


Article

Few-Shot Classification with Dual-Model Deep Feature Extraction and Similarity Measurement

Jing-Ming Guo ^{1,2,*} , Sankarasrinivasan Seshathiri ^{1,2} and Wen-Hsiang Chen ^{1,2}

¹ Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei 106335, Taiwan

² Advanced Intelligent Image and Vision Technology Research Center, National Taiwan University of Science and Technology, Taipei 106335, Taiwan

* Correspondence: jmguo@mail.ntust.edu.tw; Tel.: +886-2-29558168

Abstract: From traditional machine learning to the latest deep learning classifiers, most models require a large amount of labeled data to perform optimal training and obtain the best performance. Yet, when limited training samples are available or when accompanied by noisy labels, severe degradation in accuracy can arise. The proposed work mainly focusses on these practical issues. Herein, standard datasets, i.e., Mini-ImageNet, CIFAR-FS, and CUB 200, are considered, which also have similar issues. The main goal is to utilize a few labeled data in the training stage, extracting image features and then performing feature similarity analysis across all samples. The highlighted aspects of the proposed method are as follows. (1) The main self-supervised learning strategies and augmentation techniques are exploited to obtain the best pretrained model. (2) An improved dual-model mechanism is proposed to train the support and query datasets with multiple training configurations. As examined in the experiments, the dual-model approach obtains superior performance of few-shot classification compared with all of the state-of-the-art methods.



Citation: Guo, J.-M.; Seshathiri, S.; Chen, W.-H. Few-Shot Classification with Dual-Model Deep Feature Extraction and Similarity Measurement. *Electronics* **2022**, *11*, 3502. <https://doi.org/10.3390/electronics11213502>

Academic Editors: Leonardo Galteri, Claudio Ferrari and Stefanos Kollias

Received: 12 October 2022

Accepted: 26 October 2022

Published: 28 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: dual-model; few-shot learning; few-shot classification; feature matching; self-supervised learning

1. Introduction

In the deep learning domain, whether it is image classification [1], object detection [2], or segmentation [3], the quantity and quality of the dataset itself are critical for the training of the model. Specifically, in the case of general image classification, the quantity and quality of the datasets are critical in determining the performance of the model. However, the generation of large and complex image datasets through manual labelling results in huge labor costs and also involves significant labelling and curation time. Recently, the alternate approach for labelled dataset generation is performed using web crawling techniques to collect numerous images from the Internet along with the associated text descriptions. This can save tremendous labor costs in image generation, yet the collected images may suffer from massive mislabeling, leading to quality degradation of the overall dataset. Moreover, though many open datasets are available, their application scope is limited, and even in a dataset, there exist multiple issues such as out-of-distribution data and data versatility. For example, in the case of industrial defect detection, the open datasets cannot be used directly. Sometimes, techniques such as transfer learning are not applicable as the source and test datasets have different distributions and impose a strong domain shifting problem. Hence, dataset generation is tricky and tedious as the labelling team has to be incorporated with the production team and decide the product label based on multiple factors. Hence, it is indeed challenging to obtain sufficient annotated data, and if the dataset is small in quantity, many general supervised learning approaches cannot be trained properly, resulting in poor classifiers.

Therefore, the main objective of the proposed work focuses on the development of a few-shot-based classification algorithm, which is suited for practical applications, including small or incorrect data involvement. As in Figure 1, the few-shot training involves fewer images than that of the standard classifiers and still can achieve the classification of new classes in the testing stage [4]. The first challenge is to figure out how to learn effectively with a small amount of training data, and the key is to enable the model to extract effective information from the small dataset and, accordingly, improvised classification performance. The second direction is to maximize the image information, and thus, the model can learn efficiently and be trained with a few labelled data. Considering the objectives, in this work, a multi-backbone model is proposed to yield good feature extraction and obtain superior performance compared to the state-of-the-art methods.



Figure 1. Example of few-shot training and classification.

This manuscript is organized as follows: Section 2 covers the literature review on the existing works and limitations and main contributions of the proposed work. Section 3 briefly describes the datasets used for the model training and evaluation. The detailed description of the proposed model is provided in Section 3. The comprehensive experimental analysis on three standard datasets and the overall summary are provided in Sections 5 and 6, respectively.

2. Related Work

Some state-of-the-art works in the few-shot framework are introduced as follows.

2.1. Conventional Approaches

The primary approach to few-shot learning uses the matching nets (MNs) [5], which are based on the metric learning approach to extract feature embedding and use machine learning (ML) classifiers such as weighted nearest neighbors. The approach comprises multiple gradient update computations, limiting its utility to small datasets. Subsequent research works predominantly focused on the meta-learning-based approach [6,7], in which the model is trained on multiple learning tasks with an assumption that it can be easily extended to solve scenarios with limited data. The algorithm is inspired by how humans are trained in multiple skill sets, which is applicable to solving decision-making problems. Hence, the key objective is to obtain a more generalized model, which can be easily fine-tuned even with limited labelled data.

2.2. Meta Learning Models

In contrast to the conventional deep learning techniques, which require huge data, the meta learning models are trained using the randomly sampled subset from the large dataset. The training process is defined as N-way K-shot learning, where N-way usually refers to the number of categories of training and testing data and K-shot indicates how many pieces of data are included in each category [8].

For example, a 5-way 1-shot task is performed as shown in Figure 2. In the meta training stage, the model distinguishes the test images, which are in the same category as the randomly selected images by N-way K-shot. Thereafter, multiple random samplings and pairings are performed to train the model to learn and predict unknown labels with a small amount of data. In the meta testing stage, the trained few-shot training model can be applied to the few-shot prediction tasks of different categories from the training stage using the prior knowledge previously learned during the meta training stage. As the models are trained on limited data, there is a serious possibility of overfitting.

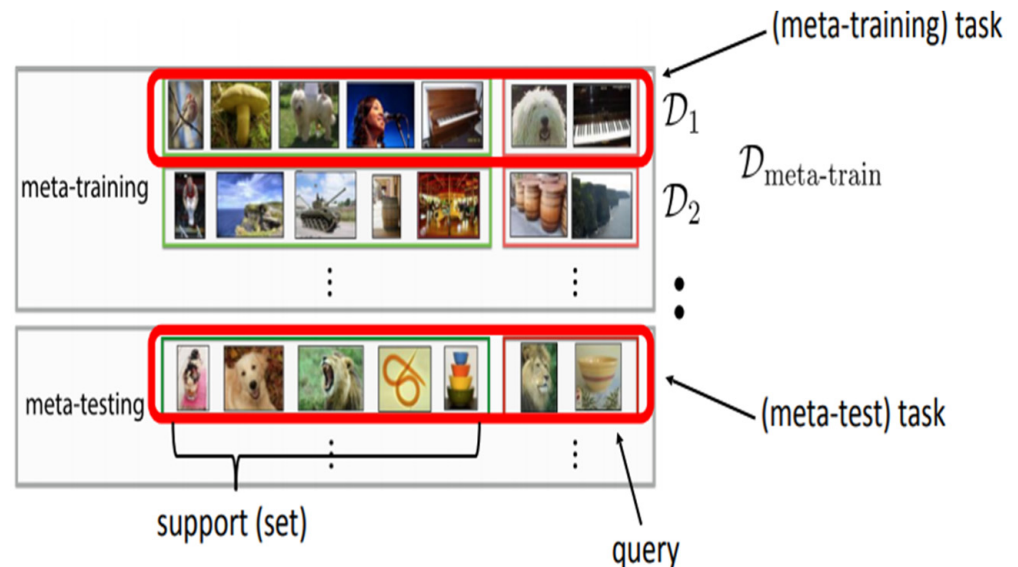


Figure 2. Concept of meta learning.

To solve the overfitting issue, the prototypical network (PN) [9] is introduced as shown in Figure 3. The central premise of the PN is that there exists an embedding in which points cluster around a single prototype representation for each class. To achieve that, the deep neural networks are used for the non-linear mapping of the input into an embedding space with the class's prototype to be its mean of the support set in the embedding space. Specifically, the feature embedding of the support set is used to find the most representative features of each category. It can be noticed in Figure 3 that there are three different types of feature cluster points in the support set, and they correspond to the features representing three different categories. The feature embedding of the query set is used for similarity calculation and for category prediction. Each feature point of the whole query set is sequentially compared with all the support set feature points based on the distance metrics. The closer the feature points of the query set are, the higher the probability that it belongs to that specific category. By designing a stable similarity calculation system, the PN not only makes full use of each selected image feature information in the training phase, but also introduces training stability, maintaining a good feature space projection relationship. Moreover, as in previous meta learning, the method also utilizes the concept of random sample training, which involves taking up a small subset, and then, it subdivides the selected samples into a support set and a query set. In handling new category labels in the testing phase, the model can perform more effectively since it has learned how to distinguish different categories of data from a few randomly selected samples.

2.3. Self-Supervised Learning Models

Recent methods mainly focus on self-supervised-learning (SSL)-based [10–14] techniques. In the latest, based on the combination of SSL and meta learning, a new approach is proposed, termed AmdimNet [15], as shown in Figure 4. As the SSL techniques are formulated based on contrastive learning, there is no requirement for labelled data to obtain

the pretrained model, and it is ideally suited to handle the few-shot learning problems. The model architecture is divided into upper and lower parts, in which the upper part comprises the pretraining process using self-supervised learning (SSL) [12] and the lower part is the fine-tuning process using the few-shot training architecture. In the pretraining process, unlabeled images are fed as the input and each image under different augmentations to guide the network to learn unique features for that particular category, i.e., the model learns the class-consistent features. In the subsequent stage, the pretrained model, which has deep prior knowledge, can be easily fine-tuned with just a few image samples. Hence, the resultant model can produce promising feature embedding, which is a good representation of pretraining, as well as few-shot data. This combined approach using SSL and meta learning results in a better model than a single-stage training with small data. However, many self-supervised learning methods and the augmentation schemes are not fully exploited. Hence, in this work, a detailed work was carried out to understand the various SSL networks and its usage in few-shot learning.

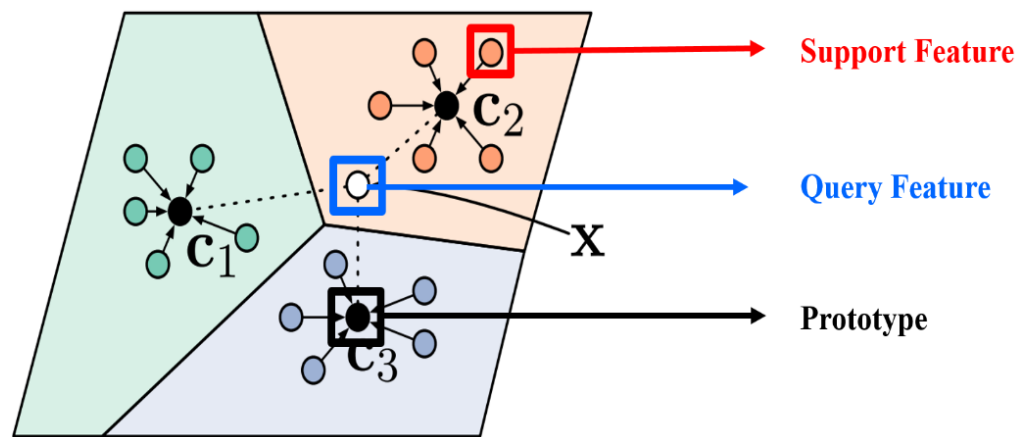


Figure 3. Prototypical network.

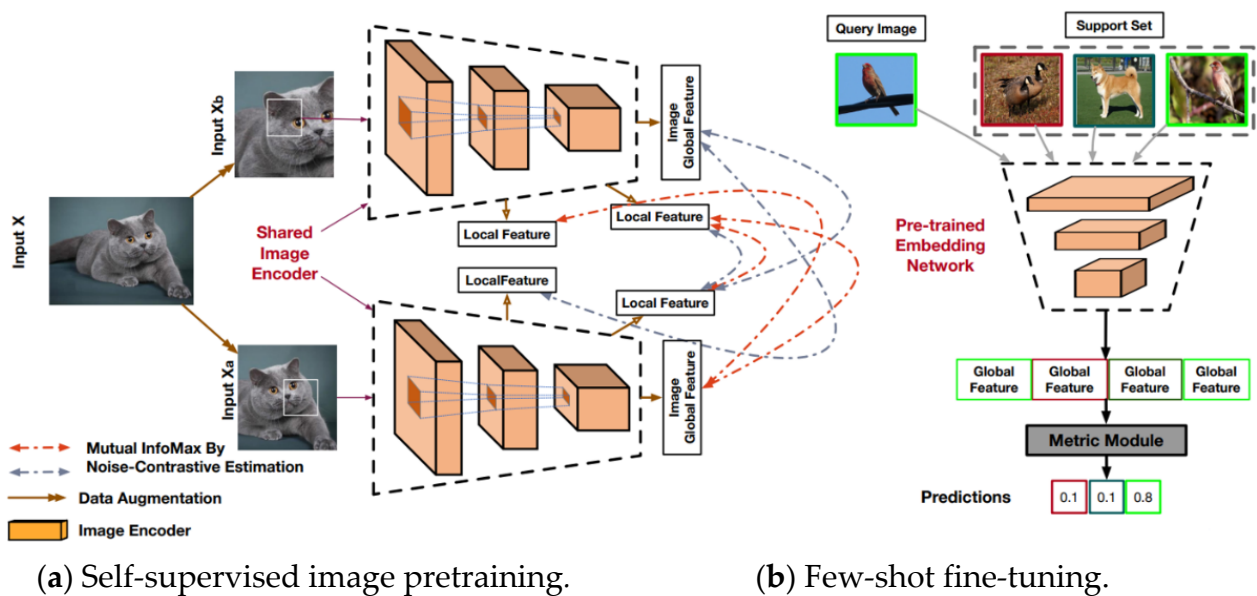


Figure 4. AmdimNet.


Considering the overall limitations, the proposed work emphasizes the development of an improved few-shot classification model with an optimal pretrained model and dual-architecture for better learning. The main contributions are as follows.

- (1) As the few-shot classification relies on learning from a few ground-truth data, the proposed work focusses on the development of optimal pretrained models, which can be generalized and fine-tuned to any datasets with limited training. In this work, four prominent SSL techniques such as SimCLR, SimSiam, BYOL, and BTs were trained and analyzed to obtain best pretrained backbone.
- (2) For further improvisation, more augmentation techniques such as random jigsaw and random patch swap were added to obtain more diversity and robustness, during the pretraining stages.
- (3) From the model perspective, the proposed work is based on the latest ConvNeXt backbone, and a new dual-model configuration is proposed with different depths, complementing the few-shot training. The new training strategies practiced in the latest vision transformer and convolution models were also integrated.
- (4) Finally, a new training approach is proposed, in which the distance between the feature embedding of the query set and the most representative feature vector of each category is used to determine the query set category. In addition, the progressive model training was performed using multiple few-shot extraction and feature similarity assessment.


3. Few-Shot Learning Datasets

The proposed work was tested on three public standard few-shot learning image datasets, as shown in Figure 5. The Mini-ImageNet dataset [16], which contains 60,000 images, was collected from ImageNet. It has a total of 100 categories, including 64 categories of training sets, 16 categories of validation sets, and 20 categories of testing sets, and each category contains 600 images. To verify the ability of the model to classify a new category with a small number of samples, the training set and the test set in the Mini-ImageNet dataset were divided into distinct categories without any overlapping.


Number of data (Categories)		Number of data (Categories)		Number of data (Categories)	
Total Number	60,000 (100)	Total Number	60,000 (100)	Total Number	11,788 (200)
Training Set	38,400 (64)	Training Set	38,400 (64)	Training Set	4796 (160)
Validation Set	9600 (16)	Validation Set	9600 (16)	Validation Set	1198 (40)
Testing Set	12,000 (20)	Testing Set	12,000 (20)	Testing Set	5794 (200)



(a) Mini-ImageNet.



(b) CIFAR-FS.



(c) CUB 200.

Figure 5. Few-shot classification datasets.

The second few-shot learning dataset was CIFAR-FS [17], containing 60,000 images collected from CIFAR100. The dataset comprises 100 categories, which are divided into 64 categories in the training sets, 16 categories in the validation sets, and 20 categories in the test sets. Each category contains 600 images, and the image size is set at 84×84 . Similar to the previous dataset, the training set and the testing set were divided in a specific way to avoid any overlapping. The third dataset used in this work was the Caltech-UCSD Birds-200-2011 (CUB 200) [18], containing 11,788 bird images with 200 categories. This dataset is the most popular in fine-grained visual classification tasks. As opposed to the category settings of the other two few-shot training datasets, the categories of the training and test sets in CUB 200 still overlap, but the amount of data is far less than the other datasets.

4. Proposed Method

The present work comprises three main elements involving self-supervised learning, a dual-mode backbone network, and feature assessment. In this section, the detailed description of all are provided in the following three subsections.

4.1. Self-Supervised Learning

The main strategy of the proposed work is to exploit the advantage of the SSL methods in obtaining the best pretrained model and then improve the performance using more effective backbone networks with better training strategies. The detailed elaboration of the SSL methods is provided below.

The self-supervised learning models work under the common objective of learning the representations that are invariant under various distortions. In general, different distorted input images are fed through the variant of the Siamese network and a specific loss function is minimized. The most challenging factor is to avoid the model collapse, leading the encoder network to generate constant or non-informative vectors. To begin with, the well-known framework based on the contrastive learning of visual representations, termed SimCLR [18], was utilized. Given any training image x , the module produces two correlated views of the same image, which are denoted as \tilde{x}_i and \tilde{x}_j , and this forms a positive pair. Among many data augmentation approaches, crop and resize, crop and flipping, rotation, cutout, Gaussian noise, and color jitter were adopted in this study for training. The model optimization involves minimizing and maximizing the distances amongst features in the intra- and inter-classes, respectively, and the distance metrics was based on the contrastive loss function.

For a given image set $\{\tilde{x}_k\}$, the positive pairs are generated as \tilde{x}_i and \tilde{x}_j . The main objective of the contrastive prediction loss is to obtain a similar \tilde{x}_j in $\{\tilde{x}_k\}_{k \neq i}$ for the given \tilde{x}_i . For each mini-batch of N examples, as two augmentations are carried out for each example at a time, $2N$ data points are generated. Herein, the normalized temperature-scaled cross-entropy loss (NT-Xent) was adopted as the loss function, defined as follows.

$$l_{i,j} = -\log \left(\frac{\exp\left(\frac{\text{sim}(x_i, x_j)}{\tau}\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(\frac{\text{sim}(x_i, x_k)}{\tau}\right)} \right), \quad (1)$$

where $\mathbf{1}_{[k \neq i]} \in \{0, 1\}$; the output is 1 if $k \neq i$, and τ and $\text{sim}(z_i, z_k)$ denote the temperature parameter and cosine similarity, respectively. The temperature parameter is useful to widen the range of cosine similarity $[-1, 1]$ according to the user preference. Herein, τ was set at 0.1, which expands the cosine similarity with the range from $\exp(0.1)$ to $\exp(10)$, and it helps to better separate the positive and negative examples. The consolidated loss is computed across all the positive pairs of both (i, j) and (j, i) for the mini-batch. The SimCLR model has two main limitations: First, it requires a large amount of contrastive learning pairs, which is not feasible for small/medium datasets. Second, to obtain the optimal performance, it requires training with large batch sizes (up to 4096 or 8192), which requires multiple graphical processing units (GPUs) or tensor processing units (TPUs), and these are highly expensive and hard to realize in many real-time applications. Another important problem is the model collapse, which results in a poor encoder model. To tackle this issue, the subsequent models were based on the distillation methods such as simple Siamese representation learning (SimSiam) [19] and bootstrap your own latent (BYOL) [20]. The architecture and parameter updates were modified to bring asymmetry in the network. The model parameters were only updated using the distorted version of the input, and the other distorted version was used as a fixed target. Though the model avoids collapse, it is not certain how it will avoid collapse. More recently, another approach based on H. Barlow's redundancy reduction principle was proposed, as demonstrated in Figure 4b, which was applied to the pair of identical networks as in SSL models. The method is termed Barlow twins (BTs) [21] and can perform well with reduced batch sizes, a deeper

projector head, a large embedding, etc. The main contribution is the introduction of a new loss function, termed the Barlow twins (BTs) loss.

$$L_{BT} \triangleq \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2, \tag{2}$$

where λ is a positive constant to balance the tradeoff between the invariance and redundancy reduction loss; the notation C refers to the cross-correlation matrix, which is the output of the two identical networks for each batch; the notation \triangleq signifies equal by definition.

$$C_{ij} \triangleq \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b (z_{b,i}^A)^2} \sqrt{\sum_b (z_{b,j}^B)^2}}, \tag{3}$$

where b refers to the batch samples and i, j are the indices of the vector dimension of the network output. The C is the square matrix and is of a size of the dimensionality of the network output. The BTs loss function is very effective in eliminating model collapse and can provide the best feature learning. Though BTs aims to reduce the redundancy at the embedding vector level, there is still a possibility that the input images may have correlated patterns. In this work, the pretraining was carried out using SimCLR, SimSiam, BYOL, and BTs with the additional augmentation used in the fine-grained classification problems. An improved pretrained model was obtained through this study, and the detailed comparative results are presented in the Results Section.

4.2. Dual-Model Architecture

Herein, the detailed description of the proposed model using the dual-model architecture is provided. As shown in Figure 6, the overall training structure and process can be divided into two parts, i.e., few-shot data extraction and feature similarity assessment. At the beginning of each few-shot training, a small subset of data is randomly selected from each category and is divided as the support set and the query set. For this randomly selected data, the categories in the support set and the query set are the same. Subsequently, the data of the support set and the query set are passed through different feature extraction networks to obtain the feature embedding for each image.

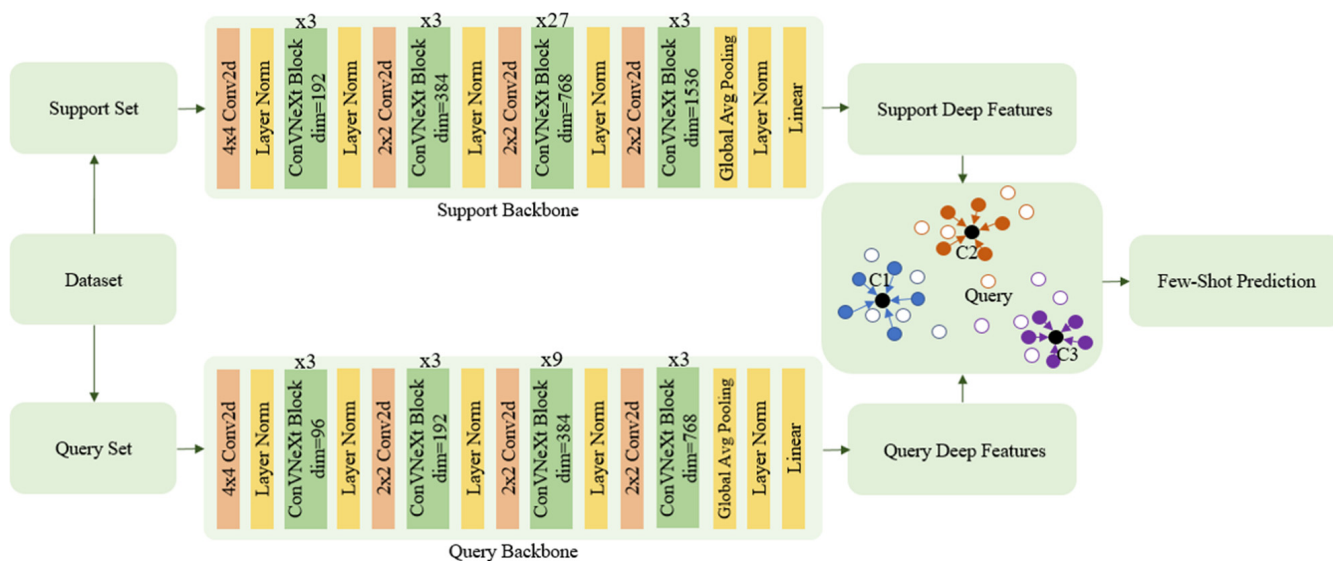


Figure 6. Few-shot classification with dual-model deep feature extraction and similarity measurement mechanism.

In the subsequent stage of model similarity learning and calculation, the feature embedding of the support set is used to build the most representative features for each category. On the other hand, the feature embedding of the query set is used for similarity calculation, which is similar to the classifier function of general supervised learning. In this way, the model can learn to understand the similarity in the feature embedding of each category from the few samples. The solid circles of different colors in the feature space represent the feature embedding of different categories in the support set. The black solid circles represent the most representative feature embedding of each category, and the white solid circles represent the feature embedding of the query set, respectively. Through the distance estimation between the feature embedding of the query set (white solid circles) and the most representative feature embedding of each category (black solid circles), the model can determine all of the data in the query set. The overall training achieves the improved few-shot classification through multiple few-shot extraction training and similarity judgment.

Figure 7 shows a schematic diagram of the selection method of few-shot training. For instance, let us assume the tasks as a 3-way 5-shot with 5-query task, in which 3-way refers to the number of categories selected each time before training and testing and 5-shot refers to number of data samples in the support set. The 5-query refers to the amount of data in each category of the query set. Hence, at the beginning of each few-shot training, different types of data are randomly selected as the new support and query set. This helps to increase the generalization ability of the model to various types of data, which is the key objective of few-shot learning.

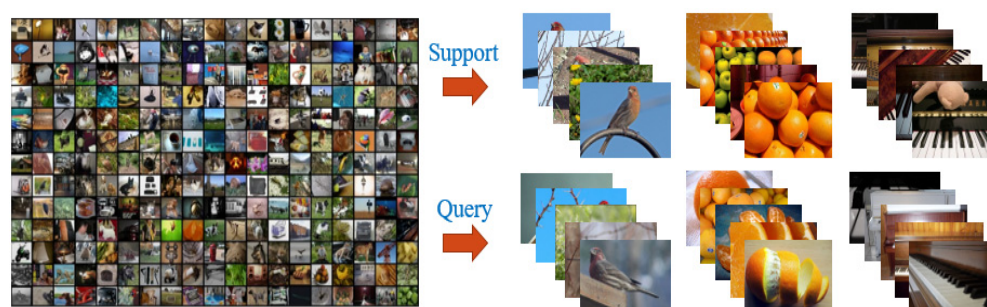


Figure 7. Schematic diagram of the selection method.

4.3. Feature Extraction and Similarity Assessment

Figures 8 and 9 show the backbone network of our feature extraction model for the support set and query set. The main architecture of the models was inspired by the design of the ConvNeXt [22] block, as shown in Figure 10. Compared with the standard ResNet [23] as a baseline, the ConvNeXt block combines the advantages of many models to optimize the performance of its own feature extraction. For example, it draws depthwise convolution to improve the learnable features of each channel and adapt the design concept of various vision transformers (ViT) [24] such as Swin-transformer (Swin-T) [25] to improve the learning performance of the CNN. Because of the different task orientations of the support set and the query set, the support set data are more influential than the query set in few-shot learning. From Figure 8, it can be seen that the number of convolutional filters used in the third convolutional block for the support set feature is three-times bigger than that of the query set backbone. As the network learns the query set through limited images, the backbone can be relatively small. However, in comparison with the single-model backbone in existing networks, this dedicated backbone for the support and query set is with more advantages.

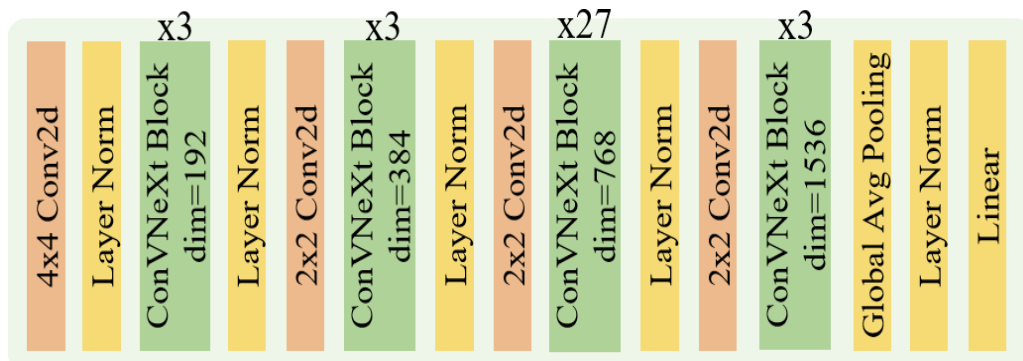


Figure 8. Support set feature extraction backbone.

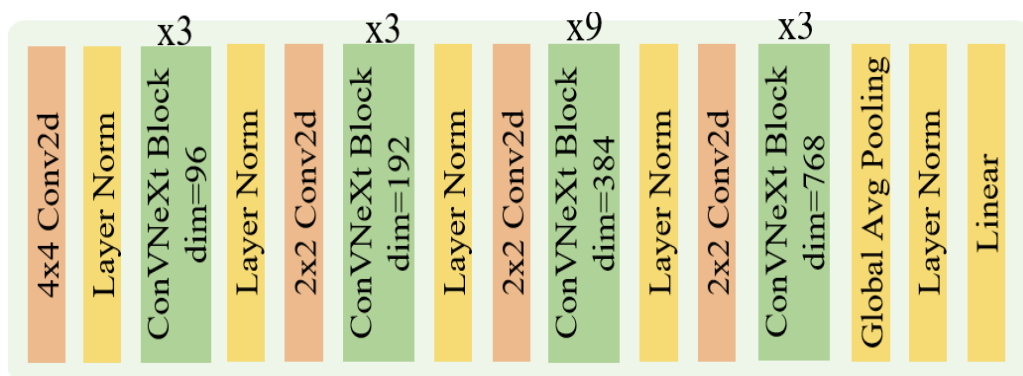


Figure 9. Query set feature extraction backbone.

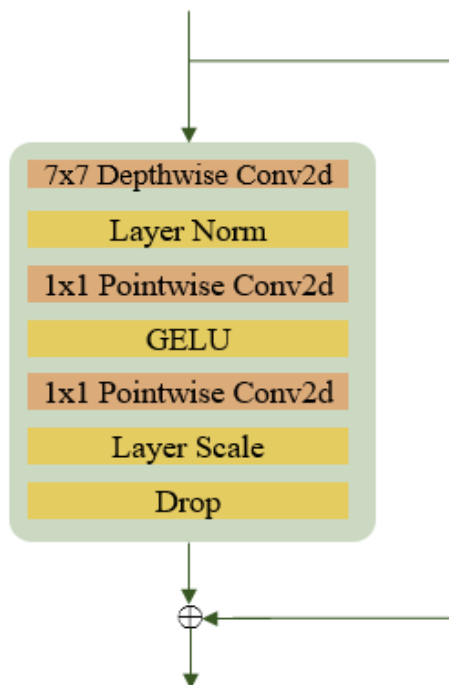


Figure 10. ConvNext block.

The feature extraction process of the support set and query set is shown in Figures 11 and 12. The feature embedding of each datum in the support set is extracted through the support set

feature extraction model, and the most representative feature of each category is computed using Equation (1). The term C_n refers to the center point within each category cluster.

$$C_n = \frac{1}{S} \sum_{x_d \in S} \text{Backbone}(x_d) \tag{4}$$

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2} \tag{5}$$

$$L(x) = p(y = n|x) = -\log \frac{\exp(-d(\text{backbone}(x), C_n))}{\sum_{n'} \exp(-d(\text{backbone}(x), C_{n'}))} \tag{6}$$

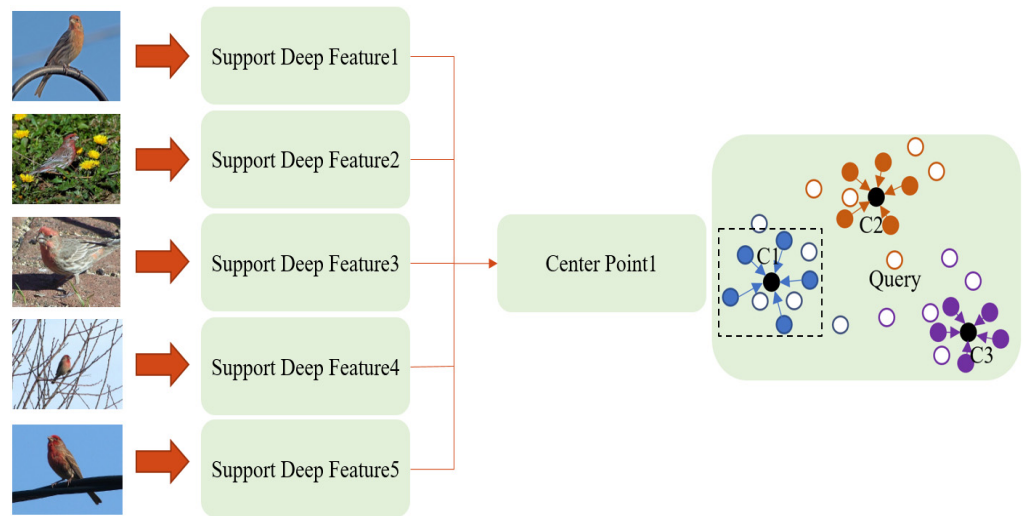


Figure 11. Support set deep feature extraction process.

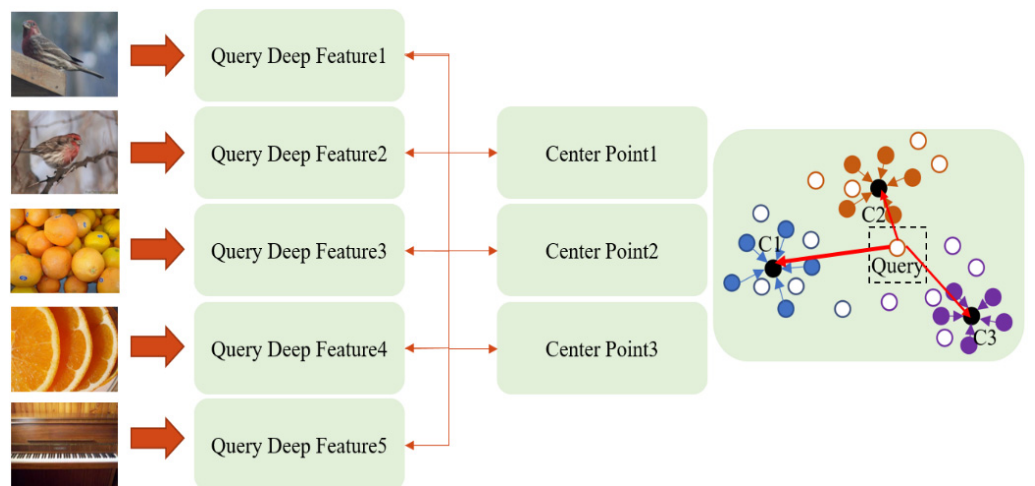


Figure 12. Query set deep feature extraction process.

An illustrative example is provided in Figure 13; it can be seen that, since the feature embedding of the query set is closer to the depth feature of C2, the category of the data is predicted as C2.

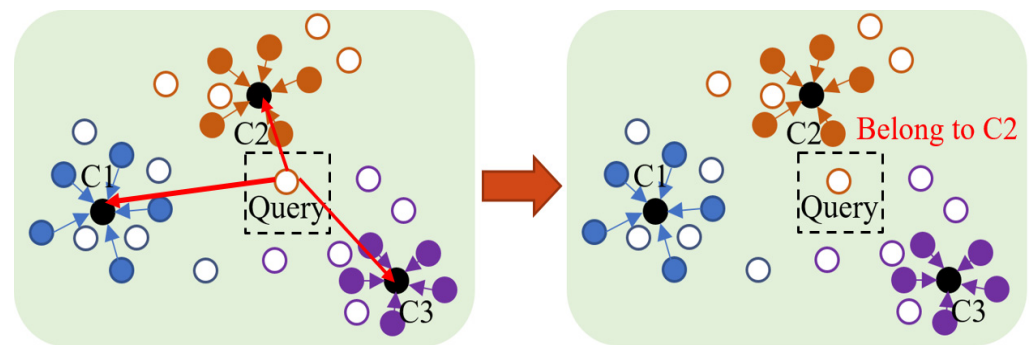


Figure 13. Query set deep feature classification.

5. Results and Analysis

To perform the comprehensive model evaluation, the three standard datasets, i.e., Mini-ImageNet, CIFAR-FS, and CUB 200, were used and compared with many state-of-the-art models. The final model predicts each image in the test dataset, and the class with the highest score is selected. The final evaluation was carried out by comparing the predicted class with the ground-truth label. If the two labels are consistent, it is a correct case; otherwise, it is an incorrect case. The accuracy rate ($Acc.$) was used as the evaluation criterion, which is defined as follows.

$$Acc = \frac{N_{correct}}{N_{all}}, \quad (7)$$

where $N_{correct}$ refers to the number of correctly classified images and N_{all} represents the total number of images in the test set.

5.1. Pretrained Model Optimization

As the few-shot learning method was trained on minimal images, the pretrained model can significantly affect the classification performance. In this work, four prominent SSL methods such as SimCLR, BYOL, SimSiam, and BTs were considered. The objective was to identify the SSL method that can produce the optimal pretrained model. The backbone model was ConvNeXt, and each model was trained for 50 epochs with a batch of size 256. For the experiments, the Mini-ImageNet dataset was used, in which 40% of the images were used to conduct training without labels, 10% of the labelled images were considered for fine-tuning, and 1000 images were used for testing. The general classification performance of these methods was tested, and the best approach was selected to obtain the pretrained model for the few-shot classification. In addition to the standard augmentation such as crop, resize, flipping, rotation, cutout, gaussian, and color jitter, we also exploited some additional augmentation that are popular in fine-grained classifiers such as random patch swap (RPS) and random jigsaw (RJ) [26]. In fine-grained learning, these augmentations are very useful in learning features of different granularities and also help the network localize in fine-grained regions. It can be also seen from Table 1 that the new augmentations also improve the general classification accuracy, which corresponds to an improved pretrained model.

Overall, the BTs with additional augmentation attained the best classification performance. Hence, instead of considering the pretrained model trained on ImageNet directly, it was further fine-tuned using BTs with an additional augmentation technique on 30% of the training images (un-labelled) for all datasets. This process can provide a good generalized pretrained model to each of the datasets and also boost the few-shot classification performance.

Table 1. Comparisons of SSL methods.

Method	Augmentations	Accuracy
SimCLR		63.5%
SimSiam	Set 1: crop, resize, flipping, rotation, cutout, gaussian and color jitter	64.21%
BYOL		66.72%
BTs		67.85%
BTs	Set 1 + RPS + RJ	68.9%

5.2. Model Ablation Studies

In this study, a detailed ablation study was carried out to determine the number of convolutional blocks used in the support and query feature extractors. It can be seen from Figures 8 and 9 that the numbers of convolutional blocks used for the support and query set were different, and the accuracies for different combination are presented below. All this testing was performed on the Mini-ImageNet dataset.

To begin with, as in Cases 1 and 2, the backbone models of the support and query set were provided with more convolutional layers from the existing setup [23]. It can be seen that the accuracy was significantly less for both cases, and it was also inferred that, when the convolutional blocks kept for the query set were fewer, the model delivered better accuracy. This reduction in accuracy was due to the overfitting issue with respect to very deep configurations and limited training data. This is evident from the fact that, in Cases 1 and 2, the model provided high accuracy on the training sets and low accuracy on the test sets, whereas in Cases 4 and 5, the training and test accuracy were nearly equal. Overall, Case 4 was considered for our final training due it having the best accuracy, and the query set backbone model is maintained to have less depth than the support set. We observed this type of configuration to be also effective in avoiding the overfitting or underfitting issues. Moreover, k-fold cross-validation was performed considering different model configurations and image categories. For the experimentation, five-fold cross-validation was considered using the training set images with different numbers of class labels, as in Figure 14. It can be seen that, in agreement with Table 2, the Case 4 model had the best average accuracy and least variance, whereas Case 3 showed significant degradation in performance with respect to the number of class labels. The comprehensive results of the Case 4 model on the standard few-learning datasets and its comparison analysis are provided in the next subsection.

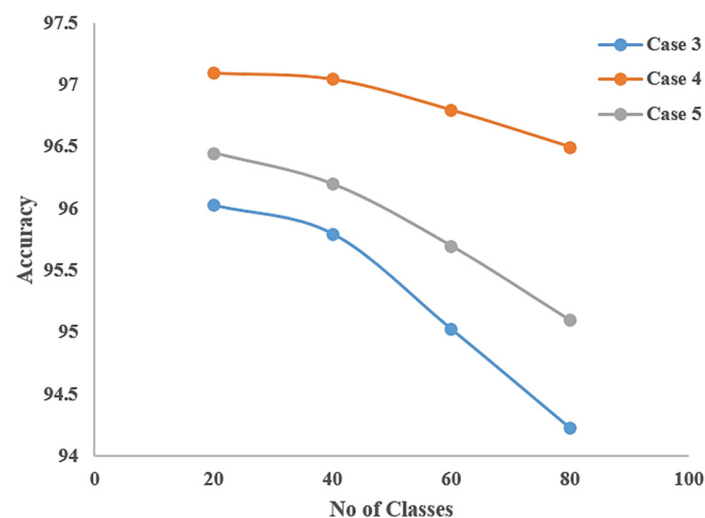
**Figure 14.** Five-fold cross-validation.

Table 2. Convolutional blocks for support and query set.

Case	Support Set	Query Set	Accuracy (Training)	Accuracy (Test)
1	6,6,27,6	6,6,27,6	96.28	91.63
2	6,6,27,6	6,6,9,6	96.53	92.63
3	3,3,27,3	3,3,27,3	95.78	93.05
4	3,3,27,3	3,3,9,3	96.10	95.50
5	3,3,9,3	3,3,27,3	94.35	93.15

5.3. Few-Shot Classification Results

Few-shot classification and few-shot learning are technically the same process. At the beginning of each small-sample test, several pieces of data are randomly selected from each category of the overall test dataset as a support set and a query set for prediction, and then, the query set is used for prediction. The feature embedding of each image is obtained from the data of the support and query set through the dual-backbone network. Subsequently, the distance is evaluated between the feature embedding of the query set and the most representative deep feature of each category (calculated from the depth feature of the support set). Based on the least distance, the class labels are estimated, and the results for various datasets are shown below.

Table 3 shows the few-shot classification results of Mini-ImageNet (5-way 5-shot). In this experiment, five categories of test images were randomly selected from the 20 categories of the Mini-ImageNet dataset as the test dataset, and 5 images in each category were randomly selected as the support set and 15 images as the query set for the prediction. The number of few-shot classifications in the testing phase was performed 10,000 times, and the average accuracy is listed. From the results, it can be seen that the proposed model performed the best among the existing works, and the overall classification accuracy was 3.11% higher than the current state-of-the-art method. Table 3 also shows the few-shot classification results of Mini-ImageNet (5-way 1-shot). In each few-shot classification, five categories of test images were randomly selected from the 20 categories in the Mini-ImageNet dataset as the test dataset, and each category would randomly select 1 image as the support set and 15 images as the query set for the prediction. The number of few-shot classifications in the testing phase was performed 10,000 times, and its average was estimated. As the data usage was with only 1-shot, thus the model accuracy was slightly lower than that of the 5-shot classification. However, the proposed method still outperformed the state-of-the-art methods by 2.31%.

Similar experiments for the CIFAR-FS dataset are presented in Table 4, and it can be seen that the proposed model performed 1.96% and 1.46% higher than the existing models for 5-way 5-shot and 5-way 1-shot, respectively.

The classification accuracy for the CUB 200 dataset is shown in Table 5, in which the proposed model performed 1.23% and 1.06% higher than the existing models for 5-way 5-shot and 5-way 1-shot, respectively.

Finally, the classification results of the single-model and dual-model on each small dataset are provided in Table 6. It can be seen from the results that the dual-model proposed in this work showed significant improvement in the classification results on the three few-shot datasets compared to the single-model. It can also be observed that the classification results of the dual-model in the three few-shot datasets were at least 1% higher than the accuracy of the single-model methods. The results of the ablation study also verified that the dual mode feature extraction architecture achieved superior classification results in the few-shot classification tasks.

Table 3. Classification accuracy on Mini-ImageNet.

Method	Accuracy (5-Way 5-Shot)	Accuracy (5-Way 1-Shot)
Matching Nets [5]	60	46.6
MAML [4]	63.1	48.7
Relation Network [6]	65.32	49.42
Prototypical Networks [9]	68.2	50.44
PT + MAP [8]	88.82	76.82
Sill-Net [7]	89.14	79.9
EASY 3xResNet12 [13]	89.14	82.99
AmdimNet [15]	90.98	84.04
SOT [12]	91.34	84.81
CNAPS + FETI [14]	91.5	85.54
PEMnE-BMS * [11]	91.53	85.59
BAVARDAGE [27]	91.65	84.80
TRIDENT [28]	95.95	86.11
Dual-Model (Proposed)	94.64	88.3
Dual-Model (Proposed) + BTs-Pretrained Model + Set 1	95.83	88.91
Dual-Model (Proposed-Final) + BTs-Pretrained Model + Set 1 + RPS + RJ	95.98	88.96

Table 4. Classification accuracy on CIFAR-FS (5-way 5-shot).

Method	Accuracy (5-Way 5-Shot)	Accuracy (5-Way 1-Shot)
EASY 3xResNet12 [13]	90.47	87.16
PT + MAP [8]	90.68	87.69
LST + MAP [10]	90.73	87.73
Sill-Net [7]	91.09	87.79
PEMnE-BMS * [11]	91.86	88.44
SOT [12]	92.83	89.94
Dual-Model (Proposed)	94.74	91.4
Dual-Model (Proposed-Final) + BTs-Pretrained Model + Set 1 + RPS + RJ	95.16	92.35

5.4. Case Studies

To understand the robustness of the model with respect to various image attacks or variants, such as cropping, scaling, illumination, color, background, etc., detailed case studies considering images from all three datasets were conducted, as in Table 7. It can be seen that, though the Sample 1 and Sample 2 images were taken from the same category, it is visually challenging to classify them because of the huge variations or diversity. However, the proposed model was very successful in correctly classifying such images, and this clearly demonstrated the model's capability in handling large intra-class variations and that it is ideal for many real-time applications.

Table 5. Classification accuracy on CUB-200 (5-way 5-shot).

Method	Accuracy (5-Way 5-Shot)	Accuracy (5-Way 1-Shot)
Relation Network [6]	65.32	50.44
AmdimNet [15]	89.18	77.09
EASY 3xResNet12 [13]	91.93	90.56
PT + MAP [8]	93.99	91.68
LST + MAP [10]	94.09	94.73
Sill-Net [7]	96.28	94.78
PEMnE-BMS * [11]	96.43	95.48
SOT [12]	97.12	95.8
Dual-Model (Proposed)	98.35	96.82
Dual-Model (Proposed-Final) + BTs-Pretrain Model + Set 1 + RPS + RJ	98.56	97.23

Table 6. Classification accuracy comparison of single-model and dual-model.

Framework	Datasets	Accuracy	
		5-Way 1-Shot	5-Way 5-Shot
Single-Model (Query Backbone)	Mini-ImageNet	82.81	89.89
	CIFAR-FS	85.68	90.44
	CUB 200	92.56	93.99
Single-Model (Support Backbone)	Mini-ImageNet	86.31	93.83
	CIFAR-FS	89.68	93.72
	CUB 200	95.1	97.43
Dual-Model (Proposed Method-Final)	Mini-ImageNet	88.96	95.98
	CIFAR-FS	92.35	95.16
	CUB 200	97.23	98.56

Table 7. Case studies on model robustness.



Dataset	Sample 1	Sample 2	Variations
Mini-ImageNet	 <p>Ground-Truth: Trifle Model Predicted: Trifle</p>	 <p>Ground-Truth: Trifle Model Predicted: Trifle</p>	<ul style="list-style-type: none"> • Shooting Angle • Illumination • Image Content

Table 7. Cont.



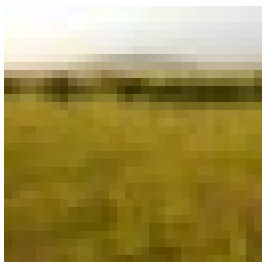
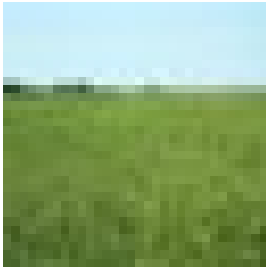
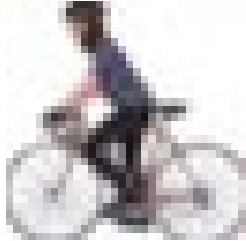
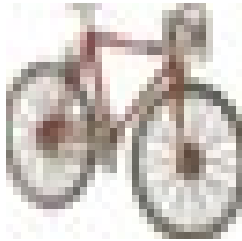




Dataset	Sample 1	Sample 2	Variations
	 <p>Ground-Truth: Scoreboard Model Predicted: Scoreboard</p>	 <p>Ground-Truth: Scoreboard Model Predicted: Scoreboard</p>	<ul style="list-style-type: none"> • Crop • Scaling
CIFAR-FS	 <p>Ground-Truth: Plain Model Predicted: Plain</p>	 <p>Ground-Truth: Plain Model Predicted: Plain</p>	<ul style="list-style-type: none"> • Color • Texture • Background
	 <p>Ground-Truth: Bicycle Model Predicted: Bicycle</p>	 <p>Ground-Truth: Bicycle Model Predicted: Bicycle</p>	<ul style="list-style-type: none"> • Multiple Objects • Crop
CUB 200	 <p>Ground-Truth: Baird Sparrow Model Predicted: Baird Sparrow</p>	 <p>Ground-Truth: Baird Sparrow Model Predicted: Baird Sparrow</p>	<ul style="list-style-type: none"> • Image Background • Pose Variation

Table 7. Cont.

Dataset	Sample 1	Sample 2	Variations
			<ul style="list-style-type: none"> • Background • Scaling • Crop
	Ground-Truth: Fox Sparrow Model Predicted: Fox Sparrow	Ground-Truth: Fox Sparrow Model Predicted: Fox Sparrow	

6. Conclusions

A new few-shot classification approach was proposed by integrating self-supervised learning, a hybrid convolutional neural network, and progressive training with multiple subsets. Four prominent SSL frameworks, i.e., SimCLR, SimSiam, BYOL, and BTs, were evaluated, and the BTs trained with additional fine-grain augmentation was found to obtain the best generalized pretrained model. A new hybrid architecture involving a dual-CNN model with the vision-transformer-based augmentation technique was developed. The few-shot training was conducted using multiple subsets and similarity estimation to obtain the best feature embeddings for the query and sample set. Extensive experiments were conducted on the three standard few-shot datasets, Mini-ImageNet, CIFAR-FS, and CUB 200. Moreover, a detailed evaluation was carried out to validate the diversity and robustness of our method. As examined from the results, the proposed method outperformed the existing state-of-the-art methods on all datasets and set a new benchmark accuracy in few-shot classification.

Author Contributions: Conceptualization, J.-M.G.; methodology, W.-H.C.; software, W.-H.C.; validation, S.S.; formal analysis, S.S.; investigation, J.-M.G.; resources, S.S.; data curation, W.-H.C.; writing—original draft preparation, S.S.; writing—review and editing, S.S.; visualization, W.-H.C.; supervision, J.-M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 770–778.
2. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 779–788.
3. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [[CrossRef](#)]
4. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning. PMLR, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
5. Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; Wierstra, D. Matching networks for one shot learning. *Adv. Neural Inf. Process. Syst.* **2016**, 3637–3645. [[CrossRef](#)]
6. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.S. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1199–1208.

7. Zhang, H.; Cao, Z.; Yan, Z.; Zhang, C. Sill-net: Feature augmentation with separated illumination representation. *arXiv* **2021**, arXiv:2102.03539.
8. Chen, X.; Wang, G. Few-shot learning by integrating spatial and frequency representation. In Proceedings of the 18th Conference on Robots and Vision (CRV), Burnaby, BC, Canada, 26–28 May 2021; pp. 49–56.
9. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2017**, 4080–4090. Available online: <https://dl.acm.org/doi/10.5555/3294996.3295163> (accessed on 25 October 2022).
10. Chobola, T.; Vařata, D.; Kondik, P. Transfer learning based few-shot classification using optimal transport mapping from preprocessed latent space of backbone neural network. *AAAI Workshop Meta-Learn. Meta-DL Chall. PMLR* **2021**, 29–37. [[CrossRef](#)]
11. Hu, Y.; Pateux, S.; Gripon, V. Squeezing Backbone Feature Distributions to the Max for Efficient Few-Shot Learning. *Algorithms* **2022**, *15*, 147. [[CrossRef](#)]
12. Bateni, P.; Barber, J.; Van de Meent, J.W.; Wood, F. Enhancing few-shot image classification with unlabelled examples. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, New Orleans, LA, USA, 18–24 June 2022; pp. 2796–2805.
13. Bendou, Y.; Hu, Y.; Lafargue, R.; Lioi, G.; Padeloup, B.; Pateux, S.; Gripon, V. EASY: Ensemble Augmented-Shot Y-shaped Learning: State-Of-The-Art Few-Shot Classification with Simple Ingredients. *arXiv* **2022**, arXiv:2201.09699.
14. Shalam, D.; Korman, S. The Self-Optimal-Transport Feature Transform. *arXiv* **2022**, arXiv:2204.03065.
15. Chen, D.; Chen, Y.; Li, Y.; Mao, F.; He, Y.; Xue, H. Self-supervised learning for few-shot image classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 1745–1749.
16. Ravi, S.; Larochelle, H. Optimization as a Model for Few-Shot Learning. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
17. Bertinetto, L.; Henriques, J.F.; Torr, P.H.; Vedaldi, A. Meta-learning with differentiable closed-form solvers. *arXiv* **2018**, arXiv:1805.08136.
18. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 12–18 July 2020; pp. 1597–1607.
19. Chen, X.; He, K. Exploring Simple Siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15750–15758.
20. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Daniel Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent—a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2021**, 21271–21284. Available online: <https://dl.acm.org/doi/abs/10.5555/3495724.3497510> (accessed on 25 October 2022).
21. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the International Conference on Machine Learning, Seoul, Korea, 18–24 July 2021; pp. 12310–12320.
22. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
23. Wightman, R.; Touvron, H.; Jégou, H. Resnet strikes back: An improved training procedure in timm. *arXiv* **2021**, arXiv:2110.00476.
24. Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Jiang, Z.; Hou, Q.; Feng, J. Deepvit: Towards deeper vision transformer. *arXiv* **2021**, arXiv:2103.11886.
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
26. Breiki, F.A.; Ridzuan, M.; Grandhe, R. Self-Supervised Learning for Fine-Grained Image Classification. *arXiv* **2021**, arXiv:2107.13973.
27. Hu, Y.; Pateux, S.; Gripon, V. Adaptive Dimension Reduction and Variational Inference for Transductive Few-Shot Classification. *arXiv* **2022**, arXiv:2209.08527.
28. Singh, A.; Jamali-Rad, H. Transductive Decoupled Variational Inference for Few-Shot Classification. *arXiv* **2022**, arXiv:2208.10559.