APPENDIX FOR DISTRIBUTION CALIBRATION FOR FEW-SHOT LEARNING BY BAYESIAN RELATION INFERENCE

A APPENDIX

A.1 EXPERIMENT CONFIGURATION AND DETAILS

A.1.1 BASELINE METHODS DETAILS

To validate the effectiveness of the proposed model, we compare it with different baseline methods for few-shot learning. The baseline methods we compare are shown as follows:

- MAML Finn et al. (2017): An algorithm for meta-learning which is model-agnostic.
- Prototypical Networks(PN) Snell et al. (2017): A classical Metric-Based meta-learning method.
- Matching Networks(MN) Vinyals et al. (2016): A classical Metric-Based meta-learning method which use LSTM to augment the network.
- Distribution Calibration(DC) Yang et al. (2021): A method for distribution calibration based on manually set Euclidean distances.
- PatchProto + tSF(tSF) Lai et al. (2022): A transformer-based semantic filter with PatchProto network for few-shot classification.
- GAP Kang et al. (2023): A meta-learning method with a geometry-adaptive preconditioner.

A.1.2 TRAINING DETAILS

The details of the resources for training and the versions of the software are provided in Table 1.

Table 1: The hardware and software configuration for training.

Software	Python	3.7.11
	PyTorch	1.10.0
	numpy	1.21.2
Hardware	CPU	Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
	RAM	128 GB
	GPU	GeForce RTX 2080

A.1.3 IMPLEMENTATION DETAILS

For the training stage, we use the Adam optimizer and set the learning rate to 1**e**-4. We train the network for 500 epochs and save the best performing model on the validation set for testing. We conduct experiments on the 5-way-K-shot setting. The average accuracy of 150 episodes is reported as the final result. The details of the hyperparameters are provided in Table 2.

A.2 CODE AND DATASET

The code of BDC and the code to preprocess the Dermnet dataset is available at: https://anonymous.4open.science/r/BDC-F873.

Table 2: The configuration of hyper-parameters for training.

Hyper-parameter	Value
N_Gaus	1000
Edge_Dim	256
Epoch	500
Batch_Size	64
Learning_Rate	1**e**-4
Lambda1	5**e**-5
Lambda2	4**e**-5
Weight_Decay	3**e**-5

We divide the images in Dermnet dataset[1] into secondary classes based on their names and removed all classes with less than 10 images to ensure that the 5way-5shot task could be completed. Finally, we divide the dataset into 344 classes with 17,206 images. We select the classes with the highest number of images from each of the 23 broad classes as the base classes, and the remaining 321 classes are randomly divided into training set, validation set and test set in the ratio of 7:1.5:1.5.

A.3 Proofs of Key Theorems

A.3.1 Theorem 1

Let $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes a Gaussian distribution with $\mu < 1/2$, and let $\mathcal{B}(n, \lambda)$ denotes a Binomial distribution with $n \to +\infty$ and $\lambda \to 0$, where n is increasing while λ is decreasing. There exists a real constant m such that if $m = n\lambda$ and if we define:

$$
\begin{aligned}
f_1(x) &= \mathrm{KL}\left(\mathcal{N}(x, x(1-x)) \| \mathcal{N}\left(\mu, \sigma^2\right)\right) \\
f_2(x) &= \mathrm{KL}(\mathcal{N}(x, x(1-x)) \| \mathcal{N}(n\lambda, n\lambda(1-\lambda)) \\
f_2^* &= \min_x f_2(x), \text{ where } x \in (0,1)
\end{aligned}
$$

according to exist works Huang et al. (2020), we have that: $f_1(x)$ attains its minimum on the interval $(0,1)$ and $f_2(x) - f_2^*$ is bounded on the interval $(0, \sqrt{2}/2 - 1/2)$, with:

$$
x = m = \frac{1+l-\sqrt{1+l^2}}{2}, \text{ where } l = \frac{2\sigma^2}{1-2\mu}
$$

Suppose we are given a Gaussian distribution $\mathcal{N}\left(\tilde{\mu}_i, \tilde{\sigma}_i^2\right)$, whose parameter $\tilde{\mu}_i$ is specifically parameterized by the neural network that can guarantee that $\tilde{\mu}_i < 1/2$. By De Moivre Laplace theorem, we have that $\mathcal{N}\left(n\lambda_i, n\lambda_i\left(1-\lambda_i\right)\right)$ is a good approximation for $\mathcal{B}\left(n, \lambda_i\right)$. They are asymptotically equivalent as n increases. Let $m_i = n\lambda_i$, direct parameterization of both the infinite parameter n and the near-zero parameter $\lambda_{i,j}$ can be avoided by adopting a re-parametrization trick Kingma & Welling (2013). This trick draws samples from such Binomial distribution via its Gaussian proxy $\mathcal{N}\left(m_i, m_i\left(1-m_i\right)\right)$.

A.4 Theorem 2

Suppose we are given two Binomial distributions, $\mathcal{B}(n, \lambda) and \mathcal{B}\left(n, \lambda^0\right)$ with $n \to +\infty, \lambda^0 \to 0$ and $\lambda \to 0$, where n is increasing while λ and λ^0 are decreasing. There exists a real constant m and another real constant $m^{(0)}$, such that if $m = n\lambda and m^{(0)} = n\lambda^{(0)}$ and if $\lambda > \lambda^{(0)}$, we have:

[1] https://dermnet.com

$$\mathrm{KL}\left(\mathcal{B}(n,\lambda)\|\mathcal{B}\left(n,\lambda^{0}\right)<m\log\frac{m}{m^{(0)}}\right.$$
$$\left.+(1-m)\log\frac{1-m+m^{2}/2}{1-m^{(0)}+m^{(0)^{2}}/2}\right.$$

By Theorem 2 which is proofed in previous work Huang et al. (2020), we have a closed-form solution that is irrelevant to n for the ELBO.

A.5 Model Architectures of Self-distribution Calibration

Self-distribution Calibration (SDC) is a variant of Bayesian Distribution Calibration (BDC) used in ablation experiment. It attempts to fit the distribution of its class from the input data itself. The specific calculation flow chart of SDC is as figure 1.

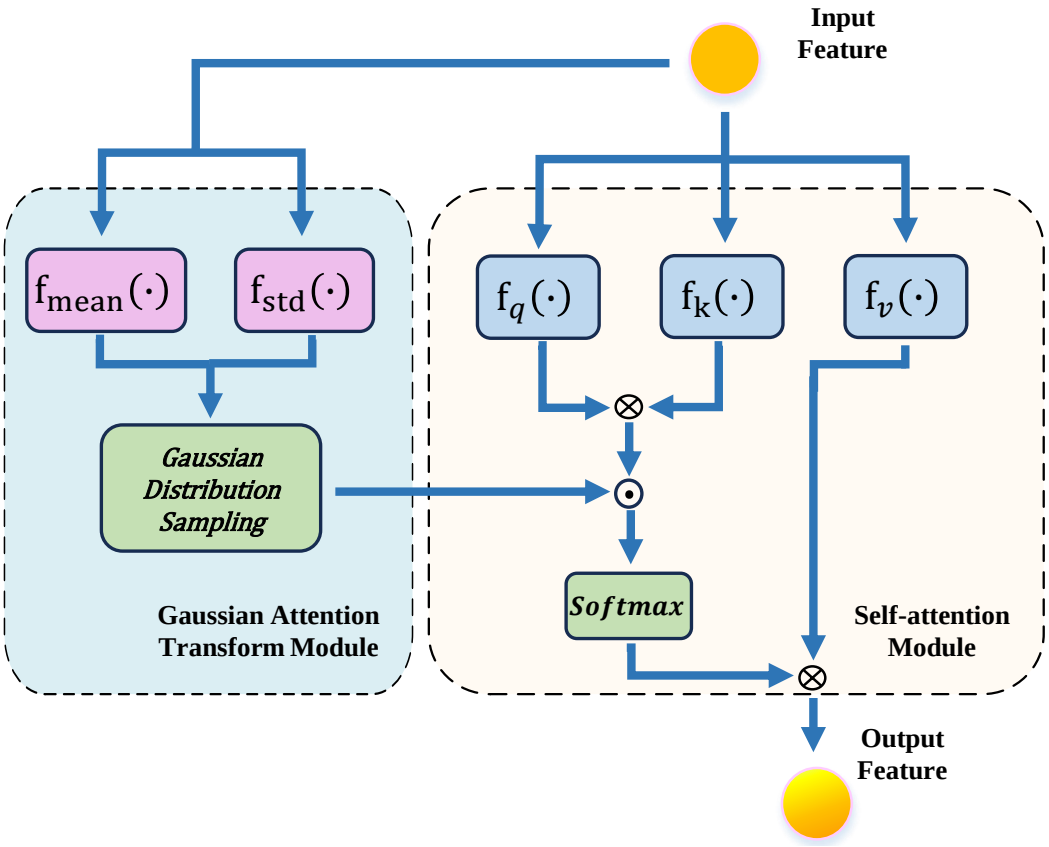

Figure 1: The specific calculation flow chart of SDC. Specifically, we use three f_q, f_k, f_v projection heads to project the input feature x into the query $\mathbf{q}=f_q(\mathbf{x})$, key $\mathbf{k}=f_k(\mathbf{x})$, value $\mathbf{v}=f_v(\mathbf{x})$. The attention matrix $\mathbf{Att}=\mathbf{q}\mathbf{k}^{\top}$. Meanwhile, we propose Gaussian attention transform operation to ensure that a sufficient number of output features can be generated. Specifically, we use two linear neural networks f_{mean}, f_{std} fitting a Gaussian distribution with mean and standard deviation. Then we generate a Gaussian random matrix $\mathbf{G}$ of the same dimension as the attention matrix based on this mean and standard deviation. The output feature $\hat{x}=\mathrm{softmax}\left(\mathbf{Att}\odot\frac{\mathbf{q}\mathbf{k}^{\top}}{\sqrt{d}}\right)\mathbf{v}$. Here operation $\odot$ denotes dot product and operation $\otimes$ denotes multiplication. With the self-attention module as well as the Gaussian attention transform module, we can obtain a sufficient number of output features.

A.6 PIPELINE GRAPH OF BAYESIAN DISTRIBUTION CALIBRATION MODEL

Figure 2 shows the overall of our Bayesian Distribution Calibration (BDC). The input of BDC is image data. The images first pass through the Backbone Network to generate features. Then the generated features pass through the Bayesian Relation Inference module to infer the relations between the input features and base node features. For the training phase, we generate a single fusion feature for each input feature for classification, and for the validation and testing phases, we generate a large number of fusion features for distribution calibration through multi-view Gaussian graph generation method.

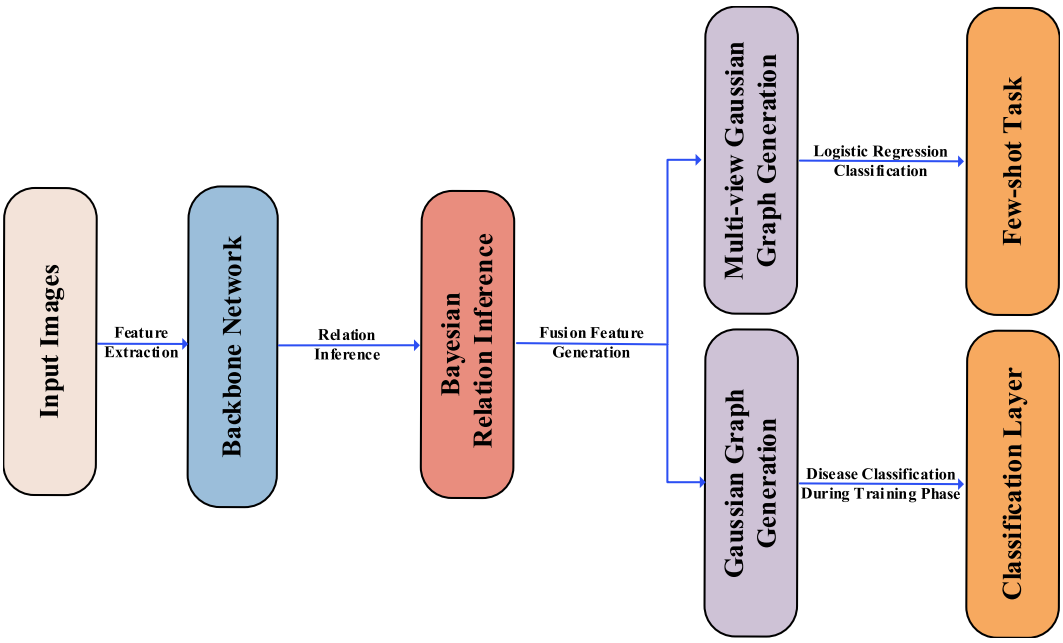

Figure 2: The pipeline graph of Bayesian Distribution Calibration.

.

A.7 SUPPLEMENTARY EXPERIMENT RESULTS AND FURTUER DISCUSSION

A.7.1 MORE INTUITIVE VISUAL ANALYSIS

In the proposed model, we use classes with a large amount of data as the base classes, which have the advantage that the features of the base classes can portray the overall distribution of the classes well. However, since there are also large differences between the images within each category (e.g., differences in the site of onset or even symptoms), this approach to base class selection results in the generated relation between target and base classes being less intuitively, and some relation ambiguity may occur. To solve this problem, we consider setting the base classes to some specific images to obtain more intuitive inter-class relations. Specifically, we selected the most representative image from each of the 23 classes with more obvious differences from other classes as the base classes, and infer the relation through these images. For any input image of the target class, the generated relation intensity graph is the relation between that image and the 23 base class images, which can be more intuitively expressed through visual analysis of the relations found by the Bayesian relation inference module. The visualization result is shown in Figure 3 and the accuracy result on 5-way 1-shot and 5-way 5-shot tasks is shown in table 3.

In the upper part of Figure 3, we show the results of the visual analysis using single image as a base classes. We average the multi-view Gaussian graphs and visualize the relationship between the target image and different base class images in the form of heatmaps, and select three images with strong positive and negative correlations respectively for further visualization analysis. It is worth noting that the target picture and the picture with serial number 16 belong to the same category in

the Dermnet dataset, which indicates that the proposed Bayesian distribution calibration method can effectively capture the potential relations between different objects. As can be seen from Table 3, using a single image as the base classes on the 5-way 1-shot and 5-way 5-shot tasks has a small decrease in accuracy (about 1% at 5-way 1-shot and 2% at 5-way 5-shot) compared to using a large number of images as base classes, which indicates that the proposed method does not require a large number of images for base classes. A small number of base class images can also provide a good distributional calibration for the target imags.

A.7.2 Visual Analysis on Robustness

In the proposed model, we use base classes that are strongly related to the task (e.g. for the Dermnet skin disease dataset, we use some of the dermatology classes in the dataset as base classes). In practice, there may not be enough data to be used as base classes, so the performance of the model in the absence of data strongly related to the task as base classes is important. To explore the robustness of the Bayesian Distribution Calibration model, we replace the base classes with animal data that are not relevant to skin diseases to explore the ability of the Bayesian relation inference component to infer potential relations between target class and base classes that is very different from the target class. The visualization result is shown in Figure 3 and the accuracy result on 5-way 1-shot and 5-way 5-shot tasks is shown in table 3.

The bottom half of Figure 3 shows the visualization result using animal image data as base classes. From the result, it is seen that the categories that have strong correlations with the target images are cat, horse and tiger. Intuitively, these three base categories have strong visual similarities, where the two categories of cat and tiger belong to the same family of felines, which can prove that the proposed Bayesian distribution calibration model is able to capture potential relations of the different categories. As can be seen from Table 3, the results using animal data as base classes still achieve high level of accuracy, which can prove that the proposed model is robust to the selection of base classes.

Table 3: Performance of Bayesian distribution calibration(BDC) on Dermnet dataset with various base classes

Method	5way1shot(%)	5way5shot(%)
BDC + Single image	49.56	68.04
BDC + Animal image	48.99	67.25
BDC(Ours)	**50.59**	**70.03**

A.7.3 Experiment Results on miniImageNet Dataset

The task of skin disease classification is famous and important. However, there are other datasets concerning ImageNet, Food-101 and so on. We further perform experiments on miniImageNet for few-shot classification tasks to validate the effectiveness of the proposed Bayesian relational inference model.

Table 4: Comparison of Bayesian distribution calibration(BDC) and baselines on miniImageNet dataset

Method	5way1shot(%)	5way5shot(%)
MN	49.02	70.11
PN	48.26	69.24
DC	68.01	82.45
tSF	68.84	84.38
GAP	69.35	83.85
BDC(Ours)	**70.08**	**84.52**

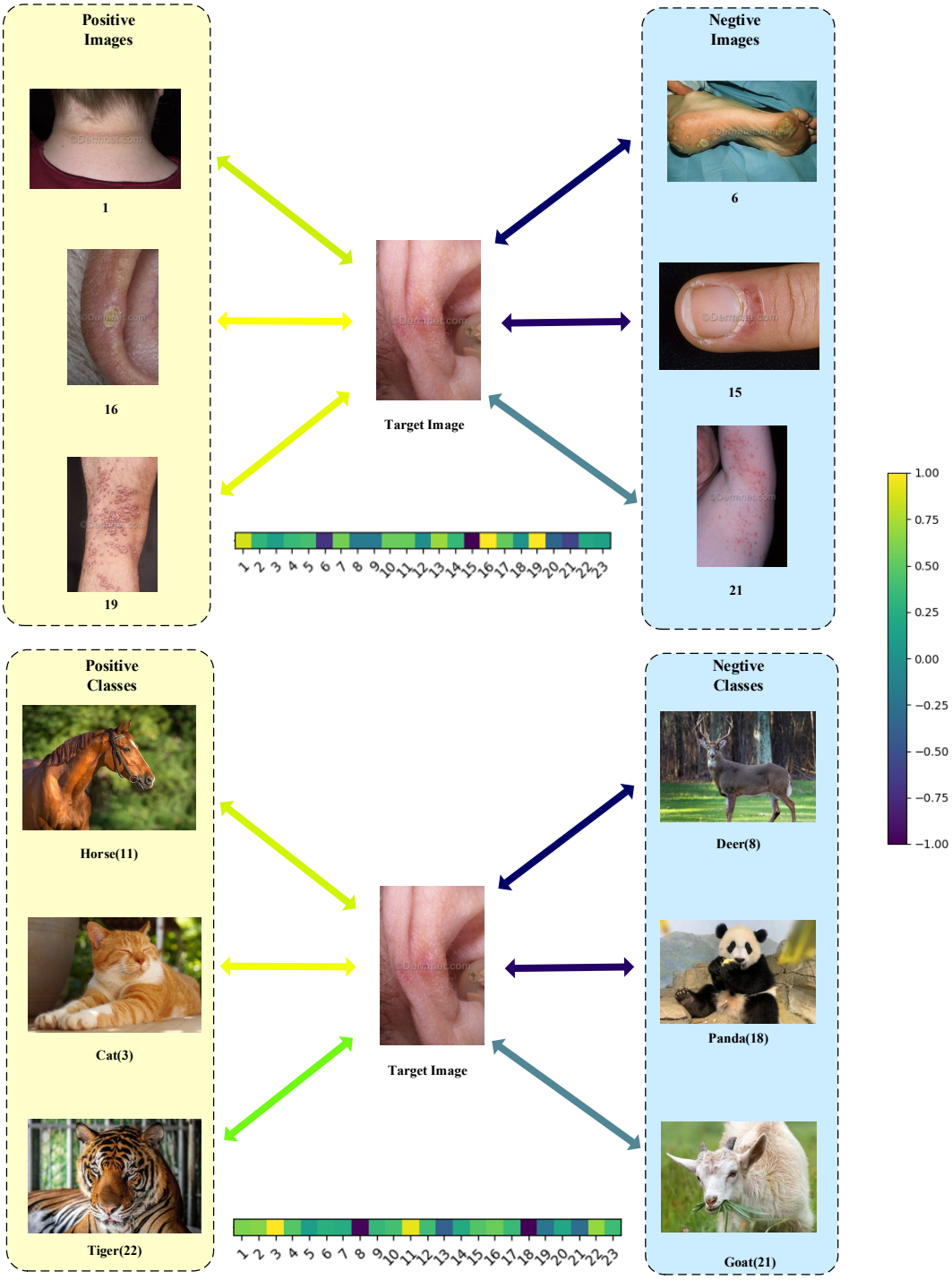

Figure 3: Schematic representation of the results of relation inference obtained using single image or animal data.

.

A.7.4 Comparison of Conventional Algorithms and Few-Shot Learning Algorithms

In few-shot learning algorithms, training a model using conventional algorithms can be difficult due to the large number of categories and the fact that the data for most of the categories is scarce. We divide the test set of the Dermnet dataset in a ratio of 8:2 into a new training and test set. Then we generate a conventional algorithm model by replacing the few-shot classification part of the Bayesian relational inference model(Multi-view Gaussian graph generation component and logistic regression classifier) with a linear classification head. We freeze the Bayesian relation inference module and fine-tune the classification head on the new training set and test it on the new test. We compare the accuracy of this algorithm with that of the few-shot learning algorithms on the 5-way 1-shot task. The experiment result is shown in Table 5.

Table 5: Comparison of few-shot algorithms and conventional algorithm on Dermnet dataset

Method	**Acc(%)**
MAML	44.05
PN	43.76
MN	44.23
DC	48.99
tSF	49.38
GAP	48.92
BDC + Conventional Algorithms	*43.50*
BDC(Ours)	50.59

Specifically, we adopt a three-layer artificial neural network as the linear classification head of the conventional algorithm, trained for 1500 epochs using the Adam optimizer with the learning rate of 0.0008. The result shows that the conventional algorithm's accuracy is similar to that of the early few-shot algorithms on the 5-way 1-shot task. It is worth noting that traing the conventional algorithm is time-consuming and overall performs less well than the few-shot algorithms.

REFERENCES

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Hengguan Huang, Fuzhao Xue, Hao Wang, and Ye Wang. Deep graph random process for relational-thinking-based speech recognition. In *International Conference on Machine Learning*, pp. 4531–4541. PMLR, 2020.

Suhyun Kang, Duhun Hwang, Moonjung Eo, Taesup Kim, and Wonjong Rhee. Meta-learning with a geometry-adaptive preconditioner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16080–16090, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jinxiang Lai, Siqian Yang, Wenlong Liu, Yi Zeng, Zhongyi Huang, Wenlong Wu, Jun Liu, Bin-Bin Gao, and Chengjie Wang. tsf: Transformer-based semantic filter for few-shot learning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 1–19, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20044-1.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395*, 2021.