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

Yi Zhang*[1,2], Ce Zhang*[3], Zihan Liao[2], Yushun Tang[2], Zhihai He[2,4]

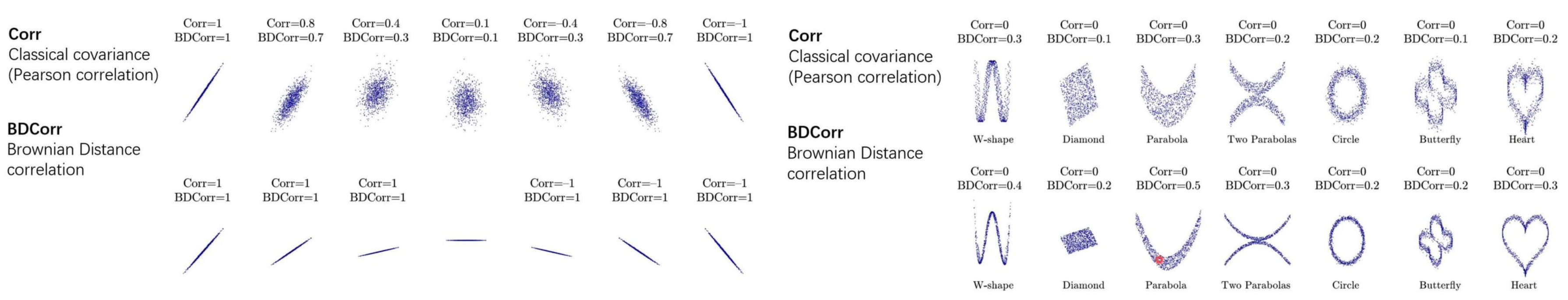
[1]Harbin Institute of Technology [2]Southern University of Science and Technology
[3]Carnegie Mellon University [4]Pengcheng Laboratory

I. 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.

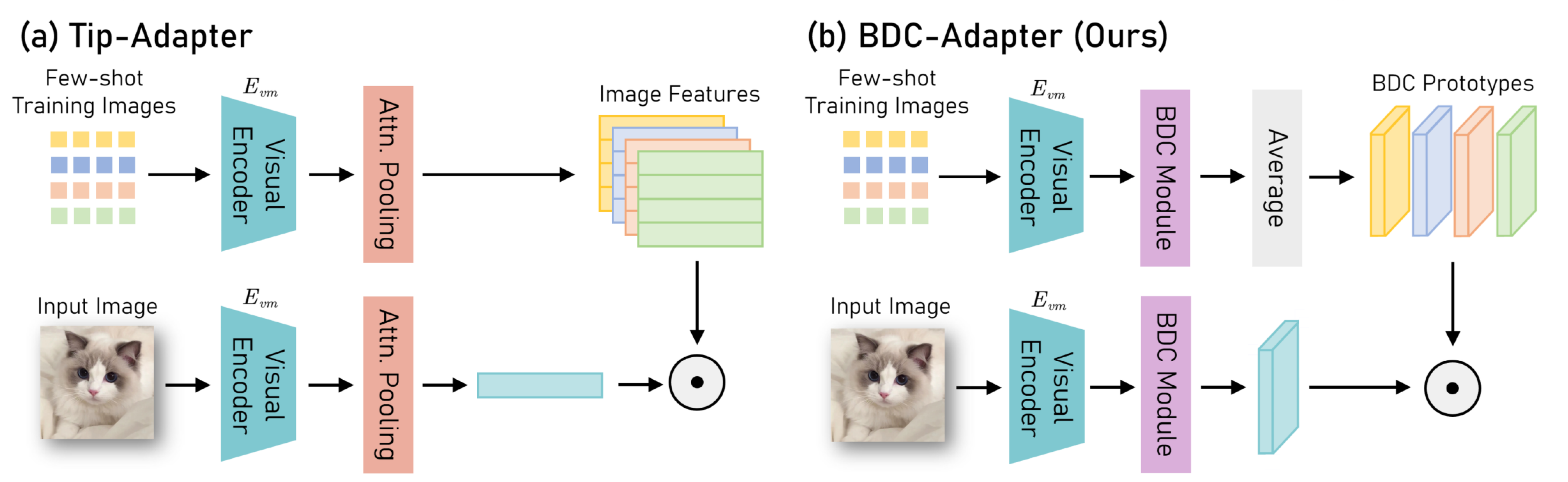
II. Motivation

- The current state-of-the-art Tip-Adapter method, 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 simply considers the covariance between each image feature pair, which only measures marginal distributions and captures linear relations.
- 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.



III. Method

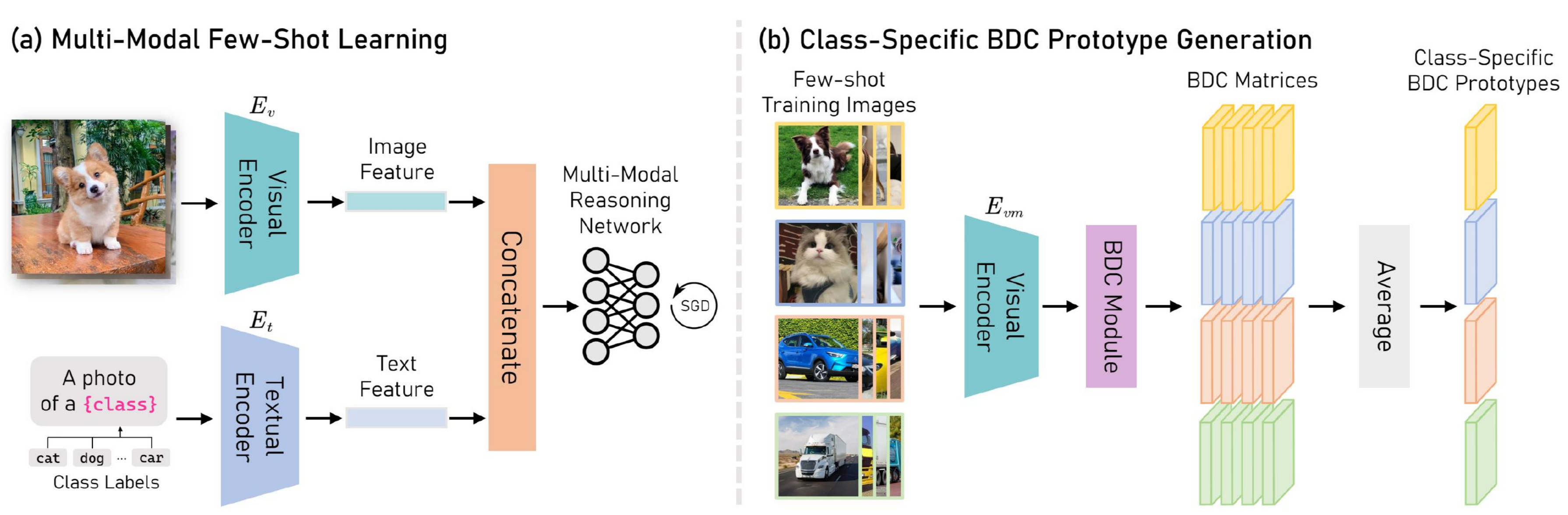
- Differences with Tip-Adapter. Tip-Adapter can only capture linear relations. Our BDC-Adapter represents each image by a BDC matrix, which considers the joint distributions and measures non-linear dependence during inference.



- Multi-Modal Few-Shot Learning. After feature extraction, we concatenate the image and text features and use this joint features f_i to train a one-layer multi-modal reasoning network ψ by cross-entropy loss:

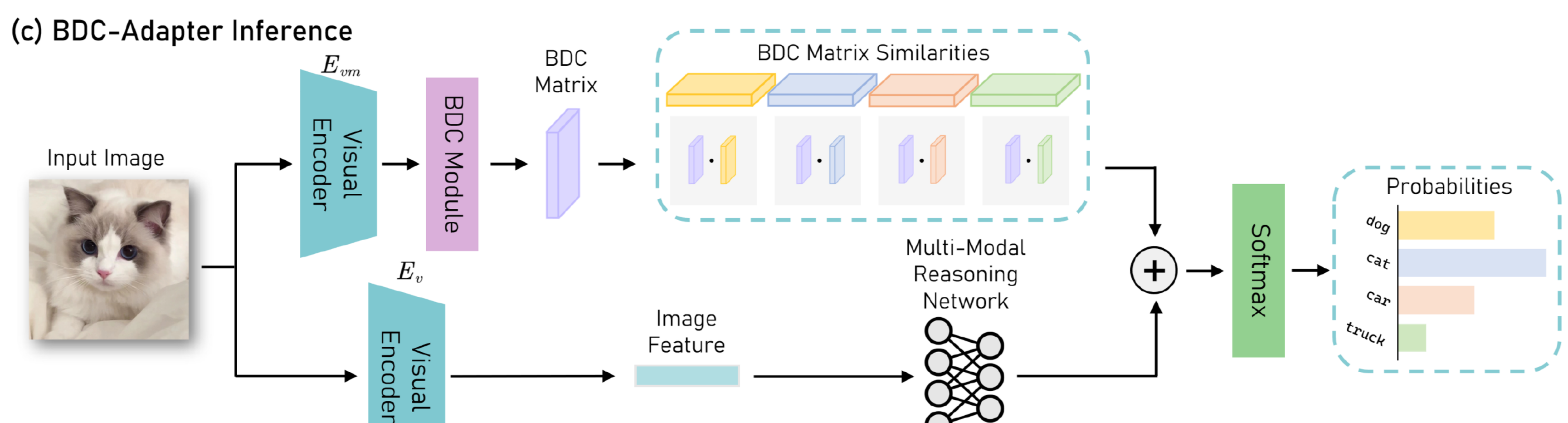
$$\mathcal{L}_{CE} = \sum_{i=1}^{n} H\left(y_i, \psi\left(f_i\right)\right) = -\sum_{i=1}^{n} \log\left(\frac{e^{w_{y_i} \cdot f_i}}{\sum_{y'} e^{w_{y'} \cdot f_i}}\right).$$

- Class-Specific BDC Prototype Generation. Given all the BDC matrices of M images within class y, we define the prototype of class y to be the average of the BDC matrices, denoted as $P_y = \frac{1}{M}\sum_{m=1}^{M} B_y(x_m)$.



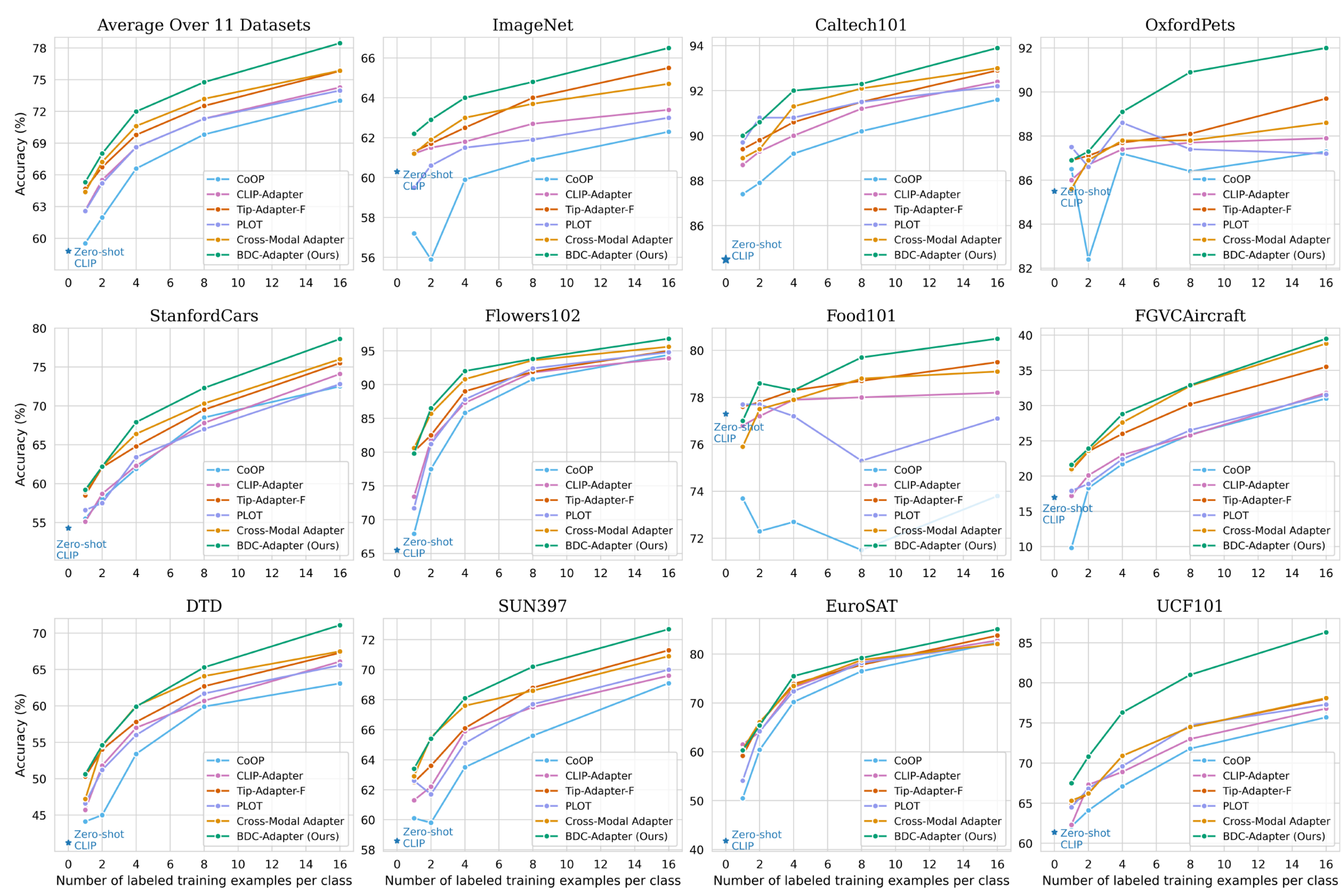
- BDC-Adapter Inference. During inference, BDC-Adapter integrates BDC prototype similarity reasoning and multi-modal reasoning network prediction to perform classification tasks, denoted as

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



IV. Experimental Results

- Performance comparisons on few-shot learning on 11 datasets.



- Performance comparisons on robustness to natural distribution shifts.

Method	Source	Target				
	ImageNet	-V2	-Sketch	-A	-R	Avg.
Zero-Shot CLIP [39]	60.33	53.27	35.44	21.65	56.00	41.59
Linear Probe CLIP [39]	56.13	45.61	19.13	12.74	34.86	28.09
CoOp [68]	63.33	55.40	34.67	23.06	56.60	42.43
CoCoOp [67]	62.81	55.72	34.48	23.32	57.74	42.82
ProGrad [70]	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 [42]	60.74	54.70	35.09	26.67	59.11	43.89
TPT + CoOp [42]	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**

- A few-shot learning instance from the Bongard-HOI.

- Visual reasoning performance comparisons on the Bongard-HOI dataset.

Method	Test Splits				
	Seen act. Seen obj.	Unseen act. Seen obj.	Seen act. Unseen obj.	Unseen act. Unseen obj.	Avg.
CNN-Baseline [35]	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 [72]	59.50	64.38	63.10	62.87	62.46
TPT (RN50) [42]	66.39	68.50	65.98	65.48	66.59
BDC-Adapter (RN50)	**68.36**	**69.15**	**67.67**	**67.82**	**68.25**

- Ablation study on 16-shot ImageNet.

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**

- Efficiency comparison.

Method	Epochs	Training	GFLOPs	Param.	Acc.
CoOp [68]	200	15 h	>10	**0.01M**	62.95
CLIP-Adapter [18]	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**

V. Contributions

- We introduce Brownian Distance Covariance to the field of vision-language reasoning to provide a robust metric for measuring feature dependence.
- Based on this, we propose a novel approach called BDC-Adapter that leverages BDC to enhance vision-language reasoning ability, which integrates BDC prototype similarity reasoning and multi-modal reasoning network prediction to perform classification tasks.
- Our extensive experimental results show that BDC-Adapter outperforms the current state-of-the-art methods by large margins.