Bayesian Cross-Modal Alignment Learning for Few-Shot Out-of-Distribution Generalization

Lin Zhu, Xinbing Wang, Chenghu Zhou, Nanyang Ye*

Shanghai Jiao Tong University, Shanghai, China
zhulin_sjtu@sjtu.edu.cn, xwang8@sjtu.edu.cn, zhouchsjtu@gmail.com, ynylincoln@sjtu.edu.cn

Abstract

Recent advances in large pre-trained models showed promising results in few-shot learning. However, their generalization ability on two-dimensional Out-of-Distribution (OoD) data, i.e., correlation shift and diversity shift, has not been thoroughly investigated. Researches have shown that even with a significant amount of training data, few methods can achieve better performance than the standard empirical risk minimization method (ERM) in OoD generalization. This few-shot OoD generalization dilemma emerges as a challenging direction in deep neural network generalization research, where the performance suffers from overfitting on few-shot examples and OoD generalization errors. In this paper, leveraging a broader supervision source, we explore a novel Bayesian cross-modal image-text alignment learning method (Bayes-CAL) to address this issue. Specifically, the model is designed as only text representations are fine-tuned via a Bayesian modelling approach with gradient orthogonalization loss and invariant risk minimization (IRM) loss. The Bayesian approach is essentially introduced to avoid overfitting the base classes observed during training and improve generalization to broader unseen classes. The dedicated loss is introduced to achieve better image-text alignment by disentangling the causal and non-casual parts of image features. Numerical experiments demonstrate that Bayes-CAL achieved state-of-the-art OoD generalization performances on two-dimensional distribution shifts. Moreover, compared with CLIP-like models, Bayes-CAL yields more stable generalization performances on unseen classes. Our code is available at https://github.com/LinLLLL/BayesCAL.

.

Introduction

Few-shot learning is an emerging research topic that aims to generalize from only a few training samples (Wang et al. 2020). Despite the success of recent few-shot learning methods on independent and identically distributed (I.I.D) settings (Finn, Abbeel, and Levine 2017; Rusu et al. 2018; Sung et al. 2018; Vuorio et al. 2019; Fan et al. 2021), these few-shot learning methods would suffer significant performance drop in the presence of domain differences between source and target datasets (Chen et al. 2019; Guo et al. 2020). This domain shift issue commonly exists in real scenarios, especially in the few-shot setting. For example, it is difficult to construct large training datasets for rare species, there can be huge differences between the training and test environments due to the large randomness of the few-shot setting.

To address this domain shift problem, i.e., training and test data access different conditional distributions, there have been plenty of methods proposed to achieve domain generalization in few-shot learning scenarios(Tseng et al. 2020; Liu et al. 2021; Liang et al. 2021; Zhou and Tan 2021). However, most of these studies focused on this issue without considering the different characteristics of distribution shifts between training and test domains. In this paper, we consider a more practical setting of *few-shot Out-of-Distribution (OoD) generalization*, focusing on the few-shot image recognition task of generalizing under two major types of distribution shifts (as described in OoD-Bench (Ye et al. 2021)). Specifically, models usually have access to multi-domain "K-shot N-way" training samples of OoD data

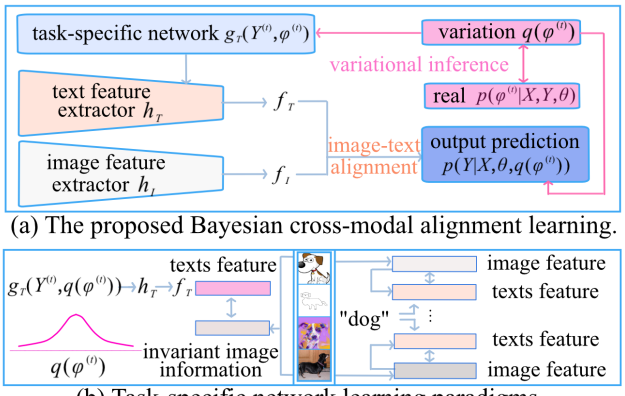


Figure 1: Illustration of the proposed Bayesian cross-modal alignment learning paradigms. To avoid overfitting on the few-shot training samples, $\varphi^{(t)}$ is modelled by a variational distribution $q(\varphi^{(t)})$ to approximate its posterior distribution.

*Corresponding author. The full Appendix is available on Arxiv.

dominated by diversity shift or correlation shift. According to (Ye et al. 2021), the diversity shift is defined by the support set's differences on latent environment's distributions (e.g., changes in the image style), and the correlation shift is defined by the probability density functions' differences caused by spurious correlations. Under this condition, models are expected to learn feature distributions only from the seen domains and generalize to unseen domains within the same downstream task.

In recent few-shot learning methods, studies(Rahman, Khan, and Barnes 2020; Zhou et al. 2022b; Zhang et al. 2021a; Zhu et al. 2021) based on large-scale vision-language pre-trained models achieved striking performances in various downstream tasks. The conventional visual deep learning models that only focus on closed-set visual concepts, are susceptible to overfitting on the pre-defined list of classes due to their limited supervision source (Zhou et al. 2022a). In contrast, vision-language models leverage a broader source of supervision coming from natural language, which has been proven effective in learning transferable representations (Jia et al. 2021; Radford et al. 2021).

A widely utilized method in vision-language pre-trained models is the cross-modal representation learning (Fang et al. 2022; Li et al. 2020; Wehrmann, Kolling, and Barros 2020; Zheng et al. 2020). Inspired by the advantages of the cross-modal alignment and Bayesian methods (Lin et al. 2022) for alleviating overfitting, instead of focusing on improving learning algorithms for learning OoD generalizable image features, we propose the Bayesian cross-modal alignment learning method (Bayes-CAL) to achieve few-shot OoD generalization. Since fine-tuning the entire model is impractical and might damage the well-learned representation space (Zhou et al. 2022a) and adjusting the text representations are more flexible, as shown in Figure 1 (a), we design the model architecture as only a task-specific network in text feature extractor is tuned in each specific downstream task.

In this paper, we instantiate the text representation learning of Bayes-CAL in three methods–prompt leaning (Ding et al. 2021), directly utilizing learnable vectors, and Word2Vector (Mikolov et al. 2013) to detail how Bayes-CAL works on few-shot OoD generalization. Our key contributions are as follows:

1. We propose a Bayesian treatment for cross-modal alignment learning for few-shot OoD generalization. The superiority of the Bayesian method is demonstrated by stable generalization performances on unseen classes. We have also carefully designed experiments to gain insight into the superiority of image-text alignment learning.
2. Under the proposed architecture (see Figure 2), gradient orthogonalization loss is introduced to achieve better alignment learning by disentangling image features. Invariant risk minimization (IRM) loss is utilized to improve OoD generalization ability further.
3. Bayes-CAL has achieved state-of-the-art performances on OoD-Bench datasets with both diversity shift and correlation shift, especially 10%-20% performance improvements compared with algorithms in OoD-Bench (Ye et al. 2021). Moreover, it outperforms CLIP-like solid models by more stable generalization performance on both I.I.D and OoD unseen classes.

Related Work

Foundation Models

In this paper, we focus on foundation models of large-scale vision-language pre-training, which have recently emerged (Gu et al. 2021; Dai et al. 2021; Radford et al. 2021) for various image-text retrieval tasks. Especially for image recognition, a contrastive language image pre-training model (CLIP, (Radford et al. 2021)) is proposed. In CLIP, images and texts are encoded to the feature space. Then, the model is optimized to maximize the similarity of image features and texts features. There are also many efficient CLIP-based models to enhance generalization performance via prompt tuning or image feature adapters. Prompt tuning is a type of method to get better vision-language alignment via only fine-tuning the input prompts, such as CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), DPLCLIP (Zhang et al. 2021c), etc. An alternative path is to conduct fine-tuning with image feature adapters on visual feature space, like CLIP-Adapter (Gao et al. 2021) and Tip-Adapter (Zhang et al. 2021b). In this paper, we mainly focus on the methods that fine-tune on the semantic space of texts. For example, with only few-shot samples for learning, CoOp improved significantly in generalization ability over intensively-tuned manual prompts via prompt learning. However, a critical problem of CoOp is identified as its learned context is not generalizable to unseen classes (Zhou et al. 2022a). Motivated by learning generalization prompts, CoCoOp is proposed to achieve generalization on unseen classes via conditional prompt learning. Nevertheless, CoCoOp requires for each image an independent forward pass of instance-specific prompts, which significantly decreases its training efficiency. Another method Domain Prompt Learning (DPLCLIP), is proposed to guide CLIP to domain transfer learning by domain inference in the form of prompt generating. It captures domain shifts by extracting information from image features, which limits its generalization ability in distribution shifts that can not be extracted efficiently from few-shot images. For more information on foundation models, we refer readers to this survey (Du et al. 2022).

Out-of-Distribution Generalization Algorithms

Plenty of methods for OoD generalization have been proposed recently. Typically, they can be categorized into three types—1) Invariant learning-based methods, such as invariant risk minimization (IRM, (Arjovsky et al. 2019)), invariant risk minimization games (IRM-Games (Ahuja et al. 2020)), etc. 2) Domain generalization methods, such as the Jigsaw method (Jigsaw, (Carlucci et al. 2019)), the representation self-challenging method (RSC, (Huang et al. 2020)), etc. For more detailed information, we refer readers to this survey (Wang et al. 2022). 3) Stable learning methods, such as the sample reweighting method (Shen et al. 2020). Although these methods yield OoD generalization performance improvement to some extent, it has been recently

demonstrated that they are pretty hard to systematically beat the standard ERM method (Gulrajani and Lopez-Paz 2020; Ye et al. 2021). Furthermore, learning from the few-shot OoD samples is much more challenging due to their large randomness, and the few-shot OoD generalization under two major distribution shifts is rarely understood.

Methodology

We propose a novel Bayesian cross-modal alignment learning method (Bayes-CAL) for few-shot OoD generalization. Unlike CoCoOp and DPLCLIP that fine-tune task-specific parameters by incorporating the conditional information extracted from image features, we fine-tune on the semantic space by enforcing domain-invariant alignment under the proposed regularization terms. Moreover, the Bayesian treatment is specially introduced to substantially alleviate overfitting. Based on the domain-invariant information disentangled from the image features, the distributions of the task-specific parameters are estimated. Without a query of a large amount of GPU memory like CoCoOp in every run, the proposed Bayes-CAL is simple yet efficient, making fine-tuning on few-shot samples practical in the few-shot OoD setting. An overview of the Bayes-CAL is shown in Figure 2.

Preliminary on Few-Shot Cross-Modal Alignment Learning

In a specific downstream classification task t, models are supposed to learn from the few-shot OoD training data $D^{(t)} = \left\{X^{(t)}, Y^{(t)}\right\}$ ($X^{(t)}$ denotes the input image and $Y^{(t)}$ denotes the category label) and generalize well on new domains within the same task t.

As shown in Figure 1 (a), incorporated with task-specific network $g_T(Y^{(t)}, \varphi^{(t)})$, the text feature extractor $h_T(\cdot)$ parameterized with θ_T is utilized to extract texts features, the image feature extractor $h_I(\cdot)$ parameterized with θ_I is used to extract image features. As discussed in the Introduction, the parameters $\theta = (\theta_I, \theta_T)$ are shared across all tasks. Given the downstream classification task t, with both $h_T(\cdot)$ and $h_I(\cdot)$ fixed, the task-specific parameter $\varphi^{(t)}$ is optimized to achieve few-shot OoD generalization.

In contrast with previous works that fine-tune text representations by conditional prompt learning, our method does not require additionally interacting with image features. As shown in Figure 1 (b), the conditional prompt learning paradigm generates input-specific or domain-specific prompts by learning from both the label name "dog" and the corresponding image feature. However, this paradigm would suffer a significant performance drop when conditional information cannot be efficiently extracted (especially for data dominated by complex correlation shifts).

In this paper, instead of generating conditional prompts, we fine-tune text representations based on the domain-invariant image information to achieve better alignment. To further avoid overfitting on the pre-defined class observed during training, we model $\varphi^{(t)}$ by a variational distribution $q(\varphi^{(t)})$ to approximate its posterior distribution $p(\varphi^{(t)}|X^{(t)}, Y^{(t)}, \theta)$. Given the learned invariant image information, Bayes-CAL estimates the distribution of $\varphi^{(t)}$ across domains, which incorporates richer semantic information compared with the determined values adopted in previous works. Hence, the probability distribution of a sample $(x_n^{(t)}, y_n^{(t)})$ can be represented as:

$$\begin{aligned} p(y_n^{(t)}|x_n^{(t)}, \theta) \propto p(y_n^{(t)}|x_n^{(t)}, \theta, g_T(y_n^{(t)}, q(\varphi^{(t)}))) \\ p(g_T(y_n^{(t)}, q(\varphi^{(t)}))|\theta, D^{(t)}) \end{aligned} \tag{1}$$

Bayesian Cross-Modal Alignment Learning

In the following contents, we omit t from the corresponding mathematical expressions for convenience.

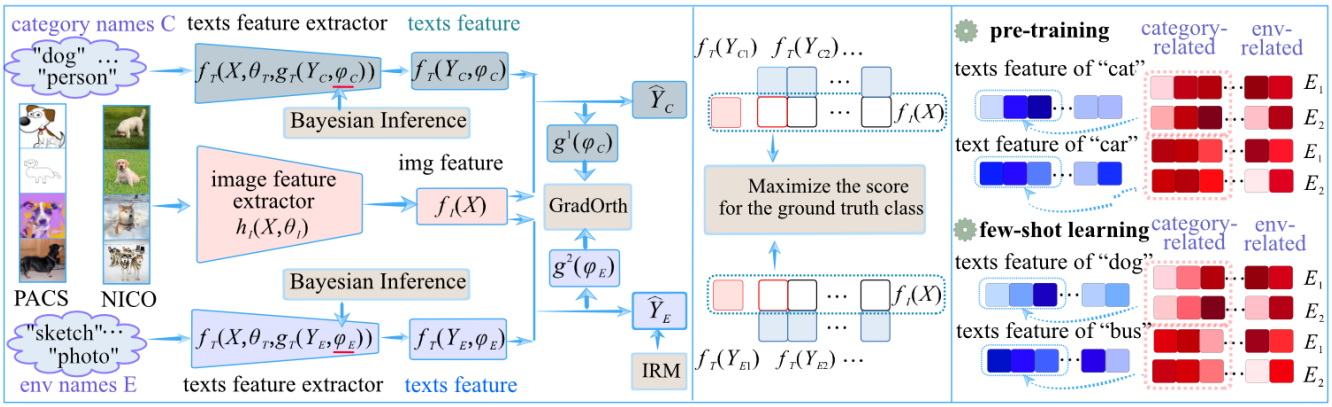
As illustrated in Figure 2 (a), to disentangle the causal and non-casual parts from the image feature $\boldsymbol{f_I}$, the category-related text branch $h_T(X, \theta_T, g_T(Y, \varphi_C))$ and the environment-related text branch $h_T(X, \theta_T, g_T(Y, \varphi_E))$ are introduced to learn category-related information and environment-related information, respectively. Then the similarity between the texts feature (category-related texts feature $f_T(Y_C, \varphi_C)$ or environment-related texts feature $f_T(Y_E, \varphi_E)$) and the image feature $\boldsymbol{f_I}$ are measured to calculate the final output of category prediction probability $\hat{\boldsymbol{Y}}_{\boldsymbol{C}}$ and environment prediction probability $\hat{\boldsymbol{Y}}_{\boldsymbol{E}}$.

Figure 2 (b) illustrates a showcase of the image-text alignment process. In this paper, we compute the cosine similarity for each image-text alignment and then input them into classical cross-entropy loss. And thus, the cross-modal alignment is achieved by maximizing the cosine similarity of the image feature and the ground-true texts feature.

Figure 2 (c) shows the model's working mechanism in the few-shot OoD generalization learning. In the pre-training process, alignment learning usually makes visually similar classes semantically embedded more closely (see Appendix C). The image feature extracted by h_I may contains category-related information $\boldsymbol{f}_{\boldsymbol{I_C}}$ as well as environment-related information $\boldsymbol{f}_{\boldsymbol{I_E}}$. The $\boldsymbol{f}_{\boldsymbol{I_C}}$ is invariant across domains, but $\boldsymbol{f}_{\boldsymbol{I_E}}$ may be various, resulting in previous models hard to achieve efficient image-text alignment.

Gradient Orthogonal Loss To disentangle the category-related and context-related information from the image feature, we introduce an orthogonalization regularization, i.e., the two gradients (w.r.t. the image feature $\boldsymbol{f_I}$) of losses for predicting category and environment labels should be orthogonal.

As shown in Figure 2 (b), the cosine similarities between the image features $\boldsymbol{f_I}$ and the category-related texts features $\boldsymbol{f}_{\boldsymbol{T_C}}$ generated by $h_T(X, \theta_T, g_T(Y_C, \varphi_C))$ are calculated to obtain the category prediction probability $\hat{\boldsymbol{Y}}_{\boldsymbol{C}}$. We perform similar operations to get the environment prediction probability $\hat{\boldsymbol{Y}}_{\boldsymbol{E}}$. And then, the image feature $\boldsymbol{f_I}$ can be disentangled into the category-related and environment-related parts by orthogonalizing the two gradients of the cross-entropy losses $\ell(\hat{\boldsymbol{Y}}_{\boldsymbol{C}}, Y_C)$ and $\ell(\hat{\boldsymbol{Y}}_{\boldsymbol{E}}, Y_E)$ with respect to $\boldsymbol{f_I}$. Therefore, the direction that changes the category loss most quickly will not change the environment loss from $\boldsymbol{f_I}$ and vice versa. In other words, the directions in which the gradients of the two losses change fastest are not on the



(a) Pipeline of the proposed Bayes-CAL. (b) Image-text alignment. (c) Few-shot OoD generalization learning in downstream tasks.

Figure 2: (a) Pipeline of the proposed Bayes-CAL. The top and the bottom branches are used to generate category-related texts feature and environment-related texts feature, respectively. (b) The image-text alignment process. The cross-modal alignment is achieved by maximizing the cosine similarities between the image feature $f_I(X)$ and its corresponding texts feature $f_T(Y_{gt})$ (where Y_{gt} is the ground-true class name of X). (c) Few-shot OoD generalization learning in downstream tasks. The embeddings of new classes "dog" and "bus" are very close to pre-training classes "cat" and "car" in the natural language models, respectively. While in the few-shot OoD setting, the image feature of the same category extracted by different environments may be various. This results in previous models hard to achieve image-text alignment since conditional information (especially for data dominated by complex correlation shift) can not be efficiently extracted from the image features. Benefiting from the image-text alignment incorporating the proposed loss, regardless of various environmental information, texts features can be aligned with the disentangled image features by a few tuning steps of texts representations. Note that the element positions of the category-related and environment-related features can be random.

same hyperplane, which enforces the two branches to pay attention to different parts of the image feature $\boldsymbol{f_I}$ if it contains both category and environment information. Specifically, let $\mathcal{G}^1(q(\varphi_C)) = \nabla_{\boldsymbol{f_I}}\ell(\hat{\boldsymbol{Y}}_{\boldsymbol{C}}, Y_C)$ and $\mathcal{G}^2(\varphi_E) = \nabla_{\boldsymbol{f_I}}\ell(\hat{\boldsymbol{Y}}_{\boldsymbol{E}}, Y_E)$ be the gradients of the category prediction loss and the environment prediction loss with respect to $\boldsymbol{f_I}$, respectively. Thus, the following gradient orthogonal regularization term is introduced:

$$\mathcal{L}_{\text{orth}}(\varphi_C, \varphi_E) = \left(\frac{\mathcal{G}^1(\varphi_C)}{\|\mathcal{G}^1(\varphi_C)\|} \cdot \frac{\mathcal{G}^2(\varphi_E)}{\|\mathcal{G}^2(\varphi_E)\|}\right)^2 \quad (2)$$

Invariant Risk Minimization Invariant risk minimization is proposed to achieve invariant predictions across different environments (Ahuja et al. 2020). It aims to learn a stable data representation $\varphi : \mathcal{X} \to \mathcal{H}$ and an invariant predictor $\omega : \mathcal{H} \to \mathcal{Y}$ across for all training distributions $\mathcal{E}_{\text{tr}}$. Mathematically, it can be written as:

$$\begin{aligned} &\min_{\substack{\varphi:\mathcal{X}\to\mathcal{H} \\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\text{tr}}} \ell^e(w\circ\varphi) \\ &\text{subject to} \quad w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} \ell^e(\bar{w}\circ\varphi), \text{ for all } e \in \mathcal{E}_{\text{tr}} \end{aligned} \quad (3)$$

where ℓ^e is the ERM risk of the training environment e. Since each constraint calls an inner optimization routine in Equation (3), the author proposes to instantiate the challenging bi-leveled optimization problem by adding a gradient norm penalty $\left\|\nabla_{w|w=1.0}\ell^e(w\cdot\varphi)\right\|^2$ as regularization term. So, the optimization problem is changed into the following form:

$$\min_{\varphi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\text{tr}}} \ell^e(\varphi) + \lambda \cdot \left\|\nabla_{w|w=1.0}\ell^e(w\cdot\varphi)\right\|^2 \quad (4)$$

In Bayes-CAL, we utilize IRM to train the task-specific network $g_T(Y_C, \varphi_C)$ suitable for all training distributions to learn domain-invariant category-related text representations. Given the category prediction probability $\hat{\boldsymbol{Y}}_{\boldsymbol{C}}$, the IRM regularization term is defined as:

$$\mathcal{L}_{\text{IRM}}(\varphi_C) = \sum_{e\in\mathcal{E}_{\text{tr}}} \|\nabla_{w|w=1}\ell_e(w * \hat{\boldsymbol{Y}}_{\boldsymbol{C}}, Y_C)\|^2 \quad (5)$$

Bayesian Learning To further enhance the generalization ability of the proposed method, Bayesian inference is utilized to overcome overfitting on the pre-defined clasess observed during training. It is known that variational inference is appealing when dealing with overfitting in few-shot learning since applying Monte Carlo sampling to the Bayesian neural network can be computationally infeasible. From Bayesian perspective, these unknown parameters φ_C and φ_E can be viewed as latent variables that follow some prior distribution $p(\varphi_C)$ and $p(\varphi_E)$, respectively. In order to infer φ_C and φ_E, training samples containing the information about the unknown parameters are utilized to approximate the posterior distribution $p(\varphi_C, \varphi_E|X, Y_C, Y_E)$.

In this paper, under the mean-field assumption partitioning the variables into independent parts, we have,

$$p(\varphi_C, \varphi_E) = \prod_{i=1}^{K} \prod_{j=1}^{K} p_i\left(\varphi_{C_i}\right) p_j\left(\varphi_{E_j}\right) \tag{6}$$

where K is the total number of learnable parameters φ_C or φ_E. And we assume each φ_{C_i} follows a Gaussian distribution $\mathcal{N}(\mu_{i1}, \sigma_{i1})$ with the posterior distribution of $\mathcal{N}(\mu_{i2}, \sigma_{i2})$. For each φ_{E_i}, we have the same assumption.

Variational inference seeks for a variational distribution $q(\varphi_C, \varphi_E)$ to approximate $p(\varphi_C, \varphi_E|X, Y_C, Y_E)$ by minimizing the Kullback-Leibler divergence $\mathbb{D}_{\mathrm{KL}}[q(\varphi_C, \varphi_E)\|p(\varphi_C, \varphi_E|X, Y_C, Y_E)]$ between them. This can be equivalent to minimizing the negative value of the evidence lower bound (ELBO):

$$\begin{aligned}\mathcal{L}_{\text{Bayes}} = -\mathbb{E}_{q(\varphi_C, \varphi_E)}\{[\log p(Y_C, Y_E \mid X, \varphi_C, \varphi_E)] \\ +\mathbb{D}_{\mathrm{KL}}[q(\varphi_C, \varphi_E)\|p(\varphi_C, \varphi_E \mid X, Y_C, Y_E)]\}\end{aligned} \tag{7}$$

The first term $\mathcal{L}_1$ in Equation (7) is the expectation of negative log-likelihood. According to Monte Carlo sampling and adding the regularization terms to the negative log-likelihood, we approximate $\mathcal{L}_1$ by:

$$\begin{aligned}\mathcal{L}_1 \approx \frac{1}{N}\sum_{i}^{N}(\ell\left(\hat{\boldsymbol{Y}}_{\boldsymbol{C}}, Y_C\right) + \lambda_1 \ell\left(\hat{\boldsymbol{Y}}_{\boldsymbol{C}}, Y_E\right) \\ + \lambda_2 \mathcal{L}_{\text{IRM}}(\varphi_C) + \lambda_3 \mathcal{L}_{\text{orth}}\left(\varphi_C, \varphi_E\right))\end{aligned} \tag{8}$$

According to the Mean-Field assumption, the second term $\mathcal{L}_2$ in Equation (7) can be calculated as:

$$\mathcal{L}_2 = \mathbb{D}_{\mathrm{KL}}(\varphi_C) + \mathbb{D}_{\mathrm{KL}}(\varphi_E) \tag{9}$$

where based on the Gaussian distribution assumption, $\mathbb{D}_{\mathrm{KL}}(\theta)$ is simplified as:

$$\mathbb{D}_{\mathrm{KL}}(\varphi_C) = \sum_{i=1}^{K} \log \frac{\sigma_{i1}}{\sigma_{i2}} + \frac{1}{2}\left(\sigma_{i2}^2 + \left(\mu_{i2} - \mu_{i1}\right)^2\right)/\sigma_{i1}^2 \tag{10}$$

and $\mathbb{D}_{\mathrm{KL}}(\varphi_E)$ can be computed similarly.

By integrating Equation (8) and Equation (9) into Equation (7), we obtain the final learning objective as follows:

$$\begin{aligned}\mathcal{L}_{\text{Bayes}} = \frac{1}{N}\sum_{i}^{N}(\ell\left(\hat{Y}_C, Y_C\right) + \lambda_1 \ell\left(\hat{Y}_E, Y_E\right) \\ + \lambda_2 \mathcal{L}_{\text{IRM}}(\varphi_E) + \lambda_3 \mathcal{L}_{\text{orth}}\left(\varphi_C, \varphi_E\right)) \\ + \mathbb{D}_{\mathrm{KL}}(\varphi_C) + \mathbb{D}_{\mathrm{KL}}(\varphi_E)\end{aligned} \tag{11}$$

Experiment Results

We evaluate Bayes-CAL in the following three settings: 1) First, without loss of generality, we instantiate Bayes-CAL by the prompt learning method of CLIP as a showcase, and compare it with OoD generalization algorithms in OoD-Bench (Ye et al. 2021) and several powerful CLIP-like models. 2) Then, we do a series of ablation studies and evaluate its base-to-new generalization ability. 3) Furthermore, to validate its working mechanism for few-shot OoD generalization, we instantiate the text branches of Bayes-CAL with other methods. Despite the strong Transformer-based pre-trained models of the two text branches, we substitute them directly with learnable vectors (LV for short) or word embeddings from Word2Vector (W2V for short). We conduct fair comparisons with the conventional visual deep network on convergence speed and OoD generalization performances.

Datasets We evaluate Bayes-CAL on datasets that cover both diversity shift and correlation shift: datasets dominated by correlation shift (NICO (He, Shen, and Cui 2021) and ColoredCatsDogs), and datasets dominated by diversity shift (PACS (Li et al. 2017) and VLCS (Torralba and Efros 2011)). ColoredCatsDogs (CCD for short) has spurious correlations with the background color (green or red), constructed in a similar principle as ColoredMNIST (Arjovsky et al. 2019) but with images of cats and dogs disturbed by Gaussian noise to increase complexity.

Following CLIP, we train models on a few-shot training data and evaluate the original test set. Due to the difficulty of the task and the large randomness of OoD data, for each category, we randomly sample a 16-shot training set and a 16-shot validation set (an 8-shot training set and a 64-shot validation set for NICO due to the validation accuracy up to 100% if the validation set is too small) from each domain. Note that the 16-shot training set is equally sampled from every domain. For all experiments, we set the max epoch as 30 unless otherwise specified and conduct three independent experiments with random seeds (1, 2, 3) to exclude the effects of randomness.

Experiment Protocol Implementing experiments based on CoOp's code, we first give Bayes-CAL's hyper-parameter setting introduced from CoOp. Throughout the experiments, the ResNet-50 model (He et al. 2016) is used as the vision backbone. The number of context tokens is set as 16, the class token position (CTP) is a hyper-parameter that can be set as "end" or "middle", and the class-specific context (CSC) can be set as "True" or "False". The three additional hyper-parameters introduced by our paradigm are λ_1, λ_2, and λ_3, corresponding to the coefficients of the three regularization terms. For fair comparisons, a similar model evaluation protocol as OoD-Bench is used—a 20-times random search for each of 3 pairs of weight initialization and training-validation data. Finally, we report the mean and standard error values on the original test set.

Competitors We compare Bayes-CAL with the **top three** algorithms for each dataset in OoD-Bench on the four datasets. Since the proposed method is instantiated by the large pre-trained model of CLIP, we mainly evaluate the OoD generalization performance of CLIP-like alignment learning methods based on text representation fine-tuning, i.e., CLIP, CoOp, CoCoOp, and DPLCLIP. The performances of the three CLIP-based methods are evaluated under the same experimental setting. Note that the OoD-Bench results on CCD and VLCS are reproduced by ourselves.

Algorithm	NICO	Algorithm	CCD
Bayes-CAL	**98.33(0.6)**	**Bayes-CAL**	**69.00(0.8)**
CLIP	96.75(0.0)	CLIP	65.00(0.0)
CoOp	95.92(0.7)	CoOp	55.21(10.6)
CoCoOp	98.00(0.8)	CoCoOp	56.25(6.8)
DPLCLIP	97.42(1.4)	DPLCLIP	57.29(10.3)
ANDMask	72.20(1.2)	IRM	51.72(0.5)
GroupDRO	71.83(0.8)	ERDG	51.71(2.0)
ERM	71.44(1.3)	SagNet	44.41(0.0)

Table 1: Performances on NICO and CCD.

Algorithm	PACS	Algorithm	VLCS
Bayes-CAL	**91.82(0.9)**	**Bayes-CAL**	**78.06(1.5)**
CLIP	90.76(0.0)	CLIP	75.07(0.0)
CoOp	91.47(0.6)	CoOp	72.79(5.4)
CoCoOp	91.81(0.6)	CoCoOp	76.25(1.0)
DPLCLIP	89.58(0.7)	DPLCLIP	74.25(6.1)
RSC	82.82(0.4)	RSC	77.14(0.5)
VREx	81.78(0.1)	ANDMask	77.13(0.4)
MMD	81.72(0.2)	MLDG	76.25(0.6)

Table 2: Performances on PACS and VLCS.

OoD Generalization on 2D Distribution Shifts

Experiment on the Correlation Shift Datasets NICO and CCD are typical datasets with correlation shift. Following the same OoD validation for NICO and test-domain validation for CCD as in OoD-Bench. As shown in Table 1, Bayes-ACL obtains far higher accuracy compared with the methods in OoD-Bench, with more than 20% improvements for NICO and more than 15% improvements for CCD with statistical significance. This demonstrates the superiority of Bayes-CAL incorporating the large-scale pre-trained vision-language model. Moreover, Bayes-CAL outperforms CLIP-like models and obtains far better results on CCD with strong spurious correlations. As shown in Figure 3 (a), the test accuracy on CCD of Bayes-CAL is stable as the training epoch increases. The result of CoOp, however, shows an obvious trend of overfitting on training data, since its test accuracy gradually decreases during the training process.

Experiment on the Diversity Shift Datasets PACS and VLCS are two common domain generalization benchmark datasets with diversity shift as shown by OoD-Bench. We follow the same training-domain validation experimental protocol as in OoD-Bench. The mean accuracy of the four tests on different domains is shown in Table 2. It is shown that our method outperforms all models in OoD-Bench on PACS and VLCS, especially with an improvement of around 10% on PACS. Bayes-CAL gets significant improvements on VLCS by more than 5%, especially with CoOp. This further validates the effectiveness of Bayes-CAL in handling both diversity shift and correlation shift.

Ablation Studies

The Effectiveness of the Proposed Components For ablation studies, we remove each regularization term from our paradigm respectively and keep other components unchanged to check each component's role for OoD generalization. The results are shown in Table 3. We can see that Bayes-CAL obtains the best result on the average accuracy of the four datasets, demonstrating synergistic benefits between the regularization terms to learn invariant alignment. Especially on CCD and VLCS, after removing the $\mathcal{L}_{\text{IRM}}$ and $\mathcal{L}_{\text{orth}}$, the performance will drop significantly, which means better alignment can be achieved by disentangling the image features. Note that the performance of Bayes-CAL on NICO and PACS is not as significant as performance on the CCD and VLCS, we have done Wilcoxon test to demonstrate the statistical significance of our results (see Appendix F).

Data	Removed Component			Bayes-CAL
	$\mathcal{L}_{\text{E}}$	$\mathcal{L}_{\text{IRM}}$	$\mathcal{L}_{\text{orth}}$	
NICO	96.58	98.00	98.25	98.33
CCD	68.67	61.33	64.33	69.00
PACS	91.56	91.77	91.20	91.82
VLCS	75.43	73.05	75.90	78.06
Average	83.06	81.04	82.42	**84.30**

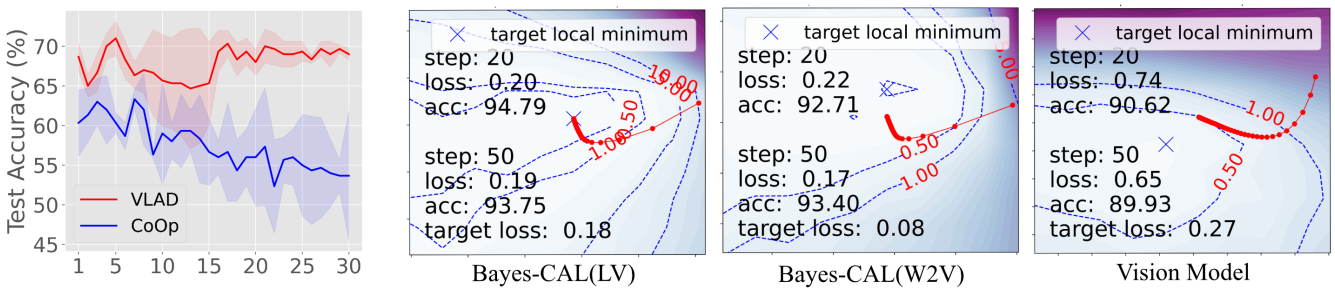
Table 3: Ablation study results.

Data	Algorithm	OoD Results		I.I.D Results	
		Acc	Acc*	Acc	Acc*
NICO	CoOp	72.10	88.30	90.55	96.41
	CoCoOp	75.15	88.41	**93.64**	**96.93**
	DPLCLIP	46.06	77.21	66.83	86.02
	CAL	72.12	89.05	83.29	95.13
	Bayes-CAL	**76.36**	**93.26**	89.45	96.27
PACS	CoOp	55.18	87.78	72.39	89.83
	CoCoOp	58.16	83.87	59.51	75.97
	DPLCLIP	35.96	53.67	47.98	70.06
	CAL	57.30	**90.73**	71.50	91.19
	Bayes-CAL	**60.40**	89.85	**72.97**	**92.54**

Table 4: Base-to-new generalization performances. (Acc denotes test accuracy, and Acc* denotes the proportion of correctly identified samples with high prediction confidence. The higher the ACC*, the higher the security of the method.)

Base-to-New Generalization Performance

To evaluate the effectiveness of the Bayesian method, we compare Bayes-CAL, CAL (the model that removes the Bayesian method from Bayes-CAL), CoOp, CoCoOp, and DPLCLIP in the base-to-new generalization setting on NICO and PACS. In this paper, we randomly split each dataset into the base and new sets. Following CoCoOp, the prompts are learned from the base classes (16 shots, CTP as "end", CSC as "False", and 4 context tokens), and the learnable contexts are initialized as "a photo of a". Under a 20-times random search for each of the 3 different data trials for additional hyper-parameters of Bayes-CAL and CAL, the base-to-new generalization performances are evaluated on in-distribution (I.I.D) new classes and out-of-distribution (OoD) new classes. We also utilize the thresholding-based



(a) Test accuracy on CCD at different epochs.

(b) Visualization of the optimization landscape.

Figure 3: (a) Test accuracy on CCD at different epochs. The test accuracy of Bayes-CAL is more stable than CoOp as the training epoch increases. (b) Visualization of the optimization landscape. The x-axis and y-axis represent the first and the second principal components of the parameters, respectively. The red lines are the corresponding optimization trajectories. Without the pre-trained text encoder, Bayes-CAL(LV) and Bayes-CAL(W2V) still see much faster convergence and lower target loss compared to the conventional visual model.

Indicator	CoOp	PL	LV	W2V	Vision
Final Loss	0.24	0.28	**0.18**	0.21	0.74
Target Loss	0.23	0.26	0.15	**0.08**	0.29
Test Acc	95.58	**96.08**	91.42	89.33	92.92

Table 5: Results of few-shot learning on NICO's subclasses. Bayes-CAL instantiated by prompt learning, learnable vectors, and Word2Vector is abbreviated as PL, LV, and W2V, respectively. Final Loss is the training loss at epoch 50.

method to detect examples with low prediction confidence, where the threshold is selected by 95% correctly classified validation examples are detected into examples with high prediction confidence. We remove predictions that were misclassified with low confidence, i.e., those with probabilities lower than the selected threshold, and then recalculate the test accuracy (denoted as Acc*). The results can be seen in Table 4. It is shown that Bayes-CAL significantly outperforms CAL on the test accuracy of new classes, especially for NICO. The results strongly justify the generalizability of the Bayesian method. Moreover, Bayes-CAL yields more stable I.I.D and OoD base-to-new generalization results than CoCoOp, which is dedicated to improving generalization to unseen classes. More details are in Appendix D.

Insight into the Superiority of Alignment Learning

We have illustrated the working mechanism of alignment learning on few-shot OoD data in Figure 2 (c). In this section, we elaborate on it by analyzing its convergence speed to gain insight into the superiority of alignment learning.

We still incorporate the image encoder of CLIP into the image feature extractor, while substituting the text branches with two methods: directly utilizing learnable vectors (LV) and Word2Vector with a two-layer multi-layer perception (W2V). We compare them with the Bayes-CAL instantiated by prompt learning (Bayes-CAL(PL)) and the conventional visual paradigm. Based on the same pre-trained model of CLIP's image encoder, the task-specific layers in the conventional visual model are instantiated by a three-layer multi-layer perception.

The subclasses in NICO are mostly unseen in the pre-training data of CLIP. Setting the hyper-parameters of Bayes-CAL as (0.1, 0, 0.1) and the max epoch as 50, we conduct a 19-way 16-shot training on NICO with 19 subclasses and 4 environments from training domains. We report the corresponding training performance (final training loss at epoch 50) and the test accuracy in Table 5. It is shown that: 1) Instantiated by LV and W2V, Bayes-CAL achieves lower final training loss. It indicates that the two instances of Bayes-CAL can still achieve fast convergence of the alignment learning without the pre-trained text encoder. 2) A wider minimum (lower target loss) can be achieved after reshaping the semantic space by LV and W2V, which also demonstrates the generality of our framework.

Visualization of the Optimization Landscape We show the optimization trajectories and loss landscapes for Bayes-CAL(LV) and the conventional visual model in Figure 3. It is observed that without the pre-trained text encoder, the instances of Bayes-CAL are close to the local minimum at epoch 50, while the conventional visual model is far from the target local minimum. Note that, Bayes-CAL(LV) has much fewer parameters (0.04 million) than the conventional model (0.28 million) to optimize. So we can say the fast learning ability is essentially caused by alignment learning.

Conclusion and Discussion

In this paper, we have proposed the Bayes-CAL method with invariant risk minimization and gradient orthogonalization loss under the Bayesian framework to tackle the few-shot OoD generalization dilemma. Numerical results not only show Bayes-CAL achieves robust OoD generalization performances under two-dimensional distribution shifts, but also reach more stable generalization performance on unseen classes under the Bayesian framework. To the best of our knowledge, it is the first work that investigates few-shot OoD generalization by Bayesian alignment learning. This may serve as a foothold for future research in foundation models for few-shot OoD generalization.

Acknowledgements

Nanyang Ye was supported in part by National Natural Science Foundation of China under Grant No.62106139, in part by National Key R&D Program of China 2022YFB3904204, in part by National Natural Science Foundation of China under Grant (No.42050105, 61960206002, 62061146002, 62020106005).

References

Ahuja, K.; Shanmugam, K.; Varshney, K.; and Dhurandhar, A. 2020. Invariant risk minimization games. In *International Conference on Machine Learning*, 145–155. PMLR.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2229–2238.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2019. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*.

Dai, Z.; Liu, H.; Le, Q. V.; and Tan, M. 2021. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34: 3965–3977.

Ding, N.; Chen, Y.; Han, X.; Xu, G.; Xie, P.; Zheng, H.-T.; Liu, Z.; Li, J.; and Kim, H.-G. 2021. Prompt-learning for fine-grained entity typing. *arXiv preprint arXiv:2108.10604*.

Du, Y.; Liu, Z.; Li, J.; and Zhao, W. X. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

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

Fang, Z.; Zhu, X.; Yang, C.; Han, Z.; Qin, J.; and Yin, X.-C. 2022. Learning Aligned Cross-Modal Representation for Generalized Zero-Shot Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6605–6613.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Gu, Y.; Han, X.; Liu, Z.; and Huang, M. 2021. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*.

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

Guo, Y.; Codella, N. C.; Karlinsky, L.; Codella, J. V.; Smith, J. R.; Saenko, K.; Rosing, T.; and Feris, R. 2020. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, 124–141. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Y.; Shen, Z.; and Cui, P. 2021. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110: 107383.

Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, 124–140. Springer.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

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

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137. Springer.

Liang, H.; Zhang, Q.; Dai, P.; and Lu, J. 2021. Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9424–9434.

Lin, Y.; Dong, H.; Wang, H.; and Zhang, T. 2022. Bayesian Invariant Risk Minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16021–16030.

Liu, Y.; Lee, J.; Zhu, L.; Chen, L.; Shi, H.; and Yang, Y. 2021. A multi-mode modulator for multi-domain few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8453–8462.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Rahman, S.; Khan, S.; and Barnes, N. 2020. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11932–11939.

Rusu, A. A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; and Hadsell, R. 2018. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*.

Shen, Z.; Cui, P.; Zhang, T.; and Kunag, K. 2020. Stable learning via sample reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5692–5699.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1199–1208.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528. IEEE.

Tseng, H.-Y.; Lee, H.-Y.; Huang, J.-B.; and Yang, M.-H. 2020. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv preprint arXiv:2001.08735*.

Vuorio, R.; Sun, S.-H.; Hu, H.; and Lim, J. J. 2019. Multimodal model-agnostic meta-learning via task-aware modulation. *Advances in Neural Information Processing Systems*, 32.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Yu, P. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, Y.; Yao, Q.; Kwok, J. T.; and Ni, L. M. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3): 1–34.

Wehrmann, J.; Kolling, C.; and Barros, R. C. 2020. Adaptive cross-modal embeddings for image-text alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12313–12320.

Ye, N.; Li, K.; Hong, L.; Bai, H.; Chen, Y.; Zhou, F.; and Li, Z. 2021. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021a. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.

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

Zhang, X.; Iwasawa, Y.; Matsuo, Y.; and Gu, S. S. 2021c. Amortized Prompt: Guide CLIP to Domain Transfer Learning. *arXiv preprint arXiv:2111.12853*.

Zheng, Y.; Huang, R.; Han, C.; Huang, X.; and Cui, L. 2020. Background learnable cascade for zero-shot object detection. In *Proceedings of the Asian Conference on Computer Vision*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, Y.; and Tan, C. 2021. Investigating the Effect of Natural Language Explanations on Out-of-Distribution Generalization in Few-shot NLI. *arXiv preprint arXiv:2110.06223*.

Zhu, C.; Chen, F.; Ahmed, U.; Shen, Z.; and Savvides, M. 2021. Semantic relation reasoning for shot-stable few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8782–8791.