Solving Semi-Supervised Few-Shot Learning from an Auto-Annotation Perspective

Tian Liu[1], Anwesha Basu[1], James Caverlee[1], and Shu Kong[2,3,✉]

[1]Texas A&M University, [2]University of Macau, [3]Institute of Collaborative Innovation
website and code: https://tian1327.github.io/SWIFT

Abstract. Semi-supervised few-shot learning (SSFSL) resembles real-world applications such as "auto-annotation", as it aims to learn a model from a few labeled and abundant unlabeled task-specific examples to annotate the unlabeled ones. Despite the availability of powerful open-source Vision-Language Models (VLMs) and open-world data, existing SSFSL literature largely neglects these resources. In contrast, the related area few-shot learning (FSL) has already exploited them to boost performance. Arguably, to solve real-world auto-annotation, SSFSL should leverage such open resources. To bridge this gap, we explore established SSL methods to finetune a VLM. Unexpectedly, they significantly underperform FSL baselines that do not use unlabeled data. Our in-depth analysis reveals the root cause of failure: VLMs produce "flat" distributions of softmax probabilities, resulting in zero utilization of unlabeled data and weak supervision signals. To address this challenge, we propose an embarrassingly simple solution that uses temperatures to sharpen the softmax output, which not only increases the confidence scores of pseudo-labels to improve the utilization of unlabeled data, but also strengthens training supervision for effective finetuning. Furthermore, we exploit task-relevant open data, e.g., those retrieved from VLMs' publicly available pretraining set. To mitigate the imbalance and domain gaps in retrieved data, we employ a stage-wise training strategy. Building on the successful finetuning of VLMs and the exploitation of open data, we present a simple yet effective SSFSL method, **S**tage-**Wi**se **F**inetuning with **T**emperatures (**SWIFT**). Across five benchmarks, SWIFT outperforms recent FSL and SSL methods by ∼5 accuracy points. SWIFT even rivals supervised learning, which finetunes a VLM assuming unlabeled data having ground-truth labels!

Keywords: Semi-Supervised Learning · Few-Shot Learning · Vision-Language Models · Retrieval Augmented Learning

1 Introduction

Semi-Supervised Few-Shot Learning (SSFSL) [15, 36, 38, 70] is well-suited for real-world applications such as "auto-annotation" [40, 42, 43]. It aims to learn a model utilizing a small set of task-specific labeled data and a much larger set of unlabeled data to produce reliable annotations for the unlabeled examples.

Status Quo. SSFSL extends few-shot learning (FSL) by leveraging unlabeled examples for training, alongside a small labeled set. Modern FSL methods

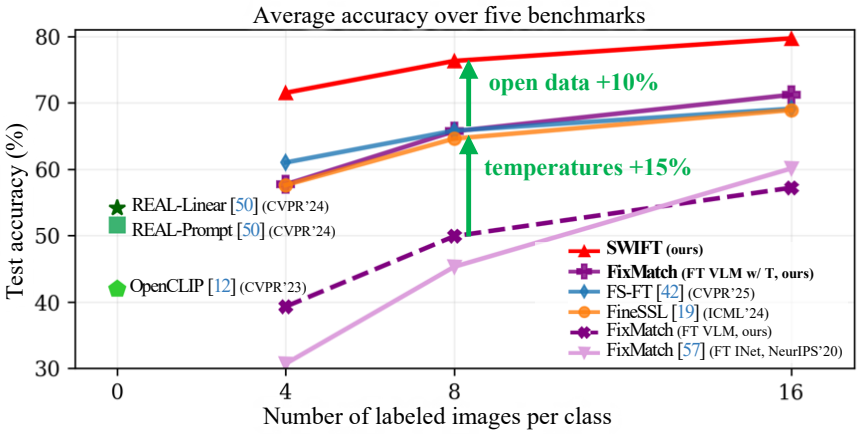

Fig. 1: Benchmarking SSFSL methods across five datasets. First, traditional SSL methods e.g., FixMatch [55], which finetune ImageNet-pretrained backbones [22], underperform recent VLM-based zero-shot methods like REAL [48]. Second, SOTA SSL method FineSSL [17] exploits a frozen VLM via prompt learning, achieving notable gains yet underperforming recent few-shot method FS-FT [40], which finetunes a VLM without using unlabeled data. Third, we are thus motivated to finetune a VLM using established FixMatch (denoted "FT VLM") and find that it fails, due to the "flat" softmax probabilities of VLMs (Fig. 3). Fourth, to address this challenge, we adopt simple temperatures to finetune VLM (denoted "FT VLM w/ T"), which significantly improves accuracy by >15%, outperforming FineSSL [17]. Finally, our method SWIFT further exploits open data via stage-wise training, achieving further 10% accuracy gains.

[40, 54, 64] have achieved significant progress by finetuning pretrained Vision-Language Models (VLMs) and retrieving task-relevant open data (e.g., from VLM's publicly available pretraining set) to augment training data. In contrast, contemporary SSFSL and general Semi-Supervised Learning (SSL) methods largely overlook these powerful resources. Some of them either train models from scratch [2,3,55,79] or finetune ImageNet-pretrained backbones [9,56,65,66]. Other works [15,36,38,70] adopt a simulated setup, first pretraining a ResNet-12 model on a large "base" dataset and then semi-supervised finetuning on domain-specific data. A few recent SSL methods exploit a frozen VLM with prompt learning rather than finetuning [17, 45, 81, 82]. These design choices significantly limit the progress of SSFSL. To solve real-world applications like auto-annotation, we study the underexplored paradigm that explores finetuning VLMs and retrieving open data for SSFSL (Fig. 2).

Motivation. SSFSL frames the important application "auto-annotation" as both aim to exploit limited labeled data and abundant unlabeled data to produce labels for unlabeled examples. Auto-annotation also inspires recent FSL research, leading to a rigorous and practical paradigm that prioritizes prediction accuracy over parameter efficiency [40, 43] and eschews the unrealistic assumption of accessible large validation sets for hyperparameter tuning [37, 40, 54]. Moreover,

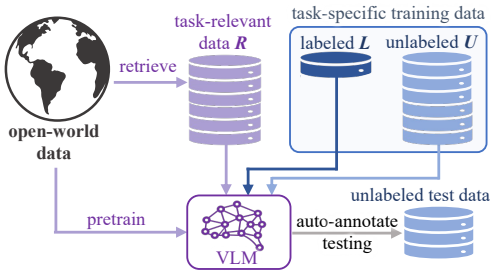

Fig. 2: A realistic setup for SSFSL embraces open-source VLMs and open-world data. It allows retrieving task-relevant data R to augment the task-specific labeled and unlabeled training data L and U. Applying the learned model to test data amounts to using it to auto-annotate the unlabeled test data.

concurrent advances in zero-shot learning (ZSL) [28,39,48,50,62] and FSL [40,64] have demonstrated that exploiting not only VLMs but also open data (e.g., VLM's pretraining data) can substantially improve performance. To bridge the research gap, we rigorously explore SSFSL by finetuning open-source VLMs and exploiting open data (Fig. 2).

Challenges and Insights. Inspired by the significant improvement in FSL by finetuning a VLM on few-shot labeled data [40], we explore established SSL methods to finetune a VLM by further utilizing unlabeled data. Unexpectedly, such direct adoption underperforms FSL and even ZSL baselines (Fig. 1), aligning with observations in prior work [82]. We identify the root cause: *VLMs produce "flat" distribution of softmax probabilities* (Fig. 3A), which results in zero utilization of unlabeled data (Fig. 3B) and weak learning signals that prevent effective finetuning (Fig. 3C). To address this, we propose an embarrassingly simple technique that uses temperatures to sharpen softmax probabilities, significantly improving SSFSL performance (Tab. 2). Moreover, we retrieve task-relevant open data and employ stage-wise training to mitigate the inherent noise, imbalance, and out-of-distribution (OOD) nature of retrieved data. Integrating these insights, our method, **S**tage-**Wi**se **F**inetuning with **T**emperatures (**SWIFT**), achieves state-of-the-art (SOTA) performance across five SSFSL benchmarks (Fig. 1).

Contributions. We make three key contributions.

- From an auto-annotation perspective, we explore an underexplored SSFSL paradigm by exploiting open-source models and open data. Our realistic setup (Fig. 2) significantly advances SSFSL performance.
- We identify why existing SSL methods fail to finetune VLMs: VLMs produce "flat", low confidence scores that lead to zero utilization of unlabeled data and weak supervision. We address this by leveraging temperatures.
- We present a simple yet effective method, **SWIFT**, which outperforms prior FSL and SSL approaches by >5 accuracy points across five SSFSL datasets, even rivaling fully supervised learning.

2 Related Works

Semi-Supervised Few-Shot Learning (SSFSL) [15,36,38,70] is a special case of SSL [7] in which the amount of unlabeled data is extremely limited. Classic SSL methods build on several core ideas: consistency regularization [69,76,85],

which enforces prediction consistency between strong and weak augmentations of unlabeled examples; pseudo-labeling [1, 6, 35, 73], which uses a teacher model's predictions to supervise the student model; and transfer learning [11, 21], where a model is first self-supervised, pretrained on large unlabeled datasets, and then adapted to a downstream task with labeled data. Among various SSL methods, FixMatch [55] stands out for its simplicity and strong performance. Subsequent works improve FixMatch by selecting confident pseudo-labels with adaptive thresholds [4, 9, 16, 20, 27, 67, 75, 79], or enhancing pseudo-label quality through logit adjustment [46, 65] or teacher ensembling [5, 34, 59]. Recent SSL methods further leverage zero-shot predictions from pretrained VLMs as auxiliary supervisions [13, 65], or adopt prompt tuning with a frozen VLM [17, 45, 81, 82]. However, no existing works have successfully finetuned VLMs for SSL. The recent work [82] even asserts that tailored SSL methods are required to finetune VLMs effectively. Our study, however, reveals the root cause of such failures is the "flat" softmax outputs from VLMs. We address this by leveraging temperatures in learning. For the first time, we enable successful finetuning of VLM for SSFSL.

Open-World Data and Open-Source Models. The abundance of open-world data not only facilitates training more robust and generalizable models [24, 31, 53], but also enables pretraining foundational models [12, 30, 49, 58, 74]. By leveraging pretrained VLMs, recent methods in ZSL [39, 48, 62] and FSL [18, 37, 54, 64, 80] have achieved substantial improvements. These paradigms [28, 39, 40, 50, 62] further enhance performance by retrieving task-relevant open data, such as those from VLMs' pretraining datasets [51, 52]. In contrast, existing SSL [4, 55, 65, 79] and SSFSL [15, 38, 41, 70] research largely overlooks these powerful resources. Current frameworks typically rely on training from scratch, finetuning ImageNet-pretrained backbones, or employing frozen VLMs without finetuning [17, 45, 81, 82]. Intuitively, combining VLM finetuning with open-world data retrieval would significantly enhance SSFSL. However, this potential remains largely untapped. We propose an SSFSL method that integrates both VLM finetuning and open data retrieval, thereby substantially advancing SSFSL research.

Temperature is a hyperparameter used in the softmax function to control the sharpness of the resulting class probability distributions [25, 29]. It has been applied in different tasks for various purposes, such as softening the teacher model's logits to enhance knowledge distillation [25], calibrating model confidence [19, 60, 61], adjusting sampling diversity in language generation [26], and scaling contrastive loss to facilitate large-scale pretraining of foundation models [10, 21, 63, 72]. In VLMs' pretraining [12, 49, 74], temperature is set as a learnable parameter in the contrastive loss, initialized to 0.07 [72] and eventually clipped at 0.01 to stabilize training [49]. With pretrained VLMs, recent methods commonly inherit the fixed temperature of 0.01 when initializing classifier weights in ZSL [71] or when computing cross-entropy loss in FSL for learning the classifier or prompt [18, 37, 54, 81, 84, 86]. The significance of temperature in VLM finetuning remains unexplored. Our work, for the first time, demonstrates that temperature is crucial for successful SSFSL with VLMs.

3 Problem Formulation and Methods

In this section, we first formalize our SSFSL setup. Then, we repurpose representative SSL and FSL methods as baselines to finetune a VLM for SSFSL and analyze their failures. Lastly, we present a simple temperature-based solution and derive our final SSFSL method.

Problem Setup and Notations. A downstream C-way image classification task provides a small set of labeled images $L = \{(\mathbf{I}_i, y_i)\}_{i=1}^{N_l}$, where $y_i \in \mathbb{N}^C$, and a large set of unlabeled images $U = \{\mathbf{I}_j\}_{j=1}^{N_u}$ with $N_l \ll N_u$. Each class in L contains $K \in \{4, 8, 16\}$ labeled images, formalizing a K-shot setting with $N_l = K * C$. A pretrained VLM and an open data pool (e.g., VLM's pretraining dataset) P are provided. The VLM consists of a visual encoder $\mathbf{V}(\cdot)$ and a text encoder $\mathbf{T}(\cdot)$, which transform an input image $\mathbf{I}$ into visual embedding $\mathbf{V}(\mathbf{I}) \in \mathbb{R}^d$ and its class name into text embedding $\mathbf{T}(y) \in \mathbb{R}^d$, respectively. From P, one can retrieve a set of task-relevant examples $R = \{(\mathbf{I}_i, y_i)\}_{i=1}^{N_r}$. These examples naturally follow imbalanced distributions and contain noisy labels $y_i \in \mathbb{N}^C$, exhibiting distribution shifts relative to the task-specific ones in L and U [40].[1] Our SSFSL exploits data in L, U, and R to finetune the VLM and evaluate on the held-out test set (Fig. 2).

Validation-free Evaluation Protocol. Given the scarcity of labeled data, our SSFSL setup adopts a realistic *"validation-free protocol"* that eschews a validation set for hyperparameter tuning. Our approach is consistent with recent works in FSL [40, 54] and SSL [56, 82] that emphasize real-world applicability. Specifically, these works employ a rigorous "cross-dataset tuning" strategy, in which hyperparameters (e.g., learning rate, weight decay) are tuned on a single source dataset and then directly applied to all other datasets. Prior work [40] shows that such protocol yields robust performance, without suffering from overfitting even in the low-data regime. Our experiment follows this rigorous protocol by directly adopting hyperparameters from [40] and applying them throughout our experiments on all datasets, without any tuning.

3.1 Baseline Results and Failure Diagnosis

We experiment with representative FSL and SSL methods as baselines to finetune VLM for SSFSL, aiming to diagnose their failures and derive our method. Results presented below are averaged over five benchmarks using the open-source VLM OpenCLIP ViT-B/32, as detailed in Sec. 4.

FSL Method. Few-shot finetuning (FS-FT) the VLM's visual encoder $\mathbf{V}(\cdot)$ is a competitive validation-free FSL method [40]. First, it initializes a classifier $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^C] \in \mathbb{R}^{d \times C}$ using the text embeddings of the C class names. For an input image $\mathbf{I}$, the visual encoder encodes the image and passes it to the classifier, producing a logit vector $\mathbf{q} = \mathbf{W}^T\mathbf{V}(\mathbf{I}) \in \mathbb{R}^{C \times 1}$. A temperature parameter τ is then applied to the logits to compute the softmax probability

[1] As we focus on VLM finetuning rather than data retrieval, we adopt the method proposed in [40, 48] to retrieve task-relevant data.

Table 1: Established SSL methods fail to finetune VLM. While directly running FixMatch [55] and DebiasPL [65] to finetune the VLM OpenCLIP ViT-B/32 [12] improves performance with the ImageNet-pretrained ResNet-50 backbone [56] under 4- and 8-shot settings, doing so yields lower accuracy in the 16-shot setting. Notably, it significantly underperforms FS-FT, which does not use unlabeled data for training. Red superscripts denotes accuracy degradation w.r.t. FS-FT [40].

method	training data	backbone	mean acc. over five datasets		
			4-shot	8-shot	16-shot
FS-FT [40]	L	VLM-OpenCLIP	61.0	65.8	69.1
FixMatch [55]	$L+U$	INet-pretrained	30.7	45.3	60.1
FixMatch	$L+U$	VLM-OpenCLIP	39.3$^{-21.7}$	49.9$^{-15.9}$	57.2$^{-11.9}$
DebiasPL [65]	$L+U$	INet-pretrained	34.5	49.0	62.7
DebiasPL	$L+U$	VLM-OpenCLIP	39.6 $^{-21.4}$	49.8 $^{-16.0}$	57.1 $^{-12.0}$

vector $\mathbf{s}$, where $s^c = \frac{\exp(q^c/\tau)}{\sum_{j=1}^{C}\exp(q^j/\tau)}$. Finally, $\mathbf{V}(\cdot)$ and $\mathbf{W}$ are jointly finetuned by minimizing the cross-entropy (CE) loss over the few-shot labeled set L: $\mathcal{L}_l = \frac{1}{|L|}\sum_{(\mathbf{I},y)\in L}\ell(\mathbf{W}^T\mathbf{V}(\mathbf{I})/\tau, y)$.

SSL Methods. The influential FixMatch [55] generates pseudo-labels for unlabeled data and selects high-confident ones for training (Fig. 4). For each unlabeled image $\mathbf{I}\in U$, FixMatch generates a weakly augmented view $\mathbf{I}^w$ and passes it through the training model to produce a C-dim logits $\mathbf{q}^w = \mathbf{W}^T\mathbf{V}(\mathbf{I}^w)$. Based on the logits $\mathbf{q}^w$, the softmax probabilities are calculated with a default temperature of 1.0: $\mathbf{s}^w = \text{softmax}(\mathbf{q}^w)$. Then, the pseudo-label $\hat{y} = \arg\max_c(\mathbf{s}^w)$ is derived. Notably, to mitigate noisy pseudo-labels, FixMatch only utilizes pseudo-labels with softmax confidence higher than a threshold, i.e., $\max_c(\mathbf{s}^w) \geq \sigma$ (default $\sigma = 0.8$ [56]). Based on the selected pseudo-labeled data, FixMatch finetunes the model on a strongly augmented view $\mathbf{I}^s$ by minimizing $\mathcal{L}^s = \frac{1}{|U|}\sum_{\mathbf{I}\in U}\mathbb{1}[\max_c(\mathbf{s}^w)\geq\sigma]\cdot\ell(\mathbf{W}^T\mathbf{V}(\mathbf{I}^s),\hat{y})$. $\mathcal{L}_l$ and $\mathcal{L}^s$ are jointly used in learning. DebiasPL [65] extends FixMatch by adjusting the logits of unlabeled data and incorporating adaptive class margins to mitigate the imbalanced distribution of pseudo-labels. Importantly, both FixMatch [55] and DebiasPL [65] typically finetune an ImageNet-pretrained backbone. We conduct experiments and analyze their performance when finetuning a VLM, and compare them with the FSL method FS-FT [40] introduced above.

Failure Diagnosis. Counterintuitively, our experiments in Tab. 1 show that finetuning VLM using FixMatch and DebiasPL significantly underperforms the simple FSL method that does not exploit unlabeled data. Our results align with the observation in recent work [82], although the underlying reason remains uninvestigated. Remarkably, our in-depth analysis identifies the root cause: *VLMs produce "flat" distributions of softmax probabilities* (Fig. 3A), which:

1. yields low-confidence pseudo-labels, resulting in zero utilization of unlabeled data with a default high confidence threshold (Fig. 3B);
2. provides weak training supervision, preventing effective finetuning (Fig. 3C).

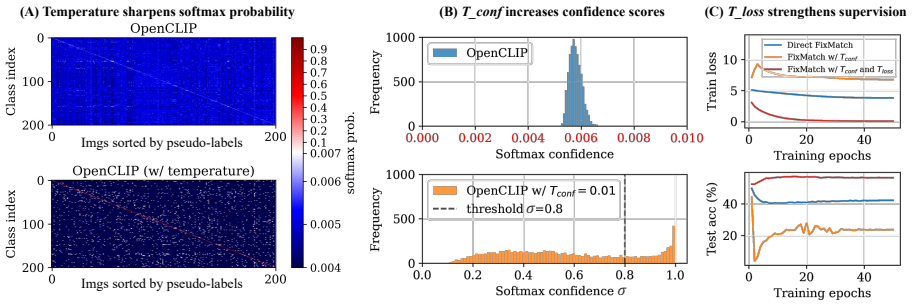

Fig. 3: Challenges in SSFSL with VLMs and our solution via temperatures. **(A-top)** We obtain softmax probabilities of the unlabeled images from semi-Aves dataset by zero-shot prompting the OpenCLIP ViT-B/32 model with its 200 class names. VLMs produce a "flat" distribution of softmax probabilities (200-dim column vector per unlabeled example), resulting in low confidence scores **(B-top)** and weak training supervision that prevents effective finetuning **(C)**. We solve this by using simple temperatures to sharpen the softmax distributions, as illustrated by the red diagonal line in **(A)**-bottom. Concretely, we apply a temperature T_{conf} (e.g., 0.01) to increase pseudo-label confidence, enabling the utilization of unlabeled data using the default threshold of 0.8 **(B-bottom)**. We also use a temperature T_{loss} to compute CE loss, enhancing training loss reduction and improving test accuracy (red lines in **(C)**).

3.2 Temperature is the Remedy

To address the above issues, we present an embarrassingly simple technique by using temperatures, which have not been explored in the SSFSL literature.[2]

Confidence Temperature Enables Utilization of Unlabeled Data. To address the small confidence scores from VLM, we use a confidence temperature T_{conf} to sharpen the softmax probability distributions for the weakly augmented images $\mathbf{I}^w$: $\mathbf{s}^w = \text{softmax}(\mathbf{W}^T\mathbf{V}(\mathbf{I}^w)/T_{\text{conf}})$. Lowering the temperature increases pseudo-label confidence, enabling the utilization of unlabeled data with the default high-confidence threshold (Fig. 3B-bottom). One may argue that an alternative approach is to lower the confidence threshold σ. However, setting an appropriate threshold requires careful tuning on a validation set [82], which is impractical under the realistic "validation-free" setup and nearly infeasible given the narrow range of confidence scores (Fig. 3B-top). In contrast, our analysis demonstrates that using a confidence temperature is much more robust, where setting T_{conf} within a broad range [0.001, 0.05] consistently yields substantial performance gains (Fig. 7).

Loss Temperature Strengthens Training Supervision. While T_{conf} enables utilization of unlabeled data, applying it alone yields much higher training loss and lower test accuracy (Fig. 3C), due to the weak training supervision. To strengthen the supervision from selected pseudo-labeled data, we apply a loss temperature T_{loss} when computing the CE loss between pseudo-labels and strongly

[2] It is worth noting that the recent work [82] uses FixMatch to finetune the VLM CLIP, achieving worse performance than finetuning on labeled data only. Importantly, their released code [83] shows that they do not use temperatures.

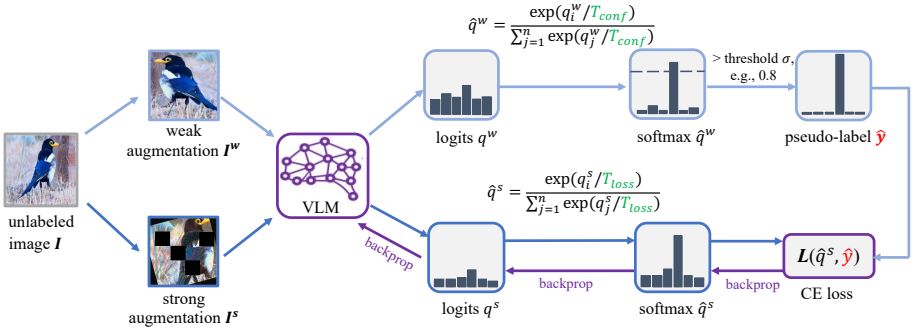

Fig. 4: Illustration of our temperature-based solution for finetuning a VLM with FixMatch. First, T_{conf} sharpens softmax probabilities and increases confidence scores of pseudo-labels (Fig. 3B), increasing the utilization of unlabeled data. Second, T_{loss} strengthens the supervision from CE loss, enabling effective VLM finetuning. We provide PyTorch-style pseudo-code in the Supplement.

augmented images $\mathbf{I}^s$: $\mathcal{L}_u = \frac{1}{|U|}\sum_{\mathbf{I}\in U} \mathbb{1}[\max_c(s^c) \geq \sigma] \cdot \ell(\mathbf{W}^T\mathbf{V}(\mathbf{I}^s)/T_{\text{loss}}, \hat{y})$. Similarly, we apply T_{loss} when calculating the CE loss for labeled data: $\mathcal{L}_l = \frac{1}{|L|}\sum_{(\mathbf{I},y)\in L} \ell(\mathbf{W}^T\mathbf{V}(\mathbf{I})/T_{\text{loss}}, y)$. Crucially, rather than treating T_{loss} as a static hyperparameter, we demonstrate that learning T_{loss} dynamically during training further enhances SSFSL performance (Fig. 6).

Fig. 4 illustrates our temperature-based solution applied to FixMatch. Fig. 3C confirms that applying T_{conf} and T_{loss} accelerates learning and improves test accuracy. Notably, our simple solution is compatible with many subsequent FixMatch-based SSL methods, e.g., DebiasPL [65] (Tab. 2). Post-analyses in Figures 6 and 7 demonstrate SWIFT's robustness to temperature settings.

3.3 Exploiting Open Data via SWIFT

Building on our successful finetuning of VLMs, we further exploit open data to enhance SSFSL. Retrieval augmentation (RA) has been studied in ZSL [28, 39, 50, 62] and FSL [40] literature, but remains unexplored in SSL. Specifically, prior ZSL and FSL works retrieve task-relevant data from VLMs' publicly available pretraining sets, e.g., LAION [51, 52]. For fair comparisons with prior methods and to ensure reproducibility, we follow [40, 48] to retrieve from LAION-400M [52] and construct a set of noisy labeled data R (Fig. 2).

With R, we employ a CE loss: $\mathcal{L}_r = \frac{1}{|R|}\sum_{(\mathbf{I},y)\in R} \ell(\mathbf{W}^T\mathbf{V}(\mathbf{I})/T_{\text{loss}}, y)$, in addition to the losses $\mathcal{L}_l$ and $\mathcal{L}_u$. A straightforward approach is to finetune the VLM using all three losses $\mathcal{L}_l$, $\mathcal{L}_u$, and $\mathcal{L}_r$ in a single stage. However, such single-stage training is suboptimal because the retrieved data R exhibit imbalanced class distributions, noisy labels, and distribution shifts relative to task-specific images L and U [40] (see examples in Supplementary Fig. 8). To address these issues, we propose a stage-wise SSFSL method, **SWIFT**: **S**tage-**Wi**se **F**inetuning with **T**emperatures (Fig. 5), consisting of three training stages, *where our employed temperatures consistently yield remarkable improvements*:

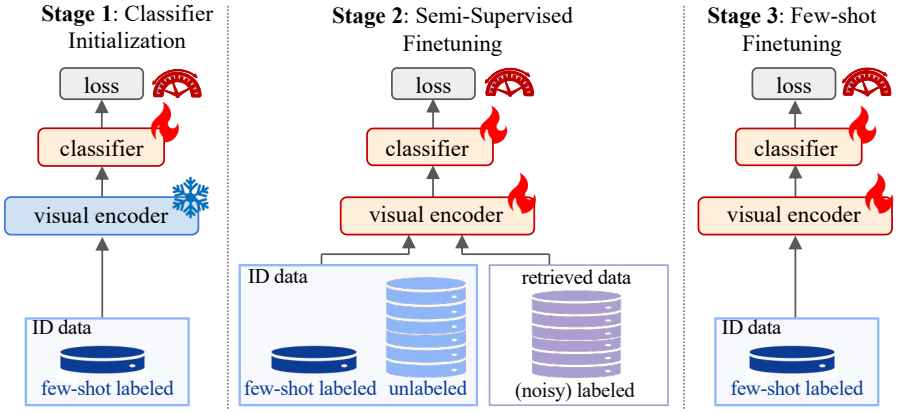

Fig. 5: Illustration of our SWIFT training pipeline for SSFSL. SWIFT adopts a stage-wise training approach, applying temperature in all three stages to enable effective VLM finetuning. SWIFT exploits not only task-specific few-shot labeled data L and abundant unlabeled data U but also large amounts of retrieved open data R.

- *Stage-1 Classifier Initialization.* While existing FSL methods [37, 40, 54] typically initialize the classifier weights $\mathbf{W}$ using text embeddings of class names, we find that further learning the classifier on few-shot labeled data L with a loss temperature T_{loss} notably improves subsequent finetuning (Tab. 4).
- *Stage-2 Semi-Supervised Finetuning.* We finetune both visual encoder $\mathbf{V}(\cdot)$ and classifier $\mathbf{W}$ using a SSL method (e.g., FixMatch [55]) combining three losses $\mathcal{L}_l$, $\mathcal{L}_u$, and $\mathcal{L}_r$. Crucially, we use a confidence temperature T_{conf} and a loss temperature T_{loss} to ensure successful SSL (Fig. 4).
- *Stage-3 Few-shot Finetuning.* We finetune $\mathbf{V}(\cdot)$ and classifier $\mathbf{W}$ on few-shot labeled data L again, with a loss temperature T_{loss}. Doing so effectively mitigates domain gaps, class imbalance, and noisy labels in R [40].

4 Experiments and Results

We validate SWIFT by benchmarking it against recent FSL and SSL methods. We also conduct post-analyses to demonstrate the robustness of SWIFT to temperature settings.

4.1 Experimental Setup

Datasets and Metrics. Following the auto-annotation perspective [40] and recent SSL studies [82], we evaluate on challenging fine-grained recognition datasets where pretrained VLMs, such as OpenCLIP [49], yield low zero-shot accuracy. This also suggests that their pretraining data are less aligned with these datasets, thereby reducing the risk of data leakage when adopting retrieval augmentation (RA). The datasets include semi-Aves [57] (CC-BY 4.0 License),

	method		mean acc. over five datasets		
			4-shot	8-shot	16-shot
FSL	CoOp [86] IJCV'22	PL	48.8	54.6	59.5
	PLOT [8] ICLR'23	PL	50.3	54.9	59.3
	Linear Probing [49] ICML'21	AL	57.0	61.2	64.7
	CLIP-Adapter [18] IJCV'23	AL	48.6	55.2	60.0
	Tip-Adapter (f) [84] ECCV'22	AL	50.0	52.9	58.0
	TaskRes(e) [77] ECCV'22	AL	53.0	58.0	62.0
	CMLP [37] CVPR'23	AL	54.4	59.4	64.2
	CLAP [54] CVPR'24	AL	56.9	61.3	65.5
	FS-FT [40] CVPR'25	FT	61.0	65.8	69.1
	SWAT [40] CVPR'25 (w/ RA)	FT	67.4	71.0	74.0
SSFSL	FixMatch [55] NeurIPS'20	FT	39.3	49.9	57.2
	FixMatch w/ Temp. (ours)	FT	57.7$^{+18.4}$	65.7$^{+15.8}$	71.2$^{+14.0}$
	DebiasPL [65] CVPR'22	FT	39.6	49.8	57.1
	DebiasPL w/ Temp. (ours)	FT	60.3$^{+20.7}$	67.6$^{+17.8}$	73.2$^{+16.1}$
	FineSSL [17] ICML'24	PL	57.6	64.6	68.9
	SWIFT (ours)	FT	**71.5**$^{+4.1}$	**76.3**$^{+5.3}$	**79.7**$^{+5.7}$
Ref.	Fully Supervised	FT	74.8	75.4	76.0
	Fully Supervised w/ RA	FT	76.0	76.8	77.2

Table 2: SWIFT achieves SOTA SSFSL performance. We compare our methods with recent FSL and SSL methods that adapt the VLM OpenCLIP ViT-B/32 through prompt learning (PL), adapter learning (AL), and finetuning (FT). Notably, our temperatures significantly improve FixMatch [55] and DebiasPL [65] by 14-20% accuracy. Crucially, SWIFT outperforms SOTA FSL method SWAT [40] that also exploits retrieval augmentation (RA) by 5 accuracy points (superscripts). SWIFT even rivals fully supervised references, which uses ground-truth labels for unlabeled data.

FGVC-Aircraft [44], Stanford Cars [32] (Custom Non-commercial License), EuroSAT [23] (MIT License), and DTD [14] (Custom Research-Only License). These datasets span bird species, aircraft and car models, land-use and cover in satellite imagery, and texture types (details and examples in Supplementary Tab. 5 and Fig. 8). For each method, we report its mean accuracy across five benchmark test sets in the main paper and provide detailed results in the Supplement. For RA, we retrieve data for each dataset from the publicly available LAION-400M [52].

Compared Methods and Models. In addition to FixMatch and DebiasPL presented in Section 3, we compare our SWIFT with the state-of-the-art (SOTA) SSL method FineSSL [17], which adopts prompt learning with a frozen VLM. We also compare with various FSL methods, including popular ones that learn either a lightweight adapter [18, 37, 49, 54, 78, 84] or prompts [8, 36, 68] with a frozen VLM. In particular, we compare SWIFT with the SOTA FSL methods that finetune a VLM's visual encoder, including FS-FT [40] that finetunes on few-shot labeled data only, and SWAT [40] that also retrieves open data to augment the training set. For fair comparison with recent VLM-based FSL and SSL methods, we mainly study the open-source VLM OpenCLIP ViT-B/32 [12]. We also demonstrate that SWIFT generalizes to other pretrained foundation models such as the DINOv2 ViT-B/14 [47]. Supplement Sec. C provides analysis of additional backbones, such as ResNet-50 and ViT, pretrained on ImageNet. As a reference, we report results from supervised learning that finetunes the visual encoder on both labeled and unlabeled data (with ground-truth labels), optionally with retrieved data.

Implementation Details. For each dataset, we randomly sample 4-, 8-, and 16-shot labeled data from the official training set and repurpose the remaining training data as unlabeled data. An exception is semi-Aves [57], where we use its official unlabeled in-domain data as the unlabeled set (details in Supplementary

Table 3: Ablation study of SWIFT. Each stage of SWIFT (Fig. 5) contributes significant accuracy gains when compared with directly applying FixMatch [55] to finetune the VLM OpenCLIP [12]. Notably, SWIFT also generalizes to the DINOv2 model, which is pretrained on large-scale unlabeled images using a self-supervised loss [47]. Superscripts mark the incremental improvements relative to the previous row. We refer the reader to Supplementary Tab. 10 and Tab. 13 for per-dataset analyses on the stage-wise design of SWIFT.

method	training data	mean accuracy over five datasets					
		OpenCLIP ViT-B/32			DINOv2 ViT-B/14		
		4-shot	8-shot	16-shot	4-shot	8-shot	16-shot
FS-FT [40] CVPR'25	L	61.0	65.8	69.1	56.5	71.1	79.6
FixMatch [55] direct	L+U	39.3	49.9	57.2	50.2	66.3	77.1
+ stage-1: Classifier Init.	L+U	56.3$^{+17.0}$	59.8$^{+9.9}$	62.2$^{+5.0}$	56.7$^{+6.5}$	71.5$^{+5.2}$	80.3$^{+3.1}$
+ stage-2: Semi-Sup. FT	L+U+R	68.8$^{+11.1}$	73.3$^{+7.6}$	77.3$^{+6.1}$	76.3$^{+19.6}$	82.0$^{+10.5}$	86.3$^{+6.0}$
+ stage-3: FS-FT (SWIFT)	L+U+R	**71.5**$^{+2.7}$	**76.3**$^{+3.0}$	**79.7**$^{+2.4}$	**78.2**$^{+1.9}$	**84.4**$^{+2.4}$	**87.8**$^{+1.5}$

Tab. 5). Following [40,48], we retrieve 500 images per class from LAION-400M [52] for RA. Adhering to the realistic "validation-free protocol" (Sec. 3), we directly adopt hyperparameters (e.g., learning rate, weight decay, batch size) from [40] for all datasets. Notably, for temperatures, we follow [56] to search on semi-Aves, leading to the configurations: initializing $T_{\text{loss}} = 0.07$ and learning it during VLM finetuning, and setting $T_{\text{conf}} = 0.01$. We apply them throughout experiments on all datasets. All experiments are run on a single NVIDIA RTX 4090 (24GB) GPU with 50 GB disk space for data storage.

4.2 Experimental Results

SWIFT achieves state-of-the-art SSFSL performance. As shown in Tab. 2, our simple temperature-based remedy significantly improves existing SSL methods, e.g., FixMatch [55] and DebiasPL [65], enabling successful VLM finetuning. Our improved versions even surpass the SOTA SSL method FineSSL [17], reinforcing our motivation that finetuning VLM performs better than prompting a frozen VLM. Importantly, SWIFT outperforms the SOTA FSL method SWAT [40], which also exploits retrieved open data for VLM finetuning. The strong accuracy gains demonstrate our successful exploitation of unlabeled data by using temperatures. Furthermore, SWIFT even rivals fully supervised learning, which finetunes VLM using few-shot labeled data, unlabeled data (with ground-truth labels), and retrieved open data in a single training stage. The results highlight the advantages of SWIFT's multi-stage training pipeline.

Ablation study validates the generalization of SWIFT. Tab. 3 highlights the significant progressive performance gains by each stage in SWIFT. Importantly, SWIFT not only improves VLM finetuning, but also generalizes to self-supervised pretrained vision foundation models, e.g., DINOv2 [47], achieving even better SSFSL performance. In addition, we show in the Supplement Tab. 9 that SWIFT also improves the stronger SSL method, DebiasPL [65], serving as a plug-and-play framework for improving existing SSL methods.

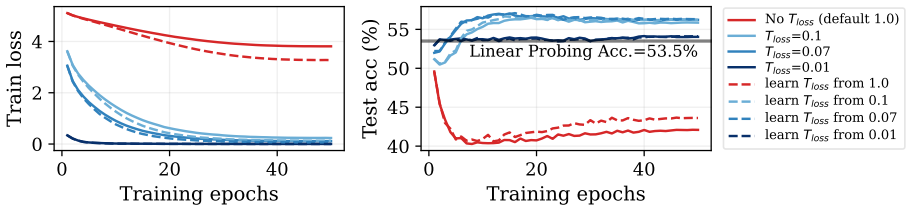

Fig. 6: Loss temperature T_{loss} strengthens training supervision. We compare the training loss and test accuracy under different settings of T_{loss} when finetuning the OpenCLIP on 16-shot labeled examples from the semi-Aves dataset [57]. Notably, compared to finetuning without temperature (default value of 1.0), using a small loss temperature (e.g., 0.1 or 0.07) effectively accelerates the training loss reduction, improving test accuracy over the few-shot linear probing baseline. In addition, learning the T_{loss} dynamically during the training (dashed lines) outperforms using a fixed T_{loss} (solid lines), due to increased flexibility. Supplementary Fig. 14 studies learning T_{loss} from various initial values with three random seeds, confirming our observations.

Using a small loss temperature T_{loss} enables effective VLM finetuning. Fig. 6 studies the effect of loss temperature when finetuning the VLM on few-shot labeled data. The results clearly show that adopting T_{loss} significantly accelerates the reduction of the training loss and improves test accuracy. Supplementary Tab. 13 provides per-dataset accuracy improvements, confirming the robustness of our proposed temperature-based solution.

Using a confidence temperature T_{conf} enables robust utilization of unlabeled data. Fig. 7 studies the utilization rate, pseudo-label accuracy, and test accuracy when finetuning VLM using FixMatch under various confidence temperatures. The results highlight the cruciality of tuning the confidence threshold σ for utilizing unlabeled data when no confidence temperature is used (red curves), which is impractical in the realistic "validation-free" (Fig. 2). In contrast, confidence temperature provides a robust solution, as evidenced by the consistent and strong accuracy gains when setting T_{conf} across a broad range (darker blue lines). The Supplementary Tab. 13 provides more detailed results on each dataset, further validating the robustness of using temperatures.

Temperature enables learning a better classifier, which improves subsequent semi-supervised finetuning. Tab. 4 compares different classifier initialization for stage-1 and stage-2 training of SWIFT. Results show that, while using text embedding [48] outperforms random initialization [55, 82], further adding our loss temperature notably improves performance. In addition, our SWIFT adopts the stage-1 learned classifier for the subsequent stage-2 finetuning, achieving better accuracy than using classifier initialized with text-embedding. Our results are similar to the observations in [33], which find that learning the classifier improves robust VLM finetuning, despite they did not explore temperatures.

Further analysis. We provide additional in-depth analyses in the Supplement. Below, we summarize our key findings. Supplementary Fig. 9 and Fig. 10 compare the distribution of confidence scores of more pretrained models (ImageNet-pretrained vs. OpenCLIP and CLIP), and different architectures (ResNet vs.

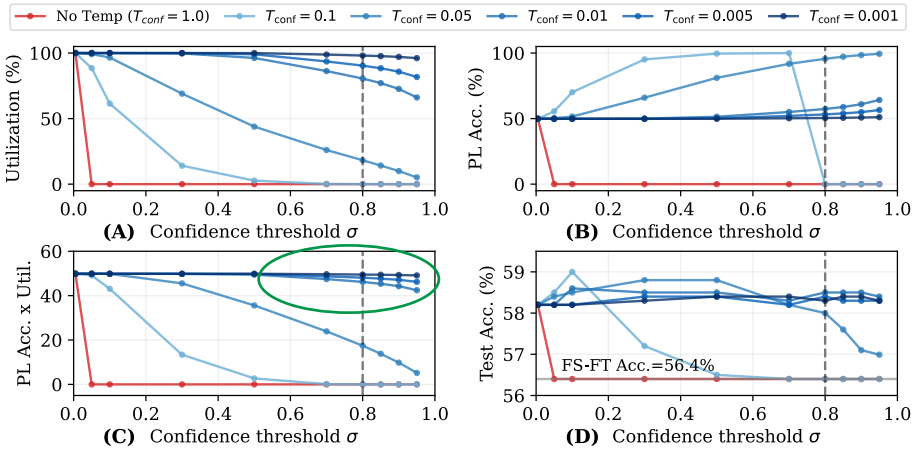

Fig. 7: Confidence temperature T_{conf} enables robust utilization of unlabeled data. Building on the T_{loss} in Fig. 6, we experiment with FixMatch on OpenCLIP with 16-shot data from the semi-Aves dataset under various T_{conf} settings. **(A)**: small T_{conf} consistently improves utilization rate (percentage of selected pseudo-labels with confidence $>$ a threshold σ, averaged over 50 training epochs); **(B)**: the pseudo-label accuracy increases with a large confidence threshold and small temperatures; **(C)** shows the product of the utilization rate and pseudo-label accuracy, highlighting the benefits of using small T_{conf} for selecting a greater quantity of correct pseudo-labels for SSL (cf. green circle); **(D)**: setting the T_{conf} in a broad range within $[0.001, 0.05]$ significantly improves test accuracy over the FS-FT baseline [40]. Importantly, using a small T_{conf} (e.g., 0.01) yields consistently strong accuracy gains across various thresholds.

ViT). The results confirm that the observed "flat softmax probabilities" arise from the contrastive pretraining objective rather than from architectural differences. Supplementary Fig. 11 visualizes the distribution of softmax probabilities after stage-1 training, confirming that logits remain flat and temperatures are needed for subsequent semi-supervised finetuning. In addition, Supplementary Fig. 12 studies using a small confidence threshold rather than using a confidence temperature, and Fig. 13 further experiments on the adaptive threshold SSL method, FreeMatch [67]. The results show that while these methods improve the utilization of unlabeled data, they yield poor test accuracy due to the lack of training supervision, highlighting the importance of using a loss temperature. Supplementary Fig. 14 compares different initial value for learning the loss temperature T_{loss}. The results show that using a small T_{loss} yields consistent and strong accuracy gains with small standard deviations across three random seeds. Supplementary Tab. 7 and Tab. 8 extensively study the effect of loss temperature across different learning paradigms (linear probing, few-shot and semi-supervised finetuning), different pretrained models (ImageNet-pretrained and VLMs), and different model architectures (ResNet vs. ViT), validating the importance of loss temperature when adapting a VLM across all settings. Moreover, Supplementary Tab. 11

Table 4: Comparison of classifier initialization methods for stage-1 linear probing and stage-2 semi-supervised finetuning (SS-FT) in SWIFT. We experiment with the VLM OpenCLIP ViT-B/32 across five datasets. **Left:** compared to prior linear probing methods that either use random initialization [49] or text embedding [48] to initialize the classifier weights, our approach, which adopts a learnable T_{loss} initialized to 0.07, achieves 2-4% average accuracy gains. The results highlight the benefits of using temperature when adapting VLMs. **Right:** In contrast to prior SSFSL methods that either randomly initialize the classifier weights [55, 82] or use text embeddings [40], SWIFT uses the stage-1 learned classifier for semi-supervised finetuning, achieving over 2% accuracy gains, especially in lower shots. The results validate our stage-wise design. Superscripts denote improvements over the text embedding.

	stage-1 classifier init.	mean acc. over five datasets		
		4-shot	8-shot	16-shot
Linear probing	random [49]	22.5	36.4	47.3
	text [48]	54.9	57.6	59.9
	text + T_{loss} (ours)	**57.0**$^{+2.1}$	**61.2**$^{+3.6}$	**64.7**$^{+4.8}$

	stage-2 classifier init.	mean acc. over five datasets		
		4-shot	8-shot	16-shot
SS-FT	random [82]	65.3	60.9	66.8
	text [40]	66.0	71.4	77.0
	stage-1 learned (ours)	**68.8**$^{+2.8}$	**73.3**$^{+1.9}$	**77.3**$^{+0.3}$

studies the impact of the number of retrieved images, demonstrating improved performance with more retrieved data.

5 Impacts, Limitations, and Future Work

Broad Impacts. SSFSL has promising real-world applications, and our work provides a simple yet effective solution by leveraging open-source pretrained foundation models and open data. However, we acknowledge potential societal risks. First, using retrieved open data may cause the finetuned model to overgeneralize to open-set or anomalous inputs. Second, as with other work that builds on foundation models, our approach may inherit biases from the pretrained VLM, potentially raising fairness concerns.

Limitations and Future Work. While our work advances SSFSL within a realistic setup and resolves the failures of established SSL methods in VLM finetuning, several avenues for improvement remain. First, although the validation-free setup realistically simulates real-world scenarios where scarce labeled data do not support validation, future research could explore unlabeled and retrieved data to aid validation so to allow hyperparameter tuning for specific tasks. Moreover, while the datasets used are standard in the literature, they may not fully reflect the extreme natural class imbalances encountered in real-world applications. Future work could construct more diverse datasets to facilitate SSFSL research.

6 Conclusions

We study semi-supervised few-shot learning (SSFSL) from a realistic "auto-annotation" perspective, which motivates us to exploit VLMs and open data to improve performance. Our extensive experiments reveal that representative SSL methods fail to finetune VLM effectively. We identify the root cause of

failures: VLMs produce a "flat" distribution of softmax probabilities, resulting in zero utilization of unlabeled data and weak training supervision that hinders effective finetuning. To address these challenges, we propose a simple technique that uses temperatures to sharpen softmax output. In addition, we further exploit retrieved open data to boost SSFSL performance. Building on these insights, we present a simple SSFSL method, SWIFT, which effectively finetunes a VLM with stage-wise training. Extensive experiments demonstrate that SWIFT significantly outperforms existing methods by 5 average accuracy points, establishing a new state-of-the-art across five benchmarks.

Acknowledgements

This work was supported by the Science and Technology Development Fund of Macau (0058/2025/RIA2, 0067/2024/ITP2), the University of Macau (SRG2023-00044-FST), the Institute of Collaborative Innovation, and CK Foundation. The authors thank Pardis Taghavi for discussions, the CSE Department at Texas A&M University, and the advanced computing resources and consultation provided by Texas A&M High Performance Research Computing (HPRC). Part of this work used the Delta system at the National Center for Supercomputing Applications [award OAC 2005572] through allocation [CIS250837, CIS250928] from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN) (2020)
2. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In: International Conference on Learning Representations (ICLR) (2020)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019)
4. Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., Kurakin, A.: Adamatch: A unified approach to semi-supervised learning and domain adaptation. In: The Eleventh International Conference on Learning Representations (ICLR) (2022)
5. Cai, Z., Ravichandran, A., Maji, S., Fowlkes, C., Tu, Z., Soatto, S.: Exponential moving average normalization for self-supervised and semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
6. Cascante-Bonilla, P., Tan, F., Qi, Y., Ordonez, V.: Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2021)

7. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks **20**(3), 542–542 (2009)
8. Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: Plot: Prompt learning with optimal transport for vision-language models. In: International Conference on Learning Representations (ICLR) (2023)
9. Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., Savvides, M.: Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In: The Eleventh International Conference on Learning Representations (ICLR) (2023)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML) (2020)
11. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in Neural Information Processing Systems (NeurIPS) (2020)
12. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
13. Chung, J., Chen, I.Y.: Enhancing semi-supervised learning with zero-shot pseudolabels. arXiv preprint arXiv:2502.12584 (2025)
14. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
15. Dong, X., Ouyang, T., Liao, S., Du, B., Shao, L.: Pseudo-labeling based practical semi-supervised meta-training for few-shot learning. IEEE Transactions on Image Processing (TIP) (2024)
16. Du, C., Han, Y., Huang, G.: Simpro: A simple probabilistic framework towards realistic long-tailed semi-supervised learning. arXiv preprint arXiv:2402.13505 (2024)
17. Gan, K., Wei, T.: Erasing the bias: Fine-tuning foundation models for semi-supervised learning. Forty-first International Conference on Machine Learning (ICML) (2024)
18. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision (IJCV) **132**(2), 581–595 (2024)
19. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning (ICML) (2017)
20. Guo, L.Z., Li, Y.F.: Class-imbalanced semi-supervised learning with adaptive thresholding. In: International Conference on Machine Learning (ICML) (2022)
21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
23. Helber, P., Bischke, B., Dengel, A., Borth, D.: Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE (2018)

24. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: ICLR (2019)
25. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
26. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: The Eleventh International Conference on Learning Representations (ICLR) (2020)
27. Huang, K., Geng, J., Jiang, W., Deng, X., Xu, Z.: Pseudo-loss confidence metric for semi-supervised few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
28. Iscen, A., Caron, M., Fathi, A., Schmid, C.: Retrieval-enhanced contrastive vision-text models. In: International Conference on Learning Representations (ICLR) (2024)
29. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: ICLR (2017)
30. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning (ICML) (2021)
31. Kong, S., Ramanan, D.: Opengan: Open-set recognition via open data generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 813–822 (2021)
32. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: IEEE International Conference on Computer Vision (ICCV) Workshops (2013)
33. Kumar, A., Raghunathan, A., Jones, R.M., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (ICLR) (2022)
34. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: The Eleventh International Conference on Learning Representations (ICLR) (2017)
35. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, International Conference on Machine Learning (ICML) (2013)
36. Li, X., Sun, Q., Liu, Y., Zhou, Q., Zheng, S., Chua, T.S., Schiele, B.: Learning to self-train for semi-supervised few-shot classification. Advances in Neural Information Processing Systems (NeurIPS) (2019)
37. Lin, Z., Yu, S., Kuang, Z., Pathak, D., Ramanan, D.: Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
38. Ling, J., Liao, L., Yang, M., Shuai, J.: Semi-supervised few-shot learning via multi-factor clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
39. Liu, H., Son, K., Yang, J., Liu, C., Gao, J., Lee, Y.J., Li, C.: Learning customized visual models with retrieval-augmented knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
40. Liu, T., Zhang, H., Parashar, S., Kong, S.: Few-shot recognition via stage-wise retrieval-augmented finetuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
41. Ma, C., Dong, W., Xu, C.: Tenet: Beyond pseudo-labeling for semi-supervised few-shot learning. Machine Intelligence Research pp. 1–13 (2025)

42. Ma, Y., Hua, W., Kong, S.: Towards auto-annotation from annotation guidelines: A benchmark through 3d lidar detection. arXiv preprint arXiv:2506.02914 (2025)
43. Madan, A., Peri, N., Kong, S., Ramanan, D.: Revisiting few-shot object detection with vision-language models. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets & Benchmark Track (2024)
44. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv:1306.5151 (2013)
45. Menghini, C., Delworth, A., Bach, S.: Enhancing clip with clip: Exploring pseudolabeling for limited-label prompt tuning. Advances in Neural Information Processing Systems (NeurIPS) (2023)
46. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: International Conference on Learning Representations (ICLR) (2021)
47. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (TMLR) (2024)
48. Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., Caverlee, J., Kong, S.: The neglected tails in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
50. Saha, O., Horn, G.V., Maji, S.: Improved zero-shot classification by adapting vlms with text descriptions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
51. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems (NeurIPS) (2022)
52. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
53. Shen, Q., Zhao, Y., Kwon, N., Kim, J., Li, Y., Kong, S.: Solving instance detection from an open-world perspective. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9901–9910 (2025)
54. Silva-Rodriguez, J., Hajimiri, S., Ayed, I.B., Dolz, J.: A closer look at the few-shot adaptation of large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
55. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems (NeurIPS) (2020)
56. Su, J.C., Cheng, Z., Maji, S.: A realistic evaluation of semi-supervised learning for fine-grained classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
57. Su, J.C., Maji, S.: The semi-supervised inaturalist-aves challenge at fgvc7 workshop. arXiv:2103.06937 (2021)

58. Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-clip: A clip model focusing on wherever you want. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR) (2024)
59. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in Neural Information Processing Systems (NeurIPS) (2017)
60. Tu, W., Deng, W., Campbell, D., Gould, S., Gedeon, T.: An empirical study into what matters for calibrating vision-language models. In: International Conference on Machine Learning (ICML) (2024)
61. Tu, W., Deng, W., Gedeon, T.: A closer look at the robustness of contrastive language-image pre-training (clip). In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
62. Wallingford, M., Ramanujan, V., Fang, A., Kusupati, A., Mottaghi, R., Kembhavi, A., Schmidt, L., Farhadi, A.: Neural priming for sample-efficient adaptation. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
63. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM International Conference on Multimedia (2017)
64. Wang, H., Liu, T., Kong, S.: Enabling validation for robust few-shot recognition. arXiv preprint arXiv:2506.04713 (2025)
65. Wang, X., Wu, Z., Lian, L., Yu, S.X.: Debiased learning from naturally imbalanced pseudo-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
66. Wang, Y., Chen, H., Fan, Y., SUN, W., Tao, R., Hou, W., Wang, R., Yang, L., Zhou, Z., Guo, L.Z., Qi, H., Wu, Z., Li, Y.F., Nakamura, S., Ye, W., Savvides, M., Raj, B., Shinozaki, T., Schiele, B., Wang, J., Xie, X., Zhang, Y.: USB: A unified semi-supervised learning benchmark for classification. In: Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2022)
67. Wang, Y., Chen, H., Heng, Q., Hou, W., Fan, Y., Wu, Z., Wang, J., Savvides, M., Shinozaki, T., Raj, B., Schiele, B., Xie, X.: Freematch: Self-adaptive thresholding for semi-supervised learning. In: The Eleventh International Conference on Learning Representations (ICLR) (2023)
68. Wang, Y., Duan, M., Kong, S.: Attention to the burstiness in visual prompt tuning! In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4253–4263 (2025)
69. Wei, T., Gan, K.: Towards realistic long-tailed semi-supervised learning: Consistency is all you need. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
70. Wei, X.S., Xu, H.Y., Zhang, F., Peng, Y., Zhou, W.: An embarrassingly simple approach to semi-supervised few-shot learning. Advances in Neural Information Processing Systems (NeurIPS) (2022)
71. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust fine-tuning of zero-shot models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
72. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

73. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
74. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. In: International Conference on Learning Representations (ICLR) (2024)
75. Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.F., Sun, B., Li, H., Jin, R.: Dash: Semi-supervised learning with dynamic thresholding. In: International Conference on Machine Learning (ICML). PMLR (2021)
76. Yang, L., Zhao, Z., Zhao, H.: Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). IEEE (2025)
77. Yu, T., Lu, Z., Jin, X., Chen, Z., Wang, X.: Task residual for tuning vision-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
78. Yu, T., Lu, Z., Jin, X., Chen, Z., Wang, X.: Task residual for tuning vision-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
79. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Advances in Neural Information Processing Systems (NeurIPS) (2021)
80. Zhang, C., Stepputtis, S., Sycara, K., Xie, Y.: Enhancing vision-language few-shot adaptation with negative learning. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025)
81. Zhang, J., Wei, Q., Liu, F., Feng, L.: Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data. Forty-first International Conference on Machine Learning (ICML) (2024)
82. Zhang, P., Mai, Z., Nguyen, Q.H., Chao, W.L.: Revisiting semi-supervised learning in the era of foundation models. Advances in Neural Information Processing Systems (NeurIPS) (2025)
83. Zhang, P., Mai, Z., Nguyen, Q.H., Chao, W.L.: Revisiting semi-supervised learning in the era of foundation models. `https://github.com/OSU-MLB/SSL-Foundation-Models/blob/3b4eb48d4986456b6f3c6c5576ff25aff781026f/semilearn/core/criterions/cross_entropy.py#L23` (2025), commit 3b4eb48
84. Zhang, R., Fang, R., Gao, P., Zhang, W., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaptation of clip for few-shot classification. In: European Conference on Computer Vision (ECCV) (2022)
85. Zheng, M., You, S., Huang, L., Wang, F., Qian, C., Xu, C.: Simmatch: Semi-supervised learning with similarity matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
86. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) (2022)