LoRA Recycle: Unlocking Tuning-Free Few-Shot Adaptability in Visual Foundation Models by Recycling Pre-Tuned LoRAs

Zixuan Hu[1] Yongxian Wei[2] Li Shen[3*] Chun Yuan[2] Dacheng Tao[1]
[1]Nanyang Technological University, Singapore;
[2]Tsinghua University, China; [3]Shenzhen Campus of Sun Yat-sen University, China.
ZIXUAN014@e.ntu.edu.sg; weiyx23@mails.tsinghua.edu.cn;
mathshenli@gmail.com; yuanc@sz.tsinghua.edu.cn; dacheng.tao@gmail.com

Abstract

Large Language Models (LLMs) such as ChatGPT demonstrate strong few-shot adaptability without requiring fine-tuning, positioning them ideal for data-limited and real-time applications. However, this adaptability has not yet been replicated in current Visual Foundation Models (VFMs), which require explicit fine-tuning with sufficient tuning data. Besides, the pretraining-finetuning paradigm has led to the surge of numerous task-specific modular components, such as Low-Rank Adaptation (LoRA). For the first time, we explore the potential of reusing diverse pre-tuned LoRAs without accessing their original training data, to achieve tuning-free few-shot adaptation in VFMs. Our framework, LoRA Recycle, distills a meta-LoRA from diverse pre-tuned LoRAs with a meta-learning objective, using synthetic data inversely generated from pre-tuned LoRAs themselves. The VFM, once equipped with the meta-LoRA, is empowered to solve new few-shot tasks in a single forward pass, akin to the in-context learning of LLMs. Additionally, we incorporate a double-efficient mechanism, accelerating the data-generation and meta-training process while maintaining or even improving performance. Extensive experiments across various few-shot classification benchmarks across both in- and cross-domain scenarios demonstrate the superiority of our framework. Code is available at https://github.com/Egg-Hu/LoRA-Recycle.

1. Introduction

Large Language Models (LLMs) like ChatGPT demonstrate a profound capacity to solve few-shot tasks without the necessity for fine-tuning [4, 40], making them ideal for data-limited and real-time applications. However, this adaptability has not yet been replicated by current Visual Foundation Models (VFMs), which typically require explicit fine-tuning with sufficient tuning data.

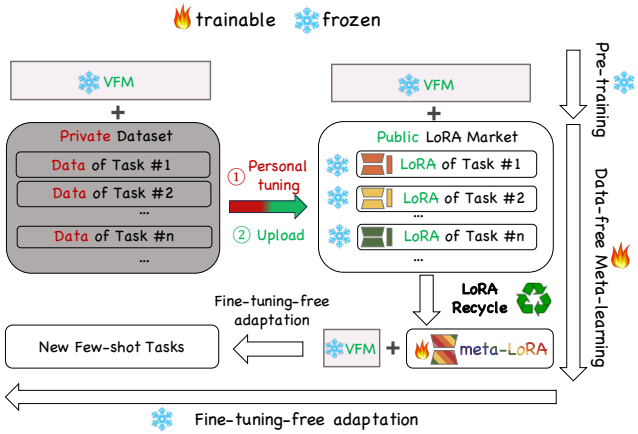

Figure 1. Concept of LoRA Recycle: Thanks to the modularity of LoRA, users can upload locally tuned LoRAs to public repositories without exposing original training data. LoRA Recycle distills a meta-LoRA from these LoRAs without needing their original training data. The VFM, once equipped with the meta-LoRA, is empowered to solve new few-shot tasks in a single forward pass without further fine-tuning.

Low-Rank Adaptation (LoRA) [26] has emerged as a prominent fine-tuning approach, offering strong performance with sufficient tuning data while requiring updates only to a small subset of additional parameters—specifically, trainable rank decomposition matrices—rather than the entire model. While promising, (i) explicit fine-tuning is often prohibitive for applications requiring real-time responses, and (ii) fine-tuning with limited data is extremely unstable. As shown in Tab. 1, fine-tuning with limited data makes performance highly sensitive to choices like the optimizer, learning rate, and step size. Besides, it introduces unacceptable latency for applications requiring real-time responses.

In this paper, we explore the potential of reusing diverse

*Corresponding author

pre-tuned LoRAs without accessing their original training data, to achieve tuning-free few-shot adaptation in VFMs (see Fig. 1). Our inspiration comes from the concept of *LoRA Market* [33], where diverse pre-tuned LoRAs are publicly accessible without exposing original training data due to privacy concerns. For task-specific reuse, users can download and then insert a task-specific LoRA of interest into the open-source VFM, to obtain a personalized VFM. Moving beyond task-specific reuse, we seek to leverage the vast availability and diversity of these LoRAs from a novel perspective, leading to our central research question: *Is it feasible to reuse diverse pre-tuned LoRAs without accessing their original training data, to achieve tuning-fee few-shot adaptation in VFMs?* This offers new insights into leveraging the vast accessibility and diversity of LoRAs beyond task-specific reuse, and avoids the need to access original training data, which is often restricted by privacy concerns.

To answer this question, we propose a framework named **LoRA Recycle** (see Fig. 2). To replace the unavailable original data, we propose *LoRA Inversion* that employs reverse engineering to generate synthetic data from pre-tuned LoRAs. Using these synthetic data, a *meta-LoRA* is distilled from the pre-tuned LoRAs with a meta-learning objective, explicitly learning how to adapt to diverse tasks without fine-tuning. Thanks to the meta-learning objective, the VFM, once equipped with the meta-LoRA, is empowered to adapt to new few-shot tasks in a single forward pass without further fine-tuning. The core idea of LoRA Recycle is to reshape the VFMs' prior over a distribution of expected tasks—represented by pre-tuned LoRAs—via meta-learning, and such prior encoded in the meta-LoRA can facilitate learning of new tasks sampled from similar distributions. To further improve efficiency, we introduce a double-efficient mechanism. During the data-generation stage, unimportant image tokens are pruned based on self-attention weights in hidden layers, accelerating the reverse engineering process. The pruning results further guide token selection for subsequent meta-training, ensuring that only the most informative tokens in the synthetic data are used. This selective use of sparse tokens significantly accelerates the meta-training process, while maintaining or even improving performance by reducing noise from synthetic data. We summarize our contributions as follows:

- **Novel perspective:** We are the first to enable tuning-free few-shot adaptation in VFMs from the perspective of LoRA reusing, offering new insights into leveraging the vast accessibility and diversity of pre-tuned LoRAs beyond traditional task-specific reuse.
- **Unified framework:** (i) We propose LoRA Recycle, a unified framework achieving tuning-free few-shot adaptation in VFMs by reusing diverse pre-tuned LoRAs without needing their original training data. (ii) We further propose a double-efficient mechanism, significantly accelerating the data-generation and meta-training process, while maintaining or even improving performance.
- **Experiments:** Extensive experiments across various few-shot classification benchmarks, within both in-domain and cross-domain scenarios, demonstrate the effectiveness of LoRA Recycle. Notably, LoRA Recycle achieves an average 6.27% improvement over baselines in the 5-way 1-shot classification setting.

Table 1. Fine-tuning ViT-B/16 on 600 5-way 1-shot classification tasks from the meta-testing set of CIFAR-FS. We report the accuracy, throughput (tasks per second) and GPU memory usage during fine-tuning. Values highlighted in green represent the best, whereas those in red denote the worst.

Method	Optimizer	Step	Learning Rate			Throughput (tasks/s) ↑	GPU Mem (GB) ↓
			0.1	0.01	0.001		
Full Fine-Tuning	SGD	50	22.81	30.13	28.99	0.10	12.88
		5	20.56	23.69	23.85		
	Adam	50	20.04	20.43	26.64		
		5	20.00	19.96	21.09		
LoRA	SGD	50	79.29	77.07	36.61	0.13	9.54
		5	73.48	37.27	20.60		
	Adam	50	22.10	26.55	82.19		
		5	20.40	73.00	55.11		
LoRA Recycle (ours)	—	—	**89.70** (+7.51%)			**8.25** (×63)	**1.28** (-87%)

2. Related Work

2.1. LoRA & LoRA Reuse

Fine-tuning the entire foundation model results in high computational costs. To mitigate these challenges, several parameter-efficient fine-tuning (PEFT) methods [21, 22, 26, 36, 46, 77] have emerged, enabling adaptation by updating only a small subset of model parameters. LoRA [26] parallelly attaches extra low-rank decomposition matrices to original weights, while adapter tuning [1, 17, 25] sequentially appends extra layers behind the original feed-forward layers. More recently, several works [6, 19, 33, 77, 78] have investigated the potential of composing multiple pre-tuned LoRAs. However, (i) they are limited to parameter arithmetic like weight averaging, lacking precise alignment for LoRAs targeting different label spaces in the context of classification. (ii) They are not specifically designed to achieve tuning-free few-shot adaptation in VFMs. (iii) They are not applicable to reuse LoRAs with different architectures like different ranks.

2.2. Meta-Learning & Data-Free Meta-Learning

Meta-learning, also known as *learning to learn*, aims to learn prior knowledge over a distribution of tasks, enabling efficient adaptation to unseen few-shot tasks from similar distributions. Data-based meta-learning [14, 15, 29, 38, 67–69, 81] typically assumes the availability of task-specific data for each meta-training task. Recently, Data-Free Meta-Learning (DFML) [30–32, 70, 73–75] emerges as a promising solution to directly meta-learn from pre-trained mod-

els available off the shelf. However, existing methods face scalability challenges with large Vision Transformers due to the high computational cost of data generation and meta-learning. In contrast, our framework meta-trains only a lightweight meta-LoRA and incorporates a novel double-efficiency mechanism, which significantly accelerates both the data-generation and meta-training processes.

2.3. Tuning-Free Adaptation of Foundation Models

Compared to explicit fine-tuning, training-free adaptation eliminates the need for parameter updates, making it particularly well-suited for real-time applications with limited computational resources. In the case of LLMs, tuning-free adaptation is achieved through their inherent in-context learning capabilities [10], acquired via large-scale pre-training. Existing studies suggest that in-context learning implicitly performs gradient descent [9, 63], framing LLMs as meta-learning models [3]. Furthermore, these inherent in-context learning abilities can be enhanced through explicit meta-learning after pre-training [7, 50]. However, this in-context learning capability has yet to be effectively replicated in current VFMs. To bridge this gap, [13] explicitly meta-trains a sequence model with VFMs to emulate LLM-style in-context learning. Additionally, [48, 83] adapt the Segment Anything Model in a tuning-free manner using a one-shot example.

3. Preliminary & Problem Setup

Low-Rank Adaptation (LoRA) [26] enables VFM to solve a specific task by only updating lightweight extra modules. For a weight matrix $W^{(l)} \in \mathbb{R}^{d\times k}$ at the l^{th} layer within the VFM f, a LoRA module is represented as a low-rank matrix decomposition $\delta W^{(l)} = \delta W_{\mathrm{A}}^{(l)} \cdot \delta W_{\mathrm{B}}^{(l)}$, where $\delta W_{\mathrm{A}}^{(l)} \in \mathbb{R}^{d\times r}$, $\delta W_{\mathrm{B}}^{(l)} \in \mathbb{R}^{r\times k}$ and the rank $r \ll \min(d,k)$. The input $\mathbf{X}_{\mathrm{in}}$ will be processed in parallel as $\mathbf{X}_{\mathrm{out}}^{(l)} = W^{(l)}\mathbf{X}_{\mathrm{in}}^{(l)} + \delta W_{\mathrm{A}}^{(l)}\delta W_{\mathrm{B}}^{(l)}\mathbf{X}_{\mathrm{in}}^{(l)}$. When fine-tuning, it freezes the original weight matrix W while only keeping δW_{A} and δW_{B} trainable. When facing classification tasks, a classification head h is always tuned together with the LoRA modules to output the prediction distribution. We use $f_{\delta W}$ to denote the VFM equipped with the LoRA δW.

Problem setup: LoRA Recycle. We are given a transformer-based VFM f pre-trained on large-scale datasets, and multiple LoRAs with classification heads pre-tuned on diverse classification tasks. Following standard meta-learning setup [14], we assume these tasks follow an underlying task distribution $p_{\mathcal{T}}$. $(\delta W_{\mathcal{T}}, h_{\mathcal{T}}) \sim p_{\mathcal{T}}$ denotes the LoRA and classification head pre-tuned on task $\mathcal{T}$. Note that we have no access to the original training data behind the given LoRAs. Our goal is to meta-train a meta-LoRA δW^* over $p_{\mathcal{T}}$, so that the VFM f, once equipped with δW^* (*i.e.*, $f_{\delta W^*}$), can adapt to new few-shot tasks sampled from similar distributions without further fine-tuning.

Testing setup. We conduct evaluation on 600 N-way K-shot classification tasks. Note that the classes in these testing tasks have not been seen by any given LoRA. Each N-way K-shot task $\mathcal{T}$ consists of one support set $\mathcal{D}_{\mathrm{s}}^{\mathcal{T}}$ and one query set $\mathcal{D}_{\mathrm{q}}^{\mathcal{T}}$. The support set $\mathcal{D}_{s}^{\mathcal{T}}$ has N classes and K examples per class. We focus on a few-shot setting where K is small (e.g., 1 or 5), thus fine-tuning f with extremely few examples is infeasible. In contrast, we use $\mathcal{D}_{\mathrm{s}}^{\mathcal{T}}$ to adapt f in a tuning-free manner. The query set $\mathcal{D}_{\mathrm{q}}^{\mathcal{T}}$ is what we actually make predictions on. The overall accuracy is measured by averaging the accuracy across all testing tasks.

4. Methodology

In this section, we present our proposed framework LoRA Recycle (see Fig. 2 and Alg. 1). We propose LoRA Inversion that generates synthetic data from pre-tuned LoRAs as a surrogate for inaccessible original data (Sec. 4.1). Using these synthetic data, a meta-LoRA is then distilled from the pre-tuned LoRAs, with a well-designed meta-learning objective (see Sec. 4.2). Furthermore, we propose a double-efficient mechanism that accelerates the data-generation and meta-training process (Sec. 4.3).

4.1. Synthetic Data Generation via LoRA Inversion

LoRA Inversion. Given a pre-tuned LoRA δW with its classification head h, we generate its original training data by iteratively optimizing (a batch of) data $\mathbf{X}$, which is initialized as Gaussian noise. This is done by minimizing the following loss function:

$$\min_{\mathbf{X}} \mathcal{L}_{\mathrm{data}} = \mathrm{CE}\left(h \circ f_{\delta W}(\mathbf{X}), \mathbf{Y}\right) + \alpha_{\mathcal{R}}\mathcal{R}_{\mathrm{BN}}(\mathbf{X}), \quad (1)$$

where $\mathbf{Y}$ is the target label (*e.g.*, $[1,0,0]$). $\mathrm{CE}(\cdot)$ is a cross-entropy classification loss. $\mathcal{R}_{\mathrm{BN}}$ is an image regularization term with a coefficient $\alpha_{\mathcal{R}}$. Minimizing the first classification loss is to achieve label-conditional generation, ensuring $\mathbf{X}$ can be predicted by $f_{\delta W}$ as the target label $\mathbf{Y}$. The second regularization term is added to improve the realism of the synthetic data [80], which is formulated as:

$$\mathcal{R}_{\mathrm{BN}}(\mathbf{X}) = \sum_{l} \left\|\mu^{(l)}(\mathbf{X}) - \mu_{\mathrm{BN}}^{(l)}\right\|_2 + \left\|\sigma^{(l)}(\mathbf{X}) - \sigma_{\mathrm{BN}}^{(l)}\right\|_2, \quad (2)$$

where $\mu^{(l)}(\mathbf{X})$ and $\sigma^{(l)}(\mathbf{X})$ denote the mean and variance of the inputs' feature maps calculated at the l^{th} layer of the pre-trained model. $\mu_{\mathrm{BN}}^{(l)}$ and $\sigma_{\mathrm{BN}}^{(l)}$ denote the statistics initially stored in the l^{th} batch normalization (BN) layer of the pre-trained model, which is calculated with the original training data. Given that Vision Transformers do not have the BN layer, [20] suggest that we can borrow the BN statistics stored in an open-source pre-trained ResNet50. Since $\mu_{\mathrm{BN}}^{(l)}$ and $\sigma_{\mathrm{BN}}^{(l)}$ is calculated with real data, minimizing gaps

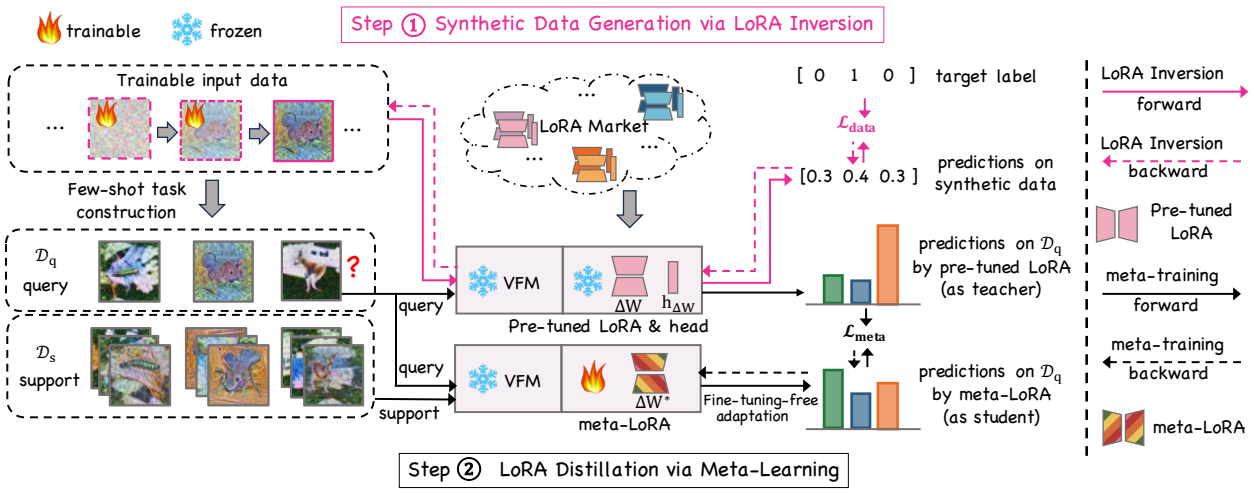

Figure 2. Pipeline of LoRA Recycle. **(i) (Pink Path)** We generate task-specific synthetic data from the pre-tuned LoRA via LoRA Inversion. The input data (attached with the fire in the left corner) is initialized as Gaussian noise and iteratively optimized by minimizing $\mathcal{L}_{\text{data}}$ (Eq. (1)). The synthetic data is then used to construct a meta-training task with one support set and one query set. **(ii) (Black Path)** We meta-train the meta-LoRA (attached with the fire in the middle) on a wide range of pre-tuned LoRAs by minimizing the meta-learning objective $\mathcal{L}_{\text{meta}}$ (Eq. (3)), explicitly teaching it how to adapt without fine-tuning.

in these statistics can align the distribution between the synthetic and real data, thus improving realism. Please refer to Fig. 5 in App. A for the ablation study.

Meta-training task construction. After generating task-specific data of task $\mathcal{T}$, we construct a few-shot task by splitting the synthetic data into one support set $\mathcal{D}_{\text{s}}^{\mathcal{T}}$ and one query set $\mathcal{D}_{\text{q}}^{\mathcal{T}}$. This constructed task serves as a meta-training task [14] for the subsequent meta-training process. In an N-way K-shot setup, the support set has N classes and K examples per class, while the query set has the same N classes but more examples per class, typically 15.

4.2. LoRA Distillation via Meta-Learning

Meta-learning objective. We distill a meta-LoRA δW^* from diverse pre-tuned LoRAs using the synthetic data. The meta-learning objective is formulated as follows:

$$\min_{\delta W^*} \mathcal{L}_{\text{meta}} = \mathbb{E}_{p_{\mathcal{T}}} \sum_{(\mathbf{X}_{\text{q}}, \mathbf{Y}_{\text{q}}) \in \mathcal{D}_{\text{q}}^{\mathcal{T}}} \text{KL}\left(P(\mathbf{Y}_{\text{pred}} | \mathbf{X}_{\text{q}}, \mathcal{D}_s^{\mathcal{T}}), h_{\mathcal{T}} \circ f_{\delta W_{\mathcal{T}}}(\mathbf{X}_{\text{q}})\right), \tag{3a}$$

$$\text{where,} \quad P(\mathbf{Y}_{\text{pred}} = i | \mathbf{X}_{\text{q}}, \mathcal{D}_{\text{s}}^{\mathcal{T}}) = \frac{\exp\left(-\| f_{\delta W^*}(\mathbf{X}_{\text{q}}) - \boldsymbol{c}_i \|_2\right)}{\sum_{i'} \exp\left(-\| f_{\delta W^*}(\mathbf{X}_{\text{q}}) - \boldsymbol{c}_{i'} \|_2\right)}. \tag{3b}$$

Here, $p_{\mathcal{T}}$ is the underlying task distribution. $(\delta W_{\mathcal{T}}, h_{\mathcal{T}}, \mathcal{D}_{\text{s}}^{\mathcal{T}}, \mathcal{D}_{\text{q}}^{\mathcal{T}})$ refer to the pre-tuned LoRA, classification head, synthetic support set, and query set of task $\mathcal{T}$, which can be viewed as sampling from the task distribution $p_{\mathcal{T}}$. The optimization in Eq. (3) involves one inner loop Eq. (3b) and one outer loop Eq. (3a).

- **Inner Loop:** We recast the inner loop Eq. (3b) as a tuning-free adaptation: we use the support set to calculate the class center $\boldsymbol{c}_i$ of each class i as the average feature embedding ($\boldsymbol{c}_i = \frac{1}{|\mathcal{D}_{\text{s},i}^{\mathcal{T}}|} \sum_{\mathbf{X} \in \mathcal{D}_{\text{s},i}^{\mathcal{T}}} f_{\delta W^*}(\mathbf{X})$). We then model the probability of a query example $\mathbf{X}_{\text{q}} \in \mathcal{D}_{\text{q}}^{\mathcal{T}}$ belonging to a class based on its Euclidean distance to the corresponding class center. This process does not involve any parameter updating, thus avoiding calculating any second-order derivatives [52].
- **Outer Loop:** In the outer loop Eq. (3a), we optimize the meta-LoRA δW^* so that it can make more accurate predictions in the inner loop across diverse tasks. Specifically, we minimize the prediction disagreements (*i.e.*, the Kullback-Leibler (KL) divergence) on the query set $\mathcal{D}_{\text{q}}^{\mathcal{T}}$ between the pre-tuned LoRA $\delta W_{\mathcal{T}}$ (as the teacher) and the meta-LoRA δW^* (as the student). The meta-LoRA δW^* is meta-trained across a wide range of pre-tuned LoRAs sampled from $p_{\mathcal{T}}$, explicitly learning to how to solve diverse tasks without fine-tuning.

Cross-task interpolation. Eq. (3) assumes a wide range of pre-tuned LoRAs sampled from the underlying task distribution $p_{\mathcal{T}}$, which is crucial for enhancing generalization of meta-learning. However, a fixed number of LoRAs might be insufficient to fully cover the support of $p_{\mathcal{T}}$, especially within the setting of limited LoRA budgets. Therefore, we propose cross-task interpolation to enhance the sampling density. We generate new tasks by combining classes from different synthetic tasks generated from pre-tuned LoRAs. For example, given LoRAs δW_i and δW_j tuned on tasks with classes (*husky*, *sparrow*) and (*golden retriever*, *wild horse*), an interpolated task might consist of data belong-

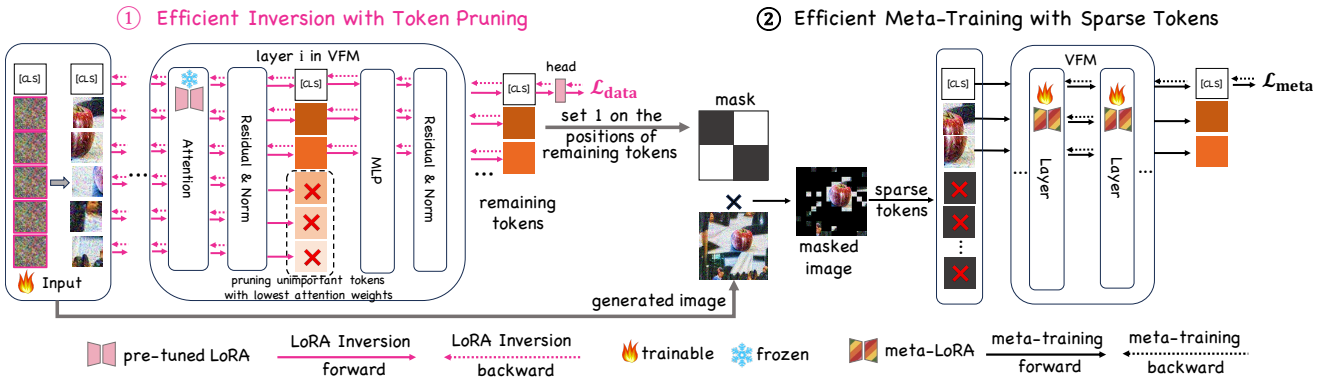

Figure 3. Double-Efficient Mechanism. **(Left: Efficient Data-Generation)** During the data-generation stage, token pruning is performed in the hidden layers by removing unimportant tokens based on self-attention weights, accelerating both forward and backward computations for reverse engineering. **(Right: Efficient Meta-Training)** To select the most informative tokens from the synthetic data for the following meta-training, we construct a mask by setting values of 1 at the positions of remaining tokens and 0 elsewhere. We multiply the mask with the synthetic image to create a masked image. We then exclusively use the unmasked tokens for meta-training. This selective use of sparse tokens significantly accelerates meta-training, while maintaining or even improving performance by reducing noise from the synthetic data.

ing to the classes (husky, golden retriever). This expands the diversity of tasks for meta-training, thus enhancing generalization. Since the interpolated task does not match the label spaces of any pre-tuned LoRAs, we modify Eq. (3a) by replacing the KL loss with the Cross Entropy (CE) loss:

$$\min_{\delta W^*} \mathbb{E}_{p_{\hat{\mathcal{T}}}} \sum_{(\mathbf{X}_\text{q},\mathbf{Y}_\text{q})\sim\mathcal{D}_\text{q}^{\hat{\mathcal{T}}}} \text{CE}\left(P(\mathbf{Y}_\text{pred}|\mathbf{X}_\text{q},\mathcal{D}_s^{\hat{\mathcal{T}}};\delta W^*),\mathbf{Y}_\text{q}\right), \tag{4}$$

where $p_{\hat{\mathcal{T}}}$ refers to the interpolated task distribution and $(\mathcal{D}_\text{s}^{\hat{\mathcal{T}}},\mathcal{D}_\text{q}^{\hat{\mathcal{T}}})$ refer to the synthetic support and query sets of the interpolated task $\hat{\mathcal{T}}$.

4.3. Double-Efficient Mechanism

Efficient inversion with token pruning. As shown in Eq. (1), the reverse engineering process for generating synthetic data involves optimizing $\mathbf{X}$ through iterative forward and backward computations. To improve efficiency, we introduce a token pruning strategy to remove unimportant image tokens during this process. This is justified by the self-attention mechanism, which inherently weights tokens according to their importance and relevance. As illustrated in the left panel of Fig. 3, at the i^{th} layer, we perform token pruning by discarding low-weight tokens, no longer processing them forward or computing backward gradients, thus significantly reducing computational complexity.

The most important tokens are those with highest attention weights in $\boldsymbol{a}_{\texttt{[CLS]}}$. Suppose we have $n+1$ tokens $[\boldsymbol{x}_{\texttt{[CLS]}},\boldsymbol{x}_1,...,\boldsymbol{x}_n]$ at the i^{th} layer, where $\boldsymbol{x}_{\texttt{[CLS]}}$ is the class token inserted before all image tokens to grasp global information. We propose to use the attention weights of the class token $\boldsymbol{x}_{\texttt{[CLS]}}$ with respect to all other tokens, as an indicator measuring each token's importance:

$$\boldsymbol{a}_{\texttt{[CLS]}} = \text{Softmax}\left(\frac{\boldsymbol{q}_{\texttt{[CLS]}}\cdot\boldsymbol{K}^\top}{\sqrt{d}}\right), \tag{5}$$

where $\boldsymbol{a}_{\texttt{[CLS]}}$ is a $(n+1)$-dimension vector, representing the attention weights from token $\boldsymbol{x}_{\texttt{[CLS]}}$ to all tokens $[\boldsymbol{x}_{\texttt{[CLS]}},\boldsymbol{x}_1,...,\boldsymbol{x}_n]$. $\boldsymbol{q}_{\texttt{[CLS]}}$ is the *query vector* of token $\boldsymbol{x}_{\texttt{[CLS]}}$. $\boldsymbol{K}=[\boldsymbol{k}_{\texttt{[CLS]}},\boldsymbol{k}_1,...,\boldsymbol{k}_n]^\top$ is the *key vectors* of all tokens. d is the dimension of the *query vector*. The $\boldsymbol{a}_{\texttt{[CLS]}}$ is then used to calculate the output of token $\boldsymbol{x}_{\texttt{[CLS]}}$ via the self-attention mechanism:

$$\boldsymbol{x}_{\texttt{[CLS]}} = \boldsymbol{a}_{\texttt{[CLS]}}\cdot\boldsymbol{V}, \tag{6}$$

where $\boldsymbol{V}=[\boldsymbol{v}_{\texttt{[CLS]}},\boldsymbol{v}_1,...,\boldsymbol{v}_n]^\top$ is the *value vectors* of all tokens. Therefore, the output of $\boldsymbol{x}_{\texttt{[CLS]}}$ can be viewed as a linear combination of all tokens' *value vectors* weighted by $\boldsymbol{a}_{\texttt{[CLS]}}$. Since the output of $\boldsymbol{x}_{\texttt{[CLS]}}$ is used for classification at the final layer, it is rational to view $\boldsymbol{a}_{\texttt{[CLS]}}$ as an indicator, measuring the extent to which each token contributes to final predictions, *i.e.*, the importance of each token. Therefore, we identify the most important tokens as those with the highest attention weights in $\boldsymbol{a}_{\texttt{[CLS]}}$. For multi-head self-attention, we compute average attention weights $\boldsymbol{a}_{\texttt{[CLS]}}$ across all heads. Note that this process requires no extra computational demands, as it is an inherent part of the forward process (see App. B for more preliminaries).

Efficient meta-training with sparse tokens. After inversion, we obtain the remaining tokens at the last layer. Since each token (except for token $\boldsymbol{x}_{\texttt{[CLS]}}$) precisely corresponds to a token in the input image, these remaining tokens can indicate the most informative regions in the synthetic data, guiding token selection for subsequent meta-training.

Algorithm 1: LoRA Recycle

INPUT The VFM f. Multiple pre-tuned LoRAs and classification heads. Coefficient $\alpha_{\mathcal{R}}$ in Eq. (1).
OUTPUT The meta-trained meta-LoRA δW^*
Randomly initialize the meta-LoRA δW^*
while *not done* **do**
  if not cross-task interpolation **then**
    Randomly sample a LoRA and head $(\delta W, h_{\delta W})$
    `// Synthetic data generation`
    Equip f with $(\delta W, h_{\delta W})$
    Generate synthetic data by minimizing Eq. (1)
    (Optional) Transform into masked versions
    Construct a meta-training task by splitting data to one support set and one query set $(\mathcal{D}_{\text{s}}^{\mathcal{T}}, \mathcal{D}_{\text{q}}^{\mathcal{T}})$
    `// LoRA distillation`
    Equip f with the meta-LoRA δW^*
    Make predictions on $\mathcal{D}_{\text{q}}^{\mathcal{T}}$ based on $\mathcal{D}_{\text{s}}^{\mathcal{T}}$ (Eq. (3b))
    Update δW^* by minimizing Eq. (3a)
  else
    `// Cross-task interpolation`
    Construct the interpolated task $(\mathcal{D}_s^{\hat{\mathcal{T}}}, \mathcal{D}_q^{\hat{\mathcal{T}}})$
    Equip f with the meta-LoRA δW^*
    Make predictions on $\mathcal{D}_{\text{q}}^{\hat{\mathcal{T}}}$ based on $\mathcal{D}_{\text{s}}^{\hat{\mathcal{T}}}$ (Eq. (3b))
    Update δW^* by minimizing Eq. (4)

Mask construction. To select the most informative tokens from the synthetic data for meta-training, we construct a mask matrix as shown in the right panel of Fig. 3. The mask matrix is constructed by setting values of 1 at the positions of remaining tokens and 0 elsewhere.

Selectively use unmasked tokens for meta-training. We multiply the mask with the synthetic image to create a masked image, reserving the most informative tokens (such as foregrounds) in the synthetic image. When meta-training the meta-LoRA, we only feed-forward the unmasked tokens. This selective use of sparse tokens significantly accelerates the meta-training process, while maintaining or even improving performance by reducing noise from the synthetic data (see Tab. 6 in App. A for analysis).

5. Experiments

In this section, we perform comprehensive experiments on various few-shot classification benchmarks, covering both in-domain (see Sec. 5.1) and cross-domain scenarios (see Sec. 5.2). We also provide comprehensive visualization results and ablation studies in Sec. 5.3 and App. A.

Setup of VFM. We select the 12-layer ViT-B/16 and ViT-B/32 pre-trained with CLIP as the pre-trained VFM, publicly available on HuggingFace. Refer to Tab. 13 in App. A for more results on more types of transformer.

Baselines. We compare LoRA Recycle against several baselines (see App. D for more implementation details).

- Multi-LoRAs reuse baselines. (a) LoRAs Avg averages all pre-tuned LoRAs into one, which is then either fine-tuned (LoRAs Avg + Linear) or used for Nearest Neighbor (LoRAs Avg + NN) inference. (b) LoRAHub [33] uses a weighted sum of pre-tuned LoRAs, with weights fine-tuned on the target task. (c) MOLE [6] fine-tunes a gating function to combine outputs from multiple LoRAs.
- Fine-tuning-free baselines. (d) Nearest Neighbor (NN) predicts based on the closest class center. (e) CAML [13] trains a sequence model to simulate in-context learning.
- Few-shot learning with foundation models. (f) P > M > F [27] is a state-of-the-art method that adapts foundation models to few-shot tasks through a pre-training, meta-training, and fine-tuning pipeline.
- Fine-tuning baselines. While our focus is on tuning-free settings, we include representative fine-tuning methods to demonstrate that tuning-free approaches can achieve comparable performance while offering advantages of stability and faster response. (g) Full Fine-Tuning updates the entire model, (h) Linear Probe updates only the classification head, and (i) LoRA + Linear [26] updates LoRA parameters alongside the classification head.

Implementation details. We fine-tune the LoRAs (rank $r = 4$) and classification heads using the Adam optimizer with a learning rate of 1×10^{-3}. LoRA performs well with sufficient data. During meta-training, the meta-LoRA is optimized with Adam at the same learning rate, following a cyclic schedule: a 25-iteration linear warm-up from 1×10^{-5} to 1×10^{-3}, followed by cosine annealing over the next 75 iterations. For LoRA Inversion, synthetic data is optimized using Adam with a learning rate of 0.25 over 2000 iterations, and synthetic images have a resolution of 224×224. We set the hyperparameter $\alpha_{\mathcal{R}} = 0.01$. Data augmentation includes random horizontal flipping and normalization in meta-training, with only normalization applied in meta-testing. Hyperparameter selections and sensitivity analysis are discussed in App. C. Unless otherwise specified, we perform token pruning in the final layer for the inversion stage to obtain masks with varying sparsity.

5.1. Recycle In-Domain LoRAs

In-domain benchmarks. For "recycle in-domain LoRAs" scenario, we collect 100 LoRAs pre-tuned on 100 5-way tasks constructed from a specific meta-training subset, including CIFAR-FS [2], MiniImageNet [62], VGG-Flower [53], or CUB [64]. Our evaluation, in contrast, is based on meta-testing tasks constructed from the corresponding meta-testing subset. Note that the meta-training and meta-testing subsets have non-overlapping label spaces but belong to the same dataset. This setup ensures the pre-tuned LoRAs and meta-testing tasks originate from the same domain but with non-overlapping label spaces.

Table 2. Recycle in-domain LoRAs. The VFM utilizes ViT-B/16 pre-trained by CLIP. **FT** refers to fine-tuning-based baselines and **FTF** refers to fine-tuning-free baselines. **LoRA Recycle**$_x$ indicates using x% token-masked images for meta-training. The superscripts represent performance gains over the best FT baselines, while the subscripts indicate gains over the best FTF baselines.

	Method	CIFAR-FS		MiniImageNet		VGG-Flower		CUB	
		5-way 1-shot	**5-way 5-shot**	**5-way 1-shot**	**5-way 5-shot**	**5-way 1-shot**	**5-way 5-shot**	**5-way 1-shot**	**5-way 5-shot**
FT	Full Finetuning	22.81	28.33	21.16	23.60	23.11	31.25	21.27	24.47
	Linear-probe	80.06	95.49	82.04	94.12	89.65	97.77	85.84	97.40
	LoRA + Linear	79.29	95.43	82.00	94.83	88.47	97.63	85.87	97.32
	P > M > F	79.54	95.62	82.77	95.12	89.32	97.65	86.12	97.38
	LoRAs Avg + Linear	80.25	96.07	83.59	95.43	90.05	97.73	87.13	97.49
	MOLE	80.31	96.11	83.53	95.41	90.14	97.68	87.07	97.21
	LoRAHub	81.23	96.24	83.68	95.72	90.89	97.75	87.22	97.51
FTF	NN	78.06	94.09	81.08	93.85	89.75	97.78	85.11	96.09
	LoRAs Avg + NN	79.37	93.45	81.72	94.64	90.08	97.92	85.16	97.23
	CMAL	81.02	93.59	81.89	94.81	91.10	97.98	86.51	97.32
	LoRA Recycle	89.69	**97.05** $^{(+0.81\%)}_{(+2.96\%)}$	**88.60** $^{(+4.92\%)}_{(+6.71\%)}$	96.12	**94.53** $^{(+3.64\%)}_{(+3.43\%)}$	98.59	91.12	97.67
	LoRA Recycle$_{25}$	**91.03** $^{(+9.80\%)}_{(+10.01\%)}$	96.53	87.51	**96.25** $^{(+0.53\%)}_{(+1.41\%)}$	94.38	98.53	90.16	97.48
	LoRA Recycle$_{50}$	90.91	96.08	87.21	95.85	94.05	98.56	90.65	97.41
	LoRA Recycle$_{75}$	89.70	96.69	87.36	96.05	94.28	**98.76** $^{(+0.99\%)}_{(+0.78\%)}$	**91.21** $^{(+3.99\%)}_{(+4.70\%)}$	**98.23** $^{(+0.72\%)}_{(+0.91\%)}$

Table 3. Recycle cross-domain LoRAs. The VFM utilizes ViT-B/16 pre-trained by CLIP.

	Method	ChestX		ISIC		EuroSAT		CropDiseases	
		5-way 1-shot	**5-way 5-shot**	**5-way 1-shot**	**5-way 5-shot**	**5-way 1-shot**	**5-way 5-shot**	**5-way 1-shot**	**5-way 5-shot**
FT	Full Finetuning	20.12	20.00	21.33	26.21	25.55	36.11	22.45	28.48
	Linear-probe	21.20	24.00	31.17	43.60	62.64	83.91	77.48	92.57
	LoRA + Linear	21.05	22.37	30.72	45.16	68.13	88.68	77.33	94.19
	P > M > F	21.12	22.21	30.77	45.54	68.51	88.71	77.65	94.21
	LoRAs Avg + Linear	21.37	20.84	30.51	45.88	68.77	88.29	78.75	94.37
	MOLE	21.24	20.67	30.61	45.79	68.84	88.42	78.81	94.40
	LoRAHub	21.45	22.61	32.11	46.12	69.45	**89.76**	79.32	94.44
FTF	NN	21.23	22.84	31.20	40.58	61.73	80.05	75.48	91.89
	LoRAs Avg + NN	20.80	23.04	29.67	39.56	62.52	78.87	78.91	91.57
	CMAL	21.26	23.24	29.97	41.27	67.69	83.87	79.71	93.38
	LoRA Recycle	22.32	24.61	33.76	47.96	66.95	85.17	83.07	95.33
	LoRA Recycle$_{25}$	22.77	24.88	33.64	**48.29** $^{(+2.41\%)}_{(+7.02\%)}$	67.65	84.73	82.41	95.40
	LoRA Recycle$_{50}$	**23.08** $^{(+1.63\%)}_{(+1.82\%)}$	**25.43** $^{(+1.43\%)}_{(+2.19\%)}$	**35.31** $^{(+3.20\%)}_{(+4.11\%)}$	47.41	**69.88** $^{(+0.43\%)}_{(+2.19\%)}$	86.72	**83.63** $^{(+4.31\%)}_{(+3.92\%)}$	**96.33** $^{(+1.89\%)}_{(+2.95\%)}$
	LoRA Recycle$_{75}$	22.99	24.91	35.16	48.25	68.00	87.98 $^{(+4.11\%)}$	81.92	95.64

Results in in-domain scenario. Tab. 2 shows the results for the "recycle in-domain LoRAs" scenario. Notable findings are as follows: (i) "LoRA Recycle" surpasses the best fine-tuning-based baselines by considerable margins, especially up to 9.80% for 1-shot learning, emphasizing its advantage of requiring no fine-tuning. It also outperforms the top fine-tuning-free baselines by up to 10.01% for 1-shot learning. (ii) "LoRA Recycle" also outperforms baselines that rely on real data (*i.e.*, CMAL and P > M > F) for few-shot generalization, highlighting its advantage of not requiring real data for meta-training. (iii) "Full fine-tuning" performs the worst, as it tends to overfit when tuning large models with extremely few examples. "LoRAs Avg" and "LoRAHub" do not ensure effective generalization to new tasks. The reason is that each pre-tuned LoRA targets different tasks, and the arithmetic operation like averaging in the parameter space lacks precise alignment among different LoRAs.

Performance gains of token pruning on synthetic data beyond acceleration. Meta-training with sparse tokens can bring performance gains up to 1.34%. This is because the inversion process typically crafts only label-relevant features (*i.e.*, the foreground) into the synthetic data, while other regions (*i.e.*, the background) often remain noisy as initialization (see Tab. 6 in App. A for evidence). Pruning background tokens helps reduce noise from the synthetic data, thereby enhancing meta-training effectiveness.

5.2. Recycle Cross-Domain LoRAs

Cross-domain benchmarks. Real-world situations might pose challenges in collecting LoRAs from the most relevant domain. For the "recycle cross-domain LoRAs" scenario, we collect 100 LoRAs pre-tuned on 100 5-way tasks constructed from four meta-training subsets, including CIFAR-FS, MiniImageNet, VGG-Flower and CUB. Our evaluation, in contrast, is based on meta-testing tasks from one specific cross-domain dataset (ChestX, ISIC, EuroSAT or CropDiseases). These meta-testing datasets cover diverse domains, including medical images (ChestX, ISIC) and specialized imagery (EuroSAT, CropDiseases), ensuring a distinct domain difference between meta-training and meta-testing. This ensures the pre-tuned LoRAs and meta-testing tasks originate from distinctly different domains (*i.e.*, datasets) and also with strictly non-overlapping label spaces.

Results in cross-domain scenario. Tab. 3 shows the results for the challenging "recycle cross-domain LoRAs". Existing methods struggle under these conditions, reflecting the inherent difficulty of this scenario. Despite this, LoRA Recycle achieves notable improvements, outperforming the

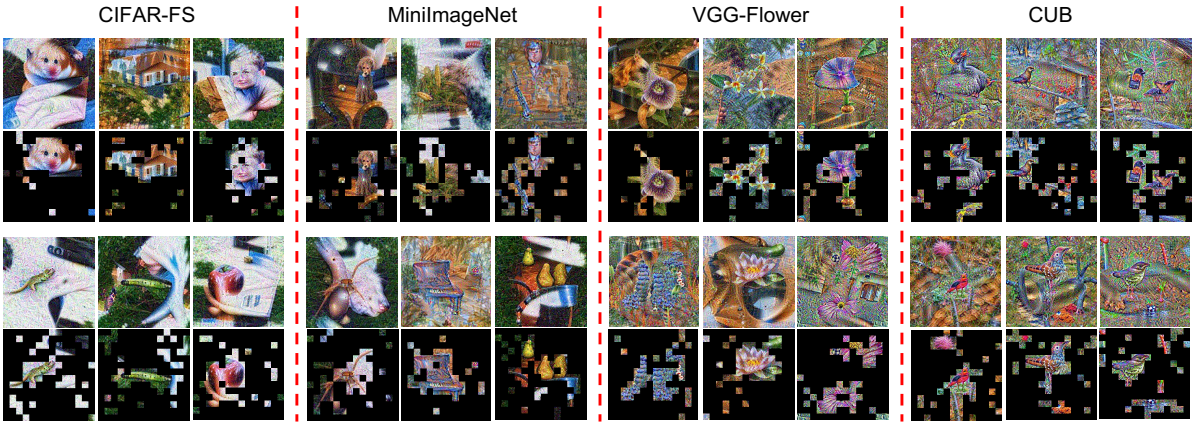

Figure 4. Visualization of synthetic images with their 75% token-masked versions.

Table 4. Complexity analysis of inversion. $\{\mathbf{x}: \mathrm{y}\}$ denotes pruning $(\mathrm{y} \times 100)\%$ tokens at the $\mathbf{x}^{\mathrm{th}}$ layer. Measurements are recorded during inversion with a batch size of 25 on CIFAR-FS.

Token Pruning Strategy	5w 1s	5w 5s	Throughput (its/s) ↑	FLOPs (G) ↓	GPU Mem (GB) ↓
{**0**: 0.0}	89.69	97.05	5.56	50.59	8.74
{**11**: 0.75}	89.43	96.72	5.81 (+4%)	48.51 (-4%)	8.63 (-1%)
{**8**: 0.75}	82.27	95.69	6.22 (+12%)	39.14 (-23%)	8.07 (-8%)
{**6**: 0.75}	81.08	95.52	7.15 (+29%)	32.89 (-35%)	7.69 (-12%)
{**3**: 0.3, **6**: 0.3, **8**: 0.3, **11**: 0.3}	84.17	96.12	6.13 (+10%)	40.00 (-21%)	8.08 (-8%)

Table 5. Effect of sparse ratio of the synthetic data for the performance and complexity of meta-training. Measurements are recorded during meta-training with a batch size of 100 on CUB.

Sparse ratio	5w 1s	5w 5s	Throughput (its/s) ↑	FLOPs (G) ↓	GPU Mem (GB) ↓
LoRA Recycle	91.12	97.67	1.76	50.59	12.86
LoRA Recycle$_{25}$	90.16	97.48	2.34 (+33%)	38.09 (-25%)	9.40 (-27%)
LoRA Recycle$_{50}$	90.65	97.41	3.63 (+106%)	25.60 (-49%)	6.23 (-52%)
LoRA Recycle$_{75}$	91.21	98.23	6.83 (+287%)	13.10 (-74%)	3.31 (-74%)

best fine-tuning-based baselines by up to 4.31% and 2.41% for 1-shot and 5-shot learning, respectively. It also exceeds the top tuning-free baselines by up to 4.11% and 7.02% for 1-shot and 5-shot learning, confirming its enhanced cross-domain robustness. However, while LoRA Recycle greatly advances performance, there remains room for improvement when facing substantial distribution shifts.

5.3. Ablation Studies

How to choose the overall pruning ratio and pruning layers during inversion? We adopt the double-efficient mechanism with two steps: (i) Set the overall pruning ratio (sparsity level) of the synthetic data. As shown in Tab. 2 and Tab. 3, pruning 50% or 75% of tokens boosts performance while preserving key foregrounds (see Fig. 4). (ii) Select the pruning layers during inversion.For instance, pruning 75% of tokens at a single layer is nearly equivalent to pruning 30% across four layers. Notably, given a fixed overall pruning ratio, the choice of pruning layers impacts only the inversion speed, whereas meta-training speed depends solely on the overall pruning ratio. The choice of pruning layers is flexible and depends on needs: with the same overall pruning ratio, Tab. 4 suggests pruning at deeper layers for better performance, or shallower layers for faster inversion. The choice of middle-layer pruning or multi-layer pruning across shallow and deep layers can balance the trade-off between efficiency and performance to some extent.

Effect of the overall sparse ratio for meta-learning. As shown in Tab. 5, discarding 75% of tokens in the synthetic data significantly accelerates meta-training by up to 3× and yields performance gains of up to +0.56% on CUB by reducing noise from the synthetic data. This benefit is even more pronounced on CIFAR-FS, where removing 25% of the tokens results in performance gains of up to +1.34%.

Visualization. As shown in Fig. 4, the synthetic data effectively retains semantic foregrounds while filtering out noisy backgrounds, resulting in high-resolution 224 × 224 images better than existing methods (see Fig. 8 in App. A).

More ablation studies are provided in App. A, including ablations on the meta-learning objective (Tab. 8), synthetic data (Tab. 9), and cross-task interpolation (Tab. 7).

6. Conclusion

In this paper, we propose LoRA Recycle, a novel meta-learning framework to achieve tuning-free few-shot adaptation in VFMs, by reusing diverse pre-tuned LoRAs without access to original training data. Additionally, we propose a double-efficient mechanism that accelerates both the data-generation and meta-training processes, while maintaining or even improving performance by reducing potential noise in the synthetic data. Experimental results across various few-shot classification benchmarks, across in-domain and challenging cross-domain scenarios, confirm the effectiveness of LoRA Recycle. Future work could explore the black-box data-free adaptation setting in the context of potential defense mechanisms [71, 72].

Acknowledgement

This project is supported by the National Research Foundation, Singapore, under its NRF Professorship Award No. NRF-P2024-001.

References

[1] Trapit Bansal, Salaheddin Alzubi, Tong Wang, Jay-Yoon Lee, and Andrew McCallum. Meta-adapters: Parameter efficient few-shot fine-tuning through meta-learning. In *International Conference on Automated Machine Learning*, pages 19–1. PMLR, 2022. 2, 18

[2] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018. 6

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 3

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[5] John Cai and Sheng Mei Shen. Cross-domain few-shot learning with meta fine-tuning. *arXiv preprint arXiv:2005.10544*, 2020. 18

[6] Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*, 2024. 2, 6, 16

[7] Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, Dublin, Ireland, 2022. Association for Computational Linguistics. 3, 18

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 18

[9] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*, 2022. 3

[10] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 3

[11] Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive model inversion for data-free knowledge distillation. *arXiv preprint arXiv:2105.08584*, 2021. 15, 17, 18

[12] Gongfan Fang, Kanya Mo, Xinchao Wang, Jie Song, Shitao Bei, Haofei Zhang, and Mingli Song. Up to 100x faster data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6597–6604, 2022. 18

[13] Christopher Fifty, Dennis Duan, Ronald G Junkins, Ehsan Amid, Jure Leskovec, Christopher Ré, and Sebastian Thrun. Context-aware meta-learning. *arXiv preprint arXiv:2310.10971*, 2023. 3, 6, 17

[14] 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. 2, 3, 4, 18

[15] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24575–24584, 2023. 2

[16] Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12799–12807, 2023. 18

[17] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 2

[18] Mozhdeh Gheini, Xuezhe Ma, and Jonathan May. Know where you're going: Meta-learning for parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11602–11612, Toronto, Canada, 2023. Association for Computational Linguistics. 18

[19] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023. 2

[20] Ali Hatamizadeh, Hongxu Yin, Holger R Roth, Wenqi Li, Jan Kautz, Daguang Xu, and Pavlo Molchanov. Gradvit: Gradient inversion of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2022. 3

[21] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 2

[22] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022. 2

[23] Junyuan Hong, Yi Zeng, Shuyang Yu, Lingjuan Lyu, Ruoxi Jia, and Jiayu Zhou. Revisiting data-free knowledge distillation with poisoned teachers. In *International Conference on Machine Learning*, pages 13199–13212. PMLR, 2023. 18

[24] Zejiang Hou, Julian Salazar, and George Polovets. Meta-learning the difference: preparing large language models for efficient adaptation. *Transactions of the Association for Computational Linguistics*, 10:1249–1265, 2022. 18

[25] 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. 2

[26] 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, 2, 3, 6, 16

[27] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9068–9077, 2022. 6, 16, 18

[28] Zixuan Hu, Yongxian Wei, Li Shen, Zhenyi Wang, Lei Li, Chun Yuan, and Dacheng Tao. Sparse model inversion: Efficient inversion of vision transformers for data-free applications. In *Forty-first International Conference on Machine Learning*. 18

[29] Zixuan Hu, Li Shen, Shenqi Lai, and Chun Yuan. Task-adaptive feature disentanglement and hallucination for few-shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):3638–3648, 2023. 2

[30] Zixuan Hu, Li Shen, Zhenyi Wang, Tongliang Liu, Chun Yuan, and Dacheng Tao. Architecture, dataset and model-scale agnostic data-free meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[31] Zixuan Hu, Li Shen, Zhenyi Wang, Yongxian Wei, Baoyuan Wu, Chun Yuan, and Dacheng Tao. Task-distributionally robust data-free meta-learning. *arXiv preprint arXiv:2311.14756*, 2023.

[32] Zixuan Hu, Li Shen, Zhenyi Wang, Baoyuan Wu, Chun Yuan, and Dacheng Tao. Learning to learn from apis: Black-box data-free meta-learning. *arXiv preprint arXiv:2305.18413*, 2023. 2

[33] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023. 2, 6, 16

[34] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10878–10887, 2023. 18

[35] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022. 18

[36] Zeyinzi Jiang, Chaojie Mao, Ziyuan Huang, Yiliang Lv, Deli Zhao, and Jingren Zhou. Rethinking efficient tuning methods from a unified perspective. *arXiv preprint arXiv:2303.00690*, 2023. 2

[37] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. *Advances in neural information processing systems*, 34: 18590–18602, 2021. 14

[38] Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[39] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1383–1392, 2024. 13

[40] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 1

[41] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. 13

[42] Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. *International Conference on Learning Representations*, 2023. 18

[43] Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *European Conference on Computer Vision*, pages 154–170, 2022. 18

[44] Zhikai Li, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Psaq-vit v2: Toward accurate and general data-free quantization for vision transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 18

[45] Zhiwei Lin, Yongtao Wang, and Zhi Tang. Training-free open-ended object detection and segmentation via attention as prompts. *arXiv preprint arXiv:2410.05963*, 2024.

[46] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022. 2

[47] He Liu, Yikai Wang, Huaping Liu, Fuchun Sun, and Anbang Yao. Small scale data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6008–6016, 2024. 18

[48] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[49] Kangyang Luo, Shuai Wang, Yexuan Fu, Xiang Li, Yunshi Lan, and Ming Gao. DFRD: Data-free robustness distillation for heterogeneous federated learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 18

[50] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States, 2022. Association for Computational Linguistics. 3, 18

[51] Ivona Najdenkoska, Xiantong Zhen, and Marcel Worring. Meta learning to bridge vision and language models for multimodal few-shot learning. In *The Eleventh International Conference on Learning Representations*, 2023. 18

[52] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 4

[53] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6

[54] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 13

[55] Renrong Shao, Wei Zhang, Jianhua Yin, and Jun Wang. Data-free knowledge distillation for fine-grained visual categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1515–1525, 2023. 18

[56] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 15

[57] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 403–412, 2019. 18

[58] 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. 14

[59] Minh-Tuan Tran, Trung Le, Xuan-May Le, Jianfei Cai, Mehrtash Harandi, and Dinh Phung. Large-scale data-free knowledge distillation for imagenet via multi-resolution data generation. *arXiv preprint arXiv:2411.17046*, 2024. 18

[60] Minh-Tuan Tran, Trung Le, Xuan-May Le, Mehrtash Harandi, Quan Hung Tran, and Dinh Phung. Nayer: Noisy layer data generation for efficient and effective data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23860–23869, 2024. 18

[61] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Jordan Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*, 2020. 14

[62] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 6

[63] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023. 3

[64] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. 2011. 6

[65] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022. 18

[66] Yuzheng Wang, Dingkang Yang, Zhaoyu Chen, Yang Liu, Siao Liu, Wenqiang Zhang, Lihua Zhang, and Lizhe Qi. De-confounded data-free knowledge distillation for handling distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12615–12625, 2024. 18

[67] Zhenyi Wang, Tiehang Duan, Le Fang, Qiuling Suo, and Mingchen Gao. Meta learning on a sequence of imbalanced domains with difficulty awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8947–8957, 2021. 2

[68] Zhenyi Wang, Li Shen, Tiehang Duan, Donglin Zhan, Le Fang, and Mingchen Gao. Learning to learn and remember super long multi-domain task sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7992, 2022.

[69] Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Donglin Zhan, Tiehang Duan, and Mingchen Gao. Meta-learning with less forgetting on large-scale non-stationary task distributions. In *European Conference on Computer Vision*, pages 221–238. Springer, 2022. 2

[70] Zhenyi Wang, Xiaoyang Wang, Li Shen, Qiuling Suo, Kaiqiang Song, Dong Yu, Yan Shen, and Mingchen Gao. Meta-learning without data via wasserstein distributionally-robust model fusion. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. 2

[71] Zhenyi Wang, Li Shen, Tongliang Liu, Tiehang Duan, Yanjun Zhu, Donglin Zhan, David Doermann, and Mingchen Gao. Defending against data-free model extraction by distributionally robust defensive training. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 8

[72] Zhenyi Wang, Yihan Wu, and Heng Huang. Defense against model extraction attack by bayesian active watermarking. In *Forty-first International Conference on Machine Learning*, 2024. 8

[73] Yongxian Wei, Zixuan Hu, Li Shen, Zhenyi Wang, Lei Li, Yu Li, and Chun Yuan. Meta-learning without data via unconditional diffusion models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[74] Yongxian Wei, Zixuan Hu, Li Shen, Zhenyi Wang, Yu Li, Chun Yuan, and Dacheng Tao. Task groupings regularization: Data-free meta-learning with heterogeneous pre-trained models. In *Forty-first International Conference on Machine Learning*, 2024.

[75] Yongxian Wei, Zixuan Hu, Zhenyi Wang, Li Shen, Chun Yuan, and Dacheng Tao. Free: Faster and better data-free meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23273–23282, 2024. 2

[76] Yongxian Wei, Zixuan Hu, Li Shen, Zhenyi Wang, Chun Yuan, and Dacheng Tao. Open-vocabulary customization from CLIP via data-free knowledge distillation. In *The Thirteenth International Conference on Learning Representations*, 2025. 18

[77] Chengyue Wu, Teng Wang, Yixiao Ge, Zeyu Lu, Ruisong Zhou, Ying Shan, and Ping Luo. pi-tuning: Transferring multimodal foundation models with optimal multi-task interpolation. In *International Conference on Machine Learning*, pages 37713–37727. PMLR, 2023. 2

[78] Xun Wu, Shaohan Huang, and Furu Wei. Mole: Mixture of lora experts. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[79] Chun-Hsiao Yeh, Bryan Russell, Josef Sivic, Fabian Caba Heilbron, and Simon Jenni. Meta-personalizing vision-language models to find named instances in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19123–19132, 2023. 18

[80] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020. 3, 15, 18

[81] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018. 2

[82] Shikang Yu, Jiachen Chen, Hu Han, and Shuqiang Jiang. Data-free knowledge distillation via feature exchange and activation region constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24266–24275, 2023. 18

[83] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. In *The Twelfth International Conference on Learning Representations*, 2024. 3