6-2024

Few-shot learner parameterization by diffusion time-steps

Zhongqi YUE

Pan ZHOU
Singapore Management University, panzhou@smu.edu.sg

Richang HONG

Hanwang ZHANG

SUN Qianru

Citation

YUE, Zhongqi; ZHOU, Pan; HONG, Richang; ZHANG, Hanwang; and SUN Qianru. Few-shot learner parameterization by diffusion time-steps. (2024). *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition Conference (CVPR), Seattle, 2024 June 17-21*. 23263-23272.
Available at: https://ink.library.smu.edu.sg/sis_research/9019

Few-shot Learner Parameterization by Diffusion Time-steps

Zhongqi Yue[1,3], **Pan Zhou**[2,3], **Richang Hong**[4], **Hanwang Zhang**[5,1], **Qianru Sun**[2]
[1]Nanyang Technological University, [2]Singapore Management University, [3]Sea AI Lab,
[4]Hefei University of Technology, [5] Skywork AI
zhongqi.yue@ntu.edu.sg, panzhou@smu.edu.sg, hongrc.hfut@gmail.com,
hanwangzhang@ntu.edu.sg, qianrusun@smu.edu.sg

Abstract

Even when using large multi-modal foundation models, few-shot learning is still challenging—if there is no proper inductive bias, it is nearly impossible to keep the nuanced class attributes while removing the visually prominent attributes that spuriously correlate with class labels. To this end, we find an inductive bias that the time-steps of a Diffusion Model (DM) can isolate the nuanced class attributes, i.e., as the forward diffusion adds noise to an image at each time-step, nuanced attributes are usually lost at an earlier time-step than the spurious attributes that are visually prominent. Building on this, we propose Time-step Few-shot (TiF) learner. We train class-specific low-rank adapters for a text-conditioned DM to make up for the lost attributes, such that images can be accurately reconstructed from their noisy ones given a prompt. Hence, at a small time-step, the adapter and prompt are essentially a parameterization of only the nuanced class attributes. For a test image, we can use the parameterization to only extract the nuanced class attributes for classification. TiF learner significantly outperforms OpenCLIP and its adapters on a variety of fine-grained and customized few-shot learning tasks. Codes are in https://github.com/yue-zhongqi/tif.

1. Introduction

Multi-modal foundation models, *e.g.*, CLIP [26], have recently demonstrated remarkable zero-shot performance in general visual classification tasks [3, 25]. They define a task by a suitable text prompt for each class (*e.g.*, "a photo of an aircraft" for the class "aircraft"), and use the model to classify an image by selecting the class prompt with the highest "image-prompt" similarity. However, the zero-shot paradigm encounters limitations when one cannot find a proper prompt to accurately describe the class of interest. This is particularly true for niche and fine-grained categories whose names may not encapsulate the subtle visual features recognizable by the model (*e.g.*, "707-320 aircraft"), or for customized categories where it is impractical to fully describe their specification through text alone (*e.g.*, a specific person). For these situations, a few-shot training set is necessary to define the classification on demand.

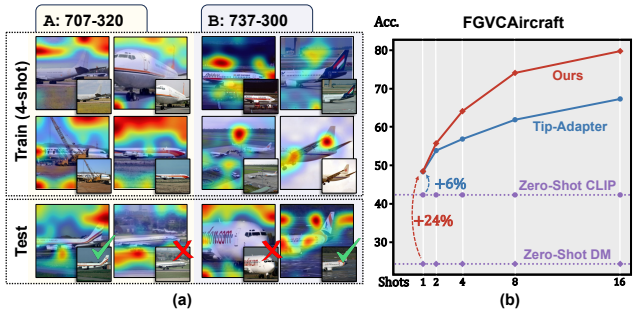

Figure 1. (a) GradCAM [34] of Tip-Adapter [44] on a 4-shot learning task from FGVCAircraft [22], where it is biased to the spurious background. (b) Comparison of few-shot learning performance. Our DM-based method significantly outperforms zero-shot OpenCLIP [14] (ViT-H/14 trained on LAION-2B [33]) and its adapter.

The most common approach is to train an auxiliary feature adapter attached after the CLIP visual encoder [8, 44] or text encoder [47, 48]. The objective is to align the adapted image feature with its prompt embedding only from the same class. However, it is well-known that such discriminative training on few-shot examples is biased to spurious correlations [41], which hinders generalization at test time. For example, in Figure 1, consider a scenario where aircraft A and B is defined by a subtle class attribute, *e.g.*, *window*, but a visually prominent attribute happens to spuriously correlate with the class labels in the few-shot training set, *i.e.*, A appearing with background *sky* and B with *ground*. The model will erroneously consider both *window* and background attributes for inaccurate prediction, *e.g.*, predicting B with *sky* as A. Frustratingly, it is impossible to isolate the nuanced class attributes from visually prominent yet spurious ones without an explicit inductive bias (*e.g.*, prior knowledge that "windows" is a class attribute) [5, 21].

To address the challenge, we turn our attention to gener-

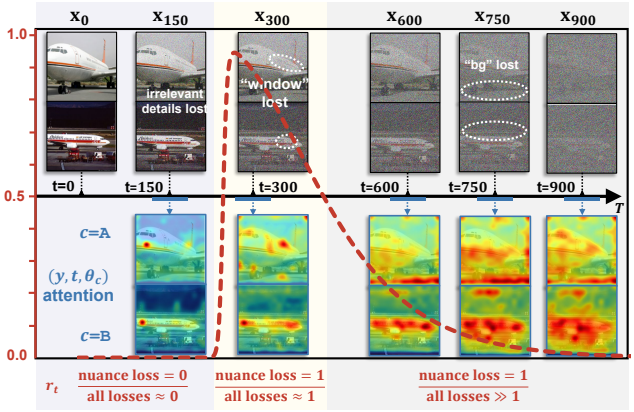

Figure 2. Top: DM forward process with attribute loss examples. Bottom: Attention map of what (y, t, θ_c) parameterizes for $c = \texttt{A}$ or $\texttt{B}$, which includes only nuances at a small t, and expands as t increases when more attributes are lost. We follow [23] to compute the average attention over a small time-step range indicated by the blue line. Details in Appendix. Red: Our proposed hyper-parameter-free weights for all time-steps.

ative classifiers [24], inspired by the recent advancement in the text-conditioned generative Diffusion Model (DM) [28]. DM has enabled extrapolation to novel attribute combinations by textual control (*e.g.*, generating "Salvador Dalí with a robotic half-face"). Hence ideally, one could prompt DM with all attribute combinations for each class (*e.g.*, "$\texttt{B}$ with sky" and "$\texttt{B}$ with ground") to synthesize a diverse training set, where no attribute is biased. Unfortunately, this approach is impractical yet, since there is no ground-truth of a complete attribute inventory.

In this paper, we introduce a practical solution by revealing that nuanced class attributes and visually prominent ones are naturally isolated by diffusion time-steps, lending itself to de-biasing. Specifically, DM defines a forward process that gradually injects Gaussian noise into each image $\mathbf{x}_0$ over T time-steps, creating a sequence of noisy images $\mathbf{x}_1, \dots, \mathbf{x}_T$ that progressively collapse to pure noise. As t increases, we show in Figure 2 (top) that more visual attributes are *lost* when they become indistinguishable. Notably, we prove in Section 3.3 that nuanced class attributes (*e.g.*, windows defining class $\texttt{A}$, $\texttt{B}$) are lost at an early time-step, while visually prominent ones (*e.g.*, the spurious background) are lost later. This motivates our approach:

(I) Parameterization. For each class c, we train a denoising network d parameterized by θ_c to reconstruct each image $\mathbf{x}_0$ in its training set from the noisy $\mathbf{x}_t$ and a text prompt y (details given below), *i.e.*, by minimizing the reconstruction error $\|d(\mathbf{x}_t, y, t; \theta_c) - \mathbf{x}_0\|^2$ for all t. When training converges with accurate reconstruction, $d(\cdot; \theta_c)$ must make up for the lost attributes in each $\mathbf{x}_t$ using y. Hence (y, t, θ_c) is essentially a *parameterization* of the lost attributes at t for class c. Particularly, we adopt a low-rank implementation of θ_c, such that (y, t, θ_c) at a small t is constrained to parameterize only the lost nuances of c, without accidentally capturing the visually prominent attributes that are not lost yet. We visualize this in Figure 2 bottom by attention map, *e.g.*, (y, t, θ_c) indeed parameterizes the visual attribute about window at $t = 300$ when it becomes lost.

(II) De-biasing by Time-steps. We classify a test image $\mathbf{x}_0$ by selecting c with the least reconstruction error $\|d(\mathbf{x}_t, y, t; \theta_c) - \mathbf{x}_0\|^2$ at a *small time-step* t. This inference rule is not affected by the spurious correlations, because from the analysis above, such (y, t, θ_c) parameterize *only* the nuanced attributes that define $\mathbf{x}_0$'s class, free from the visually prominent ones that are spurious.

We term our approach as the Time-step Few-shot (**TiF**) learner. We implement the above two points as below:

(I) We inject Low-Rank Adaptation (LoRA) matrices [13] into a pre-trained, frozen DM and denote their parameters as θ_c. We follow DreamBooth [30] to use a rare token identifier [V] (*e.g.*, "hta") to form the prompt y, such that it can be easily re-associated to the specificity of each class.

(II) To avoid searching for the best "small t", we design a hyper-parameter-free approach. We first derive an equation that measures the degree of attribute loss in Section 3.3. Then we compute a ratio r_t of nuanced attribute loss over all attribute losses at each t to calculate a weighted reconstruction error $\sum_t r_t \|d(\mathbf{x}_t, y, t; \theta_c) - \mathbf{x}_0\|^2$ for inference. As shown in Figure 2 red line, the weight reaches its peak when nuances become lost at a small t, and diminishes to 0 when more attributes become lost as t increases. Hence the weighted error accounts only for the inability of each $d(\cdot; \theta_c)$ in making up the nuanced class attribute to de-bias.

The contribution of the paper is summarized below.

- We formulate few-shot learning (FSL) within the context of recent advancements in foundation models (Section 3.1), and introduce a theoretical framework that isolates nuanced attributes from visually prominent ones by diffusion time-steps (Section 3.3).
- Motivated by our theoretical insights, we present a straightforward yet effective FSL approach that mitigates spurious correlations (Section 4).
- On fine-grained classification, Re-Identification and medical image classification, TiF learner significantly outperforms the powerful OpenCLIP and its adapters in various few-shot settings by up to 21.6%.

2. Related Works

Conventional FSL typically adopts a pre-training, meta-learning and fine-tuning paradigm [2, 43]. The first stage aims to capture rich prior knowledge as a feature backbone [6]. The second stage trains the model on "sandbox" FSL tasks to tailor it for the target task, *e.g.*, learning a classifier weight generator [9], a distance kernel func-

tion in k-NN [39], a feature space to better seprate the classes [39, 43], or even an initialization of the classifier [7]. The final stage involves training a classifier on the few-shot examples. However, multi-modal foundation models already capture extremely profound prior knowledge, hence recent works focus on few-shot adapting such models.

FSL with Foundation Models. There are two main approaches that both leverage CLIP. First is prompt tuning, which aims to learn a prompt for each class. CoOp [48] learns a continuous prompt embedding instead of using a hand-crafted prompt. CoCoOp [47] extends CoOp by learning an image conditional prompt. ProGrad [49] aligns the prompt gradient to the general knowledge of CLIP. Recent MaPLe [15] additionally fine-tune the CLIP visual encoder. The other line aims to learn a CLIP visual feature adapter. CLIP-Adapter [8] applies a lightweight residual-style adapter, followed by the training-free approach Tip-Adapter [44]. CALIP [10] proposes a parameter-free attention to improve both zero-shot and few-shot performance. The recent CaFo [45] ensembles multiple foundation models to help with feature adaptation. However, they still suffer from the spurious correlation. Besides CLIP-based approaches, recent works have explored in-context learning with vision-language models [1], yet their current classification accuracy still lags behind.

Alleviating Spurious Correlation. Previous works use knowledge from additional data. For example, IFSL [41] leverages the data in pre-training, or unsupervised domain adaptation use unlabeled data in test domain [18, 37, 42]. We leverage the time-steps of DM without such data.

3. Problem Formulations

3.1. Few-Shot Learning

We aim to solve a K-way-N-shot Few-Shot Learning (FSL) task: train a model to classify K categories using the few-shot dataset $\mathcal{D}$, where each category $c \in \{1, \dots, K\}$ has a small number of N images. In particular, we use the notation in causal representation learning [11, 32]: each image $\mathbf{x}$ from c is generated by $\Phi(\mathbf{c}, \mathbf{e})$, where Φ is the generator, $\mathbf{c}$ denotes the class attribute that define the category c (*e.g.*, "many windows" for class "707-320"), and $\mathbf{e}$ denotes other environmental attribute (*e.g.*, "sky background"). Note that c is a category index, and $\mathbf{c}$ denotes its defining attribute. Hence the crux of FSL is to pinpoint $\mathbf{c}$ for classification and discard the irrelevant $\mathbf{e}$.

Necessity of FSL. The zero-shot accuracy of multi-modal foundation models matches that of a fully-supervised model on common visual categories [3, 25] (*e.g.*, ImageNet [31]), rendering FSL unnecessary on those tasks. Specifically, such model takes an image $\mathbf{x}$ and a text prompt y_c that describes c as input, and outputs the similarity between $\mathbf{x}$ and c. For example, $\text{CLIP}(\mathbf{x}, c) := \cos\left(V(\mathbf{x}), T(y_c)\right)$, where $\cos(\cdot, \cdot)$ is the cosine similarity, V, T denotes the CLIP visual and text encoder, respectively. Therefore we can predict an image $\mathbf{x}$ in a zero-shot setting as follows

$$\hat{c} = \underset{c \in \{1,\dots,K\}}{\arg\max}\ \text{CLIP}(\mathbf{x}, c). \tag{1}$$

Hence we focus on scenarios where the zero-shot paradigm is limited—*when* $\mathbf{c}$ *is about nuanced details* on fine-grained or customized categories, it is impractical to find a prompt y_c recognizable by the model that encapsulates the intricacies of $\mathbf{c}$, thereby requiring a few-shot set to define c.

Challenge of FSL. The most direct approach appends a trainable network parameterized by θ to the CLIP visual encoder V, denoted as $V(\cdot; \theta)$, and optimizes:

$$\min_{\theta} \sum_{(\mathbf{x},c)\in\mathcal{D}} \frac{\cos\left(V(\mathbf{x};\theta), T(y_c)\right)}{\sum_{c'=1}^{K} \cos\left(V(\mathbf{x};\theta), T(y_{c'})\right)}, \tag{2}$$

which trains $V(\mathbf{x}; \theta)$ to predict its class prototype $T(y_c)$. However, such discriminative training on few-shot examples is easily biased to the spurious correlation between $\mathbf{e}$ and $\mathbf{c}$. Considering an extreme one-shot case, $\mathbf{x} = \Phi(\mathbf{c}, \mathbf{e})$ and $\mathbf{x}' = \Phi(\mathbf{c}', \mathbf{e}')$ from another class differ by $\mathbf{e} \neq \mathbf{e}'$. $V(\mathbf{x}; \theta)$ will inevitably mistake $\mathbf{e}, \mathbf{e}'$ as part of the class attribute (*e.g.*, predicting with both "window" and "background" in Figure 1), hence fail to reliably classify images not following the spurious pattern, *i.e.*, $\Phi(\mathbf{c}, \mathbf{e}')$ or $\Phi(\mathbf{c}', \mathbf{e})$.

Time-step Prior and Limitation. Our proposed TiF learner aims to circumvent the above challenge with the prior of Diffusion Model (DM) [12] time-steps: When the spurious $\mathbf{e}$ *has a larger pixel-level impact* (*i.e.*, visually prominent) than the nuanced class attribute $\mathbf{c}$, we can leverage the time-steps, introduced in Section 3.2, to isolate $\mathbf{c}$ at a small time-step in Section 3.3. However there are two limiting cases: *1) when a fine-grained* $\mathbf{e}$ *spuriously correlates with* $\mathbf{c}$, it will also be isolated at a small time-step, hence requiring additional prior to remove it; *2) when a class attribute* $\mathbf{c}$ *is coarse-grained*, this becomes a hierarchical classification task out of this paper's scope. Note that this is also unlikely on our evaluation datasets (Section 5.1).

3.2. Diffusion Model

DM is a generative model that first adds noise to images, and then learns to reconstruct them by a denoising network.

Forward Process. It adds Gaussian noise to each image $\mathbf{x}_0$ in T time-steps, producing a sequence of noisy images $\mathbf{x}_1, \dots, \mathbf{x}_T$, with the subscript denoting the time-step. Given $\mathbf{x}_0$ and a variance schedule $\beta_1, \dots, \beta_T$ (*i.e.*, how much noise is added at each time-step), $\mathbf{x}_t$ adheres to the following noisy sample distribution:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}), \tag{3}$$

where $\alpha_t := 1 - \beta_t,\ \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

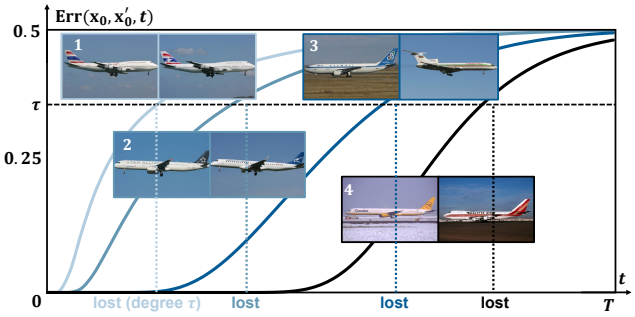

Figure 3. Plot of $\mathrm{Err}(\mathbf{x}_0, \mathbf{x}'_0, t)$ on 4 pairs of $(\mathbf{x}_0, \mathbf{x}'_0)$ with different pixel-level differences. We observe that the attribute loss for each pair is strictly increasing in t, and the fine-grained attribute that distinguishing more similar image pair is lost earlier.

Training. DM learns a denoising network d by first sampling a noisy image $\mathbf{x}_t$ at a random time-step t from the distribution in Eq. (3), and then training d to minimize the loss $\mathcal{L}_t$ of reconstructing $\mathbf{x}_0$ from $\mathbf{x}_t$:

$$\mathcal{L}_t(d, \mathbf{x}_0, y) = w_t \mathop{\mathbb{E}}_{q(\mathbf{x}_t|\mathbf{x}_0)} \|\mathbf{x}_0 - d(\mathbf{x}_t, y, t)\|^2, \quad (4)$$

where w_t is a standard weight (see Appendix), and y is an optional condition, *e.g.*, for an unconditional DM [12], $y = \emptyset$; for a text-conditioned DM [28], y is a text prompt describing $\mathbf{x}_0$. Next, we show that nuanced $\mathbf{c}$ can be separated from visually prominent $\mathbf{e}$ by diffusion time-steps.

3.3. Theory

We prove that in the forward diffusion process, each fine-grained $\mathbf{c} \in \mathcal{C}$ is lost at an earlier time-step compared to more coarse-grained $\mathbf{e} \in \mathcal{E}$, where $\mathcal{C}, \mathcal{E}$ denotes the set of all class attributes and environmental ones, respectively, and the granularity is defined by the pixel-level changes when altering an attribute. We first formalize the notion of attribute loss, *i.e.*, when an attribute becomes indistinguishable in the noisy images sampled from $q(\mathbf{x}_t|\mathbf{x}_0)$.

Definition. (Attribute Loss) *Without loss of generality, we say that the attribute $\mathbf{c} \in \mathcal{C}$ is lost with degree τ at t when $\mathbb{E}_{(\mathbf{c}',\mathbf{e})\in\mathcal{C}\times\mathcal{E}}\left[\mathrm{Err}\left(\Phi(\mathbf{c},\mathbf{e}), \Phi(\mathbf{c}',\mathbf{e}), t\right)\right] \geq \tau$, where $\mathrm{Err}(\mathbf{x}_0, \mathbf{x}'_0, t)$ is defined as*

$$\min_d \frac{1}{2}\left[\mathop{\mathbb{E}}_{q(\mathbf{x}_t|\mathbf{x}_0)} \mathbb{1}\left(d(\mathbf{x}_t) = \mathbf{x}'_0\right) + \mathop{\mathbb{E}}_{q(\mathbf{x}'_t|\mathbf{x}'_0)} \mathbb{1}\left(d(\mathbf{x}'_t) = \mathbf{x}_0\right)\right],$$

with $d : \mathcal{X} \to \mathcal{X}$ and $\mathbb{1}(\cdot)$ denoting the indicator function. We similarly define the degree of loss for each environmental attribute $\mathbf{e} \in \mathcal{E}$.

Intuitively, $\mathrm{Err}(\mathbf{x}_0, \mathbf{x}'_0, t)$ measures the smallest error for a network d to reconstruct the original image given noisy ones drawn from $q(\mathbf{x}_t|\mathbf{x}_0)$ or $q(\mathbf{x}'_t|\mathbf{x}'_0)$. Hence, when $\mathbf{x}'_0$ differs from $\mathbf{x}_0$ only by $\mathbf{c} \neq \mathbf{c}'$, a larger $\mathrm{Err}(\mathbf{x}_0, \mathbf{x}'_0, t)$ means that the attribute $\mathbf{c}$ becomes harder to distinguish from $\mathbf{c}'$, *i.e.*, more severe attribute loss. Particularly, we prove the close-form of $\mathrm{Err}(\mathbf{x}_0, \mathbf{x}'_0, t)$ in Appendix:

$$\mathrm{Err}(\mathbf{x}_0, \mathbf{x}'_0, t) = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\|\sqrt{\bar{\alpha}_t}(\mathbf{x}_0 - \mathbf{x}'_0)\|}{2\sqrt{2(1-\bar{\alpha}_t)}}\right)\right], \quad (5)$$

where $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2} dt$ denotes the error function. Hence we can compute attribute loss at each t w.r.t. the pixel-level changes $\|\mathbf{x}_0 - \mathbf{x}'_0\|$ (plotted in Figure 3), allowing us to derive a de-biasing strategy in Section 4.2.

Theorem. *1) For each $\mathbf{c} \in \mathcal{C}$, there exists a smallest time-step $t(\mathbf{c})$, such that $\mathbf{c}$ is lost with at least degree τ at each $t \in \{t(\mathbf{c}), \dots, T\}$. This also holds for each $\mathbf{e} \in \mathcal{E}$. 2) $\exists T, \{\beta_i\}_{i=1}^T$ such that $t(\mathbf{e}) > t(\mathbf{c})$ whenever $\|\Phi(\mathbf{c}', \mathbf{e}) - \Phi(\mathbf{c}', \mathbf{e}')\|$ is first-order stochastic dominant over $\|\Phi(\mathbf{c}, \mathbf{e}') - \Phi(\mathbf{c}', \mathbf{e}')\|$ with $\mathbf{c}' \sim \mathcal{C}, \mathbf{e}' \sim \mathcal{E}$ uniformly.*

Intuitively, the first part of the theorem states that a lost attribute will not be regained as time-step t increases, and there is a time-step $t(\mathbf{c})$ when $\mathbf{c}$ becomes lost (with degree τ) for the first time. The second part states that when changing the environmental attribute $\mathbf{e}$ is more likely to cause a larger pixel-level changes than changing the class attribute $\mathbf{c}$, then $\mathbf{e}$ becomes lost at a larger time-step compared to $\mathbf{c}$. See Figure 3 for an illustration of the two parts. We build on this mechanism in Section 4 to removes the visually prominent, yet spurious $\mathbf{e}$ for de-biasing in FSL.

4. Approach

We have shown in Eq. (5) that attribute loss prevents a denoising network to accurately reconstruct $\mathbf{x}_0$ from the noisy $\mathbf{x}_t$. Hence by contra-position, if we enable accurate reconstruction by minimizing Eq. (4) using a condition y, then the trained network d must associate y to the lost attributes to make up for their loss (Section 4.1). Moreover, our theorem indicates that only the nuanced $\mathbf{c}$ is lost at a small time-step t, hence y is only associated to $\mathbf{c}$ at t, allowing us to remove $\mathbf{e}$ to de-bias (Section 4.2). We detail our TiF learner below.

4.1. Parameterization

We use the reconstruction loss in Eq. (4) to train a denoising network d parameterized by θ_c on the few-shot examples of each category $c \in \{1, \dots, K\}$ by:

$$\min_{\theta_1,\dots,\theta_K} \sum_{(\mathbf{x}_0,c)\in\mathcal{D}} \sum_{t=1}^{T} \mathcal{L}_t(d(\cdot;\theta_c), \mathbf{x}_0, y). \quad (6)$$

When training converges, (θ_c, y, t) essentially parameterizes the lost attributes of category c at each time-step t. We detail our implementation below:

Choice of d. We use the text-conditioned Stable Diffusion (SD) [28], as it is trained on internet-scaled data with extensive knowledge of visual attributes. As shown in Figure 4,

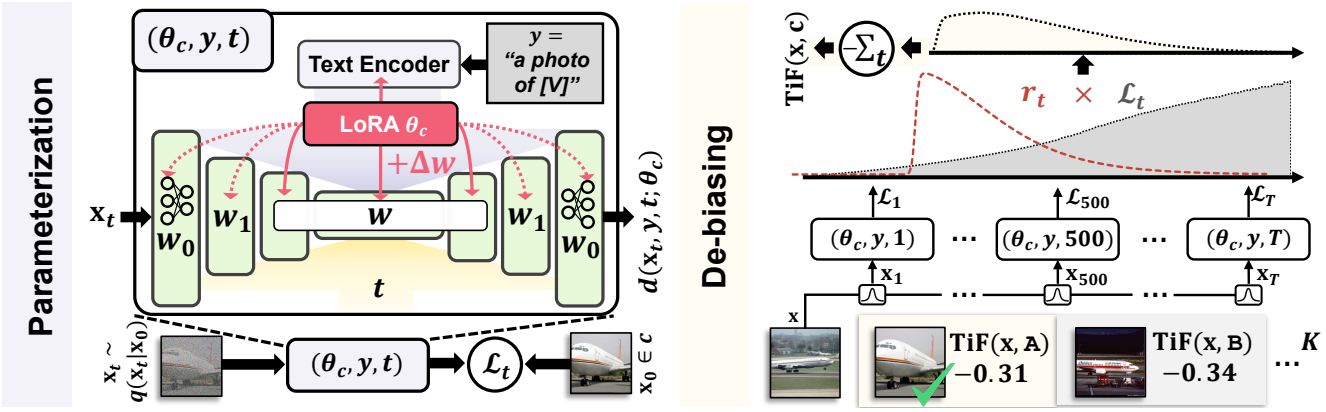

Figure 4. Overall pipeline of TiF learner. (a) Green: SD U-Net with each attention block illustrated by a rectangle. Red arrows: We inject trainable LoRA matrices θ_c to the attention blocks of U-Net and text encoder. Solid lines: always injected; dotted lines: optional (studied in ablation). We train θ_c to reconstruct $\mathbf{x}_0$ from $\mathbf{x}_t$ by $\mathcal{L}_t$. (b) Our inference rule by computing a weighted average $\mathcal{L}_t$ over time-steps.

d consists of a text encoder that extracts the text embedding from the prompt y, and a U-Net [29] that leverages text and time-step embedding to reconstruct $\mathbf{x}_0$ from $\mathbf{x}_t$. Note that SD encodes all images to a latent space. For convenience, we refer to the latent space when we say *image*, $\mathbf{x}$ or *pixel* with a slight abuse of notation.

Choice of θ_c. As shown in Figure 4, instead of fine-tuning all the parameters of d, we freeze the pre-trained weights of d and inject trainable Low-Rank Adaptation (LoRA) matrices to the linear layers in the text encoder and U-Net. For the U-Net, we use a subset of its attention blocks for injecting LoRA (ablation in Section 5.3). For each linear layer with weight $\mathbf{w}$, the new weight becomes $\mathbf{w} + \Delta\mathbf{w}$ after injection, where $\Delta\mathbf{w}$ is the injected low-rank matrix. We denote the LoRA matrices parameters for category c as θ_c. This offers several benefits: 1) The low-rank parameterization limits the model expressiveness to capture only the necessary attributes, *e.g.*, only $\mathbf{c}$ at a small t when $\mathbf{e}$ is yet lost. 2) As LoRA only affects the attention blocks, the model can only attend to the FSL task using existing visual knowledge, mitigating catastrophic forgetting [16] when adapting to the few-shot examples. 3) This also leads to efficient training and light-weight storage of θ_c.

Choice of y. We use a fixed text prompt y="a photo of [V]" for all categories, where [V] is a rare token identifier following DreamBooth [30]. The intuition is to choose [V] with a weak semantic prior, such that $d(\cdot;\theta_c)$ does not need to detach it from its existing meaning first, before associating it with the specificity of category c. Note, it is not necessary to use a different y for each c, as the LoRA matrices θ_c are class-specific and trained separately in Eq. (6).

4.2. De-biasing by Time-steps

Inference Rule. After training, the parameterization (θ_c, y, t) enables the denoising network d to make up for the lost attributes at t for category c. For a test image $\mathbf{x}$, our TiF learner defines its similarity with a category c as the ability of (θ_c, y, t) to make up for the lost *class attributes* (*i.e.*, nuances) in $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x})$ at each t, given by:

$$\mathrm{TiF}(\mathbf{x}, c) = -\sum_t r_t \mathcal{L}_t(d(\cdot;\theta_c), \mathbf{x}, y), \tag{7}$$

which enables inference by Eq. (1). Here, r_t is a ratio, namely, the degree of class attribute loss over that of all attribute losses. As shown in Figure 2, r_t highlights the ability of $d(\cdot;\theta_c)$ to make up the loss of fine-grained $\mathbf{c}$ at a small t, and essentially disregards the ability to make up the visually prominent $\mathbf{e}$ by reaching towards 0 at a larger t. Hence TiF learner essentially mitigates the influence from $\mathbf{e}$ to de-bias. We show how to compute r_t below.

Degree of Class Attribute Loss. To compute the degree of class attribute loss using Eq. (5), we need to estimate the average pixel-level changes δ^* when *only* altering the class attributes (measured by L2 distance). A naive way is to directly find the minimum L2 distance between any two images from different categories. However, it is unlikely for the few-shot training set $\mathcal{D}$ to contain two images that differ only in $\mathbf{c}$, *e.g.*, we may have class A on sky, B on ground and C with half sky and half ground. To this end, we estimate δ^* by a pixel-level approach:

$$\delta^* = \sqrt{\sum_{i=1}^{W}\sum_{j=1}^{H} \min\left\{\|\mathbf{x}_{i,j} - \mathbf{x}'_{i,j}\|^2 \mid c \neq c'\right\}}, \tag{8}$$

where (W, H) denotes the image size, and $\mathbf{x}_{i,j}$ is the pixel-level value at spatial location (i, j). In the earlier example, the sky and ground of C can be matched by min with that of A and B, respectively, hence minimizing the influence of background.

Computing r_t. After obtaining δ^*, we compute r_t by

$$r_t = \frac{1 - \mathrm{erf}\left(\gamma_t \delta^*\right)}{\int_{\delta^*}^{\infty} \left[1 - \mathrm{erf}(\gamma_t \delta)\right] d\delta}, \tag{9}$$

	Method	FGVCAircraft					ISIC2019				
		1	2	4	8	16	1	2	4	8	16
CLIP	Zero-Shot [26]			24.9					12.5		
	CoOp [48]	22.8-2.1	28.4+3.5	32.4+7.5	37.7+12.8	40.5+15.6	10.8-1.7	19.3+6.8	19.6+7.1	24.2+11.7	25.8+13.3
	Co-CoOp [47]	30.1+5.2	31.6+6.7	33.6+8.7	37.3+12.4	38.2+13.3	11.4-1.1	13.9+1.4	16.6+4.1	21.8+9.3	22.8+10.3
	MaPLe* [15]	30.1+5.2	33.0+8.1	33.8+8.9	39.4+14.5	40.7+15.8	11.8-0.7	14.5+2.0	18.0+5.5	22.6+10.1	26.8+14.3
OpenCLIP	Zero-Shot [26]			42.3					16.9		
	Linear-probe [26]	18.4-23.9	32.5-9.8	44.1+1.8	55.0+12.7	59.8+17.5	12.4-4.5	14.5-2.4	17.7+0.8	20.1+3.2	21.6+4.7
	Tip-Adapter [44]	47.7+5.4	51.6+9.3	54.7+12.4	58.4+16.0	62.2+19.9	22.6+5.7	25.3+8.4	23.6+6.7	31.1+14.2	33.9+17.0
	Tip-Adapter-F [44]	48.4+6.1	53.9+11.6	56.9+14.7	62.0+19.7	67.4+25.1	21.4+4.5	22.3+5.4	26.8+9.9	34.4+17.5	40.3+23.4
	CaFo* [45]	**51.7**+9.4	54.6+12.3	58.5+16.2	63.2+20.9	66.7+24.4	-	-	-	-	-
DM	Zero-Shot [17]			24.3					11.7		
	TiF learner w/o c	40.4+16.1	53.8+29.5	**65.0**+40.7	72.1+47.8	**80.4**+56.1	23.9+12.2	26.7+15.0	33.3+21.6	36.5+24.8	43.8+32.1
	TiF learner	48.5+24.2	**55.8**+31.5	64.2+39.9	**74.2**+49.9	79.9+55.6	**24.1**+12.4	**27.6**+15.9	**33.8**+22.1	**37.2**+25.5	**44.7**+33.0

Table 1. N-shot accuracies on FGVCAircraft and ISIC2019. The small number on the right indicates the absolute gain over its corresponding zero-shot model. MaPLe additionally tunes the visual encoder besides prompt tuning. CaFo leverages an ensemble of foundation models (see main text). Tip-Adapter has two variants, w/o fine-tuning and w/ fine-tuning (denoted as Tip-Adapter-F). We include results of additional FSL tasks in Appendix.

where $\gamma_t = \sqrt{\bar{\alpha}_t}/2\sqrt{2(1-\bar{\alpha}_t)}$ is the weight term from Eq. (5). In the denominator, we essentially use the integral to simulate diverse environmental attributes that are more coarse-grained than the nuanced class attributes. We include the computation details in Appendix.

5. Experiments

5.1. Settings

Datasets. Our experiments were conducted on four challenging and diverse datasets about fine-grained or customized classes: For **fine-grained** classification, we used: *1) FGVCAircraft* [22] consists of aircraft images from 100 classes with subtle visual differences. *2) ISIC2019* [4, 38] includes a diverse collection of dermatoscopic images with 8 types of skin lesions, serving as a testbed for our model's performance in medical image analysis. For **customized** classification, we used re-IDentification (reID) datasets, where each class is a specific identity (*e.g.*, a person or car): *3) DukeMTMC-reID* [27] contains 702 person IDs captured by 8 surveillance cameras (for evaluation), and we sub-sampled 150 IDs with the most test images for few-shot evaluation. *4) VeRi-776* [19, 20] contains 200 vehicle IDs under 20 cameras, providing a diverse range of angles and environmental conditions.

Evaluation Details. We followed the protocol in [8, 48] to directly adapt models on the K-way-N-shot few-shot training set. This setting is more versatile with two main differences from the conventional FSL: 1) We did not perform pre-training or meta-learning on a many-shot training set relating with the FSL task [2, 7, 40]. 2) Our K ranges from 100 to 200, much larger than the restricting $K = 5$ typical in the conventional setting [35, 39]. On FGVCAircraft, we sampled the few-shot set from its train split and evaluated accuracy on its test split. On ISIC2019, as its test split has no ground-truth labels, we sampled few-shot set from the train split and tested on the rest images. We used the macro F1 score to account for the class imbalance. On the reID datasets, we sampled few-shot images from their gallery set, and evaluated (rank-1) accuracy on their query set.

Implementation Details. We used SD 2.0 following [17], as we empirically found that the more commonly used SD 2.1 performs slightly worse in classification (see Appendix). For LoRA matrices rank, we used 8 on ISIC2019 and 16 on the rest datasets by visual inspection of the generation quality (see Section 5.3). We used a fixed rare token identifier [V]="hta" for all experiments. On dataset with class names, zero-shot CLIP and its adapter uses "a photo of [c], a type of [SC]" for each prompt y_c, where [c] denotes the name for class c and [SC] is a dataset-specific super-class name, *i.e.*, "aircraft" on FGVCAircraft and "skin lesion" on ISIC2019. For fair comparison, we implemented our y as "a photo of [V] [c], a type of [SC]". We also tried a variant without using class names (denoted as w/o c) by setting y as "a photo of [V], a type of [SC]". For inference, we modified the implementation in [17] to add our weight in Eq. (9). Other details in Appendix.

Baselines. We used two CLIP variants: CLIP with ViT-B/16 backbone trained on 400M image-caption pairs, and OpenCLIP with ViT-H/14 trained on LAION-2B. We compared with two lines of methods: 1) CLIP-Adapter [8], Tip-Adapter [44] and CaFo [45] based on Eq. (2). 2) Prompt-tuning methods CoOp [48], Co-CoOp [47] and MaPLe [15], which aim to learn y_c in Eq. 2. However, We did not test them on OpenCLIP, as their official implementations are specifically designed for CLIP and difficult to be fairly reproduced on OpenCLIP. We also compared with the zero-shot Diffusion Classifier [17], which is based on SD.

5.2. Main Results

Overall Results. As shown in Table 1 and 2, our TiF learner achieves state-of-the-art performance across fine-grained and reID datasets, *e.g.*, significantly improving

	Method	DukeMTMC-reID					VeRi-776				
		1	2	4	8	16	1	2	4	8	16
CLIP	CoOp [48]	8.2	10.4	17.2	29.6	33.7	11.4	13.9	18.1	30.3	34.1
	Co-CoOp [47]	9.7	12.1	20.1	32.7	44.5	30.1	31.6	33.6	37.3	38.2
	MaPLe [15]	13.5	32.9	40.7	55.7	63.0	35.2	40.7	44.4	57.5	68.1
OC	Linear-probe [26]	11.9	13.2	37.7	53.5	60.2	16.7	29.5	51.1	61.1	69.8
	Tip-Adapter [44]	28.5	36.9	46.5	59.3	66.7	36.4	47.5	59.8	71.2	80.1
	Tip-Adapter-F [44]	29.3	32.0	54.4	74.2	82.6	37.4	48.6	62.2	79.1	85.4
	TiF learner w/o c	**36.9**	**53.6**	**73.1**	**83.7**	**91.6**	**41.9**	**60.7**	**78.2**	**91.2**	**96.8**

Table 2. N-shot accuracies on DukeMTMC-reID and VeRi-776. OC is the short for "OpenCLIP". Note that the zero-shot methods are no longer applicable here, as there are no class names or ground-truth text description of the classes.

	Method	FGVCAircraft					DukeMTMC-reID				
		1	2	4	8	16	1	2	4	8	16
Subset	last	**48.5**	**55.8**	64.2	73.1	78.1	**36.9**	**53.6**	**73.1**	84.2	91.3
	last + w_1	48.2	55.7	65.0	**74.2**	78.5	28.7	47.7	69.7	**84.5**	91.1
	last + w_1+w_0	46.5	55.4	**65.1**	74.0	**79.9**	25.8	43.2	65.7	83.4	**91.6**
Weight	ELB weight	44.7	50.2	59.9	68.7	71.2	33.2	50.2	68.5	78.3	81.8
	PDAE	**49.2**	**56.7**	64.1	72.2	78.1	36.2	**53.9**	**73.4**	81.2	90.8
	Ours	48.5	55.8	**64.2**	**74.2**	**79.9**	**36.9**	53.6	73.1	**84.5**	**91.6**

Table 3. Top: Ablation on the LoRA injection subset. Last stands for the last attention block of the SD U-Net (solid lines in Figure 4). "+ w_1" and "+ w_2" stands for injecting the corresponding attention blocks in Figure 4. Bottom: Ablation on the time-step weights in inference. See main text for details.

existing methods by 13.7% on FGVCAircraft, 21.6% on DukeMTMC-reID and 16% on VeRi-776. On ISIC2019, we still achieve up to 7% absolute gain over the best-performing baseline, despite the challenges posed by diverse appearances within each class, as well as the stringent macro-F1 evaluation, which is sensitive to under-performing classes. Note that prompt tuning methods generally have non-ideal performances. This holds especially on customized reID classes with no class name to provide semantic prior. This validates the difficulties to describe the specification of fine-grained classes by prompt alone.

Low vs. High Shots. On low-shot settings (*e.g.*, 1- or 2-shot), we observe that our method may not improve significantly. However, not all foundation models are equal, and we must account for the discriminative capability of a zero-shot model. By checking the gain of each method over its zero-shot foundation model (small numbers) in Table 1, we observe that our method improves the most, *e.g.*, by +24.2% with just 1-shot on FGVCAircraft. We also highlight that CaFo leverages an ensemble of multiple models (*i.e.*, CLIP, DALL-E, GPT-3 and DINO), which provides rich prior to enable higher low-shot performance. By increasing #shots, our TiF learner's accuracy continue to grow strongly, *e.g.*, reaching near perfect predicting on the challenging 200-way-16-shot VeRi-776 task. This validates that our method can reap the benefits from well-defined class attributes using more shots. Moreover, the overall large improvements over baselines on 16-shot also highlights the persistence of spurious correlations, *i.e.*, any attribute from the diverse visual attributes set can spuriously correlate with $\mathbf{c}$ to challenge a few-shot learner.

With vs. Without [c]. On fine-grained datasets, we have access to the class name, and can include the class name in our prompt (see implementation details). Hence we compared the performance of TiF learner when using or without using class names (w/o $\mathbf{c}$) in Table 1. On FGVCAircraft, using class names significantly improves the 1- and 2-shot settings. This is because the class names can provide semantic prior to help describe the classes, which are otherwise ill-defined given the extreme low number of training images. However, on ISIC2019, using class names do not help with low-shot settings as much, and we conjecture that the prior knowledge of SD on ISIC2019 is significantly weaker compared to that on FGVCAircraft (much lower zero-shot accuracy). On higher-shot settings, adding class names overall has little to no effect, as the expanded training set can properly define each class.

5.3. Ablations

LoRA Injection Subset. In conventional few-shot learning works [36], it is common to train only the last few layers of the backbone, corresponding to low-level features. As our goal is to capture fine-grained attributes, we are motivated to inject LoRA to the last attention blocks of the SD U-Net, as shown in Figure 4. Starting from just injecting the last block (solid line), we tried expanding the injected blocks and show the results in Table 3 top. On 1-, 2- and 4-shot settings, injecting last attention block brings consistent improvements over other options. We conjecture that this provides additional inductive bias for the model to focus on

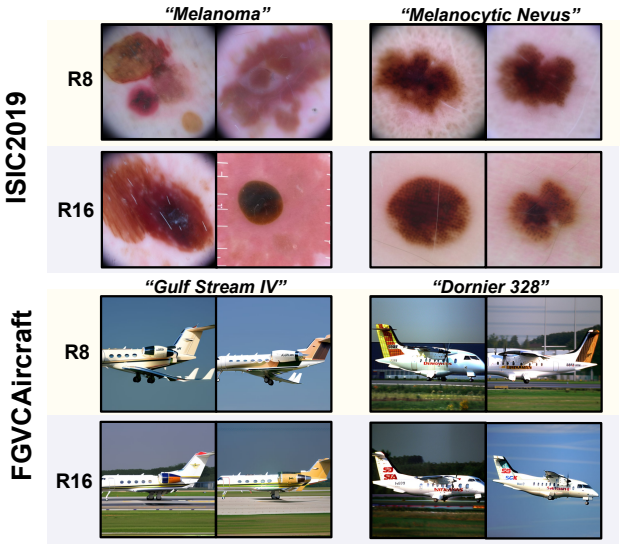

Figure 5. Comparison of synthesized images with two LoRA ranks. LoRA overfits to irrelevant details when rank is too high (top, rank 16), or fails to capture the nuances accurately when rank is too low (bottom, rank 8).

the fine-grained details (controlled by the last block). However, on 8-shot and 16-shot settings, this becomes suboptimal, and we postulate that this can limit the expressiveness of the injected SD, such that it fails to capture the specification of each class. Overall, "last + w_1" and "last + w_1 + w_2" work the best for 8- and 16-shot, respectively. Hence we use these settings for all experiments in Table 1 and 2. See additional ablation on this in Appendix.

Inference Weight r_t. In Table 3 bottom, we compared our inference weight r_t in Eq. (9) with two other weights. The ELB weight is derived from the evidence lower bound of the training data distribution (see Appendix), which respects any spurious correlation, *e.g.*, if most airplanes have sky background, the weight will encourage DM to follow this pattern in its generation. Hence it leads to sub-optimal performance when we want to suppress the spurious correlations in FSL. The PDAE weight [46] is developed to help distill visual attribute knowledge from a pre-trained DM, and also penalizes large time-steps when distillation is empirically difficult. Hence overall, it also performs reasonably well. However, it comes with a hyper-parameter γ, and when we use its default setting $\gamma = 0.1$, it does not perform well on some settings, *e.g.*, 16-shot on FGVCAircraft. In contrast, our proposed hyper-parameter-free adaptive weight is overall more stable due to its tailored design to isolate nuanced attributes from visually prominent ones.

LoRA Rank is the rank of each LoRA matrix $\Delta\mathbf{w}$. As our method is based on a generative model, we can easily choose its rank through a visual approach on the train set. Specifically, we synthesize images using the LoRA-injected SD. A desired rank should enable SD to generate the specification of each class, but still limits it from overfitting to irrelevant details. As shown in Figure 5, rank 16 on ISIC2019 is too high, as the model overfits to the scratches on the lens, while rank 8 on FGVCAircraft is too low, as the model fails to capture the wing and the vertical stabilizer. Overall, we used rank 8 for ISIC2019 with simpler visual attributes and rank 16 for the rest.

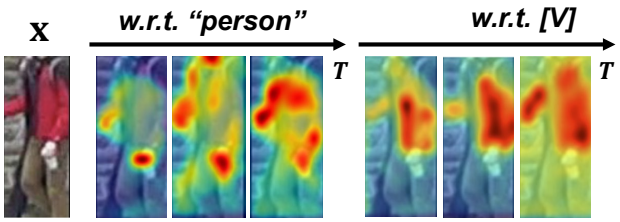

Figure 6. Attention map w.r.t. token "person" and "[V]".

Additional Attention Maps. In Figure 6, we show attention maps on DukeMTMC-reID w.r.t. different input tokens, where "person" corresponds to human parts and "[V]" is associated with the class attribute clothes that uniquely identify this person.

6. Conclusions

We presented TiF learner, a novel few-shot learner parameterization based on the Diffusion Model (DM), leveraging the inductive bias of its time-steps to isolate nuanced class attributes from visually prominent, yet spurious ones. Specifically, we theoretically show that in the forward diffusion process, nuanced attributes are lost at a smaller time-step than the visually prominent ones. Based on this, we train class-specific low-rank adapters that enables a text-conditioned DM to make up for the attribute loss by accurately reconstructing images from their noisy ones given a prompt. Hence each adapter and the prompt parameterizes only the nuanced class attributes at a small time-step, enabling a robust inference rule that focuses on the class attributes. Extensive results show that our method significantly outperforms strong baselines on various fine-grained or customized few-shot learning tasks. As future direction, we will seek additional inductive bias to tackle more complicated scenarios (*e.g.*, hierarchical classification).

7. Acknowledgement

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-022), MOE AcRF Tier 2 (MOE2019-T2-2-062), Wallenberg-NTU Presidential Postdoctoral Fellowship, the A*STAR under its AME YIRG Grant (Project No.A20E6c0101) and the Lee Kong Chian (LKC) Fellowship fund awarded by Singapore Management University. Pan Zhou was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

References

[1] Haokun Chen, Xu Yang, Yuhang Huang, Zihan Wu, Jing Wang, and Xin Geng. Manipulating the label space for in-context classification. *arXiv preprint arXiv:2312.00351*, 2023. 3

[2] 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*, 2019. 2, 6

[3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 1, 3

[4] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019. 6

[5] Alexander D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. In *AISTATS*, 2019. 1

[6] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations*, 2020. 2

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 3, 6

[8] 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. *arXiv preprint arXiv:2110.04544*, 2021. 1, 3, 6

[9] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[10] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *AAAI*, 2022. 3

[11] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 3, 4

[13] 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. In *ICLR*, 2022. 2

[14] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 1

[15] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023. 3, 6, 7

[16] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 2017. 5

[17] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, 2023. 6

[18] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. In *NeurIPS*, 2021. 3

[19] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016. 6

[20] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 2016. 6

[21] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019. 1

[22] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1, 6

[23] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2

[24] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *NeurIPS*, 14, 2001. 2

[25] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 6, 7

[27] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, 2016. 6

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 2015. 5

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2, 5

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3

[32] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021. 3

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[35] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017. 6

[36] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *CVPR*, 2019. 7

[37] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *CVPR*, 2020. 3

[38] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 2018. 6

[39] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016. 3, 6

[40] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *CVPR*, 2021. 6

[41] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *NeurIPS*, 2020. 1, 3

[42] Zhongqi Yue, Hanwang Zhang, and Qianru Sun. Make the u in uda matter: Invariant consistency learning for unsupervised domain adaptation. In *NeurIPS*, 2023. 3

[43] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. 2020. 2, 3

[44] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *ECCV*, 2022. 1, 3, 6, 7

[45] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *CVPR*, 2023. 3, 6

[46] Zijian Zhang, Zhou Zhao, and Zhijie Lin. Unsupervised representation learning from pre-trained diffusion probabilistic models. *NeurIPS*, 2022. 8

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1, 3, 6, 7

[48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 1, 3, 6, 7

[49] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. International Conference on Computer Vision, 2023. 3