CRoFT: Robust Fine-Tuning with Concurrent Optimization for OOD Generalization and Open-Set OOD Detection

Lin Zhu [1] **Yifeng Yang** [1] **Qinying Gu** [2] **Xinbing Wang** [1] **Chenghu Zhou** [1] **Nanyang Ye** [1]

Abstract

Recent vision-language pre-trained models (VL-PTMs) have shown remarkable success in open-vocabulary tasks. However, downstream use cases often involve further fine-tuning of VL-PTMs, which may distort their general knowledge and impair their ability to handle distribution shifts. In real-world scenarios, machine learning systems inevitably encounter both covariate shifts (e.g., changes in image styles) and semantic shifts (e.g., test-time unseen classes). This highlights the importance of enhancing out-of-distribution (OOD) generalization on covariate shifts and simultaneously detecting semantic-shifted unseen classes. Thus a critical but underexplored question arises: *How to improve VL-PTMs' generalization ability to closed-set OOD data, while effectively detecting open-set unseen classes during fine-tuning?* In this paper, we propose a novel objective function of OOD detection that also serves to improve OOD generalization. We show that minimizing the gradient magnitude of energy scores on training data leads to domain-consistent Hessians of classification loss, a strong indicator for OOD generalization revealed by theoretical analysis. Based on this finding, we have developed a unified fine-tuning framework that allows for concurrent optimization of both tasks. Extensive experiments have demonstrated the superiority of our method. The code is available at `https://github.com/LinLLLL/CRoFT`.

1. Introduction

Recent advances in large-scale vision-language pre-trained models (VL-PTMs), such as CLIP (Radford et al., 2021), Grounding DINO (Liu et al., 2023), MiniGPT-4 (Zhu et al., 2023b), etc., have shown promising results in visual-semantic learning. However, in real-world scenarios, machine learning models often face challenges related to out-of-distribution (OOD) data, arising from disparities in data distributions between the training and test sets (Meinshausen & Bühlmann, 2015; Koh et al., 2020). To address this issue, various paradigms (Zhou et al., 2021; Wortsman et al., 2022; Andreassen et al., 2021; Li & Zhang, 2021; Mao et al., 2023; Jiang et al., 2023; Goyal et al., 2023) have been proposed to fine-tune VL-PTMs, aiming to enhance their robustness against test-time distribution shifts. By adopting these fine-tuning techniques, the fine-tuned VL-PTMs can quickly adapt to downstream tasks even with only a few training examples (Fan et al., 2021; Nakamura & Harada, 2019; Gao et al., 2020). Furthermore, some methods have shown improved generalization performances within the same downstream tasks under OOD scenarios (Wortsman et al., 2022; Goyal et al., 2023; Zhu et al., 2023a). Nevertheless, most of these methods primarily focused on closed-set

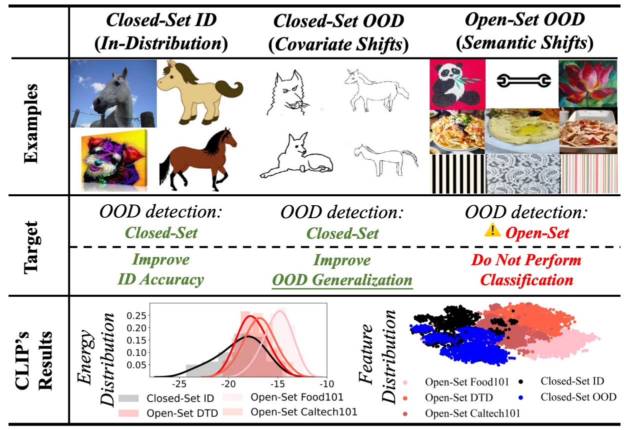

Figure 1. Illustration of typical data setting in real-world scenarios. For example, we may encounter various types of data in real-world applications: (i) closed-set ID data (e.g., dog), (ii) closed-set OOD data with covariate shifts (e.g., dog with changed image styles), and (iii) open-set OOD data with semantic shifts (e.g., panda). **The significant overlaps in energy distributions between closed-set ID and open-set OOD data pose a challenge for CLIP in detecting open-set OOD data. The notable discrepancy between the closed-set ID and closed-set OOD data also complicates achieving OOD generalization for closed-set OOD data.**

[1]Shanghai Jiao Tong University, Shanghai, China. [2]Shanghai Artificial Intelligence Laboratory, Shanghai, China.. Correspondence to: Nanyang Ye <ynylincoln@sjtu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

visual concepts, limiting the model to a pre-defined list of categories. Their discrimination ability for new categories unseen during training has not been thoroughly explored.

Unfortunately, as illustrated in Figure 1, real-world testing environments often involve known classes living in diverse environments, as well as new categories unseen during training (Wang et al., 2023a; Bai et al., 2023). It is crucial to distinguish these unknown categories from known ones, rather than blindly predicting them as known classes. The ability to detect unseen classes is particularly essential for ensuring system safety in high-risk applications, such as autonomous driving (Liu et al., 2021; Majee et al., 2021) and medical imaging (Castro et al., 2020).

Previous studies have focused on improving models' robustness when fine-tuning VL-PTMs or developing methods for unseen-class recognition independently. Consequently, existing approaches are highly specialized in one task, but not capable of handling both aspects simultaneously. This raises a critical but underexplored question:

When fine-tuning VL-PTMs to downstream tasks, how to improve models' generalization ability to closed-set OOD data, while effectively detecting open-set unseen classes during fine-tuning?

In this paper, we develop a novel fine-tuning paradigm to go beyond the limitations of previous studies that were unable to address both aspects simultaneously. Initially, leveraging the widely used energy-based function (Liu et al., 2020) for detecting unknown classes, we propose an energy distribution reshaping (EDR) loss. The proposed EDR loss aims to approach an optimal solution of minimizing energy scores on in-distribution (ID) data, which is implemented by minimizing the gradient magnitude of energy scores. This enables us to fine-tune VL-PTMs in the direction of distinguishing the energy distribution of known classes from other distributions.

Furthermore, by connecting the two challenges using Hessians, we show that the proposed EDR loss theoretically leads to domain-consistent Hessians, thereby helping to bound the generalization performance on closed-set OOD test data (Rame et al., 2022; Hemati et al., 2023). Building upon this finding, we have developed a novel Hessian-based OOD generalization bound, which is associated with model performance under worst-case OOD scenarios. Motivated by bound minimization, we introduce a unified fine-tuning framework named **CRoFT**. This framework is designed to achieve ***ro****bust* ***f****ine-****t****uning while enabling* ***c****oncurrent optimization for both aspects*. Through the use of different data settings to evaluate model performance, we have demonstrated that our CRoFT approach can obtain state-of-the-art results on both tasks, especially showcasing up to 20% improvements in detecting open-set unseen classes.

2. Problem Setting

Task definition Based on the *closed-set ID* samples, $\{\mathbf{x}^{(\mathbf{i})}, \mathbf{y}^{(\mathbf{i})}\}_{i=1}^{N}$, sampled from some source domain $\mathcal{S}$, our goal is to fine-tune a VL-PTM to obtain a robust predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$, which maps inputs $\mathbf{x} \in \mathcal{X} = \mathbb{R}^D$ to outputs $\mathbf{y} \in \mathcal{Y} = \mathbb{R}^K$ (where K is the class number and D is the dimension of $\mathbf{x}$). In open-set scenarios, different distribution shifts can occur. These scenarios involve *closed-set OOD* data that exhibit *covariate shifts* (i.e., changes in environments, while class labels remain the same as the ID data) and *open-set OOD* data with *semantic shifts* (i.e., test-time unseen classes). Therefore, our focus lies in enhancing the robustness of predictor f from two perspectives: 1) *OOD generalization*, which is related to testing the model on known classes from a new domain, $\mathcal{T}$, that exhibit covariate shifts; 2) *open-set OOD detection*, enabling the fine-tuned model to detect test-time unseen classes.

Motivated by the remarkable success of the large-scale pre-trained vision-language model CLIP (Radford et al., 2021) in learning general visual knowledge, we delve into exploring a CLIP-based framework for boosting both OOD generalization and open-set OOD detection across diverse data scenarios. It's important to note that our approach readily extends to other VL-PTMs, such as ALIGN (Li et al., 2021), BLIP-2 (Li et al., 2023), Grounding DINO (Liu et al., 2023), and MiniGPT-4 (Zhu et al., 2023b), by employing a contrastive loss technique to align image representations with the corresponding text representations (Li et al., 2022; Goel et al., 2022; Mu et al., 2022).

First look at CLIP's performance Based on the real-world data setting depicted in Figure 1, we initially assess CLIP's performance on the two challenging tasks. Details of the data setting are in Setup-II of Section 4. Utilizing the widely adopted energy score (Liu et al., 2020) for detecting unseen classes, we visualize the energy distribution of different types of data. As illustrated in Figure 1, there is a significant overlap between the energy distributions of closed-set and open-set samples. This overlap poses a challenge for the CLIP model to effectively distinguish between them, making it difficult to detect open-set OOD instances. Furthermore, we employ t-SNE (Van der Maaten & Hinton, 2008) to reduce CLIP's image features to a 2-dimensional space and provide the embedding visualization in Figure 1. It is shown that there is a noticeable discrepancy between the feature distributions of closed-set ID data and closed-set OOD data, which complicates the task of achieving OOD generalization for closed-set OOD datasets.

Therefore, addressing the problem of enabling VL-PTMs to handle various distribution shits, especially enhancing VL-PTMs' ability to detect open-set OOD data, is a key research direction that requires urgent attention. Before delving into our approach, we provide an explanation for the model assumption as described in Assumption 2.1.

Assumption 2.1. We use notation $\mathcal{D}$ to represent a distribution on input space $\mathcal{X}$. Given an empirical distribution (denoted as $\widehat{\mathcal{D}}_{\mathcal{S}} = \{\mathbf{x}^{(\mathbf{i})}\}_{i=1}^{N}$) sampled from the input space of source domain $\mathcal{S}$, we input images and closed-set class names into the CLIP model. We then obtain the zero-shot image features and text features, denoted as $\{\mathbf{z}_{\mathbf{I0}}^{(\mathbf{i})}\}_{i=1}^{N}$ and $\{\mathbf{z}_{\mathbf{T0}}^{(\mathbf{i})}\}_{i=1}^{K}$, respectively. In the fine-tuning framework, we represent the fine-tuned image features and text features as $\mathbf{z}_{\mathbf{I}}^{(\mathbf{i})} := \mathbf{z}_{\mathbf{I}}(\mathbf{x}^{(\mathbf{i})}; \theta)$ $(i = 1, \cdots, N)$ and $\mathbf{z}_{\mathbf{T}}^{(\mathbf{j})} := \mathbf{z}_{\mathbf{T}}^{(\mathbf{j})}(\theta)$ $(j = 1, \cdots, K)$, respectively. θ denotes the model parameter. Consider an bounded instance loss function ℓ such that $\mathcal{Y} \times \mathcal{Y} \to [0, c]$, and $\ell(\mathbf{y_1}, \mathbf{y_2}) = 0$ if and only if $\mathbf{y_1} = \mathbf{y_2}$ $(\mathbf{y_1} \in \mathcal{Y}, \mathbf{y_2} \in \mathcal{Y})$. The expected risk on domain $\mathcal{D}$ is represented as $\mathcal{E}_{\mathcal{D}}(\theta) := \mathbb{E}_{\mathcal{D}}(f(\mathbf{x}; \theta); \mathbf{y})$.

3. Methodology

3.1. Reshaping energy distribution for open-set OOD detection

Despite various fine-tuning strategies proposed to improve robustness in classifying closed-set data, their limitations in detecting the open-set unseen classes are largely overlooked. Motivated by the promising results of the CLIP model in learning general visual-semantic knowledge, we propose a unified fine-tuning framework to improve CLIP's OOD generalization ability while enabling the model to detect open-set unseen classes. Based on the widely used energy function (Liu et al., 2020) for OOD detection, we reshape the energy distribution of the fine-tuned model's output, which is expected to facilitate discrimination between closed-set samples and open-set samples.

The energy score (Liu et al., 2020), as commonly used for OOD detection, is trained to assign lower energy values to more plausible or confident configurations. In the regime of fine-tuning CLIP, the energy score is calculated as:

$$E_{\theta}(\mathbf{x}) = -\mathbb{E}_{\mathbf{x} \in \mathcal{X}} \log \sum_{i=1}^{K} \exp \left\langle \mathbf{z_I}(\mathbf{x}; \theta), \mathbf{z}_{\mathbf{T}}^{(\mathbf{i})}(\theta) \right\rangle \quad (1)$$

Minimizing energy scores on training data has been demonstrated as an effective approach in previous studies (Du & Mordatch, 2019; Katz-Samuels et al., 2022). In this paper, we focus on the optimization objective of $\min E_{\theta}(\mathbf{x})$. With this training criterion, we define the empirical loss $\mathcal{L}(\mathcal{D}_{\mathcal{S}}, \theta) = -\frac{1}{N} \sum_{i=1}^{N} E_{\theta}(\mathbf{x}^{(\mathbf{i})})$ with the training datasets $\mathcal{D}_{\mathcal{S}} = \{\mathbf{x}^{(\mathbf{i})}\}_{i=1}^{N}$. By utilizing the gradient decent (Amari, 1993) method, model parameter θ can be updated with the following gradient:

$$\frac{\partial \mathcal{L}(\mathcal{D}, \theta)}{\partial \theta} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial E_{\theta}(\mathbf{x}^{(\mathbf{i})})}{\partial \theta} \quad (2)$$

For large-size VL-PTMs, calculating the gradient as defined in Equation 2 along with the whole model is computationally expensive. To address this issue, we adopt lightweight fine-tuning for training efficiency. Specifically, we inject one-layer linear projections into CLIP's image encoder and text encoder, respectively, while keeping parameters (θ_0) in pre-trained encoders frozen. By doing this, we only need to calculate the gradient of loss $\mathcal{L}(\mathcal{D}_{\mathcal{S}}, \theta)$ with respect to parameters in linear projections (θ_l). Finally, we can approach the optimum in Equation 2 as illustrated in Proposition 3.1.

Proposition 3.1. ***[Energy distribution reshaping (EDR) loss]*** *Given the training data $\widehat{\mathcal{D}}_{\mathcal{S}}$, in our fine-tuning framework, we calculate the training data's energy scores based on Equation 1. To approach the solution of $\min E_{\theta}(\mathbf{x})$, i.e., $\nabla_{\theta} E_{\theta}(\mathbf{x}) \to \mathbf{0}$, we propose to minimize the squared value of Equation 2 (i.e., magnitude of the gradient vector), which is formulated as optimizing the following loss:*

$$\mathcal{L}_e = \frac{1}{N} \sum_{i=1}^{N} \left[\nabla_{\theta_l} \left(\log \sum_{j=1}^{K} \exp \left\langle \mathbf{z_I}(\mathbf{x}^{(\mathbf{i})}; \theta), \mathbf{z}_{\mathbf{T}}^{(\mathbf{j})}(\theta) \right\rangle \right) \right]^2 \quad (3)$$

where $\theta = \{\theta_0, \theta_l\}$, θ_0 is the frozen parameter in CLIP's pre-trained encoders, and θ_l is the parameter in linear projections that need to be optimized.

Unlike previous studies (Liu et al., 2020; Katz-Samuels et al., 2022; Tonin et al., 2021) that directly reduce energy scores of ID training data, **we find the proposed EDR loss not only enhances the open-set OOD detection abilities but is also secretly helping to improve OOD generalization, as demonstrated in Theorem 3.6.**

3.2. Minimizing the EDR-inspired OOD generalization bound for better concurrent optimization

In this section, we start by developing an EDR-inspired OOD generalization bound under the worst-case scenarios, where the theoretical link between the EDR loss and the generalization bound is detailed in Section 3.3. The novel bound indicates that the OOD generalization performance on target domain is related to the performance gap between the minimum risk on source domain and the emperical risk on worst-case covariate-shifted data. Motivated by reducing this performance gap, we introduce an adversarial-learning-based approach. Specifically, we iteratively generate the most challenging covariate-shifted image features, simulating the worst-case scenarios, and enhances the model's robustness for these generated OOD image features. Model assumptions in this section are presented in Assumption 3.2.

Assumption 3.2. Given empirical distributions generated with m i.i.d. samples from the source domain and target domain, denoted as $\widehat{\mathcal{D}}_{\mathcal{S}}$ and $\widehat{\mathcal{D}}_{\mathcal{T}}$, respectively, we denote the empirical risk on $\widehat{\mathcal{D}}_{\mathcal{S}}$ ($\widehat{\mathcal{D}}_{\mathcal{T}}$) as $\widehat{\mathcal{E}}_{\mathcal{S}}(\theta)$ ($\widehat{\mathcal{E}}_{\mathcal{T}}(\theta)$). Let $\widehat{\mathcal{D}}_{\mathcal{S}}^{c}$ represent the empirical distribution of the worst-case covariate-shifted OOD data, while sharing the same seman-

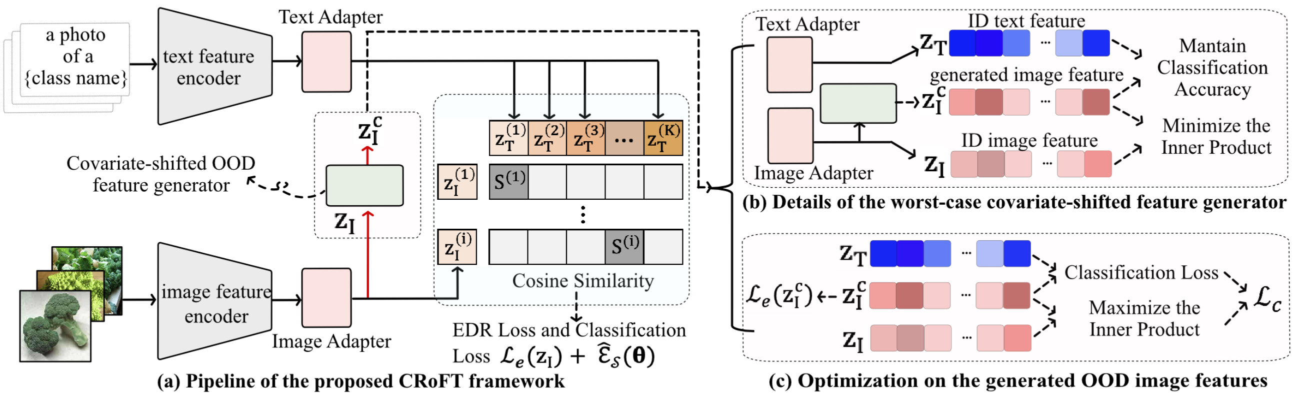

Figure 2. **Overview of our CRoFT framework**. Our theoretical analysis leads to the design of a new fine-tuning framework. As shown in Figure (a), we inject adapters, i.e., one-layer linear projections after the CLIP's pre-trained encoders. Based on the adapted image feature $\mathbf{z_I}$ and adapted text feature $\mathbf{z_T}$, we generate the most challenging covariate-shifted OOD image features $\mathbf{z_I^c}$, simulating the worst-case scenarios. The corresponding generation process, depicted in Figure (b), follows the criterion defined in Equation 5, which preserves semantic information to maintain classification accuracy but differs from the ID image feature $\mathbf{z_I}$. Finally, as shown in Figure (c), we optimize on the generated $\mathbf{z_I^c}$ using the proposed loss $\mathcal{L}_c$. Meanwhile, we minimize the classification loss (cross-entropy) on the ID image features, denoted as $\widehat{\mathcal{E}}_S(\theta)$, while reshaping the energy distribution for $\mathbf{z_I}$ and $\mathbf{z_I^c}$ through the EDR loss (i.e., $\mathcal{L}(\mathbf{z_I})$ and $\mathcal{L}(\mathbf{z_I^c})$).

tic information with $\widehat{\mathcal{D}}_\mathcal{S}$. We denote the empirical risk on the worst-case covariate-shifted OOD data as $\widehat{\mathcal{E}}^c_\mathcal{S}(\theta)$. Based on the distribution distance (Zhao et al., 2018; Cha et al., 2021): $\text{Div}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) = 2\sup_A |\Pr_{\mathcal{D}_\mathcal{S}}(A) - \Pr_{\mathcal{D}_\mathcal{T}}(A)|$, we assume that there exists $\varepsilon_c > 0$, such that $\text{Div}(\widehat{\mathcal{D}}^c_\mathcal{S}, \widehat{\mathcal{D}}_\mathcal{T}) \leq \text{Div}(\widehat{\mathcal{D}}_\mathcal{S}, \widehat{\mathcal{D}}_\mathcal{T}) + \varepsilon_c$. For each image feature $\mathbf{z_I}$ from the source domain, it is assumed that the corresponding image features $\tilde{\mathbf{z}}_\mathbf{I}$ from the target domain, which share the same label with $\mathbf{z_I}$, satisfy the condition $||\mathbf{z_I} - \tilde{\mathbf{z}}_\mathbf{I}||_2 \leq \varepsilon$, where ε is a small positive value.

Theorem 3.3. ***[Hessian-based OOD generalization bound]*** *Let v be the VC dimension of the parameter space Θ. Let $\mathbf{H}_\mathcal{S}(\theta_l)$ be the Hessian matrix of expected risk $\mathcal{E}_\mathcal{S}(\theta)$ over parameter θ_l. Utilizing Taylor expansion, we represent $\min_{\theta'} \mathcal{E}_\mathcal{T}(\theta')$ as the Hessian-based expression: $\min_{\theta'} \mathcal{E}_\mathcal{T}(\theta') \leq \max\{\lambda_\mathcal{T}, 2\lambda_\mathcal{S} - \lambda_\mathcal{T}\} + \frac{1}{2}\left|\theta^\top(\mathbf{H}_\mathcal{S}(\theta^*))\theta\right| + \varepsilon$. where θ^* is the optimal solution on the mixture of source domain and target domain, $\lambda_\mathcal{S} = \mathcal{E}_\mathcal{S}(\theta^*)$, and $\lambda_\mathcal{T} = \mathcal{E}_\mathcal{T}(\theta^*)$. This inequality is a direct result of Theorem 3.6, which we will detail later. Then for $0 \leq \delta \leq 1$, with probability at least $1 - \delta$, we have:*

$$\begin{aligned}\mathcal{E}_\mathcal{T}(\theta) \leq \widehat{\mathcal{E}}^c_\mathcal{S}(\theta) - \min_{\theta'} \widehat{\mathcal{E}}_\mathcal{S}(\theta') + \frac{1}{2}\left|\theta^\top \mathbf{H}_\mathcal{S}(\theta^*)\theta\right| \\ + \text{Div}(\widehat{\mathcal{D}}_\mathcal{S}, \widehat{\mathcal{D}}_\mathcal{T}) + \max\{\lambda_\mathcal{T}, 2\lambda_\mathcal{S} - \lambda_\mathcal{T}\} + \lambda \\ + \varepsilon + \varepsilon_c + O\left(\sqrt{\frac{v\log(m/v) + \log(1/\delta)}{m}}\right)\end{aligned} \tag{4}$$

where $\lambda = \mathcal{E}_\mathcal{S}(\theta^*) + \mathcal{E}_\mathcal{T}(\theta^*)$, ε and ε_c are small positive values as defined in Assumption 3.2.

This theorem indicates that the generalization performance on the target domain $\mathcal{T}$ is related to the performance gap between minimum risk on source domain, i.e, $\min_{\theta'} \widehat{\mathcal{E}}_\mathcal{S}(\theta')$, and empirical risk on the worst-case covariate-shifted OOD samples, i.e, $\widehat{\mathcal{E}}^c_\mathcal{S}(\theta)$. Motivated by bound minimization, we thus aim to optimize the model performances on the worst-case covariate-shifted OOD scenarios, reducing performance degradation on the target domain. To achieve this, we propose to generate the worst-case covariate-shifted image feature $\mathbf{z_I^c}$, which preserves semantic information for accurate classification but differs from the ID image feature $\mathbf{z_I}$. Formally, we formulate this procedure as below:

Proposition 3.4. ***[Optimization on the worst-case covariate-shifted OOD features]*** *Utilizing a covariate-shifted OOD feature generator $g(\cdot)$, which conducts a one-layer linear projection on ID image features, we generate the worst-case covariate-shifted image features as:*

$$\text{argmin}_g \frac{\lambda_1}{N}\sum_{i=1}^{N}\left\langle g\left(\mathbf{z_I^{(i)}}\right), \mathbf{z_I^{(i)}}\right\rangle + \widehat{\mathcal{E}}^c_\mathcal{S}(\theta) \tag{5}$$

where $g(\mathbf{z_I^{(i)}})$ is the generated worst-case OOD image feature, and λ_1 is a hyperparameter to balance covariate shifts and classification accuracy. Then we optimize performance on the generated covariate-shifted features by minimizing:

$$\mathcal{L}_c = -\frac{\lambda_1}{N}\sum_{i=1}^{N}\left\langle g\left(\mathbf{z_I^{(i)}}\right), \mathbf{z_I^{(i)}}\right\rangle + \widehat{\mathcal{E}}^c_\mathcal{S}(\theta) \tag{6}$$

3.3. Theoretical connection between the EDR loss and OOD generalization

In previous studies (Rame et al., 2022; Hemati et al., 2023), it has been discovered that matching simultaneously domain-level risks and Hessians can improve OOD generalization. Different from previous works, we do not directly regularize the model to learn domain-consistent Hessians. Instead, we theoretically demonstrate that the proposed EDR loss for open-set OOD detection can secretly lead to domain-consistent Hessinas of classification loss, thereby enabling concurrent optimization for both OOD generalization and OOD detection. Before delving into the details, we introduce an important property to show the relationship between the EDR loss and the Hessians, as illustrated in Lemma 3.5.

Lemma 3.5. ***[The proposed EDR loss leads to domain-consistent Hessians]*** *For each input $\mathbf{x}^{(\mathbf{i})}$, we represent the inner product between its image feature and the text feature of its ground-truth class name as $S^{(i)} = \left\langle \mathbf{z_I}(\mathbf{x}^{(\mathbf{i})}; \theta_l), \mathbf{z}_\mathbf{T}^{(\mathbf{i})}(\theta) \right\rangle$. Let $\widehat{\mathbf{H}}_\mathcal{D}(\theta_l)$ be the Hessian matrix of empirical risk $\widehat{\mathcal{E}}_\mathcal{D}(\theta)$ over parameter in adapters (θ_l). The local optimal solution of the EDR loss (as defined in Equation 3) leads to the following property:*

$$\begin{aligned} \widehat{\mathbf{H}}_\mathcal{D}(\theta_l) &= -\frac{1}{N}\sum_{i=1}^{N} \nabla^2_{\theta_l} S^{(i)} \\ &= -\frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \mathbf{0} & \mathbf{z}_\mathbf{I0}^{(\mathbf{i})} \mathbf{z}_\mathbf{T0}^{(\mathbf{i})\top} \\ \mathbf{z}_\mathbf{I0}^{(\mathbf{i})} \mathbf{z}_\mathbf{T0}^{(\mathbf{i})\top} & \mathbf{0} \end{bmatrix} \end{aligned} \tag{7}$$

Theorem 3.6. ***[The EDR loss bound OOD generalization]*** *By Taylor expansion, the OOD generalization gap between source domain ($\mathcal{S}$) and target domain ($\mathcal{T}$) is upper bounded by the following equation:*

$$\begin{aligned} &\max_{\{\theta:|\widehat{\mathcal{E}}_\mathcal{S}(\theta)-\widehat{\mathcal{E}}_\mathcal{S}(\theta^*)|\le\epsilon\}} |\widehat{\mathcal{E}}_\mathcal{T}(\theta) - \widehat{\mathcal{E}}_\mathcal{S}(\theta^*)| \\ &\approx \max_{\{\theta:\frac{1}{2}|\theta^\top\widehat{\mathbf{H}}_\mathcal{S}\theta|\le\epsilon\}} \left|\widehat{\mathcal{E}}_\mathcal{T}(\theta^*) + \frac{1}{2}\theta^\top\widehat{\mathbf{H}}_\mathcal{T}(\theta^*)\theta - \widehat{\mathcal{E}}_\mathcal{S}(\theta^*)\right| \\ &\le |\widehat{\mathcal{E}}_\mathcal{T}(\theta^*) - \widehat{\mathcal{E}}_\mathcal{S}(\theta^*)| + \max\frac{1}{2}|\theta^\top\widehat{\mathbf{H}}_\mathcal{S}(\theta^*)\theta| + \varepsilon \end{aligned} \tag{8}$$

where θ^* is a local minimum across all domains, i.e., $\nabla_\theta\widehat{\mathcal{E}}_\mathcal{D}(\theta^*) = \mathbf{0}$, and ε is defined in Assumption 3.2.

Therefore, by connecting the EDR loss with the Hessians of empirical classification loss, we theoretically discover that the EDR loss can lead to a bound of the performance gap between closed-set ID data and closed-set OOD data. This implies that optimizing for open-set OOD detection with EDR loss also involves optimizing for OOD generalization.

3.4. Algorithm of the proposed CRoFT

Our theoretical analysis thus leads to the design of a new fine-tuning framework with concurrent optimization for both tasks. As illustrated in Figure 2, we prioritize computational efficiency by adopting lightweight fine-tuning techniques. We employ an image adapter and a text adapter after CLIP's image encoder and text encoder, respectively, while keeping parameters in pre-trained encoders frozen. Both the image adapter and text adapter are implemented as one-layer linear projections. After that, we proceed with the adversarial learning procedure as discussed in Proposition 3.4, to generate covariate-shifted OOD image features. The generation process is illustrated in Figure 2 (b). Finally, we minimize the classification loss on the ID image features, denoted as $\widehat{\mathcal{E}}_\mathcal{S}(\theta)$, while incorporating two regularization terms. 1) We optimize on the generated OOD image features $\mathbf{z}_\mathbf{I}^\mathbf{c}$ using the proposed $\mathcal{L}_\mathrm{c}$, as defined in Equation 6. 2) Meanwhile, we employ the EDR loss to reshape the energy distribution for both ID image features and generated OOD image features, denoted as $\mathcal{L}_\mathrm{e}(\mathbf{z_I})$ and $\mathcal{L}_\mathrm{e}(\mathbf{z}_\mathbf{I}^\mathbf{c})$, respectively. Therefore, the final optimization objective is expressed as:

$$\mathcal{L}_\mathrm{CRoFT} = \widehat{\mathcal{E}}_\mathcal{S}(\theta) + \lambda_1\mathcal{L}_\mathrm{c} + \lambda_2\left(\mathcal{L}_\mathrm{e}(\mathbf{z_I}) + \mathcal{L}_\mathrm{e}(\mathbf{z}_\mathbf{I}^\mathbf{c})\right) \tag{9}$$

where λ_1 and λ_2 are hyperparameters that can be chosen based on the validation procedure. For details about the complete algorithm, please refer to Algorithm 1 in Appendix.

4. Experiments

In this section, we compare our method with competitive CLIP-based lightweight fine-tuning methods, such as CoOp (Zhou et al., 2021), CoCoOp (Zhou et al., 2022), CLIP-Adapter (Gao et al., 2023), Tip-Adapter-F (Zhang et al., 2021a), DPLCLIP (Zhang et al., 2021b) and Bayes-CAL (Zhu et al., 2023c). Additionally, we perform extensive ablation studies and visualization analyses to validate our theoretical findings. To better evaluate models' open-set OOD detection capabilities in real-world scenarios, we introduce two data settings that encompass 600 and 200+ unseen classes, respectively. Details are illustrated as follows:

Data setups 1) *Setup-I: open-set discrimination on the large-scale ImageNet dataset.* In the literature, ImageNet and its variants are commonly used for investigating fine-tuning robustness. In line with existing research, we construct datasets of Setup-I using these datasets. Specifically, we split ImageNet-1K (Deng et al., 2009) into open and closed sets w.r.t class labels. We randomly define 40% classes of ImageNet as the closed-set for training, and the remaining 60% as the open-set for testing. The samples from ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), ImageNet-Sketch (Wang et al., 2019), and ImageNet-V2 (Recht et al., 2019) with the same class labels with the closed-set are as closed-set OOD data.

2) *Setup-II: open-set discrimination on cross-dataset images.* Using cross-dataset examples as the open-set is another established protocol (Shafaei et al., 2018; Kong & Ramanan, 2021), which is introduced to reduce dataset-level bias. This protocol assesses the generalization capabilities of open-set OOD detection methods to diverse open-testing examples. In our experiment, we leverage popular datasets like PACS (Li et al., 2017) or VLCS (Li et al., 2017) for domain generalization studies as the closed-set data. We evaluate the models' ability to differentiate between closed-set OOD and cross-dataset images by utilizing different styles of datasets like Caltech101 (Bansal et al., 2021), DTD (Sharan et al.,

Table 1. **Setup-I:** Comparison with competitive fine-tuning methods based on CLIP ViT-B/16. We report the average percentage results across 3 runs, with standard errors presented in parentheses. We can observe that CRoFT surpasses the state-of-the-art methods in distinguishing between closed-set OOD and open-set OOD samples, achieving a significant reduction of up to **25.3%** on FPR95 when compared to CLIP. Meanwhile, CRoFT demonstrates comparable or even better generalization results on the closed-set OOD test sets.

DATA / Method	1 -shot				16-shot				32-shot			
	ID ACC	OOD ACC	AUROC	FPR95	ID ACC	OOD ACC	AUROC	FPR95	ID ACC	OOD ACC	AUROC	FPR95
CLIP	78.2	58.4	75.6	77.3	78.2	58.4	75.6	77.3	78.2	58.4	75.6	77.3
COOP	79.3	60.5	76.7	78.1	82.3	61.3	77.8	73.4	83.0	61.5	79.1	70.6
COCOOP	**80.0**	61.7	77.3	77.5	81.6	62.8	78.1	73.6	81.7	61.8	72.3	79.4
CLIP-Adapter	78.2	58.2	75.9	79.3	78.7	59.0	76.0	79.7	78.6	59.1	76.1	79.5
Tip-Adapter-F	79.5	**61.9**	76.3	79.6	82.3	62.5	71.7	82.9	**83.3**	62.8	68.1	85.9
DPLCLIP	79.1	60.4	77.9	77.3	82.1	61.7	78.6	72.4	83.0	61.7	77.8	72.6
Bayes-CAL	79.0	60.5	75.5	77.2	82.1	61.3	78.3	71.3	82.9	61.5	78.3	70.9
CRoFT (Ours)	79.6 (0.4)	**61.9** (0.3)	**80.5** (1.2)	**69.3** (3.5)	**82.5** (0.2)	**62.9** (0.6)	**87.2** (0.9)	**52.0** (3.3)	83.1 (0.5)	**63.1** (0.5)	**86.5** (0.3)	**55.2** (0.8)

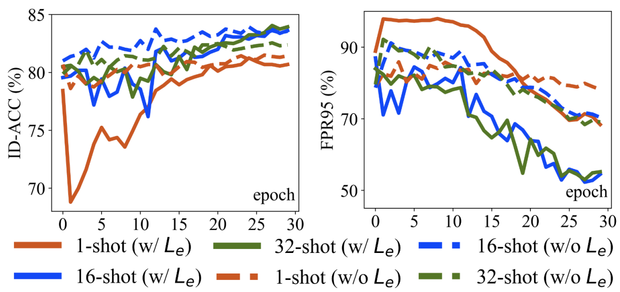

Figure 3. Ablations on the proposed EDR loss. With the proposed EDR loss $\mathcal{L}_e$, our method successfully fine-tunes CLIP's features in the direction of better-discriminating open-set and closed-set, without sacrificing the test accuracy.

2014), and Food101 (Bossard et al., 2014) as open-set examples. All overlapping classes are removed from the three open-set datasets. In the evaluation of OOD generalization performance, we utilize the leave-one-domain-out validation protocol (Gulrajani & Lopez-Paz, 2020; Cha et al., 2021) that uses three domains as closed-set ID data and the remaining one as closed-set OOD data.

4.1. Experiment results of Setup-I

Experiment details In our experiments on Setup-I, we conduct experiments based on the CLIP ViT-B/16 model. For the prompt learning methods, CoOp, CoCoOp, DPLCLIP, and Bayes-CAL, we use random initialization for context vectors and set the number of context tokens to 16. Regarding other hyperparameters, we set the class token position (CTP) as "end" and set the class-specific context (CSC) as "False". This configuration has yielded the best average performance according to CoOp's paper. We adhere to the recommended hyperparameter settings outlined in the original paper of CLIP-Adapter. In the case of Tip-Adapter-F, we perform hyperparameter searches following its original paper. Without otherwise specified, methods are trained using the SGD optimizer with a learning rate of 0.002 and batch size of 32 for fair comparisons. In our CRoFT method, we search for λ_1 in the range of $[1, 5, 10, 15, 20]$ and λ_2 in the range of $[10, 20, 30, 40, 50]$. As lightweight fine-tuning methods facilitate fast convergence, we set the maximum training epoch to 30. We compare our methods with these

Table 2. Comparison with the vanilla technique of energy minimization (named EnergyMin). We report the classification accuracy on closed-set OOD data and the FPR95 value in open-set OOD detection (OOD ACC / FPR95).

Method	1-shot	16-shot	32-shot
CLIP + EnergyMin	**80.02** / 80.50	81.91 / 69.27	81.87 / 69.12
CRoFT	79.60 / **69.31**	**82.50 / 52.01**	**83.10 / 55.19**

Table 3. Abaltion Study results on the proposed regularization loss $\mathcal{L}_c$ under 32-shot scenarios.

DATA	ID	INV2	IN-S	IN-A	IN-R	Avg OOD
w / o $\mathcal{L}_c$	82.21	75.83	42.98	69.57	60.57	62.24
w / o $\mathcal{L}_e$	81.99	76.33	43.55	70.18	61.06	62.78
w / $\mathcal{L}_c$	**83.10**	**76.64**	**44.23**	**70.30**	**61.15**	**63.07**

competitive fine-tuning methods under 1-shot, 16-shot and 32-shot scenarios. For experiments on each method, we repeat 3 times with different random splits to eliminate the effects of randomness. Finally, we report the average classification accuracy on closed-set test sets, as well as the average FPR95 and AUROC results for distinguishing between open-set OOD data and closed-set data by inferring energy score (Liu et al., 2020). For more experiment details, please refer to Appendix D.

Experiment results We present the main results in Table 1, where CRoFT establishes the overall best performance in both OOD generalization and open-set OOD detection. Notably, the proposed CRoFT method outperforms the zero-shot CLIP model by achieving an impressive improvement of up to **25.3%** on FPR95. In contrast, its competitors struggle with the open-set OOD detection task, even resulting in higher FPR95 values when compared to the CLIP model. Moreover, our method demonstrates comparable or even superior generalization results on the closed-set OOD test sets. Specifically, it achieves the best average test accuracy on closed-test OOD data compared to recent methods.

Effect of $\mathcal{L}_e$ We ablate on the proposed EDR loss $\mathcal{L}_e$ and plot the corresponding ID test accuracy and open-set OOD detection results along the fine-tuning process. As depicted in Figure 3, the FPR95 value gradually decreases with the increase of training epoch, while maintaining the ID test accuracy. In Table 2, we compare our method with the

Table 4. **Setup-II:** OOD generalization results measured by classification accuracy on closed-set OOD data (OOD ACC) and open-set OOD detection measured by AUROC and FPR95 over the mixture of closed-set OOD and open-set OOD test sets. For CLIP-Adapter and our method, the open-vs-closed discrimination results are obtained by inferring KNN distances ($k = 1$) or energy scores on the adapted image features. Due to space limitations, we only report the best results for each method. For the CLIP model, we also report its corresponding KNN-based results in CLIP (KNN). Since other methods like CoOp, CoCoOp, and Tip-Adapter-F do not fine-tune the image features, their results when inferring KNN distances are the same as CLIP (KNN).

DATA	V-Net	PACS OOD ACC	PACS vs. Open-Set (AUROC / FPR95)			VLCS OOD ACC	VLCS vs. Open-Set (AUROC / FPR95)			**AVG FPR**
Algorithm			DTD	Food101	Caltech101		DTD	Food101	Caltech101	
CLIP	RN50	90.8 (0.0)	76.6 / 80.2	94.7 / 29.2	86.8 / 52.3	75.1 (0.0)	40.6 / 94.5	82.8 / 52.8	61.5 / 85.5	65.8
	ViT16	96.1 (0.0)	82.4 / 67.6	95.9 / 26.0	86.7 / **52.2**	76.3 (0.0)	55.3 / 88.8	85.8 / 48.3	53.3 / 86.3	61.5
CLIP (KNN)	RN50	90.8 (0.0)	85.3 / 58.5	99.1 / 4.3	79.2 / 65.3	75.1 (0.0)	82.4 / 67.4	95.2 / 27.0	74.1 / 74.4	49.9
	ViT16	96.1 (0.0)	87.7 / 58.9	98.8 / 5.8	82.5 / 70.0	76.3 (0.0)	82.4 / 71.6	95.1 / 32.6	78.7 / 74.1	52.2
CoOp	RN50	91.5 (0.6)	90.1 / 43.6	97.0 / 15.5	83.1 / 56.8	72.8 (5.4)	40.4 / 96.9	54.5 / 87.0	46.3 / 93.7	65.6
	ViT16	96.3 (0.7)	85.7 / 62.7	96.2 / 22.6	80.0 / 70.9	78.3 (1.7)	47.3 / 89.2	76.3 / 62.9	46.8 / 90.0	66.4
CoCoOp	RN50	91.8 (0.6)	90.7 / 42.8	98.3 / 9.0	88.4 / 48.5	76.3 (1.0)	42.4 / 95.9	80.2 / 65.0	46.8 / 92.5	58.9
	ViT16	96.8 (0.5)	92.1 / 44.9	97.8 / 15.0	87.8 / 57.5	78.9 (0.8)	49.2 / 93.3	78.4 / 60.8	49.2 / 90.6	60.4
CLIP-Adapter	RN50	90.9 (0.1)	85.3 / 58.2	99.1 / 4.2	79.2 / 65.0	75.2 (0.1)	82.4 / 67.7	**95.3 / 26.5**	74.1 / 74.7	49.4
	ViT16	96.1 (0.0)	87.7 / 58.6	98.8 / **5.6**	82.6 / 69.6	77.3 (0.7)	82.6 / 71.2	95.1 / 32.4	78.8 / 74.2	51.9
Tip-Adapter-F	RN50	92.2 (0.4)	83.4 / 81.8	97.2 / 15.1	84.7 / 72.9	76.5 (0.2)	59.3 / 95.4	91.5 / 38.6	64.1 / 84.0	65.7
	ViT16	96.9 (0.4)	87.3 / 69.3	97.8 / 11.2	84.8 / 74.9	79.9 (0.2)	66.7 / 94.4	92.7 / 39.6	63.0 / 95.1	64.1
DPLCLIP	RN50	89.6 (1.2)	**95.4 / 25.0**	99.0 / 5.8	**97.4 / 14.1**	76.1 (0.6)	56.3 / 93.8	79.9 / 67.6	51.5 / 96.0	50.4
	ViT16	95.6 (0.2)	93.2 / 39.0	98.0 / 13.0	**94.4 / 33.5**	76.5 (1.3)	52.9 / 98.7	69.7 / 88.8	52.4 / 97.3	61.8
Bayes-CAL	RN50	91.8 (0.3)	89.7 / 47.8	97.4 / 14.8	83.5 / 70.5	78.1 (1.5)	50.9 / 94.3	69.5 / 76.0	53.4 / 97.7	66.9
	ViT16	96.6 (0.5)	79.2 / 62.6	87.0 / 38.6	71.8 / 60.3	79.6 (0.9)	44.2 / 93.6	68.1 / 72.9	34.0 / 90.7	69.8
CRoFT (Ours)	RN50	**92.5 (0.3)**	94.7 / 26.8	**99.8 / 0.8**	88.7 / 42.0	**79.5 (0.7)**	**86.6 / 58.8**	93.9 / 35.3	**77.7 / 68.1**	**38.6**
	ViT16	**97.3 (0.1)**	**94.7 / 33.2**	**99.1** / 6.1	88.9 / 52.2	**80.2 (1.0)**	**86.7 / 60.6**	**96.0 / 25.9**	**83.7 / 62.2**	**40.0**

vanilla energy reduction technique (named EnergyMin). It is shown that this method achieved lower FPR95 values in the 16-shot and 32-shot scenarios compared to CLIP, but it still lagged behind our method by about 15%. More importantly, our method demonstrated better OOD generalization results, thus validating our theoretical finding in Theorem 3.6 that the EDR loss has a positive impact on OOD generalization.

Effect of $\mathcal{L}_c$ By removing the optimization process on the generated OOD image features while keeping other components unchanged, we assess the impact of $\mathcal{L}_c$ on OOD generalization. The results of this ablation study are in Table 3. With the employment of the $\mathcal{L}_c$, our method has obtained improved OOD test accuracy, especially in the worst-case scenario (i.e., ImageNet-Sketch), which highlights the efficacy of the proposed adversarial learning procedure in learning more robust adapters.

4.2. Experiment Results of Setup-II

Experiment details In Setup-II, we keep the same hyperparameter setting as in Setup-I without further explanation. For our CRoFT method, we set λ_1 as 15 and search for λ_2 in the range of $[1, 10, 100, 1000]$. Following previous works (Zhou et al., 2022; Zhu et al., 2023c), we perform evaluations under 16-shot scenarios based on CLIP RN50 and ViT-B/16. All experiments are repeated 3 times with different random seeds. Finally, we report the average classification accuracy on closed-set OOD data, as well as the average FPR95 and AUROC for distinguishing between open-set OOD and closed-set OOD by inferring energy score (Liu et al., 2020) and KNN distances (Sun et al., 2022).

Experiment results We present the main results in Table 4, where CRoFT establishes the overall best performance in both OOD generalization and open-set OOD detection. The proposed CRoFT method outperforms the second-best FPR95 results by about **11%**, showcasing its effectiveness in detecting cross-dataset unseen classes. Consistent with the results of Setup-I, its competitors achieve even higher FPR95 results in comparison with the CLIP model. It is worth noting that all methods yield better recognition results on the Food101 dataset but significantly worse results on the Caltech101 dataset. This difference can be attributed to the fact that Caltech101 is distributed closer to the in-distribution datasets. Despite the disparity degree between closed-set and open-set data, our method consistently improves performance across various pre-trained image encoders. Moreover, our method demonstrates superior generalization results on the closed-set OOD test sets, boosting test accuracy on closed-test OOD data by an impressive **4.3%** compared to the vanilla CLIP-Adapter method.

Visualization of adapted features To comprehensively illustrate that our CRoFT method can learn more distinct closed-set image features from open-set image features, we compare it with the vanilla adapter-tuning method (CLIP-Adapter) in open-set OOD detection by inferring KNN distances. As shown in Figure 4 (a), our method consistently outperforms the CLIP-Adapter across various settings of k, showcasing the effectiveness and robustness of our method in learning discriminative features for open-vs-closed discrimination. Moreover, we visualize the distribution of image features in Figure 4 (b)-(c). It is shown that im-

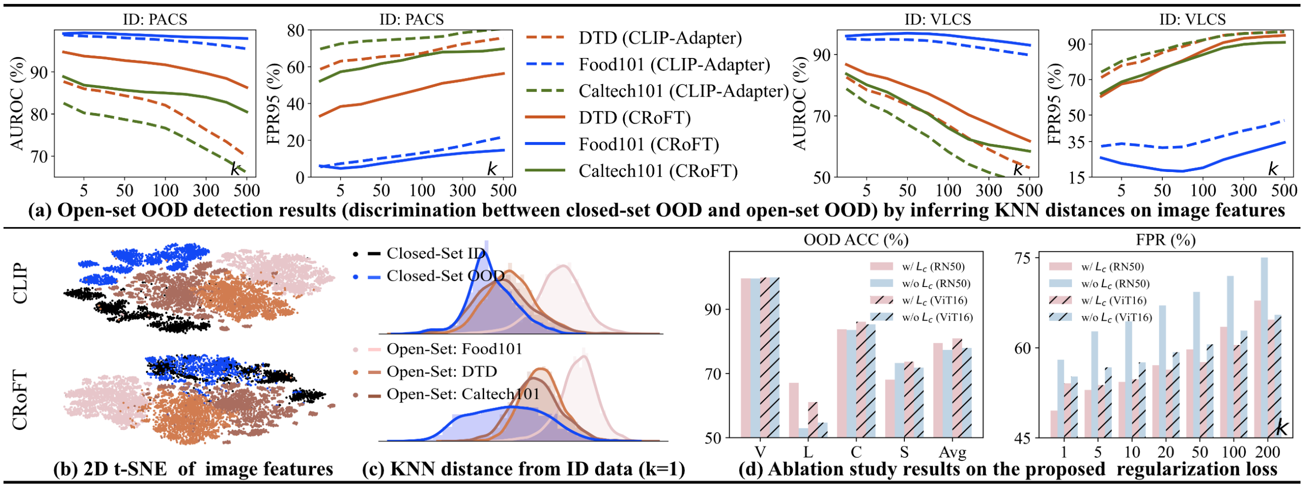

Figure 4. Ablation Study results of Setup-II. (a): Comparison with CLIP-Adapter in open-set OOD detection by referring KNN distances on the adapted image features. (b): t-SNE visualization of image features. (c): Average KNN distance between OOD features and ID features. (We use VLCS as the closed-set data in (b) and (c).) (d): Experiment results on VLCS for CRoFT with $\mathcal{L}_c$ vs. without $\mathcal{L}_c$.

Table 5. Ablation study results on $\mathcal{L}_c$ and $\mathcal{L}_e$. The best overall performances of our CRoFT validate the result of Theorem 3.6.

DATA	Method	w / o $\mathcal{L}_c$		w / o $\mathcal{L}_e$		Ours	
	V-Net	RN50	ViT16	RN50	ViT16	RN50	ViT16
PACS	OOD ACC	92.29	96.67	92.02	96.71	**92.45**	**97.26**
	AUROC	88.28	90.43	91.49	93.63	**94.41**	**94.25**
VLCS	OOD ACC	77.34	77.95	77.70	77.56	**79.47**	**80.21**
	AUROC	84.89	86.50	85.34	88.31	**86.06**	**88.79**

age features from the zero-shot CLIP also contain domain-related information, which hinders the discrimination between closed-set OOD and open-set OOD data. In contrast, our method can learn open-vs-closed discriminated image features, achieving much smaller distances between the ID data and covariate-shifted OOD data.

Ablate on $\mathcal{L}_c$ and $\mathcal{L}_e$ We conduct ablation studies on the proposed $\mathcal{L}_c$. As depicted in Figure 4 (d), we observe that the OOD generalization results, especially for challenging examples (i.e., domain "L"), are significantly enhanced by incorporating $\mathcal{L}_c$. Moreover, the open-set OOD detection results are also promoted, which is caused by the improved quality of the adapted image features. Furthermore, we conduct additional ablation studies to validate the positive impact of the EDR loss on OOD generalization. We remove the EDR loss for all image features and evaluate the OOD generalization results as reported in Table 5. It is shown that incorporating the EDR loss $\mathcal{L}_e$ leads to enhanced OOD generalization, highlighting its concurrent optimization effect on both OOD generalization and open-set OOD detection.

5. Related Works

CLIP-based fine-tuning methods For training efficiency, there have been many lightweight CLIP-based fine-tuning methods to enhance generalization performance via prompt tuning (Singha et al., 2023; Huang et al., 2022; Khattak et al., 2023; Wang et al., 2023b; Wasim et al., 2023; Zhu et al., 2023c) or adapter tuning (Gondal et al., 2024; Zhang et al., 2023; Zhu et al., 2024). Prompt tuning methods aim to get better vision-language alignment via only fine-tuning the input prompts. For example, with only few-shot samples for learning, CoOp (Zhou et al., 2021) improved significantly in generalization ability over intensively-tuned manual prompts via prompt learning. Motivated by learning generalization prompts, CoCoOp (Zhou et al., 2022) is proposed to achieve generalization on unseen classes via conditional prompt learning. Adapter-tuning is another popular lightweight fine-tuning method, like CLIP-Adapter (Gao et al., 2023) and Tip-Adapter-F (Zhang et al., 2021a). Both of them inject a lightweight bottleneck architecture after the image encoder and perform residual-style feature blending with the original pre-trained embeddings.

Open-set OOD detection There are multiple lines of work addressing open-set OOD detection, such as anomaly detection (Zong et al., 2018; Liang et al., 2017), outlier detection (Bevandić et al., 2021; Saito et al., 2021), and open-set OOD recognition (Kong & Ramanan, 2021; Geng et al., 2020; Scheirer et al., 2014). These methods can be categorized into two main groups: post hoc methods (Zhu et al., 2022; Liu et al., 2020; Sun et al., 2021; Hendrycks & Gimpel, 2016; Liang et al., 2017; Wang et al., 2022; Sun et al., 2022) and training-time regularization (Narayanaswamy et al., 2023; Bai et al., 2023; Malinin & Gales, 2018; Du et al., 2022b;a; Ming et al., 2022). The former typically resort to post-hoc functions to recognize open-set without altering the DNN training process, like density estimation (Zhang et al., 2020), uncertainty modeling (Gal & Ghahramani, 2016), and input image reconstruction (Pidhorskyi et al., 2018; Sun et al., 2020). On the other hand, regularization-based methods aim to rectify the training process, compelling models to provide predictions with lower confidence.

6. Conclusion

By connecting OOD generalization and OOD detection using Hessians, we have discovered that the proposed EDR loss not only approaches to minimizing energy scores on training data, but also leads to domain-consistent Hessians, thus enabling concurrent optimization for both OOD generalization and OOD detection. Building upon this finding, we introduce a novel Hessian-based OOD generalization bound. From the perspective of bound minimization, we have developed a unified fine-tuning framework named CRoFT, aiming to enhance OOD generalization and open-set OOD detection simultaneously. Extensive experiments under different setups have demonstrated the superiority of our method.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant (No.62106139) and State Key Laboratory of High Performance Computing, National University of Defense Technology (No.202401-11).

Impact Statement

This paper presents work whose goal is to advance the field of robust fine-tuning of VL-PTMs. We aim to improve the model's capacity to address various distribution shifts in real-world applications. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Amari, S.-i. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.

Andreassen, A., Bahri, Y., Neyshabur, B., and Roelofs, R. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

Bai, H., Canal, G., Du, X., Kwon, J., Nowak, R. D., and Li, Y. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, pp. 1454–1471. PMLR, 2023.

Bansal, M., Kumar, M., Sachdeva, M., and Mittal, A. Transfer learning for image classification using vgg19: Caltech-101 image data set. *Journal of ambient intelligence and humanized computing*, pp. 1–12, 2021.

Bevandić, P., Krešo, I., Oršić, M., and Šegvić, S. Dense outlier detection and open-set recognition based on training with noisy negative images. *arXiv preprint arXiv:2101.09193*, 2021.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 446–461. Springer, 2014.

Castro, D. C., Walker, I., and Glocker, B. Causality matters in medical imaging. *Nature Communications*, 2020.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Du, X., Gozum, G., Ming, Y., and Li, Y. Siren: Shaping representations for detecting out-of-distribution objects. *Advances in Neural Information Processing Systems*, 35: 20434–20449, 2022a.

Du, X., Wang, Z., Cai, M., and Li, Y. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022b.

Du, Y. and Mordatch, I. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

Fan, Z., Ma, Y., Li, Z., and Sun, J. Generalized few-shot object detection without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4527–4536, 2021.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pp. 1–15, 2023.

Gao, T., Fisch, A., and Chen, D. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

Geng, C., Huang, S.-j., and Chen, S. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.

Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., and Grover, A. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

Gondal, M. W., Gast, J., Ruiz, I. A., Droste, R., Macri, T., Kumar, S., and Staudigl, L. Domain aligned clip for few-shot classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5721–5730, 2024.

Goyal, S., Kumar, A., Garg, S., Kolter, Z., and Raghunathan, A. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19338–19347, 2023.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Hemati, S., Zhang, G., Estiri, A., and Chen, X. Understanding hessian alignment for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19004–19014, 2023.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Huang, T., Chu, J., and Wei, F. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

Jiang, J., Liu, Z., and Zheng, N. Correlation information bottleneck: Towards adapting pretrained multimodal models for robust visual question answering, 2023.

Katz-Samuels, J., Nakhleh, J. B., Nowak, R., and Li, Y. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pp. 10848–10865. PMLR, 2022.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv:2012.07421*, 2020.

Kong, S. and Ramanan, D. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 813–822, 2021.

Li, D. and Zhang, H. Improved regularization and robustness for fine-tuning in neural networks. *Advances in Neural Information Processing Systems*, 34:27249–27262, 2021.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, 2022.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Liu, S., Tang, Y., Tian, Y., and Su, H. Visual driving assistance system based on few-shot learning. *Multimedia systems*, pp. 1–11, 2021.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Majee, A., Agrawal, K., and Subramanian, A. Few-shot learning for road object detection. In *AAAI Workshop on Meta-Learning and MetaDL Challenge*, pp. 115–126. PMLR, 2021.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Mao, X., Chen, Y., Jia, X., Zhang, R., Xue, H., and Li, Z. Context-aware robust fine-tuning. *International Journal of Computer Vision*, pp. 1–16, 2023.

Meinshausen, N. and Bühlmann, P. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4), aug 2015. doi: 10.1214/15-aos1325. URL https://doi.org/10.1214%2F15-aos1325.

Ming, Y., Fan, Y., and Li, Y. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pp. 15650–15665. PMLR, 2022.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pp. 529–544. Springer, 2022.

Nakamura, A. and Harada, T. Revisiting fine-tuning for few-shot learning. *arXiv preprint arXiv:1910.00216*, 2019.

Narayanaswamy, V., Mubarka, Y., Anirudh, R., Rajan, D., and Thiagarajan, J. J. Exploring inlier and outlier specification for improved medical ood detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 4589–4598, October 2023.

Pidhorskyi, S., Almohsen, R., and Doretto, G. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems*, 31, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Rame, A., Dancette, C., and Cord, M. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377. PMLR, 2022.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Saito, K., Kim, D., and Saenko, K. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. *Advances in Neural Information Processing Systems*, 34:25956–25967, 2021.

Scheirer, W. J., Jain, L. P., and Boult, T. E. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014.

Shafaei, A., Schmidt, M., and Little, J. J. A less biased evaluation of out-of-distribution sample detectors. *arXiv preprint arXiv:1809.04729*, 2018.

Sharan, L., Rosenholtz, R., and Adelson, E. H. Accuracy and speed of material categorization in real-world images. *Journal of vision*, 14(9):12–12, 2014.

Singha, M., Pal, H., Jha, A., and Banerjee, B. Ad-clip: Adapting domains in prompt space using clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4355–4364, 2023.

Sun, X., Yang, Z., Zhang, C., Ling, K.-V., and Peng, G. Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13480–13489, 2020.

Sun, Y., Guo, C., and Li, Y. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Tonin, F., Pandey, A., Patrinos, P., and Suykens, J. A. Unsupervised energy-based out-of-distribution detection using stiefel-restricted kernel machine. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022.

Wang, H., Zhao, C., Guo, Y., Jiang, K., and Chen, F. Towards effective semantic ood detection in unseen domains: A domain generalization perspective. *arXiv preprint arXiv:2309.10209*, 2023a.

Wang, Z., Liang, J., He, R., Xu, N., Wang, Z., and Tan, T. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3032–3042, 2023b.

Wasim, S. T., Naseer, M., Khan, S., Khan, F. S., and Shah, M. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23034–23044, 2023.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998. PMLR, 2022.

Zhang, H., Li, A., Guo, J., and Guo, Y. Hybrid models for open set recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 102–117. Springer, 2020.

Zhang, R., Fang, R., Gao, P., Zhang, W., Li, K., Dai, J., Qiao, Y., and Li, H. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021a.

Zhang, X., Iwasawa, Y., Matsuo, Y., and Gu, S. S. Amortized prompt: Guide clip to domain transfer learning. *arXiv preprint arXiv:2111.12853*, 2021b.

Zhang, Y., Zhang, C., Hu, X., and He, Z. Unsupervised prototype adapter for vision-language models. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 197–209. Springer, 2023.

Zhao, H., Zhang, S., Wu, G., Moura, J. M., Costeira, J. P., and Gordon, G. J. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15659–15669, 2023a.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023b.

Zhu, L., Wang, X., Zhou, C., and Ye, N. Bayesian cross-modal alignment learning for few-shot out-of-distribution generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11461–11469, 2023c.

Zhu, L., Yin, W., Yang, Y., Wu, F., Zeng, Z., Gu, Q., Wang, X., Zhou, C., and Ye, N. Vision-language alignment learning under affinity and divergence principles for few-shot out-of-distribution generalization. *International Journal of Computer Vision*, pp. 1–33, 2024.

Zhu, Y., Chen, Y., Xie, C., Li, X., Zhang, R., Xue, H., Tian, X., Chen, Y., et al. Boosting out-of-distribution detection with typical features. *Advances in Neural Information Processing Systems*, 35:20758–20769, 2022.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.

A. Proof of Theorem 3.3

Before the proof, we first give an explanation of some important notations for clarity.

Notation A.1. We use domain $\mathcal{D}$ to represent a distribution on input space $\mathcal{X}$. Let $\widehat{\mathcal{D}}_{\mathcal{T}}$ and $\widehat{\mathcal{D}}_{\mathcal{S}}$ be the empirical distributions generated with m i.i.d. samples from the source domain (training domain) and the target domain $\mathcal{D}_{\mathcal{S}}$ and training sets, respectively. Consider an bounded instance loss function ℓ such that $\mathcal{Y} \times \mathcal{Y} \rightarrow [0, c]$, and $\ell(\mathbf{y_1}, \mathbf{y_2}) = 0$ if and only if $\mathbf{y_1} = \mathbf{y_2}$ ($\mathbf{y_1} \in \mathcal{Y}, \mathbf{y_2} \in \mathcal{Y}$). We set h as the true label function which generates the ground-truth label of inputs, i.e., $\mathbf{y} = h(\mathbf{x})$. Then we define functional error $\mathcal{E}_{\mathcal{D}}(f; h) := \mathbb{E}_{\mathcal{D}}(f(\mathbf{x}; \theta); h(\mathbf{x}))$, where θ ($\theta \in \Theta$) is the parameter in predictor f. Without introducing any ambiguity, we abbreviate the functional error as $\mathcal{E}_{\mathcal{D}}(\theta)$. Correspondingly, the empirical error calculated based on the empirical distribution $\widehat{\mathcal{D}}_{\mathcal{S}}$ ($\widehat{\mathcal{D}}_{\mathcal{T}}$) is denoted as $\widehat{\mathcal{E}}_{\mathcal{S}}(\theta)$ ($\widehat{\mathcal{E}}_{\mathcal{T}}(\theta)$); the empirical error on the covariate-shifted OOD data is $\widehat{\mathcal{E}}^{c}_{\mathcal{S}}(\theta)$ ($\widehat{\mathcal{E}}^{c}_{\mathcal{T}}(\theta)$).

Theorem A.2. ***[Restatement of Theorem 3.3]*** *Let a VC dimension of the parameter space Θ be v, i.e., $VCdim(\Theta) = v$. If $\widehat{\mathcal{D}}_{\mathcal{T}}$ and $\widehat{\mathcal{D}}_{\mathcal{S}}$ are the empirical distributions generated with m i.i.d. samples from the source domain and target domain, respectively. We denote the corresponding empirical error as $\widehat{\mathcal{E}}_{\mathcal{S}}(\theta)$ and $\widehat{\mathcal{E}}_{\mathcal{T}}(\theta)$. The functional error on the covariate-shifted data of the source domain is represented as $\mathcal{E}^{c}_{\mathcal{S}}(\theta)$. We denote the empirical distributions of covariate-shifted OOD data as $\widehat{\mathcal{D}}^{c}_{\mathcal{S}}$, which is assumed to be near to the ID distributions, i.e, $d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}^{c}_{\mathcal{S}}, \widehat{\mathcal{D}}_{\mathcal{T}}) \leq d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}}, \widehat{\mathcal{D}}_{\mathcal{T}}) + \varepsilon_c$. Then for $0 \leq \delta \leq 1$, with probability at least $1 - \delta$, we have:*

$$\begin{aligned} \mathcal{E}_{\mathcal{T}}(\theta) - \min_{\theta'} \mathcal{E}_{\mathcal{T}}(\theta') &\leq \widehat{\mathcal{E}}^{c}_{\mathcal{S}}(\theta) - \min_{\theta''} \mathcal{E}_{\mathcal{S}}(\theta'') + \lambda + \varepsilon_c \\ &+ d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}}, \widehat{\mathcal{D}}_{\mathcal{T}}) + O\left(\sqrt{\frac{v\log(m/v) + \log(1/\delta)}{m}}\right) \end{aligned} \tag{10}$$

where $\lambda = \mathcal{E}_{\mathcal{S}}(\theta^*) + \mathcal{E}_{\mathcal{T}}(\theta^*)$ and θ^* is the optimal solution on the mixture of source domain and target domain. Utilizing Taylor expansion, we represent $\min_{\theta'} \mathcal{E}_{\mathcal{T}}(\theta')$ as Hessian-based expression, i.e., $\min_{\theta'} \mathcal{E}_{\mathcal{T}}(\theta') \leq \mathcal{E}_{\mathcal{S}}(\theta^*) + \frac{1}{2}\left|\theta^{\top}\mathbf{H}_{\mathcal{S}}(\theta^*)\theta\right| + \varepsilon$., which gives:

$$\begin{aligned} \mathcal{E}_{\mathcal{T}}(\theta) \leq \widehat{\mathcal{E}}^{c}_{\mathcal{S}}(\theta) &- \min_{\theta'} \widehat{\mathcal{E}}_{\mathcal{S}}(\theta') + \frac{1}{2}\left|\theta^{\top}\mathbf{H}_{\mathcal{S}}(\theta^*)\theta\right| + d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}}, \widehat{\mathcal{D}}_{\mathcal{T}}) \\ &+ \lambda_{\mathcal{S}} + \varepsilon + \max\{\lambda_{\mathcal{T}}, 2\lambda_{\mathcal{S}} - \lambda_{\mathcal{T}}\} + O\left(\sqrt{\frac{v\log(m/v) + \log(1/\delta)}{m}}\right) \end{aligned} \tag{11}$$

where $\lambda = \mathcal{E}_{\mathcal{S}}(\theta^*) + \mathcal{E}_{\mathcal{T}}(\theta^*)$, ε and ε_c are small positive values as defined in Assumption 3.2, $\lambda_{\mathcal{S}} = \mathcal{E}_{\mathcal{S}}(\theta^*)$, and $\lambda_{\mathcal{T}} = \mathcal{E}_{\mathcal{T}}(\theta^*)$.

Proof: The proof of Theorem 3.3 consists of three parts. First, we show that the following inequality holds with high probability.

$$\mathcal{E}_{\mathcal{T}}(\theta) \leq \widehat{\mathcal{E}}_{\mathcal{S}}(\theta) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}} + O\left(\sqrt{\frac{v\log(m/v) + \log(1/\delta)}{m}}\right) + \lambda \tag{12}$$

Secondly, we correlate the inequality with covariate-shifted OOD scenarios, which gives Equation 10. Finally, we incorporate the Hessian-based inequality (as defined in Equation 8) into Equation 10.

To be specific, the first part makes use of the following simple inequality which is straightforward to prove, the detailed proof of which can be seen in Lemma 1 of previous work (Cha et al., 2021).

$$|\mathcal{E}_{\mathcal{S}}(f, h) - \mathcal{E}_{\mathcal{T}}(f, h)| \leq \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}) \quad d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}) = 2\sup_A|\Pr_{\mathcal{D}_{\mathcal{S}}}(A) - \Pr_{\mathcal{D}_{\mathcal{T}}}(A)|\,. \tag{13}$$

By triangle inequality, we have the following inequality with probability at least $1-\delta$:

$$\begin{aligned} \mathcal{E}_{\mathcal{T}}(f,h^*) &\leq \mathcal{E}_{\mathcal{T}}(h^*,h)+\mathcal{E}_{\mathcal{T}}(f,h^*) \leq \mathcal{E}_{\mathcal{T}}(h^*,h)+\mathcal{E}_{\mathcal{S}}(f,h^*)+|\mathcal{E}_{\mathcal{T}}(f,h^*)-\mathcal{E}_{\mathcal{S}}(f,h^*)| \\ &\leq \mathcal{E}_{\mathcal{T}}(h^*,h)+\mathcal{E}_{\mathcal{S}}(f,h^*)+\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{T}}) \\ &\leq \mathcal{E}_{\mathcal{T}}(h^*,h)+\mathcal{E}_{\mathcal{S}}(f,h^*)+\mathcal{E}_{\mathcal{S}}(h^*,h)+\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\mathcal{S}},\mathcal{D}_{T}) \\ &= \mathcal{E}_{\mathcal{S}}(f,h^*)+\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\mathcal{S}},\mathcal{D}_{\mathcal{T}})+\lambda \\ &\leq \widehat{\mathcal{E}}_{\mathcal{S}}(f,h^*)+\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}},\widehat{\mathcal{D}}_{\mathcal{T}})+\sqrt{\frac{v(\log(m/v)+1)+\log(1/\delta)}{2m}}+\lambda \end{aligned} \tag{14}$$

where $\lambda=\mathcal{E}_{\mathcal{S}}(h^*,h)+\mathcal{E}_{\mathcal{T}}(h^*,h)=\mathcal{E}_{\mathcal{S}}(\theta^*)+\mathcal{E}_{\mathcal{T}}(\theta^*)$ and $h^*=f(\cdot,\theta^*)$ is the optimal solution on the mixture of the source domain and target domain. The last step in the above inequality is an application of Lemma 2 of (Blitzer et al., 2007). Considering the proposed adversarial-learning-based fine-tuning procedure, where $d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}^c_{\mathcal{S}},\widehat{\mathcal{D}}_{\mathcal{T}})\leq d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}},\widehat{\mathcal{D}}_{\mathcal{T}})+\varepsilon_c$, it is easy to prove that:

$$\mathcal{E}_{\mathcal{T}}(\theta)\leq\widehat{\mathcal{E}}^c_{\mathcal{S}}(\theta)+\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}},\widehat{\mathcal{D}}_{\mathcal{T}})+O\left(\sqrt{\frac{v\log(m/v)+\log(1/\delta)}{2m}}\right)+\lambda+\varepsilon_c \tag{15}$$

For the second part, we set $\bar{\theta}\in\operatorname{argmin}_{\theta\in\Theta}\mathcal{E}_{\mathcal{T}}(\theta)$. Once again using Lemma 2 of (Blitzer et al., 2007), the following inequality holds with probability at most δ:

$$|\widehat{\mathcal{E}}_{\mathcal{S}}(\bar{\theta})-\mathcal{E}_{\mathcal{S}}(\bar{\theta})|>O\left(\sqrt{\frac{v\log(m/v)+\log(1/\delta)}{2m}}\right) \tag{16}$$

Furthermore, we have the following inequality :

$$|\mathcal{E}_{\mathcal{T}}(\theta)-\mathcal{E}_{\mathcal{S}}(\theta)|\leq\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}},\widehat{\mathcal{D}}_{\mathcal{T}})+O\left(\sqrt{\frac{v\log(m/v)+\log(1/\delta)}{2m}}\right) \tag{17}$$

Then, with the probability greater than $1-\delta$, we have:

$$\begin{aligned} \min_{\theta'}\widehat{\mathcal{E}}_{\mathcal{S}}(\theta')\leq\widehat{\mathcal{E}}_{\mathcal{S}}(\bar{\theta})\leq\mathcal{E}^c_{\mathcal{S}}(\bar{\theta})+O\left(\sqrt{\frac{v\log(m/v)+\log(1/\delta)}{m}}\right) \\ \leq\mathcal{E}_{\mathcal{T}}(\bar{\theta})+\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}},\widehat{\mathcal{D}}_{T})+O\left(\sqrt{\frac{v\log(m/v)+\log(1/\delta)}{m}}\right) \\ =\min_{\theta'}\mathcal{E}_{\mathcal{T}}(\theta')+\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}},\widehat{\mathcal{D}}_{\mathcal{T}})+O\left(\sqrt{\frac{v\log(m/v)+\log(1/\delta)}{m}}\right) \end{aligned} \tag{18}$$

Incorporating Equation 18 into Equation 15, we can formulate the OOD generalization bound as below:

$$\begin{aligned} \mathcal{E}_{\mathcal{T}}(\theta)-\min_{\theta'}\mathcal{E}_{\mathcal{T}}(\theta')\leq\widehat{\mathcal{E}}^c_{\mathcal{S}}(\theta)-\min_{\theta'}\widehat{\mathcal{E}}_{\mathcal{S}}(\theta')+\lambda \\ +d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}},\widehat{\mathcal{D}}_{\mathcal{T}})+O\left(\sqrt{\frac{v\log(m/v)+\log(1/\delta)}{m}}\right) \end{aligned} \tag{19}$$

Finally, utilizing Taylor expansion, we denote $\min_{\theta'}\mathcal{E}_{\mathcal{T}}(\theta')$ as the Hessian-based expression, $\min_{\theta'}\mathcal{E}_{\mathcal{T}}(\theta')\leq\max\{\mathcal{E}_{\mathcal{T}}(\theta^*),2\mathcal{E}_{\mathcal{S}}(\theta^*)-\mathcal{E}_{\mathcal{T}}(\theta^*)\}+\frac{1}{2}\left|\theta^\top\mathbf{H}_{\mathcal{S}}(\theta^*)\theta\right|+\varepsilon$. This inequality is a direct result of Theorem 3.6. We will detail it in Section C.

Then the OOD generalization bound as illustrated in Equation 19 can be further expressed as below:

$$\begin{aligned}\mathcal{E}_{\mathcal{T}}(\theta) \leq \widehat{\mathcal{E}}^{c}_{\mathcal{S}}(\theta) - \min_{\theta'} \widehat{\mathcal{E}}_{\mathcal{S}}(\theta') + \frac{1}{2}\left|\theta^{\top}(\mathbf{H}_{\mathcal{S}}(\theta^{*}))\theta\right| + \varepsilon + \varepsilon_c \\ + d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_{\mathcal{S}}, \widehat{\mathcal{D}}_{\mathcal{T}}) + \lambda + \max\{\lambda_{\mathcal{T}}, 2\lambda_{\mathcal{S}} - \lambda_{\mathcal{T}}\} + O\left(\sqrt{\frac{v\log(m/v) + \log(1/\delta)}{m}}\right)\end{aligned} \tag{20}$$

B. Proof of Lemma 3.5

Lemma B.1. ***[Restatement of Lemma 3.5]*** *Let $\mathbf{z} = \{\{\mathbf{z_I}(\mathbf{x^{(i)}})\}_{i=1}^{N}, \{\mathbf{z_T^{(i)}}\}_{i=1}^{K}\}$. The empirical classification loss is denoted as $\widehat{\mathcal{E}}_{\mathcal{D}}(\theta)$, with training data sampled from domain $\mathcal{D}$ and model parameter θ. For each input $\mathbf{x^{(i)}}$, we represent the inner product between its image feature and the text feature of its ground-truth class name as $S^{(i)} = \left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(i)}}(\theta)\right\rangle$. Let $\widehat{\mathbf{G}}_{\mathcal{D}}(\theta_l)$ and $\widehat{\mathbf{H}}_{\mathcal{D}}(\theta_l)$ be the gradient vector and Hessian matrix of empirical risk $\widehat{\mathcal{E}}_{\mathcal{D}}(\theta_l)$ over parameter θ_l, respectively. The local optimum solution of Equation 3 would lead to the following property:*

$$\widehat{\mathbf{H}}_{\mathcal{D}}(\theta_l) = -\frac{1}{N}\sum_{i=1}^{N}\nabla^2_{\theta_l} S^{(i)} = -\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix}\mathbf{0} & \mathbf{z_{I0}^{(i)}}\mathbf{z_{T0}^{(i)}}^{\top} \\ \mathbf{z_{I0}^{(i)}}\mathbf{z_{T0}^{(i)}}^{\top} & \mathbf{0}\end{bmatrix} \tag{21}$$

Proof: To approach the optimal solution of $\min E_{\theta}(\mathbf{x})$ i.e., $\nabla_{\theta} E_{\theta}(\mathbf{x}) \rightarrow \mathbf{0}$, we propose to minimize the squared value of Equation 2 (i.e., magnitude of the gradient vector as defined in Equation 2), which can be approximated by:

$$\mathcal{L}_{\mathrm{e}} = \frac{1}{N}\sum_{i=1}^{N}\left[\nabla_{\theta_l}\left(\log\sum_{j=1}^{K}\exp\left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(j)}}(\theta)\right\rangle\right)\right]^2 \tag{22}$$

where $\theta = \{\theta_0, \theta_l\}$, θ_0 is the frozen parameter in pre-trained encoders and θ_l is the parameter in linear projections that need to be optimized. Now we expand $\mathcal{L}_{\mathrm{e}}$ as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{e}} &= \frac{1}{N}\sum_{i=1}^{N}\left[\nabla_{\theta_l}\left(\log\sum_{j=1}^{K}\exp\left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(j)}}(\theta)\right\rangle\right)\right]^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\nabla_{\theta_l}\sum_{j=1}^{K}\exp\left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(j)}}(\theta)\right\rangle}{\sum_{j=1}^{K}\exp\left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(j)}}(\theta)\right\rangle}\right]^2\end{aligned} \tag{23}$$

We denote the above expression as the magnitude of vector $\mathbf{a}$, i.e., $\mathcal{L}_{\mathrm{e}} = |\mathbf{a}|^2$ and we have

$$\mathbf{a} = -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\nabla_{\theta_l}\sum_{j=1}^{K}\exp\left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(j)}}(\theta)\right\rangle}{\sum_{j=1}^{K}\exp\left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(j)}}(\theta)\right\rangle}\right]$$

.

The empirical classification loss, $\mathcal{E}_{\mathcal{D}}(\mathbf{z};\theta)$, can be calculated as:

$$\begin{aligned}\mathcal{E}_{\mathcal{D}}(\mathbf{z};\theta) &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp S^{(i)}}{\sum_{j=1}^{K}\exp\left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(j)}}(\theta)\right\rangle} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\log\sum_{j=1}^{K}\exp\left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(j)}}(\theta)\right\rangle - S^{(i)}\right]\end{aligned} \tag{24}$$

Accordingly, the gradient vector of empirical risk $\widehat{\mathcal{E}}_{\mathcal{D}}(\mathbf{z};\theta)$ with respect to parameter θ_l is represented as:

$$\widehat{\mathbf{G}}_{\mathcal{D}}(\theta_l) = \nabla_{\theta_l}\widehat{\mathcal{E}}_{\mathcal{D}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\nabla_{\theta_l}\sum_{j=1}^{K}\exp\left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(j)}}(\theta)\right\rangle}{\sum_{j=1}^{K}\exp\left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(j)}}(\theta)\right\rangle} - \nabla_{\theta_l}S^{(i)}\right] = -\mathbf{a} - \frac{1}{N}\sum_{i=1}^{N}\nabla_{\theta_l}S^{(i)} \tag{25}$$

And the Hessian matrix of empirical risk $\mathcal{E}_{\mathcal{D}}(\theta)$ with respect to parameter θ_l is calculated as:

$$\widehat{\mathbf{H}}_{\mathcal{D}}(\theta_l) = \nabla^2_{\theta_l}\widehat{\mathcal{E}}_{\mathcal{D}}(\theta) = -\nabla_{\theta_l}\mathbf{a} - \frac{1}{N}\sum_{i=1}^{N}\nabla^2_{\theta_l}S^{(i)} \tag{26}$$

We further calculate $\nabla_\theta \mathcal{L}_\text{e}$ as below:

$$\nabla_{\theta_l}\mathcal{L}_\text{e} = 2\cdot(\nabla_{\theta_l}\mathbf{a})\cdot\mathbf{a} \tag{27}$$

The local optimum solution of Equation 3, i.e., $\nabla_\theta\mathcal{L}_\text{e} = \mathbf{0}$, gives the following equation:

$$\nabla_{\theta_l}\mathbf{a} = -\widehat{\mathbf{H}}_{\mathcal{D}}(\theta_l) - \frac{1}{N}\sum_{i=1}^{N}\nabla^2_{\theta_l}S^{(i)} = \mathbf{0} \tag{28}$$

In the regime of adapter tuning on CLIP's pre-trained features, the inner product between the adapted image feature and the adapted text feature of its ground-truth class name is calculated as $S^{(i)} = \left\langle \mathbf{z_I}(\mathbf{x^{(i)}};\theta), \mathbf{z_T^{(i)}}(\theta)\right\rangle = \left\langle \theta_\mathbf{I}\mathbf{z_{IO}^{(i)}}, \theta_\mathbf{T}\mathbf{z_{TO}^{(i)}}\right\rangle$. Note that $\theta_\mathbf{I}\in\mathbb{R}^{d\times d}$ and $\theta_\mathbf{T}\in\mathbb{R}^{d\times d}$ is the parameter in image adapter and text adapter, respectively. And $\theta_l\in\mathbb{R}^{2d^2}$ is the parameter vector contains all elements in $\theta_\mathbf{I}$ and $\theta_\mathbf{T}$. Finally, we can conclude that the local optimum solution of Equation 3 leads to the following property:

$$\widehat{\mathbf{H}}_{\mathcal{D}}(\theta_l) = -\frac{1}{N}\sum_{i=1}^{N}\nabla^2_{\theta_l}S^{(i)} = -\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix}\mathbf{0} & \mathbf{z_{IO}^{(i)}}\mathbf{z_{TO}^{(i)}}^\top \\ \mathbf{z_{IO}^{(i)}}\mathbf{z_{TO}^{(i)}}^\top & \mathbf{0}\end{bmatrix} \tag{29}$$

C. Proof of Theorem 3.6

Proof: Let θ^* be the local minimum across all domains, i.e., $\nabla_\theta\widehat{\mathcal{E}}_{\mathcal{D}}(\theta^*) = \mathbf{0}$, and $\mathcal{D} = \{\mathcal{S},\mathcal{T}\}$. By Taylor expansion, the OOD generalization gap between source domain ($\mathcal{S}$) and target domain ($\mathcal{T}$) is upper bounded as shown in the following equation:

$$\begin{aligned}
&\max_{\{\theta:|\widehat{\mathcal{E}}_{\mathcal{S}}(\theta)-\widehat{\mathcal{E}}_{\mathcal{S}}(\theta^*)|\leq\epsilon\}} |\widehat{\mathcal{E}}_{\mathcal{T}}(\theta) - \widehat{\mathcal{E}}_{\mathcal{S}}(\theta^*)| \\
&\approx \max_{\{\theta:\frac{1}{2}|\theta^\top\widehat{\mathbf{H}}_{\mathcal{S}}(\theta^*)\theta|\leq\epsilon\}} \left|\widehat{\mathcal{E}}_{\mathcal{T}}(\theta^*) + \frac{1}{2}\theta^\top\widehat{\mathbf{H}}_{\mathcal{T}}(\theta^*)\theta - \widehat{\mathcal{E}}_{\mathcal{S}}(\theta^*)\right| \\
&\lesssim |\widehat{\mathcal{E}}_{\mathcal{T}}(\theta^*) - \widehat{\mathcal{E}}_{\mathcal{S}}(\theta^*)| + \max_{\{\theta:\frac{1}{2}|\theta^\top\widehat{\mathbf{H}}_{\mathcal{S}}(\theta^*)\theta|\leq\epsilon\}} \frac{1}{2}\left|\theta^\top\widehat{\mathbf{H}}_{\mathcal{T}}(\theta^*)\theta\right| \\
&\lesssim |\widehat{\mathcal{E}}_{\mathcal{T}}(\theta^*) - \widehat{\mathcal{E}}_{\mathcal{S}}(\theta^*)| + \max_{\{\theta:\frac{1}{2}|\theta^\top\widehat{\mathbf{H}}_{\mathcal{S}}(\theta^*)\theta|\leq\epsilon\}} \frac{1}{2N}\sum_{i=1}^{N}\left|\theta^\top\nabla^2_\theta S^{(i)}_{\mathcal{T}}\theta\right| \\
&\lesssim |\widehat{\mathcal{E}}_{\mathcal{T}}(\theta^*) - \widehat{\mathcal{E}}_{\mathcal{S}}(\theta^*)| + \max\frac{1}{2}|\theta^\top\widehat{\mathbf{H}}_{\mathcal{S}}(\theta^*)\theta| + \varepsilon
\end{aligned} \tag{30}$$

The last two lines in the above inequality follow:

$$\left|\theta^\top\widehat{\mathbf{H}}_{\mathcal{D}}(\theta)\theta\right| = \left|\theta^\top\frac{1}{N}\sum_{i=1}^{N}\nabla^2_\theta S^{(i)}\theta\right| \leq \frac{1}{N}\sum_{i=1}^{N}\left|\theta_l^\top\begin{bmatrix}\mathbf{0} & \mathbf{z_{IO}^{(i)}}\mathbf{z_{TO}^{(i)}}^\top \\ \mathbf{z_{IO}^{(i)}}\mathbf{z_{TO}^{(i)}}^\top & \mathbf{0}\end{bmatrix}\theta_l\right| \tag{31}$$

For each image feature $\mathbf{z_I}$ from the source domain, the image features $\tilde{\mathbf{z}}_\mathbf{I}$ from the target domain, which share the same label with $\mathbf{z_I}$, is assumed to satisfy: $||\mathbf{z_I} - \tilde{\mathbf{z}}_\mathbf{I}||_2 \leq \varepsilon$. Since we optimize the EDR loss both on ID image features and the generated OOD image features, we have the following approximation:

$$\begin{aligned}\left|\theta^\top \widehat{\mathbf{H}}_{\mathcal{T}}(\theta)\theta\right| &\lesssim \frac{1}{N}\sum_{i=1}^{N}\left|\theta_l^\top \begin{bmatrix} \mathbf{0} & (\mathbf{z_{I0}^{(i)}}+\varepsilon)\mathbf{z_{T0}^{(i)}}^\top \\ (\mathbf{z_{I0}^{(i)}}+\varepsilon)\mathbf{z_{T0}^{(i)}}^\top & \mathbf{0}\end{bmatrix}\theta_l\right| \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\left|\theta_l^\top \begin{bmatrix} \mathbf{0} & \mathbf{z_{I0}^{(i)}}\mathbf{z_{T0}^{(i)}}^\top \\ \mathbf{z_{I0}^{(i)}}\mathbf{z_{T0}^{(i)}}^\top & \mathbf{0}\end{bmatrix}\theta_l\right| + \frac{1}{N}\sum_{i=1}^{N}\left|\theta_l^\top \begin{bmatrix} \mathbf{0} & \varepsilon\mathbf{z_{T0}^{(i)}}^\top \\ \varepsilon\mathbf{z_{T0}^{(i)}}^\top & \mathbf{0}\end{bmatrix}\theta_l\right| \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\left|\theta_l^\top \begin{bmatrix} \mathbf{0} & \mathbf{z_{I0}^{(i)}}\mathbf{z_{T0}^{(i)}}^\top \\ \mathbf{z_{I0}^{(i)}}\mathbf{z_{T0}^{(i)}}^\top & \mathbf{0}\end{bmatrix}\theta_l\right| + O(\varepsilon) = \left|\theta^\top \widehat{\mathbf{H}}_{\mathcal{S}}(\theta)\theta\right| + O(\varepsilon)\end{aligned} \tag{32}$$

It should be noted that Equation 30 also holds in expectation, which gives: $\min_{\theta'} \mathcal{E}_{\mathcal{T}}(\theta') \leq \mathcal{E}_{\mathcal{T}}(\theta) \leq \max\{\mathcal{E}_{\mathcal{T}}(\theta^*), 2\mathcal{E}_{\mathcal{S}}(\theta^*) - \mathcal{E}_{\mathcal{T}}(\theta^*)\} + \frac{1}{2}\left|\theta^\top \mathbf{H}_{\mathcal{S}}(\theta^*)\theta\right| + \varepsilon$.

Algorithm 1 Algorithm of the proposed CRoFT

```
Input: ID data x^(i) (i ∈ 1, ··· , N), ID class names of the K-way classification, hyperparameter λ1 and λ2, maximum
epoch T.
for t = 1 to T do
    Calculate the pre-trained ID image features z_I0 and pre-trained language features z_T0 based on the zero-shot CLIP;
    Calculate the adapted image features z_I(x^(i), θ) and adapted text features z_T^(j)(θ);
    Generate the worst-case covariate-shifted OOD image features by Equation 5;
    Minimize the classification loss of ID image features and the generated OOD image features while reshaping their
    energy distribution by Ê_S(θ) + λ1 L_c + λ2(L_e(z_I) + L_e(z_I^c));
end for
Output: Adapters' parameters θ_l.
```

D. More Experiment Details

Based on the code of CoOp (Zhou et al., 2021), we train all models with SGD optimizer with a learning rate of $2e-2$. The batch size is set to 32 except for the experiments of Tip-Adapter-F on Setup-I. Following the original paper of Tip-Adapter-F, we set the batch size as 256. For the specific hyperparameter for each method, we follow the setting of the original paper.

For the prompt learning methods, CoOp (Zhou et al., 2021), CoCoOp (Zhou et al., 2022), DPLCLIP (Zhang et al., 2021b), and Bayes-CAL (Zhu et al., 2023c), we use random initialization for context vectors and set the number of context tokens as 16, set the class token position (CTP) as "end", and set the class-specific context (CSC) as "False". This configuration has shown the best average performance according to CoOp's paper. For the DPLCLIP (Zhang et al., 2021b) method, we set the additional hyper-parameters of DPLCLIP (Zhang et al., 2021b) as: "mlp_depth=3", "mlp_width=512", and "mlp_dropout=0.1".

For the CLIP-Adapter (Gao et al., 2023) method, we adopt image adapter only with the residual ratio of 0.2, and we use the bottleneck adapter with a hidden dimension that is 1/4 of the original feature dimension. This hyperparameter configuration has been demonstrated as the most effective for generic image datasets, such as ImageNet, in the original research (Gao et al., 2023).

For the Tip-Adapter-F (Zhang et al., 2021a) method, in Setup-I, we conduct the hyperparameter search on the learning rate and the additional hyperparameter of Tip-Adapter-F, i.e., α and β in Tip-Adapter-F's paper. The corresponding hyperparameter-search results are in Table 6. According to Table 6, we select the learning rate as 0.0001, α as 0.5, and β as 5.5. In Setup-II, we set the initial α as 1, while the initial beta value is searched within the range of [1, 5, 10]. This search is conducted using the validation sets to find the optimal value for β.

E. More Experiment Results

Sensitivity analysis of hyperparameters Based on 32-shot samples in Setup-I, we provide the ablation study results on the hyperparameter λ_1 and λ_2 in Table 7. The sensitivity analysis of hyperparameters on Setup-I in our paper once again

Table 6. Tip-Adapter-F's ID accuracy at different hyperparameter settings.

bath_size=32, lr=0.001								bath_size=32, $\alpha = 0.5, \beta = 5.5$			
(α, β)	Shot = 1	Shot = 16	Shot = 32	(α, β)	Shot = 1	Shot = 16	Shot = 32	lr	Shot = 1	Shot = 16	Shot = 32
(1.0, 1.5)	79.77±0.34	80.37±0.18	80.45±0.88	(0, 5.5)	79.44±0.11	79.76±0.43	79.85±0.47				
(1.0, 3.5)	80.03±0.15	81.11±0.60	80.40±0.59	(0.5, 5.5)	79.92±0.20	82.40±0.88	82.24±0.36	0.0001	79.45 ± 0.15	82.30 ± 0.93	83.28 ± 0.25
(1.0, 5.5)	79.85±0.29	81.65±0.46	81.62±0.58	(1, 5.5)	79.85±0.29	81.65±0.46	81.62±0.58	0.0005	79.91 ± 0.20	82.18 ± 0.28	82.46 ± 0.30
(1.0, 7.5)	79.88±0.17	81.57±0.14	82.24±0.18	(2, 5.5)	79.95±0.17	80.32±0.32	80.41±0.88	0.0010	79.92 ± 0.20	82.40 ± 0.88	82.24 ± 0.36
(1.0, 9.5)	79.72±0.24	81.94±0.74	82.14±0.35	(3, 5.5)	79.88±0.28	79.76±0.81	79.57±0.55	0.0015	79.80 ± 0.07	81.71 ± 0.43	82.08 ± 0.21
(1.0, 11.5)	79.65±0.36	81.50±0.34	82.38±0.41	(4, 5.5)	79.91±0.19	79.51±0.63	80.21±0.42	0.0020	79.44 ± 0.45	81.40 ± 0.31	82.25 ± 0.05

Table 7. The ablation study results on the hyperparameter λ_1 and λ_2. All results are based on 32-shot samples in Setup-I.

λ_2 ($\lambda_1 = 0$)	0	5	10	15	20
AUROC	77.25	**85.18**	78.56	80.14	78.41
Worst-Case ACC	42.51	42.84	42.71	42.91	**43.16**
OOD ACC	61.96	62.37	62.30	62.40	**62.55**
λ_1 ($\lambda_2 = 0$)	0	5	10	15	20
AUROC	77.25	87.07	87.32	87.38	**87.66**
Worst-Case ACC	42.51	43.65	43.71	**43.92**	43.61
OOD ACC	61.96	62.75	62.73	62.75	**63.03**

demonstrates strong evidence of CRoFT's concurrent optimization for both tasks. Our results show that the incorporation of each regularization term leads to improvements in both tasks, highlighting their effectiveness in concurrent optimization.

Visualization on the improved open-set OOD detection In this section, we present more visualization results to intuitively illustrate the effectiveness of our CRoFT method in facilitating open-set OOD detection compared to the zero-shot CLIP model. As shown in Figure 5, CLIP's energy distributions on different types of data are highly overlapped, thus CLIP may fail to identify opens-set OOD examples. Our method significantly reduced the overlap between the energy distributions of closed-set and open-set. For Setup-II, we visualize the distribution of image features in Figure 5 (e), where all features are reduced to 1 dimension by t-SNE. We can observe that CLIP obtained significantly different closed-set image features, despite sharing the same categories, hindering the discrimination between the closed-set OOD and open-set OOD data. In contrast, our method can learn more discriminated image features between closed-set and open-set, improving the FPR and AUROC results when discriminating the two types of OOD data by around 20% as shown in Figure 5 (f).

CRoFT's effectiveness on robust few-shot learning We conducted additional experiments on benchmark datasets commonly used for CLIP-based methods, including Calteh101, Oxford Flowers, DTD, FGVCAircraft, UCF101, Oxford Pets,

Table 8. Ablation study results on $\mathcal{L}_c$ and $\mathcal{L}_e$. We report the OOD test accuracy for each domain in this table. The best overall performances on both OOD generalization and open-set OOD detection validate the theoretical results of Theorem 3.3 and 3.6.

DATA	Method	w / o $\mathcal{L}_\mathrm{c}$		w / o $\mathcal{L}_\mathrm{e}$		Our	
	V-Net	RN50	ViT16	RN50	ViT16	RN50	ViT16
PACS	P	99.42	99.85	**99.46**	99.82	99.45	**99.88**
	A	94.64	98.35	94.71	98.68	**94.82**	**98.94**
	C	93.04	97.29	93.34	**98.00**	**93.38**	97.66
	S	82.08	91.18	80.56	90.32	**82.16**	**92.57**
	Avg	92.29	96.67	92.02	96.71	**92.45**	**97.26**
VLCS	V	**99.60**	**99.93**	98.37	99.72	98.94	99.88
	L	52.99	54.76	62.52	55.31	**67.09**	**61.11**
	C	83.54	85.30	80.38	84.79	**83.75**	**86.14**
	S	**73.25**	71.80	69.52	70.41	68.08	**73.73**
	Avg	77.34	77.95	77.70	77.56	**79.47**	**80.21**

Table 9. Comparison with state-of-the-art methods on few-shot benchmark datasets with CLIP RN50 pre-trained model.

Method	OxfordPets	EuroSAT	Caltech101	DTD	FGVCAircraft	Flowers102	UCF101	Food101	SUN397	StanfordCars	Imagenet	Average
ZS CLIP	85.77	37.56	86.29	42.32	17.28	66.14	61.46	77.31	58.52	55.74	60.32	56.69
Tip,shots=1	86.10	54.38	87.18	46.22	19.05	73.12	62.60	77.42	61.30	57.54	60.70	62.33
Tip,shots=2	**87.03**	61.68	88.44	49.47	21.21	79.13	64.74	77.52	62.70	57.93	60.96	64.62
Tip,shots=4	86.45	65.32	89.39	53.96	22.41	83.80	66.46	77.54	64.15	61.45	60.98	66.54
Tip,shots=8	87.03	67.95	89.83	58.63	25.59	87.98	68.68	77.76	65.62	62.93	61.45	68.50
Tip,shots=16	88.14	70.54	90.18	60.93	29.76	89.89	70.58	77.83	66.85	66.77	62.01	70.32
Tip-F,shots=1	**87.00**	59.53	89.33	49.65	20.22	79.98	64.87	77.51	62.50	58.86	61.13	64.60
Tip-F,shots=2	**87.03**	66.15	89.74	53.72	23.19	82.30	66.43	77.81	63.64	61.50	61.69	66.65
Tip-F,shots=4	87.54	74.12	90.56	57.39	25.80	88.83	70.55	78.24	66.21	64.57	62.52	69.67
Tip-F,shots=8	88.09	77.93	91.44	62.71	30.21	91.51	74.25	78.64	68.87	69.25	**64.00**	72.45
Tip-F,shots=16	89.70	**84.54**	92.86	66.55	35.55	94.80	78.03	79.43	71.47	75.74	**65.51**	75.83
Adapter,shots=1	85.99	61.40	88.60	45.80	17.49	73.49	62.20	76.82	61.30	55.13	61.20	62.67
Adapter,shots=2	86.73	63.90	89.37	51.48	20.10	81.61	67.12	77.22	63.29	58.74	61.52	65.55
Adapter,shots=4	87.46	73.38	89.98	56.86	22.59	87.17	69.05	77.92	65.96	62.45	61.84	68.61
Adapter,shots=8	87.65	77.93	91.40	61.00	26.25	91.72	73.30	78.04	67.50	67.89	62.68	71.40
Adapter,shots=16	87.84	84.43	92.49	65.96	32.10	93.90	76.76	78.25	69.55	74.01	63.59	74.44
CoOp,shots=1	85.89	50.63	87.53	44.39	9.64	68.12	61.92	74.32	60.29	55.59	57.15	59.59
CoOp,shots=2	82.64	61.50	87.93	45.15	18.68	77.51	64.09	72.49	59.48	58.28	57.81	62.32
CoOp,shots=4	86.70	70.18	89.55	53.49	21.87	86.20	67.03	73.33	63.47	62.62	59.99	66.77
CoOp,shots=8	85.32	76.73	90.21	59.97	26.13	91.18	71.94	71.82	65.52	68.43	61.56	69.89
CoOp,shots=16	87.01	83.53	91.83	63.58	31.26	94.51	75.71	74.67	69.26	73.36	62.95	73.42
Ours,shots=1	86.43	**64.64**	**89.49**	**50.18**	**21.81**	**81.45**	**66.30**	**77.36**	**63.03**	**58.86**	**61.61**	**65.56**
Ours,shots=2	86.86	**67.60**	**90.47**	**55.32**	**24.51**	**85.91**	**69.07**	**77.90**	**65.87**	**62.17**	**62.31**	**68.00**
Ours,shots=4	**88.28**	**75.15**	**91.44**	**59.34**	**27.06**	**91.47**	**71.72**	**78.35**	**68.05**	**66.61**	**62.88**	**70.94**
Ours,shots=8	**88.96**	**78.23**	**92.17**	**63.24**	**33.39**	**94.03**	**75.50**	**78.36**	**69.79**	**70.17**	63.78	**73.42**
Ours,shots=16	**89.97**	**84.54**	**93.27**	**67.26**	**38.85**	**95.66**	**78.43**	**79.44**	**71.90**	**76.09**	65.20	**76.42**

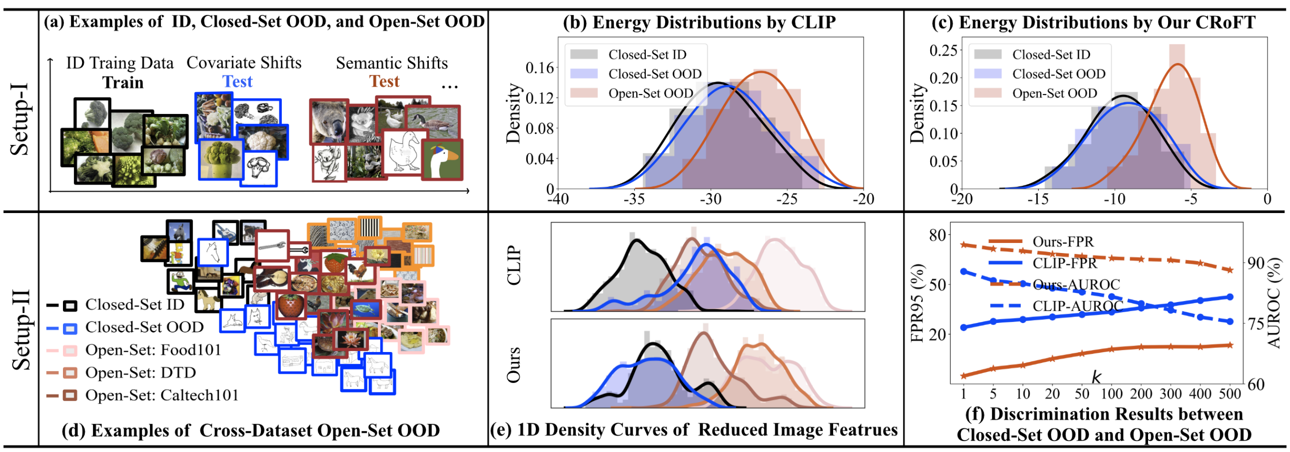

Figure 5. (a): Examples of three types of data in Setup-I: (i) closed-set ID data (e.g., broccoli), (ii) closed-set OOD data (e.g., broccoli with changed image styles), and (iii) open-set OOD data (e.g., goose). (b): CLIP's energy distributions on different types of data. (c): CRoFT's energy distributions on different types of data. (d): Examples of the closed-set data and cross-dataset open-set OOD data in Setup-II, where we use the PACS dataset as the closed-set data. (e): Visualization of image features. All image features are reduced to the 1-dimensional space by t-SNE. (f): The FPR95 and AUROC results in discriminating closed-set OOD and open-set OOD data.

Table 10. Energy score for the closed-set OOD, closed-set ID, generated closed-set OOD, and open-set OOD data. Based on Setup-I and our 32-shot fine-tuned CRoFT model, we present the 5th, 25th, 50th, 75th, and 95th percentiles of energy scores.

DATA	0.05	0.25	0.50	0.75	0.95
ID	-17.36	-13.59	-12.08	-10.04	-8.00
Closed-Set OOD	-16.90	-13.34	-11.77	-9.79	-7.75
Generated Closed-Set OOD	-17.19	-13.42	-11.89	-9.84	-7.78
Open-Set OOD	-12.76	-9.99	-8.55	-6.88	-4.97

and EuroSAT. Based on CLIP RN50, the 1-shot, 2-shot, 4-shot, and 16-shot results of our method and the SOTA method Tip-Adapetr-F are reported in Table 9. The results show that our method achieves higher test accuracy in most cases and obtains the best average results under various training shots. CRoFT also outperforms Tip-Adapetr-F stably. It is evident that CRoFT also shows stable improvements under very-low shot conditions, with about a 1.5% boost on average under the 1-shot and 2-shot settings.

Energy distribution of the generated worst-case covariate-shifted OOD features We present the energy distribution of the generated worst-case covariate-shifted OOD features to highlight the quality of these OOD features. Based on Setup-I and our 32-shot fine-tuned CRoFT model, we present the 5th, 25th, 50th, 75th, and 95th percentiles of energy scores for the generated closed-set OOD, closed-set ID, closed-set OOD, and open-set OOD across 3 runs. As shown in Table 10, our experiment results demonstrate that the energy distribution of the generated worst-case covariate-shifted OOD feature closely resembles that of the closed-set OOD data. By optimizing these high-quality generated OOD features, we can enhance the model's generalization capacity for closed-set OOD data, thereby improving the model's ability to discriminate between closed-set OOD data and open-set OOD data.