This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DPA: Dual Prototypes Alignment for Unsupervised Adaptation of Vision-Language Models

Eman Ali[1,2] Sathira Silva[1] Muhammad Haris Khan[1]
[1] Mohamed Bin Zayed University of Artificial Intelligence [2] Alexandria University
{eman.ali, sathira.silva, muhammad.haris}@mbzuai.ac.ae

Abstract

Vision-language models (VLMs), e.g., CLIP, have shown remarkable potential in zero-shot image classification. However, adapting these models to new domains remains challenging, especially in unsupervised settings where labeled data is unavailable. Recent research has proposed pseudo-labeling approaches to adapt CLIP in an unsupervised manner using unlabeled target data. Nonetheless, these methods struggle due to noisy pseudo-labels resulting from the misalignment between CLIP's visual and textual representations. This study introduces DPA, an unsupervised domain adaptation method for VLMs. DPA introduces the concept of dual prototypes, acting as distinct classifiers, along with the convex combination of their outputs, thereby leading to accurate pseudo-label construction. Next, it ranks pseudo-labels to facilitate robust self-training, particularly during early training. Finally, it addresses visual-textual misalignment by aligning textual prototypes with image prototypes to further improve the adaptation performance. Experiments on 13 downstream vision tasks demonstrate that DPA significantly outperforms zero-shot CLIP and the state-of-the-art unsupervised adaptation baselines.

1. Introduction

Context and background: Vision-language models (VLMs) [20, 27, 38, 56] have shown promising potential in zero-shot image classification. One notable example is CLIP [38]. Despite the impressive zero-shot capabilities of CLIP, its performance can be impacted by the discrepancy between the pretraining image-text pairs and the downstream task images [1, 38]. To address this limitation, several studies have attempted to enhance CLIP transfer performance on downstream tasks by leveraging limited labeled samples from the target domains [21, 22, 59–61]. However, in many practical applications, such as security and medical diagnostics, collecting labeled samples can be particularly challenging due to high labeling costs and data privacy concerns. In such contexts, unsupervised learning presents a promising alternative.

Several contributions for adapting CLIP to a target domain using an unlabeled dataset have recently been introduced [17, 18, 25, 33, 37, 45]. A common approach involves leveraging CLIP to generate pseudo-labels, which are then used to fine-tune CLIP in an unsupervised manner. However, this method encounters significant challenges due to the noisy pseudo-labels generated by CLIP. Despite efforts by recent methods for the unsupervised adaptation of CLIP, a significant modality gap persists between the text and vision representations [17, 29], leading to inaccurate pseudo-labels causing confirmation bias during adaptation.

Motivation: Recent studies highlight a significant factor behind the performance issues of CLIP in unsupervised adaptation scenarios: the visual domain gap between the source images used to train CLIP and the target images typically occurs when the target samples originate from an uncommon domain [17]. As shown in Fig. 1-(a), the t-SNE projection of visual and textual embeddings on the EuroSAT dataset [12] reveals a significant misalignment, resulting in misclassifications. Contemporary methods (e.g., [17]) attempt to address this misalignment by learning a projection space to mitigate the issue and using label propagation [19] to improve pseudo-labels. However, they could be limited as they are either computationally expensive and/or show suboptimal performance in an inductive setting. Further, on datasets with a large number of classes, they struggle to perform well, necessitating turning off label propagation and using model predictions. Effectively addressing the challenges of unsupervised adaptation of CLIP in an inductive setting is crucial for improving CLIP's performance in unsupervised adaptation scenarios.

Our proposal: To address these challenges, we propose DPA, a novel unsupervised adaptation method for VLMs (see Fig. 2). DPA aims at addressing the domain gap between visual and textual representations in downstream tasks. DPA introduces the idea of dual prototypes, namely image and textual prototypes, which act as distinct classi-

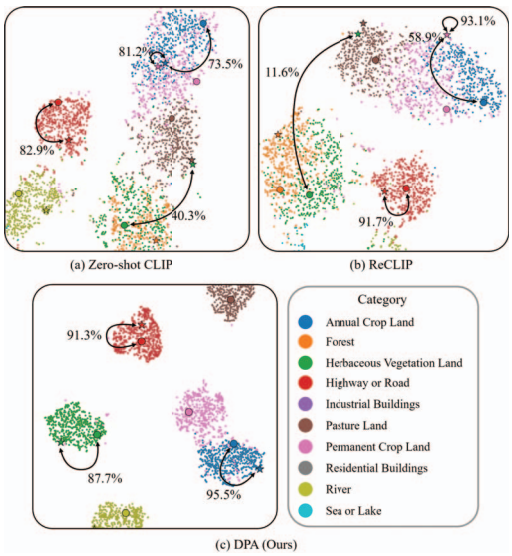


Figure 1. Comparison of t-SNE projections for zero-shot CLIP [38], ReCLIP [17], and DPA visual embeddings, along with their corresponding visual (circle ●) and textual (star ★) prototypes on the EuroSAT dataset. The visual prototypes are computed as the mean of each cluster. For clarity, class-agnostic and redundant features are removed from all embeddings using a fixed projection prior to applying the t-SNE projection, following [17]. The cosine similarities between visual and textual prototypes and the well-separated visual clusters, as illustrated in (c), demonstrate the superior performance of our method in capturing both inter-modal and intra-modal alignment compared to the existing approaches depicted in (a) and (b).

fiers, and their outputs are fused via a convex combination towards generating accurate pseudo-labels. There are two main reasons for introducing dual prototypes. Firstly, the misalignment between the image representation and its textual representation in zero-shot CLIP, due to domain shift, often leads to inaccurate pseudo-labels (see Fig. 1-(a)). Secondly, image prototypes, which tend to be less affected by noise, as shown in Fig. 1, are generally closer to the true image representation than textual prototypes. Additionally, we tackle the challenge of misalignment between visual and textual embeddings by aligning textual prototypes with image prototypes. This alignment process further enhances the performance of unsupervised CLIP adaptation.

Contributions: 1) We propose a novel dual-modality prototypic alignment framework, namely DPA for adapting VLMs in an unsupervised manner. Our approach introduces a dual-prototype pseudo-labeling strategy, where the visual prototypes serve as non-parametric classifiers to bring target visual embeddings closer to their respective class prototypes. The learnable textual prototypes, initialized by ensembling different prompt templates using zero-shot CLIP, provide rich task-specific semantics that CLIP visual embeddings typically lack. **2)** We propose ranking pseudo-labels in the classification loss to mitigate noise, particularly during early training. Furthermore, we address visual-textual misalignment by aligning textual prototypes with image prototypes, ensuring the separation of different semantics while aligning with relevant ones. **3)** Extensive experiments on 13 downstream vision tasks demonstrate consistent and significant performance enhancements over zero-shot CLIP and the state-of-the-art baselines.

2. Related Work

Large-scale Vision-Language (VL) Models: Among other vision-language models [6, 8, 16, 20, 38, 41, 56], CLIP [38] stands out as a pioneering example, aligning visual and textual features via a contrastive objective on a vast web-crawled collection of image-text pairs. This alignment empowers CLIP to generalize to diverse downstream classification tasks effectively. Building on CLIP's success, subsequent research has explored methods for efficiently transferring the pre-trained model to handle diverse downstream tasks with limited labeled target data. These approaches often leverage vision-specific adapters [10, 58, 59, 63], language-specific adapters [22, 60, 61], or both [21]. However, these techniques require a minimum number of labeled samples, which poses challenges due to data scarcity, annotation expenses, privacy issues, and practical limitations. Recent advancements have seen the emergence of unsupervised adaptation techniques, focusing on tailoring CLIP to target tasks using unlabeled datasets [17, 18, 25, 33, 37, 45]. However, these methods still depend on additional supervision signals, such as fine-tuning classifiers with a large language model such as GPT-3, as seen in LaFTer [33]. They also often require significant computational resources, as exemplified by MUST [25]. ReCLIP [17] tackles misaligned embeddings through source-free domain learning, utilizing a projection space to align these embeddings and employing pseudo-labels for self-training. Despite progress, there remains significant potential to improve the performance of unsupervised adaptation techniques across various image classification benchmarks. In contrast, our work introduces a fully label-free approach for adapting VLMs to a target task by generating more accurate pseudo-labels grounded on image and text prototypes.

Prototype-based Learning: Prototype-based learning uses a fixed set of distinctive prototypes representing the data, and then the learning is achieved by comparing the test samples directly with these prototypes. Prototype-based learning has been extensively studied in various contexts, including few-shot classification [28, 42], unsupervised learning [50, 54] and supervised classification [32, 55]. Our method innovatively incorporates two types of prototypes, image prototypes, and textual prototypes, to harness both prototypes to ultimately generate accurate pseudo-labels and effectively adapt CLIP to the target dataset.

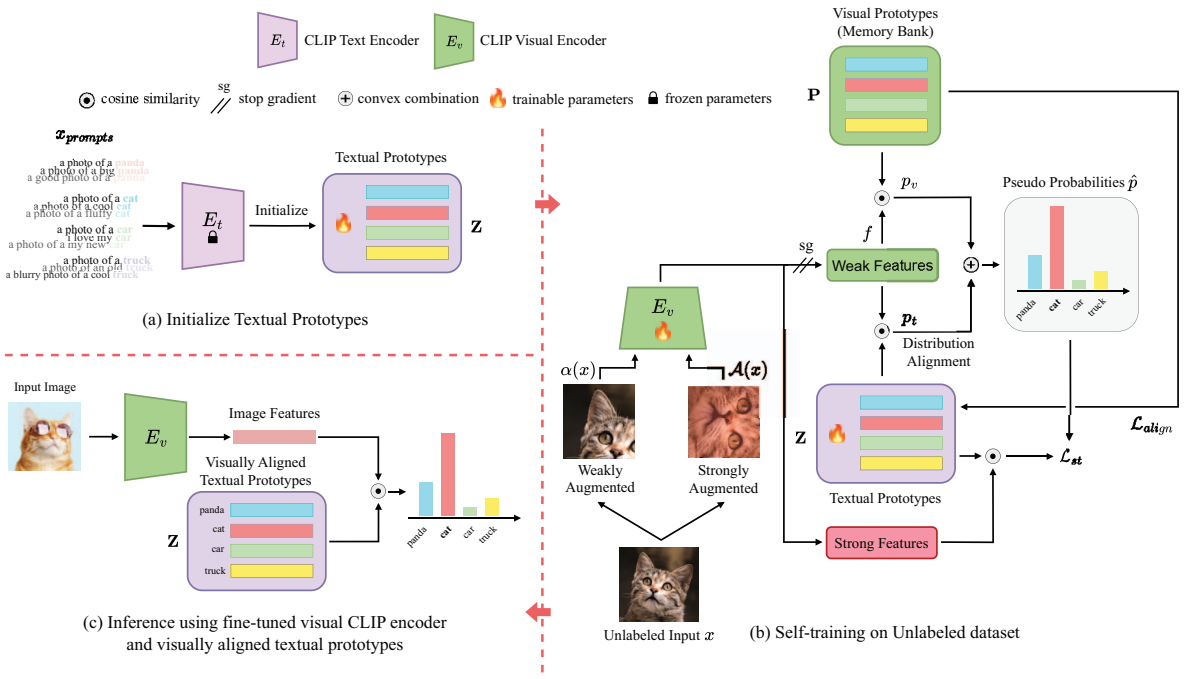


Figure 2. The overall framework of DPA. (a) Given a target dataset, DPA utilizes a set of carefully designed prompts to initialize the textual prototypes using the CLIP's zero-shot textual encoder E_t. (b) To achieve effective self-training, DPA introduces dual prototypes, namely image and textual prototypes, that behave like two distinct classifiers, and it fuses their outputs via convex combination to form accurate pseudo-labels (PLs). Moreover, it ranks PLs for classification loss to alleviate noisy PLs impact during early self-training. Finally, DPA aligns textual and visual prototypes to adeptly adjust to the target feature semantic relations learned by the visual encoder. (c) During inference, DPA discards the visual prototypes and relies solely on the textual prototypes for prediction.

Pseudo-labeling: Pseudo-labeling is popular in various domains, including vision [40, 57] and vision-language tasks [17, 18, 25, 33]. The pseudo-labeling pipeline of DPA leverages consistency regularization [43] to produce consistent predictions across different data augmentations, thereby enhancing pseudo-labeling accuracy.

3. Methodology

This study addresses the task of unsupervised adaptation of vision-language models (i.e., CLIP) for image classification. Pre-trained CLIP consists of visual encoder E_v and textual encoder E_t. The target dataset $\mathcal{D}_t = \{\mathcal{X}_t\}$ consists of unlabeled images $\{x_i\}_{i=1}^N$, where $x_i \in \mathcal{X}_t$, and we have unique class names $\mathcal{C} = \{c_j\}_{j=1}^C$ for the unlabeled target dataset. We use a pre-trained CLIP to process the unlabeled target data $\mathcal{D}_t$. Initially, the source model assigns a pseudo-label to each unlabeled target image x_i. During the adaptation phase, we employ two prototypes to generate pseudo-labels for adapting CLIP to the target domain.

3.1. Zero-shot Visual Classification in CLIP

CLIP achieves impressive zero-shot performance on image classification tasks by learning a joint embedding space that aligns image and text representations. During pre-training, CLIP minimizes a symmetric contrastive loss between semantically similar image-text pairs $(x_i^s, \mathbf{t}_i^s)_{i=1}^n$ sampled from a large-scale source dataset $\mathcal{D}_s$:

$$\sum_i -\log \frac{\exp(f_i^{sT} \cdot z_i/\tau)}{\sum_j^m \exp(f_i^{sT} \cdot z_j/\tau)} - \log \frac{\exp(z_i^T \cdot f_i^s/\tau)}{\sum_j^m \exp(z_j^T \cdot f_i^s/\tau)}, \tag{1}$$

where $f_i^s = E_v(x_i^s)$ and $z_i = E_t(\mathbf{t}_i^s)$, with both $f_i^s \in \mathbb{R}^{1\times d}$ and $z_i \in \mathbb{R}^{1\times d}$. The parameter τ represents a learned temperature, while m indicates the mini-batch size. For inference, given an unlabeled target image x and a list of class names $\mathcal{C}$, we leverage the pre-trained textual encoder E_t to generate text embeddings for each class. Specifically, a series of well-crafted prompts are fed into E_t, producing a set of text embeddings for each class, denoted as $\mathbf{Z}$ where $\mathbf{Z} \in \mathbb{R}^{C\times d}$. The zero-shot prediction $\hat{y}$ is then obtained as follows:

$$\hat{y} = \arg\max_c (f \cdot \mathbf{Z}^T) \tag{2}$$

3.2. Proposed Framework (DPA)

Motivation: While zero-shot CLIP shows impressive performance across diverse domains, it still exhibits notable limitations compared to models adapted under supervised conditions due to the misalignment between image and corresponding textual representations [38]. As depicted in

Fig. 1-(a), this discrepancy is evident in CLIP when applied to uncommon domains. Such disparities can lead to inaccurate pseudo-labels, especially when utilizing zero-shot CLIP to generate pseudo-labels for the unsupervised adaptation [17]. Our objective is to adapt CLIP to the target domain using its unlabeled samples by bridging the gap between image representations and their textual representations, all without dependence on ground-truth labels. Simultaneously, we seek to achieve efficient parameter-based unsupervised adaptation for CLIP. Our objective relies on two key components: (1) ensuring the accuracy of the pseudo-labels generated for the unlabeled samples and (2) aligning textual representations with their corresponding visual representations. To generate the pseudo-labels for adapting CLIP to the target domain, we introduce dual prototypes that act as two different classifiers and fuse their complementary outputs via a convex combination. Furthermore, we propose visual-textual prototype alignment towards further improving the adaptation performance.

Textual Prototypes Construction: To create textual prototypes, we obtain the textual representation for each class in the target domain. This is achieved by incorporating each class name c_j into a predefined prompt template. It is then processed by CLIP's text encoder to produce a textual representation z_j. We utilize k distinct prompts for each class, resulting in k corresponding textual representations $z_j \in \mathbb{R}^{k\times d}$, where d is the dimensionality of the embedding space. The textual prototype $\mathbf{Z}_j$ for class c_j is constructed by averaging these k individual textual representations:

$$\mathbf{Z}_j = \frac{1}{k}\sum_{i=1}^{k} z_{ji} \tag{3}$$

where $\mathbf{Z}_j \in \mathbb{R}^{1\times d}$. Finally, we define the textual prototypes for the set of C classes as $\mathbf{Z} \in \mathbb{R}^{C\times d}$. This approach enables us to capture a comprehensive and robust representation of class semantics by consolidating textual representations derived from multiple prompts. Finally, these textual prototypes act as the initialization for a parametric classifier that is updated during training alongside CLIP parameters.

Image Prototypes Construction: To construct the image prototypes, CLIP's image encoder processes a weakly-augmented unlabeled image $\alpha(x)$ and generates a feature representation $f = E_v(\alpha(x))$, where $f \in \mathbb{R}^{1\times d}$. Without class labels, we utilize Eq. (1) and Eq. (2) to generate pseudo-labels $\hat{y}$ from zero-shot CLIP. To obtain the image prototype $\mathbf{P}_j$ for a class c_j, we average the feature representation f of the unlabeled images assigned to class c_j as follows:

$$\mathbf{P}_j = \frac{1}{N_j}\sum_{i=1}^{N} \mathbb{I}\{\hat{y}_i = j\}\, f_i \tag{4}$$

where $N_j = \sum_{i=1}^{N}\mathbb{I}\{\hat{y}_i = j\}$. For the set of C classes, we define the image prototypes as $\mathbf{P} \in \mathbb{R}^{C\times d}$—the image prototypes function as a non-parametric classifier. Given that the pseudo-labels generated by zero-shot CLIP are initially noisy, our approach, inspired by previous works [49,53,62], employs a memory bank $\mathbf{MB}$. This memory bank facilitates the incremental update of image prototypes in a non-parametric way throughout the training process. Specifically, $\mathbf{MB}$ stores the feature representations of all unlabeled target samples along with their corresponding pseudo-labels: $\mathbf{MB} = \{(f_i, \hat{y}_i)\}_{i=1}^{N}$. As the image prototypes are non-parametric, we iteratively refine them using the memory bank as follows: (1) We initially calculate the image prototypes using Eq. (4), utilizing the image features and the pseudo-labels generated from zero-shot CLIP across the entire target dataset. (2) During training, we continuously update the memory bank by storing the image representations alongside their corresponding pseudo-labels. (3) At the end of each epoch, we update the prototypes $\mathbf{P}_j$ using Eq. (4), incorporating the image representations and their corresponding pseudo-labels from $\mathbf{MB}$. (4) We repeat steps (2) and (3) until the completion of the training process. DPA maximizes the utilization of all available training samples to enhance the image prototypes. This iterative process ensures that each prototype continuously evolves, progressively assimilating relevant knowledge from individual training samples throughout the model training phase while mitigating the error propagation commonly associated with pseudo-labels.

Pseudo-labels Generation: Both image and textual prototypes serve as distinct classifiers for the unlabeled target dataset. We generate a pseudo-label (PL) for the weakly-augmented unlabeled target image $\alpha(x)$ from these two prototypes. Initially, we determine the similarity between the image features and the textual prototypes as follows:

$$p_t = f \cdot \mathbf{Z}^T \tag{5}$$

Then, we find the similarity between the image features and the image prototypes as follows:

$$p_v = f \cdot \mathbf{P}^T \tag{6}$$

Finally, we fuse p_t and p_v to form $\hat{p}$ as follows:

$$\hat{p} = \beta \cdot DA(p_t) + (1-\beta)\cdot p_v \tag{7}$$

Following [26], we perform distribution alignment (DA) to prevent the model's prediction from collapsing onto specific classes. Here, $DA(p_t) = p_t/\bar{p}_t$, where $\bar{p}_t$ represents a running average of p_t during training, and β is a hyper-parameter. Finally, the PL is obtained as $\hat{y} = \arg\max_c \hat{p}$. This process of generating pseudo-labels improves the accuracy of the PL during training, indicating that these PLs may initially be noisy (see Fig. 4). Therefore, assigning uniform weights to all pseudo-labels could hinder the adaptation process. To address this challenge, we propose adjusting the classification loss weight for the pseudo-label $\hat{y}$

based on its similarity to the textual and image prototypes, as follows:

$$w_x = \langle f, \mathbf{Z}(\hat{y}) \rangle \langle f, \mathbf{P}(\hat{y}) \rangle \tag{8}$$

$\langle \cdot, \cdot \rangle$ represents the cosine similarity between the image features and the corresponding prototypes derived from the PLs. Finally, we utilize the pseudo-label $\hat{y}$, obtained from a weakly-augmented image $\alpha(x)$, as self-supervision for the strongly-augmented counterpart $\mathcal{A}(x)$ as follows:

$$\mathcal{L}_{st} = - \mathbb{E}_{x \in \mathcal{X}_t} \left[w_x \cdot \sum_{j=1}^{C} \mathbb{I}\{\hat{y} = j\} log \left(p_{\mathcal{A}(x)} \right) \right], \tag{9}$$

where w_x represents the weight calculated from Eq. (8), and $p_{\mathcal{A}(x)} = E_v(\mathcal{A}(x)) \cdot \mathbf{Z}^T$ represents the probabilistic output for the strongly-augmented image $\mathcal{A}(x)$. To alleviate the confirmation bias induced by CLIP [48], we adopt "fairness" regularization as suggested in [25] as follows:

$$\mathcal{L}_{reg} = -\frac{1}{C} \sum_{j=1}^{C} \log \left(\bar{p}_{\mathcal{A}(x_j)} \right), \tag{10}$$

where $\bar{p}_{\mathcal{A}(x)}$ represents the model's average prediction from the strongly augmented images across the batch. By incorporating this fairness regularization, we aim to encourage the model to make more uniform predictions across classes, reducing the tendency to overfit the PLs and promoting a more balanced adaptation to the target domain. Eq. (9) and Eq. (10) are used to update both the image encoder and the textual prototypes.

Prototypes Alignment: During training, the textual prototypes undergo continual updates to align with the associated image representation, leveraging pseudo-labels using Eq. (9) and Eq. (10). Conversely, the refinement of the image prototypes occurs through averaging image features, as detailed in Section 3.2. We exploit the image prototypes to provide additional supervision for updating textual prototypes beyond pseudo-label alone. These image prototypes, less influenced by erroneous pseudo-labels and closer to image features (as illustrated in Fig. 1), offer extra guidance for refining textual prototypes. Our primary objective is to optimize the relationship between image prototypes $\mathbf{P}$ and textual prototypes $\mathbf{Z}$ by maximizing the (cosine) similarity between these feature sets. To accomplish this optimization, we employ the InfoNCE loss [35] to directly assess the alignment between the corresponding feature vectors in $\mathbf{P}$ and $\mathbf{Z}$, as follows:

$$\mathcal{L}_{\text{align}} = - \sum_{j=1}^{C} \log \frac{\exp(\mathbf{P}_j \cdot \mathbf{Z}_j^T / \tau)}{\sum_r \exp(\mathbf{P}_j \cdot \mathbf{Z}_r^T / \tau)} \tag{11}$$

With this alignment, we maximize the (cosine) similarity between the image prototypes $\mathbf{P}$ and the textual prototypes $\mathbf{Z}$, encouraging the model to learn more closely aligned representations. The overall loss function used for training on the target data is, $\mathcal{L} = \lambda_1 \mathcal{L}_{st} + \lambda_2 \mathcal{L}_{reg} + \lambda_3 \mathcal{L}_{align}$, where $\lambda_1, \lambda_2, \lambda_3$ are non-learnable parameters.

4. Experiments

Datasets and Baselines: We extensively evaluate on 13 diverse datasets: ImageNet [7], Caltech101 [9], DTD [4], EuroSAT [12], FGVCAircraft [31], Food101 [3], Flowers102 [34], OxfordPets [36], SUN397 [52], StandfordCars [23], CIFAR10\100 [24], and UCF101 [44]. We perform comparative analysis with four state-of-the-art (SOTA) unsupervised adaptation methods: CLIP [38], UPL [18], POUF [45], and LaFTer [33].

Implementation Details: For all experiments, unless otherwise specified, we utilize a ViT/B-32 CLIP pre-trained by OpenAI [38]. In the context of unsupervised fine-tuning, we specifically target the layer-normalization weights of the image encoder and the textual prototypes. This strategy is effective and stable for adapting models with noisy supervision [2,46]. The text encoder of CLIP will be excluded after constructing the textual prototypes. Input images are standardized to a size of 224 × 224. During training, we apply a series of data augmentation techniques, including RandomResizedCrop, Flip, RandAugment [5] as a strong augmentation and Resize+RandomCrop for generating pseudo-labels. For testing, we extract a Center crop after resizing the image to 224. We utilize AdamW optimizer [30] with a cosine learning rate schedule. To ensure fair comparisons, we reproduced the results of UPL, POUF, and LaFTer using their publicly available code. Most state-of-the-art (SOTA) methods use different backbones, and UPL's reported results were only for 16 shots, unlike our approach, which provides results for the entire training dataset. For additional details on the implementation, see the supplementary materials.

Results: Table 1 shows that our approach consistently outperforms zero-shot CLIP, improving top-1 accuracy by $+7.83\%$ on average. Compared to UPL, which optimizes prompts, DPA still achieves consistent improvements with an average gain of $+8.24\%$ with minimal additional computational overhead despite slightly larger parameters. We observe that UPL's performance on the entire unlabeled training dataset is inferior to that of zero-shot CLIP. This discrepancy may stem from the fact that UPL generates offline pseudo-labels using zero-shot CLIP, which can lack confidence, particularly when dealing with fine-grained classes in certain datasets [33]. DPA also surpasses POUF and LaFTer, achieving average increases of $+7.47\%$ and $+3.93\%$, respectively, across the 13 datasets without requiring additional text corpora or further pre-training. We note significantly larger performance improvement on the EuroSAT dataset compared to other datasets. We believe this

Method	ImgNet	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	SUN	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP [38]	63.30	90.69	44.42	43.84	19.50	82.40	66.46	87.50	61.99	58.74	89.80	65.10	64.20	64.46
UPL [18]	58.22	92.36	45.37	51.88	17.07	84.25	67.40	83.84	62.12	49.41	91.26	67.41	62.04	64.05
POUF [45]	52.20	94.10	46.10	62.90	18.20	82.10	67.80	87.80	60.00	57.70	90.50	62.00	61.20	64.82
LaFTer [33]	61.63	94.39	50.32	69.96	19.86	82.45	72.43	84.93	65.87	57.44	94.57	69.79	65.08	68.36
DPA	**64.64**	**96.06**	**55.69**	**80.04**	**20.67**	**84.76**	**75.56**	**90.71**	**68.13**	**62.62**	**95.97**	**76.47**	**68.49**	**72.29**

Table 1. Comparison of state-of-the-art unsupervised adaptation methods using the ViT-B/32 backbone.

is because zero-shot CLIP exhibits limitations when applied to specialized, complex, or abstract tasks, such as satellite image classification [38]. This suggests there is potential for improvement through additional supervision. This behavior aligns with findings from previous studies. LaFTer achieved a remarkable +26.12% improvement over zero-shot CLIP. ReCLIP reported gains exceeding +26.96% (see Tab. 8). The 10% improvement by our method (DPA) over LaFTer highlights the effectiveness of our pseudo-labeling approach.

For Ablation Studies, we use 11 out of 13 datasets, excluding ImageNet and SUN397 due to their large size. Excluding these large-scale datasets allows us to perform more extensive experiments and analyses while maintaining computational feasibility.

Sensitivity Analyses of Hyperparameters: The sensitivity analyses of the hyperparameters λ_2, λ_3, and β on the EuroSAT dataset are presented in Fig. 3.

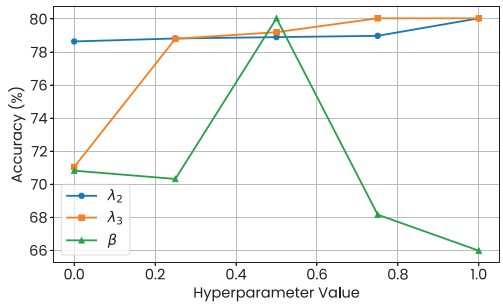


Figure 3. Sensitivity analysis of the hyperparameters λ_2, λ_3, and β on the accuracy (%) of DPA. We set $\lambda_1 = 1$ across all datasets (here EuroSAT is used) since $\mathcal{L}_{st}$ is our primary loss.

Results on Different Backbones We compare different backbones (ViT-B/16) in Tab. 2 [1].

Analyses of Model Components: We investigate the impact of different components through the training of four distinct models: **Base**: This model iteratively generates pseudo-labels using textual prototypes and incorporates regularization loss through self-training. **Center**: CLIP undergoes training with PLs generation using image and text prototypes, as detailed in Section 3.2, in addition to

[1]A fair comparison using RN50 is not possible, as our relevant baselines were specifically designed for the ViT image encoder architecture (For example, LaFTer [33] / ReCLIP [17])

Method	Caltech	DTD	ESAT	FGVCA	Flower	OxPets	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP	92.60	44.70	49.00	23.97	70.89	89.00	64.70	90.80	68.22	69.10	66.30
UPL	95.14	45.90	55.36	22.53	73.93	87.98	60.33	92.92	70.41	67.43	67.19
POUF	95.40	48.60	59.50	24.40	72.10	91.80	63.50	92.30	66.30	71.50	68.54
LaFTer	95.92	**54.79**	72.10	22.38	75.15	85.28	64.72	**96.50**	76.30	67.20	71.03
DPA	**96.09**	50.69	**81.22**	**25.14**	**78.68**	**93.19**	**68.45**	96.43	**78.10**	**74.62**	**74.26**

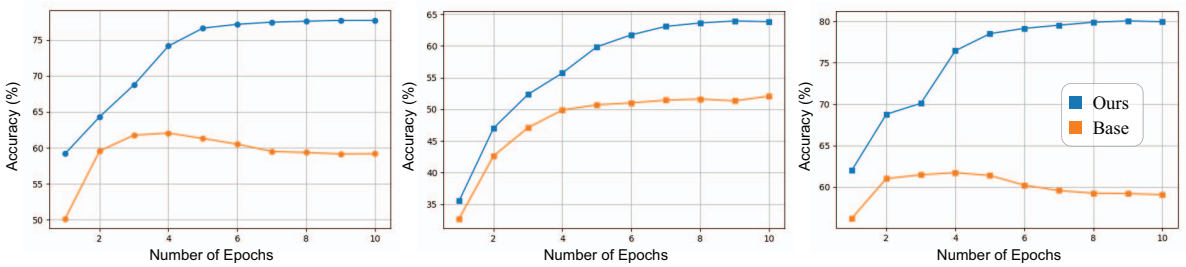
Table 2. Top-1 accuracy (%) comparison of DPA with SOTA unsupervised adaptation methods using CLIP-ViT-B/16.

building upon the foundation established in the Base model. **Center+w**: introduces the weighting strategy for **Center** model. **Center+w+Align**: introduces the alignment loss, as detailed in Section 3.2. As shown in Tab. 3, **Base** model, which relies solely on self-training, demonstrates a notable improvement compared to the zero-shot CLIP. However, as training progresses, the impact of confirmation bias becomes evident, leading to a decrease in the accuracy of the pseudo-labels, as illustrated in Fig. 4-(a). To mitigate the effects of confirmation bias, we introduce the pseudo-label generation strategy in the **Center** model. This approach enhances the quality of pseudo-labels throughout the training process, leading to a significant gain of $+7.28\%$ over zero-shot CLIP in average accuracy. Incorporating the weighting strategy in **Center+w** model results in an additional increase of $+8.01\%$ over zero-shot CLIP in accuracy. By assigning appropriate weights to the self-training loss, the model can better balance the contributions of high-confidence and low-confidence pseudo-labels, leading to more stable and practical training. Finally, including the alignment loss in **Center+w+Align** model leads to a further increase of $+8.58\%$ over zero-shot CLIP in average accuracy. By explicitly encouraging the alignment between image and text prototypes, the model learns more robust and transferable representations, which are crucial for effective unsupervised adaptation. Finally, we show a detailed analysis of the evolution of pseudo-label accuracy during training epochs for our method (DPA) compared to **Base** model in Fig. 4-(a). See more discussions in suppl.

Incorporating Image Prototypes as Pseudo-Label Source: We analyze incorporating image prototypes as an additional classifier. We exclude Eq. (7) and generate PLs solely based on text prototypes while preserving the alignment of image-text prototypes and the weighting strategy. We update $\mathbf{MB}$ solely using the pseudo-labels derived from the textual prototypes. Table 4 shows that the alignment and weighting strategies generally perform well

Method	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP	90.69	44.42	43.84	19.50	82.40	66.46	87.50	58.74	89.80	65.10	64.20	64.79
Base	93.57	48.99	61.94	19.20	84.10	68.45	90.24	59.25	95.95	73.55	65.45	69.15
Center	95.44	55.53	70.56	19.80	84.65	75.27	**90.71**	61.53	95.96	76.01	67.30	72.07
Center+w	95.46	54.54	**80.06**	19.56	84.63	75.44	90.49	61.19	**95.97**	75.92	67.51	72.80
Center+w+Align (DPA)	**96.06**	**55.69**	80.04	**20.67**	**84.76**	**75.56**	**90.71**	**62.62**	**95.97**	**76.47**	**68.49**	**73.37**

Table 3. Ablation study. See text for details.



(a) Pseudo-label Accuracy (b) Training Accuracy (c) Testing Accuracy

Figure 4. **Base** vs. DPA in (a) PLs, (b) training, and (c) testing accuracy on EuroSAT [12] dataset.

across datasets, with minor differences compared to the entire model, except for EuroSAT and Flowers-102.

Impact of Image Prototype Initialization: We evaluate three initializations for image prototype methods: mean, weighted mean, and similarity weighted mean [28] (see Tab. 5). The results demonstrate significant improvements over the baselines for all initialization strategies.

Robustness Analysis under Distribution Shifts: To evaluate robustness against natural distribution shifts, we show results on ImageNet-Sketch [47], ImageNet-Adversarial [14], and ImageNet-Rendition [13]. Initially, we evaluate a fine-tuned ViT-B/32 model on ImageNet using our method, achieving an accuracy of 64.64% on the ImageNet validation set. We explore transductive transfer learning [39, 51] with our approach. Transductive learning involves assuming access to unlabeled test images from the new distribution (see Tab. 6).

Evaluation of Fine-Tuning Techniques: Table 7 shows the results of various fine-tuning techniques for CLIP, including LoRA [15], K-Adaptation [11], and full model fine-tuning, and our approach, which involves fine-tuning only the layer normalization weights (see Tab. 7). The last row of the table shows the PLs of zero-shot CLIP used at the start of the training for constructing the initial image prototypes. In datasets characterized by highly noisy PLs, Tab. 7 underscores the significance of exclusively fine-tuning layer-normalization weights during adaptation. This approach helps to mitigate overfitting and enhances overall performance. Our findings indicate that fine-tuning solely the layer normalization weights is effective and stable for adapting models under noisy supervision.

Data Scale and Downstream Task Performance: We conducted an experiment using the CIFAR-10 dataset to evaluate the scalability of our method (DPA), compared to LaFTer, as illustrated in Fig. 5. To assess the robustness of our approach, we tested various portions of the dataset: 20%, 40%, 60%, and 80%. Our findings indicate that the DPA method consistently performs well across different dataset sizes, demonstrating its scalability regardless of the amount of data used.

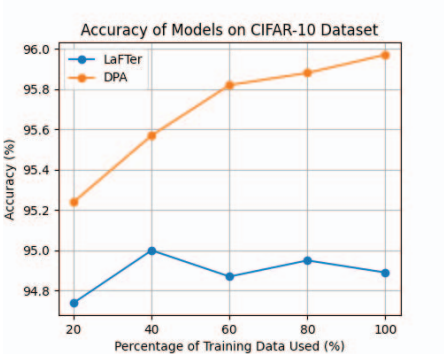


Figure 5. Top-1 accuracy (%) for the ViT-B/32 backbone on the CIFAR-10 dataset using 20%, 40%, 60%, 80%, and 100% of the training data for DPA.

Comparison with ReCLIP: We provide a comparison with a very recent method for unsupervised adaptation of VLMs, namely ReCLIP [17] (see Tab. 8). DPA consistently outperforms ReCLIP [17] in both inductive and transductive settings. Also, DPA has superior training efficiency and performance with fewer parameters and reduced computational complexity comparing the baselines (see Fig. 6).

Method	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP	90.69	44.42	43.84	19.50	82.40	66.46	87.50	58.74	89.80	65.10	64.20	64.79
DPA (w/o PLs Generation)	95.41	54.20	61.94	19.44	84.49	69.22	**91.17**	61.95	95.93	76.12	67.96	70.71
DPA	**96.06**	**55.69**	**80.04**	**20.67**	**84.76**	**75.56**	90.71	**62.62**	**95.97**	**76.47**	**68.49**	**73.37**

Table 4. Image prototypes vs. text prototypes for pseudo-label generation.

Method	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP	90.69	44.42	43.84	19.50	82.40	66.46	87.50	58.74	89.80	65.10	64.20	64.79
DPA (Weighted Mean)	96.00	54.52	68.34	20.31	84.70	73.81	90.49	**63.14**	95.96	75.81	68.04	71.92
DPA (Similarity Weighted Mean)	94.52	55.16	79.76	**21.12**	84.70	**77.06**	**90.84**	62.98	95.93	76.33	68.28	73.33
DPA (Mean)	**96.06**	**55.69**	**80.04**	20.67	**84.76**	75.56	90.71	62.62	**95.97**	**76.47**	**68.49**	**73.37**

Table 5. Performance of DPA with three different initializations for image prototypes.

Method	ImageNet	ImageNet-S	ImageNet-R	ImageNet-A
Zero-shot CLIP	63.30	42.10	**68.60**	32.10
CoOp (M =16)	**66.85**	40.44	64.45	30.62
CoOp (M =4)	66.34	41.48	65.78	31.34
DPA (Transductive)	64.64	**42.60**	68.40	**32.20**

Table 6. Transductive adaptation to distribution shift.

Method	Caltech	DTD	ESAT	FGVCA	Flower	OxPets	StCars	Avg
Zero-shot CLIP	90.69	44.42	43.84	19.50	66.46	87.50	58.74	58.74
DPA (LoRA [15])	90.87	50.32	65.16	19.53	66.87	87.71	57.59	62.58
DPA (K-Adaptation [11])	92.49	52.50	69.80	19.38	68.09	89.04	60.81	64.59
DPA (Full Tuning)	94.57	53.03	62.92	17.37	75.23	**91.14**	54.06	64.05
DPA (LN)	**96.06**	**55.69**	**80.04**	**20.67**	**75.56**	90.71	**62.62**	**68.76**
DPA (PLs accuracy of Zero-shot CLIP)	88.83	43.54	44.90	17.79	66.53	83.80	57.81	57.6

Table 7. Top-1 accuracy (%) comparison of DPA using different parameter-efficient fine-tuning methods. All experiments are conducted with ViT-B/32 as the backbone.

Method	Caltech	DTD	ESAT	FGVCA	Flower	OxPets	StCars	UCF	Avg
Zero-shot CLIP	90.69	44.42	43.84	19.50	66.46	87.50	58.74	64.20	59.42
Inductive									
ReCLIP† [17]	95.94	53.88	70.80	18.87	72.63	87.49	59.22	67.01	65.73
DPA	**96.06**	**55.69**	**80.04**	**20.67**	**75.56**	**90.71**	**62.62**	**68.49**	**68.73**
Transductive									
ReCLIP [17]	**92.43**	52.50	59.30	20.34	70.65	88.42	59.06	**69.13**	63.98
DPA	91.62	**57.13**	**71.84**	**20.76**	**74.14**	**91.11**	**62.58**	68.44	**67.20**

Table 8. Comparison of top-1 classification accuracy (%) of DPA with ReCLIP [17], a source-free unsupervised domain adaptation method for CLIP. † indicates that the model was originally trained in a transductive setting, but we trained it in an inductive setting for fair comparison.

Meth.	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	StCars	CIF.10	CIF.100	UCF	Avg
ZS	90.69	44.42	43.84	19.50	82.40	66.46	87.50	58.74	89.80	65.10	64.20	64.79
DPA	**96.06**	55.69	**80.04**	20.67	**84.76**	75.56	**90.71**	62.62	**95.97**	**76.47**	68.49	73.37
8-shot												
CoOp	93.96	61.82	77.17	26.73	78.84	88.71	85.61	66.20	94.63	75.91	77.66	75.20
MaPLe	93.23	34.99	65.25	21.96	80.51	68.78	81.85	58.51	88.23	71.14	69.18	66.69
16-shot												
CoOp	95.09	**67.55**	76.78	**30.93**	79.12	**93.34**	86.70	**70.73**	94.50	75.78	**79.51**	**77.28**
MaPLe	93.35	53.43	71.49	21.48	81.01	72.47	86.54	60.28	91.94	71.74	70.26	70.36

Table 9. Comparison with few-shot methods across 11 different datasets. DPA deliver competitive performance or outperform few-shot methods in 6 of 11 datasets.

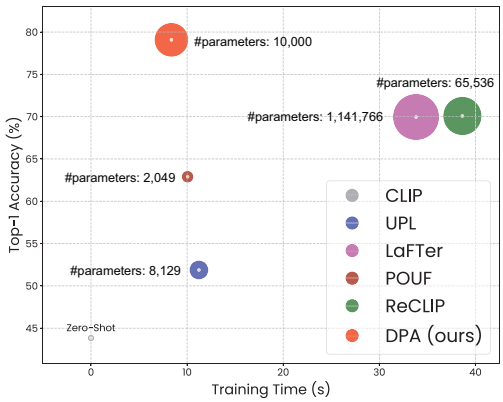


Figure 6. Efficiency comparison of DPA with baselines. The radius of each circle represents trainable parameters in each method.

Comparison with Few-Shot methods: Table 9 provides a performance comparison of DPA with the few-shot methods (CoOp [61] and MaPLe [21]). DPA show competitive performance or surpasses few-shot methods in 6 of 11 datasets.

5. Conclusion

We present DPA, an unsupervised domain adaptation method for VLMs that aims at bridging the domain gap between visual and textual representations. DPA introduces a novel idea of dual prototypes, which work as two distinct classifiers, and their outputs are fused via a convex combination. It also ranks pseudo-labels for robust self-training during early training. DPA also enhances the alignment of visual-textual prototypes to improve adaptation performance further. DPA shows improved pseudo-labeling accuracy, which leads to notable performance improvements on various downstream tasks. DPA significantly outperforms zero-shot CLIP and the state-of-the-art methods across 13 downstream vision tasks.

References

[1] Bang An, Sicheng Zhu, Michael-Andrei Panaitescu-Liess, Chaithanya Kumar Mummadi, and Furong Huang. PerceptionCLIP: Visual classification by inferring and conditioning on contexts. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

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

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 5

[6] Quan Cui, Boyan Zhou, Yu Guo, Weidong Yin, Hao Wu, Osamu Yoshie, and Yubo Chen. Contrastive vision-language pre-training with limited resources. In *European Conference on Computer Vision*, pages 236–253. Springer, 2022. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[8] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 2

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 5

[10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. 2

[11] Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. Parameter-efficient model adaptation for vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 817–825, 2023. 7, 8

[12] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 1, 5, 7

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 7

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 7

[15] 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 *International Conference on Learning Representations*, 2022. 7, 8

[16] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 2

[17] Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, et al. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2994–3003, 2024. 1, 2, 3, 4, 6, 7, 8

[18] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 1, 2, 3, 5, 6

[19] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2

[21] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 8

[22] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15190–15200, October 2023. 1, 2

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on computer vision workshops*, pages 554–561, 2013. 5

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[25] Junnan Li, Silvio Savarese, and Steven C. H. Hoi. Masked unsupervised self-training for label-free image classification. In *ICLR*, 2023. 1, 2, 3, 5

[26] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9475–9484, 2021. 4

[27] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. 1

[28] Kevin J Liang, Samrudhdhi B Rangrej, Vladan Petrovic, and Tal Hassner. Few-shot learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9089–9098, 2022. 2, 7

[29] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 1

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

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

[32] Pascal Mettes, Elise Van der Pol, and Cees Snoek. Hyperspherical prototype networks. *Advances in neural information processing systems*, 32, 2019. 2

[33] M. Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Mateusz Kozinski, Horst Possegger, Rogerio Feris, and Horst Bischof. Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 3, 5, 6

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

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[36] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 5

[37] Qi Qian, Yuanhong Xu, and Juhua Hu. Intra-modal proxy learning for zero-shot visual categorization with clip. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[38] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5, 6

[39] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. *Advances in neural information processing systems*, 26, 2013. 7

[40] Attaullah Sahito, Eibe Frank, and Bernhard Pfahringer. Better self-training for image classification through self-supervision. In *Australasian Joint Conference on Artificial Intelligence*, pages 645–657. Springer, 2022. 3

[41] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision*, pages 153–170. Springer, 2020. 2

[42] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2

[43] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 3

[44] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11):1–7, 2012. 5

[45] Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *International Conference on Machine Learning*, pages 33816–33832. PMLR, 2023. 1, 2, 5, 6

[46] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 5

[47] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[48] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14647–14657, 2022. 5

[49] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*, pages 499–515. Springer, 2016. 4

[50] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[51] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591, 2017. 7

[52] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on computer vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 5

[53] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*, pages 5423–5432. PMLR, 2018. 4

[54] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980, 2020. 2

[55] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Robust classification with convolutional prototype learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3474–3482, 2018. 2

[56] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 1, 2

[57] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 3

[58] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222, 2023. 2

[59] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022. 1, 2

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 2

[61] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 8

[62] Lihua Zhou, Nianxin Li, Mao Ye, Xiatian Zhu, and Song Tang. Source-free domain adaptation with class prototype discovery. *Pattern recognition*, 145:109974, 2024. 4

[63] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2605–2615, 2023. 2