

# DOMAIN RECENTERING AND CONFIDENCE-WEIGHTED PRIOR CALIBRATION FOR VISION-LANGUAGE MODELS

Youngeun Seol\*

Jimin Shin\*

Heeseo Yoon\*

Uiwon Hwang<sup>†</sup>

Department of Computer Science and Engineering, Ewha Womans University

## ABSTRACT

Vision-language models such as CLIP achieve strong zero-shot classification, yet under distribution shift, visual embeddings drift from fixed text embeddings. Training-free calibration avoids the per-sample optimization of prompt learning, but prior feature calibration gives each image the full bias of one hard cluster. We propose **Domain Recentering with Confidence Calibration (DRC)**, a training-free method adapting CLIP from a set of unlabeled target images. DRC fits a Gaussian mixture once and subtracts from each embedding a posterior-weighted average of component means. It then removes residual class preference with a log-prior correction, estimating the prior from confidence-weighted predictions. Among compared methods, DRC achieves the highest average accuracy on cross-domain datasets, exceeding zero-shot CLIP by 4.13 and 5.07 points with ViT-B/16 and ResNet-50, with gains over CLIP also holding under ImageNet distribution shifts.

**Index Terms**— Vision-language models, domain recentering, confidence-weighted calibration

## 1. INTRODUCTION

Vision-language models have emerged as powerful foundation models for visual recognition. CLIP [1] learns a multi-modal embedding space for images and text from natural language supervision at scale, enabling zero-shot classification by aligning an image’s embedding with the text embeddings of class descriptions. However, when target images differ from the pretraining distribution, their visual embeddings can become misaligned with the fixed text embeddings, degrading classification accuracy.

Recent works have adapted CLIP under distribution shift along two broad directions. One line performs prompt learning with few-shot labeled data [2, 3] or test-time entropy minimization [4, 5], requiring labels or per-image backpropagation. Another line calibrates features without training, using unlabeled target data. UMFC [6] corrects visual encoder bias by subtracting from each feature the mean of its hard-assigned

cluster. However, a feature near a cluster boundary then receives one cluster’s full mean, so a small change in the feature can flip its correction. Correcting the image features alone also leaves the text classifier unchanged, so any preference it holds for certain classes persists.

We propose Domain Recentering with Confidence Calibration (DRC), which models the structure of unlabeled CLIP visual features using a Gaussian mixture model. Rather than assigning each image to a single cluster, DRC uses mixture-component posterior probabilities to construct a soft, image-specific bias vector as a weighted combination of component means, so the correction changes continuously as the feature moves between components. This vector is subtracted from the image feature to suppress domain-related variation while preserving class-discriminative content. DRC further corrects residual class preferences through confidence-weighted prior calibration, without any backpropagation or parameter updates. Among the compared methods, DRC achieves the highest average accuracy on ten cross-domain datasets with both ResNet-50 and ViT-B/16, and it also improves the average accuracy of zero-shot CLIP over four ImageNet distribution shifts.

## 2. RELATED WORK

Prompt learning adapts CLIP [1] to downstream data by tuning its textual prompts. CoOp [2] and CoCoOp [3] learn these prompts from labeled few-shot data. Test-time adaptation instead uses unlabeled test samples to improve robustness under distribution shifts. TPT [4] optimizes prompts for each test sample, and DiffTPT [5] extends TPT with diffusion-based augmentation. Both repeat backpropagation for every test sample, which makes inference costly and motivates training-free adaptation that leaves the pretrained model unchanged.

Training-free methods instead correct the systematic biases that zero-shot vision-language models exhibit in representations and predictions [6, 7]. UMFC [6] subtracts each hard-assigned cluster’s mean, giving all images within a cluster the same correction. Label-free logit adjustment [8] offsets label bias in predictions. Our method instead assigns each image its own bias by weighting Gaussian mixture component means with the image’s soft assignment. It then estimates the class prior from the recentered predictions with confidence

\* Equal contribution, listed in alphabetical order

<sup>†</sup>Corresponding author: Uiwon Hwang (uiwon.hwang@ewha.ac.kr)

weighting, so that the prior targets the bias left after feature correction.

### 3. METHOD

#### 3.1. Background

**CLIP.** CLIP [1] maps images and text into a shared embedding space via a visual encoder  $E_v$  and a text encoder  $E_t$ . For zero-shot classification, each class name  $y_c$  is inserted into prompt templates and encoded by  $E_t$ , and the averaged embedding serves as the classifier weight  $\mathbf{t}_c$ . Given an image  $x_i$ , its visual feature  $\mathbf{v}_i = E_v(x_i)$  is compared to each  $\mathbf{t}_c$  via scaled cosine similarity

$$\ell_{ic} = \exp(\tau) \left( \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2} \right)^\top \frac{\mathbf{t}_c}{\|\mathbf{t}_c\|_2}, \quad (1)$$

where  $\exp(\tau)$  is the learned logit scale of CLIP, clipped at 100. CLIP predicts the class with the highest logit  $\ell_{ic}$ , and under domain shift,  $\mathbf{v}_i$  can become misaligned with the fixed  $\mathbf{t}_c$ .

**Problem Setting.** We adapt a frozen CLIP model to a target domain whose images may differ from the pretraining distribution. An unlabeled adaptation set  $\mathcal{D}_u = \{x_i\}_{i=1}^N$  of target-domain images is given. DRC estimates its statistics once from  $\mathcal{D}_u$  without any target labels and then classifies each test image with these fixed statistics. The visual encoder  $E_v$ , the text encoder  $E_t$ , and the classifier weights  $\mathbf{t}_c$  remain unchanged throughout. Figure 1 illustrates the DRC pipeline, which recenters each image feature with a Gaussian mixture fitted on  $\mathcal{D}_u$  and then calibrates the logits with a confidence-weighted class prior.

#### 3.2. Domain Recentering

DRC models the features of  $\mathcal{D}_u$  with a Gaussian mixture and estimates the bias of each image from all mixture components, weighting each component by how strongly the image belongs to it.

Each image, whether in  $\mathcal{D}_u$  or at test time, is represented by averaging its  $\ell_2$ -normalized feature with that of its horizontal flip, followed by re-normalization

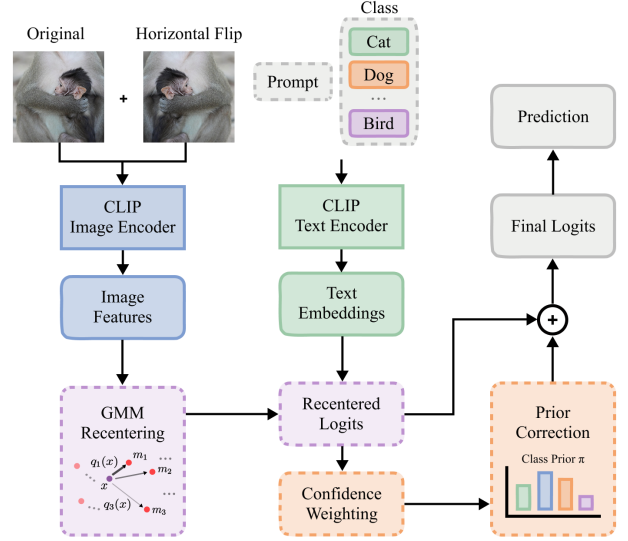
$$\mathbf{f}_i = \text{norm} \left( \text{norm}(E_v(x_i)) + \text{norm}(E_v(\text{Flip}(x_i))) \right). \quad (2)$$

With the mean  $\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{f}_i$  over  $\mathcal{D}_u$ , the centered features are projected via PCA onto  $\min(16, D, N)$  dimensions, where  $D$  is the CLIP embedding dimension, so that the mixture is fitted on the dominant directions of variation

$$\mathbf{z}_i = \mathbf{P}(\mathbf{f}_i - \boldsymbol{\mu}). \quad (3)$$

A diagonal-covariance GMM with  $K$  components,

$$p(\mathbf{z}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{z}; \boldsymbol{\nu}_k, \boldsymbol{\Sigma}_k), \quad (4)$$



**Fig. 1.** Overview of DRC. DRC recenters CLIP image features with a GMM-based soft bias correction and calibrates the resulting logits with a confidence-weighted class prior.

is fitted to  $\{\mathbf{z}_i\}$ , where  $\pi_k$ ,  $\boldsymbol{\nu}_k$ , and  $\boldsymbol{\Sigma}_k$  denote the mixture weight, mean, and diagonal covariance of component  $k$ . Unlike hard clustering such as K-means, the GMM assigns each feature softly to the  $K$  components through its posterior  $q_{i,k} = p(k | \mathbf{z}_i)$ .

Each component mean is computed in the original embedding space from the features whose most probable component is  $k$ ,  $\mathcal{I}_k = \{i | k = \arg \max_j q_{i,j}\}$ , so that it summarizes the features belonging mainly to that component

$$\mathbf{m}_k = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \mathbf{f}_i. \quad (5)$$

For an input feature  $\mathbf{f}$ , DRC estimates its bias by weighting these means with the posterior

$$\mathbf{b}(\mathbf{f}) = \sum_{k=1}^K q_k(\mathbf{f}) \mathbf{m}_k, \quad q_k(\mathbf{f}) = p(k | \mathbf{P}(\mathbf{f} - \boldsymbol{\mu})). \quad (6)$$

Since  $q_k(\mathbf{f})$  varies continuously with  $\mathbf{f}$ , the bias changes gradually as a feature moves between components, whereas a hard assignment would switch it abruptly at component boundaries.

The recentered feature is obtained by subtracting this bias and re-normalizing

$$\hat{\mathbf{f}} = \frac{\mathbf{f} - \beta \mathbf{b}(\mathbf{f})}{\|\mathbf{f} - \beta \mathbf{b}(\mathbf{f})\|_2}, \quad (7)$$

where  $\beta$  controls the strength of recentering.

### 3.3. Confidence-Weighted Prior Calibration

Recentring acts only on the image features and leaves the classifier weights  $\mathbf{t}_c$  unchanged, so the class preference of the zero-shot classifier [8, 7] can persist in the recentered predictions. DRC estimates this preference from predictions on  $\mathcal{D}_u$  and removes it with a log-prior adjustment, weighting each prediction by its confidence.

For each  $x_i \in \mathcal{D}_u$ , let  $\hat{\ell}_{ic}$  be the logit from the recentered feature. The class probability is then

$$p_i(c) = \frac{\exp(\hat{\ell}_{ic})}{\sum_{c'} \exp(\hat{\ell}_{ic'})}. \quad (8)$$

Since an uncertain prediction provides weak evidence of class preference, each prediction is weighted by one minus its normalized entropy

$$H_i = - \sum_c p_i(c) \log p_i(c), \quad w_i = 1 - \frac{H_i}{\log C}, \quad (9)$$

where  $C$  is the number of classes, so that near-uniform predictions receive weights close to zero. The confidence-weighted class prior is

$$\pi_c^{\text{Conf}} = \frac{\sum_{i=1}^N w_i p_i(c)}{\sum_{i=1}^N w_i}. \quad (10)$$

The computed distribution summarizes the class preference reflected in reliable predictions on the adaptation set. A large  $\pi_c^{\text{Conf}}$  indicates that class  $c$  receives a relatively large amount of prediction mass, while a small value denotes that the class is less frequently preferred by the model. Confidence weighting reduces the influence of uncertain predictions that could otherwise distort this estimate.

The log-prior correction and its zero-centered form are

$$r_c = -\log(\pi_c^{\text{Conf}} + \epsilon), \quad \tilde{r}_c = r_c - \frac{1}{C} \sum_{c'=1}^C r_{c'}, \quad (11)$$

where  $\epsilon$  is a small constant for numerical stability. A test image with recentered feature  $\hat{\mathbf{f}}$  is classified with the calibrated logits

$$\ell'_c = \hat{\ell}_c + \tilde{r}_c, \quad \hat{\ell}_c = \exp(\tau) \hat{\mathbf{f}}^\top \frac{\mathbf{t}_c}{\|\mathbf{t}_c\|_2}. \quad (12)$$

This adjustment lowers the logits of classes that receive excessive prediction mass on  $\mathcal{D}_u$  and raises those of less preferred classes.

## 4. EXPERIMENTS

### 4.1. Experimental Setup

**Datasets.** We evaluate DRC on the cross-domain benchmark and the out-of-distribution (OOD) benchmark. For

cross-domain evaluation, we use the same ten datasets as TPT [4], which are Aircraft, Caltech101, Cars, DTD, EuroSAT, Flower102, Food101, Pets, SUN397, and UCF101. The OOD benchmark measures robustness to distribution shifts with four variants of ImageNet [9], ImageNet-A [10], ImageNet-V2 [11], ImageNet-R [12] and ImageNet-Sketch [13].

**Implementation details.** Text embeddings obtained from dataset-specific prompt templates are averaged into a single vector for each class, and image features are extracted from each image and its horizontal flip. DRC uses PCA and a diagonal-covariance GMM ( $n_{\text{init}} = 5$ , seed 42). The adaptation set  $\mathcal{D}_u$  is the validation split of each cross-domain dataset and the unlabeled evaluation images for ImageNet and its variants.  $K$  and  $\beta$  are selected by accuracy on the validation split of each cross-domain dataset, and on the ImageNet validation set for the OOD benchmark, with the ImageNet values reused for all four variants.

### 4.2. Comparison with State-of-the-Art

**Cross-Domain Benchmark.** Table 1 evaluates DRC across ten datasets with disjoint class spaces. DRC achieves the highest average accuracy with both CLIP-ViT-B/16 and CLIP-ResNet-50, reaching 68.72% and 61.70%, respectively. These results improve over zero-shot CLIP by 4.13 and 5.07 points, and DRC ranks first on 7 and 6 of the 10 datasets, respectively. On EuroSAT, DRC improves over the best competing method by 10.75 and 3.80 points, and on UCF101, the gains are 3.52 and 2.33 points, respectively.

**OOD Benchmark.** Table 2 further evaluates DRC on the OOD benchmark. With CLIP-ViT-B/16, DRC achieves the highest Average and OOD Average of 62.79% and 60.94%. With CLIP-ResNet-50, DRC achieves 48.53% and 45.14%, improving over zero-shot CLIP by 2.10 and 2.05 points, respectively. The gains are particularly pronounced under strong appearance shifts. DRC achieves the highest accuracy on ImageNet-R/S for both backbones (61.39%/39.94% for CLIP-ResNet-50, 78.14%/52.09% for CLIP-ViT-B/16).

### 4.3. Ablation Study

We perform ablation studies on the cross-domain benchmark with CLIP-ViT-B/16 to examine the effectiveness of domain recentering and confidence-weighted prior calibration.

As shown in Table 3, both components improve over the CLIP baseline. Soft recentering outperforms hard recentering both without and with prior calibration. Confidence-weighted prior calibration provides the largest gain, increasing the average accuracy to 68.35%. Combining both components yields the best result, with DRC achieving an average accuracy of 68.72%.

| Method            | Aircraft     | Caltech101   | Cars         | DTD          | EuroSAT      | Flower102    | Food101      | Pets         | SUN397       | UCF101       | Average      |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| CLIP-ResNet-50    | 16.11        | 87.26        | 55.89        | 40.37        | 25.79        | 62.77        | 74.82        | 82.97        | 60.85        | 59.48        | 56.63        |
| CoOp              | 15.12        | 86.53        | 55.32        | 37.29        | 26.20        | 61.55        | 75.59        | 87.00        | 58.15        | 59.05        | 56.18        |
| CoCoOp            | 14.61        | <u>87.38</u> | 56.22        | 38.53        | 28.73        | <u>65.57</u> | 76.20        | <b>88.39</b> | 59.61        | 57.10        | 57.23        |
| TPT               | 17.58        | 87.02        | 58.46        | 40.84        | 28.33        | 62.69        | 74.88        | 84.49        | 61.46        | 60.82        | 57.66        |
| DiffTPT           | 17.60        | 86.89        | <b>60.71</b> | 40.72        | 41.04        | 63.53        | <b>79.21</b> | 83.40        | <b>62.72</b> | 62.67        | <u>59.85</u> |
| UMFC              | <u>17.76</u> | 85.48        | 56.25        | <u>42.85</u> | <u>37.63</u> | <u>65.57</u> | 77.47        | 86.10        | 60.49        | <u>63.39</u> | 59.30        |
| <b>DRC (Ours)</b> | <b>18.81</b> | <b>87.99</b> | <u>59.66</u> | <b>44.33</b> | <b>44.84</b> | <b>66.30</b> | <u>78.83</u> | <u>88.01</u> | <u>62.46</u> | <b>65.72</b> | <b>61.70</b> |
| CLIP-ViT-B/16     | 23.22        | 93.55        | 66.11        | 45.04        | 50.42        | 66.99        | 82.86        | 86.92        | 65.63        | 65.16        | 64.59        |
| CoOp              | 18.47        | 93.70        | 64.51        | 41.92        | 46.39        | 68.71        | 85.30        | 89.14        | 64.15        | 66.55        | 63.88        |
| CoCoOp            | 22.29        | <u>93.79</u> | 64.90        | 45.45        | 39.23        | <u>70.85</u> | 83.97        | <u>90.46</u> | <u>66.89</u> | 68.44        | 64.63        |
| TPT               | 24.78        | <b>94.16</b> | 66.87        | <b>47.75</b> | 42.44        | 68.98        | 84.67        | 87.79        | 65.50        | 68.04        | 65.10        |
| DiffTPT           | <u>25.60</u> | 92.49        | <u>67.01</u> | 47.00        | 43.13        | 70.10        | <b>87.23</b> | 88.22        | 65.74        | 68.22        | 65.47        |
| UMFC              | <u>25.47</u> | 91.44        | <u>66.35</u> | 45.57        | <u>50.98</u> | 70.60        | 85.99        | 88.91        | 65.73        | <u>68.91</u> | <u>66.00</u> |
| <b>DRC (Ours)</b> | <b>27.39</b> | 92.17        | <b>68.82</b> | <u>47.46</u> | <b>61.73</b> | <b>71.78</b> | <u>86.80</u> | <b>91.11</b> | <b>67.52</b> | <b>72.43</b> | <b>68.72</b> |

**Table 1.** Results on the Cross-Domain Benchmark with CLIP-ResNet-50 and CLIP-ViT-B/16. CoOp and CoCoOp serve as supervised transfer baselines trained on ImageNet with 16 labeled samples per class. CLIP, CoOp, CoCoOp, and TPT results are taken from the original TPT paper [4], and DiffTPT results are taken from the original DiffTPT paper [5].

| Method            | ImageNet     | -A           | -V2          | -R           | -S           | Average      | OOD Average  |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| CLIP-ResNet-50    | 59.81        | 23.24        | 52.91        | 60.72        | 35.48        | 46.43        | 43.09        |
| CoOp              | <b>63.33</b> | 23.06        | 55.40        | 56.60        | 34.67        | 46.61        | 42.43        |
| CoCoOp            | <u>62.81</u> | 23.32        | <u>55.72</u> | 57.74        | 34.48        | 46.81        | 42.82        |
| TPT               | 60.74        | <u>26.67</u> | 54.70        | 59.11        | 35.09        | 47.26        | 43.89        |
| DiffTPT           | 60.80        | <b>31.06</b> | <b>55.80</b> | 58.80        | <u>37.10</u> | <b>48.71</b> | <b>45.69</b> |
| UMFC              | 59.70        | 23.47        | 52.85        | <u>60.94</u> | 35.82        | 46.56        | 43.27        |
| <b>DRC (Ours)</b> | 62.12        | 23.77        | 55.44        | <b>61.39</b> | <b>39.94</b> | <u>48.53</u> | <u>45.14</u> |
| CLIP-ViT-B/16     | 68.34        | 49.89        | 61.88        | 77.65        | 48.24        | 61.20        | 59.42        |
| CoOp              | <b>71.51</b> | 49.71        | <u>64.20</u> | 75.21        | 47.99        | 61.72        | 59.28        |
| CoCoOp            | <u>71.02</u> | 50.63        | 64.07        | 76.18        | 48.75        | 62.13        | 59.91        |
| TPT               | 68.98        | <u>54.77</u> | 63.45        | 77.06        | 47.94        | 62.44        | 60.81        |
| DiffTPT           | 70.30        | <b>55.68</b> | <b>65.10</b> | 75.00        | 46.80        | 62.28        | 60.52        |
| UMFC              | 68.21        | 50.33        | 61.99        | <u>77.98</u> | <u>48.77</u> | 61.46        | 59.77        |
| <b>DRC (Ours)</b> | 70.17        | 49.71        | 63.83        | <b>78.14</b> | <b>52.09</b> | <b>62.79</b> | <b>60.94</b> |

**Table 2.** Results on the OOD Benchmark with CLIP-ResNet-50 and CLIP-ViT-B/16. CoOp and CoCoOp serve as supervised transfer baselines trained on ImageNet with 16 labeled samples per class. CLIP, CoOp, CoCoOp, and TPT results are taken from the original TPT paper [4], and DiffTPT results are taken from the original DiffTPT paper [5].

| Recentring | Prior calibration |              |
|------------|-------------------|--------------|
|            | without           | with         |
| None       | 64.59             | 68.35        |
| Hard       | 66.89             | 68.57        |
| Soft       | 67.01             | <b>68.72</b> |

**Table 3.** Ablation on the cross-domain benchmark with CLIP-ViT-B/16, reporting the average accuracy over the ten datasets. The bottom-right entry is DRC.

## 5. CONCLUSION

In this paper, we proposed DRC, a training-free method that improves the robustness of CLIP under distribution shift without any backpropagation or parameter updates. DRC fits a Gaussian mixture to unlabeled target features and subtracts from each feature a soft, image-specific bias, so that the correction varies continuously across mixture components. It then corrects class preferences remaining in the recentered predictions with a confidence-weighted class prior. DRC achieves the highest average accuracy among the compared methods on the cross-domain benchmark and improves the OOD Average of zero-shot CLIP with both backbones.

## 6. ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00561169), Global – Learning & Academic research institution for Master’s · PhD students, and Postdocs (G-LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25442252), and Institute of Information & communications Technology Planning & Evaluation (IITP) under the Leading Generative AI Human Resources Development (IITP-2027-RS-2026-25546026) grant funded by the Korea government (MSIT).

## 7. REFERENCES

- [1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.
- [2] Kaiyang Zhou, Jingkan Yang, Chen Change Loy, and Ziwei Liu, “Learning to prompt for vision-language models,” *International journal of computer vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [3] Kaiyang Zhou, Jingkan Yang, Chen Change Loy, and Ziwei Liu, “Conditional prompt learning for vision-language models,” in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 2022, pp. 16795–16804.
- [4] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao, “Test-time prompt tuning for zero-shot generalization in vision-language models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 14274–14289, 2022.
- [5] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo, “Diverse data augmentation with diffusions for effective test-time prompt tuning,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 2704–2714.
- [6] Jiachen Liang, Ruibing Hou, Minyang Hu, Hong Chang, Shiguang Shan, and Xilin Chen, “Umfc: Unsupervised multi-domain feature calibration for vision-language models,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 114072–114093, 2024.
- [7] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong, “The neglected tails in vision-language models,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 12988–12997.
- [8] Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang, “Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 2001–2025, 2024.
- [9] 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*. Ieee, 2009, pp. 248–255.
- [10] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song, “Natural adversarial examples,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15262–15271.
- [11] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar, “Do imagenet classifiers generalize to imagenet?,” in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.
- [12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al., “The many faces of robustness: A critical analysis of out-of-distribution generalization,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 8320–8329.
- [13] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing, “Learning robust global representations by penalizing local predictive power,” *Advances in neural information processing systems*, vol. 32, 2019.