Incorporating Geo-Diverse Knowledge into Prompting for Increased Geographical Robustness in Object Recognition

Kyle Buettner[1], Sina Malakouti[2], Xiang Lorraine Li[1,2], Adriana Kovashka[1,2]
[1]Intelligent Systems Program, [2]Department of Computer Science, University of Pittsburgh, PA, USA
{buettnerk, sem238}@pitt.edu, {xianglli, kovashka}@cs.pitt.edu
https://krbuettner.github.io/GeoKnowledgePrompting

Abstract

Existing object recognition models have been shown to lack robustness in diverse geographical scenarios due to domain shifts in design and context. Class representations need to be adapted to more accurately reflect an object concept under these shifts. In the absence of training data from target geographies, we hypothesize that geographically diverse descriptive knowledge of categories can enhance robustness. For this purpose, we explore the feasibility of probing a large language model for geography-based object knowledge, and we examine the effects of integrating knowledge into zero-shot and learnable soft prompting with CLIP. Within this exploration, we propose geography knowledge regularization to ensure that soft prompts trained on a source set of geographies generalize to an unseen target set. Accuracy gains over prompting baselines on DollarStreet while training only on Europe data are up to +2.8/1.2/1.6 on target data from Africa/Asia/Americas, and +4.6 overall on the hardest classes. Competitive performance is shown vs. few-shot target training, and analysis is provided to direct future study of geographical robustness.

1. Introduction

The performance of object recognition models degrades when tested in new geographies (e.g., cities, countries, continents) [7, 21, 33, 39, 43]. Numerous factors contribute to the challenging problem of *geographical domain shift*, such as cross-geography changes in object design/parts, materials, and context. These changes in turn may be due to cultural, climate, or economic differences around the world. Recent work has shown standard adaptation techniques fail when used for geographical domain shifts [21, 33], but there has yet to be significant progress in the creation of techniques that improve geographical robustness. Such progress is necessary to ensure equitable use of AI in the future.

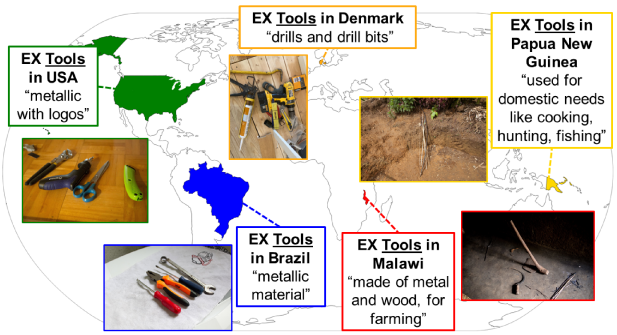


Figure 1. **Descriptive knowledge can address concept shifts across geographies.** Observe the wide range of object designs and contexts in the DollarStreet [11] category *tools* around the world. Our work's premise is that textual representations for classes in vision-language models can be enhanced to better suit diverse object representations across geographies. Map made with [16].

Overall, models need representations that adequately capture a category's various forms around the world. A natural solution is to collect training data of objects from different regions. However, this approach is expensive, takes significant effort, and is difficult for regions with limited Internet access. Fortunately, geographical shifts have a unique property compared to other common domain shifts (*e.g.* ones due to artistic style or weather changes)—they can be addressed with *descriptive knowledge* about concept changes. In other words, it is possible to describe the features of an object in a region and use this information to adapt a model's default representation. For instance, as shown in Fig. 1, for rural areas in Papua New Guinea, *tools* can be described as being used for "cooking, hunting, and fishing", and for rural areas in Malawi, *tools* may often be "made of metal and wood, for farming". Models should account for diverse presentations and contexts of a category and not be limited to biased presentations (*e.g.* if the model learns *tools* as just being "metallic with logos").

We examine the effects of probing geo-diverse knowledge in two ways. First, we analyze whether a vision-

language model (VLM, i.e. CLIP [36]) has encoded categories in a geo-specific manner, such that adding a country's name to a prompt (*e.g.* "A photo of a house in China") elicits knowledge that improves recognition. Second, we probe a large language model (LLM, i.e. GPT-3 *davinci-003*) for geography-specific knowledge to obtain visual feature descriptors for an object in different locations. We analyze results in zero-shot inference on geographically and socioeconomically diverse data (DollarStreet [11]), finding the combination of knowledge to often be complementary.

We further consider a practical scenario where CLIP is optimized with soft prompting, using only a "source" geography with easy-to-access data (*e.g.* Europe), while the model is applied downstream on "target" data from other parts of the world (*e.g.* Africa, Asia, Americas). We propose *geography knowledge regularization*, which uses knowledge ensembled over countries to enable soft prompts to achieve geographically generalizable class representations. We test our method on the recent DollarStreet and GeoNet [21] datasets. Our regularization boosts performance over baseline soft prompting methods, and has benefits with respect to few-shot target-specific training (a 16-shot-per-class regularized model without any target data outperforms a 12-shot-per-class target-trained model on DollarStreet).

Our method is the first to effectively address geo shifts in object recognition. It outperforms zero-shot CLIP (assumed to have some robustness) by 10.3% on Africa, CoOp [52] by 3.3%, and the best baseline by 4.6% on the hardest classes.

To summarize, we answer the following questions: (1) Does adding geographical context (i.e. country names) to CLIP prompts improve recognition across geographies? (2) Can an LLM provide useful geographical descriptive knowledge to improve recognition? (3) How can we optimize soft prompts for CLIP using an accessible data source with consideration of target geographies not represented in the training set? (4) Where can soft prompts enhanced with geographical knowledge provide the most benefits?

2. Related Work

Geographical domain shifts occur when the target setting is in a different geography (*e.g.* continent, country, city) than where the source data was acquired. Shifts involve changes in object design (*e.g.* differences in house architecture) and context (i.e. background/co-occurring objects vary). Datasets tailored to cross-country/continent object recognition have recently been proposed, e.g. DollarStreet [11], GeoNet [21], GeoDE [37], GeoYFCC [9], and OpenImages-Extended [5]. Interestingly, [21, 33] demonstrate that *traditional methods in unsupervised domain adaptation* [10, 19, 20, 27, 28, 38, 44, 45, 50] which seek to bridge gaps based on visual features alone, *do not effectively address geographical domain shift*. They achieve negligible gains (e.g. 0.14 for [10] in [33]) or often drops in performance (e.g. all methods tested in [21]), compared to just using the source model. Attempts to specifically address geographical robustness are limited: [43] corrects for differences in the sizes of cars, [9] proposes a discriminative domain embedding from target data, and GiVL [48] pretrains with knowledge from Wikipedia. In contrast, our descriptive knowledge regularization works for different categories (not just cars); we do not require target domain data to achieve gains cross-geography; we explore the strong capabilities of LLMs to gather relevant knowledge; and we propose lightweight adaptation through soft prompting (unlike GiVL's expensive pretraining).

Vision-language (VL) models [17, 25, 26, 36, 49] excel on a variety of tasks. CLIP [36] shows impressive zero-shot object recognition across different settings. Yet *its performance given geographical shift is less apparent*. GeoNet [21] only shows finetuned performance, which is expensive given CLIP's large scale. GeoDE [37] only shows zero-shot inference with CLIP's default prompts. Neither work evaluates descriptive knowledge or soft prompting.

Learning soft textual prompts. Several recent works to adapt CLIP have focused on parameter and data efficiency using linear probing [36] and prompting [18, 23, 52]. Soft textual prompting (*e.g.* CoOp [52]) is notable as it optimizes class text embeddings (without manual tuning), which we hypothesize is critical to adequately adapt for geographical robustness. As CoOp overfits on base (seen) classes, CoCoOp [51] proposes to condition prompts on the image for better generalizability. KgCoOp [47] alternatively guides learned prompt embeddings towards CLIP's manual prompt embeddings through a distance constraint to avoid degradation on unseen classes. Our approach also uses a distance constraint, but it differs from [47] with the purpose of regularizing learned prompt representations for cross-*geography* generalization instead of the base-to-new-*class* setting. We also show novel benefits of regularization when used with an *ensemble* of CLIP's internal geographical knowledge and *external* geographical descriptive knowledge. Our approach notably outperforms each of CoOp, CoCoOp, and KgCoOp by at least +2.8 accuracy on target countries in Africa in DollarStreet. External knowledge aids unseen classes in KAPT [22], but *not with respect to geographical knowledge*. Prompt tuning for adaptation has been tested in [12, 40], but not with descriptive knowledge.

Knowledge probed from large language models like [4, 6, 31, 32] has been used for visual reasoning [46], embodied agent planning [15, 41], and to generate additional context for VLM class prompts in object recognition [30, 34]. We uniquely probe LLMs for *distinguishable visual descriptions for the same object class across different geographical regions*. We are also the first to incorporate geographical knowledge from LLMs into soft prompting.

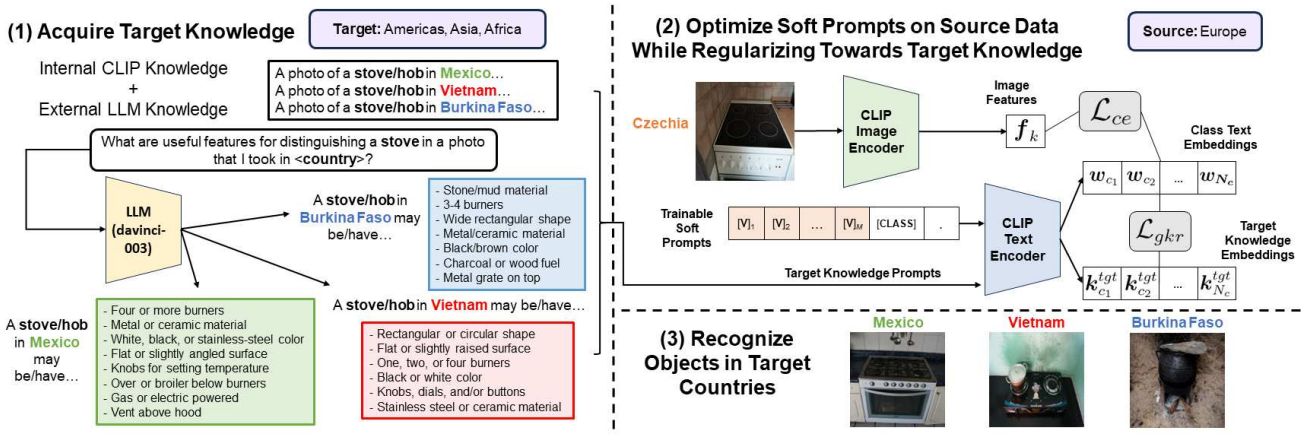


Figure 2. **Geography knowledge regularization.** To ensure robustness in soft prompt learning, we (1) incorporate knowledge internal to CLIP and externally obtained from an LLM. (2) This descriptive knowledge regularizes class representations when training on a specific source geography (*e.g.* Europe), thus (3) increasing robustness when generalizing to target geographies (*e.g.* Vietnam).

3. Approach

We investigate geographical shift in object recognition with VLMs. We posit that the manner in which classes are *described* is critical due to cross-geography shifts in design and context. We also hypothesize that CLIP's default class representations elicited through "a photo of a/an <object>" prompts may not adequately represent classes around the world. Instead, they may be more aligned to high-resource geographies due to Internet-based training data. Optimizing representations (with soft prompts) on a specific geography (*e.g.* Europe) may exacerbate a lack of robustness. Our main idea (Fig. 2) is to incorporate object-related geographical knowledge into prompting to ensure model robustness in different regions. We outline our mechanism to obtain geography-specific context by probing CLIP's internal knowledge and an external LLM's descriptive knowledge. We further propose geography knowledge regularization to ensure soft prompts do not overfit when training data is limited to certain geographies.

Preliminaries. We consider object recognition on a dataset $\mathcal{S}$ containing a class set $\mathcal{C}$ (size N_c) over a set of geographies $\mathcal{G}$. We consider a geography g to be either a country or continent. Our VLM is CLIP [36], with an image encoder f and language encoder h. We incorporate knowledge of each geography g into prompting using (1) zero-shot inference or (2) soft textual prompting. Prompts are defined as $\boldsymbol{t}$ (each is a set of tokens), and class embeddings $\boldsymbol{w}$ are calculated as $h(\boldsymbol{t})$. We refer to CLIP's default prompt "a photo of a/an <object>" for a class c as $\boldsymbol{t}_c^{\text{default}}$.

3.1. Geographical Knowledge Probing

Probing CLIP's internal geographical knowledge. Our first strategy of investigation is to augment CLIP's manual prompts to include country names, as we surmise that some of the resulting class representations may be better aligned to how categories present in different regions. [3] inspires this hypothesis, showing that adding country names to image generation prompts can achieve gains in geographical representativeness. However, it is an open question whether adding country names in prompts improves recognition. We define the setting **CountryInPrompt**, using the prompt $\boldsymbol{t}_c^{\text{CountryInPrompt}}$ with template "a photo of a/an <object> in <country>", *e.g.* "a photo of a stove in Burundi."

Probing external LLM geographical knowledge. As CLIP may not have sufficient knowledge of objects in some regions, we consider further augmenting prompts with *external* knowledge. Motivated by probing LLMs for general attribute-based object descriptions [30, 34] (*e.g.* a *tiger* with "stripes and sharp teeth"), we probe GPT-3 (*davinci-003*) for *geography-specific* descriptions of object styles, contexts, and materials.[1] We reason that since LLMs are trained on large information sources (*e.g.* CommonCrawl [1], WebText [35], Wikipedia [2]), they may have knowledge about how an object presents in a region due to climate, economics, and/or cultural factors. For instance, *roofs* may sometimes be "thatched" in tropical and temperate climates, and *cutlery* may sometimes be made of "bamboo" in areas with bamboo forests. Our goal is unique vs. [30, 34] in that we explore descriptive knowledge differences for the same class to address domain shifts across regions.

Acquiring knowledge. We follow [30], but instead of gathering one set of feature descriptors $\mathcal{D}(c)$ for each c, we collect sets *per country*. For each class c and geography g, we prompt the LLM to generate descriptor lists $\mathcal{D}_g(c)$, using a template consisting of an example question, answer, and format. We use 1-shot prompting to show how to capture geographically representative object designs and contexts.

[1] We found ChatGPT to perform worse than GPT-3, also found in [34].

Our prompt exemplifies this below, using the descriptors for Japanese *ofuro* (お風呂, *bathtub*):

> **Q:** What are useful features for distinguishing a bathtub in a photo that I took in Japan?
> **A:** There are several useful visual features to tell there is a bathtub in a photo that I took in Japan:
> - short in length and deep
> - square shape
> - wooden, plastic, or steel material
> - white or brown color
> - benches on side
> - next to shower
>
> **Q:** What are useful features for distinguishing <category> in a photo that I took in <country>?
> **A:** There are several useful visual features to tell there is/are <category> in a photo that I took in <country>:

Using knowledge. To convert LLM outputs to CLIP prompts, each descriptor d in $\mathcal{D}_g(c)$ serves in a prompt $\boldsymbol{t}_{c,d}$. The format of $\boldsymbol{t}_{c,d}$ is "a photo of a/an <object> which (is/has/etc.) <descriptor>". The setting where geography-specific LLM descriptors are used in prompting is referred to as **CountryLLM** (prompts $\boldsymbol{t}_{c,d}^{\text{CountryLLM}}$), while [30] is **GeneralLLM** (prompts $\boldsymbol{t}_{c,d}^{\text{GeneralLLM}}$). To perform zero-shot inference on an image I, each class score $s(c, I)$ is computed using the average of CLIP logits $\phi(I, d)$ over each d in the set $\mathcal{D}$. For GeneralLLM, the score is calculated as:

$$s(c, I) = \frac{1}{|\mathcal{D}(c)|} \sum_{d \in \mathcal{D}(c)} \phi(I, d) \tag{1}$$

For CountryLLM, we use the geo-specific set:

$$s(c, I, g) = \frac{1}{|\mathcal{D}_g(c)|} \sum_{d \in \mathcal{D}_g(c)} \phi(I, d) \tag{2}$$

The argmax of s with respect to c is taken as the prediction. Due to averaging over descriptor scores, not every descriptor needs to strongly activate in a correct prediction. The model therefore can account for diverse features of objects within a geography. These descriptors effectively serve as *complements* to CLIP's default knowledge of class names.

Combining knowledge. Our third method of exploration, **CountryInPrompt+LLM**, combines both CLIP's internal knowledge and LLM external knowledge. The prompt template $\boldsymbol{t}_{c,d}^{\text{CountryInPrompt+LLM}}$ is "a photo of a/an <object> in <country> which (is/has/etc.) <descriptor>".

3.2. Regularizing Soft Prompts via Geo Knowledge

Adaptation scenario. In practice, one may want to further optimize a VLM for a downstream task. To update a model effectively, one promising strategy is soft textual prompting. It is parameter-efficient [52] and avoids feature distortion unlike finetuning [24]. Its mechanism is to learn *context* parameters that directly change the *class text embeddings* used in inference. We posit that learning context on a dataset with limited diversity (*e.g.* just Europe) may tailor these class representations to the region and overfit. To investigate cross-geography generalization when using soft prompting, we pose a domain generalization scenario where we aim to learn only from a high-resource *source set* of countries and generalize to a *target set* of countries at inference time. A method that performs well in this setting could provide a viable alternative to few-shot target training when acquiring target data for training is not feasible.

Soft prompts. Our idea is to learn soft prompts while constraining the class text embeddings to be close to geographical knowledge of objects *outside* of source geographies. In this way, we hope to learn class representations that are more applicable to the rest of the world. Building from CoOp [52], we assume there is a text prompt $\boldsymbol{t}_c$ for each class c. All prompts share M context vectors (each denoted $[\text{V}]_m$), which are the same size as the word embeddings (i.e. 512-D) and precede a class name token $[\text{CLASS}_c]$:

$$\boldsymbol{t}_c = [\text{V}]_1[\text{V}]_2 ... [\text{V}]_M [\text{CLASS}_c] \tag{3}$$

The respective class text embedding $\boldsymbol{w}_c$ is produced as $h(\boldsymbol{t}_c)$, forwarding the prompt through the text encoder. Learning proceeds by minimization of cross-entropy, for image k with features $\boldsymbol{f}_k$, using ground-truth *source* labels $y_{k,c}$ and temperature τ:

$$\mathcal{L}_{ce} = -\sum_{c=1}^{N_c} y_{k,c} \log \frac{\exp(\cos(\boldsymbol{w}_c, \boldsymbol{f}_k)/\tau)}{\sum_{j=1}^{N_c} \exp(\cos(\boldsymbol{w}_j, \boldsymbol{f}_k)/\tau)} \tag{4}$$

Geography knowledge regularization (gkr). We minimize the cosine distance of normalized class embedding $\boldsymbol{w_c}$ and overall target class knowledge $\boldsymbol{k}_c^{tgt}$, over all c:

$$\mathcal{L}_{gkr} = 1 - \frac{1}{N_c} \sum_{c=1}^{N_c} \cos(\boldsymbol{w_c}, \boldsymbol{k}_c^{tgt}) \tag{5}$$

Geo knowledge ensemble. To define $\boldsymbol{k}_c^{tgt}$, we identify that a model may be deployed in various locations. Therefore, we define a *target* geography set $\mathcal{G}_t$, which can practically be thought of as the countries that a model may be deployed in that are not in the training set $\mathcal{D}$ (*e.g.* Africa, Asia, Americas in $\mathcal{G}_t$ if only Europe in $\mathcal{D}$). Then for each geography g in $\mathcal{G}_t$, we define the corresponding class knowledge $\boldsymbol{k}_c^g$ as:

$$\boldsymbol{k}_c^g = \frac{1}{|\mathcal{D}_g(c)|} \sum_{d \in \mathcal{D}_g(c)} \boldsymbol{w}_{c,d}^{\text{CountryInPrompt+LLM}} \tag{6}$$

This is defined analogously for CountryInPrompt and CountryLLM. The final regularization target $\boldsymbol{k}_c^{tgt}$ for class c aggregates the set's geographical knowledge:

$$\boldsymbol{k}_c^{tgt} = \frac{1}{|\mathcal{G}_t|} \sum_{g \in \mathcal{G}_t} \boldsymbol{k}_c^g \tag{7}$$

While the loss formulation includes cosine distance like KgCoOp [47], it serves a different purpose: we regularize for *cross-geography domain generalization*, while KgCoOp regularizes for base-class-to-new-class inference. Our method outperforms KgCoOp in cross-geography generalization due to its use of geo-specific knowledge.

Overall loss. The final loss $\mathcal{L}$ for learning soft prompts, where λ controls the strength of regularization, is:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{gkr} \tag{8}$$

4. Experimental Setup

Datasets. We use *DollarStreet* [11], which has 38,479 images of household objects across regions (Africa, South/Central/North America, Asia, Europe) and incomes. The classes may represent abstract concepts (*e.g. most loved toys*), so we narrow focus to 95 object classes. We merge especially close categories (*light sources by bed/in living room*) and ignore multi-label examples, resulting in 23,114 total images. For zero-shot inference, the entire set is used. For training, the source is Europe, and the target is Americas, Asia, and Africa. 20% of source data, stratified based on class proportions, is heldout for testing; target evaluation is on all data from target continents. To set up $\boldsymbol{k}_c^{tgt}$, the 49 target countries in DollarStreet make up $\mathcal{G}_t$. We also use the GeoImNet benchmark of *GeoNet* [21], comprised of 171,692 images across 600 objects from the USA (source) and 78,358 images across the same number in Asia (target). We use existing train-test splits for soft prompt training. For GeoNet, given the relatively large number of categories and inference costs of *davinci-003*, $\boldsymbol{k}_c^{tgt}$ and $\mathcal{G}_t$ use the top 10 most frequent countries in the GeoNet set.

Baselines. We evaluate geography knowledge regularization vs. CoOp [52], CoCoOp [51] and KgCoOp [47]. For zero-shot inference, we evaluate CLIP with default prompts and the classification via description method of [30].

Metrics. We report balanced accuracy, which is the average of per-class recall scores. We use this metric to account for class imbalance in both DollarStreet and GeoNet. For zero-shot inference, we also show top-3 accuracy as some similar categories exist (*e.g. cooking utensils*, *cutlery*).

Experimental details. For all soft prompting experiments, models are trained with 16 shots, context length $M = 4$, and for 100 epochs, unless otherwise stated. The class token position follows the soft prompts, and class-shared context is used. Our method uses a batch size of 128 (same as KgCoOp), while the batch sizes for CoOp and CoCoOp follow [47] (i.e. 32, and 1 for CoCoOp due to memory limitations). The encoders used for training include ViT-B/16 [8] and ResNet-50 (RN50) [14] as reported in [47]. Both our method and KgCoOp use a regularization weight λ. We set $\lambda = 4$ for DollarStreet, and compare to KgCoOp at $\lambda = 4$ (which performs better than KgCoOp's default $\lambda = 8$). For GeoNet, we use $\lambda = 8$. Training is performed on 1 NVIDIA Quadro RTX A5000 GPU with 24 GB of memory. All reported soft prompt results are averages over 3 runs. For experiments in the zero-shot setting, results are shown over ViT-B/16, ViT-B/32, and RN50 encoders. LLM descriptors for all experiments are generated from the *davinci-003* version of GPT-3, with max tokens 100 and temperature 0.7.

5. Results

5.1. Zero-shot CLIP Inference with Geo Knowledge

We gauge the effectiveness of three zero-shot strategies: (1) CountryInPrompt (including countries in prompts to probe CLIP's knowledge), (2) CountryLLM (gathering descriptive knowledge of objects with *davinci-003*), and (3) CountryInPrompt+LLM (using country names and LLM knowledge). We compare to [30] (GeneralLLM) and CLIP with manual prompts (i.e. "a photo of a/an <object>"). Results on DollarStreet are shown in Table 1.

Including country names in prompts can improve object recognition, especially in Africa and Asia. This observation is supported by gains for CountryInPrompt vs. Zero-Shot CLIP, especially in Africa and Asia (up to +5.4 and +2.6 top-1 gains for RN50, resp.). Such differences may occur as country-specific context can align representations closer to these regions, while default prompts do not adequately capture objects around the world (esp. from non-Western regions). In Americas/Europe, adding country names leads to gains with RN50, but slight drops with ViT-B/16 and ViT-B/32. We reason that CLIP's default prompts may be already well-aligned to countries in these regions for those encoders due to overrepresentation in training.

Prompting with country-specific descriptive knowledge from LLMs outperforms general object knowledge. We observe this from CountryLLM's larger gains over default CLIP than GeneralLLM's for almost all encoders, regions, and metrics. The largest top-1 difference is with ViT-B/32 (in Total, 52.6% for CountryLLM vs. 51.4% for GeneralLLM). In top-3 accuracy, the differences for CountryLLM/GeneralLLM in Total are 74.6/73.0 for ViT-B/32, 78.8/77.9 for ViT-B/16, 70.0/68.6 for RN50. These suggest that default non-country-specific knowledge is less adequate for various countries. The gains of CountryLLM vs. Zero-Shot CLIP are generally largest on Africa and Asia, as countries in these regions may have greater shifts vs. the default prompts, but CountryLLM also performs well on Europe. LLM description in general is less effective in Americas, though Americas has a large proportion of USA images, for which default CLIP may be well-aligned.

There are complementary effects when using CLIP's internal and external LLM geo knowledge. This observation is supported by CountryInPrompt+LLM, the combina-

Encoder	Prompting Method	*Top-1 Accuracy*										*Top-3 Accuracy*									
		Europe		**Africa**		**Asia**		**Americas**		**Total**		**Europe**		**Africa**		**Asia**		**Americas**		**Total**	
		Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ
ViT-B/32	Zero-Shot CLIP [36]	59.1	-	43.7	-	50.8	-	**55.3**	-	51.7	-	81.1	-	64.8	-	72.3	-	**77.4**	-	73.7	-
	GeneralLLM [30]	57.3	-1.8	44.3	+0.6	50.9	+0.1	54.6	-0.7	51.4	-0.3	78.8	-2.3	64.5	-0.3	72.1	-0.2	75.7	-1.7	73.0	-0.7
	CountryInPrompt	57.5	-1.6	45.2	+1.5	51.9	+1.1	55.0	-0.3	52.1	+0.4	80.2	-0.9	65.5	+0.7	73.3	+1.0	76.9	-0.5	73.9	+0.2
	CountryLLM	59.4	+0.3	45.2	+1.5	52.1	+1.3	**55.3**	0.0	52.6	+0.9	80.9	-0.2	66.4	+1.6	**73.6**	+1.3	**77.4**	0.0	74.6	+0.9
	CountryInPrompt+LLM	**60.8**	+1.7	**45.3**	+1.6	**52.2**	+1.4	55.0	-0.3	**52.8**	+1.1	**81.5**	+0.4	**67.4**	+2.6	**73.6**	+1.3	76.7	-0.7	**74.7**	+1.0
ViT-B/16	Zero-Shot CLIP [36]	64.3	-	46.9	-	53.9	-	**60.1**	-	55.5	-	84.3	-	69.3	-	75.9	-	81.1	-	77.2	-
	GeneralLLM [30]	64.2	-0.1	48.8	+1.9	**56.0**	+2.1	58.5	-1.6	56.8	+1.3	83.9	-0.4	71.1	+1.8	76.3	+0.4	80.4	-0.7	77.9	+0.7
	CountryInPrompt	63.9	-0.4	49.6	+2.7	55.7	+1.8	59.3	-0.8	56.6	+1.1	84.0	-0.3	71.3	+2.0	76.5	+0.6	80.0	-1.1	77.7	+0.5
	CountryLLM	65.2	+0.9	49.6	+2.7	55.6	+1.7	59.7	-0.4	57.0	+1.5	84.3	0.0	71.8	+2.5	**77.5**	+1.6	**81.5**	+0.4	**78.8**	+1.6
	CountryInPrompt+LLM	**65.5**	+1.2	**50.8**	+3.9	**56.0**	+2.1	59.7	-0.4	**57.4**	+1.9	**85.5**	+1.2	**72.5**	+3.2	77.0	+1.1	80.9	-0.2	78.7	+1.5
RN50	Zero-Shot CLIP [36]	53.0	-	38.0	-	44.4	-	49.8	-	45.7	-	76.5	-	60.2	-	66.4	-	**72.7**	-	68.1	-
	GeneralLLM [30]	55.5	+2.5	40.9	+2.9	46.9	+2.5	50.3	+0.5	47.9	+2.2	76.0	-0.5	61.2	+1.0	67.7	+1.3	71.1	-1.6	68.6	+0.5
	CountryInPrompt	54.5	+1.5	**43.4**	+5.4	47.0	+2.6	50.8	+1.0	48.4	+2.7	76.0	-0.5	**64.0**	+3.8	68.7	+2.3	**72.7**	0.0	**70.0**	+1.9
	CountryLLM	56.2	+3.2	41.1	+3.1	47.3	+2.9	50.4	+0.6	48.3	+2.6	**77.2**	+0.7	62.5	+2.3	**68.8**	+2.4	72.4	-0.3	**70.0**	+1.9
	CountryInPrompt+LLM	**56.4**	+3.4	43.0	+5.0	**48.0**	+3.6	**50.9**	+1.1	**49.1**	+3.4	76.7	+0.2	63.1	+2.9	68.3	+1.9	71.1	-1.6	69.4	+1.3

Table 1. **Zero-shot CLIP inference with descriptive knowledge prompts, top-1/3 balanced accuracy (Acc) on DollarStreet.** Strategies to capture CLIP's internal country knowledge (CountryInPrompt), external LLM country knowledge (CountryLLM), and their combination (CountryInPrompt+LLM), often improve vs. the zero-shot CLIP baseline (prompt "a photo of a/an <object>"), especially on Africa and Asia; gains in green, drops in red. CountryLLM notably outperforms the GeneralLLM [30] baseline.

Encoder	Prompting Method	*Source*		*Target*							
		Europe		**Africa**		**Asia**		**Americas**		**Total**	
		Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ
ViT-B/16	CoOp [52]	72.2	-	53.9	-	61.5	-	68.6	-	61.7	-
	CoCoOp [51]	73.2	-	54.3	-	61.2	-	68.3	-	61.4	-
	KgCoOp [47]	73.1	-	54.4	-	62.6	-	68.7	-	62.4	-
	CountryInPrompt Reg	71.8	*-1.4*	56.8	+2.4	63.0	+0.4	69.8	+1.1	63.5	+1.1
	CountryLLM Reg	73.2	*0.0*	55.6	+1.2	63.0	+0.4	70.0	+1.3	63.2	+0.8
	CountryInPrompt+LLM Reg	**73.6**	*+0.4*	**57.2**	+2.8	**63.8**	+1.2	**70.3**	+1.6	**64.0**	+1.6
RN50	CoOp [52]	64.6	-	45.2	-	51.6	-	59.5	-	52.2	-
	CoCoOp [51]	62.9	-	44.5	-	51.0	-	58.3	-	51.4	-
	KgCoOp [47]	63.5	-	46.3	-	53.9	-	**60.5**	-	53.9	-
	CountryInPrompt Reg	63.5	*-1.1*	48.0	+1.7	53.9	0.0	60.3	-0.2	54.3	+0.4
	CountryLLM Reg	64.5	*-0.1*	47.4	+1.1	54.2	+0.3	59.9	-0.6	54.3	+0.4
	CountryInPrompt+LLM Reg	**65.5**	*+0.9*	**48.1**	+1.8	**54.5**	+0.6	60.4	-0.1	**54.8**	+0.9

Table 2. **Regularizing soft prompts with geographical knowledge, top-1 bal. acc. on DollarStreet.** We emphasize that our regularization aims to improve **target** performance, rather than source (gray, *italicized*). Gains/drops are shown vs. the *best* of soft prompt baselines (shaded). CountryInPrompt+LLM Reg achieves notable gains in target, especially on Africa. Methods use 16 shots per class.

tion of CountryInPrompt and CountryLLM, achieving the best Total top-1 performance for every encoder. The Total gains vs. default CLIP are as large as +3.4 (RN50). While CLIP has internal knowledge of country-specific categories, it may be incomplete and imprecise due to limited representation in the image-text training corpus. Adding LLM knowledge, trained on a purely textual corpus, may address some gaps. CountryInPrompt+LLM is notably the top setting in 3/4 regions for each encoder in top-1 accuracy.

5.2. Soft Prompting

We next evaluate geography knowledge regularization (Sec. 3.2), our method *to improve target performance* by ensuring that soft prompts do not overfit class text representations to a source dataset with limited geographical representativeness (*e.g.* only data from Europe). We compare regularization with ensembles of CountryInPrompt, CountryLLM, and CountryInPrompt+LLM prompts vs. state-of-the-art soft prompting methods in Tables 2/3.

Regularizing soft prompts with target geographical knowledge reduces overfitting to source geographies. Our method effectively improves the ability of CLIP, *with prompts trained only on images from Europe*, to generalize to target countries. This observation is supported by Total Target gains for CountryInPrompt, CountryLLM, and CountryInPrompt+LLM Reg on DollarStreet (+1.1/0.8/1.6 over the *best soft prompt baseline* for ViT-B/16). Improvements are notable in Africa: CountryInPrompt+LLM achieves +2.8 for ViT-B/16 and +1.8 for RN50. The effectiveness extends to GeoNet in Table 3: target gains are +1.3

Encoder	Method	*Source* USA Acc	Δ	*Target* Asia Acc	Δ
ViT-B/16	CoOp [52]	**58.7**	-	51.2	-
	CoCoOp [51]	57.7	-	52.6	-
	KgCoOp [47]	58.2	-	52.6	-
	CIP Reg	57.5	*-1.2*	53.5	+0.9
	LLM Reg	58.5	*-0.2*	53.1	+0.5
	CIP+LLM Reg	57.6	*-1.1*	**53.9**	+1.3
RN50	CoOp [52]	51.4	-	45.6	-
	CoCoOp [51]	51.1	-	46.3	-
	KgCoOp [47]	**51.8**	-	46.9	-
	CIPReg	50.6	*-1.2*	47.6	+0.7
	LLMReg	**51.8**	*0.0*	47.4	+0.5
	CIP+LLMReg	51.1	*-0.7*	**48.3**	+1.4

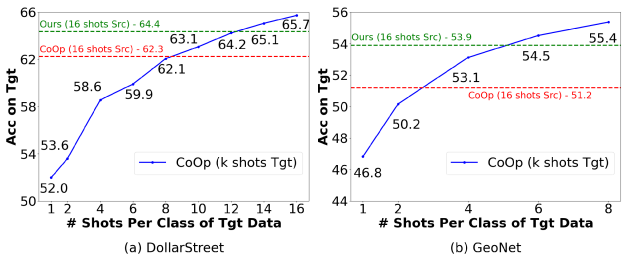
Table 3. **Regularizing soft prompts with geographical knowledge, top-1 bal. accuracy on GeoNet.** The regularization method accomplishes our goal to increase *target* performance in GeoNet's USA-to-Asia transfer setting. CIP = CountryInPrompt, LLM = CountryLLM, CIP+LLM = CountryInPrompt+LLM.

Method	Threshold t (# Classes) <40% (13)	Δ	<60% (45)	Δ	<80% (77)	Δ	≤100% (95)	Δ
CoOp [52]	31.2	-	45.6	-	55.6	-	61.7	-
CoCoOp [51]	32.8	+1.6	45.4	-0.2	55.2	-0.4	61.4	-0.3
KgCoOp [47]	35.3	+4.1	47.9	+2.3	56.7	+1.1	62.4	+0.7
CIPReg	37.5	+6.3	48.9	+3.3	57.8	+2.2	63.5	+1.8
LLMReg	36.8	+5.6	48.1	+2.5	57.2	+1.6	63.2	+1.5
CIP+LLMReg	**39.9**	+8.7	**49.5**	+3.9	**58.2**	+2.6	**64.0**	+2.3

Table 4. **Performance on DollarStreet classes with less than t% recall in CoOp**, with ViT-B/16. Gains w.r.t. CoOp of our geography knowledge regularization are especially large for CoOp's difficult classes (+8.7 in <40%), compared to KgCoOp's (+4.1 in <40%, i.e. a **4.6** difference from ours). CIP = CountryInPrompt, LLM = CountryLLM, CIP+LLM = CountryInPrompt+LLM.

for ViT-B/16 and +1.4 for RN50. The combined strategy works best on target, showing the value of incorporating descriptive knowledge. Since regularization prevents overfitting and potentially optimal source performance, we naturally observe *source* drops for CountryInPrompt and CountryLLM in Tables 2/3. However, CountryInPrompt+LLM in Table 2 even offers source gains. It is also notable that there are small drops in Americas for RN50, but upon inspection, countries in North America overall have a -0.9 drop, while ones in Central/South America have a +0.8 gain. These results concur with our hypothesis that CLIP is already aligned to countries like the USA. More is in supp., along with experiments varying the source and ensemble.

Regularization helps significantly on difficult classes. As certain objects may be especially sensitive to geographical domain shift, we break down classwise performance on DollarStreet in Table 4, using a stratification of class difficulty based on default soft prompting performance (CoOp). *The CountryInPrompt+LLM strategy achieves significant gains on the classes most difficult with respect to the CoOp baseline*. In particular, gains of +8.7% in balanced accuracy are achieved for classes with <40% baseline recall, while the highest achieved by KgCoOp is 4.1%. Example classes in this subset are *snacks*, *clothes*, and *makeup*. The DollarStreet classes with greatest improvement, independent of original CoOp accuracy are: *piercings*, *clothes*, *homes*, *medication*, and *refrigerators* (all at least +14% over CoOp). In GeoNet, *dome*, *goby* (fish), *eland* (antelope), and *gloriosa* (flower) have >20 samples and >30% improvement. A total of 64/95 classes in DollarStreet and 209/344 in GeoNet improve vs. CoOp, showing broad coverage.



(a) DollarStreet (b) GeoNet

Figure 3. **Geography knowledge-regularized soft prompts trained on source data (ours, green line) vs. few-shot soft prompts trained on target data (blue curve).** (a) Src=Europe, Tgt=Africa,Asia,Amer.; (b) Src=USA,Tgt=Asia. Our 16-shot model trained on only source data (green) outperforms a model with prompts trained on 12 or 4 shots *per class* of target data (on DollarStreet&GeoNet, resp.), which is 1140&2400 images total.

Regularized source-only prompts outperform few-shot target-trained prompts. Given that soft prompts can show strong performance in few-shot settings, a potential alternative to regularizing soft prompts on source data is to directly acquire a few examples of target data for training. We evaluate this setting by splitting target data into train/test, and training CoOp at varying # of shots of *target* data for GeoNet and DollarStreet, shown in Fig. 3. Notably, training on 16 shots per class of *source* data with *our regularization method* outperforms using 12 & 4 shots per class of *target* data on DollarStreet & GeoNet. This is a vast amount of target data overall (*e.g.* 12 shots x 95 classes = 1140 target samples in DollarStreet, 4 shots x 600 classes = 2400 samples in GeoNet). The baseline CoOp trained on 16 shots of source data only outperforms an 8-shot/2-shot target-trained CoOp model (DollarStreet/GeoNet). Our strategy is thus more compelling in the absence of a lot of target data.

Performance by income. DollarStreet provides estimated monthly income of the household in which an image was captured. We evaluate with the delineation of low, medium, and high-income buckets from [13]. Compared to CoOp/KgCoOp, CountryInPrompt+LLM gains are +2.5/+3.4 in low, +2.4/+1.5 in medium, and +2.1/+0.7 in high. Thus our method especially improves in low-income areas, though it helps across levels. The table is in supp.

Statistic	CIP	CountryLLM	CIP+LLM
GDP Per Capita (US $)	**0.219**	0.063	**0.217**
Human Devel. Index (HDI)	**0.439**	**0.385**	**0.451**
Land Area (km^2)	-0.072	0.050*	-0.046*
Population (#)	-0.131	0.077	-0.123
Population Density (#/km^2)	0.103	0.158	0.081
% Agricultural Land	0.139	0.070	0.122
% Forest Area	0.191	0.087	**0.201**
Avg. Yearly Temp. ($^\circ$C)	**0.380**	**0.256**	**0.391**
Avg. Yearly Precip. (mm/year)	**0.236**	0.124	**0.230**

Table 5. **Correlation (Pearson's ρ) of avg. CLIP class text embedding distance and country statistic difference**, *e.g.* economic (GDP per capita, HDI) and climate factors (temperature, precipitation, forest area). We use the 63 countries in DollarStreet (1,953 pairs). **Bold** values have ρ>0.2, * means not significant (α=0.01).

Figure 4. **Qualitative analysis.** We show examples where geography-specific descriptors improve/hurt vs. general descriptors in zero-shot inference. We highlight the prediction's descriptors, bolding the highest activating one. Encoder=RN50.

5.3. Further Analysis

Are descriptions correlated with key country statistics? For a pair of countries, we compute two values. We measure the distance between each class embedding and take the average overall distance as one value. We also take the absolute difference between statistics for those countries (from [2, 42], *e.g.* difference in avg. yearly temperature) as the other value. We compute the correlation between these two values over every unique country pair in DollarStreet, showing results in Table 5. We find that the strongest correlation across each prompt type is with HDI, which summarizes human development. It is notable that factors like yearly temperature and precipitation also show moderate correlations, indicating a potential role of climate. Future work may further explore how object differences present with respect to these factors. It will also be critical to ensure that differences between countries are representative and not exaggerated in embeddings.

Descriptor topics. We show a UMAP [29] visualization comparing CountryLLM text embeddings for the category *homes* across geographies in Fig. 5. Countries tend to group by continent, showing the representations may capture similarity in features like climate and/or economics. We examine a few topics mentioned in the CountryLLM descriptors for *homes*. While "stone" is described across continents, "bright colors" and "mud" are mentioned mostly in Africa, and "balcony" in Europe and Asia. We show more in supp.

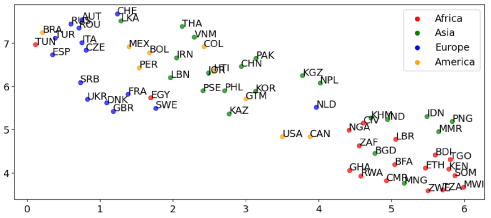


Figure 5. **UMAP [29] plot for CountryLLM and the category *homes* in DollarStreet.** Country-specific descriptors are often close to those of other countries intra-continent, likely due to similar weather, environment, and/or economic conditions.

Success and failure examples. We provide examples of CountryLLM vs. GeneralLLM in Fig. 4 on DollarStreet. The model captures geographical descriptive knowledge like "sandy colored stucco walls" for *homes* in Jordan, a feature which may be less common for Western homes. Sometimes the model may be too attentive to attributes, leading to confusion (*e.g.* choosing *rug* over *bed*). Future work that enhances alignment in VLMs can likely improve results.

6. Conclusion

In this work, we bring attention to how various strategies to prompt CLIP affect recognition performance across geographies. In addition, through soft prompting with descriptive knowledge, we provide a mechanism to achieve a more geo-generalizable set of class representations across regions. Our work is only a first step in this important area.

Limitations and ethical considerations. While our method's proof of concept is demonstrated in a positive effort to debias CLIP's default representations through diversity, due to the biased worldview of the Internet, CLIP's representations are likely inadequate, exaggerated, and/or not fully representative for some countries. While we expect quality LLM knowledge to guide better representations, LLM knowledge can also be incorrect (e.g., through hallucination), imprecise, or biased.

Future work. For the above reasons, our future efforts aim to ensure more representative VLM/LLM knowledge. We strongly advocate for the community to seek communication with diverse groups within all countries (i.e. to capture areas that range from low to high income) to ensure better representation and fairness in AI technology use. There are notable continent-level disparities to still improve upon.

Acknowledgement. This work was supported by National Science Foundation Grants No. 2006885 and 2329992.

References

[1] Common Crawl, 2023. https://commoncrawl.org/. 3

[2] Wikipedia, 2023. https://en.wikipedia.org/. 3, 8

[3] Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5136–5147, 2023. 3

[4] 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. 2

[5] Pei-Yu Peggy Chi, Matthew Long, Akshay Gaur, Abhimanyu Deora, Anurag Batra, and Daphne Luong. Crowdsourcing images for global diversity. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–10, 2019. 2

[6] Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. ChatGPT goes to law school. *Available at SSRN*, 2023. 2

[7] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021. 5

[9] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14340–14349, 2021. 2

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015. 2

[11] William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The Dollar Street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Advances in Neural Information Processing Systems*, pages 12979–12990, 2022. 1, 2, 5

[12] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022. 2

[13] Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 70–88, 2022. 7

[14] 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. 5

[15] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022. 2

[16] Plotly Technologies Inc. Plotly: The python graphing library, 2023. https://plotly.com/python/. 1

[17] 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

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2

[19] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision (ECCV)*, pages 464–480. Springer, 2020. 2

[20] Tarun Kalluri, Astuti Sharma, and Manmohan Chandraker. Memsac: Memory augmented sample consistency for large scale domain adaptation. In *European Conference on Computer Vision*, pages 550–568. Springer, 2022. 2

[21] Tarun Kalluri, Wangdong Xu, and Manmohan Chandraker. Geonet: Benchmarking unsupervised adaptation across geographies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15368–15379, 2023. 1, 2, 5

[22] Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15670–15680, 2023. 2

[23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2

[24] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *International Conference on Learning Representations (ICLR)*, 2022. 4

[25] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021. 2

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, 2023. 2

[27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation net-

works. In *International Conference on Machine Learning*, pages 97–105. PMLR, 2015. 2

[28] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[29] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018. 8

[30] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *International Conference on Learning Representations, ICLR*, 2023. 2, 3, 4, 5, 6

[31] OpenAI. ChatGPT: Optimizing language models for dialogue, 2022. 2

[32] OpenAI. GPT-4 technical report, 2023. 2

[33] Viraj Prabhu, Ramprasaath R Selvaraju, Judy Hoffman, and Nikhil Naik. Can domain adaptation make object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3981–3988, 2022. 1, 2

[34] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? Generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 2, 3

[35] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[36] 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. 2, 3, 6

[37] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. GeoDE: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 2023. 2

[38] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2

[39] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NeurIPS 2017 Workshop: Machine Learning for the Developing World*, 2017. 1

[40] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 2

[41] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. LLM-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[42] The World Bank. World Bank - Indicators, 2023. https://data.worldbank.org/indicator/. 8

[43] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in Germany, test in the USA: Making 3D object detectors generalize. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11713–11723, 2020. 1, 2

[44] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zhibo Chen. Toalign: Task-oriented alignment for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:13834–13846, 2021. 2

[45] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1426–1435, 2019. 2

[46] Zhengyuan Yang*, Linjie Li*, Jianfeng Wang*, Kevin Lin*, Ehsan Azarnasab*, Faisal Ahmed*, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-ReAct: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2

[47] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 2, 5, 6, 7

[48] Da Yin, Feng Gao, Govind Thattai, Michael Johnston, and Kai-Wei Chang. Givl: Improving geographical inclusivity of vision-language models with pre-training methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10961, 2023. 2

[49] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 2

[50] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019. 2

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16816–16825, 2022. 2, 5, 6, 7

[52] 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. 2, 4, 5, 6, 7