ReCLIP: Refine Contrastive Language Image Pre-Training with Source Free Domain Adaptation

Xuefeng Hu[1] Ke Zhang[2] Lu Xia[2] Albert Chen[2] Jiajia Luo[2] Yuyin Sun[2]
Ken Wang[2] Nan Qiao[2] Xiao Zeng[2] Min Sun[2] Cheng-Hao Kuo[2] Ram Nevatia[1]
[1]University of Southern California [2]Amazon
[1]{xuefengh,nevatia}@usc.edu
[2]{kezha,luxial,aycchen,lujiajia,yuyinsun,zixiaow,qiaonan,zenxiao,minnsun,chkuo}@amazon.com

Abstract

Large-scale pre-trained vision-language models (VLM) such as CLIP [32] have demonstrated noteworthy zero-shot classification capability, achieving 76.3% top-1 accuracy on ImageNet without seeing any examples. However, while applying CLIP to a downstream target domain, the presence of visual and text domain gaps and cross-modality misalignment can greatly impact the model performance. To address such challenges, we propose ReCLIP, a novel source-free domain adaptation method for VLMs, which does not require any source data or target labeled data. ReCLIP first learns a projection space to mitigate the misaligned visual-text embeddings and learns pseudo labels. Then, it deploys cross-modality self-training with the pseudo labels to update visual and text encoders, refine labels and reduce domain gaps and misalignment iteratively. With extensive experiments, we show that ReCLIP outperforms all the baselines significantly and improves the average accuracy of CLIP from 69.83% to 74.94% on 22 image classification benchmarks.

1. Introduction

Large-scale pre-training vision-language models (VLM) such as CLIP [32] have emerged recently and have formed a new paradigm in the task of image classification. Instead of matching images with category abstraction (label index), vision-language models match images towards text embeddings from their category names. With semantic relationship from text and large-scale pre-training over 400 million image-caption pairs, CLIP is capable of performing accurate image classification on novel target domains requiring zero training samples but only a dictionary of potential category names.

However, we still observe domain gaps from both image and text input that impact CLIP performance. The existence of visual domain gap between source and target images has been a challenge for computer vision models [8, 43]. CLIP has been observed to have limitations on visual embedding when data comes from less common domains, *e.g.* PatchCamelyon [40], CLEVR [17], *etc*. On the other hand, the domain gap in text is also a challenge for vision-language models. The performance of CLIP is often limited by the text embeddings rather than the visual embeddings, especially on fine-grained datasets *e.g.* RESISC45 [5], Birdsnap [2], where CLIP is able to create distinctive visual embeddings but the text embeddings from class names fail to capture discriminative information.

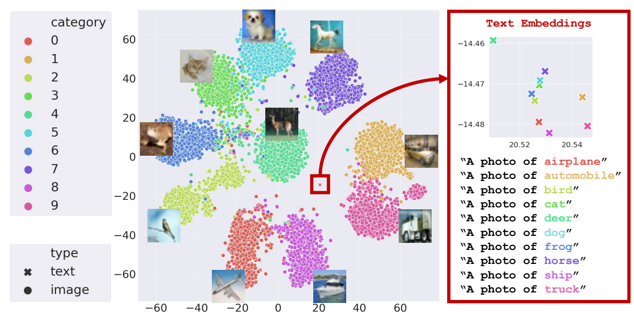

Figure 1. the t-SNE plot of visual and text embeddings from CLIP on CIFAR10 [21] test set. It is clear to see the misalignment in the vision-language space: the text embedding of a class name is adjacent to ones of other classes, but distant from image embeddings in the same class.

In addition to the gaps in the visual and text domains, we have identified significant misalignment between visual and text embeddings across most datasets. Some recent studies [26, 38] have also observed similar modality gaps across various contrastive-learned visual-language models. Figure 1 provides examples of this issue on the widely used benchmark CIFAR10. We believe that there are two primary reasons for these misalignments. Firstly, text embeddings may be redundant, as CLIP was trained to work with millions of captions and concepts, whereas target domain categories might only activate limited feature dimensions, leaving the remaining ones inactive and redundant; these redundant dimensions can dominate the similarity calculation. Secondly, visual embeddings may contain a significant amount of class-agnostic information; since CLIP uses real captions for training, it preserves rich information, such as lighting, color, texture, and relationship, but only a small portion of this information is crucial for classification.

Therefore, adaptation on both visual and text representations, and re-alignment between visual and text embeddings are crucial in improving the target domain performance of vision-language models like CLIP. However, traditional domain adaptation methods have significant limitations in this context. One major challenge is that these methods either require target domain labeled examples (e.g. semi-supervised domain adaptation [9, 33, 48]), or source domain examples (e.g., unsupervised domain adaptation [18, 29, 34]). Nonetheless, typical use cases of CLIP only have access to unlabeled target images, which requires source-free unsupervised domain adaptation that does not need source data or labeled target data. Another challenge is that existing methods assume conditions that may not hold for vision-language models. For instance, most existing methods [25, 42, 47] assume a lightweight classifier, while a vision-language model uses a large text encoder to generate classification weights based on category descriptions. Such modules add flexibility and complexity to adaptation. Thus, the lack of labeled data from source and target domains and the presence of multiple adaptable modules make it essential to develop a novel source-free domain adaptation algorithm for vision-language models.

More recently, POUF [38] also proposes to address the misaligned embeddings of a vision-language model through source-free adaptation. However, the unsupervised objective of POUF considers each target example independently, instead of taking advantages from the neighboring relationship over the entire embedding space. Moreover, POUF cannot leverage multiple template augmented text embeddings as used in CLIP and our proposed method, which limited its performance during the adaptation.

To take advantage of the unified vision-language space, and address the challenges on the visual and text domain gaps and cross-modality misalignment, we propose ReCLIP, a novel source-free domain adaptation method to **Re**fine **CLIP** models. Firstly, ReCLIP addresses the misalignment of visual and text embeddings from CLIP by learning a projection subspace that removes redundant dimensions and class-agnostic information, and realigns embeddings. ReCLIP then utilizes the neighboring relationship between aligned embeddings, and employs label propagation to produce accurate pseudo-labels in the target domain. Secondly, ReCLIP leverages cross-modality self-training with high-confidence pseudo labels to iteratively refine embedding spaces and label assignments. Two parallel components are deployed to update the text and visual encoders. The first component fine-tunes the text encoder while freezing the visual to pull the text embedding of a label closer to the embeddings of images assigned the label. Meanwhile, the second component fine-tunes the visual encoder so that the images under the same label get closer to each other and to the multi-template augmented text embedding of the label. During fine-tuning, each component learns cross-modality consistency in the target domain, leading to new label assignments. ReCLIP selects labels agreed upon by both components as high-confidence ones for the next iteration. This iterative process improves the quality of visual and text embeddings and significantly enhances the assignment of pseudo labels.

Our contributions are summarized in the following:

- We proposed ReCLIP, a novel source-free domain adaptation method for vision-language model, which enhances the CLIP's classification ability towards target domains without labeled data;
- We identified the cross-modality misalignment issue between CLIP's visual and language embeddings, and address the issue with an efficient projection-based component in ReCLIP;
- We proposed a novel cross-modality self-training algorithm with high quality commonly agreed pseudo labels leveraging cross-modality consistency to mitigate domain gaps from both visual and text inputs;
- With extensive experiments and ablation studies, ReCLIP produces consistent and significant improvements over CLIP and other baseline methods; ReCLIP improves the average accuracy of CLIP from 69.83% to 74.94% on 22 datasets.

2. Related Works

2.1. Large-Scale Vision-Language Models

Many large-scale pre-training vision-language models have been recently proposed and demonstrate impressive zero-shot classification ability, such as CLIP [32], ALIGN [16] that perform large-scale contrastive training for strong generalization ability, and DeCLIP [8], SLIP [28] that focus on efficient training with additional self-supervised objectives. In this work, we adopt CLIP as our main base model, as it is still the most representative vision-language model with outstanding zero-shot classification performance and publicly available model weights. In addition, we will also demonstrate the effectiveness of our method with different base models in ablation studies.

Augmented prompts through multiple templates. CLIP makes classification prediction by matching the visual embeddings of query images with the text embeddings of categories names (wrapped in template text such as "`A photo` of a {}"), and selects the category with the highest cosine similarity as prediction (please refer to the supplementary materials for more details on CLIP and VLM).

To further align these text embeddings with the pretraining distribution generated from real captions, CLIP prepares a long list of templates with various contexts for each of the 27 benchmarks it evaluated on. Instead of using just one template, CLIP reported scores are produced with the averaged text embeddings from a list of templated prompts for each category to boost performance.

Limitations of CLIP. We observe the following conditions where CLIP's performance could be improved. **1) Inaccurate Text Description.** The accuracy of CLIP can sometimes be drastically improved when the classification weights are fully-supervised fine-tuned, e.g., On EuroSAT, accuracy of CLIP improved from 59.9% to 98.2% [32]. This indicates that CLIP has good quality default visual representations, but the zero-shot performance is limited by the quality of text-generated classification weights. This is often observed on fine-grained datasets (e.g., AID [45], FGVC [27], EuroSAT [14], *etc*.), where the class names can not fully capture the visual differences between classes (e.g., "737-200" and "747-200" as class names from FGVC); **2) Visual Gap.** On some datasets, there are clear gaps for CLIP to be further improved even after the fully supervised fine-tuning on classification weight. For example, fine-tuned CLIP achieves only 42.9% on Country211 [32], and 85.97% on PatchCamelyon [40] (a binary classification task with state-of-the-art system achieves 97.50%). This indicates that the visual encoder of CLIP can also be further improved. **3) Visual-Text Misalignment.** Recent studies [26, 39] have also shown that the modality gap between visual and text embeddings caused by contrastive pretraining could also limit the performance of CLIP. By modifying contrastive temperature during pre-training [26], or by minimizing the gap during few-shot fine tuning [39], these works suggest that mitigating the modality gap can benefit the classification ability of CLIP.

2.2. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) is a task aimed at improving target domain performance of models that were pre-trained on a related but different source domain. Many techniques have been developed [18,29,34,37,44], including a recent method designed for visual-language models [22]. However, most of these techniques are not ideal for the purpose of improving CLIP's zero-shot performance, as they often require access to source domain data, while we do not require access to CLIP's training data.

Source-Free Adaptation defines a more challenging setting than UDA, where training examples are not available in both the source and target domains. SHOT [25] is one of the first Source-Free Adaptation (SFDA) methods. SHOT updates the feature extractor with cluster-based pseudo labels and information entropy loss, while maintaining the classifier frozen. AaD [47] improves SHOT by replacing the information entropy loss with a novel Attracting-and-Dispersing (AaD) loss. This simple but effective approach achieves state-of-the-art performance on the task of SFDA.

More recently, POUF [38] also proposes to mitigate the misalignment embeddings through source-free domain adaptation for vision-language models. But the optimization objective of POUF has limited its performance in two ways: 1) the training of POUF imposes dependency on the number of text encoder inputs (prompts), which limits POUF from using multiple templates to boost performance, especially on datasets with large number of classes; 2) the training objectives consider each image separately and fail to leverage neighboring relationships.

3. Method

We describe our method **ReCLIP**, which **Re**fines **CLIP**'s classification performance by accessing only to the pre-trained model and the following target domain data:

- Pre-trained vision-language model $M = \{M_T, M_V\}$, with text encoder M_T and visual encoder M_V,
- Unlabeled target images $X = \{x_1, x_2, ..., x_n\}$,
- Target class names $C = \{c_1, c_2, ..., c_m\}$.

Our goal is to increase the classification accuracy of M on target data X. As the first method that studies the source-free adaptation problem for vision-language model, we approach this problem in two steps: (1) How to align visual and text embeddings by removing class-agnostic and redundant information in a learned projection space (Section 3.1). Then we show how to assign pseudo labels for images in the projection space via label propagation (Section 3.2); (2) How to utilize the pseudo labels to further mitigate the visual and text domain gaps by efficiently updating both visual and text encoders, we propose a cross-modality self-training algorithm which updates embeddings and pseudo labels in a iterative fashion (Section 3.3).

3.1. Projection Space to Align Visual and Text

Figure 1 demonstrates the misalignment issue of text and visual embeddings from CIFAR10 [21], which we have also observed over all the ablation datasets. The plot indicates that the text embeddings of different class names are closer to each other than to images in the corresponding categories. We also validate the misalignment with quantitative statistics, as shown in Figure 2. The average cosine similarity between text embeddings is 82%, while the average similarity between visual and text embeddings from the same category is only 23%. This indicates that the unified vision-language space of CLIP is far from well aligned.

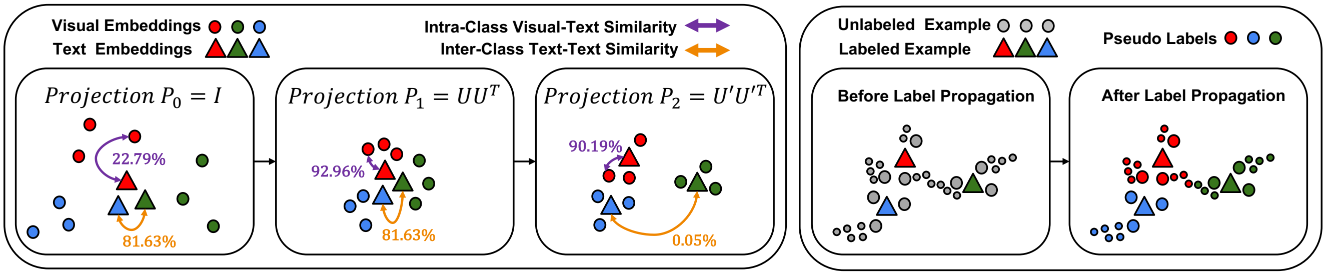

Figure 2. Demonstration on Feature Redundancy Removal (left) and Label Propagation (right). **Left**: P_0 shows the original distribution of visual and text embeddings of CLIP, where text embeddings are close to each other distant from visual embeddings; $P_1 = UU^\top$ removes the class agnostic information from visual embeddings, and has pulled closer visual and text embeddings. $P_2 = U'U'^\top$ separates the text embeddings away by removing the redundant information from them. Similarity values demonstrated in this example is calculated based on average statistics from CIFAR10 test set; **Right**: the Label Propagation process generates pseudo labels for unlabeled training images by propagating label information from labeled text embeddings (categories names) to unlabeled visual embeddings (training images) through nearest-neighbor connections.

As highlighted in Section 1, although the visual and text embeddings from CLIP convey rich information, much of them could be redundant and class-agnostic to target classification tasks. This redundancy can result in misalignment between the text and visual embeddings. We hence propose a projection-based method to eliminate the redundancy from both visual and text embeddings.

Remove class-agnostic information from visual embeddings. A straightforward way to remove class-agnostic information from visual features is just to project all the visual embeddings onto the span of text embeddings. Assuming we have a d dimensional representation space $\mathcal{R}^d$, and we have m classes whose text embeddings are $T = [t_1, ..., t_m] \in \mathcal{R}^{m \times d}$, where $t_i = M_t(c_i)$ for $i \in \{1, 2, ..., m\}$. With Singular Value Decomposition

$$U, S, V = svd(T)$$

we get $U = [e_1, e_2, ..., e_m]$ as the orthonormal basis of the span of T, which defines a projection matrix $P_1 = UU^\top$. Then, $\forall f \in \mathcal{R}^d$, we can calculate $f' = fP_1$ with

$$e_k \cdot (f - f') = 0, \forall k \in \{1, ...m\}$$

where $f - f'$ is the class-agnostic information that does not contribute to the classification. As shown in Figure 2, P_1 increases the average similarity between images and text embeddings from the same category to 92.96% on CIFAR10.

Remove redundant information from text embeddings. As suggested in Principal Component Analysis, the first dimension e_1 of the outer-space basis U will be the major component that most $\{t_1, ..., t_m\}$ overlap on. Removing the major component e_1 will make all text embeddings nearly perpendicular to each other. Therefore, with $U' = [e_2, e_3, ..., e_m]$ we define a new projection matrix $P_2 = U'U'^\top$. As shown in Figure 2, P_2 successfully separates the text embeddings from different classes to an average cosine similarity of 0.05%, while maintaining high intra-class visual-text similarity at 90.19% on CIFAR10.

In addition to the improvement of CIFAR10 statistics, experiments on pseudo label generation also indicate the effectiveness of embedding space induced by P_2 in improving clustering performance, as demonstrated in Section 4.3.2.

3.2. Pseudo Label Generation for VLM

The projection matrix P_2 removes the redundancies and aligns visual and text embeddings, which enables the generation of pseudo labels through Label Propagation [15], which is a semi-supervised learning method that propagates label information from labeled to unlabeled data points through nearest neighbor connections, as demonstrated in Figure 2. Although in source-free adaptation we do not have access to labeled data points, the embedding alignment through P_2 has enabled us to treat text embeddings from class names as labeled points, and visual embeddings from images as unlabeled points.

With labeled examples $\{\hat{t}_i\}_{i=1}^m$ (class name embeddings) and unlabeled examples $\{\hat{v}_j\}_{j=1}^n$ (image visual embeddings), we make the union set L:

$$L = [\hat{t}_1, \hat{t}_2, ..., \hat{t}_m, \hat{v}_1, \hat{v}_2, ..., \hat{v}_n] \in \mathcal{R}^{d \times (m+n)}$$

Following Label Propagation [15], we first produce affinity matrix A_k through $k-$nearest neighbor affinity ranking $A_k = top^k(L^\top L)$ where $top^k(\cdot)$ is an operation that keeps the top k highest value per row from the full affinity matrix $L^\top L$. Then, with normalization and symmetrization, we have:

$$\mathcal{W} = D^{-\frac{1}{2}}(A_k + A_k^\top)D^{-\frac{1}{2}}$$

where $D := diag(W\mathbf{1}_{m+n})$ is the degree matrix, $\mathbf{1}_{m+n}$ is the all-ones $(m+n)-$vector, and $\mathcal{W}$ is the normalized adjacency matrix that defines the random walk probability. With an label matrix $Y \in \mathcal{R}^{(m+n) \times m}$ is defined with elements

$$Y_{ji} := \begin{cases} 1, & \text{if } j = i, j \leq m \\ 0, & \text{otherwise} \end{cases}$$

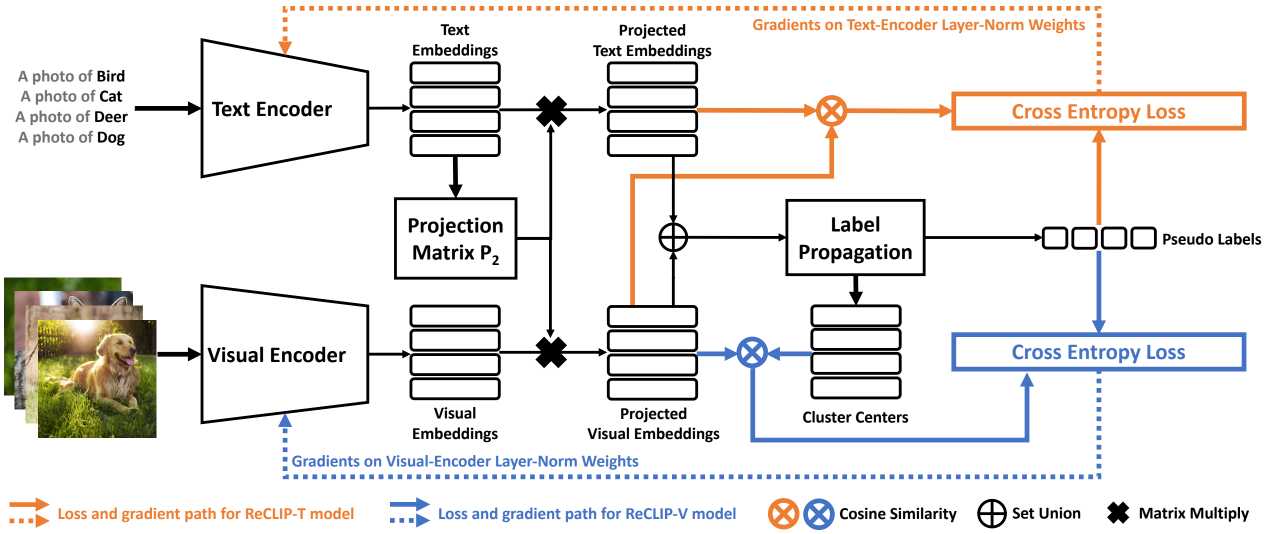

Figure 3. Flow Chart of ReCLIP-V and ReCLIP-T. Orange symbols describe the loss and gradients path of ReCLIP-V, and blue symbols describe the loss and gradients path of ReCLIP-T. Black symbols describe the common steps that both ReCLIP-V and ReCLIP-T have.

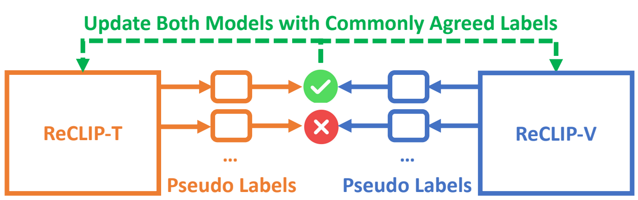

Figure 4. Flow Chart of Pseudo Labels Sharing. The cross-modality self-training algorithm merges the pseudo labels from ReCLIP-T and ReCLIP-V at the end of each epoch and updates the encoders only on high-confidence pseudo labels agreed by both.

where Y_{ji} is 1 for the text embedding entries at the corresponding column, and 0 otherwise. Then, the pseudo label vector Z can be estimated by solving the random walk problem with initial state Y, propagation probability matrix $\mathcal{W}$ and diffusion magnitude α:

$$Z := (\mathbf{I} - \alpha\mathcal{W})^{-1}Y \tag{1}$$

where $(\mathbf{I} - \alpha\mathcal{W})^{-1}$ is the closed-form solution to the random walk problem. As $(\mathbf{I} - \alpha\mathcal{W}) \in \mathcal{R}^{m+n}$ is not sparse, and therefore the calculation of its inverse matrix is very time consuming, we use conjugate gradient (CG) to approximately solve Equation 1, following the suggestion from [15]. Finally, with Equation 1 solved, the pseudo label can be given by

$$\tilde{y}_j := \arg\max_i z_{m+j,i}$$

where $\tilde{y}_j$ is the pseudo label of image x_j, and z_{ji} is the (j,i) element of matrix Z.

3.3. Source-Free Adaptation for Vision-Language Model via Cross-Modality Self-Training

Vision-language models present a new challenge to adaptation algorithms, where both visual and text encoders need to be adapted. In this section, we discuss how to mitigates the domain gaps of visual and text domains, and propose a cross-modality self-training algorithm with pseudo labels from 3.2 to iteratively update the label assignments, and the visual and text encoders.

The self-training algorithm of ReCLIP consists of two parallel components: ReCLIP-T aims at closing the text domain gap by pushing text embeddings towards visual embeddings of the same class, by fine-tuning the text encoder with the visual encoder frozen. ReCLIP-V aims at closing the visual domain gap by pushing visual embeddings of the same class closer to each other, by fine-tuning the visual encoder with the text encoder frozen. On top of ReCLIP-V and ReCLIP-T, we integrate the commonly-agreed pseudo labels to produce high-confidence training signals. For inference, we add the prediction logits from both ReCLIP-V and ReCLIP-T to make the final prediction.

ReCLIP-T: Text Encoder Training. We optimize the text encoder M_t with simple cross-entropy loss $Loss^T := CE(\hat{Y}^T, \tilde{Y})$ between pseudo label $\tilde{Y}$ and cosine similarity prediction logits $\hat{Y}^T = [\hat{v}_1, ..., \hat{v}_n]^\top[\hat{t}_1, ..., \hat{t}_m]$. The objective of adaptation on the text encoder is to push text embeddings $\{\hat{t}_i\}$ closer to the image embeddings $\{\hat{v}_j\}$ from the same class based on pseudo label assignments $\tilde{Y}^T$. In Figure 3 we present the details of ReCLIP-T, the detailed algorithm is provided in the supplementary materials.

ReCLIP-V: Visual Encoder Training. The goal of visual encoder adaptation is to push visual embeddings $\{\hat{v}_j\}$ from the same class to be closer to each other, to form a better feature space for classification. As contrastive loss is expensive and applying constraints on batch size, we have instead chosen to push visual embeddings closer to the center of its class instead of other visual embeddings as an alternative resort. To be specific, in ReCLIP-V we optimize the visual encoder M_v with cross-entropy loss $Loss^V := CE(\hat{Y}^V, \tilde{Y})$ between pseudo label $\tilde{Y}$ and cosine similarity logits $\hat{Y}^V = [\hat{v}_1, ..., \hat{v}_n]^\top [\hat{w}_1, ..., \hat{w}_m]$, where $\hat{w}_1, ..., \hat{w}_m$ are the class centers calculated based on $\tilde{Y}$. In Figure 3 we present the details of ReCLIP-V, the detailed algorithm is provided in the supplementary materials.

High-Confidence Pseudo Labels Sharing. ReCLIP-V updates the similarities among visual embeddings with $Loss^V$, while ReCLIP-T updates the projection matrix and text embeddings with $Loss^T$. As these two modules separately optimize the visual and text encoders with different objectives, their pseudo labels may start to diverge after a certain number of epochs, resulting in different views where only the commonly agreed samples are likely to be correctly classified. As such, ReCLIP collects pseudo labels from both ReCLIP-V and ReCLIP-T at the end of each epoch, and updates both models with only the commonly agreed pseudo labels $\tilde{Y}$, as illustrated in Figure 4. The detailed algorithm is provided in the supplementary materials.

4. Experiment and Results

Baselines We use the following methods for comparison:
1) CLIP [32]: State-of-the-art zero-shot image classification model. We choose CLIP with ViT/L-14 architecture as the main baseline model for comparison and adaptation. We report both published results from Radford *et al*. [32] and our reproduction, denoted as *report* and *multi* respectively. Both *report* and *multi* are prepared with the official prompt template lists provided by Radford *et al*. [32]. In addition, we also report the results we reproduced with a single template (“`A photo of a {}`”), denoted as *single*;
2) AaD [47]: State-of-the-art SFDA method. We adapt the official code to apply it on CLIP and our benchmarks;
3) POUF [38]: A recent SFDA method that also aims to mitigate misaligned visual and text embedding spaces. Since POUF does not report on the benchmarks where CLIP has published scores, we produce its results on these benchmarks using its official code. We report the best performing version of POUF which fine-tunes the entire model.

Evaluation and Datasets. **1) Main Results:** for SFDA comparison between ReCLIP, POUF, AaD and base model CLIP, we use an abundant and comprehensive list of 21 common image classification benchmarks out of the 27 benchmarks from Radford *et al*. [32], except the 6 datasets where CLIP are evaluated on the custom splits or protocols which are not released at the time of this submission (KITTI [13], UCF101 [35], VOC2007 [12], Kinetics700 [4], HatefulMemes [19], CLEVR [17]). In addition to the ablation dataset AID [45] we use for hyper-parameters selection, SFDA evaluation is performed on 22 benchmarks in total. **2) Comparison with POUF:** For additional comparison with POUF on its published scores, we evaluate ReCLIP on Office-Home [41], which contains four different domains: Art (Ar), Clipart (Cl), Product (Pr) and Real-World (Rw). **3) Ablation Studies:** we choose AID, CIFAR10, CIFAR100 and SUN397 as ablation datasets to represent datasets with different sizes and characteristics. For more details on evaluation datasets, please refer to supplementary materials.

For SFDA evaluation in Section 4.1, AaD and ReCLIP use CLIP-*multi* as base model, and POUF uses CLIP-*single* due to its design. For experiments on Office-Home, both ReCLIP and POUF use CLIP-*single* as base model.

Unless otherwise specified, we perform our experiments in transductive manner, where SFDA methods ReCLIP, POUF and AaD first perform adaptation on the unlabeled test data of each dataset, and then the adapted models are evaluated on the same test data following the standard CLIP inference protocol. For all benchmarks, we use top-1 classification accuracy as our metric,

Implementation Details For the self-training of ReCLIP, we fine-tune the layer-normalization [1] weights with other weights frozen, as it is shown to be one of the most effective and stable option to adapt models with noisy supervision [42]. For the SFDA evaluation, we use AID [45] to select the best hyper-parameter for ReCLIP, POUF and AaD. We then use the same set of hyper-parameters for all 22 datasets during the evaluation. We match the maximum adaptation steps for all methods to be the same, as $\min\{5000 \text{ iterations}, 50 \text{ epochs}\}$. For the evaluation on Office-Home, we select the hyper-parameter on the Real-World (Rw) domain and use the same hyper-parameters across all domains for evaluation. For details on exact hyper-parameters used in experiments, ablation studies on choices of learnable modules, and the setup of Label Propagation, please refer to supplementary materials.

4.1. Main Results

In Table 1 we present the SFDA accuracy of ReCLIP, AaD and POUF over 22 datasets. Besides the accuracy from the final epoch of self-training, we report the accuracy from the peak-performing epoch for AaD, POUF and ReCLIP as well, denoted as *peak*.

ReCLIP achieves consistent and significant improvements over CLIP on 21 datasets and comparable performance on Country 211. ReCLIP improves the averaged top-1 accuracy of CLIP by 5.11% and 6.02% at the *final* and *peak* epochs respectively over the 22 datasets without accessing any labeled data, which outperforms both baseline adaptation methods AaD, POUF by clear margin.

	Avg Acc	AID [45]	Birdsnap [2]	Caltech101 [23]	CIFAR10 [21]	CIFAR100 [21]	Country211 [32]	DTD [6]	EuroSAT [14]	FER2013 [49]	FGVC [27]	Flowers [30]	Food101 [3]	GTSRB [36]	ImageNet [10]	MNIST [11]	Oxford Pet [31]	PCam [40]	SST2 [32]	RESISC45 [5]	Cars [20]	STL10 [7]	SUN397 [46]
CLIP-*report*	70.08	-	48.30	92.6*	96.20	77.90	**32.70**	55.30	59.90	57.50	36.1*	78.7*	92.90	50.30	75.30	87.20	93.50	58.80	64.00	71.60	77.3*	99.30	67.70
CLIP-*single*	65.53	61.30	51.88	92.02	95.19	77.18	25.78	52.50	56.03	52.22	30.18	74.19	92.56	45.57	73.46	52.63	93.21	57.75	52.39	63.29	76.45	99.47	66.42
CLIP-*multi*	69.83	68.73	52.48	91.63	95.60	78.22	31.84	55.37	60.00	56.39	31.59	79.04	93.08	50.59	75.52	76.23	93.62	62.43	68.92	69.66	77.88	99.36	67.97
AaD	46.53	69.83	52.42	91.45	96.54	80.18	0.47	55.43	11.12	16.91	32.37	78.61	0.99	51.26	0.11	89.81	93.62	49.95	49.92	2.51	0.52	99.41	0.25
AaD *peak*	71.79	70.33	52.58	91.93	96.55	80.46	31.90	55.59	76.18	55.67	32.43	79.22	93.04	52.83	75.53	91.95	93.73	64.03	68.97	71.01	77.96	99.42	67.96
POUF	69.73	64.83	52.91	92.97	96.06	80.39	28.19	56.65	67.95	55.92	32.88	75.62	92.71	51.47	73.05	91.22	94.20	66.57	48.22	67.54	76.72	99.50	68.38
POUF *peak*	69.76	64.87	52.96	92.97	96.06	80.39	28.22	56.75	67.95	55.92	32.91	75.62	92.73	51.47	73.06	91.22	94.20	66.75	48.60	67.54	76.72	99.53	68.38
ReCLIP	74.94	77.97	52.96	93.02	96.95	82.32	31.92	60.85	78.75	58.07	36.63	82.05	94.15	66.81	75.81	90.88	95.61	70.15	73.48	78.41	77.96	99.58	74.41
ReCLIP *peak*	**75.85**	**79.27**	**53.28**	**93.10**	**97.04**	**83.42**	31.95	**61.38**	**79.94**	**58.29**	**38.70**	**83.14**	**94.18**	**69.14**	**76.01**	**97.11**	**96.05**	**70.56**	**73.48**	**79.31**	**79.26**	**99.59**	**74.53**

Table 1. Classification accuracies (%) on 22 benchmarks. * on FGVC, Caltech101, Oxford-IIIT Pet and Flowers102, CLIP reported mean-class-accuracy. All other scores in this table are top-1 accuracy.

	Avg	Ar	Cl	Pr	Rw
CLIP *single*	82.45	82.70	68.10	89.10	89.90
POUF-prompt	84.28	83.70	71.20	91.40	90.80
POUF	86.10	**86.20**	73.80	92.70	91.70
Label Propagation	84.94	83.27	73.49	91.89	91.09
ReCLIP	**87.00**	86.11	**75.97**	**93.90**	**92.01**

Table 2. Comparison of ReCLIP and published scores from POUF [38] on Office-Home [41], both use CLIP-*single* as base model.

	CIFAR10	CIFAR100	AID	SUN397
Vanilla CLIP	95.54	76.48	64.87	67.25
Label Propagation	96.38	80.66	74.73	70.54
ReCLIP-V	96.69	80.84	79.47	67.15
ReCLIP-T	96.50	81.10	79.07	70.12
ReCLIP (w/o Label Sharing)	97.40	82.80	80.01	71.10
ReCLIP (w/ Label Sharing)	**97.48**	**84.14**	**82.53**	**71.34**

Table 3. Comparison of classification accuracy with different version ReCLIP on ablation datasets. ReCLIP with Label Sharing (Figure 4) is shown to be most effective compared to ReCLIP-V, ReCLIP-T (Figure 3) and their simply assembled predictions (ReCLIP w/o Label Sharing).

AaD achieves 1.96% improvements over CLIP at its *peak* epochs. However, it encounters drastic performance drops at *final* epochs that lose 25.26% of the averaged accuracy, due to collapsed unsupervised training on target datasets such as Food101, SUN397, ImageNet, etc. Meanwhile, ReCLIP maintains the performance at *final* epochs, with only 0.91% differences from the *peak* epochs. These results suggest the effectiveness of the high-quality commonly agreed pseudo labels of ReCLIP in stabilizing the noisy self-training and preventing model collapse.

POUF achieves 4.20% improvement over its base model CLIP-*single*. However, such improvement is counteracted by the inability to employ multiple prompts to enhance text embedding quality, as suggested by CLIP [32]. Multiple templates create a large number of prompts, which are not likely to fit in the same mini-batch for text encoder optimization. ReCLIP also experiences this limitation when fine-tuning the text encoder. However, thanks to the dual-component structure of ReCLIP, although ReCLIP-T also only use single template for text-encoder optimization, ReCLIP-V can still take advantage of the multiple template augmented text embeddings and provides better pseudo labels to ReCLIP-T through pseudo-label sharing. In addition to the advantage brought by multi-template augmented text embeddings, ReCLIP also takes advantage from the neighboring relationships over the entire visual-text embedding space, while POUF does not, which has also contributed to the better performance of ReCLIP. More evidence and discussion on this are covered in Section 4.2.

Country211 is designed to predict geo-location based on visual appearance, while CLIP might tend to describe the image from actual content and texture. As shown in [32], CLIP can only achieve 42.9% after its classifier is fine-tuned in the fully supervised way. Therefore, it is challenging to obtain improvement during source-free domain adaptation.

4.2. Comparison with POUF

In Table 2 we present the comparison between the published scores of POUF and ReCLIP on the Office-Home, where both methods use CLIP-*single* (ViT/B-16) as base model. We also include the Label Propagation pseudo label accuracy generated on our projected CLIP embeddings prior to any updates on the base model. It is shown that the Label Propagation accuracy already outperforms POUF-prompt, which fine-tunes the learnable text prompt. Moreover, ReCLIP achieves clear improvement over POUF over most of the domains, with 2.17%↑ on Cl, 1.20%↑ on Pr, 0.31%↑ on Rw and on-par performance on Ar. These results indicate that ReCLIP can still outperform POUF without using multi-template augmented embeddings.

4.3. Ablations Studies

In this section, we present the ablation studies on comparison of different ReCLIP versions, pseudo label gener-

	AID	CIFAR10	CIFAR100	SUN397
Vanilla CLIP	68.80	95.59	78.21	67.97
Hierarchical Clustering	55.20	36.52	9.27	46.93
Spectrum Clustering	68.10	61.25	57.35	27.45
k-means Clustering	72.73	95.07	49.43	43.66
k-NN Classifier (P_0)	72.30	93.74	69.46	60.72
k-NN Classifier (P_1)	72.76	95.77	77.81	63.07
k-NN Classifier (P_2)	72.43	95.76	78.19	63.29
Label Propagation (P_0)	60.80	94.01	63.58	51.77
Label Propagation (P_1)	60.43	96.23	45.41	33.41
Label Propagation (P_2)	**76.36**	**96.31**	**81.56**	**70.44**

Table 4. Pseudo label accuracy with different methods. Label Propagation on projection space P_2 is shown to be the most effective and stable method in generating accurate pseudo labels.

ation, and ablation on various VLMs as base models. We use AID, CIFAR10, CIFAR100 and SUN397 as our ablation datasets, and the test set of each dataset is equally split into two fixed partitions. We report the ablation results in an inductive manner where models are first adapted on partition 1 and then evaluated on partition 2. Note that results in this section are not directly comparable to 4.1 because the different evaluation partition.

4.3.1 Effectiveness of ReCLIP Components

In Table 3 we present the comparison between different versions of ReCLIP. As shown, Label Propagation can create pseudo labels with significantly improved accuracy compared to vanilla CLIP. On the top of Label Propagation, both ReCLIP-V and ReCLIP-T (Figure 3) are shown to be effective in providing further improvements. In ReCLIP(w/o Label Sharing) we present the result by simply assembling predictions from separately trained ReCLIP-V and ReCLIP-T at inference time. Comparing the last two rows of Table 3 we observe that ReCLIP (w/ Label Sharing) has clear improvement over ReCLIP (w/o Label Sharing), which indicates that the commonly agreed pseudo-labels stabilizes the noisy adaptation process and improved both ReCLIP-V and ReCLIP-T to achieve better performance.

4.3.2 Comparison on Pseudo Label Generations

In Table 4, we compare methods in pseudo label generation. For clustering based methods, we assign the same pseudo labels for examples from the same cluster, based on the in-cluster majority vote; For k-NN Classifier and Label Propagation methods, we experiment them on original CLIP feature space P_0, and on projection spaces P_1, P_2 as described in Figure 2. For k-NN Classifiers, we assign each example with the major vote prediction within its k-nearest-neighborhood, with k equal to the average sample count per class. For Label Propagation on P_0, we select the example with the highest confidence from each class as the labeled example to perform label propagation as a baseline. Label Propagation on P_1, P_2 are as described in Section 3.1.

	CIFAR10	CIFAR100	AID	SUN397
	Init $\rightarrow$ Adapt	Init $\rightarrow$ Adapt	Init $\rightarrow$ Adapt	Init $\rightarrow$ Adapt
SLIP (ViT-L/16)	89.45 $\rightarrow$ **91.80**	56.69 $\rightarrow$ **67.61**	48.13 $\rightarrow$**64.07**	55.56 $\rightarrow$ **65.28**
DeCLIP (ViT-B/32)	90.57 $\rightarrow$ **94.50**	66.58 $\rightarrow$ **77.10**	53.53 $\rightarrow$**65.93**	63.05 $\rightarrow$ **66.90**
CLIP (RN50)	71.46 $\rightarrow$ **82.73**	42.32 $\rightarrow$ **53.15**	53.43 $\rightarrow$**65.97**	59.76 $\rightarrow$ **65.38**
CLIP (ViT-B/32)	89.83 $\rightarrow$ **92.15**	65.25 $\rightarrow$ **71.09**	60.83 $\rightarrow$**76.80**	62.96 $\rightarrow$ **68.30**

Table 5. Ablation Studies on the effectiveness of ReCLIP on different model architecture and pre-training strategies.

Table 4 indicates k-NN based methods achieve better performance on projection spaces P_1 are P_2, which indicates the effectiveness of P_1, P_2 in refining CLIP's visual embeddings. On Label Propagation methods, P_2 gives a significant improvement over P_0, P_1, indicating its effectiveness in aligning CLIP's visual and text embeddings.

4.3.3 Comparison on other Vision-Language Models

ReCLIP is designed to improve the classification performance of visual-language models in general, not only on CLIP. We tested the effectiveness of ReCLIP on SLIP [28] and DeCLIP [24], both of these improved CLIP by adding self-supervision learning objectives during pre-training. We have also tested ReCLIP on other versions of CLIP with smaller architectures. As shown in Table 5, ReCLIP demonstrates steady and significant improvements on various vision-language models and architectures.

4.3.4 Runtime and Inductive Performance

Self-training of ReCLIP is very efficient, which completes adaptation in only 0.5 to 5 GPU-Hour on a single V100 GPU, depending on the target dataset size. Note that this adaptation time is a one-time effort on each target domain and ReCLIP can then inference on unseen data from the same domain without re-training. For complete runtime of ReCLIP over each benchmarks and more inductive evaluation results, please refer to the supplementary materials.

5. Conclusion

In this paper, we introduce ReCLIP, a novel solution on source-free domain adaptation for vision-language models. ReCLIP first uses a novel designed projection space to re-aligns visual and text embeddings and to generate dependable pseudo labels for target classification tasks. ReCLIP further applies cross-modality self-training with pseudo labels, which iteratively enhances label assignments and visual and text embeddings. Compared to the previous methods AaD and POUF, ReCLIP provides an effective and stable solution to the source-free adaptation problem of vision-language models. ReCLIP significantly improves CLIP, increasing the average accuracy from 69.83% to 74.94% across 22 datasets.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 6

[2] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2018, 2014. 1, 7

[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. 7

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[5] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 1, 7

[6] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

[7] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 7

[8] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017. 1, 2

[9] Hal Daumé III, Abhishek Kumar, and Avishek Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59, 2010. 2

[10] 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. 7

[11] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 7

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 6

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 6

[14] 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. 3, 7

[15] 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*, pages 5070–5079, 2019. 4, 5

[16] 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. 2

[17] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 1, 6

[18] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 2, 3

[19] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020. 6

[20] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 7

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 3, 7

[22] Zhengfeng Lai, Sol Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. *ICCV*, 2023. 3

[23] Li, Andreeto, Ranzato, and Perona. Caltech 101, Apr 2022. 7

[24] 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. *arXiv preprint arXiv:2110.05208*, 2021. 8

[25] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 2, 3

[26] 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, 3

[27] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 3, 7

[28] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 529–544. Springer, 2022. 2, 8

[29] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1103, 2021. 2, 3

[30] 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. 7

[31] 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. 7

[32] 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, 6, 7

[33] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8050–8058, 2019. 2

[34] Astuti Sharma, Tarun Kalluri, and Manmohan Chandraker. Instance level affinity-based transfer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5361–5371, 2021. 2, 3

[35] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[36] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, (0):–, 2012. 7

[37] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019. 3

[38] Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. *arXiv preprint arXiv:2305.00350*, 2023. 1, 2, 3, 6, 7

[39] Vishaal Udandarao. *Understanding and Fixing the Modality Gap in Vision-Language Models*. PhD thesis, Master's thesis, University of Cambridge, 2022. 3

[40] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018. 1, 3, 7

[41] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 6, 7

[42] 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. 2, 6

[43] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 1

[44] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16643–16653, 2021. 3

[45] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 3, 6, 7

[46] 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. 7

[47] Shiqi Yang, Shangling Jui, Joost van de Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35:5802–5815, 2022. 2, 3, 6

[48] Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2142–2150, 2015. 2

[49] Lutfiah Zahara, Purnawarman Musa, Eri Prasetyo Wibowo, Irwan Karim, and Saiful Bahri Musa. The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi. In *2020 Fifth international conference on informatics and computing (ICIC)*, pages 1–9. IEEE, 2020. 7