Context-Aware Multimodal Pretraining

Karsten Roth[1,2,*] Zeynep Akata[2,3] Dima Damen[4] Ivana Balažević[4,†] Olivier J. Hénaff[4,†]

[1]Tübingen AI Center [2]Munich Center for ML [3]Helmholtz Munich, TU Munich [4]Google DeepMind

[*]Work done while author was at Google DeepMind. [†]Equal senior contribution.

Abstract

Large-scale multimodal representation learning successfully optimizes for zero-shot transfer at test time. Yet the standard pretraining paradigm (contrastive learning on large amounts of image-text data) does not explicitly encourage representations to support few-shot adaptation. In this work, we propose a simple, but carefully designed extension to multimodal pretraining which enables representations to accommodate additional context. Using this objective, we show that vision-language models can be trained to exhibit ***significantly increased*** *few-shot adaptation: across 21 downstream tasks, we find up to four-fold improvements in test-time sample efficiency, and average few-shot adaptation gains of over 5%,* ***while retaining*** *zero-shot generalization performance across model scales and training durations. In particular, equipped with simple, training-free, metric-based adaptation mechanisms, our representations easily surpass more complex and expensive optimization-based schemes, vastly simplifying generalization to new domains.*

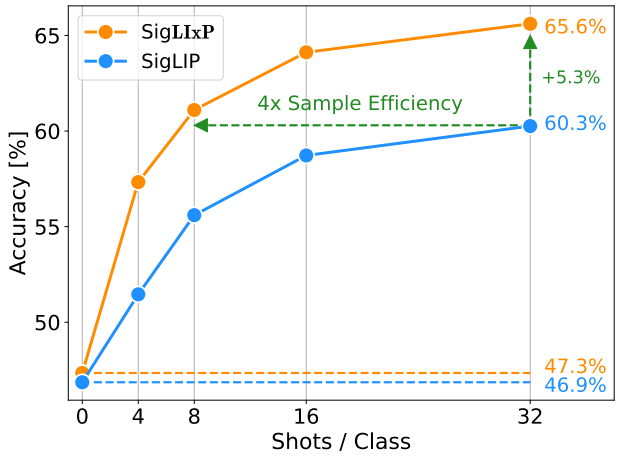

Figure 1. **Context-aware multimodal pretraining facilitates few-shot transfer.** Applying Tip-Adapter [101] on a ViT-S/16 pretrained with and without our contextualized pretraining objective (here modifying SigLIP [99]) showcases increases in test-time sample efficiency and overall few-shot performance while maintaining the underlying zero-shot transfer performance.

1. Introduction

Contrastive language-image pretraining [33, 65, 99]—where dual vision and text encoders must align both modalities while ensuring unrelated embeddings remain dissimilar—has yielded a new class of models with remarkable zero-shot transfer capabilities [50, 65, 69, 84, 99]. Yet even large-scale pretraining can lack out-of-distribution generalization if downstream distributions diverge from pretraining data [18, 26, 73, 84, 85, 90]. As a result, additional data at test time is often essential to adapt to more severe visual distribution shifts [24, 40, 68, 103], as well as more fine-grained context [41, 69, 70, 84, 91]. Such post-hoc adaptation can involve all manners of post-training optimization—including model finetuning [26, 70, 77], prompt tuning [81, 105] or training of adapters [59, 90]—though generally more costly, and often prone to instability and overfitting if the number of available adaptation samples is low [10, 17, 26, 84]. At the same time, representation learning for few-shot learning on smaller models [7, 48, 82] has shown that more involved optimization-based objectives at test time and meta-training objectives during model training [44, 55, 96] are often matched or outperformed [7, 48, 82] by simple, *training-free* metric- or distance-based methods (such as prototypical or nearest-neighbor classifiers [22, 24, 54, 76, 87]) operating on robust visual representations.

Training-free, metric-based model adaptation of large-scale vision representation models has seen widespread adoption across few- and many-shot classification, segmentation or retrieval tasks [1, 22, 24, 68, 84, 101, 105], as it exhibits both rapid and easy scalability over both small and large numbers of test-time examples [1, 22, 63], high computational efficiency e.g. via approximate nearest neighbor search [34, 49], and a great degree of flexibility in incorporating application constraints [22]. However, large-scale image-text pretraining does not explicitly account for this form of training-free model re-use at test time, instead assuming that models optimized for

zero-shot generalization will inherently transfer well to few-shot scenarios [7, 48, 82, 101]. In this work we question this assumption, and showcase that multimodal models can be trained to be significantly more amenable to training-free few-shot adaptation, with no reduction in their zero-shot transfer capabilities. More precisely, this work aims to answer the question: *"Can large-scale contrastive vision-language pretraining **better support** downstream few- and many-shot transfer **without impacting** zero-shot generalization performance?"*

To achieve this, we introduce a simple, yet carefully designed *context-aware* extension to contrastive language-image pretraining, **LIxP** (for **L**anguage-**I**mage Conte**x**tual **P**retraining). LIxP augments standard language-image contrastive objectives with cross-attention-based contextualization during training, preparing its representations for metric-based adaptation. Moreover, we carefully consider its inclusion into the image-text contrastive training process in order to maintain base zero-shot capabilities, leveraging particular choices in overall loss design and the use of individually learnable temperatures. In doing so, we show that across 21 few- and many-shot downstream classification tasks LIxP objectives enable up to **four-fold** sample-efficiency gains and over **5%** average performance improvements, *while retaining* the original zero-shot transfer performance. In doing so, LIxP closes the gap between training-free and optimization-based adaptation methods, dramatically simplifying generalization to new domains.

2. Related Works

Contrastive image-text pretraining has become the *de facto* standard when training large-scale general visual representation models. Developed initially in CLIP [65] and ALIGN [33] using an InfoNCE-style training objective [57] (stemming from augmentation-based self-supervised learning such as in [5]), several works have since augmented these in order to improve their zero-shot transfer capabilities [23, 52, 62, 80]. To enable more efficient language-image pretraining at scale, SigLIP leverages only pairwise sigmoidal losses for pretraining while achieving comparable or improved transfer performance [99]. SigLIP and correspondingly pretrained vision-language models have seen notable adoption [15, 16]. Other approaches [14, 28, 30, 95] show how external support data or training memory can be utilized to explicitly encourage increased zero-shot generalization in InfoNCE-style training. In our work, we showcase how a particular buffer and objective design choices can instead encouraging the emergence of improved few-shot capabilities while maintaining zero-shot performance.

Meta- and few-shot learning study pretraining and method design choices to facilitate adaptation to often labeled "shots" of new data at test time, in order to both rapidly and effectively account for changes in downstream data and label distributions. Approaches are generally separated into *optimization-based* variants, which operate under the premise of full or partial model optimization over new data at test time [20, 44, 55, 67, 72, 83, 96, 110], and *metric-based* methods, which aim to meta-learn representation spaces equipped with a pre-defined metric [76, 79, 87, 98, 100] to then be re-used at test time, often without any additional training. While approaches are numerous, [7, 45, 82, 89] highlight that on smaller scale, many complex meta-learning approaches are often matched or outperformed by simple metric- or regression based approaches (like prototypical or nearest neighbor classification) operating on top of a general representation backbone [48]. Nevertheless, at larger scale and with increased test-time compute budgets, optimization-based approaches generally tend to produce highest performances [48, 92, 101, 102, 108]. In this work, we ask whether well-chosen modifications to the large-scale pretraining paradigm can reduce or even close the gap between simple, cost-effective training-free and sophisticated optimization-based methods - while maintaining general representations capable of zero-shot transfer.

Post-hoc adaptation of large image and text representation models such as CLIP or SigLIP has seen extensive adoption, incorporating different modifications to enhance base zero-shot transferability through architectural changes and language model extensions [25, 50, 56, 64, 69, 103, 107], as well as test-time training through finetuning of the entire model [70, 77, 78] or prompt tuning [37, 105, 106], as well as inclusion of a learnable adapter [59, 92, 101], new components [12, 19, 32] or generative models [38] to account for new data available at test time. More recently, a large number of works have studied the ability to conduct such adaptation efficiently and *training-free*, such as TiP-Adapter [101], SuS-X [84] (extending TiP-Adapter with support set retrieval), Visual Memory [22] utilizing large-scale visual retrieval at test time, Dual Memory [102] retrieving from a dual feature cache, Ming and Li [51], Zhu et al. [108] using prior refinement and retrieval augmentation against support and training data, alongside other similar extensions [27, 46, 90, 109]. These methods rely on metric-based training-free classification mechanisms such as nearest neighbor retrieval based on the existence of a well-defined similarity metric to effectively adapt to new examples. In this work, we show how our context-aware LIxP can modify fundamental vision-language pretraining to facilitate such general, training-free post-hoc adaptation.

3. Method

We first describe the basic contrastive language-image pretraining setup and different post-hoc, metric-based training-free adaptation mechanisms, before introducing our context-aware pretraining extension LIxP.

3.1. Background

3.1.1. Contrastive Image-Text Pretraining

For a large pretraining dataset $\mathcal{D}$ containing image-text pairs $(I_i, T_i) \in \mathcal{D}$, the softmax-training objective $\mathcal{L}_{\text{CLIP}}$ for a dual image ϕ_I [13] and text ϕ_T [86] encoder is defined as

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{\mathcal{B}}\left(\log\frac{e^{\tau_1 x_i t_i}}{\sum_{j=1}^{|\mathcal{B}|} e^{\tau_1 x_i t_j}} + \log\frac{e^{\tau_1 x_i t_i}}{\sum_{j=1}^{|\mathcal{B}|} e^{\tau_1 x_j t_i}}\right) \quad (1)$$

given a batch $\mathcal{B}$ of image-text pairs and normalized image representations $x_i = \hat{x}_i/\|\hat{x}_i\|$ with $\hat{x}_i = \phi_I(I_i)$ and textual counterparts $t_i = \phi_T(T_i)/\|\phi_T(T_i)\|$. We use $\mathbf{X}_{\mathcal{B}} = \phi_I(\mathcal{B}_I)$ to denote normalized image representations over the full batch of images $\mathcal{B}_I$, and $\mathbf{T}_{\mathcal{B}} = \phi_T(\mathcal{B}_T)$ for corresponding textual batch representations. We follow [99] and parameterize $\tau_1 = \exp(\tau_1')$ with learnable temperature τ_1'. Equation 1 uses normalization over the entire input batch $\mathcal{B}$, which significantly impacts scalability [99]. Instead, Zhai et al. [99] propose a pairwise sigmoid objective, which treats each image-text pair within a batch independently

$$\mathcal{L}_{\text{SigLIP}} = -\frac{1}{|\mathcal{B}|}\sum_{i,j=1}^{|\mathcal{B}|}\log\frac{1}{1+e^{\mathbb{I}_{i=j}(-(\tau_1 x_i t_j + b_1))}}, \quad (2)$$

with an indicator $\mathbb{I}_{i=j} = 1$ for $i = j$, and -1 otherwise, and an additional learnable bias initialized at $b = -10$ [99].

3.1.2. Metric-based Few-Shot Image Classification

We consider a few- or many-shot classification task with N target classes and a *support set* $\mathcal{I}_{\text{spt}}$ of K image examples per class. Setting the normalized embedded image support set as $\mathbf{X}_{\text{spt}} \in \mathbb{R}^{N\cdot K\times d}$ (with embedding dimensionality d), their corresponding labels $\mathbf{L}_{\text{spt}}$ and a similarity metric $s(\cdot,\cdot)$, one can define different metric-based approaches for few-shot image classification. As contrastive image-text pretraining generally operates on normalized representations, we simply set $s(\cdot,\cdot)$ to the cosine similarity.

Prototypical classifier [76]. We define a prototype representation for each class $c \in \{1, ..., N\}$ as

$$\mu_c = \frac{1}{|I^c_{\text{spt}}|}\sum_{i\in I^c_{\text{spt}}}\mathbf{X}_{\text{spt},i}, \quad (3)$$

where I^c_{spt} *indexes* the images in the support set belonging to class c. Given these prototypes for each class $\mathbf{C} = [\mu_1, ..., \mu_N]$, the classification logits for the test image representation x_{test} are computed as $x_{\text{test}}\mathbf{C}^T$.

Tip-Adapter. Given text representations for all classes $\mathbf{T} \in \mathbb{R}^{N\times d}$, the Tip-Adapter logits are computed via

$$\text{logits} = x_{\text{test}}\mathbf{T}^T + \alpha\exp\left(-\beta(1 - x_{\text{test}}\mathbf{X}^T_{\text{spt}})\right)\mathbf{L}_{\text{spt}} \quad (4)$$

with weighting and modulation hyperparameters α and β. While the first term computes the standard zero-shot classification logits over textual representations, the second term reweighs those against support-set image features. By default we use original [101] hyper-parameter values $\alpha = 1.0$ and $\beta = 5.5$. We also include a $3-$fold cross-validation on $\mathbf{X}_{\text{spt}}$ to select the most favorable α, β combination, refered to as Cross-Validated Tip-Adapter or CV-Tip. In initial tests, we find CV-Tip to often outperform even recent training-free clip-adapter methods such as APE [108] or DMN-TF [102] across most datasets.

Nearest-neighbor voting classifier. Given the full support set $\mathbf{X}_{\text{spt}}$ and a test image representation x_{test}, each support example votes for its class c_i with a weight w_i which is proportionate to its similarity with x_{test}. Plurality-, softmax-, and rank-based voting [4, 22, 54] each have their own way of deriving the weights w_i, see Appendix B. Given the vector of weights w, the logits are then computed as $w\mathbf{L}_{\text{spt}}$.

3.2. Context-Aware Vision-Language Pretraining

All methods listed in the previous section rely on metric-based aggregation and voting over a support set of normalized representations. However such re-use of supports sets for downstream metric-based classification is not explicitly accounted for in the standard contrastive pretraining setup.

Representation contextualization. To tackle this, we introduce key and value context buffers $\mathcal{M}_K$ and $\mathcal{M}_V$, which provide *a proxy for test-time context during pretraining*. For a normalized training image representation x_i as defined in Sec. 3.1.1 we define its contextualized counterpart x^{ctx}_i by cross-attending over the contextualization buffer

$$x^{\text{ctx}}_i = \sigma\left(\frac{x_i\cdot\mathcal{M}_K^T}{\tau_{\text{ctx}}\sqrt{d}}\right)\mathcal{M}_V, \quad (5)$$

where σ denotes the softmax operation across query-key similarities. We use $\mathbf{X}^{\text{ctx}}_{\mathcal{B}}$ to denote the contextualized batch representations. In all our experiments, we find it important to define $\tau_{\text{ctx}} = \exp(\tau'_{\text{ctx}})$ with learnable temperature τ'_{ctx}.

Contextualization while preserving zero-shot generalization. While $\mathbf{X}^{\text{ctx}}_{\mathcal{B}}$ can be directly plugged into the contrastive loss, we find downstream few-shot adaptation gains to be limited and zero-shot transfer capabilities reduced. This can be attributed to the external memory minimizing the need to learn robustly transferable image-text representations. Thus, we explicitly separate context re-use from the representation learning objective, yielding our context-aware pretraining objective with weighting α:

$$\mathcal{L}_{\textbf{LIxP}} = \alpha\mathcal{L}_{\text{LIP}}(\mathbf{X}_{\mathcal{B}}, \mathbf{T}_{\mathcal{B}}, \tau_1) + (1-\alpha)\mathcal{L}_{\text{LIP}}(\mathbf{X}^{\text{ctx}}_{\mathcal{B}}, \mathbf{T}_{\mathcal{B}}, \tau_2). \quad (6)$$

Using Eq. (6), we derive contextualized SigLIP and CLIP as SigLIxP setting $\mathcal{L}_{\text{LIP}} = \mathcal{L}_{\text{SigLIP}}$, and CLIxP via $\mathcal{L}_{\text{LIP}} =$

$\mathcal{L}_{\text{CLIP}}$. Importantly, we introduce another, separately learnable temperature $\tau_2 = \exp(\tau_2')$ to uncouple the two training objectives, which we show to be essential for optimizing for both zero- and few-shot transfer. Together, this gives three distinctly trainable temperatures: τ_1, τ_2 and τ_{ctx} (Eq 8).

Contextualization buffer design. Given our contextual mechanism and objective, the quality of the final learned representation is driven by the exact composition of the contextualization buffers $\mathcal{M}_K$ and $\mathcal{M}_V$. While the product between base image representation x_i and $\mathcal{M}_K$ mimics the nearest neighbor retrieval process at test time, $\mathcal{M}_V$ determines the gains from context usage during training. These can be set as e.g. image or text representations of the current batch, as well as representations from previous iterations. We opt for a simple image-only context, consistent with the usage of these representations at test time:

$$\mathcal{M}_K = \phi_I(\mathcal{B}_I)/\|\phi_I(\mathcal{B}_I)\|, \quad \mathcal{M}_V = \phi_I(\mathcal{B}_I) \tag{7}$$

While $\mathcal{M}_K$ contains normalized entries as expected for retrieval at test-time, $\mathcal{M}_V$ utilizes *non-normalized* embeddings to leverage the additional degree of freedom in the norm [39, 74] to provide an additional value signal leverageable during training. In Sec. 4.2, we thoroughly evaluate *alternative choices for buffer and contextualization design*, and show that this simple scalable setup outperforms alternatives. Moreover, having equivalence between the buffer and the training batch notably improves compute efficiency. Our extensive experiments across model sizes reveal that for metric-based downstream adaptation, allowing the model to jointly populate and backpropagate through the buffer with image representations is key to encourage re-use at test time. Deviations from this basic setup consistently fail to strike a suitable tradeoff between maintaining zero-shot generalization performance and improving few-shot adaptation performance.

Overall objective. Put together, our training objective is defined as the single-stage application of Eq. (5) with $\mathcal{M}_K = \mathbf{X}_\mathcal{B} = \phi_I(\mathcal{B}_I)/\|\phi_I(\mathcal{B}_I)\|$ and $\mathcal{M}_V = \hat{\mathbf{X}}_\mathcal{B} = \phi_I(\mathcal{B}_I)$, s.t. for batch $\mathcal{B}_I$ we have

$$\mathbf{X}_\mathcal{B}^{\text{ctx}} = \frac{\hat{\mathbf{X}}_\mathcal{B}^{\text{ctx}}}{\|\hat{\mathbf{X}}_\mathcal{B}^{\text{ctx}}\|}, \quad \hat{\mathbf{X}}_\mathcal{B}^{\text{ctx}} = \sigma\left(\frac{\mathbf{M} \odot \mathbf{X}_\mathcal{B}\mathbf{X}_\mathcal{B}^T}{\tau_{\text{ctx}}\sqrt{d}}\right)\hat{\mathbf{X}}_\mathcal{B} \tag{8}$$

where $\odot$ denotes element-wise multiplication. Importantly, $\mathbf{M} = \mathbf{1} - \mathbf{I}_\infty$, with $\mathbf{I}_\infty$ denoting diagonal "∞" entries, defines a ones-mask with $-\infty$-diagonals to avoid a representation attending to itself (i.e. setting softmax entries for self-attention to 0) and reverting to the standard image-text training objective. For a large enough batch-size, this effectively creates implicit per-iteration episodic training; where each non-masked batch entry constitutes support samples used to predict respective textual embeddings.

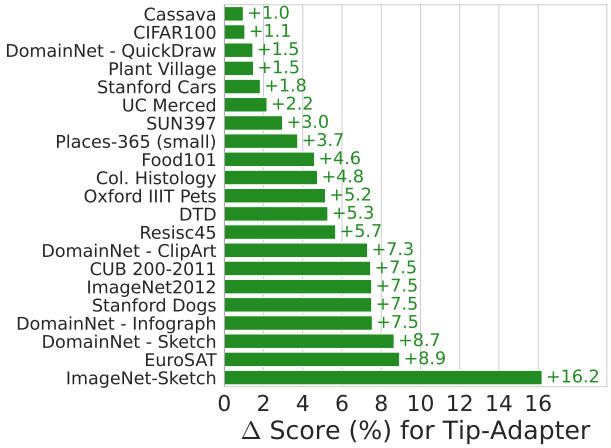

Figure 2. **Dataset-level performance breakdown (32-shot, Tip-Adapter [101])** for ViT-S/16 shows gains up to +16.2% on **all 21** benchmarks; with each dataset improving by at least +1.0%.

4. Experiments

Our pretraining pipeline follows the SigLIP [99] protocols for training ViT-(S/16, B/16, L/16) image encoders and correspondingly sized BERT-{S,B,L} text encoders. To evaluate both zero-shot as well as few-shot transfer capabilities, we measure performance on 21 diverse datasets commonly used for few-shot and domain adaptation (see Appendix A). The majority of our experiments use the SigLIP pretraining objective due to its superior computational efficiency, but we show all our conclusions hold for CLIP in Tab. 3.

4.1. Context-Aware Training

Impact on downstream few-shot adaptation. To study the impact of **LIxP** on downstream few-shot tasks, we apply SigLIxP to a ViT-S/16 model (57.2M parameters) trained over 1.5B WebLI examples and evaluate its few-shot adaptation performance with the Tip-Adapter (see Eq. (4)). Results are shown in Fig. 1, where we compare zero- to 32-shot adaptation performance for SigLIxP against an equivalently pretrained SigLIP model. Walltime changes when using Eq. (8) are negligible. While zero-shot performance remains comparable (+0.4% improvements), we find significant gains in few-shot performance (e.g. +5.3% for 32-shots) and a 4× increase in sample efficiency (SigLIxP achieving 61.1% 8-shot vs SigLIP 60.3% 32-shot). Looking at a dataset-level breakdown in Fig. 2 we find that all 21 datasets exhibit over +1% gains, with a maximum gain of +16.2% on ImageNet-Sketch. To further showcase the *robustness to the choice of the few-shot adaptation method*, we consider all six metric-based adaptation methods described in Sec. 3.1.2 (see Fig. 3). In all cases, improvements remain high, ranging from a +1.7% gain for *distance-free* rank-voting to +5.4% for Tip-Adapter. These results indicate that **LIxP** offers a general solution to facilitate training-

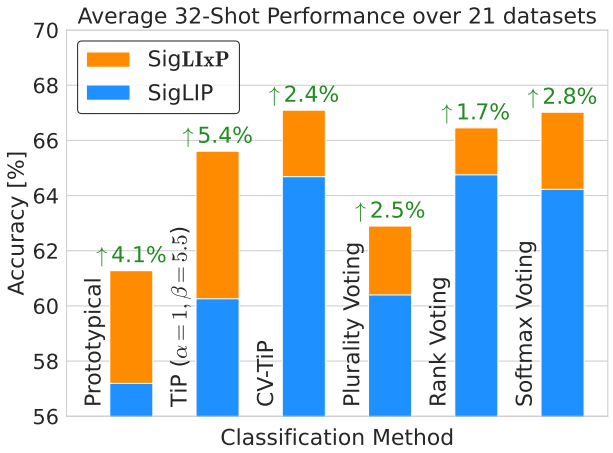

Figure 3. **Significant gains across metric-based few-shot classifiers.** Applying prototypical classification [76], Tip-Adapters [101] and nearest neighbor classifiers [22, 54] on vision-backbones using context-aware pretraining significantly boosts 32-shot results across the board (here ViT-S/16, 1.5B samples).

Model → / **Examples →**	ViT-S/16 1.5B	→ 6B	ViT-B/16 6B	→ 15B	ViT-L/16 8B
ZeroShot	46.9	52.1	60.3	62.5	64.1
	+0.4	−0.2	−0.4	−0.5	−0.1
Prototypical	57.2 ± 0.2	60.8 ± 0.3	66.8 ± 0.2	67.4 ± 0.3	70.7 ± 0.3
	+4.1	+3.4	+3.4	+3.9	+3.3
Default Tip	60.3 ± 0.1	63.6 ± 0.3	69.5 ± 0.2	70.2 ± 0.3	73.2 ± 0.3
	+5.4	+4.8	+4.3	+4.3	+4.0
XVal Tip	64.7 ± 0.2	67.8 ± 0.3	73.8 ± 0.2	74.7 ± 0.2	77.0 ± 0.3
	+2.4	+2.3	+1.6	+1.6	+1.4
Plurality NN	60.4 ± 0.1	63.8 ± 0.2	69.1 ± 0.2	69.6 ± 0.3	72.6 ± 0.2
	+2.5	+1.9	+2.3	+2.6	+1.8
Rank NN	64.8 ± 0.1	68.1 ± 0.1	73.2 ± 0.1	74.1 ± 0.2	76.5 ± 0.2
	+1.7	+1.3	+1.8	+1.8	+1.1
Softmax NN	64.2 ± 0.1	67.5 ± 0.2	72.6 ± 0.2	73.3 ± 0.3	75.9 ± 0.2
	+2.8	+2.4	+2.6	+2.7	+1.8
Average Gain	**+3.2**	**+2.7**	**+2.7**	**+2.8**	**+2.2**

Table 1. **Context-aware pretraining for SigLIP holds across model size and data scale.** Benefits of contextualized pretraining across larger architectures, longer pretraining runs, and combinations of both (here shown for 32-shot classification for three metric-based classifiers) show that our contextualized pretraining objective boosts performance across model sizes and training duration; with no or at most very marginal drops in zero-shot transfer.

free adaptation at test time, allowing practitioners to choose adaptation methods based on individual constraints.

Scaling model and data size. To ensure these insights hold across increased model and training data sizes, we further consider ViT-S/16 trained on 6B examples, ViT-B/16 on 6B and 15B examples, and ViT-L/16 on 8B examples in Tab. 1. Improvements remain high across the board, showcasing the *robustness of our approach to the size of the model and the amount of training data used*. For example, training ViT-S/16 for 4× the amount of examples improves base zero-shot performance by 5.2%, while gains from switching to context-aware pretraining remain high and stable (e.g. +2.4%→2.3% for the cross-validated

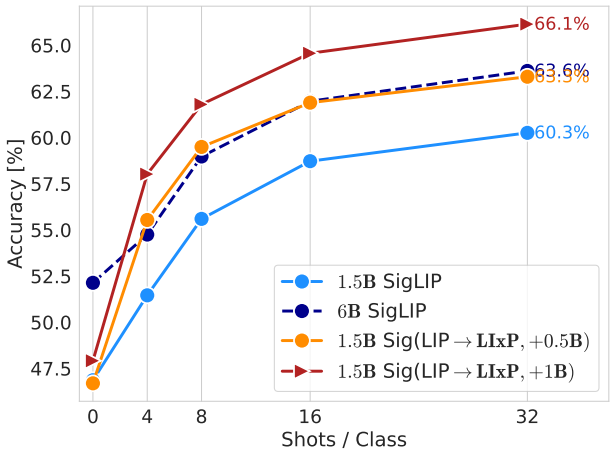

Figure 4. **Context-aware post-training.** We apply Sig**LIxP** on an already pretrained ViT-S/16 (1.5B examples), finetuning for +0.5B and +1B examples. We contrast the performance against a 6B ViT-S/16. Results indicate that context-aware finetuning can match much longer base pretraining with only +0.5B examples, and noticeably outperform it with just +1B examples - even if the base zero-shot transfer performance of the 6B reference model is much higher. Visualized results use Tip-Adapter for classification.

Tip-Adapter). Similarly, training a ViT-B/16 for 15B versus 6B training examples again maintains relative gains (e.g. +2.6%→2.7% when used with softmax-voted nearest neighbor classification). When scaling model and data further to ViT-L/16 (8B), improvements remain high (e.g. +1.8% for SNN and +4.0% for Tip-Adapter), with slightly lower absolute gains in parts attributable to several datasets reaching classification performance plateaus (e.g. 95.2% on `UC Merced`): When excluding datasets with base 32-shot performances over e.g 70%, the average gain for example on Tip-Adapter improves from +4.0% to +5.9%; with the same performance thresholding on the smallest ViT-S/16 (1.5B) giving only a +5.4%→+5.8% jump.

Context-aware post-training. Finally, we study the possibility of contextualized post-training, where SigLIxP is applied after base SigLIP pretraining. Results on a ViT-S/16 model pretrained on 1.5B WebLI examples are shown in Fig. 4. When finetuning on +0.5B additional examples using SigLIxP, zero-shot performance is retained but 32-shot performance increases from 60.3% to 63.3%, even approaching the 6B example baseline (trained on 3× more examples). When finetuning on +1B examples, we outperform the 6B SigLIP baseline (63.6% vs 66.1%). These results clearly indicate that post-hoc contextualized finetuning is possible and, measured by number of contextualized training iterations, more sample-efficient. Moreover, contextualized finetuning can significantly improve on few-shot adaptation performance compared to baseline pretraining which uses more than twice the number of examples.

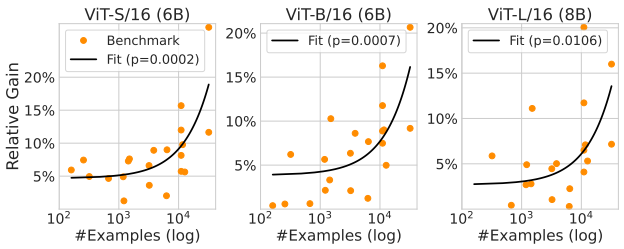

Figure 5. **Contextualized pretraining particularly benefits many-shot transfer.** For all 21 evaluation benchmarks, we plot the absolute number of examples (shots/class × #classes) against the relative gain when switching to SigLIxP. Results shown are for 32 shots/class. We find consistent relative gain in all scenarios, which become higher as absolute example counts increases.

Method	Train-free	IN-1K	DTD	Food101	Pets	Cars
Linear Probe [92]	✗	67.3	70.0	82.9	85.3	80.4
TIP-X [84]	✓	71.1	-	-	-	-
APE [108]	✓	72.1	-	-	-	-
DMN-TF [102]	✓	72.6	71.9	86.0	92.9	78.4
Clip-Adapter [21]	✗	71.1	-	-	-	
MaPLe [36]	✗	72.3	71.3	85.3	92.8	83.6
PromptSRC [37]	✗	73.2	72.7	87.5	93.7	83.8
Tip-Adapter-F [101]	✗	73.7	-	-	-	-
APE-T [108]	✗	74.3	-	-	-	-
CasPL [92]	✗	74.2	75.1	88.4	94.1	86.7
DMN [102]	✗	74.7	75.0	87.1	94.1	85.3
SigLIxP	✓	**77.9**	**76.7**	**92.6**	**94.4**	**92.8**
CLIxP	✓	77.6	76.0	91.4	94.1	91.8

Table 2. **Comparison against optimization-based methods in literature.** On five standard benchmarks, we compare Sig**LIxP** with simple *training-free*, softmax-voted nearest-neighbor classifier on top of the base zero-shot classification against e.g. fine-tuned Tip-Adapter [101] or SOTA prompt-learning methods [37, 92] (16-shot) on the same backbone (ViT-B/16). We show that through context-aware pretraining, simple *training-free* metric classifiers can be made highly competitive, strongly outperforming specifically tuned optimization approaches.

Benefits across few- and many-shots. Since each dataset has a different number of classes N, the actual number of support examples $N \times K$ varies per dataset. To account for this, we visualize relative gains (SigLIxP − SigLIP) / SigLIP over baseline training as a function of absolute number of examples provided in Figure 5. The linear fit (visualized semi-logarithmically) and p-value estimates are computed using [75]. Even when taking certain confounds (such as high base performance) into account, we find that performance gains increase with the absolute number of examples and reach their maximum in the many-shot setup.

Comparison with state-of-the-art. In Tab. 2, we compare SigLIxP results to state-of-the-art finetuning and prompt-learning methods on the same ViT-B/16 model. These methods specifically finetune and optimize for the downstream domains at hand (on 16-shots/class). In contrast, our SigLIxP backbone equipped with a simple and highly scalable *training-free* mechanism (softmax

Model → / **Examples →**	ViT-S/16 1.5B	→ 6B	→ 15B	ViT-B/16 6B	→ 15B
ZeroShot	47.9	52.1	53.3	60.8	61.9
	−0.2	0.0	−0.1	−0.2	+0.5
Prototypical	57.0 ± 0.3	59.5 ± 0.3	61.5 ± 0.3	66.3 ± 0.3	67.3 ± 0.2
	+4.0	+5.0	+4.0	+3.7	+3.4
Default Tip	60.0 ± 0.2	62.1 ± 0.2	63.8 ± 0.2	68.6 ± 0.3	69.5 ± 0.2
	+5.4	+6.2	+5.4	+4.8	+4.5
Softmax NN	64.5 ± 0.2	66.8 ± 0.1	68.2 ± 0.2	72.0 ± 0.2	73.1 ± 0.1
	+2.7	+2.3	+2.8	+2.8	+2.3

Table 3. **Context-aware pretraining for CLIP.** Benefits of contextualized pretraining also directly transfer to pretraining using the CLIP (softmax-based) objective [65], CLIxP, here shown for 32-shot classification for three metric-based classifiers.

NN + zero-shot logits) strongly outperforms these state-of-the-art optimization-based adaptation schemes, achieving 77.9% on ImageNet 16-shot (versus state-of-the-art DMN [102] 74.7%), or 76.7% on DTD versus state-of-the-art CasPL [92] 75.1%. Our context-aware pretraining *closes the gap between optimization-based and training-free methods*, where the latter generally lag behind the more sophisticated and expensive optimization-based ones, thereby notably simplifying adaptation to new domains, as a single unmodified model can be now used cheaply across domains.

Robustness to choice of pretraining objective. While we conduct most of our experiments using the more scalable SigLIP [99] objective, Tab. 3 shows that all the performance gains hold when switching the underlying pretraining loss to CLIP [65]. For each architecture and example count, we directly transfer the same hyperparameters to context-aware pretraining using CLIxP (Eq. (6)) with $\alpha = 0.9$ for 1.5B and 6B training examples, and $\alpha = 0.95$ for 15B runs. Similarly to the SigLIxP results, we see significant improvements when varying backbones, training duration and adaptation methods — e.g. +6.2% for a Tip-Adapter applied on a ViT-S/16 (6B) model or +2.8% for a softmax-voted nearest neighbor classifier on a ViT-S/16 (15B) model — while maintaining the base zero-shot transfer performance. This highlights the *generality of context-aware pretraining* for different types of vision-language contrastive pretraining.

Contextualized Pretraining Dynamics. To understand contextualized pretraining dynamics, we compare the dynamics of the base SigLIP training loss $\mathcal{L}_{\text{SigLIP}}$ and the contextualized SigLIP counterpart $\mathcal{L}^{\text{ctx}}_{\text{SigLIP}}$ (Eq. (6), first and second terms). Figure 6 (left) visualizes this against the changes in contextualization temperature τ_{ctx}. During early stages of training (around the warmup period), no context use is learned. Only once the base objective reaches a certain threshold with sufficient representation quality does the model start to leverage available context following Eq. (8). This coincides with τ_{ctx} reaching a value sufficiently low to encourage strong separation of buffer entries. During this joint inflection point, we often find an increase in the

Method	Zero-Shot	16-Shot
SigLIP	51.0	60.1 ± 0.4
Sig**LIxP**	50.5	64.1 ± 0.5
No masking	50.9	60.5 ± 0.4

(a) **Self-Attention Masking**

Method	Zero-Shot	16-Shot
$\alpha = 0.95$	50.8	62.5 ± 0.3
$\alpha = 0.9$	50.5	64.1 ± 0.5
$\alpha = 0.8$	50.0	63.8 ± 0.3
$\alpha = 0.6$	48.7	61.5 ± 0.4

(b) **Relative weighting**

Method	Zero-Shot	16-Shot
Single-Stage	50.5	64.1 ± 0.5
Residual	49.2	59.2 ± 0.4
Multimodal	47.9	61.4 ± 0.4
Two-Stage	50.5	62.0 ± 0.3

(c) **Contextualization Type**

Method	Zero-Shot	16-Shot
$\tau_1, \tau_2, \tau_{\text{ctx}}$ learnable	50.5	64.1 ± 0.5
$\tau_1 = \tau_2, \tau_{\text{ctx}}$	47.8	61.8 ± 0.4
$\tau_1, \tau_2 = \tau_{\text{ctx}}$	50.4	60.2 ± 0.6
τ_{ctx} frozen	47.1	59.4 ± 0.5

(d) **Uncoupled, learnable temperatures**

Method	Zero-Shot	16-Shot
Full back-propagation	50.5	64.1 ± 0.5
Stop Gradient: $\{K\}$	49.6	62.7 ± 0.4
Stop Gradient: $\{V\}$	43.8	58.7 ± 0.2
Stop Gradient: $\{K, V\}$	44.4	56.5 ± 0.4

(e) **Back-propagation through memory buffer**

Table 4. **Context-Aware Pretraining Ablations.** We ablate the context-aware training objective in Eq. (6) and Eq. (8) on a ViT-S/16 (1.5B), evaluated with prototypical classifiers on a subset of our evaluation benchmarks. We report average zero- and 16-shot performances.

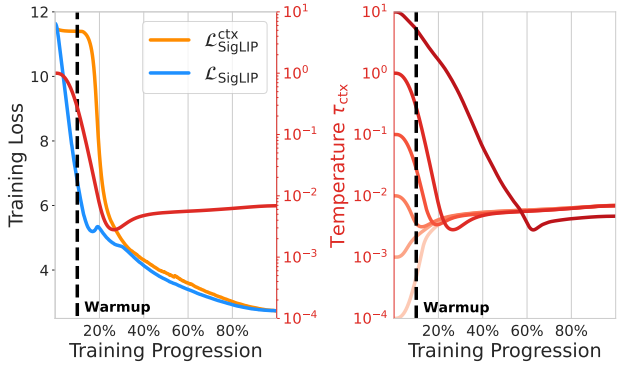

Figure 6. **Training Dynamics.** *(Left)* Relation between the base SigLIP training objective $\mathcal{L}_{\text{SigLIP}}$ and its contextualized counterpart $\mathcal{L}^{\text{ctx}}_{\text{SigLIP}}$ (here referring to the second term in Eq. (6)) for a ViT-B/16 model trained on 6B WebLI examples using SigLIxP. We jointly visualize the learned contextualization temperature τ_{ctx} (note the log-scale). **(Right)** Progression of τ_{ctx} for different initializations. We find that context-usage is emergent during training and dependent on a suitable learned context temperature τ_{ctx}, which fortunately is very robust across initializations.

base siglip loss, followed by a slight increase in τ_{ctx} for less aggressive context selection to better align the base representation pretraining and the contextualization paradigm. We find that ensuring τ_{ctx} to be exponentially learnable following the formulation using in Zhai et al. [99] is crucial to effectively trade-off retention of base zero-shot transfer capabilities and the improvement in few-shot adaptability. This exponential treatment also introduces robustness towards τ_{ctx} starting values (Fig. 6, right); with temperatures converging to similar inflection points irrespective of initializations across orders of magnitude (10^{-4} to 1).

4.2. Ablations

The context-aware pretraining framework introduced in Sec. 3.2 is the best performing, most scalable instance of a larger contextualization framework which we explored. In this section, we characterize the design space for our context-aware pretraining objective in Tab. 4, and for particular buffer design choices $\mathcal{M}_K$ and $\mathcal{M}_V$ in Tab. 5. For more details, see supplementary.

Mitigating shortcut solutions. We first show the importance of correctly masking out self-attention entries in the cross-attention formalism in Eq. (8) via $\mathbf{M}$ (Tab. 4a). We can see that without masking out self-attention in Eq. (8), no context re-use emerges with results matching the baseline SigLIP (first row), as the model simply learns to attend to itself (having the highest similarity by default), consequently having no incentive to leverage other context.

Loss balancing. We ablate the relative weighting α between base image-text contrastive objective and its contextualized counterpart (Eq. (6)) in Tab. 4b, observing significant gains in few-shot transfer and retention of base zero-shot performance for $\alpha \in [0.95, 0.8]$. For larger α, training regresses back to the base image-text contrastive objective, while smaller α skew too heavily towards context usage at increased cost to zero-shot transfer.

Contextualization objective. In Tab. 4c, we investigate several modifications to the contextualization objective. **(a)** Instead of separating training into base and context loss (c.f. Eq. (6)), we instead utilize a residual connection to Eq. (5) and train on a single objective $\mathcal{L}_{\text{SigLIP}}(\alpha x_i + (1-\alpha)x^{\text{ctx}}_i, t_i, \tau_1)$. **(b)** Multimodal contextualization by setting $\mathcal{M}_V$ to use textual representations $\mathbf{T}$. **(c)** Conducting two successive contextualization steps via

$$x^{\text{ctx}}_{i,m} = \psi_{\text{ctx}}\left(f_{\text{ctx}}(x^{\text{ctx}}_{i,m-1}, \mathcal{M}^{m-1}_K, \mathcal{M}^{m-1}_V)\right) \quad (9)$$

with $m \in \{1, ..., M\}, M = 2$, and f_{ctx} our contextualization objective from Eq. (8). ψ_{ctx} defines an optimal learnable map, which we set to a linear one for our two-stage contextualization. Note that for $M = 1$ and ψ_{ctx} the identity function we recover our base objective in Eq. (8). First,

Method	Zero-Shot	16-Shot
SigLIP	51.0	60.1 ± 0.4
Sig**LIxP**	50.5	64.1 ± 0.5
Separate Batch	48.8	63.2 ± 0.3

(a) **Context Batch Separation**

Method	Zero-Shot	16-Shot
None	50.5	64.1 ± 0.5
linear	50.2	62.8 ± 0.2
2-layer MLP	49.8	61.1 ± 0.2
3-layer MLP	49.1	60.8 ± 0.3

(b) **Value Heads for $\mathcal{M}_V$**

Method	Zero-Shot	16-Shot
SigLIP	51.0	60.1 ± 0.4
Sig**LIxP**	50.5	64.1 ± 0.5
No Normalization	48.9	61.7 ± 0.3

(c) **QK Normalization**

Method	Zero-Shot	16-Shot
None	50.5	64.1 ± 0.5
LayerNorm $\{K\}$	46.8	57.1 ± 0.4
LayerNorm $\{V\}$	48.5	62.3 ± 0.5
LayerNorm $\{K, V\}$	48.3	58.2 ± 0.3

(d) **Layer Normalization on $\mathcal{M}$**

Method	Zero-Shot	16-Shot
None	50.5	64.1 ± 0.5
+ Stale (32K)	45.9	59.7 ± 0.3
+ Stale (128K)	46.9	60.5 ± 0.5
+ Stale (512K)	47.2	60.7 ± 0.4

(e) **Inclusion of Stale Buffer**

Method	Zero-Shot	16-Shot
Full (32k)	50.5	64.1 ± 0.5
Subset (1k)	50.7	59.9 ± 0.5
Subset (2k)	50.5	62.9 ± 0.4
Subset (8k)	50.3	63.9 ± 0.3

(f) **Reduced Active Buffer Size**

Table 5. **Context Buffer Ablations.** Following Tab. 4, we study context buffer design $\mathcal{M}_K$ and $\mathcal{M}_V$. Our results clearly motivate our most scalable, and yet best performing variant in Eq. (8) which directly reuses batch-level image representations $\phi_I(\mathcal{B}_I)$.

we find that for residual connection in **(a)**, the optimal α was $\alpha = 1$, i.e. effectively retaining the base contrastive training objective. For slightly smaller values, i.e capping α at 0.9, both zero-shot transfer and few-shot adaptability (51.0%$\rightarrow$ 49.2% and 60.1%$\rightarrow$ 59.2%) are negatively impacted, as it effectively mitigates effectiveness of the base SigLIP objective. This showcases the *explicit need for separation of base and contextualization losses* alongside distinctly learnable temperatures. Utilizing multimodal context **(b)** reduces transfer, albeit maintaining slight gains in few-shot capabilities, and only works if masking $\mathbf{M}$ in Eq. (8) is set, as otherwise a simple weight-copying shortcut is found. We find two-stage contextualization **(c)** reduces gains compared to *single-stage contextualization*.

Individual, learnable temperatures. Having three distinct temperatures (τ_1, τ_2, τ_{ctx}) is crucial for optimal performance as shown in Tab. 4d. We ablate **(a)** sharing the temperature across all constituents of Eq. (6), **(b)** sharing the SigLIxP and contextualization temperature, and **(c)** freezing τ_{ctx} to either $\{0.01, 0.1, 1\}$ and reporting the best results. We find that allowing τ_{ctx} to adapt to be crucial for few-shot gains and restricting $\tau_1 = \tau_2$ to negatively impact the zero-shot transfer ability. Consequently, our results strongly advocate for a *full set of independently learnable* temperatures.

Key and value backpropagation. Eq. (8) allows the model backpropagate through key and value entries $\mathcal{M}_K$ and $\mathcal{M}_V$. To understand which component is most crucial for peak performance, we ablate the gradient flow through $\mathcal{M}_K$ and $\mathcal{M}_V$ in Tab. 4e. Interestingly, freezing $\mathcal{M}_V$ and only optimizing for the retrieval component (i.e. entries within $\sigma(\cdot)$ in Eq. (8), directly mimicking retrieval operations at test-time) is highly detrimental (16-shot score 64.1% $\rightarrow$ 58.7%), whereas freezing $\mathcal{M}_K$ has a much smaller negative impact (62.7% versus 60.1%). For contextualization to function for pretraining, allowing the model to optimize for both key, but especially value entries is thus important to reach optimal context usage.

General Buffer Design. To understand the right way to structurally design the contextualization buffer (see Table 5), we investigate using examples from a separate batch $\mathcal{B}_I$ to populate $\mathcal{M}$ (Tab. 5a), learnable value heads over $\mathcal{M}_V$ (Tab. 5b), normalization of query-key representations in Eq. (8) (Tab. 5c), layer normalization of buffer entries (Tab. 5d), the inclusion of stale memories from previous iterations (Tab. 5e), and finally the reduction of utilized $\mathbf{X}_I$ in Eq. (8) (Tab. 5f). Table 5a shows that populating $\mathcal{M}_K$ and $\mathcal{M}_V$ with a separately embedded batch results in a small decrease in both zero-shot and few-shot performance. From Table 5b and Table 5d, we conclude that no value heads or layer normalization are needed - most closely aligning with the direct re-use of learned embeddings downstream. Similarly, Tab. 5c shows that normalized embeddings for QK weighting in Eq. (8) are essential. Extending $\mathcal{M}$ with stale representations from previous iterations in Tab. 5e offers no benefits, though on the other hand, maintaining a large enough active buffer $|\mathcal{M}| = |\mathcal{B}|$ in **(f)** is crucial to reliably benefit from contextualized training.

5. Conclusion

We introduced a context-aware pretraining objective for large-scale vision-language representation learning that facilitates few- and many-shot visual context use in a training-free, metric-based manner at test time. Importantly, we demonstrate the possibility of significantly boosting visual adaptation performance with no compromises in underlying zero-shot transfer capabilities. Extensive evaluations on 21 diverse visual adaptation tasks show up to *four-fold* gains in test-time sample efficiency and improvements in average few-shot performance by often over 5%, closing the gap to more complex optimization-based strategies. These gains hold across model and data scales, highlighting the potential of simple and scalable pretraining strategies to augment test-time adaptation beyond simple zero-shot transfer.

Acknowledgements

The authors would like to thank Nikhil Parthasarathy, Relja Arandelović and Vishaal Udandarao for helpful feedback. KR thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS), the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program and the ELISE Mobility Grant for support. ZA acknowledges the support from the German Research Foundation (DFG): SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, project number: 276693517 and ERC Grant DEXIM, project number: 853489. All authors thank Google DeepMind for providing the resources and high-quality environment for this research.

References

[1] Ivana Balazevic, David Steiner, Nikhil Parthasarathy, Relja Arandjelovic, and Olivier J Henaff. Towards in-context scene understanding. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 1, 2

[3] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. 1

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3, 1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[6] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[7] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9042–9051, 2021. 1, 2

[8] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 1, 2

[9] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[10] 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 (CVPR) Workshops*, pages 4950–4958, 2022. 1

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1, 2

[12] Kun Ding, Qiang Yu, Haojian Zhang, Gaofeng Meng, and Shiming Xiang. Calibrated cache model for few-shot vision-language model adaptation. *arXiv preprint arXiv:2410.08895*, 2024. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3, 1

[14] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9588–9597, 2021. 2

[15] Talfan Evans, Shreya Pathak, Hamza Merzic, Jonathan Schwarz, Ryutaro Tanno, and Olivier J Henaff. Bad students make great teachers: Active learning accelerates large-scale visual understanding. *arXiv preprint arXiv:2312.05328*, 2023. 2

[16] Talfan Evans, Nikhil Parthasarathy, Hamza Merzic, and Olivier J Henaff. Data curation via joint example selection further accelerates multimodal learning. *arXiv preprint arXiv:2406.17711*, 2024. 2, 1

[17] 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). In *ICML*, pages 6216–6234. PMLR, 2022. 1

[18] Benjamin Feuer, Ameya Joshi, and Chinmay Hegde. Caption supervision enables robust learners. *arXiv preprint arXiv:2210.07396*, 2022. 1

[19] Christopher Fifty, Dennis Duan, Ronald Guenther Junkins, Ehsan Amid, Jure Leskovec, Christopher Re, and Sebastian Thrun. Context-aware meta-learning. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[20] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks.

In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 1126–1135. JMLR.org, 2017. 2

[21] 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. *Int. J. Comput. Vision*, 132(2):581–595, 2023. 6

[22] Robert Geirhos, Priyank Jaini, Austin Stone, Sourabh Medapati, Xi Yi, George Toderici, Abhijit Ogale, and Jonathon Shlens. Towards flexible perception with visual memory. *arXiv preprint arXiv:2408.08172*, 2024. 1, 2, 3, 5

[23] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. In *Advances in Neural Information Processing Systems*, pages 6704–6719. Curran Associates, Inc., 2022. 2

[24] Zhongrui Gui, Shuyang Sun, Runjia Li, Jianhao Yuan, Zhaochong An, Karsten Roth, Ameya Prabhu, and Philip Torr. kNN-CLIP: Retrieval enables training-free segmentation on continually expanding large vocabularies. *Transactions on Machine Learning Research*, 2024. 1

[25] 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. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2023. 2

[26] Jinwei Han, Zhiwen Lin, Zhongyisun Sun, Yingguo Gao, Ke Yan, Shouhong Ding, Yuan Gao, and Gui-Song Xia. Anchor-based robust finetuning of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26919–26928, 2024. 1

[27] Zongbo Han, Jialong Yang, Junfan Li, Qinghua Hu, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. *arXiv preprint arXiv:2409.19375*, 2024. 2

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019. cite arxiv:1911.05722Comment: CVPR 2020 camera-ready. Code: https://github.com/facebookresearch/moco. 2

[29] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 1, 2

[30] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23369–23379, 2023. 2

[31] David P. Hughes and Marcel Salathé . An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. *CoRR*, abs/1511.08060, 2015. 1, 2

[32] Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Retrieval-enhanced contrastive vision-text models. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[33] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2

[34] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. 1

[35] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Z"ollner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6:27988, 2016. 1, 2

[36] 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 (CVPR)*, pages 19113–19122, 2023. 6

[37] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15190–15200, 2023. 2, 6

[38] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Datadream: Few-shot guided dataset generation. In *European Conference on Computer Vision*, pages 252–268. Springer, 2025. 2

[39] M. Kirchhof, K. Roth, Z. Akata, and E. Kasneci. A non-isotropic probabilistic take on proxy-based deep metric learning. In *Computer Vision - ECCV 2022 - 17th European Conference, Proceedings, Part XXVI*, pages 435–454. Springer, 2022. 4

[40] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021. 1

[41] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 1, 2

[42] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 1, 2

[43] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71,

Brussels, Belgium, 2018. Association for Computational Linguistics. 1

[44] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[45] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representation learning from multiple domains for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9526–9535, 2021. 2

[46] Yaohui Li, Qifeng Zhou, Haoxing Chen, Jianbing Zhang, Xinyu Dai, and Hao Zhou. The devil is in the few shots: Iterative visual knowledge completion for few-shot learning. *arXiv preprint arXiv:2404.09778*, 2024. 2

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[48] Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. A closer look at few-shot classification again. In *Proceedings of the 40th International Conference on Machine Learning*, pages 23103–23123. PMLR, 2023. 1, 2

[49] Yury A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1

[50] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[51] Yifei Ming and Yixuan Li. Understanding retrieval-augmented task adaptation for vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 35719–35743. PMLR, 2024. 2

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

[53] Ernest Mwebaze, Timnit Gebru, Andrea Frome, Solomon Nsumba, and Jeremy Tusubira. icassava 2019 fine-grained visual categorization challenge. *arXiv preprint arXiv:1908.02900*, 2019. 1, 2

[54] Kengo Nakata, Youyang Ng, Daisuke Miyashita, Asuka Maki, Yu-Chieh Lin, and Jun Deguchi. Revisiting knn-based image classification system with high-capacity storage. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, page 457–474, Berlin, Heidelberg, 2022. Springer-Verlag. 1, 3, 5

[55] A Nichol. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 1, 2

[56] Zachary Novack, Julian Mcauley, Zachary Chase Lipton, and Saurabh Garg. CHiLS: Zero-shot image classification with hierarchical label sets. In *Proceedings of the 40th International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023. 2

[57] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[58] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 1

[59] Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. Svl-adapter: Self-supervised adapter for vision-language pretrained models. In *British Machine Vision Conference (BMVC)*, 2022. 1, 2

[60] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 2

[61] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 1, 2

[62] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning. *Neurocomput.*, 555(C), 2023. 2

[63] Ameya Prabhu, Zhipeng Cai, Puneet Dokania, Philip Torr, Vladlen Koltun, and Ozan Sener. Online continual learning without the storage constraint. *arXiv preprint arXiv:2305.09253*, 2023. 1

[64] 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 (ICCV)*, pages 15691–15701, 2023. 2

[65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6

[66] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 1

[67] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 2

[68] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, 2022. 1

[69] Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15746–15757, 2023. 1, 2

[70] Karsten Roth, Vishaal Udandarao, Sebastian Dziadzio, Ameya Prabhu, Mehdi Cherti, Oriol Vinyals, Olivier Hénaff, Samuel Albanie, Matthias Bethge, and Zeynep Akata. A practitioner's guide to continual multimodal pretraining. *arXiv preprint arXiv:2408.14471*, 2024. 1, 2

[71] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 2

[72] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. 2

[73] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a study on representation learning. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[74] Tyler R. Scott, Andrew C. Gallagher, and Michael C. Mozer. von mises-fisher loss: An exploration of embedding geometries for supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10612–10622, 2021. 4

[75] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. 6

[76] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4080–4090, Red Hook, NY, USA, 2017. Curran Associates Inc. 1, 2, 3, 5

[77] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100, Dublin, Ireland, 2022. Association for Computational Linguistics. 1, 2

[78] Zafir Stojanovski, Karsten Roth, and Zeynep Akata. Momentum-based weight interpolation of strong zero-shot models for continual learning. *arXiv preprint arXiv:2211.03186*, 2022. 2

[79] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 2

[80] Ajinkya Tejankar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv preprint arXiv:2112.13884*, 2021. 2

[81] Lukas Thede, Karsten Roth, Olivier J Hénaff, Matthias Bethge, and Zeynep Akata. Reflecting on the state of rehearsal-free continual learning with pretrained models. *arXiv preprint arXiv:2406.09384*, 2024. 1

[82] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, pages 266–282. Springer, 2020. 1, 2

[83] Eleni Triantafillou, Hugo Larochelle, Richard Zemel, and Vincent Dumoulin. Learning a universal template for few-shot dataset generalization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10424–10433. PMLR, 2021. 2

[84] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *ICCV*, 2023. 1, 2, 6

[85] Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 3, 1

[87] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 1, 2

[88] 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. 1, 2

[89] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens Van Der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 2

[90] Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[91] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 1, 2

[92] Ge Wu, Xin Zhang, Zheng Li, Zhaowei Chen, Jiajun Liang, Jian Yang, and Xiang Li. Cascade prompt learning for vision-language model adaptation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part L*, page 304–321, Berlin, Heidelberg, 2024. Springer-Verlag. 2, 6

[93] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 1

[94] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. 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, 2010. 1, 2

[95] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19265–19274, 2023. 2

[96] Jin Xu, Jean-Francois Ton, Hyunjik Kim, Adam Kosiorek, and Yee Whye Teh. MetaFun: Meta-learning with iterative functional updates. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10617–10627. PMLR, 2020. 1, 2

[97] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS)*, 2010. 1, 2

[98] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. TapNet: Neural network augmented with task-adaptive projection for few-shot learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7115–7123. PMLR, 2019. 2

[99] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 1, 2, 3, 4, 6, 7

[100] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12200–12210, 2020. 2

[101] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, page 493–510, Berlin, Heidelberg, 2022. Springer-Verlag. 1, 2, 3, 4, 5, 6

[102] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 2, 3, 6, 1

[103] Zhaoxiang Zhang, Hanqiu Deng, Jinan Bao, and Xingyu Li. Dual-image enhanced clip for zero-shot anomaly detection. *arXiv preprint arXiv:2405.04782*, 2024. 1, 2

[104] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1, 2

[105] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16795–16804, 2022. 1, 2

[106] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 2

[107] Yifei Zhou, Juntao Ren, Fengyu Li, Ramin Zabih, and Ser-Nam Lim. Test-time distribution normalization for contrastively learned visual-language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[108] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2605–2615, 2023. 2, 3, 6, 1

[109] Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[110] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7693–7702. PMLR, 2019. 2