ProTeCt: Prompt Tuning for Taxonomic Open Set Classification

Tz-Ying Wu* Chih-Hui Ho* Nuno Vasconcelos
University of California, San Diego
{tzw001, chh279, nvasconcelos}@ucsd.edu

Abstract

Visual-language foundation models, like CLIP, learn generalized representations that enable zero-shot open-set classification. Few-shot adaptation methods, based on prompt tuning, have been shown to further improve performance on downstream datasets. However, these methods do not fare well in the taxonomic open set *(TOS) setting, where the classifier is asked to make prediction from label set across different levels of semantic granularity. Frequently, they infer incorrect labels at coarser taxonomic class levels, even when the inference at the leaf level (original class labels) is correct. To address this problem, we propose a prompt tuning technique that calibrates the hierarchical consistency of model predictions. A set of metrics of hierarchical consistency, the Hierarchical Consistent Accuracy (HCA) and the Mean Treecut Accuracy (MTA), are first proposed to evaluate TOS model performance. A new* Prompt Tuning for Hierarchical Consistency *(ProTeCt) technique is then proposed to calibrate classification across label set granularities. Results show that ProTeCt can be combined with existing prompt tuning methods to significantly improve TOS classification without degrading the leaf level classification performance. The code is available at* https://github.com/gina9726/ProTeCt.

1. Introduction

Vision-language foundation models (FMs) have opened up new possibilities for image classification. They are large models, trained on large corpora, to learn aligned representations of images and text. For example, CLIP [30] combines text and image encoders trained with 400M image-text pairs in an open vocabulary fashion, using a contrastive loss [3, 4, 34, 35]. Zero-shot classification can then proceed by leveraging the feature alignments. Each class name is first converted to a text prompt, e.g., "a photo of [CLASS]," which is fed to the text encoder. The resulting text feature is then used as the parameter vector of a softmax classifier of image feature vectors. Since the training does not emphasize any particular classes, CLIP supports open set classification. Several works [17, 42, 45, 46] have shown that classification accuracy can be enhanced by fine-tuning the FM on the few-shot setting (i.e. few examples per class). To adapt the model and maintain image-text alignment, these works augment the FM with a few learnable prompts [17, 42, 45, 46]. The model parameters are then frozen and only the prompts are optimized. This process is known as **prompt tuning** and can outperform zero-shot performance, on the dataset of interest.

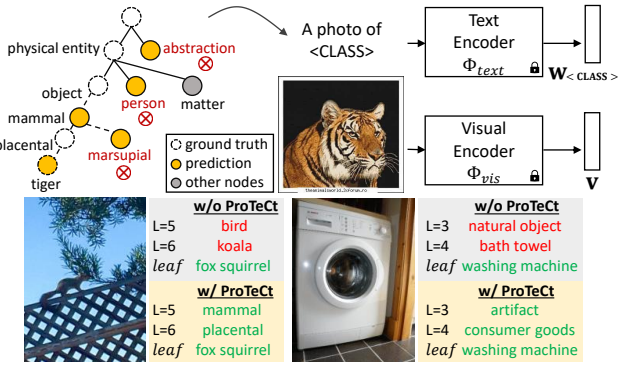


Figure 1. (Top) An example of class hierarchy, where CLIP predicts the tiger image as "person" at the internal hierarchy level. (Bottom) Correct/incorrect model predictions (green/red) of CoOp w/ and w/o ProTeCt on ImageNet variants. L denotes the tree level.

Method	Acc_{leaf}	HCA	MTA
CLIP [30]	68.36	3.32	48.21
CoOp [46]	71.23	2.99	46.98
MaPLe [17]	70.70	4.15	48.29

Table 1. TOS classification performance of CLIP-based classifiers.

While prompting enables classifiers to be designed for virtually any classes with minimal dataset curation effort, it should not compromise the open set nature and generality of the FM representation. In this work, we consider the setting where "open set" means the ability to refer to concepts at different levels of granularity. Consider, for example, an educational application in biology. While at grade school level it will teach students to classify animals into ("cat", "dog", "lizard"), at the high-school level the **exact same images** should be classified into much more detailed classes, e.g. ("iguana", "anole", "komodo", etc.) for lizards. A classifier that classifies an image as a "komodo" lizard for high schoolers but "dog" for gradeschoolers is not useful and trustworthy. Advanced biology students should

even learn about the taxonomic relations between different species. This requires a representation that supports hierarchical classification [20, 28, 39, 41], where the classifier understands the relations between the superclasses and subclasses that compose a class hierarchy, and provides correct predictions *across* hierarchy levels.

Fig. 1 shows an example hierarchy built from ImageNet [5] classes, according to the WordNet [11]. When faced with a tiger image, the classifier should provide a correct prediction under the label sets $\mathcal{Y}_1$ =("dog", "cat", "**tiger**"), $\mathcal{Y}_2$ = ("person", "**animal**", "insect") or $\mathcal{Y}_3$ =("**physical entity**", "abstraction"), where the correct one is shown in bold. Note that, given a classifier with this property, teachers have the ability to define different classification problems, for many levels of granularity, tailoring the same app to different uses. We refer to this setting as **taxonomic open set** (TOS) classification. In many real-world applications, support for this restricted form of open set classification is much more important than support for unbounded open set classification. In the example above, biology teachers do not really care if the classifier can still discriminate between cars and trucks, or soda cans and wine cups. Hence, these classes are irrelevant to the app developer.

In principle, TOS should be trivially supported by FMs. Even at zero-shot level, it should suffice to specify [CLASS] names at the desired levels of granularity. However, our experiments show that this does not work because the representation of most FMs fails to capture taxonmic relations. This is illustrated for CLIP in Fig. 1. While the model knows that the object is a tiger, it fails to know that it is "a physical entity" and not an "abstraction" or that it is a "placental mammal" and not a "marsupial," indicating that it only understands class relations locally. It can perform well for the leaf class label set $\mathcal{Y}_1$, but cannot reason across abstraction levels, and can thus not support TOS classification. To enable TOS, we introduce the notion of **hierarchical consistency**, and a new *hierarchical consistency accuracy* (HCA) metric, where classification is defined with respect to a taxonomic tree and its success requires the correct prediction of all superclasses (e.g., mammal, object and physical entity) of each ground truth leave class (e.g., tiger). This is complemented by the notion of **TOS classification**, where classifiers can have any set of nodes in the class hierarchy as the label set, and a new *mean treecut accuracy* (MTA) metric, which estimates classification accuracy in this setting.

Our experiments show that neither CLIP nor existing prompt tuning methods [17, 45, 46] perform well under the HCA and MTA metrics of the TOS setting. Fig. 1 illustrates the problem and the *inconsistent* CLIP class predictions (orange dots) across hierarchy levels. Table 1 compares the standard (leaf) accuracy of the model with HCA/MTA, under both the zero-shot and two prompt-tuning settings. While the leaf accuracy is quite reasonable, hierarchical consistency is very poor. To address this problem, we propose a novel prompt-tuning procedure, denoted *Prompt Tuning for Hierarchical Consistency* (ProTeCt), that explicitly targets the TOS setting. Given a dataset of interest, a class hierarchy is extracted from the associated metadata, a generic public taxonomy (e.g. WordNet [11]), or a special purpose taxonomy related to the application (e.g. scientific taxonomies). Since FMs support classification with open vocabulary, any node in the hierarchy can be used in the label set of the classifier. Prompts are then learned with the help of two new regularization losses that encourage hierarchical consistency. A *dynamic treecut loss* (DTL) encourages correct classification at all tree levels by sampling random tree cuts during training. A *node-centric loss* (NCL) contributes additional supervision to each internal tree node to increase classification robustness for all granularities of the hierarchy.

Experiments show that ProTeCt significantly improves the performance of prompt tuning methods, like CoOp [46] and MaPLe [17], under TOS setting. Fig. 1 shows the predictions of CoOp at different hierarchy levels before/after adding ProTeCt. Under the HCA/MTA metrics, the improvement can be more than 15/25 points on Cifar100, SUN and ImageNet datasets. Following [17, 45, 46], we show that these gains hold for zero-shot domain generalization to several variants of ImageNet [14, 15, 31, 36], showing that hierarchical consistency transfers across datasets. Furthermore, ablations show that ProTeCt can be used with different CLIP architectures, parameter tuning methods and taxonomies.

Overall, this work makes four contributions. First, we introduce the TOS setting, including two novel metrics (HCA and MTA) that evaluate the consistency of hierarchical classification. Second, we show that neither zero-shot CLIP nor existing prompting methods fare well in this setting. Third, we propose a novel prompt-tuning method for the TOS setting, ProTeCt, which improves hierarchical consistency by combining DTL and NCL losses. The former relies on a dynamic stochastic sampling of label sets involving multiple levels of the hierarchy, while the latter regularizes the classification of every node in the hierarchy. Finally, ProTeCt is shown to outperform vanilla prompt tuning methods on three datasets with different hierarchies. Extensive ablations demonstrate that ProTeCt is applicable to different parameter tuning methods, CLIP architectures, taxonomies and the learned hierarchical consistency transfers to unseen datasets from different image domains.

2. Related Work

Prompt Tuning of Vision-Language Models. Many large vision-language FMs have been proposed recently [10, 37, 43]. Despite their promising zero-shot performance, several works [16, 17, 45, 46] have shown that their few-shot finetuning with a dataset from the target application can further improve performance. Unlike conventional finetuning methods that optimize the entire model, these methods are

designed to (a) be parameter efficient and (b) maintain the general purpose feature representation of the FM. Several such tuning methods have been proposed for CLIP [30]. Inspired by prompt tuning techniques from the language literature [21–23], CoOp [46] inserts learnable prompts at the CLIP text input. CoCoOp [45] further learns a meta-network to generate an image-conditioned prompt. The idea of connecting image and text prompts is further extended by UPT [42] and MaPLe [17]. The former learns a unified transformer for generating an image and text prompt, the latter learns a coupling function to generate image prompts from text prompts. LASP [2] proposed a text-to-text cross-entropy loss to regularize the distribution shift when different prompts are used. Unlike these works, we investigate the TOS problem, where labels can be drawn from any level in a class taxonomy, and propose prompting techniques to improve hierarchical classification consistency. This is shown to be compatible with several of the above prompt-tuning methods without degrading their leaf classification accuracy.

Hierachical Classifiers. Hierarchical classification aims to predict labels at different levels of a class hierarchy. Early works [6, 7, 28, 32, 33, 44] date back to the era before deep learning and are not directly applicable to deep learning-based models. Several works [1, 13, 18, 24, 41, 47] propose hierarchical classifiers for CNN-based deep models. For example, [13, 24, 47] use additional convolutional modules to learn a hierarchical feature space. It is unclear how these approaches generalize to the recent transformer-based architectures [8, 25, 26]. Furthermore, prior works [1, 13, 24, 39, 41, 47] finetune the entire model, which requires substantial data and computation, especially at the FM scale. In this work, we study the problem of hierarchical consistency for foundational vision-language models (e.g., CLIP). While CLIP-based classifiers [30, 45, 46] have outstanding zero/few-shot performance, we show that they produce inconsistent predictions for label sets of different granularity and cannot be used in the TOS setting. We propose an efficient prompt tuning method to address this.

3. Preliminaries

Foundation Models (FMs). Visual-language FMs are composed by a text Φ_{text} and a visual Φ_{vis} encoder, which extract features from text and images, respectively. The two encoders are optimized by contrastive training [3, 4, 34, 35] to create a joint representation for the two modalities. Since the encoders are learned from a large-scale corpus of image-text pairs, the features are general and support various downstream tasks, e.g., image classification [17, 42, 45, 46] and segmentation [27, 38]. While in this work we use the CLIP [30], ProTeCt should generalize to other FMs.

Image Classification with FMs. Given a label set $\mathcal{Y} = \{t_y\}_{y=1}^{C}$, a zero-shot classifier can be designed in the FM representation space by introducing a weight vector $\mathbf{w}_y$ per class y. These weight vectors are obtained by simply using the class name t_y (e.g., “dog") as a text encoder prompt, i.e., $\mathbf{w}_y = \Phi_{text}(Emb_t(t_y)) \in \mathbb{R}^k$, where $Emb_t(\cdot)$ is a word embedding. Given these weight vectors, an image classifier of label set $\mathcal{Y}$ can be implemented by computing class posterior probabilities with

$$p(t_y|\mathbf{x};\mathcal{Y}) = \frac{\exp\left(\cos(\mathbf{w}_y, \mathbf{v})/\tau\right)}{\sum_{t_j \in \mathcal{Y}} \exp\left(\cos(\mathbf{w}_j, \mathbf{v})/\tau\right)}, \quad (1)$$

where $p(t_y|\mathbf{x};\mathcal{Y})$ is the probability of class label t_y given image $\mathbf{x}$, $\mathbf{v} = \Phi_{vis}(Emb_v(\mathbf{x})) \in \mathbb{R}^k$ the visual feature vector, $Emb_v(\cdot)$ an image embedding, $cos(\cdot,\cdot)$ the cosine similarity metric, and τ a temperature hyperparameter. Classification performance can usually be improved by inferring the classifier parameters $\mathbf{w}_y$ from multiple text prompts, e.g. by including context words such as a prompt prefix p =“a photo of", or p =“a drawing of", computing $\mathbf{w}_y = \Phi_{text}(Emb_t(\{p, t_y\}))$, and ensembling the vectors $\mathbf{w}_y$ obtained from multiple prompts [30, 46]. This, however, requires multiple forward passes through Φ_{text} during inference and can be undesirable for downstream applications.

More efficient inference can be achieved with prompt tuning [17, 42, 45, 46], which leverages a set of learnable parameters $\{\mathbf{c}_m^t\}_{m=1}^{M}$ as context features. These are prepended to each class name embedding $Emb_t(t_y)$ as text prompts, to produce the weight vectors $\mathbf{w}_y = \Phi_{text}(\{\mathbf{c}_1^t, \dots \mathbf{c}_M^t, Emb_t(t_y)\})$. Note that each $\mathbf{c}_i^t$ has the same dimension as the word embedding. Given a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, context features can be end-to-end optimized with the cross-entropy loss

$$L_{\mathcal{Y}}(\mathbf{C}^t) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t_j \in \mathcal{Y}} -\mathbb{1}(t_j = t_{y_i}) \log p(t_j|\mathbf{x}_i;\mathcal{Y},\mathbf{C}^t) \quad (2)$$

for the classifier of (1), where $\mathbb{1}(\cdot)$ is the indicator function, and $\mathbf{C}^t$ the matrix of context features. Similarly, learnable prompts $\mathbf{c}_i^v$ can be inserted into the image branch, i.e. $\mathbf{v} = \Phi_{vis}(\{\mathbf{c}_1^v, \dots, \mathbf{c}_M^v, Emb_v(\mathbf{x})\})$, for better visual adaptation [16, 17, 42]. To prevent compromising the generalization of the FM embeddings, the parameters of the two encoders (i.e., Φ_{text}, Φ_{vis}) are frozen in the few-shot setting. In this paper, we consider two prompt tuning variants, CoOp [46] and MaPLe [17], the former using learnable prompts in the text branch, and the latter on both branches.

Class Taxonomy. A class taxonomy $\mathcal{Y}^{tax}$ organizes classes into a tree where classes of similar semantics are recursively assembled into superclasses, at each graph node (e.g. “dog" is a superclass of “Chihuahua" and “Corgi"). For a tree hierarchy, $\mathcal{T}$, each node $n \in \mathcal{N}$ has a single parent and multiple child nodes $Chd(n)$, where $\mathcal{N}$ is the set of tree nodes. Given a set of classes $\{t_y\}_{y=1}^{C}$, a tree hierarchy

$\mathcal{T}$ can be built by treating $\{t_y\}_{y=1}^{C}$ as leaf nodes (where $Chd(t_y) = \emptyset$), i.e., $Leaf(\mathcal{T}) = \{t_y\}_{y=1}^{C}$, and recursively grouping classes in a bottom-up manner until a single root node is created, according to the similarity relationships defined by the taxonomy $\mathcal{Y}^{tax}$. For example, ImageNet [5] classes are organized into a tree of 1,000 leaf nodes derived from the WordNet [11] taxonomy. Nodes that are not at the leaves are denoted as internal nodes $\mathcal{N}^{int} = \mathcal{N} \setminus Leaf(\mathcal{T})$.

4. Taxonomic Open Set Classification

Definition. A significant advantage of FMs for practical applications is their support for open set classification. Since the classifier of (1) can be implemented with any class names t_y, and the FM is trained with an open vocabulary, it is possible to perform classification for virtually any class. Prompting methods improve the classification of the classes defined by the label set $\mathcal{Y}$, but attempt to maintain this generality. However, for most applications "open set" does not mean the ability to recognize "any possible word." On the contrary, the whole point of prompt tuning is to enhance the FM for a given application *context*. This context defines what "open set" truly means for the application. In practice, it frequently means "all the possible ways" to refer to the classes in $\mathcal{Y}$.

One important component of this requirement is the ability to describe classes at different levels of granularity. For example, while user A (a car mechanic) may need to know if an image depicts a "Fan Clutch Wrench" or a "Box-Ended Wrench," user B (a retail store worker) may need to know if the exact same image depicts a "a mechanic's tool" or a "plumber's tool." A FM-based classification app should be deployable in both the car garage or the retail store. However, because the app is a tool classification app, the prompted model does not need to be good at recognizing "lollipops," which are beyond the context of the app. On the other hand, it is undesirable to have to prompt-tune the app for every specific use or user group. Ideally, it should be possible to prompt tune the FM *once,* with respect to the *entire* class taxonomy $\mathcal{Y}^{tax}$ of tools. The app can then be deployed to each user base without any retraining, by simply drawing the most suitable class names t_y from $\mathcal{Y}^{tax}$. We refer to this problem as **Taxonomic Open Set** (TOS) classification and introduce a formal definition in the remainder of this section.

Datasets. Most existing classification dataset can be used to study the TOS problem, since the very nature of taxonomies is to group objects or concepts into semantic classes of different levels of granularity. Hence, most vision datasets are already labeled taxonomically or adopt classes defined by a public taxonomy, usually WordNet [11]. We consider three popular datasets: Cifar100 [19], SUN [40] and ImageNet [5]. ImageNet is complemented by the ImageNetv2 [31], ImageNet-S [36], ImageNet-A [15] and ImageNet-R [14] to enable the study of generalization across image domains. For each dataset, the K-shot setting is considered, where K images per class are sampled for training. We consider $K = \{1, 2, 4, 8, 16\}$.

Label sets. Given a dataset $\mathcal{D}$ and class hierarchy $\mathcal{Y}^{tax}$ a label set $\mathcal{Y}$ is defined at each level of granularity, according the latter. The leaf label set $\mathcal{Y}_{leaf}$ is defined as the set of classes of $\mathcal{D}$ and the class hierarchy $\mathcal{T}$ is build recursively, denoting by $\mathcal{Y}_n = Chd(n)$ the set of class labels for the children of node n. In our experiments, we adopt the default hierarchy of the SUN dataset and use WordNet [11] to build the hierarchy for Cifar100 and ImageNet. The resulting class hierarchies are as follows. Cifar100 [19] contains 100 leaf nodes and 48 internal nodes. SUN contains 324 leaf nodes and 19 internal nodes (after pruning 73 leaf classes that have confusing superclasses). ImageNet [5], ImageNetv2 [31] and ImageNet-S [36] share a class hierarchy of 1,000 leaf nodes and 368 internal nodes. ImageNet-A [15] and ImageNet-R [14] only contain 200 subclasses and the corresponding internal nodes from the ImageNet hierarchy.

Metrics: Given a classifier

$$\hat{y}(\mathbf{x}; \mathcal{Y}) = \arg\max_{t_y \in \mathcal{Y}} p(t_y|\mathbf{x}; \mathcal{Y}) \tag{3}$$

using a label set $\mathcal{Y}$, several metrics are proposed to evaluate TOS performance.

Leaf Accuracy is defined as

$$Acc_{leaf} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\hat{y}(\mathbf{x}_i; \mathcal{Y}_{leaf}) = t_{y_i}] \tag{4}$$

and measures the classification accuracy at the leaves of the taxonomic tree (usually defined as the "dataset classes"). This enables comparison of hierarchical classifiers to standard, or *flat*, classifiers which only consider the leaf classes.

Hierarchical Consistent Accuracy (HCA) is defined as

$$HCA = \frac{1}{N} \sum_{i=1}^{N} (\mathbb{1}[\hat{y}(\mathbf{x}_i; \mathcal{Y}_{leaf}) = t_{y_i}] \prod_{n \in \mathcal{A}(t_{y_i})} \mathbb{1}[\hat{y}(\mathbf{x}_i; \mathcal{Y}_n) \in \mathcal{A}(t_{y_i}) \cup \{t_{y_i}\}]), \tag{5}$$

where $\mathcal{A}(n)$ denotes all the ancestors of node n, and t_{y_i} is the leaf node corresponding to class label y_i. While Acc_{leaf} considers successful any correct classification at the leaf level of the tree, the *HCA* is stricter. It declares a success only when all the ancestors of the leaf node are correctly classified. In other words, each sample needs to be classified correctly at each tree level to be viewed as correctly classified under the *HCA*. Acc_{leaf} is an upper bound for the *HCA*.

Mean Treecut Accuracy (MTA) estimates the expected accuracy under the TOS classification setting. It computes the average accuracy over a set of treecuts $\mathcal{T}_c \in \Omega$,

$$MTA = \frac{1}{|\Omega|} \sum_{\mathcal{T}_c \in \Omega} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\hat{y}(\mathbf{x}_i; \mathcal{Y}_{\mathcal{T}_c}) = t_{y_i}] \,, \tag{6}$$

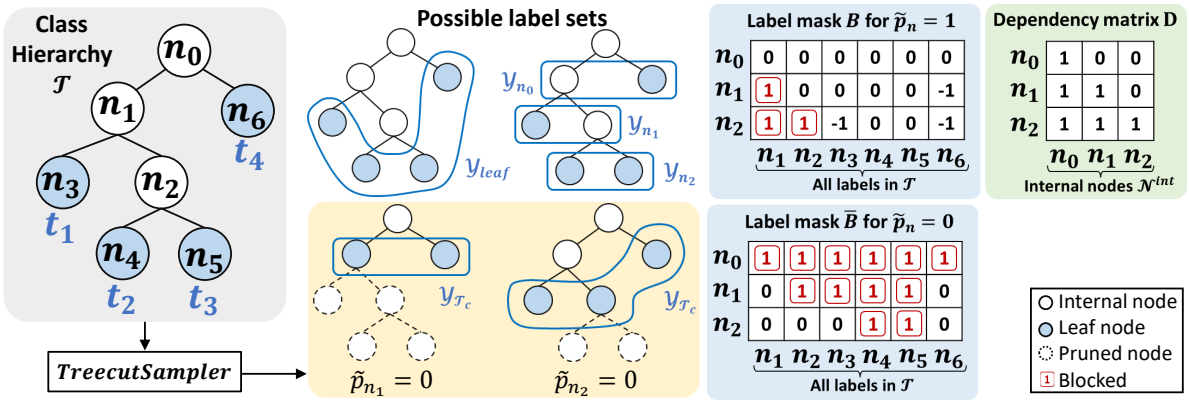


Figure 2. (Left) Multiple possible label sets are available in a class hierarchy. The label set can cover nodes at same level or across different hierarchy levels. (Right) Predefined matrices for efficient treecut sampling used in Algorithm 1.

where $\mathcal{Y}_{\mathcal{T}_c} = Leaf(\mathcal{T}_c)$. However, as shown by the following lemma (see appendix for proof), the set of all possible tree cuts in the hierarchy $\mathcal{T}$ is usually very large.

Lemma 4.1. *For a balanced M-ary tree with depth L (root node is excluded and is at depth 0), the number of all valid treecut is* $L + \sum_{l=2}^{L} \sum_{k=1}^{N-1} \frac{N!}{k!(N-k)!}|_{N=M^{l-1}}$.

For example, a tree with $M = 2$ and $L = 6$ has more than 4 billion treecuts. For a dataset like ImageNet ($L = 15$), this number is monumental. Thus, we randomly sampled $|\Omega| = 25$ treecuts from $\mathcal{T}$ in all experiments and showed that it is already fairly stable.

State-of-the-art. To test TOS performance of the CLIP with existing prompting techniques, we performed an experiment on ImageNet. Table 1 summarizes the performance of the different methods under the three metrics. Two conclusions are possible. First, the sharp drop from Acc_{leaf} to HCA shows that none of the methods make consistent predictions across the class hierarchy. Second, the low MTAs show that the expected accuracy of TOS classification is dramatically smaller than that of flat classification (leaf classes).

5. Prompt Tuning for Hierarchical Consistency

To enhance TOS performance of FMs, we propose *Prompt Tuning for Hierarchical Consistency* (ProTeCt). ProTeCt can be implemented with many existing prompt tuning methods (e.g., CoOp, MaPLe). These methods optimize context prompts using the cross-entropy loss of (2) with leaf label set $\mathcal{Y}_{leaf}$. While this optimizes leaf accuracy Acc_{leaf}, it is not robust to label set changes, even for label sets comprised of superclasses of $\mathcal{Y}_{leaf}$. A simple generalization would be to replace (2) with $\mathcal{L}(\mathbf{C}^t) = \sum_{\mathcal{Y}_p \in \mathcal{T}} L_{\mathcal{Y}_p}(\mathbf{C}^t)$, i.e., to consider all the partial label sets $\mathcal{Y}_p$ of the tree $\mathcal{T}$. However, for sizeable taxonomies, this involves a very large number of label sets and is not feasible. ProTeCt avoids the problem by dynamically sampling label sets from $\mathcal{T}$ during training, with a combination of two learning objectives, a *node-centric loss* (NCL) and a *dynamic tree-cut loss* (DTL).

Node-Centric Loss (NCL). NCL is the aggregate cross-entropy loss of (2) over all node-centric label sets $\mathcal{Y}_n = Chd(n)$ defined by each internal node $n \in \mathcal{N}^{int}$ of the hierarchy, i.e.,

$$\mathcal{L}_{NCL}(\mathbf{C}^t) = \frac{1}{|\mathcal{N}^{int}|} \sum_{n \in \mathcal{N}^{int}} L_{\mathcal{Y}_n}(\mathbf{C}^t) \,. \tag{7}$$

NCL optimization encourages prompts that robustify the classification at different granularities. For example, "Corgi" should be classified as "mammal" within the animal label set $\mathcal{Y}_{n_1}$ ={mammal, reptile, bird}, as a "dog" in the mammal label set $\mathcal{Y}_{n_2}$ ={dog, cat, elephant, tiger}, and so forth.

Dynamic Treecut Loss (DTL). While NCL calibrates node classification, guaranteeing consistency within each node, the label sets of TOS classification can also span different sub-trees of the hierarchy, including nodes at different levels, e.g., $\mathcal{Y}$ ={dog, cat, elephant, tiger, reptile, bird}. DTL seeks to calibrate such label sets, by aggregating the cross-entropy loss of (2) dynamically, i.e., on an example basis, over randomly sampled label sets $\mathcal{Y}_{\mathcal{T}_c} = Leaf(\mathcal{T}_c)$ comprised of the leaves of the tree cuts $\mathcal{T}_c$ (sub-trees) of $\mathcal{T}$. At each training iteration, a random tree cut $\mathcal{T}_c$ is sampled with the $TreeCutSampler$ procedure of Algorithm 1, as illustrated on the middle of Fig. 2, to define the loss

$$\mathcal{L}_{DTL}(\mathbf{C}^t) = L_{\mathcal{Y}_{\mathcal{T}_c}}(\mathbf{C}^t) \quad \mathcal{T}_c \sim TreecutSampler(\mathcal{T}, \beta), \tag{8}$$

where $\beta \in [0, 1]$ is a rate of tree dropout. For this, a Bernoulli random variable $P_n \sim Bernoulli(\beta)$ of dropout rate β is defined for each internal node $n \in \mathcal{N}^{int} \setminus n_0$. The algorithm descends the tree $\mathcal{T}$, sampling a binary drop-out variable p_n at each node. If $p_n = 1$, node n is kept in the pruned tree $\mathcal{T}_c$. Otherwise, the sub-tree of $\mathcal{T}$ rooted with n is dropped from $\mathcal{T}_c$. The parameter β controls the degree of pruning. Larger β induces the pruning of more tree nodes, while $\beta = 0$ guarantess that $\mathcal{Y}_{\mathcal{T}_c} = \mathcal{Y}_{leaf}$. The root node n_0 is excluded, as $p_{n_0} = 0$ would imply discarding the whole $\mathcal{T}$.

The $TreeCutSampler$ algorithm is an efficient procedure to sample tree cuts $\mathcal{T}_c$ from $\mathcal{T}$. It starts by sampling a vector

Algorithm 1 Treecut Sampler

Input: The tree hierarchy $\mathcal{T}$ of the dataset, tree dropout rate β
Output: The treecut label set $\mathcal{Y}_{\mathcal{T}_c}$

```
// sampling p for internal nodes; prune the
   sub-tree rooted at n if p_n = 0
p_{n_0} ← 1 ;                    // always keep the root node
for n ∈ N^{int} \ n_0 do
    p_n ← Bernoulli(β)
p ← (p_{n_1^{int}}, ..., p_{n_K^{int}})
// correct p based on the node dependency
p̃ ← p ⊗ 𝟙[Dp = D1]
// obtain blocked labels with predefined masks
   and the sampled p̃
b ← min(B, 0)^T p̃ + B̄^T (1 − p̃)
// gather available (unblocked) labels as the
   sampled label set
Y_{T_c} ← {n_j : n_j ∈ N \ n_0, b_j = 0}
return Y_{T_c}
```

$\mathbf{p} = (p_{n_1^{int}}, ..., p_{n_K^{int}})$, where n_i^{int} denotes the i-th internal node and $K = |\mathcal{N}^{int}|$, containing pruning flags p_n for all internal nodes $n \in \mathcal{N}^{int}$. The next step is to enforce consistency between these flags, according to the tree structure. If any node in $\mathcal{A}(n)$ is pruned, then node n should be pruned even if $p_n = 1$. This is efficiently enforced across all the flags by defining a dependency matrix $\mathbf{D} \in \{0,1\}^{K\times K}$ where $\mathbf{D}_{ij} = \mathbb{1}[n_j^{int} \in \mathcal{A}(n_i^{int}) \cup \{n_i^{int}\}]$ indicates whether the i-th internal node n_i^{int} is a child of the j-th internal node n_j^{int}. An example is provided on the right of Fig. 2 for the tree on the left. The sampled flags are then corrected by computing $\tilde{\mathbf{p}} = \mathbf{p} \otimes \mathbb{1}[\mathbf{Dp} = \mathbf{D1}]$, where $\mathbf{1}$ is the vector of K ones and $\otimes$ the Hadamard product. Note that both $\mathbf{D}$ and $\mathbf{D1}$ are pre-computed, making the complexity of this step roughly that of one matrix-vector multiplication.

To identify the leaves of the sampled treecut ($\mathcal{Y}_{\mathcal{T}_c} = Leaf(\mathcal{T}_c)$) efficiently, a mask $\mathbf{B} \in \{0, 1, -1\}^{K\times|\mathcal{N}\setminus\{n_0\}|}$ is defined, where each row corresponds to an internal node, and the columns contain all possible labels in $\mathcal{T}$, i.e., all nodes except the root n_0. Entry B_{ij} flags that n_j cannot appear in the sampled label set, given that $n_i \in \mathcal{N}^{int}$ has not been pruned (i.e., $\tilde{p}_{n_i^{int}} = 1$), as follows

$$B_{ij} = \begin{cases} 1, \text{ if } n_j \in \mathcal{A}(n_i^{int}) \cup \{n_i^{int}\} & (n_j \textit{ is an ancestor of } n_i^{int}) \\ 0, \text{ if } n_i^{int} \in \mathcal{A}(n_j) & (n_j \textit{ is a descendant of } n_i^{int}) \\ -1, \text{ otherwise} & (n_j \textit{ is outside of the sub-tree rooted at } n_i^{int}) \end{cases} \tag{9}$$

Similarly, a matrix $\bar{\mathbf{B}}$, of entries $\bar{B}_{ij} = 1 - |B_{ij}|$, is defined to flag that n_j cannot appear in the label set, given that $n_i \in \mathcal{N}^{int}$ has been pruned, i.e. $\tilde{p}_{n_i^{int}} = 0$. A mask of the nodes unavailable to the label set is then computed by accumulating the masks corresponding to the values of $\tilde{\mathbf{p}}$,

$$\mathbf{b} = min(\mathbf{B}, 0)^T \tilde{\mathbf{p}} + \bar{\mathbf{B}}^T (\mathbf{1} - \tilde{\mathbf{p}}) \,, \tag{10}$$

where the mask in $min(\mathbf{B}, 0)$ is selected if $\tilde{p}_n = 1$, and that in $\bar{\mathbf{B}}$ if $\tilde{p}_n = 0$. Note that $min(\mathbf{B}, 0)$ clips $B_{ij} = -1$ to 0. The mask $\mathbf{b}$ can then be used to obtain $\mathcal{Y}_{\mathcal{T}_c} = Leaf(\mathcal{T}_c) = \{n_j : n_j \in \mathcal{N} \setminus n_0, b_j = 0\}$. Fig. 2 gives an example. When $\tilde{\mathbf{p}} = (\tilde{p}_{n_0}, \tilde{p}_{n_1}, \tilde{p}_{n_2}) = (1, 0, 0)$, then $\mathbf{b} = min(\mathbf{B}_1, 0) + \bar{\mathbf{B}}_2 + \bar{\mathbf{B}}_3 = (0, 1, 1, 2, 2, 0)$, signaling that only n_1 and n_6 are available to the label set (as $b_1, b_6 = 0$), resulting in $\mathcal{Y}_{\mathcal{T}_c} = \{n_1, n_6\}$. More detailed examples are given in the appendix.

Optimization. The overall loss used for prompt tuning is a combination of the two losses

$$\mathcal{L}(\mathbf{C}^t) = \mathcal{L}_{DTL}(\mathbf{C}^t) + \lambda \mathcal{L}_{NCL}(\mathbf{C}^t) \tag{11}$$

where λ is a hyperparameter. Note that, like previous prompting approaches, ProTeCt optimizes the learnable prompts $\{\mathbf{c}_m\}_{m=1}^M$ while keeping the parameters of Φ_{text}, Φ_{vis} frozen.

6. Experiments

In this section, we discuss experiments for evaluating the effectiveness of ProTeCt. To demonstrate that ProTeCt is a plug-an-play method, it was applied to two SOTA prompt tuning methods: CoOp [46] and MaPLe [17]. Each experiment is averaged over 3 runs and full tables with error bars are shown in the appendix for brevity. All experiments were conducted on a single Nvidia A10 GPU, using Pytorch [29]. Please see the appendix for more training details and results. ProTeCt code builds on the publicly available codebases for CoOp and MaPLe and will be released upon publication.

Metrics: Acc_{leaf} of (4), HCA of (5) and MTA of (6) are considered. MTA uses 5 tree dropout rates ($\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$) to sample treecuts of various granularities. For each β, T treecuts are sampled without repetition to obtain a total of $5T$ treecuts. MTA($5T$) indicates the result is averaged over these $5T$ treecuts. We ablate $T = 5$ and $T = 20$ on Cifar100 and use $T = 5$ for all datasets by default.

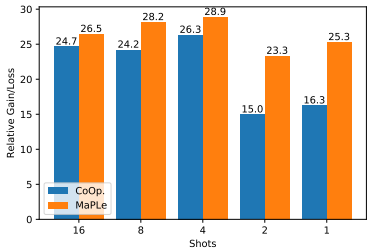
Training Details: All vanilla prompt-tuning and their ProTeCt counterparts are trained under the same setting. The following configuration is used unless noted. All experiments use SGD optimizer and the learning rate is set to 0.02 with a cosine learning rate scheduler. By default, a pretrained ViT-B/16 CLIP model is used as initialization. For Cifar100 and SUN, we train both CoOp and MaPLe prompts for 200 epochs, using a batch size of 128 and 32, respectively. For ImageNet, CoOp is trained for 30 epochs with a batch size of 8, while MaPLe is trained for 10 epochs with a batch size of 2. Note that the setting is slightly different from the original paper due to our GPU availability.

6.1. TOS Classification Performance

Table 2 shows that vanilla CoOp and MaPLe have reasonable leaf accuracy for both 1-shot and 16-shot classification on Cifar100, SUN, and ImageNet. However, their very low HCA shows that their predictions are not consistent over the class hierarchy. As a result, their TOS classification performance (MTA) is much weaker than their leaf accuracy.

Method	K-Shot	w/ ProTeCt	Cifar100 Acc_{leaf}	Cifar100 HCA	Cifar100 MTA (25)	Cifar100 MTA (100)	SUN Acc_{leaf}	SUN HCA	SUN MTA (25)	ImageNet Acc_{leaf}	ImageNet HCA	ImageNet MTA (25)
CoOp	16		72.88	10.04	50.64	51.14	73.82	38.28	52.99	71.23	2.99	46.98
	16	✓	72.94	56.85	87.69	87.30	74.59	62.94	83.51	69.92	37.74	88.61
			(+0.06)	(+46.81)	(+37.05)	(+36.16)	(+0.77)	(+24.66)	(+30.52)	(-1.31)	(+34.75)	(+41.63)
	1		65.03	7.81	41.78	44.17	63.65	33.36	51.20	63.67	1.59	40.52
	1	✓	66.88	41.01	81.64	81.01	63.79	49.62	76.25	66.11	25.79	86.14
			(+1.85)	(+33.2)	(+39.86)	(+36.84)	(+0.14)	(+16.26)	(+25.05)	(+2.44)	(+24.2)	(+45.62)
MaPLe	16		75.01	17.54	52.21	50.82	71.86	33.25	54.29	70.70	4.15	48.29
	16	✓	75.34	61.15	88.04	88.33	72.17	59.71	82.27	69.52	31.24	87.87
			(+0.33)	(+43.61)	(+35.83)	(+37.51)	(+0.31)	(+26.46)	(+27.98)	(-1.18)	(+27.09)	(+39.58)
	1		68.75	4.65	50.60	54.99	63.98	25.15	50.31	68.91	2.97	48.16
	1	✓	69.33	48.10	83.36	83.78	64.29	50.45	76.73	66.16	20.44	85.18
			(+0.58)	(+43.45)	(+32.76)	(+28.79)	(+0.31)	(+25.30)	(+26.42)	(-2.75)	(+17.47)	(+37.02)

Table 2. TOS performance w/ and w/o ProTeCt on Cifar100 ($\lambda = 0.5$), SUN ($\lambda = 0.5$) and ImageNet ($\lambda = 1$). $\beta = 0.1$ for all datasets.



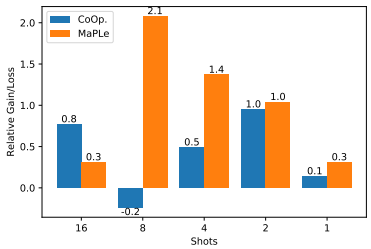
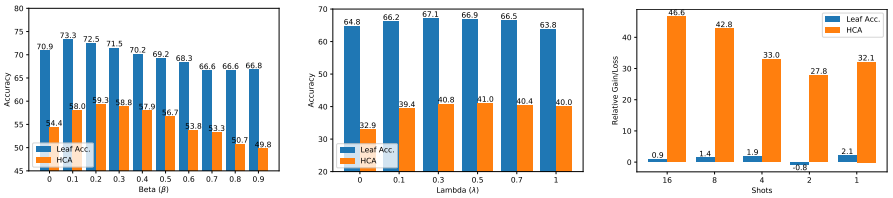


Figure 3. Relative gain/loss after adding ProTeCt to CoOp and MaPle, respectively. (Top) HCA ; (Bottom) Acc_{leaf}.



(a) (b) (c)

Figure 4. Ablation of (a) tree dropout rate β, (b) NCL strength λ and (c) CLIP ViT B32 architecture.

(a): [Boxer, Person] (b): [Trolleybus, Animal] (c):[Lion, Koala] (d): [Cheeseburger, Ice cream]

Figure 5. ProTeCt correctly predicts examples from ImageNet (a,b) and its variants (c,d) at all levels. [GT, Prediction] shows the groundtruth and incorrect prediction by vanilla prompt tuning.

For example, 16-shot classification with CoOp on ImageNet has a leaf accuracy of 71.23, but expected TOS accuracy of 46.98. This is explained by the very low HCA of 2.99. Similar observations hold for different few-shot configurations. In all cases, ProTeCt (results on rows with a checkmark) significantly improves HCA and MTA(25). For example, it boosts the HCA of 16-shot classification with CoOp on ImageNet by 34.75 (2.99 vs 37.74), leading to an increase of MTA(25) of 41.63 (46.98 to 88.61).

Note that, in all cases, MTA(25) after ProTeCt training is *higher* than leaf accuracy. This is expected for a well-calibrated classifier, since decisions at intermediate levels of the tree are coarser-grained than those at the leaves, which can require very fine class distinctions. These results show that ProTeCt robustifies the model for use in the TOS classification setting. The table also shows that ProTeCt maintains leaf accuracies comparable to those of the vanilla methods. Furthermore, the MTA results when 25 and 100 treecuts are sampled (corresponding to $T = 5$ and $T = 20$), are compared on Cifar100. It can be seen that the performances are similar, showing that sampling 25 treecuts is sufficient to achieve good estimation. Fig. 3 compares the **relative** gains in HCA and leaf accuracy of training with ProTeCt, as compared to vanilla prompt tuning. These gains are shown for both CoOp and MaPLe, under several few shot configurations, on SUN dataset. In all cases, ProTeCt increases HCA by more than 15 points, while maintaining a leaf accuracy comparable to that of vanilla CoOp/MaPLe. Similar results for Cifar100 and ImageNet can be found in appendix.

6.2. Domain Generalization of TOS Classification

We investigate whether TOS classification performance generalizes across datasets, following the domain generalization setting of [17, 42, 45, 46]. The CLIP model with ProTeCt prompts trained on ImageNet (source) is applied to 4 ImageNet variants (target) with visual domain shift: ImageNetv2 [31], ImageNet-Sketch [36], ImageNet-A [15] and ImageNet-R [14]. Table 3 summarizes the three metrics on these datasets for CoOp and MaPLe. Similarly to Table 2, ProTeCt enables significant gains in HCA and MTA(25) over the baselines for all datasets. Note that since ImageNet-A and ImageNet-R only contain 200 ImageNet subclasses, their

Method	K-Shot	w/ ProTeCt	ImageNetv2 [31]			ImageNet-S [36]			ImageNet-A [15]			ImageNet-R [14]		
			Acc_{leaf}	HCA	MTA (25)	Acc_{leaf}	HCA	MTA (25)	Acc_{leaf}	HCA	MTA (25)	Acc_{leaf}	HCA	MTA (25)
CoOp	16		64.01	2.31	43.74	47.82	1.39	38.58	50.28	2.97	52.56	75.83	18.49	64.13
	16	✓	62.60	32.84	86.66	46.80	20.73	82.60	49.08	22.45	78.21	74.94	31.18	75.59
			(-1.41)	(+30.53)	(+42.92)	(-1.02)	(+19.34)	(+44.02)	(-1.20)	(+19.48)	(+25.65)	(-0.89)	(+12.69)	(+11.40)
	1		56.43	1.51	38.27	41.38	1.11	33.61	45.92	1.76	47.54	69.84	11.74	55.31
	1	✓	60.16	22.95	84.38	44.75	13.88	80.64	48.95	20.52	76.95	74.26	27.46	76.48
			(+3.73)	(+21.44)	(+46.11)	(+3.37)	(+12.77)	(+47.03)	(3.03)	(+18.76)	(+29.41)	(+4.42)	(+15.72)	(+21.17)
MaPLe	16		64.15	1.97	45.93	48.97	1.58	43.37	50.61	2.31	54.88	76.61	20.67	63.06
	16	✓	62.77	27.86	86.14	47.47	17.77	82.52	47.41	19.75	77.46	75.70	32.58	77.99
			(-1.38)	(+25.89)	(+40.21)	(-1.50)	(+16.19)	(+39.15)	(-3.20)	(+17.44)	(+22.58)	(-0.91)	(+11.91)	(+14.93)
	1		61.78	2.18	45.50	46.79	1.70	45.26	47.55	3.52	55.48	74.55	18.85	62.48
	1	✓	59.14	17.89	83.27	44.92	11.24	79.94	47.15	16.03	76.81	74.60	25.20	75.72
			(-2.64)	(+15.71)	(+37.77)	(-1.87)	(+9.54)	(+34.68)	(-0.40)	(+12.51)	(+21.33)	(+0.05)	(+6.35)	(+13.24)

Table 3. The gain of hierarchical consistency after adding ProTeCt generalizes across datasets in unseen domains. All methods are fine-tuned on ImageNet and evaluated on its 4 variants.

DTL	NCL	16-shot			1-shot		
		Acc_{Leaf}	HCA	MTA (25)	Acc_{Leaf}	HCA	MTA (25)
		72.88	10.04	50.64	65.03	7.81	41.78
✓		72.81	47.97	87.32	64.77	32.93	81.38
	✓	64.20	51.69	79.44	61.22	38.02	62.16
✓	✓	**72.94**	**56.85**	**87.69**	**66.88**	**41.01**	**81.64**

Table 4. Loss ablation with CoOp on Cifar100 dataset. Both losses improve the hierarchical consistency.

K-Shot	w/ ProTeCt	CLIP-Adapter [12]			CLIP+LORA [9]		
		Acc_{leaf}	HCA	MTA (25)	Acc_{leaf}	HCA	MTA (25)
16		71.96	5.59	42.93	70.45	4.57	47.19
16	✓	72.47	57.15	87.67	70.64	51.06	77.29
		(+0.51)	(+51.56)	(+44.83)	(+0.19)	(+46.49)	(+30.10)
1		65.35	8.35	48.25	63.57	2.89	38.63
1	✓	67.29	36.21	78.49	63.62	24.66	56.42
		(+1.94)	(+27.86)	(+30.24)	(+0.05)	(+21.8)	(+17.79)

Table 5. ProTeCt also improves adapter-based methods, including CLIP-Adapter [12] and CLIP+LORA [9] (dataset: Cifar100).

hierarchy is different from that of ImageNet. These results demonstrate the flexibility and robustness of ProTeCt, even when transferring the model to a target domain whose class hierarchy is different from that of the source domain.

6.3. Ablation Study and Visualization

In this section, we discuss the ablations of ProTeCt components and visualize the predictions (more in the appendix).

Tree Dropout Rate β: Fig. 4 (a) plots Cifar100 Acc_{leaf} and HCA as a function of the drop-out rate β, for 16-shot CoOp+ProTeCt training ($\lambda = 1$). Larger values of β reduce the likelihood of sampling the leaf nodes of the tree, resulting in shorter trees and weaker regularization. Hence, both leaf accuracy and HCA degrade for large β. However, always using the full tree ($\beta = 0$) also achieves sub-optimal results. The two metrics peak at $\beta = 0.1$ and $\beta = 0.2$, respectively. $\beta = 0.1$ is selected for all experiments.

Loss: Fig. 4(b) ablates the strength of NCL loss (i.e. λ) for ProTeCt+CoOp using 1-shot setting on Cifar100 and $\beta = 0.1$. The introduction of NCL improves leaf accuracy/HCA from 64.8/32.9 ($\lambda = 0$) to 66.9/41 ($\lambda = 0.5$). We adopt $\lambda = 0.5$ for CIFAR100 and SUN. For ImageNet, $\lambda = 0.5$ and $\lambda = 1$ have similar performance. Table 4 further summarizes the CoOp+ProTeCt performance with and without the two losses of (11). Both losses improve TOS performance individually and there is a large additional gain when they are combined. Using NCL alone can degrade leaf performance, due to the lack of regularization across different levels of the hierarchy. The combination of the two losses overcomes this problem.

Architecture: Fig. 4 (c) shows that the gains for CoOp+ProTeCt in Fig. 3 with CLIP ViT B16 also hold for ViT B32, showing the plug-and-play properties of ProTeCt.

Adapter-based tuning methods: We further use the ProTeCt losses to train the CLIP adapter of [12] and the CLIP+LORA method of [9] to test the generation of ProTeCt. Table 5 shows that this again produces large consistency gains on the TOS setting, indicating that ProTeCt losses generalize to both prompt-based and adapter-based methods.

Visualization: Fig. 5 shows examples from ImageNet (a,b) and its variants (c,d). While ProTeCt correctly classifies these examples at all hierarchy levels, vanilla prompt tuning fails at certain levels. More examples are in the appendix.

7. Conclusion

In this work, we formulated the TOS classification setting, including datasets, performance metrics, and experiments. Given a dataset, a class hierarchy is built by assigning dataset classes to leaf nodes and superclasses to internal nodes. The TOS classifier is then expected to support classification with label sets drawn throughout the taxonomy. We have shown that existing FMs and prompting methods fail under this setting and proposed ProTeCt training to enhance the TOS performance of FMs, as a plug-and-play method. ProTeCt includes two losses. A dynamic treecut loss, based on an efficient treecut sampler, dynamically regularizes labels of varying granularity. A node-centric loss encourages correct predictions at all hierarchy levels. Experiments show that ProTeCt enhances TOS performance of existing prompt-tuning techniques, and the gain generalizes across unseen domains. Finally, we show that ProTeCt is applicable to various architectures, hierarchies, and parameter-tuning methods.

Acknowledgement This work was partially funded by NSF awards IIS-2303153, and a gift from Qualcomm. We also acknowledge and thank the use of the Nautilus platform for some of the experiments discussed above.

References

[1] Karim Ahmed, Mohammad Haris Baig, and Lorenzo Torresani. Network of experts for large-scale image categorization. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[2] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision&language models. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[3] 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*. JMLR.org, 2020. 1, 3

[4] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *ArXiv*, abs/2003.04297, 2020. 1, 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 4

[6] Jia Deng, Jonathan Krause, Alexander C. Berg, and Li Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[7] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision (ECCV)*, 2014. 3

[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. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[9] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogério Schmidt Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2657–2668, 2022. 8

[10] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. In *International Joint Conference on Artificial Intelligence*, 2022. 2

[11] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998. 2, 4

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Jiao Qiao. Clip-adapter: Better vision-language models with feature adapters. *ArXiv*, abs/2110.04544, 2021. 8

[13] Wonjoon Goo, Juyong Kim, Gunhee Kim, and Sung Ju Hwang. Taxonomy-regularized semantic deep convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2020. 2, 4, 7, 8

[15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 2, 4, 7, 8

[16] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[17] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 6, 7

[18] Hyo Jin Kim and Jan-Michael Frahm. Hierarchy of alternating specialists for scene recognition. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[19] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report, Citeseer*, 2009. 4

[20] Kibok Lee, Kimin Lee, Kyle Min, Yuting Zhang, Jinwoo Shin, and Honglak Lee. Hierarchical novelty detection for visual object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3

[22] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, 2021. Association for Computational Linguistics.

[23] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland, 2022. Association for Computational Linguistics. 3

[24] Yuntao Liu, Yong Dou, Ruochun Jin, and Peng Qiao. Visual tree convolutional neural network in image classification. In *International Conference on Pattern Recognition (ICPR)*, 2018. 3

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[26] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[27] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 3

[28] Marcin Marszałek and Cordelia Schmid. Semantic hierarchies for visual object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 2, 3

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[30] 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 *International Conference on Machine Learning*, 2021. 1, 3

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019. 2, 4, 7, 8

[32] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 3

[33] Babak Shahbaba and Radford M. Neal. Improving classification when a class hierarchy is available using a hierarchy-based prior. *Bayesian Analysis, 2(1):221–238*, 2007. 3

[34] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016. 1, 3

[35] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 1, 3

[36] 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. 2, 4, 7, 8

[37] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiaoyong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *ArXiv*, abs/2302.10035, 2023. 2

[38] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 3

[39] Tz-Ying Wu, Pedro Morgado, Pei Wang, Chih-Hui Ho, and Nuno Vasconcelos. Solving long-tailed recognition with deep realistic taxonomic classifier. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[40] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 4

[41] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition. In *International Conference on Computer Vision (ICCV)*, 2015. 2, 3

[42] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *ArXiv*, abs/2210.07225, 2022. 1, 3, 7

[43] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *ArXiv*, abs/2304.00685, 2023. 2

[44] Bin Zhao, Li Fei-Fei, and Eric P. Xing. Large-scale category structure aware image categorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2011. 3

[45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 7

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 1, 2, 3, 6, 7

[47] Xinqi Zhu and Michael Bain. B-cnn: Branch convolutional neural network for hierarchical classification. *CoRR*, abs/1709.09890, 2017. 3