Open Set Recognition by Mitigating Prompt Bias

Anonymous authors
Paper under double-blind review

Abstract

Existing open set recognition (OSR) methods are usually performed on relatively small datasets by training a visual model from scratch. OSR on large-scale datasets has rarely been studied for their great complexity and difficulty. Recently, vision-language (VL) pre-training has promoted closed-set image recognition with prompt engineering on datasets with various scales. However, prompts tuned on the training data often exhibit label bias towards known classes, leading to the poor performance in recognizing unknown data in the open environment. In this paper, we aim at developing a new paradigm for OSR both on small and large-scale datasets by prompt engineering on VL models in a divide-and-conquer strategy. Firstly, the closed-set data is processed as the combination of one or more groups. Each group is devised with a group-specific prompt. Then, we propose the Group-specific Contrastive Tuning (GCTu), in which negative label words are introduced into tuning to mitigate the label bias of group-specific prompts. In inference, to achieve comprehensive predictions both on small and large-scale datasets, we propose the Group Combined Testing (GCTe). It determines the optimal prediction prompt among the multiple group-specific predictions by focusing on the group-wise closed-set probability distributions. Our method namely GCT2 achieves excellent performance on both small and large-scale OSR benchmarks. The strong and wide applicability of our method is also verified in ablation studies.

1 Introduction

Real-world image recognition often involves samples from unknown classes, which are unseen in the training stage. Accordingly, open set recognition (OSR) Scheirer et al. (2012); Bendale & Boult (2016) has been devised for classifying known classes appearing in the training set as well as detecting unknown classes. However, existing OSR methods Sun et al. (2020); Júnior et al. (2017); Oza & Patel (2019); Neal et al. (2018); Kong & Ramanan (2021); Zhou et al. (2021a) are mostly performed on small-scale datasets in terms of the number of classes, such as CIFAR10 Krizhevsky (2009) and TinyImageNet Le & Yang (2015), which include up to tens of known classes and less than two hundred unknown classes. The recognition models are commonly trained from scratch with a simple visual backbone consisting of nine convolutional layers and one full connection layer Neal et al. (2018); Zhang et al. (2020); Zhou et al. (2021a). Being far more challenging and difficult, such OSR methods can not perform well on large datasets due to their great complexity, such as ImageNet Russakovsky et al. (2015) consisting of 1000 classes. Only a few methods Yang et al. (2020); Chen et al. (2020a); Lu et al. (2022) are proposed to solve this issue with a stronger backbone ResNet50 He et al. (2016).

Recently, vision-language (VL) pre-training models Lu et al. (2019); Chen et al. (2020b); Gan et al. (2020); Li et al. (2020); Zhang et al. (2021) have shown the promising ability to benefit the downstream tasks. By redesigning the downstream tasks as pre-training tasks, prompt engineering on the VL pre-trained models Radford et al. (2021); Jia et al. (2021); Li et al. (2021) exhibits excellent potential in image recognition tasks with various scales with only a few embedding parameters optimized. However, applying prompt to recognition tasks usually obeys the closed setting. The downstream training and testing classes are the same. Because VL models have already been pre-trained on a large amount of data, the open-set concept is hard to be guaranteed if we take the pre-training data into consideration. Therefore, we refer to the setting where *the testing classes are composed of classes from the downstream training classes and classes out of downstream training classes as* *the open-set setting based on pre-trained VL models.* It rises the so-called *label bias* issue Cao et al. (2021); Zhao et al. (2021), which is defined as that the prompt tuned on limited training classes forcefully selects a known class as the predictions for unknown classes, in open-set scenarios.

A question arises that whether it is more effective to apply pre-trained VL models to OSR on datasets with both small and large numbers of classes. To this end, the main goals in this paper include: (i) exploring a new paradigm for solving the OSR problem with prompt engineering on pre-trained VL models; (ii) exploring a strong applicable strategy for OSR on small and large-scale datasets uniformly. Surpassing other state-of-the-art methods is not our goal.

Firstly, we introduce the divide-and-conquer strategy for the wide applicability on datasets with both small and large number of classes. Each dataset can be processed as the combination of one or more mutual-independent class groups. Each group is devised with a set of unused tokens, namely group-specific prompts, which will be tuned only on the classes in its corresponding group. Then, we build an open negative label pool containing thousands of label words collected from the WordNet Miller (1995). To mitigate the prompt label bias towards closed-set classes in each group, we propose the Group-specific Contrastive Tuning (GCTu). Several open negative label words irrelevant to the downstream datasets are collected from the built label pool and introduced into prompt tuning without paired images to regularize group-specific predictions.

In inference, each sample obtains multiple predictions from all the group-specific prompts. To make flexible and comprehensive decisions generalizing to both small and large-scale datasets, we propose the Group Combined Testing (GCTe). The prompt, which exhibits the highest probability within its group-specific closed-set classes, of all prompts is employed as the optimal prediction prompt for a given sample.

To our best knowledge, this is the first work applying VL models to OSR that scales up its applicability to datasets with a large number of classes by prompt engineering. Experimentally, the proposed method, which we name as GCT2, achieves excellent performance on both small and large-scale benchmarks. Extensive ablation experiments validate the effectiveness of each component. The highlights of the proposed new paradigm include:

(1) To solve the misclassification issue of prompt in the open world, we propose the Group-specific Contrastive Tuning (GCTu). It mitigates the prompt label bias by introducing open negative label words, which are irrelevant to downstream datasets, without paired images into tuning.

(2) To achieve the wide applicability on different scales of datasets, we propose the Group Combined Testing (GCTe). It determines the optimal prompt by measuring the group-wise closed-set probabilities.

2 Related Work

2.1 Open Set Recognition

Towards practical recognition, open set recognition (OSR) Scheirer et al. (2012) has made fast progress in recent years. Methods in the literature include traditional machine learning methods Zhang & Patel (2016); Rudd et al. (2017); Clifton et al. (2011); Hoffmann (2007); Scheirer et al. (2014); Jain et al. (2014); Bendale & Boult (2015); Júnior et al. (2017) as well as deep learning methods Miller et al. (2021); Geng & Chen (2020); Meyer & Drummond (2019); Oza & Patel (2019); Sun et al. (2020); Zhou et al. (2021a); Neal et al. (2018); Chen et al. (2020a; 2021), which almost perform on small-scale datasets. Specifically, CIFAR-series benchmarks Krizhevsky (2009); Neal et al. (2018) include no more than 10 closed-set classes. TinyImageNet Le & Yang (2015) is composed of 20 known and 180 unknown classes. In addition, these methods commonly train models based on a simple visual backbone Neal et al. (2018); Zhang et al. (2020); Zhou et al. (2021a) from scratch. Being far more challenging and difficult, only a few methods Yang et al. (2020); Chen et al. (2020a); Lu et al. (2022) are proposed to handle the OSR problem on ImageNet-series Russakovsky et al. (2015) benchmarks, which include hundreds classes. However, the simple backbone on small-datasets usually fail when facing the great complexity brought by the large amount of classes. The visual backbone adopted on large-scale datasets is usually stronger than that on small-scale datasets. Most similar to our method which adopts pre-trained VL models, ZOC Esmaeilpour et al. (2022) trains a text decoder based on CLIP Radford et al. (2021) using an image caption dataset

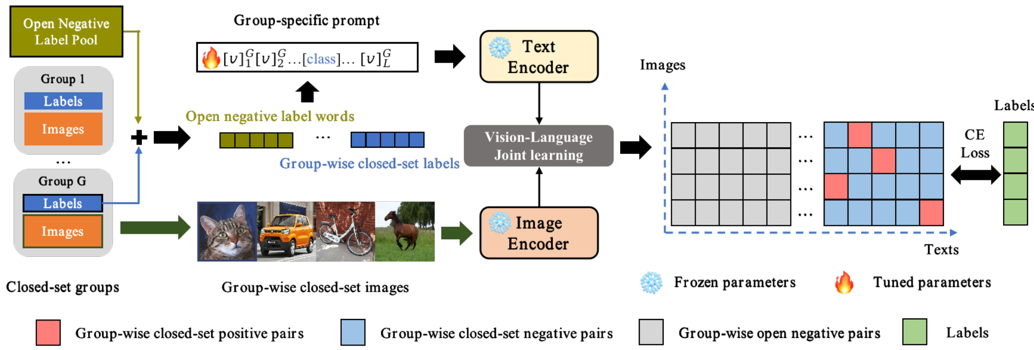


Figure 1: Framework of Group-specific Contrastive Tuning (GCTu). Group-specific prompts are tuned only on their corresponding groups. The prompt label bias is mitigated by introducing open negative label words into the tuning stage. The parameters of the pre-trained model are all kept frozen. Only the prompt embeddings are optimized.

to generate predicted category words for out-of-distribution (OOD) detection but only on small-scale datasets. Moreover, it could not gurantee the great closed-set classification performance for its being conducted in the zero-shot way.

In this paper, we propose a new paradigm, which explores the diverse knowledge of vision-language pre-trained models, to solve OSR both on small and large scale datasets uniformly with the grouping strategy.

2.2 Prompt Engineering

Prompt engineering is primarily proposed in natural language processing (NLP) Petroni et al. (2019). It redesigns downstream tasks as pre-training tasks Jiang et al. (2020); Lester et al. (2021); Li & Liang (2021); Liu et al. (2021); Poerner et al. (2019); Shin et al. (2020) and thus narrows down the gap between them, which contributes to exploring the pre-learned knowledge adequately, also in vision-language (VL) models Zhou et al. (2021b); Jia et al. (2021); Radford et al. (2021); Li et al. (2021). Three parts are usually contained in prompt engineering, namely a template, a set of training samples and their orderings. Concerning the training and testing classes, prompt engineering is now performed with the closed setting assumption, which causes the so-called label bias Cao et al. (2021); Zhao et al. (2021) in open world, whereby the model has to output a predicted class in the training set for all the testing samples. To improve the performance in out-of-distribution (OOD) detection by prompt, true label words of OOD data are introduced as prior Fort et al. (2021) into CLIP Radford et al. (2021). However, it is not applicable in the open-set scenario because we have no knowledge of the unknown classes. Instead, we propose to mitigate the label bias by introducing open negative label words, irrelevant to the downstream datasets and without paired images, into both the prompt tuning and in-context prediction stages and serve for OSR.

3 Approach

We present our new paradigm for OSR both on small and large-scale datasets with group-guided prompt engineering on pre-trained vision-language (VL) models. Specifically, the problems in the new paradigm include two folds: the group-specific prompt engineering and combined prediction. For the first fold, we propose the Group-specific Contrastive Tuning (GCTu) to learn group-specific text prompts with the label bias being mitigated, as shown in Fig. 1. Second, the Group Combined Testing (GCTe) is developed to make flexible and comprehensive decisions combining predictions on all group-specific prompts, as shown in Fig. 2.

3.1 Grouping on Closed-set Classes

To develop a widely applicable strategy for small and large-scale datasets, we divide the downstream closed-set dataset consisting of N_C classes into G groups, in which the maximum number of classes

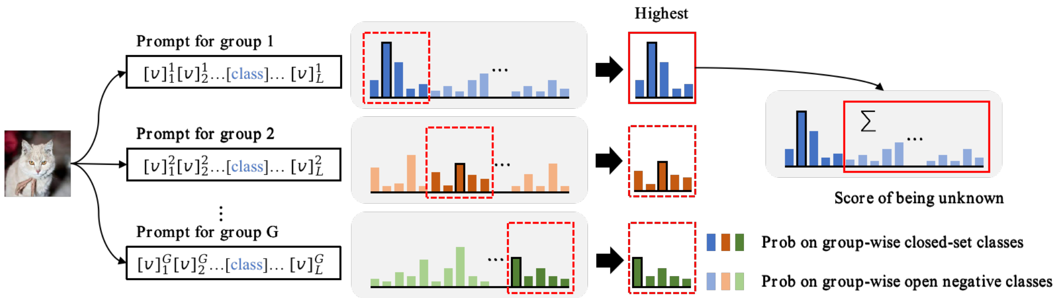


Figure 2: Framework of the proposed Group Combined Testing (GCTe). Each image obtains multiple predictions from all group-specific prompts. We only focus on the probabilities of the group-specific closed-set classes for comprehensive comparisons. The prompt which exhibits the highest closed-set probability is selected as the final prediction prompt.

is N_{max}. The numbers of classes in the first $G-1$ groups are equal. Formally, we denote the number of classes in each group as $N_g^i, i \in [1, G]$. The grouping rule is:

$$(G-1) \cdot N_{max} < N_C \leq G \cdot N_{max}, N_g^G \leq= N_g^i = N_{max}, i \in [1, G-1]. \tag{1}$$

Aiming at efficiently exploring the knowledge in pre-trained VL models with only a few parameters to be optimized and obtain the group-independent predictions without mutual impacts, we devise a set of unused tokens $[\boldsymbol{v}]$ as the group-specific continuous prompts $\boldsymbol{F}_i(CLASS), i \in [1, G]$ as Eq. 2. Each group-specific continuous prompt with length L will be tuned only on the data in its corresponding group.

$$\boldsymbol{F}_i(CLASS) = [\boldsymbol{v}]_1^i[\boldsymbol{v}]_2^i...[CLASS]...[\boldsymbol{v}]_L^i, i \in [1, G]. \tag{2}$$

3.2 Group-specific Contrastive Tuning (GCTu)

Image recognition tasks promoted by prompt engineering methods almost obely closed-setting Jia et al. (2021); Radford et al. (2021); Li et al. (2021) ignoring unknown classes in the open world. When detecting unknown classes, the significant label bias of prompts Cao et al. (2021); Zhao et al. (2021) inevitably harms the OSR that unknown data would still be predicted as the classes on which the prompts have been tuned.

The underlying rationality is that prompts tuned on closed-set classes forces the images belonging to both known and unknown classes to be predicted within the known classes with high probability. If the high probability could be regularized, by which closed-set classes could still be correctly predicted while open-set unknown images obtain much lower probabilities on known classes, the label bias would be mitigated.

To this end, we propose the Group-specific Contrastive Tuning (GCTu) as shown in Fig. 1. The open negative label pool is built by collecting thousands of label words that are irrelevant to the downstream datasets from the WordNet Miller (1995). By introducing open negative label words into tuning, the prompts are forced to make predictions on each sample not only from the group-specific closed-set classes, irrelevant label words are set to be selected as probable predictions for regularization. Therefore, the large probability of unknown data belonging to closed-set classes is avoided. The labels participated in the tuning stage of each group include two parts: (i) the group-specific closed-set labels $C_j^i, i \in [1, G], j \in [0, N_g^i - 1]$; (ii) the group-specific open negative label words $C_j^i, i \in [1, G], j \in [N_g^i, N_g^i + N_o - 1]$, in which N_o is the number of open negative label words sampled from the open negative label pool.

Given a pre-trained VL model consisting of an image encoder E_I and text encoder E_T, the probability that an image $\boldsymbol{x}$ belonging to class $C_j^i, i \in [1, G], j \in [0, N_g^i - 1] \cup [N_g^i, N_g^i + N_o - 1]$ is measured by a commonly used cosine metric $< \cdot >$ with the temperature parameter T:

$$p(y = C_j^i | \boldsymbol{x}) = \frac{\exp(< E_I(\boldsymbol{x}) \cdot E_T(\boldsymbol{F}_i(C_j^i)) > /T)}{\sum_{j=0}^{N_g^i + N_o - 1} \exp(< E_I(\boldsymbol{x}) \cdot E_T(\boldsymbol{F}_i(C_j^i)) > /T)}. \tag{3}$$

Denoting the true label encoding of an image $\boldsymbol{x}$ in the i-th group as $\boldsymbol{y}^{gt}$, we optimize the i-th group-specific prompt using cross entropy loss as:

$$L_i = -\sum_{j=0}^{N_g^i+N_o-1} y_j^{gt} \log(p(y = C_j^i|\boldsymbol{x})). \tag{4}$$

To prevent the mutual impact of the group-specific prompts and achieve group-independent predictions, when a certain prompt is being tuned, others are kept frozen together with parameters of the pre-trained VL model.

3.3 Group Combined Testing (GCTe)

In the testing phase, to preserve the generalization with the prompt bias being mitigated, open negative label words are also introduced in model inference. As the whole closed-set data is divided into groups, each image will be predicted by all group-specific prompts. Specifically, to perform comprehensive predictions combining all closed-set groups and all group-specific predictions, we propose the Group Combined Testing (GCTe), as shown in Fig. 2.

In the prediction of a group-specific prompt, only the probabilities on the corresponding group-specific closed-set classes are worth comparison for their actual meanings of how likely the test sample belongs to these classes. We define the group-specific closed-set maximum probability $p^i_{max}, i \in [1, G]$ as:

$$p^i_{max} = \max(p(y = C_j^i|\boldsymbol{x})), i \in [1, G], j \in [0, N_g^i - 1]. \tag{5}$$

As a test sample will be predicted with a high probability on its true class by the prompt corresponding to its group, we choose the optimal prompt with the group index as I_{opt}:

$$I_{opt} = \arg\max p^i_{max}, i \in [1, G]. \tag{6}$$

Considering the sum of probabilities on group-specific closed-set classes to be the score of being known. Other probabilities are therefore taken for binary detection by a defined score of being unknown S_{open} as:

$$S_{open} = 1 - \sum_{j=0}^{N_g^{I_{opt}}-1} p(y = C_j^{I_{opt}}|\boldsymbol{x}). \tag{7}$$

We set the threshold τ_{max} on the maximum probability $p^{I_{opt}}_{max}$ predicted by the optimal prompt to directly detect the unknown-class samples in OSR. Formally, the prediction is specified as:

$$pred = \begin{cases} \arg\max\limits_{j\in[0,N_g^{I_{opt}}-1]} \; p(y = C_j^{I_{opt}}|\boldsymbol{x}), \; if \; p^{I_{opt}}_{max} \geq \tau_{max} \\ \text{unknown}, else \end{cases}. \tag{8}$$

4 Experiments

4.1 Implementation Details

For data preparation, small datasets are divided into groups by their category names in order, ImageNet-serises large-scale datasets are grouped by the WordNet ID (WNID) orders. The pre-trained VL model is contrastive language-image pre-training (CLIP) Radford et al. (2021) with the base version ViT-B/32 Dosovitskiy et al. (2020) as image encoder. In the GCTu, the initial learning rate is set to $1e-5$. We apply the linear learning rate decay scheduler to the AdamW optimizer Loshchilov & Hutter (2018) as suggested by the Huggingface Transformers [1] default setup. The temperature parameter T is set to 1 for simplicity. For each divided group, we tune for 30 epochs on 4 NVIDIA Tesla V100 GPUs with batch size 64. The $[CLASS]$ is placed in the middle of prompts.

[1] https://huggingface.co/transformers/.

4.2 Unknown Detection on Small-Scale Datasets

In this part, we present the main results of binary known/unknown detection on five small-scale benchmarks. The AUROC (Area Under ROC Curve) is adopted for performance evaluation based on the defined score of being unknown S_{open} in Eq. 7.

Datasets and settings. Each dataset is split into a known and an unknown part. CIFAR10 Krizhevsky (2009) is randomly split into 6 known classes and 4 unknown classes. The 100 classes in CIFAR100 Krizhevsky (2009) are divided into 20 known classes and 80 unknown classes. For CIFAR+10/+50 Neal et al. (2018), 4 classes are selected from CIFAR10 as known, 10 or 50 classes are randomly sampled from CIFAR100 as unknown. TinyImageNet Le & Yang (2015) includes 200 classes with 20 classes set as known and the remaining 180 classes set as unknown. Experiments are all performed for five randomized trials on each benchmark.

Table 1: Unknown detection performance evaluated by AUROC on small datasets, averaged among 5 randomized trials. We use C and TinyIN to represent CIFAR and TinyImageNet respectively.

Methods	C10	C+10	C+50	TinyIN	C100
OSRCI (Neal et al.)	69.9	83.8	82.7	58.6	N.R.
CGDL (Sun et al.)	90.3	95.9	95.0	76.2	N.R.
GDFR (Perera et al.)	83.1	92.8	92.6	60.8	N.R.
C2AE (Oza & Patel)	89.5	95.5	93.7	74.8	N.R.
PROSER (Zhou et al.)	89.1	96.0	95.3	69.3	N.R.
CPN (Yang et al.)	82.8	88.1	87.9	63.9	N.R.
RPL (Chen et al.)	90.1	97.6	96.8	80.9	N.R.
ARPL+CS (Chen et al.)	91.0	97.1	95.1	78.2	N.R.
PMAL (Lu et al.)	95.1	97.8	96.9	83.1	N.R.
OpenGAN-pix (Kong & Ramanan)	97.1	N.R.	N.R.	79.5	N.R.
OpenGAN-feat (Kong & Ramanan)	**97.3**	N.R.	N.R.	**90.7**	N.R.
ZOC (Esmaeilpour et al.)	93.0	**97.8**	**97.6**	84.6	82.1
GCT2 (Ours)	96.1	96.1	96.2	88.2	**86.2**

Results comparison. In our ablation study, when N_{max} is set to 20 and L is set to 10, the results on almost all benchmarks are the best. Under this setting, we show the unknown detection results of our method compared with other existing methods in Table 1. Our method achieves excellent performance, especially on CIFAR10, TinyImageNet and CIFAR100. The most similar method to ours is the ZOC which also adopts CLIP for unknown detection. Our method surpasses it on 3 out of 5 datasets more than 3 percents. The competitive performance of our method validates that the prompt bias has been mitigated and contributes to OSR on small-scale datasets with only a few parameters to be optimized.

4.3 Unknown Detection on Large-Scale Datasets

Here, we measure the performance of our method in unknown detection on ImageNet-series benchmarks, which is more challenging and difficult than on small datasets.

Datasets and settings. Following the dataset preparation Yang et al. (2020); Chen et al. (2020a); Lu et al. (2022) on the ImageNet dataset which includes 1000 classes in total, two benchmarks namely ImageNet-100 and ImageNet-200 are constructed. The first 100 or 200 classes in ImageNet are selected as known, while the remaining 900 or 800 classes are treated as unknown to build ImageNet-100 and ImageNet-200 respectively. The other benchmark is a long-tailed dataset namely ImageNet-LT Liu et al. (2019) which includes 1000 known classes from ImageNet-2012 Russakovsky et al. (2015). The number of images in known classes ranges from 5 to 1280. Additional classes in the validation dataset of ImageNet-2010 are set as unknown.

Table 2: Unknown detection AUROC on large-scale datasets.

Methods	IN-100	IN-200	IN-LT
Softmax	79.7	78.4	53.3
CPN (Yang et al.)	82.3	79.5	54.5
RPL (Chen et al.)	81.2	80.2	55.1
PMAL (Lu et al.)	94.9	93.9	71.7
GCT2 (Ours)	**98.1**	**95.5**	**81.9**

Results comparison. When setting L to 10 and N_{max} to 20, the comparison of our method and the only three existing methods for unknown detection on large-scale datasets is shown in Table 2. The results show that by dividing large-scale datasets into small groups for independent prompt tuning with label bias being mitigated and combined prediction, our method successfully applied to large-scale datasets and achieves the best performance. Note that as stated in CLIP Radford et al. (2021) that the pre-training dataset of it does not access to the ImageNet, which will not bring explicit information of both known and unknown classes.

Table 3: Comparison to CLIP baseline in unknown detection.

Methods	CIFAR10	CIFAR+10	CIFAR+50	TinyImageNet	CIFAR100	ImageNet-100	ImageNet-200	IN-LT
CLIP+MSP	87.3	92.2	92.4	83.7	84.0	74.0	78.5	58.1
GCT2 (Ours)	**96.1**	**96.1**	**96.2**	**88.2**	**86.2**	**98.1**	**95.5**	**81.9**

4.4 Unknown Detection Baseline Comparison

Table 1 and Table 2 compare our method with existing OSR methods with different backbones. Though we do not aim to surpass them, for fair comparison and validating the effectiveness of our method, we construct a baseline adopting the maximum over softmax probabilities (MSP) Hendrycks & Gimpel (2017) for unknown detection with CLIP. The baseline setting only involves one prompt for each dataset without grouping. It also does not introduce open negative label words into tuning. The comparison to the baseline is shown in Table 3.

Our method outperforms the baseline on all datasets by a large margin. Specifically, the performance margin on large-scale datasets is much larger than that on small-scale datasets. The comparison validates the effectiveness of the two key designs: (i) introducing open negative label words into tuning to mitigate the prompt bias; (ii) adopting grouping and combined prediction strategies to achieve strong applicability, especially on large-scale datasets with great challenge and complexity.

4.5 Open-set Recognition

We evaluate the performance of closed-set classification and unknown class recognition using the macro-averaged F1-score (mF1-score). In consistent with the literature Neal et al. (2018); Yoshihashi et al. (2019); Oza & Patel (2019), we set CIFAR10 as known. ImageNet-crop, ImageNet-resize, LSUN-crop, LSUN-resize, which are cropped and resized from ImageNet Russakovsky et al. (2015) and LSUN Yu et al. (2015), are selected as 4 sets of open-set data.

Table 4: Open-set recognition on CIFAR10 evaluated by mF1-score. IN-c, IN-r, LS-c, and LS-r represent ImageNet-crop/-resize, LSUN-crop/-resize.

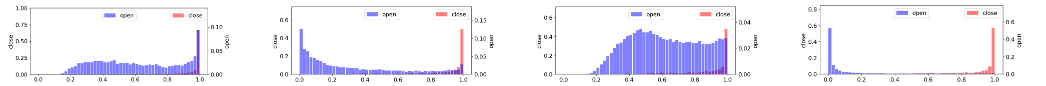
Methods	IN-c	IN-r	LS-c	LS-r
OpenMax (Bendale & Boult)	66.0	68.4	65.7	66.8
OSRCI (Neal et al.)	63.6	63.5	65.0	64.8
LadderNet+Openmax (Yoshihashi et al.)	65.3	67.0	65.2	65.9
DHRNet+Openmax (Yoshihashi et al.)	65.5	67.5	65.6	66.4
CROSR (Yoshihashi et al.)	72.1	73.5	72.0	74.9
C2AE (Oza & Patel)	83.7	82.6	80.6	80.1
CGDL (Sun et al.)	84.0	83.2	80.6	81.2
PROSER (Zhou et al.)	84.9	82.4	86.7	85.6
GCT2 (Ours)	**87.0**	**84.2**	**87.5**	**88.5**

Under the setting that $N_{max} = 10$, $N_o = 10$, $L = 10$ and $\tau_{max} = 0.90$, our method achieves excellent performance as shown in Table 4. It demonstrates that by introducing open negative label words into prompt tuning, the label bias has been mitigated with the closed-set classification ability preserved, contributing to the superior performance both on closed-set classification and unknown recognition.

4.6 Effect Study of Open Negative Label Words

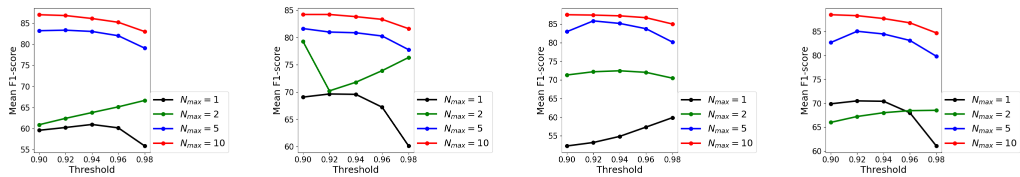
In our method, we introduce additional open negative label words into group-specific contrastive tuning to mitigate the label bias and regularize the predictions. To see how these words affect the unknown detection performance, here we show results on CIFAR100 and ImageNet-100. Results on other benchmarks are shown in the supplementary. As demonstrated in our supplementary, when N_{max} is set to 20, the best performance is achieved when $N_o = 40$ for CIFAR100 and $N_o = 20$ for ImageNet-100. By this division, CIFAR100 is composed of only 1 group, ImageNet-100 is composed of 5 independent groups. The comparison of the distributions on p^i_{max} between tuning with and without open negative label words on CIFAR100 and the first group of ImageNet-100 are shown in Fig. 3.

When N_o is 0 which stands for no open negative label word is introduced into tuning, the distributions of closed-set and open-set data exhibit severe overlap on the higher side. It shows the significant label bias of prompt that test images are prompted to be predicted as known classes with high probability, which inevitably harms the unknown detection. In contrast, when introducing additional open negative label words into tuning, the distributions of closed-set and open-set data are clearly sepa-



(a) CIFAR100, $N_o = 0$. (b) CIFAR100, $N_o = 40$. (c) ImageNet-100, $N_o = 0$ (d) ImageNet-100, $N_o = 20$.

Figure 3: Distributions of the maximal similarity p^i_{max} between each image to the group-specific closed-set classes in the unknown detection experiments. The comparisons on the CIFAR100 and the first group of ImageNet-100 are used.



(a) IN-c as the open-set. (b) IN-r as the open-set. (c) LS-c as the open-set. (d) LS-r as the open-set.

Figure 4: Ablation study on N_{max} in open-set recognition on the CIFAR10 benchmark. We use IN-c, IN-r, LS-c, and LS-r to denote ImageNet-crop/-resize, LSUN-crop/-resize.

Table 5: Ablation study by AUROC on N_{max} in unknown detection experiments on small datasets.

AUROC	CIFAR10	CIFAR+10	CIFAR+50	AUROC	TinyImageNet	CIFAR100
$N_{max} = 1$	94.9	92.3	94.4	$N_{max} = 1$	78.7	79.3
$N_{max} = 2$	93.6	92.2	92.9	$N_{max} = 5$	85.2	78.5
$N_{max} = 6$	**96.1**	**96.1**	**96.2**	$N_{max} = 10$	87.2	**86.3**
$N_{max} > 6$	**96.1**	**96.1**	**96.2**	$N_{max} = 20$	**88.2**	86.2

rated. Unknown samples are much less confident to be predicted as known, by which the label bias of prompts has been mitigated to achieve great unknown detection performance. The results reveal the effectiveness of the proposed GCTu which mitigates the label bias in prompt engineering and contributes to the great performance of unknown detection in the open world.

4.7 Ablation Study on the Maximum Number of Classes in Each Group N_{max}

Setting the length of prompt L to 10 with $[CLASS]$ placed in the medium, we investigate the impact that grouping brings to small and large-scale datasets in OSR tasks. More groups stand for more unused tokens are utilized, i.e., more embedding parameters are optimized in tuning.

Ablation study of unknown detection on small-scale datasets. In this part, we analyze how grouping affects unknown detection on small-scale datasets. As the known classes in CIFAR10 and CIFAR+10/50 are no more than 10, we set the N_{max} to 1, 2, 6 for comparison. TinyImageNet and CIFAR100 both include 20 known classes, thus we set the N_{max} to 1, 5, 10, 20 for comparison. Specifically, the group-specific prompts is also the class-specific prompts when N_{max} is set to 1. Larger N_{max} represents fewer groups that the dataset is divided into. The results in Table 5 show that when there is only 1 group, the unknown detection performance on almost all small-scale datasets is the best. It reveals that only one prompt, which corresponds to the special case that the datasets are composed of only 1 group, is enough for small-scale datasets. More than one prompt leads to more complex group-combined prediction.

Table 6: Ablation study on N_{max} in unknown detection experiments on large-scale datasets. We use IN to represent ImageNet.

AUROC	IN-100	IN-200	IN-LT
$N_{max} = 10$	96.9	94.0	78.8
$N_{max} = 20$	**98.1**	**95.5**	**81.9**
$N_{max} = 30$	97.8	94.3	79.7
$N_{max} = 40$	97.8	94.3	80.2
$N_{max} = 1000$	97.4	91.9	72.5

Ablation study of unknown detection on large-scale datasets. The unknown detection performance on large-scale datasets setting N_{max} to 10, 20, 30, 40 and 1000 is compared in Table 6. The case $N_{max} = 1000$ refers to the case that only one prompt is devised to all closed-set classes of each large-scale dataset as defined in Eq. 1. When N_{max} is set to 20, the AUROC is the highest. When the number of classes in a group increases a lot, the prediction within each group is more difficult, which leads to poorer performance. When only one prompt is devised for all classes, the results are almost the worst. In contrast, by dividing large-scale datasets into tiny groups which include 10 classes at most, the combined prediction is more complex. The results show that grouping with rational N_{max} contributes to the comprehensive performance on large-scale datasets. As a general law, when setting N_{max} as 20, the performance both on small and large-scale datasets are the best.

Ablation study on CIFAR10. To study the impact the grouping causes to the comprehensive performance of closed-set classification and unknown recognition, we set the N_{max} to 1, 2, 5, 10 on the same benchmark in Table 4. The mF1-score comparison for $\tau \in [0.90, 0.92, 0.94, 0.96, 0.98]$ are reported in Fig. 4. Obviously, the best performance is achieved when $N_{max} = 10$ and $\tau_{max} = 0.90$. From the grouping perspective, the conclusion is the same as that in ablation studies of unknown detection on small-scale datasets that dividing small-scale datasets into more than one group makes the prediction more complex and harms the performance.

Ablations in this part validates that grouping achieves the strong applicability on small and large-scale datasets. As a special case, one group is better for small-scale datasets. More groups are more suitable for large-scale datasets. *As a general law, in our paper, the best performance on small and large-scale datasets are both achieved by setting $N_{max} = 20$.* The wide applicability of our method has been verified.

4.8 The Effect of Prompt Length and Group Number for Large-Scale Datasets

In our main experiments, the length of each prompt L is set by 10. The number of adopted unused tokens increases with the number of groups, leading to more embedding parameters can be optimized. In this part, we aim at investigating whether the performance gains come from the increase of unused tokens or groups. We keep the number of unused tokens adopted for each large-scale dataset the same across all settings. The length of the group-specific prompts changes together with the number of classes in a group. More classes in a group lead to fewer groups and longer group-specific prompts. To mitigate the impact caused by the number of open negative label words and achieve a general law, the unknown detection performance is evaluated by average among 4 trials setting N_o to 10, 20, 40 and 60. The settings and results are compared in Table 7.

Table 7: Comparisons of AUROC between longer prompts and more groups in unknown detection on large-scale datasets. The numbers of unused tokens in all settings are same for each dataset.

ImageNet-100	TokenNum	N_{max}	G	L	AUROC
Setting 1	100	5	20	5	**97.1**
Setting 2	100	10	10	10	96.9
Setting 3	100	20	5	20	96.5
Setting 4	100	30	4	25	96.5
ImageNet-200	**TokenNum**	N_{max}	G	L	**AUROC**
Setting 1	200	5	40	5	**94.5**
Setting 2	200	10	20	10	93.7
Setting 3	200	20	10	20	92.0
Setting 4	196	30	7	28	90.9
ImageNet-LT	**TokenNum**	N_{max}	G	L	**AUROC**
Setting 1	500	20	50	10	**77.9**
Setting 2	500	40	25	20	75.2
Setting 3	500	50	20	25	74.3
Setting 4	500	100	10	50	74.2

Results in setting 1 are the best. The performance drops with the increase of N_{max} and L. It reveals that longer prompts are not the reason of improving unknown detection on large-scale datasets. Making predictions on large groups with more classes is hard and complex. In contrast, though the prompts in setting 1 are equipped with fewer unused tokens, the grouping strategy contributes to the excellent performance by combining predictions on multiple groups with a few classes. The results validate the effectiveness and necessity of grouping on large-scale datasets.

5 Conclusion

In this paper, we aim at exploring a new paradigm for solving the OSR problem by prompt engineering on pre-trained VL models, in which an universal data grouping strategy is devised. We firstly process the closed-set data into the combination of one or more groups. The Group-specific Contrastive Tuning (GCTu) is devised to mitigate the label bias of prompts by introducing open negative label words from the built label pool for regularizing the predictions. Then, to make comprehensive predictions combining sub-predictions of each group, the Group Combined Testing (GCTe) is developed. Our method performs competitively across datasets including ImageNet, validating the effectiveness of the proposed new paradigm for OSR.

REFERENCES

Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, pp. 1893–1902, 2015.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, pp. 1563–1572, 2016.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *ACL/IJCNLP*, pp. 1860–1874, 2021.

Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *ECCV*, pp. 507–522. Springer, 2020a.

Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE TPAMI*, 2021.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pp. 104–120. Springer, 2020b.

David Andrew Clifton, Samuel Hugueny, and Lionel Tarassenko. Novelty detection with multivariate extreme value statistics. *Journal of signal processing systems*, 65(3):371–389, 2011.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In *AAAI*, 2022.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *NIPS*, 34:7068–7081, 2021.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *NIPS*, 33:6616–6628, 2020.

Chuanxing Geng and Songcan Chen. Collective decision for open set recognition. *IEEE TKDE*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

Heiko Hoffmann. Kernel pca for novelty detection. *Pattern recognition*, 40(3):863–874, 2007.

Lalit P Jain, Walter J Scheirer, and Terrance E Boult. Multi-class open set recognition using probability of inclusion. In *ECCV*, pp. 393–409. Springer, 2014.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pp. 4904–4916. PMLR, 2021.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *TACL*, 8:423–438, 2020.

Pedro R Mendes Júnior, Roberto M De Souza, Rafael de O Werneck, Bernardo V Stein, Daniel V Pazinato, Waldir R de Almeida, Otávio AB Penatti, Ricardo da S Torres, and Anderson Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3):359–386, 2017.

Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *ICCV*, pp. 813–822, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. pp. 32–33, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pp. 3045–3059, 2021.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, volume 34, pp. 11336–11344, 2020.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NIPS*, 34, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL/IJCNLP*, pp. 4582–4597, 2021.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pp. 2537–2546, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NIPS*, 32, 2019.

Jing Lu, Yunxu Xu, Hao Li, Zhanzhan Cheng, and Yi Niu. Pmal: Open set recognition via robust prototype mining. *AAAI*, 2022.

Benjamin J Meyer and Tom Drummond. The importance of metric learning for robotic vision: Open set recognition and active learning. In *ICRA*, pp. 2924–2931. IEEE, 2019.

Dimity Miller, Niko Sunderhauf, Michael Milford, and Feras Dayoub. Class anchor clustering: A loss for distance-based open set recognition. In *WACV*, pp. 3570–3578, 2021.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *ECCV*, pp. 613–628, 2018.

Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *CVPR*, pp. 2307–2316, 2019.

Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *CVPR*, pp. 11814–11823, 2020.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *EMNLP-IJCNLP*, pp. 2463–2473, 2019.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. Bert is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised qa. *arXiv preprint arXiv:1911.03681*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Ethan M Rudd, Lalit P Jain, Walter J Scheirer, and Terrance E Boult. The extreme value machine. *IEEE TPAMI*, 40(3):762–768, 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE TPAMI*, 35(7):1757–1772, 2012.

Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE TPAMI*, 36(11):2317–2324, 2014.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, pp. 4222–4235, 2020.

Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *CVPR*, pp. 13480–13489, 2020.

Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, Qing Yang, and Cheng-Lin Liu. Convolutional prototype network for open set recognition. *IEEE TPAMI*, 2020.

Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *CVPR*, pp. 4016–4025, 2019.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

He Zhang and Vishal M Patel. Sparse representation-based open set recognition. *IEEE TPAMI*, 39 (8):1690–1696, 2016.

Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In *ECCV*, pp. 102–117. Springer, 2020.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pp. 5579–5588, 2021.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *ICML*, pp. 12697–12706. PMLR, 2021.

Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *CVPR*, pp. 4401–4410, 2021a.

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

A Unknown detection performance evaluated by AUROC on small-scale datasets

The AUROC results averaged among five randomized trials together with standard deviation are shown in Table 8. We can see that our method achieves excellent performance. It validates the efficacy of the proposed new paradigm which solves OSR on small-scale datasets by prompt engineering with the label bias being eliminated.

Table 8: Unknown detection performance evaluated by AUROC on small-scale datasets. Results are averaged among 5 randomized trials.

Methods	CIFAR10	CIFAR+10	CIFAR+50	TinyImageNet	CIFAR100
OSRCI (Neal et al.)	69.9±3.8	83.8±N.R.	82.7±N.R.	58.6±N.R.	N.R.
CGDL (Sun et al.)	90.3±0.9	95.9±0.6	95.0±0.6	76.2±0.5	N.R.
GDFR (Perera et al.)	83.1±3.9	92.8±0.2	92.6±0.0	60.8±1.7	N.R.
C2AE (Oza & Patel)	89.5±0.8	95.5±0.6	93.7±0.4	74.8±0.5	N.R.
PROSER (Zhou et al.)	89.1±1.6	96.0±0.4	95.3±0.3	69.3±0.5	N.R.
CPN (Yang et al.)	82.8±2.1	88.1±N.R.	87.9±N.R.	63.9±N.R.	N.R.
RPL (Chen et al.)	90.1±N.R.	97.6±N.R.	96.8±N.R.	80.9±N.R.	N.R.
ARPL+CS (Chen et al.)	91.0±N.R.	97.1±N.R.	95.1±N.R.	78.2±N.R.	N.R.
PMAL (Lu et al.)	95.1±N.R.	97.8±N.R.	96.9±N.R.	83.1±N.R.	N.R.
ZOC (Esmaeilpour et al.)	93.0±1.7	**97.8**±0.6	**97.6**±0.0	84.6±1.0	82.1±2.1
GCT2 (Ours)	**96.1**±0.7	96.1±0.8	96.2±0.4	**88.2**±1.4	**86.2**±1.3

Table 9: Unknown detection performance evaluated by AUROC with different numbers of open negative label words N_o.

AUROC	CIFAR10	CIFAR+10	CIFAR+50	TinyImageNet	CIFAR100	ImageNet-100	ImageNet-200	ImageNet-LT
$N_o = 5$	**96.1**	94.7	94.3	85.7	84.4	97.0	93.4	72.3
$N_o = 10$	94.9	95.1	94.1	85.8	84.9	97.4	93.9	75.1
$N_o = 20$	93.7	95.1	95.3	86.6	85.4	**98.1**	94.9	77.1
$N_o = 40$	93.9	**96.1**	**96.2**	85.1	**86.2**	97.7	**95.5**	77.7
$N_o = 60$	94.7	94.9	93.7	**88.2**	85.9	97.9	93.8	**81.9**

B Ablation Study on the Number of Open Negative Label Words

Under the setting that $N_{max} = 20$ and $L = 10$, the detailed comparison of unknown detection performance measured by AUROC with different numbers of open negative label words are shown in Table 9. In addition, the distributions of p^i_{max} on each benchmark with/without introducing open negative label words into tuning are shown in Fig. 5.

The results and distribution comparisons reveal that the significant label bias prevents recognizing unknown classes correctly when $N_o = 0$, which stands for prompt tuning without introducing open negative label words. After introducing open negative label words, the performance is improved. The difference in the distributions between closed-set and open-set data has been obviously widened after introducing the open negative label words. The effectiveness of the Group-specific Contrastive Tuning (GCTu) has been verified. It successfully addresses the label bias of prompt engineering and contributes to superior unknown detection performance.

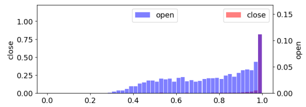
C Study on the Effect of Prompt Length and Group Number for Large-Scale Datasets

In this part, we deliver the comparison results on joint closed-set classification for supplementation. The joint closed-set classification performance evaluated by accuracy is averaged among 4 trials by setting N_o to 10, 20, 40 and 60. Results are shown in Table 10.

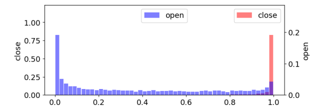
Results show that longer prompts are not the reason for improving the joint closed-set classification on large-scale datasets. After dividing the large-scale datasets into small groups, the group-specific tuning and inference are simplified for fewer classes within each group. Classification performance is better in groups with fewer classes.

D Study on the Definition of S_{open}

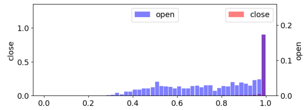
In this paper, we define the score measuring a sample being unknown as one minus the sum of closed-set probabilities. The intuition is that the labels participated into group-specific tuning include the group-wise closed-set labels and additional open negative label words, in which the sum of the



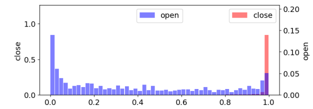
(a) CIFAR10, $N_o = 0$.



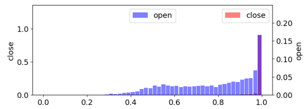
(b) CIFAR10, $N_o = 5$.



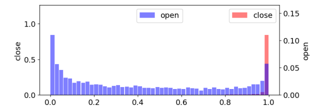
(c) CIFAR+10, $N_o = 0$.



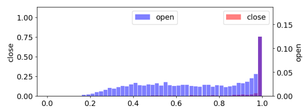
(d) CIFAR+10, $N_o = 40$.



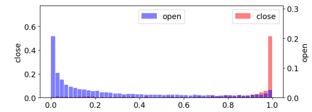
(e) CIFAR+50, $N_o = 0$.



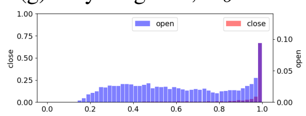
(f) CIFAR+50, $N_o = 40$.



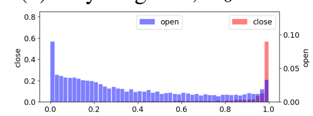
(g) TinyImageNet, $N_o = 0$.



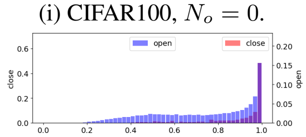
(h) TinyImageNet, $N_o = 60$.



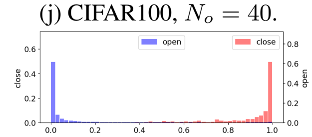
(i) CIFAR100, $N_o = 0$.



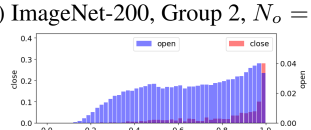
(j) CIFAR100, $N_o = 40$.



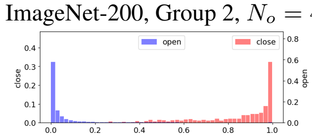
(k) ImageNet-200, Group 2, $N_o = 0$.



(l) ImageNet-200, Group 2, $N_o = 40$.



(m) ImageNet-LT, Group 13, $N_o = 0$.



(n) ImageNet-LT, Group 13, $N_o = 60$.

Figure 5: The distributions of the maximal similarity p^i_{max} between each image to the group-specific closed-set classes in the unknown detection experiments.

```
# Closed_set - List of closed-set label words
# model     - CLIP model

txt = torch.cat([clip.tokenize(f"a photo of a {c}") for c in Closed_set])
with torch.no_grad():
    txt_feats = model.encode_text(txt)

sim = (100.0 * txt_feats @ txt_feats.T).softmax(dim=-1)

semantic_order = [0]
for i in range(sim.shape[0]-1):
    if i == 0:
        sub_sim = sim[i, :]
        sub_sim[i] = -100
        semantic_order.append(torch.argmax(sub_sim))
    else:
        sub_sim = sim[semantic_order[i], :]
        sub_sim[semantic_order[i]] = -100

        while torch.argmax(sub_sim) in semantic_order:
            sub_sim[torch.argmax(sub_sim)] = -100

        if not (torch.argmax(sub_sim) in semantic_order):
            semantic_order.append(torch.argmax(sub_sim))
```

Figure 6: Torch-like pseudocode for semantic similarity sorting process.

Table 10: Comparisons of closed-set accuracy between longer prompts and more groups in unknown detection on large-scale datasets. The numbers of unused tokens are kept the same across all settings for each dataset. Results are averaged among 4 trials.

ImageNet-100	TokenNum	N_{max}	G	L	Close-Acc
Setting 1	100	5	20	5	**82.3**
Setting 2	100	10	10	10	78.2
Setting 3	100	20	5	20	76.9
Setting 4	100	30	4	25	77.5
ImageNet-200	TokenNum	N_{max}	G	L	Close-Acc
Setting 1	200	5	40	5	**85.7**
Setting 2	200	10	20	10	79.5
Setting 3	200	20	10	20	76.3
Setting 4	196	30	7	28	71.9
ImageNet-LT	TokenNum	N_{max}	G	L	Close-Acc
Setting 1	500	20	50	10	**77.97**
Setting 2	500	40	25	20	75.6
Setting 3	500	50	20	25	74.2
Setting 4	500	100	10	50	69.8

Table 11: Comparison to CLIP baseline and definition of score of being unknown in unknown detection.

Methods	CIFAR10	CIFAR+10	CIFAR+50	TinyImageNet	CIFAR100	ImageNet-100	ImageNet-200	IN-LT
CLIP+MSP	87.3	92.2	92.4	83.7	84.0	74.0	78.5	58.1
GCT2+MSP	92.4	94.3	95.0	86.3	85.3	92.2	90.8	79.5
GCT2 (Ours)	**96.1**	**96.1**	**96.2**	**88.2**	**86.2**	**98.1**	**95.5**	**81.9**

probabilities on open negative labels are naturally regarded as the score of a sample being unknown. Thus, we define the S_{open} as in Eq. 7 in the main paper.

As the supplement explanation, we conduct comparison experiments on the definition of S_{open}. The compared one is based on maximum over softmax probabilities (MSP) (Hendrycks & Gimpel, 2017) denoted as S_{open}^{MSP}:

$$S_{open}^{MSP} = 1 - p_{max}^{I_{opt}}. \tag{9}$$

Results delivered in Table 11 show that our definition is more suitable with our method. In addition, based on the same definition of the score of being open by using MSP, the results achieved by GCT2 in the second row all surpass the results achieved by vanilla OSR method built on CLIP in the first row. It further validates the efficacy of our proposed method.

E Ablation Study on the Grouping Strategy

As studied in Section 4.7, one group is better for small-scale datasets, while multiple groups with rational N_{max} (maximum number of classes per group) contributes to the great performance on large-scale datasets. Grouping on small-scale datasets result in no more than 5 classes within each group due to the small number of classes. Therefore, in this section, we only perform the ablation study on the grouping strategy on large-scale datasets. The strategies include grouping by WordNet ID (WNID) order adopted in the main experiments, grouping by randomness and grouping by semantic similarities.

Table 12: Ablation study on grouping strategies in unknown detection on large-scale datasets.

AUROC	ImageNet-100	ImageNet-200	ImageNet-LT
CLIP+MSP (Baseline)	74.0	78.5	58.1
PMAL (Lu et al.)	94.9	93.9	71.7
Random	97.4	91.9	80.7
Semantics	96.0	92.6	76.2
WNID Order	**98.1**	**95.5**	**81.9**

Table 13: Closed-set classification accuracy comparison.

Methods	CIFAR10	CIFAR+10	CIFAR+50	TinyImageNet	CIFAR100	ImageNet-100	ImageNet-200	ImageNet-LT
SoftMax	80.1	N.R.	N.R.	N.R.	N.R.	81.7	79.7	37.8
CPN (Yang et al.)	92.9	94.8	95.0	81.4	N.R.	86.1	82.1	37.1
CGDL (Sun et al.)	91.2	N.R.	N.R.	N.R.	N.R.	N.R.	N.R.	N.R.
RPL (Chen et al.)	95.1	95.5	95.9	81.7	N.R.	81.8	80.7	39.0
ARPL (Chen et al.)	87.9	94.7	92.9	65.9	N.R.	N.R.	N.R.	N.R.
PMAL (Lu et al.)	97.5	**97.8**	**98.1**	84.7	N.R.	**86.2**	84.1	42.9
GCT2 (Ours)	**97.8**	96.2	95.8	**87.3**	**87.2**	82.7	**85.2**	**78.0**

In the semantics grouping strategy, we sort the classes by the similarities on their text embeddings extracted by the text encoder of CLIP Radford et al. (2021). We select the first class in closed-set label words as the start one in the semantic order, the class with the highest similarity to the previous class is then appended to the semantic similarity sorted class list. Details of the ordering is shown in Fig. 6. Therefore, any two categories that are adjacent in the semantic order list are the ones with the highest semantic similarity.

The ablation experiments are conducted with $N_{max} = 20$. Results of unknown detection evaluated under different grouping strategies together with the results of method PMAL Lu et al. (2022) and CLIP baseline are delivered in Table 12. The ablation study show that grouping by WNID order achieves the best results. Even though random grouping and semantic grouping are inferior to grouping by WNID, the results are still competitive and far better than those of CLIP baseline. We analyze the reason as that WNID order contributes the optimal inter-class split both within and across all groups, by which the classes within a group are easily to be classified, the optimal prompts are easily to be selected without confusion. It validates the effectiveness of GCT2 and the grouping strategy guided by WNID order.

F Closed-set accuracy in unknown detection experiments

Taking the results in PMAL Lu et al. (2022), the closed-set classification accuracy in the unknown detection experiments are compared in Table 13. The results show that our method achieves competitive closed-set classification performance, which demonstrates the efficacy of solving OSR by promptv tuning with label bias mitigated.