;\mathcal{D})>\varepsilon\right)\\&\leq\sum_{k=1}^{K}\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\Theta_k}\left[\mathcal{E}(\boldsymbol{\theta};\mathfrak{D})-\hat{\mathcal{E}}(\boldsymbol{\theta};\mathcal{D})\right]>\varepsilon\right)\\&\leq\sum_{k=1}^{K}\left(\frac{eN}{v_k}\right)^{v_k}e^{-2N\epsilon^2}.\end{aligned} \tag{13}$$

Then by the definition of ε, we have

$$\begin{aligned}\mathbb{P}\left(\max_{k}\left[\mathcal{E}\left(\boldsymbol{\theta}_k;\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_k;\mathcal{D}\right)\right]>\varepsilon\right)&\leq\sum_{k=1}^{K}\left(\frac{eN}{v_k}\right)^{v_k}e^{-2N\varepsilon^2}\\&\leq\sum_{k=1}^{K}\left(\frac{eN}{v_k}\right)^{v_k}e^{-2N\varepsilon_k^2}\\&=\sum_{k=1}^{K}\frac{\delta}{N}=\delta.\end{aligned} \tag{14}$$

Since $\mathcal{E}(\boldsymbol{\theta};\mathfrak{D})-\hat{\mathcal{E}}_{\text{SAM}}(\boldsymbol{\theta};\mathcal{D})\leq\max_k\left[\mathcal{E}\left(\boldsymbol{\theta}_k;\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_k;\mathcal{D}\right)\right]$ and $\mathbb{P}\left(\max_k\left[\mathcal{E}\left(\boldsymbol{\theta}_k;\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_k;\mathcal{D}\right)\right]>\varepsilon\right)\leq\delta$, the following inequality holds with probability at least $1-\delta$:

$$\begin{aligned}\mathcal{E}(\boldsymbol{\theta};\mathfrak{D})-\hat{\mathcal{E}}_{\text{SAM}}(\boldsymbol{\theta};\mathcal{D})&\leq\max_{k}\left[\mathcal{E}\left(\boldsymbol{\theta}_k;\mathfrak{D}\right)-\hat{\mathcal{E}}\left(\boldsymbol{\theta}_k;\mathcal{D}\right)\right]\\&\leq\varepsilon=\max_{k}\sqrt{\frac{(v_k[\ln(N/v_k)+1]+\ln(K/\delta))}{2N}}.\end{aligned} \tag{15}$$

□

A.2. Proof of Lemma A.3

We prove Lemma A.3 using Lemmas A.1 and A.2 in this subsection.

Lemma A.3. *Let $\left\{\Theta_k \subset \mathbb{R}^d, k = 1, \cdots, K\right\}$ (d is dimension of Θ) be a finite cover of a parameter space Θ which consists of K closed balls with radius $\rho/2$ where $K \triangleq \lceil (\mathrm{diam}(\Theta)/\rho)^d \rceil$. Let v_k be a VC dimension of each Θ_k. Then, for any $\boldsymbol{\theta} \in \Theta$, the following bound holds with probability at least $1 - \delta$,*

$$\begin{aligned}\mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}) \leq & \hat{\mathcal{E}}_{SAM}(\boldsymbol{\theta}; \mathcal{D}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D}, \mathfrak{T}^{(i)}) \\ & + \max_k \sqrt{\frac{(v_k[\ln(N/v_k) + 1] + \ln(K/\delta))}{2N}}.\end{aligned} \tag{16}$$

Proof. We consider two cases.

Case 1: $\mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)}) \leq \mathcal{E}(\boldsymbol{\theta}; \mathfrak{D})$

As $\mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)}) - \mathcal{E}(\boldsymbol{\theta}; \mathfrak{D}) \leq 0$, Lemma A.3 automatically holds.

Case 2: $\mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)}) > \mathcal{E}(\boldsymbol{\theta}; \mathfrak{D})$

From Lemma A.1 and Eq. (9), we have

$$|\mathcal{E}(\boldsymbol{\theta}; \mathfrak{D}) - \mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)})| = \mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)}) - \mathcal{E}(\boldsymbol{\theta}; \mathfrak{D}) \leq \frac{1}{2}\mathbf{Div}(\mathfrak{D}, \mathfrak{T}^{(i)}). \tag{17}$$

Combining it with Lemma A.2, we have

$$\begin{aligned}&\mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)}) \leq \mathcal{E}(\boldsymbol{\theta}; \mathfrak{D}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D}, \mathfrak{T}) \\ &\mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)}) \leq \mathcal{E}(\boldsymbol{\theta}; \mathfrak{D}) - \hat{\mathcal{E}}_{\mathrm{SAM}}(\boldsymbol{\theta}; \mathcal{D}) + \hat{\mathcal{E}}_{\mathrm{SAM}}(\boldsymbol{\theta}; \mathcal{D}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D}, \mathfrak{T}) \\ &\mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)}) \leq \hat{\mathcal{E}}_{\mathrm{SAM}}(\boldsymbol{\theta}; \mathcal{D}) + \max_k \sqrt{\frac{(v_k[\ln(N/v_k) + 1] + \ln(K/\delta))}{2N}} + \frac{1}{2}\mathbf{Div}(\mathfrak{D}, \mathfrak{T}).\end{aligned} \tag{18}$$

□

A.3. Proof of Lemma A.4

We prove Lemma A.4 using Lemma A.3 in this subsection.

Lemma A.4. *Denote the VC dimension of Θ as v. Let $\boldsymbol{\theta}^*_{SAM}$ denote the optimal solution of the SAM loss $\hat{\mathcal{E}}_{SAM}(\boldsymbol{\theta}; \mathcal{D})$, i.e., $\boldsymbol{\theta}^*_{SAM} \triangleq \arg\min_{\boldsymbol{\theta}} \hat{\mathcal{E}}_{SAM}(\boldsymbol{\theta}; \mathcal{D})$. Then, the gap between the optimal test loss, $\min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)})$, and the test loss of SAM optimal solution on training set $\boldsymbol{\theta}^*_{SAM}$, $\mathcal{E}(\boldsymbol{\theta}^*_{SAM}; \mathfrak{T}^{(i)})$, has the following bound with probability at least $1 - \delta$:*

$$\begin{aligned}\mathcal{E}(\boldsymbol{\theta}^*_{SAM}; \mathfrak{T}^{(i)}) - \min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)}) \leq & \hat{\mathcal{E}}_{SAM}(\boldsymbol{\theta}^*_{SAM}; \mathcal{D}) - \min_{\boldsymbol{\theta}} \hat{\mathcal{E}}(\boldsymbol{\theta}; \mathcal{D}) + \mathbf{Div}(\mathfrak{D}, \mathfrak{T}^{(i)}) \\ & + \max_k \sqrt{\frac{(v_k[\ln(N/v_k) + 1] + \ln(K/\delta))}{2N}} + \sqrt{\frac{v\ln(N/v) + \ln(2/\delta)}{N}}.\end{aligned} \tag{19}$$

Proof. We first prove

$$-\min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)}) \leq -\min_{\boldsymbol{\theta}} \hat{\mathcal{E}}(\boldsymbol{\theta}; \mathcal{D}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D}, \mathfrak{T}) + \sqrt{\frac{v\ln(N/v) + \ln(2/\delta)}{N}} \tag{20}$$

holds with probability at least $1 - \delta$, then we combine it with Lemma A.3 to prove Lemma A.4.

Let $\bar{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{E}(\boldsymbol{\theta}; \mathfrak{T}^{(i)})$. From generalisation error bound of $\mathcal{E}(\bar{\boldsymbol{\theta}}; \mathfrak{D})$, the following inequality holds with probability at least $1 - \delta$,

$$\hat{\mathcal{E}}(\bar{\boldsymbol{\theta}}; \mathcal{D}) - \mathcal{E}(\bar{\boldsymbol{\theta}}; \mathfrak{D}) \leq \sqrt{\frac{v\ln(N/v) + \ln(2/\delta)}{N}}, \tag{21}$$

where v is a VC dimension of Θ.

Then, use Eq. (21) and Eq. (17), with probability at least $1-\delta$, we have

$$\begin{aligned}
\min_{\boldsymbol{\theta}} \hat{\mathcal{E}}(\boldsymbol{\theta};\mathcal{D}) &\le \hat{\mathcal{E}}(\boldsymbol{\theta};\mathcal{D}) \\
\min_{\boldsymbol{\theta}} \hat{\mathcal{E}}(\boldsymbol{\theta};\mathcal{D}) &\le \mathcal{E}(\bar{\boldsymbol{\theta}};\mathfrak{D}) + \sqrt{\frac{v\ln(N/v)+\ln(2/\delta)}{N}} \\
&\le \mathcal{E}(\bar{\boldsymbol{\theta}};\mathfrak{T}^{(i)}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D},\mathfrak{T}^{(i)}) + \sqrt{\frac{v\ln(N/v)+\ln(2/\delta)}{N}} \\
&\le \min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta};\mathfrak{T}^{(i)}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D},\mathfrak{T}^{(i)}) + \sqrt{\frac{v\ln(N/v)+\ln(2/\delta)}{N}} \\
-\min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta};\mathfrak{T}^{(i)}) &\le -\min_{\boldsymbol{\theta}} \hat{\mathcal{E}}(\boldsymbol{\theta};\mathcal{D}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D},\mathfrak{T}^{(i)}) + \sqrt{\frac{v\ln(N/v)+\ln(2/\delta)}{N}}.
\end{aligned} \tag{22}$$

Combine Eq. (22) with Lemma A.3, then with probability at least $1-\delta$ we have

$$\begin{aligned}
-\min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta};\mathfrak{T}^{(i)}) &\le -\min_{\boldsymbol{\theta}} \hat{\mathcal{E}}(\boldsymbol{\theta};\mathcal{D}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D},\mathfrak{T}^{(i)}) + \sqrt{\frac{v\ln(N/v)+\ln(2/\delta)}{N}} \\
\mathcal{E}(\boldsymbol{\theta}^*_{\text{SAM}};\mathfrak{T}^{(i)}) - \min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta};\mathfrak{T}^{(i)}) &\le \hat{\mathcal{E}}_{\text{SAM}}(\boldsymbol{\theta}^*_{\text{SAM}};\mathcal{D}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D},\mathfrak{T}^{(i)}) + \max_k \sqrt{\frac{(v_k[\ln(N/v_k)+1]+\ln(K/\delta))}{2N}} \\
&\quad - \min_{\boldsymbol{\theta}} \hat{\mathcal{E}}(\boldsymbol{\theta};\mathcal{D}) + \frac{1}{2}\mathbf{Div}(\mathfrak{D},\mathfrak{T}^{(i)}) + \sqrt{\frac{v\ln(N/v)+\ln(2/\delta)}{N}} \\
\mathcal{E}(\boldsymbol{\theta}^*_{\text{SAM}};\mathfrak{T}^{(i)}) - \min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta};\mathfrak{T}^{(i)}) &\le \hat{\mathcal{E}}_{\text{SAM}}(\boldsymbol{\theta}^*_{\text{SAM}};\mathcal{D}) - \min_{\boldsymbol{\theta}} \hat{\mathcal{E}}(\boldsymbol{\theta};\mathcal{D}) + \mathbf{Div}(\mathfrak{D},\mathfrak{T}^{(i)}) \\
&\quad + \max_k \sqrt{\frac{(v_k[\ln(N/v_k)+1]+\ln(K/\delta))}{2N}} + \sqrt{\frac{v\ln(N/v)+\ln(2/\delta)}{N}}.
\end{aligned} \tag{23}$$

□

A.4. Proof of Theorem 3.2

We prove Theorem 3.2 using Lemma A.4 in this subsection.

Proof. If we take expectation with respect to $\mathfrak{T}^{(i)} \sim \mathbb{P}(\mathfrak{T})$ on both sides of Lemma A.4, we have

$$\begin{aligned}
\mathbb{E}_{\mathfrak{T}^{(i)}\sim\mathbb{P}(\mathfrak{T})}\left[\mathcal{E}(\boldsymbol{\theta}^*_{\text{SAM}};\mathfrak{T}^{(i)}) - \min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta};\mathfrak{T}^{(i)})\right] &\le \hat{\mathcal{E}}_{\text{SAM}}(\boldsymbol{\theta}^*_{\text{SAM}};\mathcal{D}) - \min_{\boldsymbol{\theta}} \hat{\mathcal{E}}(\boldsymbol{\theta};\mathcal{D}) + \mathbb{E}_{\mathfrak{T}^{(i)}\sim\mathbb{P}(\mathfrak{T})}\left[\mathbf{Div}(\mathfrak{D},\mathfrak{T}^{(i)})\right] \\
&\quad + \max_k \sqrt{\frac{(v_k[\ln(N/v_k)+1]+\ln(K/\delta))}{2N}} + \sqrt{\frac{v\ln(N/v)+\ln(2/\delta)}{N}} \\
\mathcal{E}(\boldsymbol{\theta}^*_{\text{SAM}};\mathfrak{T}) - \min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta};\mathfrak{T}) \le \hat{\mathcal{E}}_{\text{SAM}}(\boldsymbol{\theta}^*_{\text{SAM}};\mathcal{D}) &- \min_{\boldsymbol{\theta}} \hat{\mathcal{E}}(\boldsymbol{\theta};\mathcal{D}) + \mathbb{E}_{\mathfrak{T}^{(i)}\sim\mathfrak{T}}[\mathbf{Div}(\mathfrak{D},\mathfrak{T}^{(i)})] \\
&\quad + \max_k \sqrt{\frac{(v_k[\ln(N/v_k)+1]+\ln(K/\delta))}{2N}} + \sqrt{\frac{v\ln(N/v)+\ln(2/\delta)}{N}}.
\end{aligned} \tag{24}$$

□

B. Additional Experimental Results

B.1. Backbone Selection

To investigate whether our backbone selection strategy selects a backbone that is compatible with data sets in unseen domains, we report the performance of each trained backbone and their corresponding PARC score in Tables 5 and 6 respectively. In Table 5, we observe that the ImageNet-trained backbone gives the best performance on the coloured data sets, while the Omniglot-trained backbone gives the best performance for MNIST, which is consistent with our backbone selection result in Table 6.

Table 5. We evaluate the performance of all backbones in the backbone bank on unseen domains. ImageNet-trained backbone gives the best performance on coloured data sets (Traffic Sign, MSCOCO, MNIST, CIFAR-10 and CIFAR-100) and the Omniglot-trained backbone gives the best performance on the monochrome data set (MNIST).

	Backbone trained on:							
	ILSVRC_2012	OMNIGLOT	AIRCRAFT	CU_BIRDS	DTD	QUICKDRAW	FUNGI	VGG_FLOWER
TRAFFIC_SIGN	**86.02±0.89**	49.18±1.28	58.54±1.23	62.57±1.16	69.91±1.17	80.85±1.10	63.17±1.20	61.25±1.20
MSCOCO	**57.07±1.08**	20.53±0.87	26.38±1.01	28.89±1.01	30.76±1.01	29.09±1.05	35.23±1.09	31.84±1.01
MNIST	94.35±0.56	**96.84±0.34**	86.70±0.85	90.91±0.77	91.30±0.67	96.32±0.38	92.22±0.69	90.29±0.70
CIFAR10	**80.56±0.71**	42.73±0.75	47.57±0.80	49.51±0.82	51.88±0.83	54.43±0.89	53.86±0.92	51.59±0.86
CIFAR100	**72.32±0.94**	25.47±1.01	31.88±1.13	37.44±1.17	38.16±1.19	37.27±1.22	45.24±1.24	40.29±1.18
Average unseen	78.07	46.95	50.21	53.87	56.40	59.59	57.94	55.05

Table 6. PARC scores on unseen domains in the Meta-Dataset for backbones in the backbone bank. Higher score means that the backbone is more compatible to the unseen domain.

	Backbone trained on:							
	ILSVRC_2012	OMNIGLOT	AIRCRAFT	CU_BIRDS	DTD	QUICKDRAW	FUNGI	VGG_FLOWER
TRAFFIC_SIGN	**22.22**	8.22	19.57	19.91	20.92	19.90	17.91	17.66
MSCOCO	**18.88**	7.31	10.80	11.15	10.87	11.84	11.22	12.41
MNIST	34.14	**45.05**	27.65	27.00	30.67	42.44	27.32	26.71
CIFAR10	**29.56**	7.75	14.76	12.17	15.91	18.75	13.04	15.55
CIFAR100	**18.26**	4.80	7.88	9.31	8.88	10.37	11.13	11.31

B.2. Vision Transformer Experiment

To test whether our proposed training procedure works for different model architecture, we conduct experiment on Vision Transformer (ViT) in this section. We use the ViT-small pre-trained with DINO [41] on ImageNet and fine-tune the pre-trained backbones on MetaDataset with SAM and ERM respectively. After training, we evaluate the performance by using Prototype classifier and the results are given in Table 7. Compared with ERM, SAM-trained backbones have better generalisation ability on all data sets.

Table 7. We fine-tune the pre-trained ViT with SAM and ERM respectively on MetaDataset and evaluate the trained backbone directly with Prototype classifier. SAM-trained backbones result in better generalisation than ERM.

	SAM	ERM
ILSVRC_2012	**63.97±0.98**	62.65±0.96
OMNIGLOT	**86.86±0.77**	86.32±0.81
AIRCRAFT	**84.20±0.59**	78.35±0.68
CU_BIRDS	**78.08±0.76**	73.50±0.83
DTD	**78.56±0.61**	**78.56±0.61**
QUICKDRAW	**84.13±0.53**	83.70±0.55
FUNGI	**72.46±0.91**	68.45±0.94
VGG_FLOWER	**95.20±0.32**	93.81±0.36
TRAFFIC_SIGN	**54.38±1.05**	50.46±1.09
MSCOCO	**64.74±0.88**	64.33±0.87
MNIST	**91.94±0.42**	90.72±0.47
CIFAR10	**89.27±0.46**	88.49±0.45
CIFAR100	**82.85±0.68**	82.10±0.68
Average Seen	80.43	78.17
Average Unseen	76.64	75.22

C. Implementation Details

For a fair comparison with prior work, we use RestNet-18 as the backbone in all experiments. We use NVIDIA V100 GPUs for backbone training and run the experiments on a cluster.

C.1. Implementation Details for Table 1

For ERM, we use the backbone provided by SUR where the SGD optimiser is used for all backbones. To eliminate the influence introduced by different optimiser, we adopt SGD as well for SAM backbone training. For b-SAM, we follow the optimisation algorithm proposed in the paper. For both SAM and b-SAM, we use cosine annealing learning rate decay with a restart and provide their hyperparameters in Table 8 and Table 9 respectively.

Table 8. Hyperparameters of SAM training in Table 1.

	Batch Size	Learning Rate	Total Iterations	Optimizer Restart	ρ
ILSVRC_2012	128	0.01	480000	48000	0.05
OMNIGLOT	16	0.03	50000	3000	0.01
AIRCRAFT	8	0.03	50000	3000	0.1
CU_BIRDS	16	0.03	50000	3000	0.1
DTD	32	0.03	50000	1500	0.1
QUICKDRAW	64	0.01	480000	48000	0.05
FUNGI	32	0.03	480000	15000	0.05
VGG_FLOWER	8	0.03	50000	1500	0.1

Table 9. Hyperparameters of b-SAM training in Table 1. For all data sets, we set $\rho = 0.01$, $\gamma = 0.1$, and $\beta_1 = 0.9$.

	Batch Size	Learning Rate	Prior Precision	Total Iterations	Optimiser Restart
ILSVRC_2012	500	0.1	100	250000	50000
OMNIGLOT	200	0.1	100	10000	2000
AIRCRAFT	200	0.5	10	5000	1000
CU_BIRDS	200	0.5	10	5000	1000
DTD	200	0.1	10	5000	1000
QUICKDRAW	200	0.5	50	30000	6000
FUNGI	200	0.1	50	30000	6000
VGG_FLOWER	200	0.1	10	5000	1000

C.2. Implementation Details for Table 2

We use SGD optimisers and cosine annealing learning rate decay with a restart for both vanilla fine-tuning and LoRA fine-tuning. For LoRA, we set the rank $r = 10$ for all data sets. The hyperparameters for vanilla and LoRA fine-tuning are provided in Table 10 and Table 11 respectively.

Table 10. Hyperparameters for vanilla fine-tuning in Table 2.

	Batch Size	Learning Rate	Iterations	Optimizer Restart
OMNIGLOT	16	0.01	10000	2000
AIRCRAFT	8	0.001	10000	2000
CU_BIRDS	16	0.001	10000	2000
DTD	32	0.001	10000	1000
QUICKDRAW	64	0.005	100000	5000
FUNGI	32	0.01	80000	5000
VGG_FLOWER	8	0.001	10000	2000

Table 11. Hyperparameters for LoRA fine-tuning in Table 2.

	Batch Size	Learning Rate	Iterations	Optimizer Restart
OMNIGLOT	16	0.005	40000	2000
AIRCRAFT	8	0.005	40000	2000
CU_BIRDS	16	0.005	10000	2000
DTD	32	0.001	10000	2000
QUICKDRAW	64	0.01	100000	5000
FUNGI	32	0.005	80000	5000
VGG_FLOWER	8	0.001	10000	2000

C.3. Implementation Details for Table 4

We use SGD optimisers and cosine annealing learning rate decay with a restart for SAM objective fine-tuning. The hyperparameters are given in Table 12 and we set $\rho = 0.05$ for all data sets. When combining SAM with knowledge distillation, we use the provided hyperparameter setting for backbone training in URL.

Table 12. Hyperparameters for SAM combined with vanilla fine-tuning in Table 4.

	Batch Size	Learning Rate	Iterations	Optimizer Restart
OMNIGLOT	16	0.01	20000	2000
AIRCRAFT	8	0.001	20000	2000
CU_BIRDS	16	0.001	20000	2000
DTD	32	0.001	20000	1000
QUICKDRAW	64	0.01	100000	5000
FUNGI	32	0.01	80000	5000
VGG_FLOWER	8	0.001	10000	2000