Gestalt-Guided Image Understanding for Few-Shot Learning

Kun Song, Yuchen Wu, Jiansheng Chen, Tianyu Hu, and Huimin Ma[✉]

University of Science and Technology Beijing, Beijing, China
{songkun,yuchen.wu}@xs.ustb.edu.cn, {jschen,Tianyu,mhmpub}@ustb.edu.cn

Abstract. Due to the scarcity of available data, deep learning does not perform well on few-shot learning tasks. However, human can quickly learn the feature of a new category from very few samples. Nevertheless, previous work has rarely considered how to mimic human cognitive behavior and apply it to few-shot learning. This paper introduces Gestalt psychology to few-shot learning and proposes Gestalt-Guided Image Understanding, a plug-and-play method called **GGIU**. Referring to the principle of totality and the law of closure in Gestalt psychology, we design Totality-Guided Image Understanding and Closure-Guided Image Understanding to extract image features. After that, a feature estimation module is used to estimate the accurate features of images. Extensive experiments demonstrate that our method can improve the performance of existing models effectively and flexibly without retraining or fine-tuning. Our code is released on https://github.com/skingorz/GGIU.

1 Introduction

In recent years, deep learning has shown surprising performance in various fields. Nevertheless, deep learning often relies on large amounts of training data. More and more pre-trained models are based on large-scale data. For example, CLIP [1] is trained on 400 million image-text pairs. However, a large amount of data come with extra costs in deep learning procedures, such as collection, annotation, and training. In addition, many kinds of data, such as medical image data, requires specialized knowledge to annotate. Data for some rare scenes are hard to obtain, such as car accidents. Therefore, there is a growing interest in training a better model using fewer data. Motivated by this, Few-Shot Learning (FSL) [2,3] is proposed to solve the problem of learning from small amounts of data.

The most significant obstacle to few-shot learning is the lack of data. In order to address this obstacle, existing few-shot learning approaches mainly employ metric learning, such as PN [4], and meta-learning, such as MAML [5]. Regardless of the technique, the ultimate goal is to extract more robust features for novel classes. Previous researches mainly focus on two aspects: designing a more robust feature extractor to represent the image feature better, such as meta-baseline [6], and using more dense features, such as DeepEMD [7]. However, few people consider how to mimic human learning patterns to enhance the effectiveness of few-shot learning.
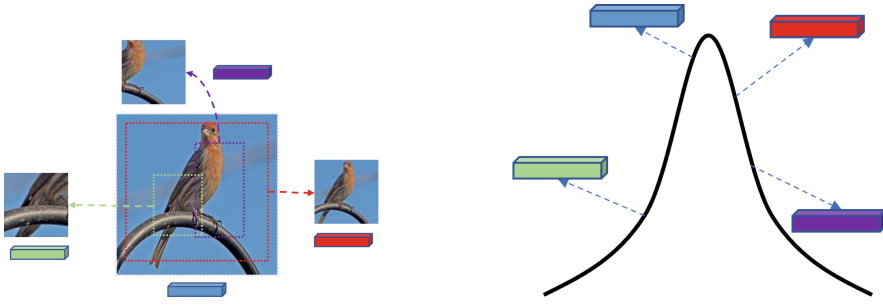
Fig. 1. Describe an image using a multivariate Gaussian distribution.

Modern psychology has extensively studied the mechanisms of human cognition. Gestalt psychology is one of them. On the one hand, Gestalt psychology states that conscious experience must be considered globally, which is the principle of totality of gestaltism. For example, as shown in the left of Fig. 1, the whole picture shows a bird standing on a tree branch. Its feature can represent the image. Meanwhile, given any patch of the image, human can easily determine that it is a part of a bird. These can be explained by the law of closure in Gestalt psychology: When parts of a whole picture are missing, our perception fills in the visual gap. Therefore, the feature of patches can also represent the image.

Motivated by Gestalt psychology, we imitate human learning patterns to redesign the image understanding process and apply it to few-shot learning. In this paper, we innovatively describe the image as a data distribution to represent the principle of totality and the law of closure in Gestalt psychology. We assume that image can be represented by a corresponding multivariate Gaussian distribution. As shown in the right of Fig. 1, the feature of the image can be considered as a sample of the potential distribution. Likewise, the feature of the largest patch, the image itself, is also a sample of the distribution. Then we design a feature estimation module with reference to Kalman filter to estimate the features of images accurately.

The main contributions of this paper are:

1. We introduce Gestalt psychology into the process of image understanding and propose a plug-and-play method without retraining or fine-tuning, called **GGIU**.
2. We innovatively propose to use multivariate Gaussian distribution to describe the image and design a feature estimation module with reference to Kalman filter to estimate image feature accurately.
3. We conduct extensive experiments to demonstrate the applicability of our method to a variety of different few-shot classification tasks. The experiment results demonstrate the robustness and scalability of our method.

2   Related Work

2.1   Few-Shot Learning

Most of the existing methods for few-shot learning is based on the meta-learning [8] framework. The motivation of meta-learning is learning to learn [9]. The network is trained on a set of meta-tasks during the training process to gain the ability to adapt itself to different tasks. The meta-learning methods for few-shot learning are mainly divided into two categories, optimization-based and metric-based.

Koch [10] apply metric learning to few-shot learning for the first time. They proposed to apply Siamese Neural Network to one-shot learning. A pair of convolutional neural networks with shared weights is used to extract the embedding features of each class separately. When inferring the category of unknown data, the unlabeled data and the training set samples are paired. The Manhattan Distance between unlabeled data and training data is calculated as the similarity, and the category with the highest similarity is used as the prediction of the samples. Matching Network [3] first conducts experiments on $mini$ImageNet for few-shot learning. It proposes an attention module that uses cosine distance to determine the similarity between the target object and each category and uses the similarity for the final classification. Prototypical Network [4] proposed the concept of category prototypes. Prototypical Network takes a class's prototype to be the mean of its support set in the embedding space. The similarity between the unknown data and each category's prototypes are measured, and the most similar category is selected as the final classification result. Satorras [11] uses graphical convolutional networks to transfer information between support and query sets and extended prototypical networks and matching networks to non-euclidean spaces to assist few-shot learning. DeepEmd [7] adopt the Earth Mover's Distance (EMD) as a metric to compute a structural distance between dense image representations to determine image relevance. COSOC [12] extracts image foregrounds using contrast learning to optimize the category prototypes. Yang [13] takes the perspective of distribution estimation to rectify the category prototype. SNE [14] encodes the latent distribution transferring from the already-known classes to the novel classes by label propagation and self-supervised learning. CSS [15] propose conditional self-supervised learning with a 3-stage training pipeline. CSEI [16] proposed an Erasing-then-Inpainting method to augment the data while training, which needs to retrain the model. AA [17] expands the novel data by adding extra "related base" data to few novel ones and fine-tunes the model.

2.2   Gestalt Psychology

Gestalt psychology is a psychology school that emerged in Austria and Germany in the early twentieth century. Gestalt principles, proximity, similarity, figure-ground, continuity, closure, and connection describe how humans perceive visuals in connection with different objects and environments. In this section, we mainly introduce the principle of totality and the law of closure.

Fig. 2. Although we see incomplete figures, our brain can easily complete them and regard them as rectangles (left) and circles (right).

Principle of Totality. The principle of totality points out that conscious experience must be considered globally. Wertheimer [18] described holism as fundamental to Gestalt psychology. Kohler [19] thinks: "In psychology, instead of being the sum of parts existing independently, wholes give their parts specific functions or properties that can only be defined in relation to the whole in question." Thus, the maxim that the whole is more than the sum of its parts is not a precise description of the Gestaltist [18].

Law of Closure. Gestalt psychologists held the view that humans often consider objects as complete rather than focusing on the gaps they may have. [20] For example, a circle has a good Gestalt in terms of completeness. However, we may also consider an incomplete circle as a complete one. This tendency to complete shapes and figures is called closure. [21] The law of closure states that even incomplete objects, such as forms, characters, and pictures, are seen as complete by people. In particular, when part of the entire picture is missing, our perception fills in the visual gaps. For example, as shown in Fig. 2, despite the incomplete shape, we still perceive a rectangle and a circle. If the law of closure did not exist, the image would depict different lines with different lengths, rotations, and curvatures. However, because of the law of closure, we perceptually combine the lines into whole shapes [22].

3   Method

3.1   Problem Definition

The dataset for few-shot learning consists of training set $\mathcal{D}_B$ and testing set $\mathcal{D}_V$ with no shared classes. We train a feature extractor $f_\theta(\cdot)$ on $\mathcal{D}_B$ containing lots of labeled data. During the evaluation, many N-way K-shot Q-query tasks $\mathcal{T} = \{(\mathcal{S}, \mathcal{Q})\}$ are constructed from $\mathcal{D}_V$. Each task contains a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{K \times N}$ and a query set $\mathcal{Q}$. Firstly, N classes in $\mathcal{D}_V$ are randomly sampled for $\mathcal{S}$ and $\mathcal{Q}$. Then we calculate a classifier for N-way classification for each task based on $f_\theta(\cdot)$ and $\mathcal{S}$. At last, we calculate the feature of the query image $x \in \mathcal{Q}$ and classify it into one of the N class.

3.2 Metric-Based Few-Shot Learning Pipeline

Calculating a better representation for all classes in metric-based few-shot learning is critical. Under most circumstances, the features of all images in each category of the support set are calculated separately. The mean of features is used as the category representation, the prototype.

$$P_n = \frac{1}{K} \sum_{(x_i, y_i) \in \mathcal{S}_n} f_\theta(x_i) \tag{1}$$

In Eq. 1, P_n represents the prototype of the n-th category and $\mathcal{S}_n$ represents the set of all images of the n-th class. When an image $x \in \mathcal{Q}$ and a distance function $d(,)$ are given, the feature extractor $f_\theta(\cdot)$ is used to calculate the feature of x. Then we calculate the distance between $f_\theta(x)$ and the prototypes in the embedding space. After that, a softmax function (Eq. 2) is used to calculate a distribution for classification.

$$p_\theta(y = n|x) = \frac{exp(-d(f_\theta(x), P_n))}{\sum_{i=1}^{N} exp(-d(f_\theta(x), P_i))} \tag{2}$$

Finally, the model is optimized by minimizing the negative log-likelihood $L(\theta) = -\log p_\theta(y = n|x)$, i.e.,

$$NLL(\theta) = d(f_\theta(x), P_n) + \log \sum_{i=1}^{k} exp(-d(f_\theta(x), P_i)) \tag{3}$$

3.3 Gestalt Guide Image Understanding

Previous works mainly calculate prototypes with the features of images. Nevertheless, due to the limit of data volume, the image feature does not represent the information of the class well. In this section, inspired by Gestalt psychology, we reconceptualize images in terms of the principle of totality and the law of closure separately.

Given one image I, many patches of different sizes can be cropped from it. As shown in the right of Fig. 1, we assume that all patches follow a potential multivariate Gaussian distribution. All the patches can be regarded as samples of this distribution. Likewise, the largest patch of this image, the image itself, is also a sample of the distribution. Therefore, this potential Gaussian distribution can describe the image. The above can be expressed as follows: given a feature extractor $f_\theta(\cdot)$, for any patch p cropped from image I, its feature follows a multivariate Gaussian distribution, i.e., $f_\theta(p) \in \mathbb{R}^D$ and $f_\theta(p) \sim N_I(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I^2})$. Next, $\boldsymbol{\mu_I}$ can represent the feature of I. Finally, we estimate the image feature by the principle of totality and the law of closure.

Totality-Guided Image Understanding. The existing image understanding processes in few-shot learning are most from the totality of the image. The image can be considered as a sample of the potential multivariate Gaussian distribution, i.e., $f_\theta(I) \sim N_I(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I^2})$. Therefore, $N_I(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I^2})$ can be estimated by $f_\theta(I)$ (Eq. 4). $\boldsymbol{\mu_t}$ represents the estimate guided by the principle of totality.

$$\boldsymbol{\mu_t} = \hat{\boldsymbol{\mu_I}} = f_\theta(I) \tag{4}$$

Closure-Guided Image Understanding. Guided by the law of closure, we randomly crop patches from images as the sample of the potential distribution of image. For any patch $p \in I$, $f_\theta(p) \in \mathbb{R}^D$, $f_\theta(p) \sim N_I(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I^2})$. The joint probability density function of the feature of the i-th patch is shown as Eq. 5.

$$p_{N_I}(f_\theta(p_i)) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(f_\theta(p_i) - \boldsymbol{\mu_I}\right)^T \boldsymbol{\Sigma}^{-1}\left(f_\theta(p_i) - \boldsymbol{\mu_I}\right)\right) \tag{5}$$

The log-likelihood function is:

$$\ell_{N_I}(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I}; p_1, \ldots, p_M) = -\frac{MD}{2}log(2\pi) - \frac{M}{2}log(|\boldsymbol{\Sigma_I}|) \\ -\frac{1}{2}\sum_{j=1}^{M}(p_i - \boldsymbol{\mu_I})^T \boldsymbol{\Sigma_I}^{-1}(p_i - \boldsymbol{\mu_I}) \tag{6}$$

Solve the following maximization problem

$$\max_{\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I}} \ell_{N_I}(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I}; p_1, \ldots, p_M) \tag{7}$$

Then we have

$$\boldsymbol{\mu_c} = \hat{\boldsymbol{\mu_I}} = \frac{1}{M}\sum_{i=1}^{M} f_\theta(p_i) \tag{8}$$

$\boldsymbol{\mu_c}$ represents the estimation guided by the law of closure.

3.4   Feature Estimation

We regard the process of estimating image feature guided by the totality and closure as two different observers following multivariate Gaussian distribution: O_t and O_c. The observations of O_t, O_c are $\boldsymbol{\mu_t} \in \mathbb{R}^{D\times 1}$, $\boldsymbol{\mu_c} \in \mathbb{R}^{D\times 1}$ and their random errors are $\boldsymbol{e_t}$, $\boldsymbol{e_c}$ respectively, where $\boldsymbol{e_t} \in \mathbb{R}^{D\times 1}$ and $\boldsymbol{e_c} \in \mathbb{R}^{D\times 1}$. $\boldsymbol{e_t}$ and $\boldsymbol{e_c}$ follow multivariate Gaussian distribution, i.e., $\boldsymbol{e_t} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_t^2})$, $\boldsymbol{e_c} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_c^2})$, where $\boldsymbol{\Sigma_t} \in \mathbb{R}^{D\times D}$ and $\boldsymbol{\Sigma_c} \in \mathbb{R}^{D\times D}$. In this section, we use Kalman filter to estimate image features $\boldsymbol{f} \in \mathbb{R}^{D\times 1}$.

For O_t and O_c, we have

$$f = \mu_t + e_t \tag{9}$$

$$\mu_c = f + e_c \tag{10}$$

The prior estimates of f under O_t and O_c are

$$\hat{f_t^-} = \mu_t \tag{11}$$

$$\hat{f_c^-} = \mu_c \tag{12}$$

The feature f can be estimate by the prior estimates under O_t and O_c, we have

$$\hat{f} = \hat{f_c^-} + \lambda(\hat{f_t^-} - \hat{f_c^-}) \tag{13}$$

i.e.,

$$\hat{f} = \mu_c + \lambda(\mu_t - \mu_c) \tag{14}$$

$\lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_D)$, where λ_i is a diagonal matrix, ranging from 0 to I. The error between $\hat{f}$ and f is

$$e = f - \hat{f} \tag{15}$$

The error e follows a multivariate Gaussian distribution, i.e., $e \sim N(\mathbf{0}, \mathbf{\Sigma}_e)$. where

$$\mathbf{\Sigma}_e = E(ee^T) \tag{16}$$

$$= E[(f - \hat{f})(f - \hat{f})^T] \tag{17}$$

$$= E[(\lambda e^- - (I - \lambda e_c))(\lambda e^- - (I - \lambda e_c))^T] \tag{18}$$

e^- represents the prior estimation of e, Since e^- and e_c are independent of each other, we have $E(e_c e^-) = E(e^-)E(e_c) = 0$. Therefore, we have

$$\mathbf{\Sigma}_e = \lambda E(e^- e^{-T})\lambda^T + (I - \lambda)E(e_c e_c^T)(I - \lambda)^T \tag{19}$$

$$= \lambda \mathbf{\Sigma}^- \lambda^T + (I - \lambda)\mathbf{\Sigma}_c(I - \lambda)^T \tag{20}$$

In order to estimate f accurately, we have to minimize e, i.e.,

$$\min_{\lambda} tr(\mathbf{\Sigma}_e) = \min_{\lambda} [tr(\lambda \mathbf{\Sigma}^- \lambda^T) - 2tr(\lambda \mathbf{\Sigma}_c) + tr(\lambda \mathbf{\Sigma}_c \lambda^T)] \tag{21}$$

We need to solve this equation:

$$\frac{\partial \mathbf{\Sigma}_e}{\partial \lambda} = 0 \tag{22}$$

We have

$$\lambda = (\mathbf{\Sigma}^- \mathbf{\Sigma}_c^{-1} + I)^{-1} \tag{23}$$

where

$$\boldsymbol{\Sigma}^- = E(\boldsymbol{e}^-\boldsymbol{e}^{-T}) \tag{24}$$

$$= E[(\boldsymbol{f} - \hat{\boldsymbol{f}})(\boldsymbol{f} - \hat{\boldsymbol{f}})^T] \tag{25}$$

$$= E[(\boldsymbol{\mu_t} + \boldsymbol{e_t} - \boldsymbol{\mu_t})(\boldsymbol{\mu_t} + \boldsymbol{e_t} - \boldsymbol{\mu_t})^T] \tag{26}$$

$$= \boldsymbol{\Sigma_t} \tag{27}$$

Therefore

$$\boldsymbol{\lambda} = (\boldsymbol{\Sigma_t}\boldsymbol{\Sigma_c}^{-1} + \boldsymbol{I})^{-1} \tag{28}$$

Although the error covariance matrix $\boldsymbol{\Sigma_t}$ and $\boldsymbol{\Sigma_c}$ of the observers O_t and O_c cannot be calculated, the relationship between $\boldsymbol{\lambda}$ and the number of patches can still be estimated, which can assist us in choosing the parameter. When the number of patches is large enough, the Closure-Guided image understanding can estimate the image features accurately. At this time, $\boldsymbol{\Sigma_c}$ is close to $\boldsymbol{0}$. According to Eq. 28, $\boldsymbol{\lambda}$ is close to $\boldsymbol{0}$. As the number of patches decreases, $\boldsymbol{\Sigma_c}$ gradually increases, and $\boldsymbol{\lambda}$ is close to $\boldsymbol{I}$.

3.5 The Overview of Our Approach
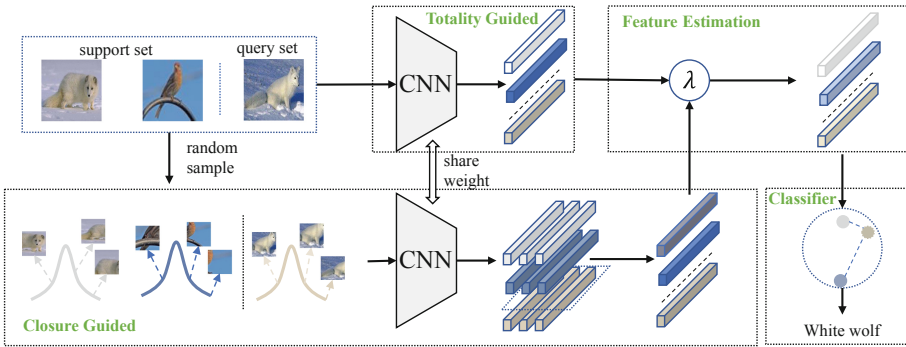
Fig. 3. There are two branches in our method. Totality-Guided module extracts the feature of the whole image. Guided by Closure-Guided module, image features are estimated from incomplete images. After that, we use feature estimation module to fuse the feature calculated by the Totality-Guided and Closure-Guided module. Finally, we classify the query image according to the image feature.

Our pipeline on 2-way 1-shot task is illustrated in Fig. 3. Given a few-shot learning task, firstly, we feed the image x into the feature extractor $f_\theta(\cdot)$ to extract the features. The prototype guided by the principle of totality $\boldsymbol{P_n^t}$ can be calculated by Eq. 29. Query features estimated by the principle of totality are the whole image feature extracted by the feature extractor (Eq. 4).

$$P_n^t = \frac{1}{K} \sum_{(x_i, y_i) \in \mathcal{S}_n} f_\theta(x_i) \tag{29}$$

Meanwhile, guided by the law of closure, we randomly crop M patches from each image and feed them into a feature extractor with shared weights to calculate feature $f_\theta(p)$. For the convenience of calculating the categories prototypes, as shown in Eq. 30, we use all patches in the same category to calculate the prototype guided by the law of closure. Query features estimated by the law of closure can be calculated by Eq. 8.

$$P_n^c = \frac{1}{K \times M} \sum_{(x_i, y_i) \in \mathcal{S}_n} \sum_{p_i \in x_i} f_\theta(p_i) \tag{30}$$

After calculating the category prototypes P_n^t and P_n^c, they are fed into the feature estimate module to calculate the category prototype P_n (Eq. 31). Query features can be re-estimated by Eq. 14. Then the distances between the query feature and the category prototypes are calculated and the query set is classified according to Eq. 2.

$$P_n = \lambda P_n^t + (I - \lambda) P_n^c \tag{31}$$

4 Experiment

4.1 Datasets

We test our method on *mini*ImageNet [3] and Caltech-UCSD Birds 200-2011 (CUB200) [23], which are widely used in few-shot learning.

miniImageNet is a subset of ILSVRC-2012 [24]. It contains 60,000 images in 100 categories, with 600 images in each category. Among them, 64 classes are used as the training set, 16 classes as the validation set, and 20 as the testing set.

Table 1. Results of the performance of different methods on *mini*ImageNet before and after adding GGIU The reported accuracy is 95% confidence interval.

Method	5-way 1-shot (%)	5-way 5-shot (%)
PN	61.59 ± 0.54	76.75 ± 0.46
PN+GGIU	64.34 ± 0.53 (↑ **2.75**)	79.49 ± 0.41 (↑ **2.74**)
CC	63.11 ± 0.74	80.43 ± 0.31
CC+GGIU	65.72 ± 0.77 (↑ **2.61**)	82.55 ± 0.29 (↑ **2.12**)
CL	63.74 ± 0.59	79.33 ± 0.31
CL+GGIU	65.50 ± 0.45 (↑ **1.76**)	80.76 ± 0.39 (↑ **1.43**)
CLIP	88.21 ± 0.33	97.47 ± 0.08
CLIP+GGIU	89.31 ± 0.33 (↑ **1.10**)	97.71 ± 0.06 (↑ **0.24**)

CUB200 contains 11788 images of birds in 200 species, which is widely used for fine-grained classification. Following the previous work [25], we split the categories into 130, 20, 50 for training, validation and testing.

4.2   Implementation Details

Since we propose a test-time feature estimation approach, we need to reproduce the performance of existing methods to validate our approach's effectiveness. Therefore, following Luo [12], we reproduce PN [4], CC [26], and CL [27]. The backbone we use in this paper is ResNet-12 [28], which is widely used in few-shot learning. We implement our method using PyTorch and test on an NVIDIA 3090 GPU. Since the authors do not provide a configuration file for the CUB200 dataset, we use the same configuration file as $mini$ImageNet. In the test phase, we randomly sampled five groups of test tasks, and each group of tasks contained 2000 episodes. Then five patches from each image are randomly cropped to rectify for prototypes and features with $\boldsymbol{\lambda} = diag(0.5, 0.5, \ldots, 0.5)$. The size of the patches is range from 0.08 to 1.0 of that of the original image.

4.3   Experiment Results

Results on In-Domain Data.   Table 1 shows our performance on $mini$Image Net. Our method can effectively improve the performance based on existing methods for different tasks. We improve the performance on PN, CC, and CL by 2.75%, 2.61%, and 1.76%, separately.

We also test the performance on CLIP [1] to explore the performance of our method on the model pre-trained on large-scale data. We use the ViT-B/32 model published by OpenAI as a feature extractor and use PN for classification. Surprising, on such a high baseline, our method can still improve the 1-shot task by 1.10% and the 5-shot by 0.23%. The experiment results also illustrate that even with such large-scale training data, **GGIU** can still estimate more accurate image features.

Similarly, to test the effectiveness of our method on fine-grained classification, we also test the performance of our method on CUB200. As shown in Table 2, our method also significantly improves the performance based on existing methods of fine-grained classification.

Table 2. Results of the performance of different methods on CUB200 before and after adding GGIU. The reported accuracy is 95% confidence interval.

Method	5-way 1-shot (%)	5-way 5-shot (%)
PN	76.13 ± 0.21	88.06 ± 0.09
PN+GGIU	78.79 ± 0.24 (↑ **2.66**)	89.69 ± 0.17 (↑ **1.63**)
CC	70.57 ± 0.35	86.65 ± 0.16
CC+GGIU	72.60 ± 0.29 (↑ **2.03**)	87.90 ± 0.27 (↑ **1.25**)
CL	72.34 ± 0.48	85.93 ± 0.25
CL+GGIU	73.64 ± 0.46 (↑ **1.30**)	87.17 ± 0.25 (↑ **1.24**)

Table 3. Results of the performance reported by other methods with the performance of ours on *mini*ImageNet. The reported accuracy is 95% confidence interval. * represents the results reported by the original paper and † represents the results that we implement.

Method	Backbone	5-way 1-shot (%)	5-way 5-shot (%)
PN* [4]	Conv-4	49.42 ± 0.78	**68.20 ± 0.66**
PN†	Conv-4	50.15 ± 0.44	65.19 ± 0.51
DC* [13]	Conv-4	**54.62 ± 0.64**	–
Spot and Learn* [29]	Conv-4	51.03 ± 0.78	67.96 ± 0.71
PN+GGIU†	Conv-4	52.55 ± 0.52	67.36 ± 0.55
PN†	ResNet-12	61.59 ± 0.54	76.75 ± 0.46
CC* [26]	ResNet-12	55.45 ± 0.89	70.13 ± 0.68
CC†	ResNet-12	63.11 ± 0.74	80.43 ± 0.31
PN+TRAML* [30]	ResNet-12	60.31 ± 0.48	77.94 ± 0.57
PN+CL* [27]	ResNet-12	59.54 ± 0.47	74.46 ± 0.52
PN+CL†	ResNet-12	63.74 ± 0.59	79.33 ± 0.31
DC*	ResNet-18	61.50 ± 0.47	–
AA* [17]	ResNet-18	58.84 ± 0.77	80.35 ± 0.73
PN+GGIU†	ResNet-12	64.34 ± 0.53	79.49 ± 0.41
CC+GGIU†	ResNet-12	**65.72 ± 0.77**	**82.55 ± 0.29**
PN+CL+GGIU†	ResNet-12	65.50 ± 0.45	80.76 ± 0.39
CLIP†	ViT-B/32	88.21 ± 0.33	97.47 ± 0.08
CLIP+GGIU†	ViT-B/32	**89.31 ± 0.33**	**97.71 ± 0.06**

In addition, we compare the performance of our method with existing methods in Table 3.

Results on Cross-Domain Data. To validate the effectiveness and robustness of our approach, we conduct experiments on cross-domain tasks. We test the cross-domain performance on *mini*ImageNet and CUB200: Table 4 shows the model trained on *mini*ImageNet and tested on CUB200; Table 5 shows the

Table 4. Results of the performance of different methods trained on *mini*ImageNet and tested on CUB200 before and after adding GGIU. The reported accuracy is 95% confidence interval.

Method	5-way 1-shot (%)	5-way 5-shot (%)
PN	40.47 ± 0.21	56.14 ± 0.20
PN+GGIU	42.61 ± 0.49 (↑ **2.14**)	58.95 ± 0.49 (↑ **2.81**)
CC	43.56 ± 0.47	61.51 ± 0.39
CC+GGIU	45.88 ± 0.48 (↑ **2.32**)	64.77 ± 0.27 (↑ **3.26**)
CL	38.65 ± 0.44	52.36 ± 0.35
CL+GGIU	39.87 ± 0.31 (↑ **1.22**)	53.74 ± 0.21 (↑ **1.38**)

Table 5. Results of the performance of different methods trained on CUB200 and tested on miniImageNet. The reported accuracy is 95% confidence interval.

Method	5-way 1-shot (%)	5-way 5-shot (%)
PN	40.24 ± 0.35	55.47 ± 0.47
PN+GGIU	43.17 ± 0.57 ($\uparrow$ **2.93**)	58.12 ± 0.49 ($\uparrow$ **2.65**)
CC	43.54 ± 0.52	60.40 ± 0.39
CC+GGIU	44.27 ± 0.42 ($\uparrow$ **0.73**)	60.94 ± 0.42 ($\uparrow$ **0.54**)
CL	44.47 ± 0.56	61.84 ± 0.44
CL+GGIU	46.42 ± 0.68 ($\uparrow$ **1.95**)	63.89 ± 0.45 ($\uparrow$ **2.05**)

results of the model trained on CUB200 and tested on *mini*ImageNet. It can be seen that our method has adequate performance improvement on the cross-domain task of few-shot learning.

4.4   Result Analysis

Ablation Study. This section performs ablation experiments on *mini*ImageNet to explore the performance impact of feature rectification on the support set and query set.

As shown in Table 6, compared with only using **GGIU** to estimate query features, the performance of estimation support features is higher. For the 5-way 5-shot task, when **GGIU** is used to estimate query features, the performance is improved more. This can be explained as follows: The more support set samples, the more accurate the representation of the category prototype is. At this time, the accuracy of query features is the bottleneck of classification performance. Similarly, when the support set samples are few, the category prototypes calculated by the support set are inaccurate. At this moment, accurate category prototypes can significantly improve classification performance.

The Influence of the Fusion Parameters In our method, the fusion parameter $\boldsymbol{\lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_D)$ is essential. This section explores the influence of $\boldsymbol{\lambda}$. We conduct three sets of experiments on PN for the different number of patches, 1, 5, and 10, respectively. As shown in Fig. 4, when $\lambda_i = 1$, image features $\hat{\boldsymbol{f}}$ are solely estimated by totality-guided image understanding (Eq. 14). The performance at this point is the baseline performance. As λ_i decreases, the influence of closure-guided image understanding increases, and the estimated image feature $\hat{\boldsymbol{f}}$ can represent the image better. So the performance gradually improves until an equilibrium point is reached. After the highest performance, continuing to decrease λ_i leads to a gradual decrease in model performance. It is worth noting that the performance will improve with a large enough number of patches even if $\lambda_i = 0$. It can be concluded that when the number of patches is large enough, closure-guided image understanding can well estimate image features. It can also be seen that the value of λ corresponding to the highest

Table 6. The results of ablation experiments. support with ✓ means adding GGIU while calculating prototypes; query with ✓ means adding GGIU while computing the feature of query images.

| Support	Query	5-way 1-shot (%)	5-way 5-shot (%)
		61.59	76.75
✓		62.80	77.01
	✓	62.50	77.90
✓	✓	64.34	79.49

performance decreases as the number of patches increases. The more patches, the more accurate the estimation guided by the law of closure. What's more, when the number of patches is 1, 5, and 10, the best λ equals 0.7, 0.5, and 0.4, respectively. The above results are consistent with the deduction in Sect. 3.4.

The Influence of the Number of Patches. The number of patches is a very important hyper-parameter, and the inference efficiency might be low if this number is too big. This section explores how the number of patches influences the performance. As shown in Fig. 5, on *mini*ImageNet, we perform a 5-way 1-shot experiment on PN with $\lambda_i = 0.5$. Suppose only one patch is cropped to estimate the closure feature, in which case, it can be seen that it will also improve the performance substantially. However, as the number increases, the rate of increase in accuracy gradually slows down, which demonstrates that too many patches might lead to a marginal effect on the correction of the distribution, especially when the patches almost cover the whole image. Therefore, too many patches do not significantly improve the model performance.
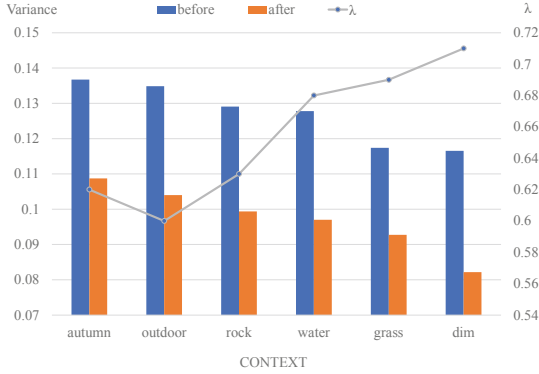
The Relationship Between Intra-class Variations and λ. In order to analyse the relationship between the optimal λ and intra-class variations, we
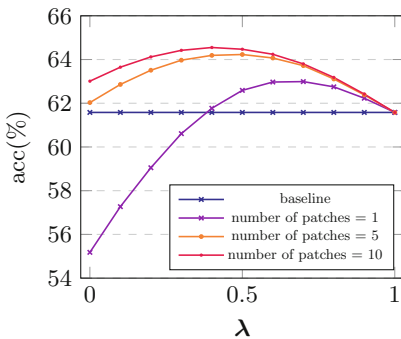
Fig. 4. Relationship between λ and performance
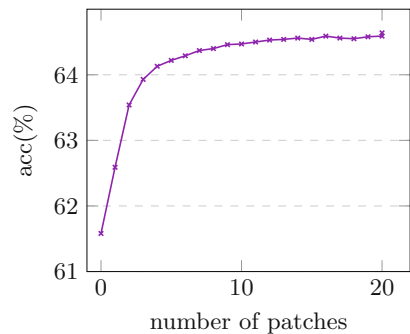
Fig. 5. The influence of the number of patches on performance

Fig. 6. The relationship between intra-class variations and $\boldsymbol{\lambda}$

conducted experiments on the NICO [31], which contains images labeled with category and context. We searched for the optimal $\boldsymbol{\lambda}$ in different contexts and calculated the intra-class variance in each context before and after using our method. As shown in Fig. 6, with the decrease of the intra-class variance, the optimal $\boldsymbol{\lambda}$ shows an increasing trend. Moreover, our method can significantly reduce the intra-class variance.

5   Conclusion

In this paper, we reformulate image features from the perspective of multivariate Gaussian distributions. We introduce Gestalt psychology into the process of image understanding to estimate more accurate image features. The Gestalt-guided image understanding consists of two modules: Totality-guided image understanding and Closure-guidied image understanding. Then we fed the features obtained from the above two modules into the feature estimation module and estimate image features accurately. We conduct many experiments on $mini$ImageNet and CUB200 for coarse-grained, fine-grained, and cross-domain few-shot image classification. The results demonstrate the effectiveness of **GGIU**. Moreover, **GGIU** even improved the performance based on CLIP. Finally, we analyze the influence of different hyper-parameters, and the results accord with our theoretical analysis.

References

1. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763 (2021)

2. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 594–611 (2006). https://doi.org/10.1109/TPAMI.2006.79

3. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2016, 5–10 December 2016, Barcelona, Spain, vol. 29, pp. 3630–3638 (2016)

4. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2017, 30 December 2017, Long Beach, CA, USA, vol. 30, pp. 4077–4087 (2017)

5. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 1126–1135 (2017)

6. Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X.: Meta-baseline: exploring simple meta-learning for few-shot learning. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021, pp. 9042–9051 (2021). https://doi.org/10.1109/ICCV48922.2021.00893

7. Zhang, C., Cai, Y., Lin, G., Shen, C.: DeepEMD: few-shot image classification with differentiable earth mover's distance and structured classifiers. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020, pp. 12200–12210 (2020). https://doi.org/10.1109/CVPR42600.2020.01222

8. Hospedales, T., Antoniou, A., Micaelli, P., Storkey, A.: Meta-learning in neural networks: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 5149–5169 (2021)

9. Thrun, S., Pratt, L.: Learning to learn: introduction and overview. In: Thrun, S., Pratt, L. (eds.) Learning to Learn, pp. 3–17. Springer, Cham (1998). https://doi.org/10.1007/978-1-4615-5529-2_1

10. Koch, G., Zemel, R., Salakhutdinov, R., et al.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop, vol. 2, Lille (2015)

11. Satorras, V.G., Estrach, J.B.: Few-shot learning with graph neural networks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018, Conference Track Proceedings (2018)

12. Luo, X., et al.: Rectifying the shortcut learning of background for few-shot learning. Adv. Neural. Inf. Process. Syst. **34**, 13073–13085 (2021)

13. Yang, S., Liu, L., Xu, M.: Free lunch for few-shot learning: distribution calibration. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021 (2021)

14. Tang, X., Teng, Z., Zhang, B., Fan, J.: Self-supervised network evolution for few-shot classification. In: Zhou, Z. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event/Montreal, Canada, 19–27 August 2021, pp. 3045–3051 (2021). https://doi.org/10.24963/ijcai.2021/419

15. An, Y., Xue, H., Zhao, X., Zhang, L.: Conditional self-supervised learning for few-shot classification. In: Zhou, Z. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event/Montreal, Canada, 19–27 August 2021, pp. 2140–2146 (2021). https://doi.org/10.24963/ijcai.2021/295

16. Li, J., Wang, Z., Hu, X.: Learning intact features by erasing-inpainting for few-shot classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 8401–8409 (2021)

17. Afrasiyabi, A., Lalonde, J.-F., Gagné, C.: Associative alignment for few-shot image classification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 18–35. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_2

18. Wagemans, J., et al.: A century of gestalt psychology in visual perception: II. conceptual and theoretical foundations. Psychol. Bull. **138**(6), 1218 (2012)

19. Henle, M.: The selected papers of Wolfgang Köhler. Philos. Phenomenol. Res. **33**(2), 270–271 (1972)

20. Hamlyn, D.W.: The Psychology of Perception: A Philosophical Examination of Gestalt Theory and Derivative Theories of Perception. Routledge, New York (2017)

21. Brennan, J.F., Houde, K.A.: History and Systems of Psychology. Cambridge University Press, Cambridge (2017)

22. Stevenson, H.: Emergence: The Gestalt Approach to Change. Unleashing Executive and Organizational Potential (2012). Retrieved July 2012

23. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset (2011)

24. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y

25. Li, W., et al.: LibFewShot: a comprehensive library for few-shot learning. arXiv preprint arXiv:2109.04898 (2021)

26. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 4367–4375 (2018). https://doi.org/10.1109/CVPR.2018.00459

27. Luo, X., Chen, Y., Wen, L., Pan, L., Xu, Z.: Boosting few-shot classification with view-learnable contrastive learning. In: 2021 IEEE International Conference on Multimedia and Expo, ICME 2021, Shenzhen, China, July 2021, pp. 1–6 (2021). https://doi.org/10.1109/ICME51207.2021.9428444

28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

29. Chu, W.H., Li, Y.J., Chang, J.C., Wang, Y.C.F.: Spot and learn: a maximum-entropy patch sampler for few-shot image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6251–6260 (2019)

30. Li, A., Huang, W., Lan, X., Feng, J., Li, Z., Wang, L.: Boosting few-shot learning with adaptive margin loss. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020, pp. 12573–12581 (2020). https://doi.org/10.1109/CVPR42600.2020.01259

31. Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., Liu, H.: Nico++: towards better benchmarking for domain generalization. arXiv preprint arXiv:2204.08040 (2022)