Efficient and Long-Tailed Generalization for Pre-trained Vision-Language Model

Jiang-Xin Shi*
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, China
shijx@lamda.nju.edu.cn

Chi Zhang*
National Key Laboratory for Novel Software Technology
Nanjing University, China
chi-zhang@smail.nju.edu.cn

Tong Wei
School of Computer Science and Engineering
Key Laboratory of Computer Network and Information Integration of Ministry of Education
Southeast University, China
weit@seu.edu.cn

Yu-Feng Li†
National Key Laboratory for Novel Software Technology
School of Artificial Intelligence
Nanjing University, China
liyf@lamda.nju.edu.cn

ABSTRACT

Pre-trained vision-language models like CLIP have shown powerful zero-shot inference ability via image-text matching and prove to be strong few-shot learners in various downstream tasks. However, in real-world scenarios, adapting CLIP to downstream tasks may encounter the following challenges: 1) data may exhibit long-tailed data distributions and might not have abundant samples for all the classes; 2) There might be emerging tasks with new classes that contain no samples at all. To overcome them, we propose a novel framework to achieve efficient and long-tailed generalization, which can be termed as **Candle**. During the training process, we propose compensating logit-adjusted loss to encourage large margins of prototypes and alleviate imbalance both within the base classes and between the base and new classes. For efficient adaptation, we treat the CLIP model as a black box and leverage the extracted features to obtain visual and textual prototypes for prediction. To make full use of multi-modal information, we also propose cross-modal attention to enrich the features from both modalities. For effective generalization, we introduce virtual prototypes for new classes to make up for their lack of training images. **Candle** achieves state-of-the-art performance over extensive experiments on 11 diverse datasets while substantially reducing the training time, demonstrating the superiority of our approach. The source code is available at https://github.com/shijxcs/Candle.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**.

KEYWORDS

long-tail learning, vision-language model, new class generalization

ACM Reference Format:
Jiang-Xin Shi, Chi Zhang, Tong Wei, and Yu-Feng Li. 2024. Efficient and Long-Tailed Generalization for Pre-trained Vision-Language Model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671945

1 INTRODUCTION

Over the past few years, the rapid development of deep learning [8, 37] and the emergence of web-scale datasets [5, 30, 31] have made large-scale pre-trained models possible. Particularly, Vision-Language (V-L) models [2, 15, 18, 27, 42] have become a recent research hype due to their strong generalization capabilities as well as promising transferability to downstream tasks. One of the most successful pre-trained V-L models is CLIP [27]. Trained on a massive dataset of 400 million image-text pairs, CLIP utilizes a contrastive objective to align the visual and textual representations and manages to establish a connection between images and natural language. During inference, CLIP can perform zero-shot image recognition by simply using the class names. For example, one can adopt a prompt template like 'a photo of a {class}' as input to the text encoder and generate the classification weights for each class. The weights can then be used to calculate cosine similarity with image features to get classification scores.

With the rise of such powerful V-L models, extensive efforts have been invested into finding potential solutions to better adapt these models to downstream tasks. For instance, several previous works including CoOp[46], CoCoOp [45] and MaPLe [16] have explored the idea of prompt learning, where a sequence of learnable context vectors is used to replace carefully-chosen hard prompts. These methods have achieved impressive improvements.

Despite delivering promising results, a number of existing works suffer from two practical limitations. a) significant performance decline under real-world long-tailed data distributions. The natural long-tail distribution [19] phenomenon brings class imbalance and makes it hard to collect data for all classes, leaving some rare classes

*Equal contribution.
†Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '24, August 25–29, 2024, Barcelona, Spain.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671945

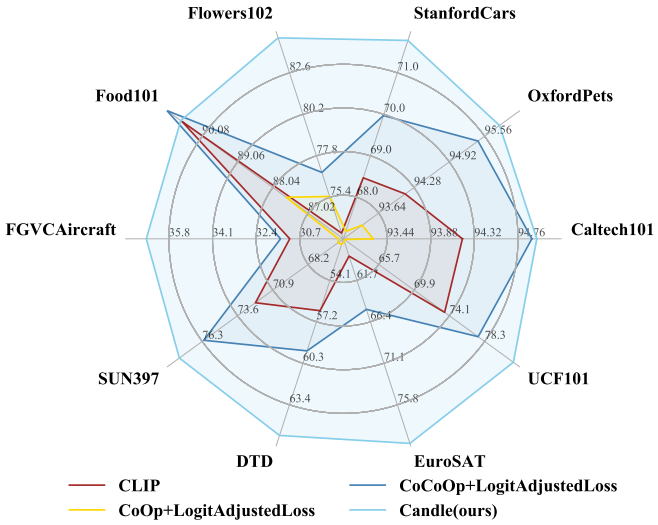

Figure 1: Candle achieves significant improvements on multiple imbalanced base-to-new generalization tasks.

Table 1: Training time and accuracy for CoOp, CoCoOp, and Candle with ViT-B/16 as the visual encoder, on a 100-shot ImageNet with an imbalance ratio of 100. The experiments are done on an RTX 3090 GPU.

Method	Training epochs	Time cost	Accuracy (%)
CoOp	50	~ 5 hours	70.7
CoCoOp	10	~ 30 hours	71.3
Candle (ours)	20	**11 min**	**71.6**

entirely void of samples. Most works fail to consider the real-world data distributions and suffer from severe performance degradation under imbalanced scenarios. They also tend to overlook the valuable label information for unseen categories, which may be a leading cause for a notable performance drop on these label-only classes. b) extensive computational overhead. Despite using fewer trainable parameters, most methods still need to calculate gradients through the model's backbone and require access to the model's weights. As the size of foundation V-L models continues to grow (*e.g.*, up to 80 billion [2]) and industry standards gradually switch to providing only API, they may become impractical for actual application.

In this paper, we aim to address the above issues and propose a novel framework to achieve efficient and long-tailed generalization which can be named as Candle. During the training process, we propose compensating logit-adjusted loss to encourage large margins of virtual prototypes and alleviate imbalance both within the base classes and between the base and new classes. For efficient adaptation, we treat the CLIP model as a black box and leverage the extracted features to obtain visual and textual prototypes for prediction. To make full use of multi-modal information, we also propose cross-modal attention to enrich the features from both modalities. For effective generalization, we introduce virtual prototypes for new classes to make up for their lack of training images. As shown in Figure 1 and Table 1, our method achieves impressive improvements over previous methods while cutting down the training time. In summary, the main contributions of this work include:

- We propose a novel framework named Candle for efficient and long-tailed generalization of CLIP. To the best of our knowledge, this is the first work to explore the adaptation of V-L models under an imbalanced setting.
- To make full use of both visual and textual information, we propose to perform *cross-modal attention* on the feature space. For better new class generalization, we introduce *virtual prototypes* and propose a novel *compensating logit adjusted loss* to simultaneously alleviate the imbalance within the base classes as well as between the base and new classes.
- Our extensive experimental results demonstrate the strength of Candle, which achieves state-of-the-art results over various settings while substantially reducing the training time.

2 RELATED WORK

Vision-Lanuage (V-L) Models. V-L foundation models have experienced a substantial surge in recent years with the emergence of different architectures such as Flamingo [2], CLIP [27], ALIGN [15], BLIP [18], CoCa [42], etc. These models are usually trained on a web-scale dataset comprised of massive image-text pairs to learn a joint embedding space. Due to their strength in understanding open-vocabulary concepts, V-L models have been widely explored in various downstream tasks, such as few-shot learning [33, 46], continual learning [6, 44] and adversarial learning [22, 35]. In this work, we focus on adapting CLIP for new class generalization.

Fine-tuning V-L Models. Despite the effectiveness of V-L models (*e.g.*, CLIP) towards generalizing to new concepts, its massive scale makes it infeasible to fine-tune the full model for downstream tasks. Linear probing [27] serves as a naive solution, while its performance deteriorates significantly under few-shot settings. CoOp [46] proposes the idea of prompt learning, which optimizes a set of context vectors instead of using the standard prompt template 'a photo of a {class}'. CoCoOp [45] aims to learn more robust prompts through image conditioning, which optimizes an instance-specific prompt by training a meta-network. CoCoOp also proposes the novel base-to-new setting for better examination of a model's generalizability. MaPLe [16] simultaneously learns the prompts for both the vision and language branches of CLIP. While these methods have achieved impressive results under few-shot settings, their training cost can be prohibitive in terms of both time and memory.

Aside from prompt learning, another line of work utilizes adapter modules for lightweight and fast adaptation. For instance, CLIP-Adapter [10] proposes to add an MLP layer after the final visual layer and mix the transformed output with the original zero-shot output via a residual connection. TIP-Adapter [43] further replaces the MLP layer with a carefully designed linear layer, whose weights are comprised of labeled visual embeddings. Although these works have significantly reduced the training cost for fine-tuning CLIP, they perform poorly under the base-to-new setting, with TIP-Adapter [43] even unable to test on new classes.

Furthermore, subsequent works have attempted to improve adaptation by leveraging multi-modal information [25] or adopting a generative approach to synthesize features for categories without

Table 2: Empirical study results for zero-shot CLIP and visual prototypes over 11 datasets, using ViT-B/16 as the visual encoder. The visual prototypes are obtained by calculating the mean value of 16-shot features for each class and used subsequently to calculate cosine similarity with image features to get the classification scores.

	CAL.	OP.	SC.	Flw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	IN.	Avg Results.
Zero-shot CLIP	89.3	**88.9**	65.6	70.4	**89.2**	27.1	65.2	46.0	54.1	69.8	**68.6**	66.7
Visual Prototypes	**93.4**	80.2	**71.7**	**95.9**	81.4	**41.3**	**69.8**	**64.2**	**75.2**	**78.1**	61.7	**73.9**
Δ	+4.1	-8.7	+6.1	+25.5	-7.8	+14.2	+4.6	+18.2	+21.1	+8.3	-6.9	+7.2

data [40]. However, they still suffer from expensive training costs or unsatisfactory new-class generalizations. In contrast to the aforementioned works, this paper presents a lightweight framework built directly upon the feature space for efficient adaptation, as well as virtual prototypes with a novel loss function for effective new class generalization.

Imbalance Learning via Pre-trained Models. Recent research has found that models pre-trained on large-scale datasets can learn more generalized representations, and can serve as an effective tool for alleviating class imbalance issues. For instance, BALLAD [20] and VL-LTR [36] fine-tunes both of the entire image and text encoder of CLIP on the downstream tasks. Wang et al. [39] proposes to ignore the text branch of CLIP and append a decoder consisting of transformer blocks after the image encoder. LPT [7] adopts a two-stage method to learn both shared prompts and group-specific prompts to capture both general and specialized knowledge. PEL [32] systematically investigates different parameter-efficient fine-tuning modules for long-tailed recognition tasks. In contrast to these previous works, this paper deals with the new class generalization setting and proposes an efficient and effective approach.

3 CLIP

3.1 Premilinaries

Contrastive Language-Image Pretraining, known as CLIP [27], is mainly comprised of an image encoder $f_I(x)$ and a text encoder $f_T(t)$, which map input from the respective modality into a joint embedding space. The image encoder can be in the form of either ResNet [11] or ViT [8], whereas the text encoder is built on top of the Transformer [37] architecture.

During training, CLIP goes through 400 million image and caption pairs, adopting a contrastive loss to pull together the corresponding image-text pairs while pushing apart unmatched ones. After training, CLIP can be readily used for downstream image classification in a zero-shot manner. Let x be the input image and $\{t_1, \cdots, t_K\}$ be the K class descriptions. These descriptions can be generated through prompt templates like 'a photo of a {class}', where the {class} token denotes the corresponding class name. Then, it extracts image features $\boldsymbol{x} = f_I(x)$, textual prototypes $\boldsymbol{T} = \{\boldsymbol{T}_1, \cdots, \boldsymbol{T}_K\} = \{f_T(t_1), \cdots, f_T(t_K)\}$, and the predicted result for x is:

$$y_{\text{pred}} = \arg\max_{i \in [K]} \cos(\boldsymbol{x}, \boldsymbol{T}_i) \tag{1}$$

In this way, CLIP turns the image classification task into an image-text matching problem.

3.2 Practical Limitations of CLIP

Despite delivering promising results, a number of existing works suffer from two practical limitations. a) significant performance decline under real-world long-tailed data distributions. The natural long-tail distribution [19] phenomenon brings class imbalance and makes it hard to collect data for all classes, leaving some rare classes entirely void of samples. Most works fail to consider the real-world data distributions and suffer from severe performance degradation under imbalanced scenarios. They also tend to overlook the valuable label information for unseen categories, which may be a leading cause for a notable performance drop on these label-only classes. b) extensive computational overhead. Despite using fewer trainable parameters, most methods still need to calculate gradients through the model's backbone and require access to the model's weights. As the size of foundation V-L models continues to grow (*e.g.*, up to 80 billion [2]) and industry standards gradually switch to providing only API, they may become impractical for actual application.

Moreover, despite the effectiveness of CLIP in general cases, its insufficient usage of visual information remains a weak point. CLIP relies highly on image-text matching for downstream zero-shot prediction, which may cause potential risks. For instance, on the FGVCAircraft [21] dataset, the class names are different numeral versions such as '737-200' and '737-300', which hardly contain any useful information; or on the UCF101 [34] dataset, the image samples consist of frames from a video and do not precisely match the prompt templates such as 'a photo of a {class}'.

Based on this motivation, we conduct our empirical study by comparing image-image matching with image-text matching. Specifically, we calculate visual prototypes as the mean value of the 16-shot image features for each class. Then, we replace the textual prototypes in zero-shot CLIP with visual prototypes for prediction. Formally, let $V = \{\boldsymbol{V}_1, \cdots, \boldsymbol{V}_K\}$ be the K visual prototypes for each class, then the predicted result is:

$$y_{\text{pred}} = \arg\max_{i \in [K]} \cos(\boldsymbol{x}, \boldsymbol{V}_i) \tag{2}$$

We compare the prediction results calculated by Equation 1 and by Equation 2, and report the empirical results in Table 2. As shown, zero-shot CLIP significantly underperforms visual prototypes on multiple datasets including Flowers102 (-25.5%), FGVCAirCraft (-14.2%), DTD (-18.2%) and EuroSAT (-21.1%), demonstrating that the class labels are not descriptive enough.

4 CANDLE: EFFICIENT AND LONG-TAILED GENERALIZATION

In this section, we aim to address the above issues and propose a novel framework to achieve efficient and long-tailed generalization

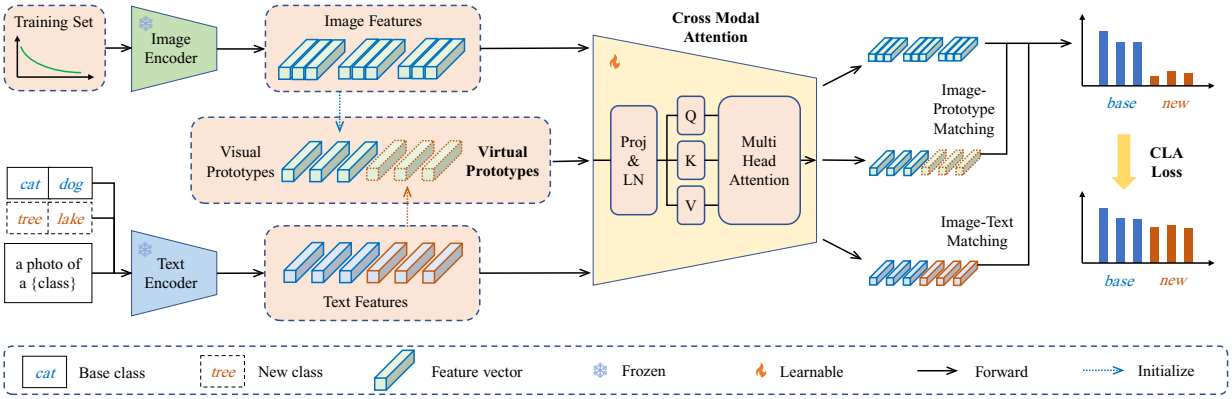

Figure 2: An overview of the proposed framework.

which can be named as Candle. During the training process, we propose compensating logit-adjusted loss to mitigate the long-tail problems, as well as to avoid the risk of neglecting new classes during the optimization. For efficient adaptation, we directly leverage the features extracted from the model, calculate the corresponding visual and textual prototypes, and propose cross-modal attention to enrich the information for both modalities. For effective generalization, we propose to generate virtual prototypes for new classes, by which we compensate for their lack of training images.

4.1 Compensating Logit-Adjusted Loss

Loss functions designed to deal with class imbalance usually do not apply to new class generalization since there is no sample for new classes. Therefore, we propose to consider new class generalization as an extreme case of class imbalance and treat visual prototypes as samples to alleviate such imbalance.

On top of this, we introduce our Compensating Logit-Adjusted Loss (CLA Loss) inspired by [29] to handle such imbalanced scenarios. Let z_i denote the predicted logit for class i, then CLA loss takes the form of :

$$\mathcal{L}_{cla}(\boldsymbol{z}, y=j) = -\log \frac{\exp(z_j + \log p(y=j))}{\sum_{k=1}^{K} \exp(z_k + \log p(y=k))} \tag{3}$$

where $p(y=j)$ is the class prior probability. Let n_i denote the number of samples for class i, $p(y=j)$ can be estimated as $n_j / \sum_{i=1}^{K} n_i$. Particularly, we suppose $n_i = 1$ for new classes and treat the corresponding visual prototypes as training samples. In this way, we have found a solution to not only deal with the imbalance within the base classes but also compensate for the imbalance between the base and new classes.

4.2 Feature-Level Cross-Modal Attention

To overcome the efficiency issue, we introduce cross-modal attention to leverage both visual and textual information. To further enhance its efficiency and practicality, we treat the model as a black box following [24] and apply optimizations directly within the feature space.

Our method can divided into the following steps. First, we precompute and save the visual and textual prototypes for each class. Then, the features together with the prototypes are fed into the corresponding linear projection layers P_I and P_T. In this way, we can further align the two modalities for downstream adaptation, and the prediction can be calculated by

$$p(y=i \mid x) = \frac{\exp(\cos(P_I(\boldsymbol{x}), P_T(\boldsymbol{T}_i))/\tau_t)}{\sum_{j=1}^{N} \exp(\cos(P_I(\boldsymbol{x}), P_T(\boldsymbol{T}_j))/\tau_t)} \tag{4}$$

where τ_t is the temperature for image-text matching. However, the scarce data in the downstream tasks still makes it difficult to achieve satisfactory adaptations.

To remedy this issue, we propose to enrich the features from both modalities by leading them to interact with each other through cross-modal attention. Specifically, we concatenate the image features, visual prototypes, and textual prototypes together and then feed them into a self-attention [37] module, considering its ability to establish connections for long-dependency embeddings. Let $\mathrm{Attn}(\cdot)$ denote the multi-head self-attention function, the output is

$$\boldsymbol{x}', \boldsymbol{V}', \boldsymbol{T}' = \mathrm{Attn}([P_I(\boldsymbol{x}), P_I(\boldsymbol{V}), P_T(\boldsymbol{T})]) \tag{5}$$

Note that the operation is permutation invariant, thereby we can slice the output to get the corresponding features and prototypes. After obtaining the enriched features, we can now predict with both visual and textual prototypes by:

$$p_V(y=i \mid x) = \frac{\exp(\cos(\boldsymbol{x}', \boldsymbol{V}'_i)/\tau_v)}{\sum_{j=1}^{K} \exp(\cos(\boldsymbol{x}', \boldsymbol{V}'_j)/\tau_v)} \tag{6}$$

$$p_T(y=i \mid x) = \frac{\exp(\cos(\boldsymbol{x}', \boldsymbol{T}'_i)/\tau_t)}{\sum_{j=1}^{K} \exp(\cos(\boldsymbol{x}', \boldsymbol{T}'_j)/\tau_t)} \tag{7}$$

where τ_v denotes the temperature for visual modality. By ensembling these two results, we are able to leverage both visual and textual information.

4.3 Virtual Prototypes for New Classes

We have discussed the weakness of CLIP and introduced visual prototypes as well as cross-modal attention as a remedial measure. However, another crucial problem emerges regarding new class generalization. As for new classes, we cannot obtain their visual prototypes during training.

To solve this issue, we introduce learnable virtual prototypes for new classes to hold the place of missing visual prototypes. Specifically, we freeze the precomputed textual prototypes as well as the visual prototypes for base classes, while treating the virtual prototypes as the corresponding visual prototypes for new classes, and optimizing them during the training stage. Other than this, the entire procedure is the same as in Section 4.2. Formally, let $\hat{\boldsymbol{V}}'$ denote the transformed virtual prototypes for new classes, then Equation 6 can be rewritten as:

$$p(y=i \mid w) = \frac{\exp(\cos(\boldsymbol{x}', \boldsymbol{V}'_i)/\tau_v)}{\sum \exp(\frac{\cos(\boldsymbol{x}', \boldsymbol{V}')}{\tau_v}) + \sum \exp(\frac{\cos(\boldsymbol{x}', \hat{\boldsymbol{V}}')}{\tau_v})} \tag{8}$$

In this way, the virtual prototypes are guided to learn a suitable representation of the new classes in the feature space and can act as an supplement to the textual prototypes.

4.4 Overall Objective

The overall loss function is given by applying CLA loss to the logits calculated by Equation 4, 6 and 7 and aggregating the results. Suppose the logits given by these equations are z_P, z_V, z_T respectively, then the loss objective $\mathcal{L}$ for optimization during training is:

$$\mathcal{L} = \mathcal{L}_{cla}(z_P, y) + \mathcal{L}_{cla}(z_V, y) + \mathcal{L}_{cla}(z_T, y) \tag{9}$$

During inference, we aggregate the logits obtained after cross-modal attention, namely $z = z_V + z_T$. The entire framework is shown in Figure 2.

5 EXPERIMENTS

5.1 Experimental Settings

We evaluate our approach Candle in the following problem settings: 1) generalization from base to new classes under imbalance and few-shot settings; 2) cross-dataset transfer and 3) domain generalization. For the imbalanced settings, all the training data is generated by down-sampling the base classes to obey an exponential decay of different ratios. Let n_i denote the number of samples in the i-th class, imbalance ratio is defined as $\max\{n_i\}/\min\{n_i\}$. The maximum number of samples per class of the generated dataset is set to either 100 (if has) or the maximum number of samples per class of the original dataset.

Datasets and Evaluation. For new class generalization and cross-dataset transfer, the experiments are conducted over a total of 11 diverse image classification datasets, including ImageNet [5] and Caltech101 [9] for generic object recognition, OxfordPets [26], StanfordCars [17], Flowers102 [23], Food101 [1] and FGVCAircraft [21] for fine-grained image recognition, SUN397 [41] for scene recognition, DTD [4] for texture classification, EuroSAT [12] for satellite image classification, and UCF101 [34] for action recognition. For domain generalization, we use ImageNet as the source dataset and four other variants that exhibit different types of domain shift as the target datasets, including ImageNet-A [14], ImageNetV2 [28], ImageNet-Sketch [38], and ImageNet-R [13].

Details of the 11 datasets used in base-to-new generalization and cross-dataset transfer, and the 4 datasets used during testing for domain generalization, are shown respectively in Table 3 and Table 4. We report mean-class accuracy for the imbalanced settings, which is different from overall accuracy for datasets with imbalanced test sets. The test set for some datasets have varying numbers of sample per class, which is indicated in the rightmost column.

Table 3: Statistics for datasets used in base-to-new generalization and cross-dataset transfer. The rightmost column indicates whether the testing set is balanced.

Dataset	Classes	Train	Test	Balanced
ImageNet [5]	1000	1.28M	50000	✓
Caltech101 [9]	100	4128	2456	✗
OxfordPets [26]	37	2944	3669	✓
StanfordCars [17]	196	6509	8041	✓
Flowers102 [23]	102	4093	2463	✗
Food101 [1]	101	50500	30300	✓
FGVCAircraft [21]	100	3334	3333	✓
SUN397 [41]	397	15880	19850	✓
DTD [4]	47	2820	1692	✓
EuroSAT [12]	10	13500	8100	✗
UCF101 [34]	101	7639	3783	✗

Table 4: Statistics for datasets used during testing for domain generalization. The rightmost column indicates whether the testing set is balanced.

Dataset	Classes	Test	Balanced
ImageNet-A [14]	200	7500	✗
ImageNetV2 [28]	1000	10000	✓
ImageNet-Sketch [38]	1000	50889	✓
ImageNet-R [13]	200	30000	✗

Following the setting in CoCoOp [45], we examine our model on a similar but more practical scenario, where the base training set follows an imbalanced distribution. We also report test results of new class generalization in the balanced few-shot form to show the robustness of our model. Note that for imbalanced scenarios, we report mean-class accuracy instead of overall accuracy.

Baselines. We compare our method to zero-shot CLIP [27], CoOp [46], CoCoOp [45] and LFA [24], which also focuses on feature-level adaptation for CLIP. For the imbalanced settings, our method is compared to CoOp and CoCoOp by switching their loss function to Logit-Adjusted (LA) Loss [29] to ensure fairness. LFA is only compared under the balanced setting because its framework is not compatible to different loss functions.

Implementation Details. We use ViT-B/16 as the vision backbone for all methods for fair comparison. Our models are trained for 10-100 epochs on each dataset and use the SGD optimizer with a batch size of 128, learning rate of 3×10^{-4}, weight decay of 5×10^{-4}, and momentum of 0.9. The temperature parameter τ_t for image-text matching is set to 0.01 following CLIP, whereas the τ_v for image-image matching is decided by searching from {0.005, 0.01, 0.02, 0.05, 0.1} on each dataset. For the baseline methods, the results are generated by following the exact setting as introduced in the original articles. All the experiments are carried out on a single NVIDIA GeForce RTX 3090.

Table 5: Harmonic mean values of base-to-new accuracy (%) of different methods on datasets with imbalance ratios 10, 20, 50. The models are trained on an imbalanced base set and then evaluated on both base and new classes. Harmonic accuracy is calculated by $\frac{2\times base\times new}{base+new}$ to highlight the generalization trade-off. The best results are presented in bold.

(a) Imbalance Ratio = 10.

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	91.78	94.10	69.23	71.86	89.25	32.12	72.15	54.88	54.42	64.74	70.66
CoCoOp + LogitAdjusted Loss	95.09	**96.69**	71.91	77.61	**91.20**	33.71	77.99	65.11	60.28	76.78	75.67
Linear Feature Alignment	95.69	94.09	72.72	84.38	90.44	34.27	78.39	67.43	69.56	82.71	76.97
Candle (Ours)	**95.89**	95.99	**74.30**	**85.03**	90.80	**37.78**	**79.26**	**68.13**	**80.51**	**83.17**	**79.34**

(b) Imbalance Ratio = 20.

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	92.65	94.15	67.39	73.72	86.38	29.33	68.93	55.18	62.64	60.12	70.08
CoCoOp + LogitAdjusted Loss	95.25	**96.64**	71.38	80.31	**91.20**	32.78	77.29	61.31	58.82	71.70	74.48
Linear Feature Alignment	95.56	90.90	70.35	84.03	89.72	33.02	77.30	66.07	68.74	81.80	75.75
Candle (Ours)	**95.84**	95.89	**73.49**	**84.92**	90.75	**38.02**	**78.53**	**67.32**	**80.96**	**82.59**	**79.08**

(c) Imbalance Ratio = 50.

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	93.30	93.34	67.18	75.45	87.65	29.20	65.91	51.42	57.35	61.62	69.92
CoCoOp + LogitAdjusted Loss	94.90	95.44	69.97	76.84	**91.10**	31.45	76.18	59.37	64.99	77.53	74.32
Linear Feature Alignment	94.23	86.76	67.95	82.81	87.73	30.75	75.13	61.78	61.91	79.49	72.85
Candle (Ours)	**94.95**	**95.83**	**71.78**	**84.62**	90.70	**36.68**	**78.05**	**65.69**	**80.17**	**81.72**	**78.20**

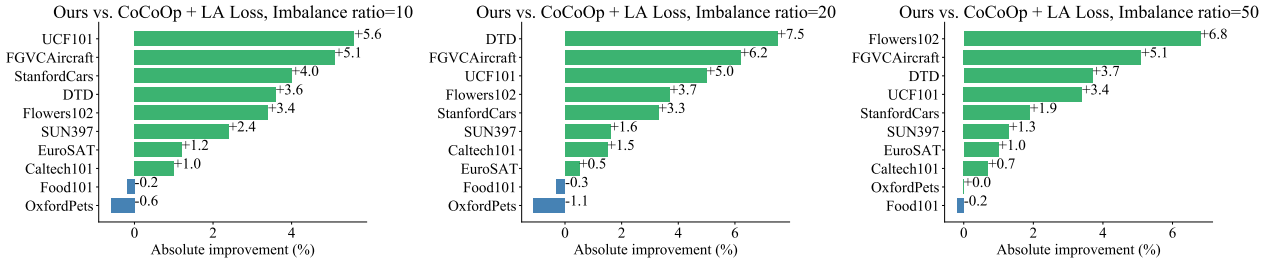

Figure 3: Absolute improvement on the base classes with imbalance ratio 10, 20, 50

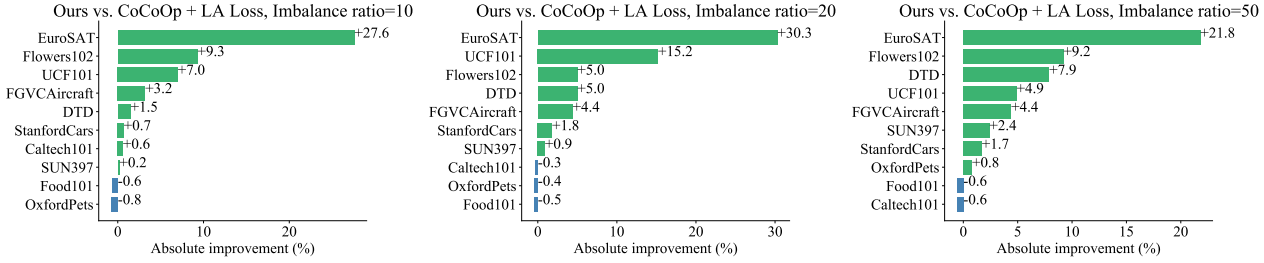

Figure 4: Absolute improvement on the new classes with imbalance ratio 10, 20, 50

5.2 Main Results

Generalization from base to new classes. For generalization from base to new classes, we partition each dataset into two disjoint subsets, namely base classes and new classes. Then, the model is trained on the imbalanced base set and subsequently tested on base and new classes to demonstrate its generalization ability. Table 5 presents the harmonic mean values for the base-to-new setting over imbalance ratios {10, 20, 50}. ImageNet is skipped here due to the extremely high training cost for CoCoOp under this setting. The results show that Candle consistently achieves state-of-the-art results across different imbalance ratios. Specifically, the harmonic mean values of Candle outperform the best previous method by an

Table 6: Comparison of different methods in 16-shot base-to-new generalization. We report the accuracy (%) on both base and new classes, as well as their harmonic mean. The best results are presented in bold.

(a) Average over 11 datasets.

	Base	New	H
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
LFA	83.62	74.56	78.83
Ours	**83.86**	**76.55**	**80.04**

(b) ImageNet.

	Base	New	H
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
CoCoOp	75.98	**70.43**	**73.10**
LFA	76.89	69.36	72.93
Ours	**76.97**	68.54	72.48

(c) Caltech101.

	Base	New	H
CLIP	96.84	94.00	95.40
CoOp	98.00	89.91	93.73
CoCoOp	97.96	93.81	95.84
LFA	98.41	93.93	96.13
Ours	**98.54**	**94.47**	**96.46**

(d) OxfordPets.

	Base	New	H
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	**97.69**	**96.43**
LFA	95.13	96.23	95.68
Ours	**95.53**	97.34	**96.43**

(e) StanfordCars.

	Base	New	H
CLIP	63.37	74.89	68.85
CoOp	78.12	60.40	68.13
CoCoOp	70.49	73.59	72.01
LFA	76.32	74.88	75.59
Ours	**79.14**	**74.92**	**76.97**

(f) Flowers102.

	Base	New	H
CLIP	72.08	**77.80**	74.83
CoOp	97.60	59.67	74.06
CoCoOp	94.87	71.75	81.71
LFA	97.34	75.44	85.00
Ours	**98.01**	77.52	**86.57**

(g) Food101.

	Base	New	H
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	**90.70**	91.29	90.99
LFA	90.52	**91.48**	**91.00**
Ours	90.52	91.23	90.87

(h) FGVCAircraft.

	Base	New	H
CLIP	27.19	36.29	31.09
CoOp	40.44	22.30	28.75
CoCoOp	33.41	23.71	27.74
LFA	41.48	32.29	36.31
Ours	**43.86**	**36.69**	**39.96**

(i) SUN397.

	Base	New	H
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
CoCoOp	79.74	76.86	78.27
LFA	**82.13**	77.20	79.59
Ours	81.64	**77.93**	**79.74**

(j) DTD.

	Base	New	H
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
CoCoOp	77.01	56.00	64.85
LFA	81.29	60.63	69.46
Ours	**81.40**	**61.35**	**69.97**

(k) EuroSAT.

	Base	New	H
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.90
CoCoOp	87.49	60.04	71.21
LFA	**93.40**	71.24	80.83
Ours	89.97	**81.33**	**85.43**

(l) UCF101.

	Base	New	H
CLIP	70.53	77.50	73.85
CoOp	84.69	56.05	67.46
CoCoOp	82.33	73.45	77.64
LFA	86.97	77.48	81.95
Ours	**87.13**	**80.51**	**83.69**

Table 7: Cross-dataset transfer learning accuracy (%) of different methods. The methods are trained on an imbalanced source dataset (ImageNet) and subsequently evaluated on the target datasets. The best results are presented in bold.

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	90.8	87.0	64.9	67.3	85.3	18.8	63.2	42.2	44.4	65.9	63.0
CoCoOp + LogitAdjusted Loss	**91.4**	88.6	**65.6**	**69.4**	**86.3**	23.0	66.0	**45.0**	42.8	**67.5**	64.6
Candle (Ours)	91.3	**88.9**	64.6	68.3	85.5	**24.2**	**66.1**	44.6	**48.4**	67.2	**64.9**

average of 3.67%, 4.60%, and 3.88% under the three imbalance ratios. We also illustrate the absolute improvements of Candle compared to the previous best method in Figure 3 and Figure 4. The results show average improvements of 2.58%, 2.79%, and 2.27% on the base classes and higher average improvements of 4.87%, 6.11%, and 5.19% on the new classes, affirming that it does indeed compensate for new classes. Together, the above results demonstrate the effectiveness of our approach in addressing imbalance both within the base classes and between the base and new classes. We present detailed results for each setting in the appendix due to the page limit.

To examine the robustness of Candle, we also report the results for 16-shot base-to-new generalization in Table 6. In this setting, Candle still achieves an improvement of 1.21% in average harmonic mean over the best previous method. Specifically, it performs comparably with LFA on the base classes (+0.14% by average) but outperforms LFA on the new classes by a large margin (+1.99% by average), thus validating its ability to help with new classes.

In addition, by taking a closer look at the results for each dataset, Candle achieves significant gains on datasets such as Flowers102, FGVCAircraft, EuroSAT, and UCF101. This is in accordance with

Table 8: Domain generalization accuracy (%) of different methods. The methods are trained on an imbalanced source dataset (ImageNet) and subsequently evaluated on the target datasets. The best results are presented in bold.

	IN.	IN-A.	INV2.	IN-S	IN-R.
CoOp + LA Loss	70.7	48.7	**63.5**	47.2	73.8
CoCoOp + LA Loss	71.3	**49.1**	63.3	47.8	74.4
Candle (Ours)	**71.6**	**49.1**	62.8	**48.3**	**75.0**

the analysis in Section 4.2 that CLIP performs poorly on datasets where textual information is relatively unreliable, and our proposed approaches alleviate this issue by leveraging both visual and textual information.

Cross dataset transfer. For cross-dataset transfer, we train the model on an imbalanced ImageNet subset with an imbalance ratio of 100 and subsequently test the model on the other 10 datasets. Table 7 presents the results for cross-dataset transfer. Candle shows similar results compared to CoCoOp with LogitAdjusted Loss across the 10 target datasets, achieving an average improvement of 0.3%. It's worth noting that the baseline methods require much more training time compared to ours. For CoCoOp, 10 epochs of training lasts for 1 day and 6 hours and inference alone takes up 3 hours, while our method only needs about 20 minutes for the whole training process. Nonetheless, our method Candle is able to deliver comparable results while significantly reducing computational cost. Similar to the base-to-new generalization task, performance gains on specific datasets can be observed in this task as well.

Domain generalization. For domain generalization, we train the model on an imbalanced ImageNet subset with an imbalance ratio of 100 and evaluate the model on four domain-shift target datasets. The results are presented in Table 8. Candle achieves improvements over the previous best method on 3 out of 4 target datasets, with an average increase of 0.15% on the target datasets, and an increase of 0.3% on the source dataset. The results demonstrate the robustness of Candle against domain shifts.

5.3 Ablation Study

Impacts of different loss functions. In the main results, we equip the baseline methods with the balanced LA loss for fair comparison. Here, we further examine the robustness of different methods against class imbalance without the assistance of such a tailored loss function. Specifically, we use cross entropy (CE) loss for all the methods and run on the imbalanced base-to-new generalization task. The results are shown in Table 9. It can be observed that CoCoOp suffers from a more severe performance drop without LA loss, i.e., a decrease of 5.78% in average harmonic value. In contrast, our method Candle manages to hold on with a drop of only 1.51%. This indicates that even without the balanced logit-adjusted loss, our model still shows potential strength in dealing with class imbalance.

Effect of cross-modal attention. We conduct ablation study on the imbalanced base-to-new generalization task to examine the effectiveness of the cross-modal attention module. For the sake of simplicity, we examine on the imbalance ratio = 50 setting. The

Table 9: Ablation on different loss functions. Δ indicates the difference in performance for the same method trained with CE loss or LA loss. Our method is the least sensitive to the change in loss function.

	Base	New	Harmonic Mean
CoOp + LA Loss	80.26	61.94	69.92
CoOp + CE Loss	76.12	61.19	67.84
Δ	-4.14%	-0.75%	-2.08%
CoCoOp + LA Loss	77.91	71.05	74.32
CoCoOp + CE Loss	72.05	65.36	68.54
Δ	-5.86%	-5.69%	-5.78%
Candle (Ours)	80.38	76.14	78.20
Candle w/ CE Loss	78.07	75.36	76.69
Δ	-2.31%	-0.78%	-1.51%

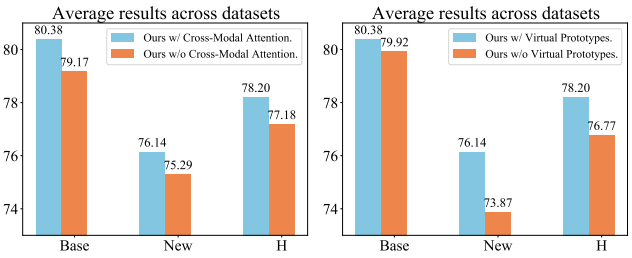

Figure 5: Ablation studies on cross-modal attention (left) and virtual prototypes (right). The experiment is conducted on the imbalanced base-to-new generalization task with an imbalance ratio of 50.

current model is compared to one with only linear projection after the extracted features, with the rest of the settings the same. The results are shown in the left part of Figure 5. Without the cross-modal attention module, the average results on the base and new classes experience a drop of 1.19% and 0.85% respectively, leading to a 1.02% decline of harmonic mean value. These figures clearly show that our cross-modal attention module acts contributes positively and significantly to the model's overall performance.

Effect of virtual prototypes. We further examine the effectiveness of the virtual prototypes. Since the removal of virtual prototypes renders the image-image matching for new classes unachievable, the model in comparison can only leverage image-text matching on the new classes. We conduct comparison experiments and report the results in the right part of Figure 5. The results show that, the performance gap on the base classes is relatively small, with our proposed model holding an average advantage of 0.46%. However, the performance on the new classes drops remarkably in response to the removal of virtual prototypes, showing an average decline of 2.27% across different datasets. The results prove that the introduction of virtual prototypes significantly helps with new class generalization.

6 CONCLUSION

In this paper, we aim to address the new class generalization problem for vision-language models under more practical scenarios, where the data may exhibit a long-tailed distribution. We propose a novel and simple framework named Candle to solve this issue in an efficient manner. Candle achieves state-of-the-art performance over extensive experiments on diverse image classification datasets, with an especially strong generalization on the new classes. Just as significantly, the proposed framework directly optimizes in the feature space and does not need access to model weights, which also contributes to its economical training cost compared to past methods. We hope our work serves as an inspiration for further advances in exploring efficient and long-tailed generalization for vision-language models.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation of China (62176118).

REFERENCES

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – Mining Discriminative Components with Random Forests. In *ECCV*.
[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.
[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *NeurIPS*.
[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing Textures in the Wild. In *CVPR*.
[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.
[6] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. 2022. Don't Stop Learning: Towards Continual Learning for the CLIP Model. *arXiv preprint arXiv:2207.09248* (2022).
[7] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. 2023. LPT: Long-tailed Prompt Tuning for Image Classification. In *ICLR*.
[8] 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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.
[9] Li Fei-Fei, R. Fergus, and P. Perona. 2004. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In *CVPR Workshops*.
[10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *arXiv preprint arXiv:2110.04544* (2021).
[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
[12] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. 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* (2019).
[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*.
[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural Adversarial Examples. In *CVPR*.
[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.
[16] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *CVPR*.
[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3D Object Representations for Fine-Grained Categorization. In *ICCV Workshops*.
[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.
[19] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. 2019. Large-Scale Long-Tailed Recognition in an Open World. In *CVPR*.
[20] Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2021. A Simple Long-Tailed Recognition Baseline via Vision-Language Model. *arXiv preprint arXiv:2111.14745* (2021).
[21] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-Grained Visual Classification of Aircraft. *arXiv preprint arXiv: 1306.5151* (2013).
[22] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. 2023. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*.
[23] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated Flower Classification over a Large Number of Classes. In *ICVGIP*.
[24] Yassine Ouali, Adrian Bulat, Brais Matinez, and Georgios Tzimiropoulos. 2023. Black Box Few-Shot Adaptation for Vision-Language Models. In *ICCV*.
[25] Jishnu Jaykumar P, Kamalesh Palanisamy, Yu-Wei Chao, Xinya Du, and Yu Xiang. 2023. Proto-CLIP: Vision-Language Prototypical Network for Few-Shot Learning. *arXiv preprint arXiv: 2306.15955* (2023).
[26] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *CVPR*.
[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
[28] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet?. In *ICML*.
[29] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. 2020. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*.
[30] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.
[31] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Workshop Datacentric AI*.
[32] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Xin-Yan Han, Jie-Jing Shao, and Yu-Feng Li. 2023. Parameter-Efficient Long-Tailed Recognition. *arXiv preprint arXiv:2309.10019* (2023).
[33] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. 2022. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190* (2022).
[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv: 1212.0402* (2012).
[35] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. 2023. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. In *CVPR*.
[36] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. 2022. VL-LTR: Learning Class-wise Visual-Linguistic Representation for Long-Tailed Visual Recognition. In *ECCV*.
[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*.
[38] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. In *NeurIPS*.
[39] Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang. 2023. Exploring Vision-Language Models for Imbalanced Learning. *IJCV* (2023).
[40] Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, and Tieniu Tan. 2023. Improving Zero-Shot Generalization for CLIP with Synthesized Prompts. In *ICCV*.
[41] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*.
[42] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).
[43] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. *arXiv preprint arXiv:2111.03930* (2021).
[44] Da-Wei Zhou, Yuanhan Zhang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. 2023. Learning without Forgetting for Vision-Language Models. *arXiv preprint arXiv: 2305.19270* (2023).
[45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional Prompt Learning for Vision-Language Models. In *CVPR*.
[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to Prompt for Vision-Language Models. *IJCV* (2022).

Table 10: Base-to-new generalization results of different methods with imbalance ratios 10, 20, 50. The best results are in bold.

(a) Imbalance Ratio = 10, base classes

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	96.5	94.7	**75.8**	**98.7**	89.4	39.0	74.5	**81.0**	**94.3**	83.9	82.8
CoCoOp + LogitAdjusted Loss	94.5	**95.6**	70.2	94.6	**90.7**	34.9	78.7	75.2	87.8	81.6	80.5
Candle (Ours)	**96.9**	95.0	74.2	98.0	90.5	**40.0**	**81.1**	80.8	89.0	**87.2**	**83.3**

(b) Imbalance Ratio = 10, new classes

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	87.5	93.5	63.7	56.5	89.1	27.3	59.1	41.5	41.2	57.0	61.6
CoCoOp + LogitAdjusted Loss	94.3	**97.8**	73.7	65.8	**91.7**	32.6	77.3	57.4	45.9	72.5	70.9
Candle (Ours)	**94.9**	97.0	**74.4**	**75.1**	91.1	**35.8**	**77.5**	**58.9**	**73.5**	**79.5**	**75.8**

(c) Imbalance Ratio = 20, base classes

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	**96.5**	**93.9**	73.1	**98.5**	88.9	37.0	77.6	77.0	**93.1**	83.7	81.9
CoCoOp + LogitAdjusted Loss	95.2	**95.9**	69.2	94.0	**90.8**	33.6	78.2	71.6	88.3	80.7	79.8
Candle (Ours)	**96.7**	94.8	72.5	97.7	90.5	**39.8**	**79.8**	**79.1**	90.5	**85.7**	**82.5**

(d) Imbalance Ratio = 20, new classes

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	89.1	94.4	62.5	58.9	84.0	24.3	62.0	43.0	47.2	46.9	61.2
CoCoOp + LogitAdjusted Loss	**95.3**	**97.4**	73.7	70.1	**91.6**	32.0	76.4	53.6	44.1	64.5	69.9
Candle (Ours)	95.0	97.0	**74.5**	**75.1**	91.0	**36.4**	**77.3**	**58.6**	**74.4**	**79.7**	**75.9**

(e) Imbalance Ratio = 50, base classes

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	93.8	92.3	**70.7**	**96.6**	88.0	35.3	**79.7**	**72.0**	**93.4**	80.8	80.3
CoCoOp + LogitAdjusted Loss	94.5	**94.6**	67.1	88.5	**90.5**	31.5	77.3	68.3	86.9	79.9	77.9
Candle (Ours)	**95.2**	**94.6**	69.0	**95.3**	90.3	**37.6**	78.6	**72.0**	87.9	**83.3**	**80.4**

(f) Imbalance Ratio = 50, new classes

	Cal.	OP.	SC.	FLw.	Food.	FA.	SUN.	DTD.	ES.	UCF.	Avg.
CoOp + LogitAdjusted Loss	92.8	94.4	64.0	61.9	87.3	24.9	65.9	40.0	38.4	49.8	61.9
CoCoOp + LogitAdjusted Loss	**95.3**	96.3	73.1	67.9	**91.7**	31.4	75.1	52.5	51.9	75.3	71.0
Candle (Ours)	94.7	**97.1**	**74.8**	**76.1**	91.1	**35.8**	**77.5**	**60.4**	**73.7**	**80.2**	**76.1**

A ADDITIONAL RESULTS

Generating imbalanced datasets. Following the method proposed by Cao et al. [3], we generate imbalanced versions from the original datasets to obey an exponential decay of a given ratio. Let n_i denote the number of samples in the i-th class, the imbalance ratio is defined as $\max\{n_i\}/\min\{n_i\}$. From Table 3, we can see that the training set for some datasets only has around 30 samples per class (FGVCAircraft) whereas some has over 1000 (ImageNet). Therefore, we set the maximum number of samples per class of the generated dataset to be either 100 (if has) or the maximum number of samples per class of the original dataset, to guarantee enough data to form a valid imbalanced distribution. Additionally, in cases where the maximum number of samples per class is lower than the imbalance ratio, we ensure there is at least 1 sample instead of 0 for the tail classes.

Imbalanced base-to-new generalization details. In Table 10, we show the full results of imbalanced base-to-new generalization, including the accuracy of different methods on base and new classes, under imbalance ratios $\{10, 20, 50\}$. On the base classes, our method exhibits a slight edge over CoOp+LA Loss with an increase of 0.3% averaging across different ratios and shows a clearer improvement (2.7%) over CoCoOp+LA Loss, which is the previous best method in harmonic mean value. On the new classes, our method far outperforms CoOp+LA Loss with an advantage of 14.4% and still leads CoCoOp+LA Loss by 5.3%. This again demonstrates that our method compensates significantly for the new classes while preserving a strong performance on the base classes.

Table 11: Results of different attention strategies compared to the original method under the imbalanced base-to-new generalization setting with imbalance ratio 50. Δ indicates the average difference in accuracy across different datasets.

Mask Type	Δ, Base Classes	Δ, New Classes
Within Visual.	-0.26%	-0.49%
Within Text.	-0.19%	-0.78%
Between Visual & Text.	-0.74%	-0.90%

Different attention strategies. As mentioned in the article, we perform cross-modal attention by concatenating image features, visual prototypes and textual prototypes together, and then feed them into a self-attention module. Here, we provide analysis of different attention strategies by adding different input masks. For the sake of simplicity, we consider the concatenated inputs to be comprised of two parts, the visual part (image features + visual prototypes) and the texual part (texual prototypes). Hence, there are three different kinds of masks to choose from, including masking attention within each part and between each part. Table 11 shows the results of comparing different attention strategies with the original method (no mask at all). The consistent decline proves the superiority of the original design. Particularly, the removal of attention between visual and textual part leads to the largest drop on both base and new classes, which goes to show that the interaction between different modalities does contribute to improving the model's performance.