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

In this Supplemental Material, we provide additional experimental results and implementation details for further understanding of our proposed BDC-Adapter method.

Contents

A Additional Experimental Results **1**
A.1 Comparison Methods 1
A.2 Specific Per-Dataset Results on Few-Shot Learning 2
A.3 Domain Generalization on DomainNet 3
A.4 More Efficiency Comparisons 3
A.5 Sensitivity of Hyper-Parameters 3

B Additional Implementation Details **3**
B.1 Implementation Details for Visual Reasoning on HOI 3
B.2 Algorithm Pseudo-Code of Our Method 5

A Additional Experimental Results

A.1 Comparison Methods

In few-shot learning task (Section 4.3.1), we compare our method with the following state-of-the-art methods: zero-shot CLIP [10], CoOp [17], PLOT [2], CLIP-Adapter [5], Cross-Modal Adapter [7], and Tip-Adapter-F [14]. For a fair comparison, we choose CoOp's best-

*Equal contribution. †Corresponding author.

performance setting (with the class token placed at the end of 16-token prompts) and the fine-tuned Tip-Adapter-F in our experiments.

In domain generalization task (Section 4.3.2), we compare our method with the following state-of-the-art methods: zero-shot CLIP [10], linear probe CLIP [10], CoOp [17], CoCoOp [16], ProGrad [18], PLOT [2], DeFo [13], TPT [11], TPT + CoOp [11].

In visual reasoning on HOI task (Section 4.3.3), we compare our method with the following state-of-the-art methods: CNN-Baseline [9], Meta-Baseline [3], ProtoNet [12], HOITrans [19], and TPT [11] based on ResNet-50 image encoder of CLIP.

A.2 Specific Per-Dataset Results on Few-Shot Learning

In Section 4.3.1, we provide the performance analysis of our method and other state-of-the-art methods on the few-shot learning task. For easier comparison with other methods, we also provide the specific numerical results in Table 1. Note that we use the ResNet-50 visual backbone for all the methods in this table. The average accuracy shown in the last column clearly demonstrates the superior performance of our proposed BDC-Adapter method. This proves the effectiveness of using Brownian Distance Covariance for vision-language reasoning. It should be noted that, our method also exhibits lower computational complexity, which will be shown in the following Section A.4.

Table 1: **Per-dataset results using ResNet-50 backbone.** We also report the performance results of the baseline methods from existing works. We **bold** the best result for each shot and each dataset, and underline the second best result.

Method	Shots	Dataset											
		Cal	Image	DTD	Euro	FGVC	Food	Flower	Pets	Cars	SUN	UCF	Avg.
CLIP [10]	0	84.5	60.3	41.2	41.8	17.0	77.3	65.5	85.5	54.3	58.6	61.4	58.8
CoOp [17]	1	87.4	57.2	44.1	50.5	9.8	73.7	67.9	86.5	55.5	60.1	62.1	59.5
	2	87.9	55.9	45.0	60.4	18.3	72.3	77.5	82.4	58.1	59.8	64.1	62.0
	4	89.2	59.9	53.4	70.2	21.7	72.7	85.8	87.2	61.9	63.5	67.1	66.6
	8	90.2	60.9	59.9	76.5	25.9	71.5	90.8	86.4	68.5	65.6	71.8	69.8
	16	91.6	62.3	63.1	82.4	31.0	73.8	94.4	87.3	72.5	69.1	75.7	73.0
CLIP-Adapter [8]	1	88.7	61.2	45.7	**61.5**	17.2	76.8	73.4	86.0	55.1	61.3	62.3	62.7
	2	89.3	61.5	51.8	64.1	20.1	77.2	81.8	86.7	58.7	62.2	67.3	65.5
	4	90.0	61.8	57.0	73.2	23.0	77.9	87.3	87.4	62.3	65.9	68.9	68.6
	8	91.2	62.7	60.7	78.3	25.8	78.0	91.8	87.7	67.8	67.5	73.0	71.3
	16	92.4	63.4	66.1	82.8	31.8	78.2	93.9	87.9	74.1	69.6	76.8	74.3
Tip-Adapter-F [24]	1	89.4	61.3	50.3	59.2	20.8	77.6	80.1	86.9	58.5	62.5	64.9	64.7
	2	89.8	61.7	54.0	65.8	23.5	77.8	82.5	87.1	62.1	63.6	66.2	66.7
	4	90.6	62.5	57.8	73.9	26.0	**78.3**	89.0	87.7	64.8	66.1	70.9	69.8
	8	91.5	64.0	62.7	77.8	30.2	78.7	91.9	88.1	69.5	68.8	74.5	72.5
	16	92.9	65.5	67.3	83.8	35.5	79.5	95.0	89.7	75.5	71.3	78.0	75.8
PLOT [2]	1	89.7	59.5	46.6	54.1	17.9	**77.7**	71.7	**87.5**	56.6	62.6	64.5	62.6
	2	**90.8**	60.6	51.2	64.2	18.9	77.7	81.2	86.6	57.5	61.7	66.8	65.2
	4	90.8	61.5	56.0	72.4	22.4	77.2	87.8	88.6	63.4	65.1	69.6	68.6
	8	91.5	61.9	61.7	78.2	26.5	75.3	92.4	87.4	67.0	67.7	74.7	71.3
	16	92.2	63.0	65.6	82.2	31.5	77.1	94.8	87.2	72.8	70.0	77.3	74.0
Cross-Modal Adapter [14]	1	89.0	61.2	47.2	60.5	21.0	75.9	**80.6**	85.6	59.0	62.9	65.3	64.4
	2	89.4	61.9	54.5	**66.1**	23.6	77.5	85.7	86.9	**62.2**	65.5	66.2	67.2
	4	91.3	63.0	**60.0**	73.5	27.6	77.9	90.8	87.8	66.4	67.6	70.9	70.6
	8	92.1	63.7	64.1	78.8	32.8	78.8	93.6	87.8	70.3	68.6	74.5	73.2
	16	93.0	64.7	67.5	82.1	38.8	79.7	95.6	88.6	76.0	70.9	78.1	75.9
BDC-Adapter (Ours)	1	**90.0**	**62.2**	**50.6**	60.3	**21.6**	77.0	79.8	86.9	**59.2**	**63.4**	**67.5**	**65.3**
	2	90.6	**62.9**	**54.6**	65.4	**23.9**	**78.6**	**86.5**	**87.3**	**62.2**	**65.4**	**70.8**	**68.0**
	4	**92.0**	**64.0**	59.9	**75.5**	**28.8**	**78.3**	**92.0**	**89.1**	**67.9**	**68.1**	**76.3**	**72.0**
	8	**92.3**	**64.8**	**65.3**	**79.2**	**32.9**	**79.7**	**93.8**	**90.9**	**72.3**	**70.2**	**81.0**	**74.8**
	16	**93.9**	**66.5**	**71.1**	**85.1**	**39.5**	**80.5**	**96.8**	**92.0**	**78.6**	**72.7**	**86.3**	**78.5**

A.3 Domain Generalization on DomainNet

We also conduct a new domain generalization experiment on DomainNet and report the performance results in Table 2. For a fair comparison, we use the same ViT-B/16 image encoder of CLIP. All performance scores are evaluated by leave-one-out cross-validation protocol. We can see that our BDC-Adapter also achieves state-of-the-art performance, which demonstrates the effectiveness of our method.

Table 2: **Domain generalization performances on DomainNet**. We use ViT-B/16 image encoder of CLIP for this experiment.

Method	ERM	MIRO [9]	DPL [35]	CAR-FT [8]	**Ours**
Accuracy	53.8	54.0	56.7	62.5	**64.8**

A.4 More Efficiency Comparisons

In our BDC-Adapter method, we only include a single linear layer to be updated during training, which makes our method highly efficient in few-shot training and reasoning. Table 3 compares the training time and accuracy of our method and other state-of-the-art methods for 16-shot recognition on ImageNet [4]. These results prove that with the lightweight and parameter-efficient design, BDC-Adapter not only exhibits better vision-language reasoning capabilities but also has lower computational complexity.

Table 3: **Efficiency and accuracy for different methods on ImageNet-16-shot.** All experiments are tested on a single NVIDIA GeForce RTX 3090 GPU.

Method	Epochs	Time	Accuracy	Gain
Zero-shot CLIP [11]	0	0	60.33	0
Linear Probe CLIP [11]	-	13min	56.13	-4.20
CoOp [12]	200	14h 40min	62.26	+1.93
ProGrad [13]	200	17hr	63.45	+3.12
CLIP-Adapter [6]	200	50min	63.59	+3.26
Cross-Modal Adapter [9]	20	**2min**	64.72	+4.39
Tip-Adapter-F [14]	20	5min	65.51	+5.18
BDC-Adapter (Ours)	20	**2min**	**66.46**	**+6.13**

A.5 Sensitivity of Hyper-Parameters

In our experiments on ImageNet [4], we set the hyper-parameters α and τ defined in Section 3 to 1.5 and 4.5, respectively. To comprehensively investigate the effects of different hyperparameters, we conducted a sensitivity experiment where we varied each hyperparameter individually and evaluated the performance on 16-shot ImageNet. The ablation results are presented in Table 4. We can see that the value of the residual ratio α has a significant impact on the model's performance, while influence of temperature hyper-parameter τ on the model's performance is minor.

Table 4: **Sensitivity of hyper-parameters**. All the results are reported on a 16-shot setting on ImageNet [4].

α	0.0	0.5	1.0	**1.5**	2.0	2.5
	63.97	64.33	65.41	**66.46**	65.27	64.44
τ	0.5	2.5	**4.5**	6.5	8.5	10.5
	65.88	66.05	**66.46**	66.17	66.03	65.73

B Additional Implementation Details

B.1 Implementation Details for Visual Reasoning on HOI

In context-dependent visual reasoning, for instance, in the Bongard-HOI task [5], a test sample consists of a query image and two sets of support images that are used to evaluate it. The mentioned sets of support images are used to demonstrate the existence or non-existence

Figure 1: **Illustration of two few-shot learning instances from the Bongard-HOI [6] benchmark**. In each instance, there exist 6 positive examples, 6 negative examples, and 1 query image.

of a human-object interaction (HOI) concept, such as "wash dog." Afterwards, the model's objective is to determine whether the HOI concept is present in the query image.In this specific task, each concept is expressed as a visual relationship, indicated as $c =< s,a,o >$, where s represents the subject (typically "human" for HOI tasks), a represents the action, and o denotes the involved object. Each test sample, denoted as X_{test}, encompasses a distinct concept by showcasing $c =< s,a,o >$ in a set of support images, which are regarded as positive instances. On the contrary, the alternate set of support images functions as negative examples, illustrating $\hat{c} =< s,\hat{a},o >$, where $\hat{a}$ differs from a. It should be noted that the object o and the action a are not explicitly given in the task. Hence, the model's reasoning ability is crucial for predicting whether the concept c is present or absent in the query image of the test sample. Previous studies [3, 9] have addressed the Bongard-HOI problem by training the model on various similar tasks using the Bongard-HOI training split. This approach allows the model to make relevant inferences on test samples during evaluation. In this task, the use of CLIP does not require additional training data since the CLIP model already possesses a comprehensive understanding of diverse visual concepts. Hence, CLIP is a suitable option for this type of visual reasoning task.

There are differences between visual reasoning in HOI and traditional few-shot image classification. In Bongard-HOI [6], the accuracy of predictions relies on the context, which consists of example images indicating the presence or absence of concept c. Since the labels are binary (either containing the concept or not), a simple prompting strategy involves manually assigning "labels" to positive and negative examples. We use "True" or "False" as labels. We create a hand-crafted prompt $\rho =$ "a photo that the person {`action`} {`object`}, it is {`class`}", where {`action`} represents actions like "wash", "run", "hug", "feed", *etc.*,

{object} represents objects in the image such as "orange", "bicycle", *etc.*, and {class} represents "true" or "false". As an example, we have the prompt: "a photo that person washes dog, it is true." The test set consists of 6 positive examples, 6 negative examples, and 1 query image. Examples of test instances in the Bongard-HOI [6] dataset are shown in Figure 1.

B.2 Algorithm Pseudo-Code of Our Method

An PyTorch-style pseudo-code for our BDC-Adapter method is shown in Algorithm 1. We show both the training stage and inference stage for clarity. Notably, in the training stage, the image and text samples do not need to be paired and one may sample different numbers of them per batch. The class-specific BDC prototype generation process defined in Section 3.3 is not presented in this pseudo-code.

Algorithm 1: Pseudo-code of our BDC-Adapter method.

```
# Training Stage
# T: the randomly chosen texts
# I: the randomly chosen images
# w:  linear layer initialized with text features
# tau:  temperature hyper-parameter
for e in Epochs:
    # Extract feature representations of each modality
    I_f = image_encoder(I)
    T_f = text_encoder(T)

    # Concatenate and L2 normalize
    # mm_features represents the multi-modal features
    mm_features = concat((I_f, T_f))
    mm_features = normalize(mm_features)
    labels = concat((I_labels, T_labels))

    # Compute loss with cross entropy
    # logits_mm represents the logits of multi-modal reasoning network
    logits_mm = w(mm_features)
    loss = cross_entropy_loss(logits_mm / tau, labels)

    # Update w by gradient descent
    loss.backward()
    optimizer.step()

# Inference Stage
# E_v is the orginial image encoder of CLIP. x is the test image.
# E_vm represents the modified image encoder of CLIP that does not
 include the last atttenion pooling layer
# B(·) stands for the BDC module defined in Section 3.2
B_x = B(E_vm(x))

# logits_bdc is the logits of BDC prototype similarity reasoning
# P is the BDC prototype set defined in Section 3.3
logits_bdc = np.dot(B(E_vm(x)), P.T)
logits_mm = w(E_v(x))

# alpha is the residual ratio to combine two predictions
logits = logits_mm + alpha * logits_bdc
```

References

[1] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *European Conference on Computer Vision*, pages 440–457. Springer, 2022.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Xiaofeng Mao, Yuefeng Chen, Xiaojun Jia, Rong Zhang, Hui Xue, and Zhao Li. Context-aware robust fine-tuning. *arXiv preprint arXiv:2211.16175*, 2022.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Xin Zhang, Shixiang Shane Gu, Yutaka Matsuo, and Yusuke Iwasawa. Domain prompt learning for efficiently adapting clip to unseen domains. *arXiv preprint arXiv:2111.12853*, 2021.

[16] 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.

[17] 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.

[18] 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.

[19] 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.