# BDC-Adapter: Brownian Distance Covariance for Better Vision-Language Reasoning

Yi Zhang*[1,2]
zhangyi2021@mail.sustech.edu.cn
Ce Zhang*[3]
cezhang@cs.cmu.edu
Zihan Liao[2]
liaozh2020@mail.sustech.edu.cn
Yushun Tang[2]
tangys2022@mail.sustech.edu.cn
Zhihai He†[2,4]
hezh@sustech.edu.cn

[1] Harbin Institute of Technology
Harbin, China
[2] Southern University of Science and Technology (SUSTech)
Shenzhen, China
[3] Carnegie Mellon University
Pittsburgh, United States
[4] Pengcheng Laboratory
Shenzhen, China

**Abstract**

Large-scale pre-trained Vision-Language Models (VLMs), such as CLIP and ALIGN, have introduced a new paradigm for learning transferable visual representations. Recently, there has been a surge of interest among researchers in developing lightweight fine-tuning techniques to adapt these models to downstream visual tasks. We recognize that current state-of-the-art fine-tuning methods, such as Tip-Adapter, simply consider the covariance between the query image feature and features of support few-shot training samples, which only captures linear relations and potentially instigates a deceptive perception of independence. To address this issue, in this work, we innovatively introduce Brownian Distance Covariance (BDC) to the field of vision-language reasoning. The BDC metric can model all possible relations, providing a robust metric for measuring feature dependence. Based on this, we present a novel method called BDC-Adapter, which integrates BDC prototype similarity reasoning and multi-modal reasoning network prediction to perform classification tasks. Our extensive experimental results show that the proposed BDC-Adapter can freely handle non-linear relations and fully characterize independence, outperforming the current state-of-the-art methods by large margins.

## 1 Introduction

Recently, large-scale pre-trained Vision-Language Models (VLMs), such as CLIP [39] and ALIGN [25], have introduced a new paradigm for generic visual recognition [24]. These VLMs jointly learn both visual and textual representations in a shared feature space through pre-training on large-scale datasets retrieved from the Internet, enabling them to recognize a wide range of visual concepts without the need for additional annotated data [17, 39].

*Equal contribution. †Corresponding author.

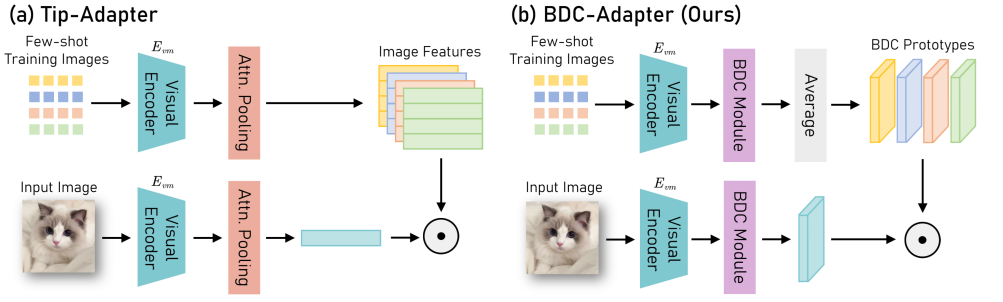


Figure 1: **A comparison on Tip-Adapter (left) vs. our proposed BDC-Adapter (right).** Our BDC-Adapter represents each image by a BDC matrix, which considers the joint distributions and measures non-linear dependence during inference. Note that in this figure, $E_{vm}$ is the modified image encoder of CLIP without the last attention pooling layer.

However, due to the massive size of VLMs, it is impractical for individuals to re-train those models. Therefore, lightweight fine-tuning techniques have become essential for adapting VLMs to downstream visual tasks, such as image classification [39, 68], object detection [13, 41], and image captioning [30, 52, 59]. One research direction focuses on the **prompt tuning method**, which aims to learn the prompt from downstream data. For instance, CoOp [68] firstly introduces the prompt tuning method to fine-tune CLIP, while CoCoOp [67] uses prompts conditioned on model inputs to address the generalization problem. Another approach is the **adapter-based method**, which directly adapts the extracted features. CLIP-Adapter [18] and Tip-Adapter [63] are examples of this approach, both of which introduce feature adapters to enhance CLIP's performance on various downstream tasks.

We notice that the current state-of-the-art Tip-Adapter [63] method, as shown in Figure 1, establishes a key-value cache model and evaluates the similarities of the query image feature and features of support few-shot training samples to perform classification. However, we recognize that Tip-Adapter [63] simply considers the covariance between each image feature pair, which only measures marginal distributions and captures linear relations. If the relation between features is non-linear [46, 47], the covariance might be zero, potentially instigating a deceptive perception of independence. This problem, if not effectively addressed, will hinder our capabilities to fine-tune VLMs.

In this paper, we introduce Brownian Distance Covariance (BDC) to the field of vision-language reasoning to provide a robust metric for measuring feature dependence. While classical covariance can only capture linear relations, Brownian covariance can model all possible relations [46, 47]. Based on this, we propose a novel approach called BDC-Adapter that leverages BDC to enhance vision-language reasoning ability. During the training stage, we first train a one-layer multi-modal reasoning network that learns from few-shot examples across different modalities (*i.e.*, vision and language). Then, we introduce a BDC module that takes feature maps as input and outputs a BDC matrix as a visual representation. Using this, we compute class-specific prototypes by averaging the BDC matrices of the few-shot image samples for each class, which act as a support set for test image classification. In Figure 1, we show the BDC prototype similarity reasoning process of our proposed BDC-Adapter. During the inference stage, we combine the BDC prototype similarity reasoning and multi-modal reasoning network prediction to perform classification tasks. To evaluate the effectiveness of our BDC-Adapter, we conduct experiments on few-shot learning, domain generalization, and visual reasoning tasks. Our extensive experimental results show that BDC-Adapter outperforms the current state-of-the-art methods by large margins.

## 2 Related Work

**Fine-Tuning Vision-Language Models.** In recent studies on VLMs, researchers have explored the semantic correspondence between the textual and visual modalities by leveraging a huge amount of image-text pairs [7, 10, 19, 25, 39, 43, 53, 55, 61, 62]. Researchers have demonstrated that with sufficient fine-tuning, the large-scale pre-trained VLMs can be transferred to various downstream tasks, such as image retrieval [14, 32], visual grounding [30, 59], semantic segmentation [33], and visual question answering [14, 29, 69].

Recent advances in fine-tuning VLMs can be classified into two major categories: *prompt tuning methods* and *adapter-based methods*. As the pioneering work in the context of prompt tuning, CoOp [68] learns a set of additional vectors to optimize the prompt context. Further, Zhou *et al*. [67] extend CoOp to generate image-conditioned vectors to tackle the generalization problem. TPT [42] can adaptively learn prompts for each test sample in the inference stage. Adapter-based methods directly adapt the extracted visual and textual representations. For example, CLIP-Adapter [18] introduces a feature adapter that generates the adapted features to enhance the performance of few-shot recognition. Further, Tip-Adapter [63] proposes a training-free scheme, which achieves higher accuracy by establishing a key-value cache model. UP-Adapter [65] proposes to generate pseudo-labels for the unannotated images, which will be used to train a prototype adapter module.

**Cross-Modal Few-Shot Image Classification.** Few-shot learning is an important problem in machine learning, which attempts to enable models' transferability to new tasks with limited labeled examples [48, 54]. Traditional few-shot learning methods typically rely on training from base classes in the source domain, which limits their generalization capabilities to the novel target domains [1, 16, 38, 60, 71]. With the help of large-scale pre-trained VLMs, an alternative direction of work focus on tackling the few-shot classification task without source-domain training [31, 39]. By freezing the pre-trained weights and training additional sets of learnable parameters for downstream tasks, these models can achieve remarkable performance with very limited training samples [39, 65, 66].

**Brownian Distance Covariance.** The BDC metric, defined as the Euclidean distance between the joint characteristic function and the product of the marginals, was first proposed in Székely *et al*. [46, 47]. While classical covariance can only model linear relations, Brownian covariance can model all possible relations. Therefore, the BDC metric has been introduced into appearance matching [2] and people recognition [3, 4] to provide more complementary information for the network model. In recent years, the BDC metric has also been applied in other computer vision applications, such as object detection [56], hyperspectral image classification [64], and few-shot learning [58] tasks. In this work, we use the BDC metric for representation learning mainly in the few-shot classification setting.

## 3 Method

### 3.1 Background

**Contrastive Language-Image Pre-Training**. CLIP [39] has demonstrated remarkable performance on visual tasks by encoding images and text descriptions onto a shared embedding space and exploiting contrastive learning on noisy image-text pairs on the Internet. We denote CLIP's encoders as $\{E_t, E_v\}$, where $E_t$ is the text encoder (typically a Transformer [49]), $E_v$ is the image encoder (typically a ResNet [20] or ViT [12]). In the zero-shot scenario, given

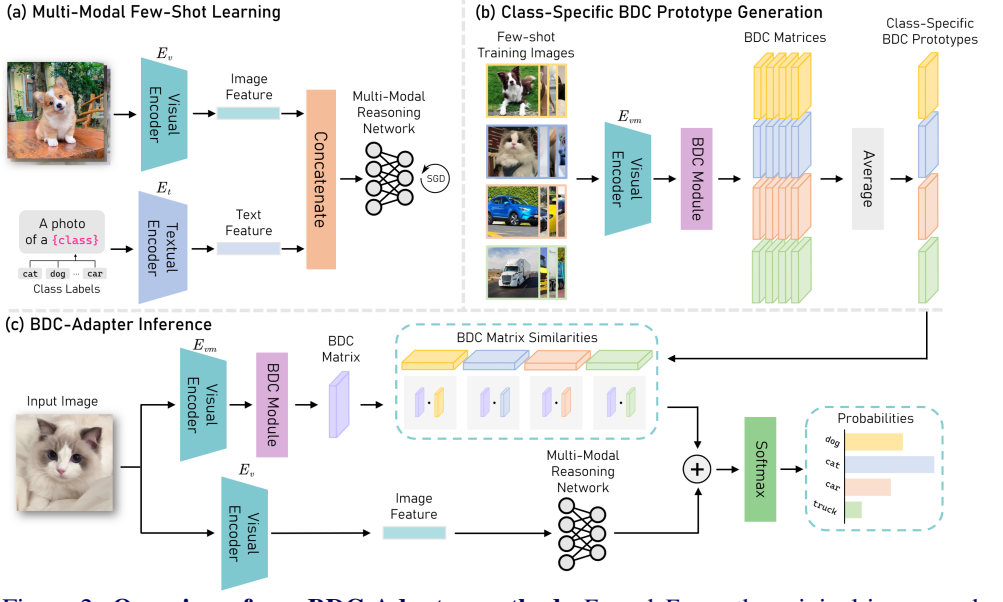


Figure 2: **Overview of our BDC-Adapter method**. $E_v$ and $E_t$ are the original image and text encoders of CLIP respectively, and $E_{vm}$ is the modified image encoder of CLIP that does not include the last attention pooling layer. **(a)** shows the multi-modal few-shot learning process, **(b)** shows the class-specific BDC prototype generation process, and **(c)** presents the whole BDC-Adapter inference process.

a test image $x_{test}$ for a $N$-class classification problem, we utilize CLIP's encoders to extract the visual feature $f_v = E_v(x_{test})$ and $N$ text features $f_{t_i} = E_t(\{\pi; y_i\})$ for all classes, where the class name $y_i$ is appended to a hand-crafted prompt $\pi$, such as "a photo of a". The prediction probability on $x_{test}$ can be computed as

$$p(y = y_i | x_{test}) = \frac{\exp\left(\text{sim}\left(f_{t_i}, f_v\right)/\tau\right)}{\sum_{i'} \exp\left(\text{sim}\left(f_{t'}, f_v\right)/\tau\right)}, \tag{1}$$

where $\tau$ is the temperature hyper-parameter of the softmax function, and $\text{sim}(\cdot,\cdot)$ indicates the cosine similarity.

**Brownian Distance Covariance**. The concept of Brownian Distance Covariance (BDC) was first formalized in the literature by Székely *et al.* [46, 47], with a foundation in characteristic function theory. Suppose $X \in \mathbb{R}^p, Y \in \mathbb{R}^q$ denote random vectors with dimensions $p$ and $q$ respectively, and $p_{XY}(\mathbf{x},\mathbf{y})$ represents their joint probability density function (PDF). With $\mathbf{t}$ standing as the characteristic of the distribution $X$ and $\mathbf{s}$ for that of $Y$, the joint characteristic functions of $(X,Y)$, expressed as $f_{XY}(\mathbf{t},\mathbf{s}) = \mathcal{F}(p_{XY}(\mathbf{x},\mathbf{y}))$, embody a collection of functions that encapsulate the interrelation within the random distributions $X$ and $Y$. Here, $\mathcal{F}$ is a mapping from the distribution space to the feature space. The selection of $f_{XY}(\mathbf{t},\mathbf{s})$ can be diverse, and in our experiments, we employ a network as the characteristic function. In accordance with the definitions of the joint and marginal characteristic functions, and assuming that random vectors $X$ and $Y$ possess finite first moments, the BDC metric, quantifying the similarity between the characteristics of distributions $X$ and $Y$, can be expressed as

$$\mathcal{V}(X,Y) = \int_{\mathbf{t}\in\mathbb{R}^p} \int_{\mathbf{s}\in\mathbb{R}^q} \frac{\|f_{XY}(\mathbf{t},\mathbf{s}) - f_X(\mathbf{t}) f_Y(\mathbf{s})\|^2}{c_p c_q \|\mathbf{t}\|^{1+p} \|\mathbf{s}\|^{1+q}} d\mathbf{s} d\mathbf{t}. \tag{2}$$

Here, $\|\cdot\|$ denotes the Euclidean norm, $c_p = \pi^{(1+p)/2}/\Gamma((1+p)/2)$, and $\Gamma$ represents the complete gamma function. Despite Equation (2) is complicated in its current form, the BDC metric possesses a closed-form expression for discrete observations as established in the work by Székely *et al*. [46], which is elaborated in the following Section 3.2.

## 3.2 BDC Module for Image Representation

Let the pair of observation data matrices $(\mathbf{T},\mathbf{S})$ represent the observation of the joint characteristic function $f_{XY}(\mathbf{t},\mathbf{s})$. Here, the $i$-th observation's $\mathbf{t}_i,\mathbf{s}_i$ constitute the $i$-th column of $\mathbf{T}$ and $\mathbf{S}$ respectively. The Euclidean distance matrix, derived from each observational pair of $X$, is denoted as $\mathbf{D}_1^{\mathbf{T}} = (d_{k,l}^{\mathbf{T}}) \in \mathbb{R}^{m\times m}$, where $d_{k,l}^{\mathbf{T}} = \|\mathbf{t}_k - \mathbf{t}_l\|$.

The Euclidean distance matrix $\mathbf{D}_1^{\mathbf{T}}$ can be calculated by first computing the squared Euclidean distance matrix $\mathbf{D}_2^{\mathbf{T}}$ and subsequently taking the square root. Through this process, we can obtain a closed-form expression of $\mathbf{D}_1^{\mathbf{T}}$ in terms of $\mathbf{T}$:

$$\mathbf{D}_1^{\mathbf{T}} = \sqrt{\mathbf{D}_2^{\mathbf{T}}}, \quad \mathbf{D}_2^{\mathbf{T}} = \mathbf{1}(\mathbf{T}^\top\mathbf{T}\circ\mathbf{I}) + (\mathbf{T}\mathbf{T}^\top\circ\mathbf{I})\mathbf{1} - 2\mathbf{T}^\top\mathbf{T}. \tag{3}$$

In this equation, $\mathbf{1}\in\mathbb{R}^{d\times d}$ is a matrix where each element equals 1, $\mathbf{I}$ signifies the identity matrix, $\circ$ denotes the Hadamard product, and $\top$ represents the matrix transpose.

The entry located at the $k$-th row and $l$-th column $r_{k,l}^{\mathbf{T}}$ of the so-called *BDC matrix*, denoted as $\mathbf{R}^{\mathbf{T}} = (r_{k,l}^{\mathbf{T}})$, is defined relative to the Euclidean distance $d_{k,l}^{\mathbf{T}}$ as

$$r_{k,l}^{\mathbf{T}} = d_{k,l}^{\mathbf{T}} - \frac{1}{m}\sum_{k=1}^{m} d_{k,l}^{\mathbf{T}} - \frac{1}{m}\sum_{l=1}^{m} d_{k,l}^{\mathbf{T}} - \frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m} d_{k,l}^{\mathbf{T}}, \tag{4}$$

where the final three terms represent the means of the $l$-th column, $k$-th row, and all entries of $\mathbf{D}_1^{\mathbf{T}}$, respectively. Consequently, we can represent $\mathbf{R}^{\mathbf{T}}$ concerning $\mathbf{D}_1^{\mathbf{T}}$ as

$$\mathbf{R}^{\mathbf{T}} = \mathbf{D}_1^{\mathbf{T}} - \frac{1}{d}\left(\mathbf{1}\mathbf{D}_1^{\mathbf{T}} + \mathbf{D}_1^{\mathbf{T}\top}\mathbf{1}\right) + \frac{1}{d^2}\mathbf{1}\mathbf{D}_1^{\mathbf{T}}\mathbf{1}. \tag{5}$$

The matrix $\mathbf{R}^{\mathbf{S}}$ can be derived analogously from $\mathbf{S}$. Subsequently, the BDC metric assumes the ensuing form, as per [46]:

$$\mathcal{V}(X,Y) = \mathrm{tr}\left(\mathbf{R}^{\mathbf{T}\top}\mathbf{R}^{\mathbf{S}}\right). \tag{6}$$

Here, $\mathrm{tr}(\cdot)$ denotes the matrix trace.

Referring to the above derivation, it is clear that the BDC metric facilitates an explicit expression in terms of the feature matrix. Subsequent to this, the construction of the BDC module proceeds as follows. Specifically, we devise a dual-layer module, which firstly reduces feature dimension and then computes the BDC matrix. **(1) Dimension Reduction:** Owing to the polynomial increase in the computation complexity of the BDC matrix with respect to the number of channels of the feature, we incorporate a convolution layer for the purpose of dimension reduction. **(2) Calculation of BDC Matrix:** Assuming that the previous layer embeds the input image $\mathbf{u}\in\mathbb{R}^{H\times W\times 3}$ into the feature represented by a $g\times d$ tensor and each column or each row of the tensor can be considered as a characteristic of observation from $X$. In the second layer, we calculate the BDC matrix according to Equation (3) and Equation (5). It should be noted that this layer contains no learnable parameters.

As a result, we characterize the BDC module as a training-free pooling layer. Deriving from Equation (3) and (5), it is obvious that the BDC matrix encapsulates non-linear interrelations among channels via the Euclidean distance. Consequently, when the relation between features is non-linear, the traditional covariance might be zero, which may potentially instigate a deceptive perception of independence. In contrast, the BDC is invariably non-negative and only amounts to zero when the features are indeed independent [46, 47]. This constitutes an advantage over conventional covariance, thereby positioning BDC as a more robust metric for evaluating dependence between features.

### 3.3 BDC-Adapter for CLIP

An overview of our proposed BDC-Adapter method is shown in Figure 2. In this section, we introduce each component of our proposed BDC-Adapter method in detail.

**Multi-Modal Few-Shot Learning.** Following prior works [31], we first construct text samples by appending the class label $y_i$ to a hand-crafted prompt such as $\pi =$ "a photo of a", then we get the text descriptions $t_i = \{\pi; y_i\}$ for each class $y_i$ in all $N$ classes. In each training batch, we randomly sample $n_1$ text descriptions $\{t_i\}_{i=1}^{n_1}$, $n_2$ image samples $\{x_i\}_{i=1}^{n_2}$ and their labels $\{y_i\}_{i=1}^{n}$, where $n = n_1 + n_2$ is the number of samples in a batch. We then extract the text feature or image feature $f_i$ for each sample, denoted as $f_i = E_v(x_i)$ for images or $f_i = E_t(t_i)$ for texts. Here we use $f_i$ for both features since the text or image samples will be projected onto the same dimensional embedding space by encoders of CLIP. Note that the multi-modal features $\{f_i\}_{i=1}^{n}$ in each batch is also L2 normalized. Based on these features, we learn a one-layer multi-modal reasoning network $\psi$ to classify the image, which can be denoted as

$$\psi(x) = W^\top x, \tag{7}$$

where $W$ is the parameter of the multi-modal reasoning network $\psi$ initialized with text features by $w_{y_i} = E_t(t_i), \forall i \in [1, N]$, where $w_{y_i}$ is the classification weight for class $y_i$ in parameter matrix $W$. The weights in this linear layer can be updated by gradient descent with the following cross-entropy loss during training:

$$\mathcal{L}_{CE} = \sum_{i=1}^{n} H\left(y_i, \psi\left(f_i\right)\right) = -\sum_{i=1}^{n} \log\left(\frac{e^{w_{y_i} \cdot f_i}}{\sum_{y'} e^{w_{y'} \cdot f_i}}\right). \tag{8}$$

**Class-Specific BDC Prototype Generation.** In a few-shot learning task, it provides $M$-shot $N$-class training samples (*i.e.* $M$ annotated images in each of the $N$ categories) in a new dataset. We can denote the $M$ images in a class as $\{x_m\}_{m=1}^{M}$ and the class labels as $\{y_n\}_{n=1}^{N}$. For each image $x$ within class $y$, we first utilize a modified visual encoder of CLIP $E_{vm}$ to generate its L2 normalized visual feature, then feed it to a BDC module to produce a BDC matrix $B_y(x)$. Given all the BDC matrix $\{B_y(x_m)\}_{m=1}^{M}$ of $M$ images within class $y$, we define the prototype of class $y$ to be the average of the BDC matrices, denoted as $P_y = \frac{1}{M}\sum_{m=1}^{M} B_y(x_m)$. Therefore, for the entire training dataset, we can build a prototype set $\mathcal{P} = \{P_{y_n}\}_{n=1}^{N}$.

**BDC-Adapter Inference.** During inference, for a test image $x_{test}$, we first utilize the visual encoder to extract its image feature $f_{test} = E_v(x_{test})$. Therefore, the prediction of the multi-modal reasoning network can be denoted as

$$p_m(y = y_n | x_{test}) = w_{y_n} \cdot f_{test}, \tag{9}$$

where $1 \leqslant n \leqslant N$ is the class index.

After the BDC prototype generation process, we have obtained the prototype of the few-shot training samples. Similarly, we utilize $E_{vm}$ to generate the image feature of $x_{test}$, then feed it into the BDC module and obtain the BDC matrix $B(x_{test})$. We can get the prediction of image $x_{test}$ via calculating the similarity between $B(x_{test})$ and the prototypes in the set $\mathcal{P}$, denoted as

$$p_b(y=y_n|x_{test}) = \exp\left(-\delta\left(1-\mathrm{vec}\left(B(x_{test})\right)\cdot\mathrm{vec}\left(P_{y_n}\right)\right)\right), \tag{10}$$

where $\mathrm{vec}(\cdot)$ denotes the vectorization of a matrix. The term $\mathrm{vec}\left(B(x_{test})\right)\cdot\mathrm{vec}\left(P_{y_n}\right)$ is equivalent to the cosine similarities between the BDC matrix of the test image $x_{test}$ and the prototype matrix $P_{y_n}$. The exponential function is adopted to convert the similarities into non-negative values with $\delta$ adjusting its sharpness.

We then combine $p_m$ and $p_b$ to get the final prediction,

$$\begin{aligned} p(y=y_n|x_{test}) &= \alpha p_b(y=y_n|x_{test}) + p_m(y=y_n|x_{test}) \\ &= \alpha\exp\left(-\delta\left(1-\mathrm{vec}\left(B(x_{test})\right)\cdot\mathrm{vec}\left(P_{y_n}\right)\right)\right) + w_{y_n}\cdot f_{test}, \end{aligned} \tag{11}$$

where $\alpha$ is the residual ratio to combine two predictions. Note that $w_{y_n}$ is the weight for class $y_n$ in the linear layer $\psi$ and can be updated by $\mathcal{L}_{CE}$ defined in Equation (8) during training. The final predicted label of test image $x_{test}$ is produced by $\hat{y} = \arg\max_{y'} p(y'|x_{test})$.

For clarity, we also analyze the sensitivity levels of the hyper-parameters and provide the pseudo-code of our method in the Supplemental Materials.

# 4 Experiment

## 4.1 Experiment Setup

To comprehensively evaluate the performance of our proposed BDC-Adapter method, we conduct experiments on few-shot image recognition, domain generalization, and visual reasoning tasks. For **few-shot image recognition**, we follow prior methods [63, 68] and adopt the common few-shot protocol to evaluate our method on 11 well-known image classification datasets, including generic object classification, fine-grained object classification, remote sensing recognition, texture classification, scene recognition, and action recognition: ImageNet [40], Caltech101 [15], OxfordPets [37], StandfordCars [28], Flowers102 [36], Food-101 [5], FGVC Aircraft [34], DTD [9], SUN397 [57], EuroSAT [21], and UCF101 [45]. These datasets provide a comprehensive benchmark to evaluate the few-shot learning performance of each method. For **domain generalization**, we evaluate the model's robustness to natural distribution shifts by training on 16-shot ImageNet [11] and testing on four variants of ImageNet: ImageNet-V2 [40], ImageNet-Sketch [51], ImageNet-A [23], and ImageNet-R [22]. Those variant datasets have been treated as out-of-distribution data for ImageNet in previous work [42, 67]. For **visual reasoning on human object interaction (HOI)**, we conduct experiments on Bongard-HOI [26] benchmark to evaluate the effectiveness of our proposed BDC-Adapter method on visual reasoning tasks.

## 4.2 Implementation Details

Our BDC-Adapter is based on CLIP [39] with ResNet-50 image encoder and Transformer text encoder. In the training stage, we freeze the weights to inherit the prior knowledge. Note that $E_{vm}$ defined in Section 3.3 is the modified image encoder of CLIP that does not

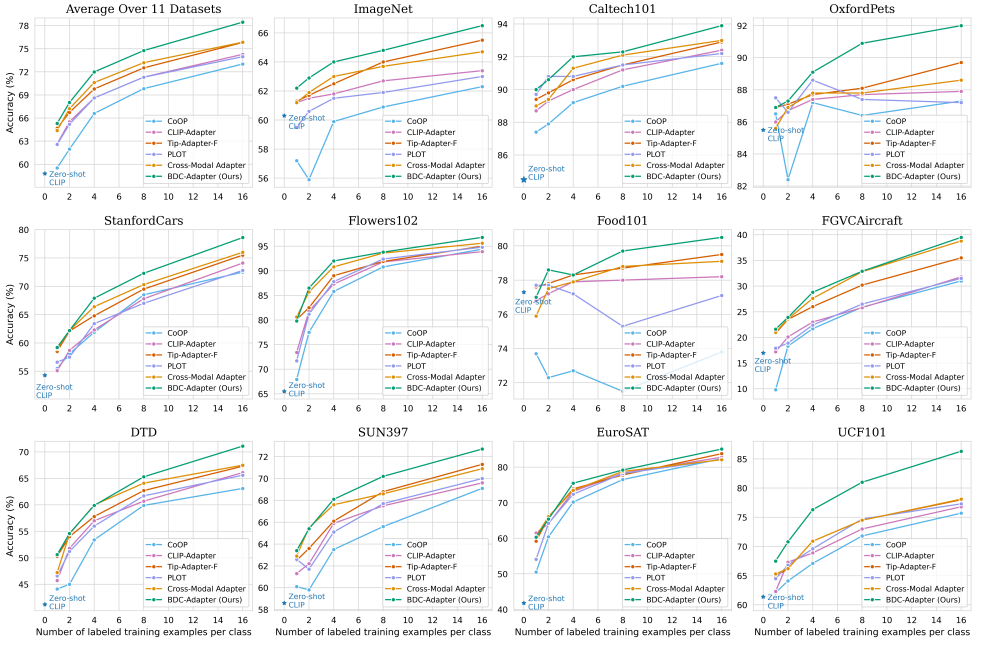


Figure 3: **Performance comparisons on few-shot learning on 11 datasets**. For each dataset, we report the accuracy on 1-/2-/4-/8-/16-shot settings. The top-left subfigure shows the average accuracy over all 11 datasets.

include the last attention pooling layer, whose output will be fed to the BDC module. For multi-modal few-shot learning introduced in Section 3.3, there are no requirements for the image and text samples to be exactly matched, and the number of samples for each modality can vary in a batch, which means $n_1$ (the number of image samples) is not always equal to $n_2$ (the number of text samples). Therein, $n_1$ is equal to the number of shots for training (*i.e.*, 1, 2, 4, 8, 16 in our experiments). Following prior methods, we apply the data pre-processing protocol in CLIP [39], such as resizing and random cropping operations, *etc*. On all the datasets in our experiments, we train our model for 30 epochs and set the initial learning rate as $1 \times 10^{-3}$. The AdamW [27] optimizer with a cosine annealing scheduler is used to optimize the parameters. Our method is parameter-efficient and lightweight, and we only use one single NVIDIA RTX 3090 GPU for training. For the visual reasoning on HOI task, we further introduce the implementation details in the Supplemental Materials.

## 4.3 Performance Analysis

### 4.3.1 Few-Shot Learning

Figure 3 compares the performance of our method with five baseline methods on all 11 datasets. We also present the average accuracy in the top-left sub-figure of Figure 3. We observe that our BDC-Adapter outperforms other state-of-the-art methods and obtains the highest average accuracy. In comparison to Tip-Adapter-F [63] (a fine-tuned version of Tip-Adapter), our method consistently outperforms it by large margins on all 11 datasets. This proves that our proposed BDC-Adapter can capture the non-linear relations ignored by Tip-Adapter-F [63] and fully characterize independence.

We notice that our method experiences a performance drop when using 4 shots on Food101, which appears to be a common overfitting challenge encountered not only by our approach but also by several existing adaptation methods like PLOT [6] and CoOp [58]. However, the overall results have demonstrated the effectiveness of our BDC-Adapter.

### 4.3.2 Domain Generalization

Table 1 summarizes the performance of our proposed BDC-Adapter and other state-of-the-art methods. For a fair comparison, we directly include the results of other baselines from their original paper. We report the classification accuracy of the source domain (ImageNet [11]), target domain (ImageNet-V2 [40], ImageNet-Sketch [51], ImageNet-A [23], and ImageNet-R [22]), and the average accuracy on out-of-distribution data. Our method outperforms the DeFo [50] method on 3 out of 4 target datasets, and surpasses all other baselines in all metrics. These results indicate that our BDC-Adapter exhibits remarkable robustness to distribution shifts.

Table 1: **Performance comparisons on robustness to natural distribution shifts**. All the experiments are conducted with ResNet-50 visual backbone. The best results are in **bold** and the second are underlined.

| Method | Source | Target | | | | |
|---|---|---|---|---|---|---|
| | ImageNet | -V2 | -Sketch | -A | -R | Avg. |
| Zero-Shot CLIP [43] | 60.33 | 53.27 | 35.44 | 21.65 | 56.00 | 41.59 |
| Linear Probe CLIP [43] | 56.13 | 45.61 | 19.13 | 12.74 | 34.86 | 28.09 |
| CoOp [58] | 63.33 | 55.40 | 34.67 | 23.06 | 56.60 | 42.43 |
| CoCoOp [57] | 62.81 | 55.72 | 34.48 | 23.32 | 57.74 | 42.82 |
| ProGrad [60] | 62.17 | 54.70 | 34.40 | 23.05 | 56.77 | 42.23 |
| PLOT [6] | 63.01 | 55.11 | 33.00 | 21.86 | 55.61 | 41.40 |
| DeFo [50] | 64.00 | **58.41** | 33.18 | 21.68 | 55.84 | 42.28 |
| TPT [45] | 60.74 | 54.70 | 35.09 | 26.67 | 59.11 | 43.89 |
| TPT + CoOp [45] | 64.73 | 57.83 | 35.86 | 30.32 | 58.99 | 45.75 |
| **BDC-Adapter (Ours)** | **66.46** | 58.05 | **36.92** | **30.77** | **59.52** | **46.31** |

### 4.3.3 Visual Reasoning on Bongard-HOI

In Figure 4, we illustrate some instances in the Bongard-HOI dataset [26]. Note that there are 6 positive examples, 6 negative examples, and 1 query image in a test instance, which is different from the illustration here. Following the experimental design outlined in Jiang *et al*. [26], the comparison is conducted on four distinct test splits of the Bongard-HOI dataset. It should be noted that the results for the other baselines are sourced directly from the research paper by Jiang *et al*. [26]. For more details on this task, interested readers are directed to this paper. We

Table 2: **Performance comparisons on the Bongard-HOI [26] dataset**. The last column shows the average accuracy. The best results are in **bold** and the second are underlined.

| Method | Test Splits | | | | |
|---|---|---|---|---|---|
| | Seen act. Seen obj. | Unseen act. Seen obj. | Seen act. Unseen obj. | Unseen act. Unseen obj. | Avg. |
| CNN-Baseline [26] | 50.03 | 49.89 | 49.77 | 50.01 | 49.92 |
| Meta-Baseline [8] | 58.82 | 58.75 | 58.56 | 57.04 | 58.30 |
| ProtoNet [44] | 58.90 | 58.77 | 57.11 | 58.34 | 58.28 |
| HOITrans [54] | 59.50 | 64.38 | 63.10 | 62.87 | 62.46 |
| TPT (RN50) [45] | 66.39 | 68.50 | 65.98 | 65.48 | 66.59 |
| **BDC-Adapter** (RN50) | **68.36** | **69.15** | **67.67** | **67.82** | **68.25** |

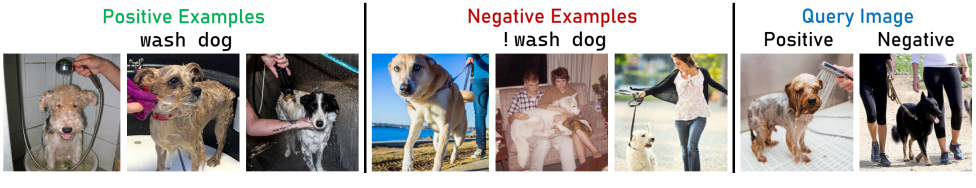


Figure 4: **Illustration of a few-shot learning instance from the Bongard-HOI [26] benchmark**. The left side shows positive images that depict the visual relationship of a person washing a dog. In contrast, negative examples do not exhibit such relationships. The right side shows query images, where the ground-truth labels are positive or negative, respectively.

compare the performance of the proposed BDC-Adapter approach with previous approaches in Table 2. Remarkably, our method outperforms the conventional methods, including ProtoNet [42] and HOITrans [72], by large margins. Even compared to the CLIP-based TPT method, BDC-Adapter still yields better performance in all 4 test splits.

## 4.4 Ablation Study

To systematically evaluate the effectiveness of our proposed BDC-Adapter, we conduct an ablation study on the ImageNet [11] dataset to analyze the impacts of different components in our BDC-Adapter. Table 3 presents the performance results, where the last row shows the accuracy of our full BDC-Adapter. We can see that both initialization of the multi-modal reasoning network and BDC prototype similarity reasoning contribute significantly to the overall performance.

Table 3: **Effectiveness of different components in our BDC-Adapter method**. MRN represents multi-modal reasoning network and BDC represents BDC prototype similarity reasoning, init. stands for initialization.

| Few-shot Setup | 1 | 2 | 4 | 8 | 16 |
|---|---|---|---|---|---|
| MRN (*w/o* init.) | 60.55 | 61.07 | 61.89 | 63.04 | 63.57 |
| MRN (*w/* init.) | 61.12 | 61.77 | 62.73 | 63.78 | 64.68 |
| **MRN + BDC (Ours)** | **62.19** | **62.91** | **63.95** | **64.83** | **66.46** |

## 4.5 Efficiency Comparison

In order to show the great fine-tuning efficiency of our BDC-Adapter, we compare the number of training epochs, training time, computational cost, and number of parameters of our method with other state-of-the-art methods on 16-shot ImageNet using a single NVIDIA RTX 3090 GPU. We report the comprehensive results in Table 4. Our BDC-Adapter has only a single linear layer for training, thereby exhibiting great efficiency in fine-tuning VLMs. With just 2 minutes of training and 1 MFLOP on a single RTX 3090, our BDC-Adapter achieves a remarkable accuracy of 66.46% on 16-shot ImageNet. In comparison, the CoOp method needs about 15 hours of training and 4 MFLOPs to achieve 62.26% accuracy; the Tip-Adapter-F method needs 5 minutes of training and 30 MFLOPs to achieve 65.51% accuracy.

Table 4: **Efficiency comparisons on 16-shot ImageNet**. We report the results using a single NVIDIA RTX 3090 GPU.

| Method | Epochs | Training | GFLOPs | Param. | Acc. |
|---|---|---|---|---|---|
| CoOp [63] | 200 | 15 h | >10 | **0.01M** | 62.95 |
| CLIP-Adapter [63] | 200 | 50 min | 0.004 | 0.52M | 63.59 |
| Tip-Adapter-F [63] | **20** | 5 min | 0.030 | 16.3M | 65.51 |
| **BDC-Adapter (Ours)** | **20** | **2 min** | **0.001** | 1.02M | **66.46** |

# 5 Conclusion

In this work, we innovatively introduce Brownian Distance Covariance to the field of vision-language reasoning, which provides a more robust metric for measuring feature dependence to enable beter generalization capability. Based on this, we present a novel method called BDC-Adapter, which takes advantage of the BDC metric in computing the similarities between the few-shot BDC prototypes and the BDC matrix of the test image. Meanwhile, BDC-Adapter only introduces a one-layer multi-modal reasoning network that learns from multi-modal few-shot instances, to adapt VLMs to downstream tasks using limited training data. Our extensive experiment results indicate the effectiveness of our proposed BDC-Adapter method for fine-tuning VLMs. With its lightweight and parameter-efficient design, BDC-Adapter not only exhibits better vision-language reasoning capabilities but also has lower computational complexity, which makes it suitable for practical applications.

# References


[1] Mohamed Afham, Salman Khan, Muhammad Haris Khan, Muzammal Naseer, and Fahad Shahbaz Khan. Rich semantics improve few-shot learning. In *Proceedings of the British Machine Vision Conference*, 2021.

[2] Sławomir Bąk, Ratnesh Kumar, and François Brémond. Brownian descriptor: A rich meta-feature for appearance matching. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 363–370. IEEE, 2014.

[3] Sławomir Bąk, Marco San Biagio, Ratnesh Kumar, Vittorio Murino, and François Brémond. Exploiting feature correlations by brownian statistics for people detection and recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(9): 2538–2549, 2016.

[4] Piotr Bilinski, Michal Koperski, Sławomir Bąk, and François Bremond. Representing visual appearance by video brownian covariance descriptor for human action recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 87–92. IEEE, 2014.

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014.

[6] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations*, 2023.

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations*, 2023.

[8] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9062–9071, 2021.

[9] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[10] Quan Cui, Boyan Zhou, Yu Guo, Weidong Yin, Hao Wu, Osamu Yoshie, and Yubo Chen. Contrastive vision-language pre-training with limited resources. In *European Conference on Computer Vision*, pages 236–253. Springer, 2022.

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

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[13] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022.

[14] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15651–15660, 2022.

[15] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 178–178, 2004.

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[17] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

[18] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[19] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. HiCLIP: Contrastive language-image pretraining with hierarchy-aware attention. In *International Conference on Learning Representations*, 2023.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[21] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[22] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[24] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9068–9077, 2022.

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916, 2021.

[26] Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19056–19065, 2022.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[29] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021.

[30] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[31] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337, 2023.

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32, pages 13–23, 2019.

[33] Chaofan Ma, Yuhuan Yang, Yanfeng Wang, Ya Zhang, and Weidi Xie. Open-vocabulary semantic segmentation with frozen vision-language models. In *Proceedings of the British Machine Vision Conference*, 2022.

[34] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[35] Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. Bongard-logo: A new benchmark for human-level concept learning and reasoning. In *Advances in Neural Information Processing Systems*, volume 33, pages 16468–16480, 2020.

[36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729. IEEE, 2008.

[37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012.

[38] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5822–5830, 2018.

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

[40] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[41] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9611–9620, 2022.

[42] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 14274–14289, 2022.

[43] Mustafa Shukor, Guillaume Couairon, and Matthieu Cord. Efficient vision-language pretraining with visual concepts and hierarchical alignment. In *Proceedings of the British Machine Vision Conference*, 2022.

[44] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 4080–4090, 2017.

[45] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[46] Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, 3:1236–1265, 2009.

[47] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35:2769–2794, 2007.

[48] Tatiana Tommasi and Barbara Caputo. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In *Proceedings of the British Machine Vision Conference*, pages 80.1–80.11, 2009.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010, 2017.

[50] Feng Wang, Manling Li, Xudong Lin, Hairong Lv, Alex Schwing, and Heng Ji. Learning to decompose visual features with latent textual prompts. In *International Conference on Learning Representations*, 2023.

[51] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, pages 10506–10518, 2019.

[52] Ning Wang, Jiangrong Xie, Hang Luo, Qinglin Cheng, Jihao Wu, Mingbo Jia, and Linlin Li. Efficient image captioning for edge devices. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2608–2616, 2023.

[53] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[54] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020.

[55] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022.

[56] Yang Wu, Hao Zhang, Lingyan Liang, Yaqian Zhao, and Kaihua Zhang. Group-wise co-salient object detection with siamese transformers via brownian distance covariance matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023.

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

[58] Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7972–7981, 2022.

[59] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021.

[60] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8808–8817, 2020.

[61] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[62] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[63] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022.

[64] Shuzhen Zhang, Ting Lu, Shutao Li, and Wei Fu. Superpixel-based brownian descriptor for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021.

[65] Yi Zhang, Ce Zhang, Xueting Hu, and Zhihai He. Unsupervised prototype adapter for vision-language models. In *Chinese Conference on Pattern Recognition and Computer Vision*, 2023.

[66] Yi Zhang, Ce Zhang, Yushun Tang, and Zhihai He. Cross-modal concept learning and inference for vision-language models. *arXiv preprint arXiv:2307.15460*, 2023.

[67] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[68] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[69] Mingyang Zhou, Licheng Yu, Amanpreet Singh, Mengjiao Wang, Zhou Yu, and Ning Zhang. Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16485–16494, 2022.

[70] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[71] Linhai Zhuo, Yuqian Fu, Jingjing Chen, Yixin Cao, and Yu-Gang Jiang. Tgdm: Target guided dynamic mixup for cross-domain few-shot learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6368–6376, 2022.

[72] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021.