SuS-X: Training-Free Name-Only Transfer of Vision-Language Models

Vishaal Udandarao
University of Cambridge
vu214@cam.ac.uk

Ankush Gupta
DeepMind, London
ankushgupta@google.com

Samuel Albanie
University of Cambridge
sma71@cam.ac.uk

Abstract

*Contrastive Language-Image Pre-training (CLIP) has emerged as a simple yet effective way to train large-scale vision-language models. CLIP demonstrates impressive zero-shot classification and retrieval performance on diverse downstream tasks. However, to leverage its full potential, fine-tuning still appears to be necessary. Fine-tuning the entire CLIP model can be resource-intensive and unstable. Moreover, recent methods that aim to circumvent this need for fine-tuning still require access to images from the target task distribution. In this paper, we pursue a different approach and explore the regime of training-free "name-only transfer" in which the only knowledge we possess about the downstream task comprises the names of downstream target categories. We propose a novel method, **SuS-X**, consisting of two key building blocks—"SuS" and "TIP-X", that requires neither intensive fine-tuning nor costly labelled data. **SuS-X** achieves state-of-the-art (SoTA) zero-shot classification results on 19 benchmark datasets. We further show the utility of TIP-X in the training-free few-shot setting, where we again achieve SoTA results over strong training-free baselines. Code is available at https://github.com/vishaal27/SuS-X.*

1. Introduction

Vision-language pre-training has taken the machine learning community by storm. A broad range of vision-language models (VLMs) [57, 42, 73, 1, 37] exhibiting exceptional transfer on tasks like classification [80, 84], cross-modal retrieval [67, 2] and segmentation [63, 27] have emerged. These models are now the *de facto* standard for downstream task transfer in the field of computer vision.

One such prominent model, CLIP [57], is trained on 400M image-text pairs using a contrastive loss that maximises the similarities of paired image-text samples. CLIP pioneered the notion of *zero-shot transfer* in the vision-language setting[1]: classification on unseen datasets. For a given classification task, CLIP converts the class labels into textual prompts (*e.g.* "A photo of a <CLASS>.", where <CLASS> represents the ground-truth text label for each class). It then computes similarities between all the class prompts and the query image, selecting the class with the highest image similarity as the predicted label (see Eq. (2)).

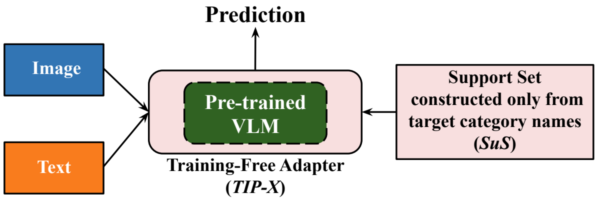

Prediction
Image
Text
Pre-trained VLM
Training-Free Adapter (*TIP-X*)
Support Set constructed only from target category names (*SuS*)

Figure 1: **Training-free name-only transfer.** We propose ***SuS-X***, a framework for enhancing the zero-shot transfer abilities of VLMs like CLIP [57], BLIP [42] and TCL [72], without training. To achieve this, we propose a novel method *TIP-X*, which adapts these VLMs using a curated *support set* (*SuS*) that is *not drawn* from the target distribution. Our *SuS* leverages one key piece of information about the task at hand: the names of the target categories.

CLIP's zero-shot performance is however limited by its pre-training distribution [24, 60, 21, 51]. If the downstream dataset diverges significantly from the pretraining image distribution, CLIP's zero-shot performance drastically drops [21]. To mitigate this, several lines of work propose to adapt CLIP on diverse downstream tasks—Tab. 1 briefly summarises these methods. Most of them employ fine-tuning on either labelled or unlabelled subsets of data from the target task. However, fine-tuning such an over-parameterised model can be unstable and lead to overfitting [15, 25]. Furthermore, having access to the true distribution of the target task can be prohibitive in data-scarce environments [12, 4, 38] and online learning settings [14, 65].

To alleviate these issues, in this paper, we aim to adapt CLIP and other VLMs for downstream classification in a *name-only* (requires only category names[2], but no samples

[1] This idea of zero-shot transfer is distinct from the traditional zero-shot classification setup introduced by Lampert et al. [41] in which the task is to generalise to classes not seen during training.

[2] We use category and class interchangeably in this paper.

Table 1: **Taxonomy of CLIP adaptation methods for downstream classification.** We underline the Zero-Shot CLIP model to signify that it is the base model that all others build on top of. *This method considers access to all test-set samples simultaneously, hence we still consider it zero-shot. †This method additionally uses class hierarchy maps.

	Method	**Does not require training**	**Does not require labelled data**	**Does not require target data distribution**
Few-shot fine-tuning methods	LP-CLIP [57]	✗	✗	✗
	CoOp [84]	✗	✗	✗
	PLOT [11]	✗	✗	✗
	LASP [9]	✗	✗	✗
	SoftCPT [19]	✗	✗	✗
	VT-CLIP [79]	✗	✗	✗
	VPT [17]	✗	✗	✗
	ProDA [45]	✗	✗	✗
	CoCoOp [83]	✗	✗	✗
	CLIP-Adapter [25]	✗	✗	✗
Intermediate methods	TIP-Adapter [80]	✓	✗	✗
	UPL [36]	✗	✓	✗
	SVL-Adapter [54]	✗	✓	✗
	TPT [48]	✗	✓	✓
	CLIP+SYN [33]	✗	✓	✓
	CaFo [78]	✗	✓	✓
Zero-shot methods	Zero-Shot CLIP [57]	✓	✓	✓
	CALIP [31]	✓	✓	✓
	CLIP+DN [85]*	✓	✓	✓
Training-free name-only transfer methods	CuPL [56]	✓	✓	✓
	VisDesc [49]	✓	✓	✓
	CHiLS [53]†	✓	✓	✓
	SuS-X (ours)	✓	✓	✓

from the target task) and *training-free* fashion. We propose ***SuS-X*** (see Fig. 1), consisting of two novel building blocks: (i) *SuS* (Support Sets), our dynamic *support set* curation strategy that forgoes the need for samples from the target task, and (ii) *TIP-X*, our main framework for performing zero-shot classification while being training-free. For a given downstream task, we first curate a *support set* by leveraging the task category labels, either in a parametric manner *i.e.*, generating images from large-scale text-to-image models (*e.g.*, Stable Diffusion [59]) or non-parametric manner *i.e.*, retrieving real-world images from a large vision-language data bank (*e.g.*, LAION-5B [61]). We then use the curated *support set* as a proxy few-shot dataset to inform our downstream predictions using *TIP-X*, in a similar vein to recent few-shot adaptation methods [25, 80].

Our extensive experiments show that ***SuS-X*** outperforms zero-shot methods on 19 benchmark datasets across three VLMs, namely, CLIP, BLIP and TCL by 4.60%, 5.97% and 11.37% absolute average accuracy respectively. We further extend the *TIP-X* framework to the few-shot regime, outperforming previous SoTA methods in the *training-free* domain. Our main contributions are three-fold: (1) We propose ***SuS-X***, a SoTA method in the *training-free name-only transfer* setting for downstream adaptation of VLMs, (2) We present *SuS*, an effective strategy for curating *support sets* using parametric or non-parametric methods to mitigate the lack of data samples available from the target task distribution, and (3) We propose *TIP-X*, a novel training-free method for adapting VLMs to downstream classification in both the *name-only* transfer and few-shot regimes.

2. Related Work

Vision-Language (VL) Foundation Models. Recent years have seen a Cambrian explosion in large-scale VL foundation models [6]. In a seminal work, Radford et al. [57] introduced CLIP, a large VLM trained on a web-scale corpus (400M image-text pairs), that exhibits strong downstream visual task performance. The introduction of CLIP inspired further development of VLMs [42, 1, 37, 18, 81, 75, 72, 10, 70, 26, 28, 43, 46, 74], each pre-trained on web-scale datasets to learn joint image-text representations. These representations can then be applied to tackle downstream tasks like semantic segmentation [63, 27], object detection [30, 20], image captioning [50, 3] and generative modelling [59, 58], In this work, we adapt such VLMs in a training-free setting to diverse downstream tasks.

Adaptation of VL models. CLIP introduces a paradigm shift with its zero-shot transfer ability [57]. In this setup, none of the target dataset classes are known *a-priori* and the task is to adapt implicitly at inference time to a given dataset. Since CLIP's training objective drives it to assign appropriate similarities to image-text pairs, it acquires the ability to perform zero-shot classification directly.

Inspired by CLIP's zero-shot success, further work has sought to improve upon its performance. In Tab. 1, we characterise some of these methods along three major axes: (i) if the method requires training, (ii) if the method requires labelled samples from the target task, and (iii) if the method requires samples from the target task distribution[3].

In this work, we focus on the *training-free name-only transfer* regime—our goal is to adapt VLMs to target tasks without explicit training or access to samples from the target distribution. Instead, we assume access only to category names of target tasks. This formulation was recently considered for semantic segmentation, where it was called *name-only transfer* [62]—we likewise adopt this terminology. To the best of our knowledge, only two other concurrent approaches, CuPL [56] and VisDesc [49], operate in this regime. They use pre-trained language models to enhance textual prompts for zero-shot classification. By contrast, ***SuS-X*** pursues a *support set* curation strategy to adapt VLMs using knowledge of category names. These approaches are complementary, and we find that they can be productively combined. Two other related works operating purely in the zero-shot setting are: (1) CALIP [31], which uses parameter-free attention on image-text features, and (2) CLIP+DN [85], which uses distribution normalisation. We compare with these four baselines in Sec. 4.

3. *SuS-X*: Training-Free Name-Only Transfer

We describe the two main building blocks of ***SuS-X***—(1) Support Set (*SuS*) construction, and (2) training-free inference using our novel *TIP-X* method. Fig. 2 depicts our overall *training-free name-only* transfer framework.

3.1. *SuS* Construction

We follow recent adaptation methods [80, 25] that use a small collection of labelled images to provide visual information to CLIP. However, differently from these methods, rather than accessing labelled images from the target distribution, we propose two methods (described next) to construct such a *support set* (*SuS*) without such access.

(I) Stable Diffusion Generation. Our first method leverages the powerful text-to-image generation model, *Stable Diffusion* [59]. We employ specific prompting strategies for generating salient and informative support images. Concretely, given a set of downstream textual class labels, $\mathcal{T} = \{t_1, t_2, \ldots, t_C\}$, where C denotes the number of categories, we prompt Stable Diffusion to generate N images per class. In this way, we construct our *support set* of size NC, with each image having its associated class label.

By default, we prompt Stable Diffusion using the original CLIP prompts, *i.e.*, "A photo of a <CLASS>.", where <CLASS> is the class text label. To further diversify the generation process, we follow CuPL [56] to first generate customised textual prompts for each class by prompting GPT-3 [8] to output descriptions of the particular class. We then feed this customised set of prompts output by GPT-3 into Stable Diffusion for generating images. For example, to generate images from the "dog" class, we prompt GPT-3 to describe "dogs", and then prompt Stable Diffusion with the resulting descriptions. In section 4.4, we compare the performance of the default (called *Photo*) and this augmented prompting procedure (called *CuPL*). Unless otherwise specified, all our experiments with Stable Diffusion *support sets* use the *CuPL* strategy.

(II) LAION-5B Retrieval. Our second method leverages the large-scale vision-language dataset, *LAION-5B* [61]. It contains 5.85 billion image-text pairs, pre-filtered by CLIP. Using LAION-5B, we retrieve task-specific images using class text prompts for constructing the *support set*. Concretely, given textual class labels, $\mathcal{T} = \{t_1, t_2, \ldots, t_C\}$, we rank all images in LAION-5B by their CLIP image-text similarity to each text class label t_i, where $i \in [1, C]$. We then use the top N image matches as our *support set* for class i, resulting in an NC-sized *support set* of images with their associated class labels. Note that curating supporting knowledge by search is a classical technique in computer vision [23] that was recently revisited in the task of semantic segmentation [63]. Here we adapt this idea to the *name-only transfer* classification setting. For efficient retrieval, we leverage the approximate nearest neighbour indices released by the authors[4]. Similar to the Stable Diffusion generation approach, we experiment with both *Photo* and *CuPL* prompting strategies for curating our LAION-5B *support set* (see Sec. 4.4). By default, we use *Photo* prompting for all our experiments with LAION-5B *support sets*.

Remark. *SuS* can be seen as a visual analogue to CuPL [56]—we augment VLMs with rich, class-specific images, instead of generating customised text descriptions.

3.2. *TIP-X* Inference

Given our *support set* from the previous section, we now leverage it in a training-free method to inform CLIP's zero-shot predictions. We first briefly review CLIP zero-shot classification and TIP-Adapter [80] (a training-free adaptation method). We then highlight a critical shortcoming in TIP-Adapter due to uncalibrated intra-modal embedding distances, which we address in our method—*TIP-X*.

Zero-shot CLIP. For classification into C classes, CLIP converts class labels into text prompts and encodes them with its text encoder. Collectively, the encoded prompt vectors can be interpreted as a classifier weight matrix $W \in$

[3] Note that (iii) subsumes (ii). (ii) refers to access to labelled data samples from the target dataset whereas (iii) refers to a more general setting where the samples from the target dataset can be unlabelled. We distinguish between the two for clarity.

[4] https://huggingface.co/datasets/laion/laion5B-index

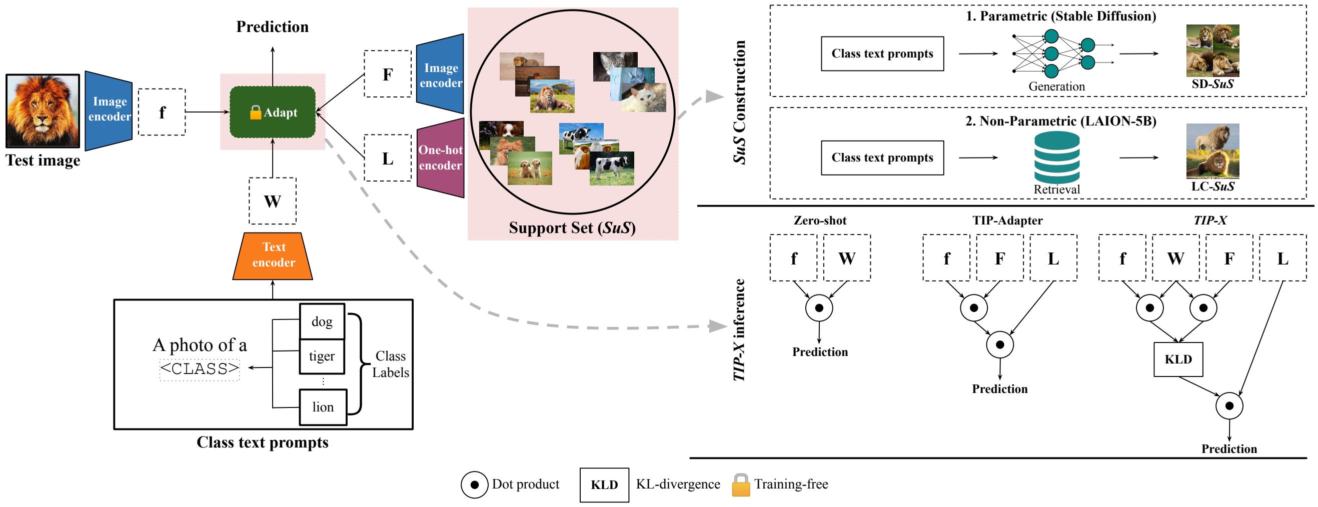


Figure 2: ***SuS-X* for training-free name-only transfer.** *SuS-X* consists of two core building blocks. (1) *SuS* (top right), a dynamic *support set* that we construct to infuse visual information into the VLM based only on knowledge of target category names. We construct support sets either in a parametric (generating images using Stable Diffusion) or non-parametric (retrieving images from LAION-5B) manner. (2) *TIP-X* (bottom right), our novel training-free method that leverages image-text distances to compute similarities between the *support set* and the test images. These similarities act as attention weights for the *support set* labels, and can directly be combined with the original logits from the VLM for classification.

$\mathbb{R}^{C\times d}$, where d is embedding dimension. For a test set $T=\{y_1, y_2, ..., y_t\}$ comprising t test images, CLIP's image encoder is applied to produce test image features:

$$\begin{gathered} f_i = \texttt{CLIPImageEncoder}(y_i), i \in [1,t], f_i \in \mathbb{R}^d \\ f = \texttt{Concat}([f_1, f_2, \dots, f_t]), f \in \mathbb{R}^{t\times d} \end{gathered} \tag{1}$$

Using W and f, CLIP performs classification by computing zero-shot logits (ZSL) via a dot product:

$$\texttt{ZSL} = fW^T \tag{2}$$

TIP-Adapter. Given a CK-sized K-shot labelled dataset $D=\{x_1, x_2, \dots, x_{CK}\}$[5] from the target domain, TIP-Adapter [80] encodes D using CLIP's image encoder:

$$\begin{gathered} F_i = \texttt{CLIPImageEncoder}(x_i), i \in [1,CK], F_i \in \mathbb{R}^d \\ F = \texttt{Concat}([F_1, F_2, \dots, F_{CK}]), F \in \mathbb{R}^{CK\times d} \end{gathered} \tag{3}$$

It then converts each of the few-shot class labels to one-hot vectors $L \in \mathbb{R}^{CK\times C}$. Next, it computes an affinity matrix to capture the similarities between F and f:

$$A = \exp(-\beta(1 - fF^T)) \tag{4}$$

where β is a hyperparameter that modulates "*sharpness*". Finally, these affinities are used as attention weights over L to produce logits that are blended with ZSL using a hyperparameter, α:

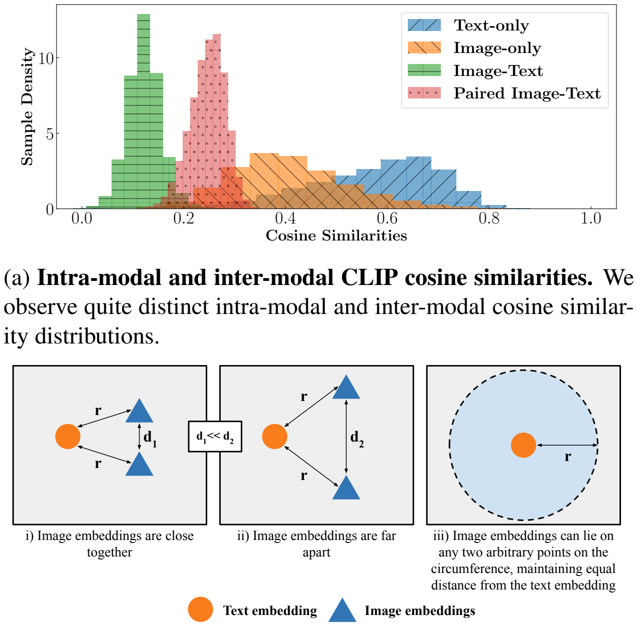
$$\texttt{TL} = \alpha AL + fW^T \tag{5}$$

Motivating *TIP-X*. TIP-Adapter benefits from the affinity computation between the test and few-shot image samples (Eq. (4)). This similarity is computed in CLIP's image space. However, prior research [76, 44, 66] has demonstrated the existence of a *modality gap* between CLIP's image and text spaces. This leads us to question if doing image-image similarity comparisons in CLIP's image space is optimal.

Fig. 3a shows the pairwise image-image, text-text and image-text cosine similarities of the ImageNet validation set CLIP embeddings. Clearly, the intra-modal and inter-modal similarities are distributed differently—the inter-modal similarities have small variance and mean, whereas the intra-modal similarities have larger means and variances. This mismatch happens because *contrastive training of CLIP maximises the inter-modal cosine similarities of paired samples without regard to intra-modal similarities*. This implies that the intra-image CLIP embedding similarities employed by TIP-Adapter may not reflect the true intra-image similarities. Fig. 3b illustrates this idea with a simple example. Consider two image embeddings that are required to be a distance r away from a particular text embedding. The two image embeddings can satisfy this condition by being very close to each other or very far apart from each other. Fig. 3b shows that this constraint can be satisfied by any two arbitrary points on a hypersphere of radius r. While we expect loose constraints to be imposed via transitivity, we nevertheless expect a lower quality of calibration in intra-modal (*e.g*., image-image) comparisons.

***TIP-X* to the rescue.** To get around the problem of un-

[5] Note that a K-shot labelled dataset for C classes has a size CK.



(a) **Intra-modal and inter-modal CLIP cosine similarities.** We observe quite distinct intra-modal and inter-modal cosine similarity distributions.

(b) **Intra-modal degrees of freedom**. Different intra-modal similarities can satisfy same inter-modal constraints, leaving room for poor calibration.

Figure 3: **Our two-fold analysis motivating *TIP-X***

calibrated intra-modal embedding distances in TIP-Adapter, we propose to use inter-modal distances as a bridge. More specifically, rather than computing similarities between the test features ($f \in \mathbb{R}^{t \times d}$) and few-shot features ($F \in \mathbb{R}^{CK \times d}$) in the image embedding space (fF^T), we use the image-text space. We first construct signatures by computing similarities of f and F with the text classifier weights W:

$$\begin{aligned} S &= \texttt{softmax}(FW^T), S \in \mathbb{R}^{CK \times C} \\ s &= \texttt{softmax}(fW^T), s \in \mathbb{R}^{t \times C} \end{aligned} \tag{6}$$

These signatures comprise probability distributions encoding inter-modal affinities between the few-shot features and class text vectors, and likewise for the test features. We then construct our affinity matrix $M \in \mathbb{R}^{t \times CK}$ by measuring the KL-divergence between the signatures as follows:

$$M_{i,j} = \texttt{KL}(s_i || S_j), i \in [1, t], j \in [1, CK] \tag{7}$$

where s_i represents the i^{th} test signature for the t test samples, and S_j represents the j^{th} few-shot signature. Since we are working with discrete probability distributions, we compute the KL-divergence as $\texttt{KL}(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}$.

The construction of the affinity matrix M can be seen as analogous to the affinity computation in TIP-Adapter (Eq. (4)). However, our affinity matrix construction removes direct reliance on the uncalibrated image-image similarities.

Finally, before using our affinity matrix M as attention weights for L (one-hot encoded class labels), we rescale (denoted by ψ) the values of M to have the same range (min, max values) as the TIP-Adapter affinities (A). Further, since our affinity matrix M consists of KL-divergence values, the most similar samples will get small weights since their KL-divergence will be low (close to 0). To mitigate this, we simply negate the values in M. We then blend our predicted logits with $\texttt{TL}$ using a scalar γ:

$$\texttt{TXL} = fW^T + \alpha AL + \gamma \psi(-M)L \tag{8}$$

The entire *TIP-X* method is shown in Fig. 2 (bottom right).

3.3. *SuS-X*: Combining *SuS* and *TIP-X*

Since our constructed *support sets* act as pseudo few-shot datasets, we directly replace the few-shot features F in the *TIP-X* framework with the features of our *support set*. We call our method ***SuS-X-LC*** if we combine *TIP-X* with the LAION-5B curated *support set*, and ***SuS-X-SD*** when combined with the Stable Diffusion generated *support set*. These methods enable *training-free name-only* adaptation of zero-shot VLMs.

4. Experiments

First, we evaluate ***SuS-X*** against strong baselines in the *training-free zero-shot/name-only* transfer regimes, across three VLMs. Next, we illustrate the adaptation of *TIP-X* into the few-shot training-free regime. Finally, we ablate and analyse our method to provide additional insights.

4.1. Training-free name-only transfer evaluation

Datasets. For a comprehensive evaluation, we test on 19 datasets spanning a wide range of object, scene and fine-grained categories: ImageNet [16], StanfordCars [39], UCF101 [64], Caltech101 [22], Caltech256 [29], Flowers102 [52], OxfordPets [55], Food101 [7], SUN397 [71], DTD [13], EuroSAT [34], FGVCAircraft [47], Country211 [57], CIFAR-10 [40], CIFAR-100 [40], Birdsnap [5], CUB [68], ImageNet-Sketch [69] and ImageNet-R [35]. Previous few-shot adaptation methods [77, 25, 82] benchmark on a subset of 11 of these 19 datasets. We report results on the 19-dataset suite in the main paper and compare results using only the 11-dataset subset in the supp. mat.

Experimental Settings. We compare against six baselines. For zero-shot CLIP, we use prompt ensembling with 7 different prompt templates following [57, 80][6]. We run CuPL[7], VisDesc[8] (*name-only* transfer) and CLIP+DN[9]

[6] The 7 prompt templates are: "itap of a `<class>`.", "a origami `<class>`.", "a bad photo of the `<class>`.", "a photo of the large `<class>`.", "a `<class>` in a video game.", "art of the `<class>`.", and "a photo of the small `<class>`.".

[7] https://github.com/sarahpratt/CuPL

[8] https://github.com/sachit-menon/classify_by_description_release

[9] https://github.com/fengyuli2002/distribution-normalization

(*zero-shot* transfer) using their official code. We also experiment with augmenting the CuPL prompts with the original prompt ensemble, and call it CuPL+e. For CALIP (*zero-shot* transfer), in the absence of public code at the time of writing, we aim to reproduce their results using our own implementation. For our proposed methods, we report results using both ***SuS-X-LC*** and ***SuS-X-SD***. For both methods, we use a fixed number of support samples per dataset (see supp. mat. for details). For CALIP and ***SuS-X***, we conduct a hyperparameter search on the dataset validation sets. In Sec. 4.4 we perform a hyperparameter sensitivity test for a fair evaluation. By default, we use the ResNet-50 [32] backbone as CLIP's image encoder for all models.

Main Results. In Tab. 2, we compare both variants of ***SuS-X*** with the baselines. We report an average across 19 datasets. We also include results on ImageNet, EuroSAT, DTD, Birdsnap, ImageNet-R and ImageNet-Sketch (results on all 19 datasets in the supp. mat.). ***SuS-X*** methods outperform zero-shot CLIP by 4.6% on average across all 19 datasets. We observe striking gains of 18%, 8% and 7% on EuroSAT, DTD and Birdsnap respectively. We also outperform the SoTA training-free adaptation methods—CuPL+ensemble and VisDesc by 1.1% and 3.1% on average respectively. To further probe where we attain the most gains, we plot the absolute improvement of our models over zero-shot CLIP in Fig. 4a. We observe large gains on fine-grained (Birdsnap, CUB, UCF101) and specialised (EuroSAT, DTD) datasets, demonstrating the utility of ***SuS-X*** in injecting rich visual knowledge into zero-shot CLIP (additional fine-grained classification analysis in supp. mat.). We further compare ***SuS-X*** to few-shot methods that use labelled samples from the true distribution in the supp. mat.—despite being at a disadvantage due to using no target distribution samples, ***SuS-X*** is still competitive with these methods.

4.2. Transfer to different VLMs

We evaluate transfer to VLMs other than CLIP, namely TCL [72] and BLIP [42]. We only retain image and text encoders of these models for computing features, while preserving all other experimental settings from Sec. 4.1. Tab. 3 shows our ***SuS-X*** methods strongly outperform all baseline methods across both VLMs—we improve on zero-shot models by 11.37% and 5.97% on average across 19 datasets. This demonstrates that our method is not specific to CLIP, but can improve performance across different VLMs.

4.3. Adapting to the few-shot regime

A key component of our ***SuS-X*** method is *TIP-X*. In the previous section, we showcased SoTA results in the training-free name-only transfer regime. Due to its formulation, *TIP-X* can directly be extended to the few-shot regime, where our *support sets* are labelled samples from the target dataset rather than curated/generated samples. To evaluate *TIP-X* on such real-world *support sets*, we conduct training-free few-shot classification using *TIP-X*. We compare against the SoTA method in this regime—TIP-Adapter [80]. We report results on the 11-dataset subset used by TIP-Adapter on five different shot settings of the K-shot classification task: 1, 2, 4, 8 and 16.

We present average accuracy results on all shots in Fig. 4b—*TIP-X* outperforms both Zero-shot CLIP and TIP-Adapter (absolute gain of 0.91% across shots). Notably, on OxfordPets, we achieve 2.1% average gain. This further demonstrates the generalisability of the *TIP-X* method in transferring to the few-shot training-free setting.

4.4. Analysis

We conduct several ablations and provide additional visualisations to offer further insight into the ***SuS-X*** method.

Component Analysis. ***SuS-X*** consists of two major building blocks—*SuS* construction and *TIP-X*. We compare the performance difference (with average accuracy across 19 datasets) of using *SuS* with TIP-Adapter instead of *TIP-X* in Tab. 4. We use both default ensemble prompts and CuPL prompts for CLIP's text classifier to break down the performance gains further. We note that both *SuS* and *TIP-X* are crucial for achieving the best results.

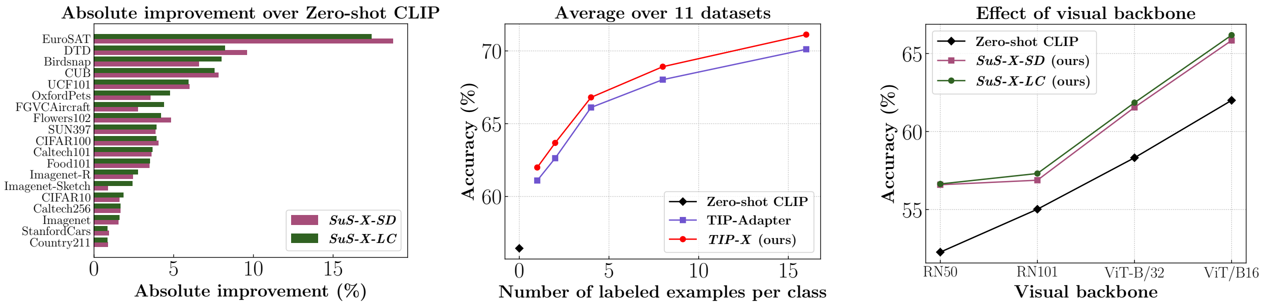
Transfer to different visual backbones. We evaluate the scalability of our model across different CLIP visual backbones— Fig. 4c shows that both ***SuS-X*** variants consistently improve upon zero-shot CLIP across ResNet and VisionTransformer backbones of varying depths and sizes.

***SuS* size.** We study the effect of varying *support set* size for *SuS-LC* and *SuS-SD*—we generate three different *support sets* with random seeds for support sizes of 1, 5, 10, 25, 50, 75 and 100 samples. From Fig. 6, we observe two broad trends—some tasks benefit (ImageNet-R, DTD) from having more *support set* samples while others do not (Country211, Flowers102). We suggest that this is connected to the domain gap between the true data distribution and *support set* samples—if the domain gap is large, it is inimical to provide a large *support set*, whereas if the domains are similar, providing more support samples always helps.

***SuS* visualisation.** We visualise samples from both *support set* construction methods on ImageNet in Fig. 5. It is hard to distinguish between the true ImageNet samples and the *SuS* samples—we can therefore construct *support sets* to mimic the true data distribution, with access to only the category names. A caveat is that the *support set* does not always capture the domain characteristics of the true distribution, leading to a domain gap (lighting conditions, diverse scene backgrounds, confounding objects etc). To fully close the gap to using true few-shot datasets as *support sets* [25, 80], further research into exact unsupervised domain matching of *support sets* and few-shot datasets is required.

Table 2: **Training-free adaptation of CLIP on 19 datasets with RN50 visual backbone**. The best and second best results for each dataset are **bolded** and underlined, respectively. Individual results for all 19 datasets are available in the supp. mat. *Average reported across 19 datasets. †Our re-implementation.

	Method	Average*	ImageNet [16]	ImageNet-R [35]	ImageNet-Sketch [69]	EuroSAT [34]	DTD [13]	Birdsnap [5]
Zero-shot	Zero-shot CLIP [57]	52.27	60.31	59.34	35.42	26.83	41.01	30.56
	CALIP [31]	–	60.57	–	–	38.90	42.39	–
	CALIP [31]†	52.37	60.31	59.33	36.10	26.96	41.02	30.68
	CLIP+DN [85]	53.02	60.16	60.37	35.95	28.31	41.21	31.23
Name-only	CuPL [56]	55.50	61.45	61.02	35.13	38.38	48.64	35.65
	CuPL+e	55.76	61.64	61.17	35.85	37.06	47.46	35.80
	VisDesc [49]	53.76	59.68	57.16	33.78	37.60	41.96	35.65
	SuS-X-SD (ours)	56.73	61.84	61.76	36.30	**45.57**	**50.59**	37.14
	SuS-X-LC (ours)	**56.87**	**61.89**	**62.10**	**37.83**	44.23	49.23	**38.50**



(a) (b) (c)

Figure 4: **(a)** Comparison of ***SuS-X*** with Zero-shot CLIP. **(b)** Results of training-free few-shot classification. **(c)** Performance comparison of ***SuS-X*** across visual backbones.

Table 3: ***SuS-X* generalises to different VLMs.** *Average reported across 19 datasets.

VLM	Method	Average*	ImageNet	EuroSAT	DTD	Birdsnap
TCL	Zero-shot	31.38	35.55	20.80	28.55	4.51
	CuPL	34.79	41.60	26.30	42.84	6.83
	CuPL+e	32.79	41.36	25.88	41.96	6.60
	VisDesc	33.94	40.40	21.27	34.28	5.69
	SuS-X-SD	41.49	52.29	28.75	**48.17**	13.60
	SuS-X-LC	**42.75**	**52.77**	**36.90**	46.63	**17.93**
BLIP	Zero-shot	48.73	50.59	44.10	44.68	10.21
	CuPL	51.11	52.96	39.37	52.95	12.24
	CuPL+e	51.36	53.07	41.48	53.30	12.18
	VisDesc	49.91	50.94	42.25	47.45	11.69
	SuS-X-SD	53.20	55.93	45.36	**56.15**	16.95
	SuS-X-LC	**54.64**	**56.75**	**51.62**	55.91	**23.78**

Table 4: **Component Analysis of *SuS-X*.**

Text Prompts	Method	*SuS*	*TIP-X*	Average Accuracy
Default	Zero-shot CLIP	✗	✗	52.27
	SuS-TIP-SD	✓	✗	53.49 (+1.22%)
	SuS-X-SD	✓	✓	53.69 (+1.42%)
	SuS-TIP-LC	✓	✗	53.83 (+1.56%)
	SuS-X-LC	✓	✓	54.20 (+1.93%)
CuPL+e	CuPL+e	✗	✗	55.76 (+3.49%)
	SuS-TIP-SD	✓	✗	56.63 (+4.36%)
	SuS-X-SD	✓	✓	56.73 (+4.46%)
	SuS-TIP-LC	✓	✗	56.72 (+4.45%)
	SuS-X-LC	✓	✓	**56.87** (+4.60%)

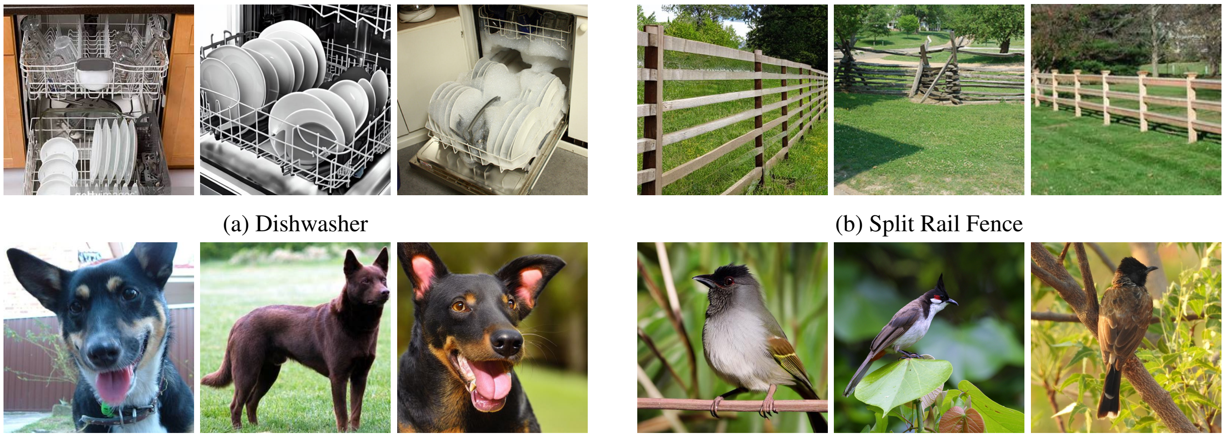
Table 5: **Prompting strategies for *SuS* construction.**

SuS method	Average Acc.		ImageNet Acc.		Diversity	
	Photo	*CuPL*	*Photo*	*CuPL*	*Photo*	*CuPL*
LC	**56.87**	56.20	**61.89**	61.79	0.28	**0.32**
SD	56.32	56.73	61.79	61.84	0.17	**0.20**

Prompting strategies for *SuS* construction. Tab. 5 depicts the performance of *Photo* and *CuPL* prompting—best results are achieved with the *LC-Photo* and *SD-CuPL* strategies. We further compare the diversity of images produced by the two strategies on ImageNet[11]—from Tab. 5, it is evident that *CuPL* prompting leads to more diverse support sets as compared to *Photo* prompting.

Hyperparameter Sensitivity. We perform a sensitivity test for our γ hyperparameter (refer Eq. 8) on ImageNet-R, OxfordPets, and DTD. We fix α and β to be 1, and run a sweep over $\gamma \in [0, 1]$. From Tab. 6, we observe that moderate values of γ are typically preferred, and the variance of the accuracy values is small. However, note that for DTD, the optimal γ is slightly larger (0.75)—this is due to its spe-

[11] We compute diversity as 1 minus the mean of the average pairwise image cosine-similarities within a class. A larger value implies low cosine similarities across images within a class, implying more diverse images. Alternatively, a smaller value implies less diverse images.



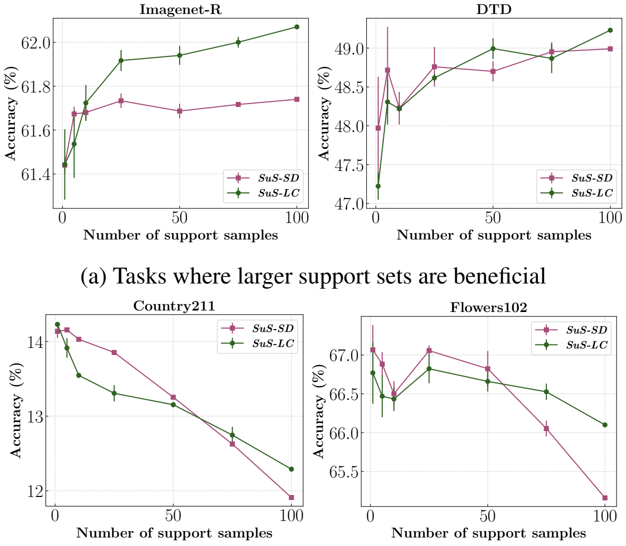
(a) Dishwasher

(b) Split Rail Fence

(c) Australian Kelpie

(d) Bulbul

Figure 5: **Support samples from the generated *SuS-SD*, retrieved *SuS-LC* and true training distribution for ImageNet.** By randomising the image order in each subfigure, we pose a challenge question—can you match the three images for each subfigure to their source *i.e. SuS-SD*, *SuS-LC* or ImageNet train set? The answers are provided at the bottom of the page[10].



(a) Tasks where larger support sets are beneficial

(b) Tasks where larger support sets are harmful

Figure 6: **Effect of support size.**

Table 6: **Hyperparameter sensitivity for γ**

Dataset	γ value						
	0	0.1	0.2	0.3	0.5	0.75	1
ImageNet-R	60.87	60.98	61.03	**61.05**	61.00	60.89	60.65
OxfordPets	76.76	77.17	**77.58**	77.44	77.17	77.17	76.90
DTD	47.16	47.16	47.51	47.69	47.87	**47.96**	47.60

cialised nature which requires more guidance from the specialised *support set* to inform pre-trained CLIP. Previous few-shot adaptation works [25, 80] observed similar results. For more hyperparameter ablations, see the supp. mat.

4.5. Limitations and broader impact

While demonstrating promising results, we note some limitations of our approach: (1) To perform *name-only* transfer, we rely on CLIP having seen related concepts during pre-training. For rare concepts not seen during pre-training, transfer might not be feasible. (2) We employ LAION-5B [61] as a source of knowledge. While reasonable for a proof of concept, this data is relatively uncurated and may contain harmful content. As such, our approach is unsuitable for real-world deployment without careful mitigation strategies to address this. Similar arguments apply to Stable Diffusion [59].

5. Conclusion

In this paper, we studied the training-free name-only transfer paradigm for classification tasks with vision-language models. We systematically curated *support sets* with no access to samples from the target distribution and showed that they help improve CLIP's zero-shot predictions by providing rich, task-specific knowledge. We further motivated the *TIP-X* framework through the observation that CLIP's intra-modal embedding spaces are not optimal for computing similarities. With these two building blocks, we demonstrated superior performance to prior state-of-the-art.

Acknowledgements. This work was supported by the Isaac Newton Trust and an EPSRC access-to-HPC grant. SA would like to acknowledge the support of Z. Novak and N. Novak in enabling his contribution. VU would like to thank Gyungin Shin, Surabhi S. Nath, Jonathan Roberts, Vlad Bogolin, Kaiqu Liang and Anchit Jain for helpful discussions and feedback.

[11] (a)LC,SD,Train,(b)SD,Train,LC,(c)Train,LC,SD,(d)SD,Train,LC

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker's guide to long video retrieval. *arXiv preprint arXiv:2205.08508*, 2022.

[3] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of clip features for image captioning: An experimental analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4662–4670, 2022.

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[5] 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.

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[7] 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.

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

[9] Adrian Bulat and Georgios Tzimiropoulos. Language-aware soft prompting for vision & language foundation models. *arXiv preprint arXiv:2210.01115*, 2022.

[10] Delong Chen, Zhao Wu, Fan Liu, Zaiquan Yang, Yixiang Huang, Yiping Bao, and Erjin Zhou. Prototypical contrastive language image pretraining. *arXiv preprint arXiv:2206.10996*, 2022.

[11] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022.

[12] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[13] 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.

[14] Andrea Cossu, Tinne Tuytelaars, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, and Davide Bacciu. Continual pre-training mitigates forgetting in language and vision. *arXiv preprint arXiv:2205.09357*, 2022.

[15] Guillaume Couairon, Matthijs Douze, Matthieu Cord, and Holger Schwenk. Embedding arithmetic of multimodal queries for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4950–4958, 2022.

[16] 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.

[17] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Variational prompt tuning improves generalization of vision-language models. *arXiv preprint arXiv:2210.02390*, 2022.

[18] 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.

[19] Kun Ding, Ying Wang, Pengzhang Liu, Qiang Yu, Haojian Zhang, Shiming Xiang, and Chunhong Pan. Prompt tuning with soft context sharing for vision-language models. *arXiv preprint arXiv:2208.13474*, 2022.

[20] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022.

[21] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). *arXiv preprint arXiv:2205.01397*, 2022.

[22] 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.

[23] Robert Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. Learning object categories from google's image search. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1816–1823. IEEE, 2005.

[24] Benjamin Feuer, Ameya Joshi, and Chinmay Hegde. Caption supervision enables robust learners. *arXiv preprint arXiv:2210.07396*, 2022.

[25] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[26] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095*, 2022.

[27] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021.

[28] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022.

[29] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[30] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[31] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.

[34] 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.

[35] 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.

[36] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

[37] 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.

[38] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[39] 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.

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

[41] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009.

[42] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[43] 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.

[44] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022.

[45] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. *arXiv preprint arXiv:2205.03340*, 2022.

[46] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. *arXiv preprint arXiv:2207.07285*, 2022.

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

[48] Shu Manli, Nie Weili, Huang De-An, Yu Zhiding, Goldstein Tom, Anandkumar Anima, and Xiao Chaowei. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022.

[49] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

[50] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[51] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *arXiv preprint arXiv:2208.05516*, 2022.

[52] 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.

[53] Zachary Novack, Saurabh Garg, Julian McAuley, and Zachary C Lipton. Chils: Zero-shot image classification with hierarchical label sets. *arXiv preprint arXiv:2302.02551*, 2023.

[54] Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. Svl-adapter: Self-supervised adapter for vision-language pretrained models. *arXiv preprint arXiv:2210.03794*, 2022.

[55] 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.

[56] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022.

[57] 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.

[58] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjarn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[60] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a controlled study for representation learning. *arXiv preprint arXiv:2207.07635*, 2022.

[61] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[62] Gyungin Shin, Weidi Xie, and Samuel Albanie. Namedmask: Distilling segmenters from complementary foundation models. *arXiv preprint arXiv:2209.11228*, 2022.

[63] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *arXiv preprint arXiv:2206.07045*, 2022.

[64] 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.

[65] Tejas Srinivasan, Ting-Yun Chang, Leticia Leonor Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. Climb: A continual learning benchmark for vision-and-language tasks. *arXiv preprint arXiv:2206.09059*, 2022.

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

[67] Vishaal Udandarao, Abhishek Maiti, Deepak Srivatsav, Suryatej Reddy Vyalla, Yifang Yin, and Rajiv Ratn Shah. Cobra: Contrastive bi-modal representation algorithm. *arXiv preprint arXiv:2005.03687*, 2020.

[68] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[69] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[70] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[71] 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.

[72] 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.

[73] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[74] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. *arXiv preprint arXiv:2207.12661*, 2022.

[75] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[76] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, JaeSung Park, Ximing Lu, Prithviraj Ammanabrolu, Rowan Zellers, Ronan Le Bras, Gunhee Kim, et al. Multimodal knowledge alignment with reinforcement learning. *arXiv preprint arXiv:2205.12630*, 2022.

[77] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[78] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *arXiv preprint arXiv:2303.02151*, 2023.

[79] Renrui Zhang, Longtian Qiu, Wei Zhang, and Ziyao Zeng. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*, 2021.

[80] 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. *arXiv preprint arXiv:2207.09519*, 2022.

[81] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

[82] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021.

[83] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. *arXiv preprint arXiv:2203.05557*, 2022.

[84] 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.

[85] Yifei Zhou, Juntao Ren, Fengyu Li, Ramin Zabih, and Ser-Nam Lim. Distribution normalization: An "effortless" test-time augmentation for contrastively learned visual-language models. *arXiv preprint arXiv:2302.11084*, 2023.