PREPRINT of:
K. Bjerge, P. Bodesheim, H. Karstoft: Deep Image Clustering with Model-Agnostic Meta-Learning.
International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), pages 286–297, 2025.

The final publication is available at https://www.scitepress.org:
https://www.scitepress.org/Papers/2025/131146/131146.pdf
https://dx.doi.org/10.5220/0013114600003912

Deep Image Clustering with Model-Agnostic Meta-Learning

Kim Bjerge[1][a], Paul Bodesheim[2][b] and Henrik Karstoft[1][c]
[1]*Department of Electrical and Computer Engineering, Aarhus University, Finlandsgade 22, 8200 Aarhus N, Denmark*
[2]*Computer Vision Group, Friedrich Schiller University, Ernst-Abbe-Platz 2, 07743 Jena, Germany*
{*kbe,hka*}*@ece.au.dk, paul.bodesheim@uni-jena.de*

Keywords: Deep Clustering, Episodic Training, Few-Shot Learning, Multivariate Loss, Semi-supervised Learning

Abstract: Deep clustering has proven successful in analyzing complex, high-dimensional real-world data. Typically, features are extracted from a deep neural network and then clustered. However, training the network to extract features that can be clustered efficiently in a semantically meaningful way is particularly challenging when data is sparse. In this paper, we present a semi-supervised method to fine-tune a deep learning network using Model-Agnostic Meta-Learning, commonly employed in Few-Shot Learning. We apply episodic training with a novel multivariate scatter loss, designed to enhance inter-class feature separation while minimizing intra-class variance, thereby improving overall clustering performance. Our approach works with state-of-the-art deep learning models, spanning convolutional neural networks and vision transformers, as well as different clustering algorithms like K-means and Spectral clustering. The effectiveness of our method is tested on several commonly used Few-Shot Learning datasets, where episodic fine-tuning with our multivariate scatter loss and a ConvNeXt backbone outperforms other models, achieving adjusted rand index scores of 89.7% on the EU moths dataset and 86.9% on the Caltech birds dataset, respectively. Hence, our proposed method can be applied across various practical domains, such as clustering images of animal species in biology.

1 INTRODUCTION

In real-world applications such as biology and ecology, training supervised networks to classify rare species from images presents significant challenges (Mora et al., 2011; Binta Islam et al., 2023). In many cases, images of these species are either scarce or non-existing, making unsupervised methods such as deep image clustering a valuable alternative.

Clustering is a core problem in unsupervised learning, with traditional methods like K-means (Macqueen, 1967) and Spectral clustering (Von Luxburg, 2007) used to group data into clusters where similar data points are close together and dissimilar points are far apart. However, these classical methods often operate in the feature space of hand-crafted features, which may not capture the underlying structure of the data, limiting their effectiveness. To overcome this, deep clustering techniques have emerged (Bo et al., 2020; Sun et al., 2022; Huang et al., 2024; Lu et al., 2024), leveraging deep neural networks to learn an embedding that better reflects the intrinsic data structure before applying clustering. For instance, deep embedded clustering uses auto-encoders and Kullback-Leibler (KL) divergence minimization to improve clustering accuracy (Xie et al., 2016), while other methods use Convolutional Neural Networks (CNN) to extract latent representations (Yang et al., 2016). Since deep clustering is an unsupervised method it still may be misled by noisy or complex data in the absence of labels.

A natural solution to this limitation is to incorporate some supervised information, leading to semi-supervised deep clustering (Ren et al., 2019; Qin et al., 2019; Cai et al., 2023). In real-world applications, individual labels are often hard to obtain, but pairwise relations between data points are more accessible. For example, in face recognition, while the identity (label) of a person may be unknown, it is often easy to determine whether two images represent the same individual (Chopra et al., 2005). The learning process minimizes a contrastive loss function that drives the similarity metric to be small for pairs of faces from the same person, and large for pairs from different persons. Semi-supervised methods use pairwise constraints — whether two data points belong to the same class — to guide clustering. For example,

[a] https://orcid.org/0000-0001-6742-9504
[b] https://orcid.org/0000-0002-3564-6528
[c] https://orcid.org/0000-0003-3739-8983

Vilhagra et al. proposed a Siamese network (Vilhagra et al., 2020), which has been applied with Spectral clustering for single-cell RNA sequencing data (Jiang et al., 2022). While this approach can enhance performance, its success largely depends on the quality of the pairwise constraints, and selecting them appropriately is a difficult challenge.

In this paper, we propose using Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017), a technique commonly applied in Few-Shot Learning (FSL) (Wang et al., 2020; Song et al., 2023), to address the challenge of clustering with limited labeled data using a set of data points. FSL, which involves training models using only a few labeled examples, employs a support and query set for episodic training (Li et al., 2019), and has gained substantial attention for its ability to enable models to learn from sparse data (Wang et al., 2020). FSL bridges the gap between traditional deep learning, which requires large labeled datasets, and the growing need for systems that can quickly adapt to new tasks or domains.

In our approach, we utilize a small amount of labeled data, such as similar images from a limited set of categories, to improve clustering performance. This method is particularly relevant in domains like biology, where clustering rare or unseen species is crucial (Bjerge et al., 2023). By merging FSL with clustering, we aim to develop models that can generalize beyond their training data, addressing the critical need for adaptability in real-world applications.

In domains where labeled data is available for some classes but absent for others, our proposed method offers a viable solution by clustering data, unlike supervised classifiers, which struggle to handle unknown data samples.

Contribution Our paper introduces a novel methodology specifically tailored for clustering samples within the framework of few-shot learning, utilizing episodic training to achieve domain generalization (Li et al., 2019; Ren et al., 2018). The approach is designed to cluster samples from novel classes not included in the FSL support set.

Our proposal uses transductive inference with the Prototypical Network during meta-learning with episodic training of models to tackle the challenges posed by limited labeled data.

To further improve clustering performance, we propose a novel multivariate scatter loss function, extending the univariate scatter loss introduced by (Bjerge et al., 2024). This function is strategically applied during episodic training to disentangle and separate the distribution of classes within the similarity space. The paper evaluates four state-of-the-art deep learning networks fine-tuned with episodic training and the multivariate scatter loss evaluated with the K-means and Spectral clustering algorithms.

2 RELATED WORK

2.1 Deep Image Clustering

Deep clustering has emerged as a powerful technique for analyzing complex, high-dimensional data. A variety of methods have been proposed to enhance clustering performance, especially in cases where labeled data is scarce or non-existent. Deep Embedded Clustering (DEC), introduced by Xie et al. (Xie et al., 2016), is one of the state-of-the-art approaches in deep clustering. However, DEC does not leverage prior knowledge to guide the learning process. To address this limitation, Ren et al. (Ren et al., 2019) propose Semi-Supervised Deep Embedded Clustering (SDEC), which incorporates labeled data to improve clustering outcomes. This highlights the importance of integrating supervision in deep clustering for improved performance. Bo et al. (Bo et al., 2020) further advanced the field by introducing the Structural Deep Clustering Network (SDCN), which integrates structural information into deep clustering. They point out that most deep clustering methods rely on the neural network ability to learn effective feature representations, often through auto-encoders.

While deep clustering methods have shown great promise, Sun et al. (Sun et al., 2022) note that the absence of labels often results in unreliable clustering. To mitigate this, they propose Deep Active Clustering, which actively selects key data points for human labeling. This novel approach improves the clustering process by intelligently choosing the most informative samples for annotation, addressing one of the limitations of conventional semi-supervised deep clustering methods that rely on fixed, pre-labeled data.

The evolution of deep clustering methods has also been tracked by Lu et al. (Lu et al., 2024), who offer a comprehensive review of prior knowledge used in deep clustering. Huang et al. (Huang et al., 2024) provide a general framework for Deep Image Clustering Networks, outlining the key stages: image preprocessing, embedding, feature processing, clustering, and result processing. The challenges posed by big data clustering are also highlighted by Fahad et al. (Fahad et al., 2014). They offer a survey of existing clustering algorithms and a comparison of their theoretical and empirical performance.

Rodriguez et al. (Rodriguez et al., 2019) performed a systematic comparison of nine well-known

clustering methods, offering insights into their relative performance across different datasets. Additionally, various classic clustering algorithms such as Density-Based Clustering (Ester et al., 1996; Ankerst et al., 1999), EM Algorithms (Dempster et al., 1977), Hierarchical Density-Based Clustering (HDBSCAN) (McInnes et al., 2017) and Spectral Clustering (Von Luxburg, 2007) have been explored in the context of clustering of large-scale data. Jain et al. (Jain, 2010) summarize these methods, discussing key challenges like feature selection, semi-supervised clustering, and clustering at scale.

2.2 Few-Shot Learning

Few-Shot Learning comprising diverse approaches tailored to address the challenges inherent in learning from limited labeled data. Few-shot learning can be categorized into two distinct branches: *inductive FSL* and *transductive FSL*. The former involves the prediction of test samples individually, while the latter addresses the prediction of test samples collectively.

For a comprehensive overview of the evolving FSL landscape, we refer to surveys such as those presented in works by Wang et al. (Wang et al., 2020) and Song et al. (Song et al., 2023). A notable category of methods relevant to our proposed approach leverage euclidean distance and cosine similarity as the fundamental measure. This includes Prototypical Networks (Snell et al., 2017), Finetune (Chen et al., 2019), Transductive Information Maximization (Boudiaf et al., 2020), and Prototypical Rectification (Liu et al., 2020).

Meta-learning is a key paradigm in the FSL landscape, and a comprehensive survey on the subject is presented by Hospedales et al. (Hospedales et al., 2022). Additionally, the integration of episodic training for domain generalization, as discussed by Li et al. (Li et al., 2019) and Model-Agnostic Meta-Learning (Finn et al., 2017), emerges as a crucial aspect in enhancing the adaptability and robustness of FSL models.

In few-shot classification, we are given a small support set of $(N \cdot K)$ labeled examples $S = \{(X_1^{(1)}, y_1), ..., (X_N^{(K)}, y_K)\}$ where each $X_n^{(k)} \in \mathbb{R}^D$ is the D-dimensional embedded feature vector of an example and $y_k \in \{1, ..., K\}$ are the corresponding labels (Bjerge et al., 2024). The set $S_k = \{(X_1^{(k)}, y_k), ..., (X_N^{(k)}, y_k)\}$ denotes the subset of examples labeled with class k and the number of class labels in the support set is denoted K-way where we have N-shots of examples in each S_k. A query set Q contains sample images q_i that belong to classes in the support set where the goal is to match the query samples to the correct class label. When training with a dataset that comprises more classes than those included in the support set, a random subset of K classes is selected for each few-shot task, which encompasses both the support and query sets. We propose to use the Prototypical Network (Snell et al., 2017), which employs the Euclidean distance to compare the center point of the support set with the query samples.

In conclusion, FSL and deep clustering has made significant strides, leveraging powerful feature extraction methods from deep learning and incorporating structural and active learning techniques to enhance clustering performance. Despite these advancements, challenges remain, particularly in large-scale data clustering and the integration of prior knowledge, which continue to drive innovation in the field.

3 METHOD

3.1 Deep Clustering with Few-Shot Learning

We propose clustering images by utilizing feature embeddings derived from the output vector of a deep learning (DL) model. The DL model is fine-tuned through episodic training, a technique commonly used in FSL (Li et al., 2019). We introduce a multivariate loss function to enhance feature clustering during episodic training. The pipeline, illustrated in Figure 1, consists of three steps.

Step 1 involves training and fine-tuning a pretrained DL model to extract features that effectively represent the images, ensuring that the embedding space is spread across diverse clusters, as visualized in Figure 2. To enhance the quality of these feature embeddings, we leverage MAML (Finn et al., 2017; Hospedales et al., 2022), a technique commonly used in FSL. MAML uses a training and validation dataset with classes of images that are not present in the test dataset, allowing the model to generalize well to new, unseen classes.

In our study, we selected four DL models that span both convolutional neural networks (CNNs) and vision transformers (ViTs). We chose ResNet50v2 (He et al., 2016) and EfficientNetB3 (Tan and Le, 2019) as they are widely used CNNs, achieving top-1 accuracies of 76.0% and 81.6% on the ImageNet dataset (Russakovsky et al., 2015), respectively. Furthermore, we included ConvNext-B (Todi et al., 2023), a recently published model that achieves an ImageNet top-1 accuracy of 85.3%, and the vision

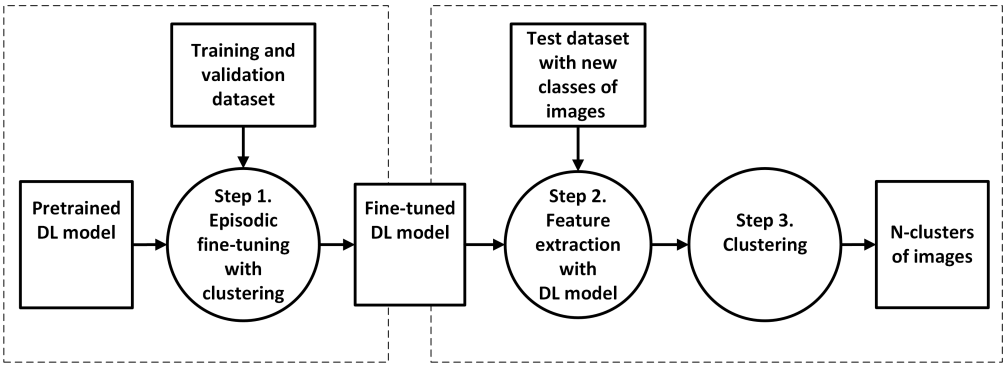
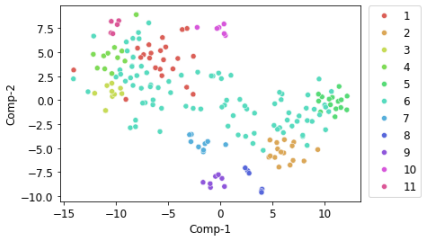
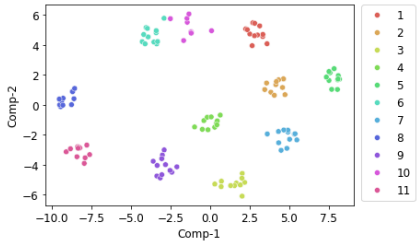


Figure 1: Illustrates the pipeline of our proposed deep clustering method for effectively grouping images in few-shot scenarios. A training and validation dataset with different classes is used to fine-tune a deep learning model for feature extraction. In step 1. the model is fine-tuned with few-shot learning and episodic training where the best model is chosen based on K-means clustering of the validation dataset. In Step 2 and 3, the extracted features are fed into a clustering algorithm, which groups a set of new test images into N clusters.



(a) Pretrained DL model



(b) Fine-tuned DL model

Figure 2: T-SNE projection of feature embeddings for clusters with a pretrained DL model and a fine-tuned DL model with episodic training and multivariate scatter loss. The scatter plot illustrates the clustering of the first and second components of 11 classes in the embedding space.

transformer ViT-B/16 (Dosovitskiy et al., 2021) with an accuracy of 88.6%.

Step 2 concerns extracting features from the fine-tuned DL model. The feature vector is extracted from the backbone of the trained DL model. The backbone architecture includes multiple convolutional layers or vision transformer layers, with the final head of fully connected layers removed. The test dataset contains image categories that are not present in the training or validation datasets. Additionally, the categories in the training and validation datasets are distinct from each other as well. The size of the feature vector depends on the used DL model.

Step 3 involves clustering of feature embeddings extracted with the DL model. Clustering analysis involves grouping a set of objects such that those within the same cluster are more similar to each other than to objects in other clusters. In our study, the objects being clustered are unlabeled images, which could include images of various unlabeled categories. In ecological studies, this could involve species such as animals, birds, or insects for which labels are sparse or entirely absent. The primary challenge is to learn effective feature embeddings of training images that can serve as input to a clustering algorithm.

For this purpose, we utilize K-means (Macqueen, 1967) and Spectral clustering (Von Luxburg, 2007), assuming the number of clusters is known. These two clustering methods were selected based on a comparative analysis of nine common clustering methods by Rodriguez et al. (Rodriguez et al., 2019). K-means is a well-known and widely used method, while Spectral clustering demonstrated superior performance, achieving an adjusted rand index of 68.16% as the best out of nine different selected clustering algorithms.

Unlike K-means, which works well for convex clusters, spectral clustering can capture more complex, non-linear relationships in the data. Spectral clustering is a graph-based approach and leverages the spectral (eigenvalue) properties of a similarity graph constructed from the data to identify clusters, rather than relying on traditional distance-based metrics like

K-means. It effectively detects clusters of arbitrary shapes and sizes, particularly in cases where data points are not well-separated by traditional distance metrics.

While K-means and Spectral clustering are central to our demonstration, other clustering methods could also be employed. These include Fast Density-Based Clustering (DBSCAN) (Ester et al., 1996), Expectation-Maximization (EM) (Dempster et al., 1977), or hierarchical methods such as agglomerative and divisive clustering, as explored by Jain et al. (Jain, 2010). For unsupervised clustering where the number of clusters is unknown, methods like Ordering Points To Identify the Clustering Structure (OPTICS) (Ankerst et al., 1999) and HDBSCAN (McInnes et al., 2017) could be utilized.

A solution to clustering of images not included in training would benefit many real-life applications where the number of samples for each class is sparse or labelled data was non-existent during training.

3.2 Episodic Training with Multivariate Scatter Loss

In few-shot learning, episodic training with domain generalization (DG) (Li et al., 2019), also known as meta-learning (Hospedales et al., 2022), involves training with a set of tasks during each epoch. Each task comprises several episodes, with each episode containing a labeled support set and a query set drawn from the training dataset. After each epoch, the model accuracy is evaluated using the validation dataset, which contains tasks with class categories different from those in the training dataset, thereby assessing the model ability to generalize across domains.

A Prototypical Network uses Euclidean distance as a similarity function to predict the relationship between query samples and class labels in the support set. Experiments have demonstrated that training with Euclidean distance, rather than cosine similarity, yields superior results (Li et al., 2019). In our approach, we use the cross entropy loss together a new a multivariate scatter loss function during training. This new loss function is designed to minimize within-class variance while simultaneously maximizing the mean separation between classes, enhancing the model's discriminative ability. The method is inspired by the work of Bodesheim et al. (Bodesheim et al., 2013) and Bjerge et al. (Bjerge et al., 2024). For an illustration of class distribution after training with multivariate scatter loss, see Figure 2. The multivariant scatter loss is defined as

$$Lm(\theta) = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N_k}(X_n - \overline{X}_k)^T(X_n - \overline{X}_k)}{\sum_{i=1}^{K}\sum_{j=i+1}^{K}(\overline{X}_i - \overline{X}_j)^T(\overline{X}_i - \overline{X}_j)} \quad (1)$$

with X_n being the embedding vector for each sample in the support set S_k of class k. N_k is the number of samples in each class of the support set and θ is the parameters of the DL model. With $\overline{X}_k$, $\overline{X}_i$, and $\overline{X}_j$ we denote the mean center point of samples in the classes k, i, and j of the support set. The numerator of Eq. (1) encourages the norm of the data points, centred with respect to their cluster, to be small, and the denominator encourages the cluster centres to be far apart.

The cross-entropy loss function in Eq. (2) ensures that a query sample q is classified correctly according to the support set during training:

$$Lc_q(\theta) = -\hat{y}_j \log\left(\frac{exp(-d_{q,k_j})}{\sum_{i=1}^{K} exp(-d_{q,k_i})}\right) \quad (2)$$

Here, d_{q,k_j} is the euclidean distance between the support center point k_j and the query sample q and $\hat{y}_j$ is the one-hot encoded vector for the correct label of the query sample. K is the number of classes (K-way) in the support set. Finally a combined loss function is defined in Eq. (3) to prioritize between the cross-entropy and scatter loss:

$$L(\theta) = \alpha Lm(\theta) + (1-\alpha)Lc(\theta) \quad . \quad (3)$$

The goal is to increase the distance between class distributions in the support set while increasing the number of correctly classified query samples related to the support classes. The loss function $L(\theta)$ will prioritize between minimizing the multivariate scatter loss (Lm) and the cross-entropy loss for the batch of query samples (Lc) by adjusting $\alpha \in [0,1]$

3.3 Fine-tuning with Episodic Training

A DL model was used as a backbone to extract feature embeddings. The outputs from the backbone of the DL model were flattened and used as embeddings, leading to D-dimensional feature vectors. In our work, we have trained larger models than used in FSL. The models were fine-tuned using four DL architectures, generating feature vectors with the following dimensions: 2048 (ResNet50v2), 1536 (EfficientNetB3), 1024 (ConvNeXt-B), and 768 (ViT-B/16). For fine-tuning, we used classical pre-trained weights from the ImageNet dataset (Russakovsky et al., 2015). These pre-trained DL models were then fine-tuned with episodic training on a new domain-specific dataset. During all episodic training sessions, a 5-shot K-way support set was utilized. The value of K ranged between 15 and 30 constrained by the DL model and a maximum of 50GB of GPU memory. Training on the datasets was performed using data augmentation, including image scaling, horizontal flip and adding color jitter for brightness, contrast and saturation.

The stochastic gradient descent optimizer (SGD) was used during episodic training and fine-tuning. The SGD was configured with the momentum of 0.9 and a weight decay of $5.0 \cdot 10^{-4}$ using a multi-step scheduler to lower the learning rate at two specified milestones specified by epochs. The first milestone was set to 3 epochs and the second to 6 epochs with a total of 9 epochs. The limited number of epochs is a result of episodic training, with each epoch comprising 300 few-shot tasks, each consisting of 6 query samples per class from the support set. SGD was tested with the initial learning rate of $1.0 \cdot 10^{-3}$ for pre-trained models during fine-tuning.

K-means clustering was performed on all images in the validation dataset after each epoch to select the best performing model. The clustering was applied to the validation datasets described in section 4.1 which included 20, 40, 50, or 97 classes. After each epoch, cluster accuracy (CA) was used to select the optimal model for clustering the validation dataset with K-means.

3.4 Performance Metrics for Clustering

In this section, we briefly outline the criteria used for performance evaluation based on commonly used clustering validation indices (Wu et al., 2019; Huang et al., 2024; Lu et al., 2024). These criteria are analogous to accuracy or recognition rate in supervised learning. Four key evaluation metrics are used include Cluster Accuracy (CA), Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), Normalized Mutual Information (NMI) (Strehl and Ghosh, 2003) and Adjusted Mutual Information (AMI) (Vinh et al., 2009). When the true class labels of a dataset are known, these metrics allow us to assess how accurately a clustering technique partitions the data relative to the correct class labels.

CA measures the percentage of correctly classified images in the clustering solution compared to predefined image class labels (Ω). We use CA as defined by (Fahad et al., 2014):

$$CA = \sum_{i=1}^{K} \frac{max(C_i|L_i)}{|\Omega|} \tag{4}$$

where C_i is the set of instances in the ith cluster. L_i is the class labels for all images in the ith cluster, and $max(C_i|L_i)$ is the number of instances with the majority label in the ith cluster (e.g. if label l appears in the ith cluster more often than any other label, then $max(C_i|L_i)$ is the number of instances in C_i with the label l).

ARI is a measure used to evaluate the similarity between two clustering results, taking into account the possibility of chance. It is an improvement over the basic Rand Index, which measures the proportion of agreement between two clusterings (i.e., the proportion of points that are either clustered together or clustered apart in both clusterings). ARI takes into account the number of instances that exist in the same cluster and different clusters.

NMI is a normalized version of Mutual Information (MI) using the upper bound to score results between 0 (no mutual information) and 1 (perfect correlation).

AMI is an adjustment of the MI score to account for chance. It accounts for the fact that the Mutual Information is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared.

4 EXPERIMENTAL SETUP

The proposed method undergoes training and evaluation using four distinct datasets for FSL that are described in Sec. 4.1. Classic training and fine-tuning of models has also been conducted to compare performances with episodic training. The best model was selected based on cluster accuracy using K-means, evaluated after each epoch for both training methods. The metrics outlined in the tables of Sec. 5 include CA, NMI, AMI, and ARI for clustering of features with classic and episodic training. Metrics in all tables are computed across 5 random runs of clustering feature vectors from fine-tuned DL models, with average (AVG) and standard deviations (SD).

4.1 Datasets

We have selected four datasets with images typically used for FSL, here we use the validation and testing dataset for clustering. These datasets were chosen over commonly used clustering datasets due to their relatively high resolution and the clear separation into distinct classes for training, validation, and testing, as demonstrated in previous studies (Ravi and Larochelle, 2017; Wang et al., 2019; Bjerge et al., 2024). Unlike FSL datasets, standard clustering datasets typically do not provide such distinct class separations.

Mini-ImageNet is a benchmark dataset and is a subset of the larger ILSVRC-12 dataset (Russakovsky et al., 2015). It has a total of 60,000 color images from

Table 1: Shows the metrics (CA, NMI, AMI, ARI) for K-means and spectral clustering of features from fine-tuned models with classic and episodic training with ConvNeXt-B on the four datasets, tested with clustering on test datasets. A best α value above zero indicates that the multivariate scatter loss improved episodic training in finding the best model. The best metric results for each clustering method and dataset is highlighted with bold.

Dataset	Cluster method	Training	Best α	CA AVG (SD)	NMI AVG (SD)	AMI AVG (SD)	ARI AVG (SD)
EU Moths	K-means	Classic	-	0.773 (0.022)	0.872 (0.009)	0.785 (0.014)	0.653 (0.026)
EU Moths	K-means	Episodic	0.1	**0.837** (0.029)	**0.912** (0.005)	**0.851** (0.008)	**0.748** (0.009)
EU Moths	Spectral	Classic	-	0.874 (0.009)	0.922 (0.003)	0.868 (0.006)	0.787 (0.009)
EU Moths	Spectral	Episodic	0.1	**0.940** (0.011)	**0.962** (0.004)	**0.935** (0.007)	**0.897** (0.012)
CUB	K-means	Classic	-	**0.844** (0.019)	0.885 (0.010)	0.864 (0.011)	**0.759** (0.026)
CUB	K-means	Episodic	0.1	0.834 (0.025)	**0.885** (0.005)	**0.864** (0.006)	0.745 (0.038)
CUB	Spectral	Classic	-	**0.936** (0.000)	**0.932** (0.000)	**0.919** (0.000)	**0.881** (0.000)
CUB	Spectral	Episodic	0.0	0.928 (0.002)	0.930 (0.002)	0.917 (0.002)	0.869 (0.004)
mini-ImageNet	K-means	Classic	-	0.895 (0.005)	0.942 (0.001)	0.941 (0.001)	0.853 (0.005)
mini-ImageNet	K-means	Episodic	0.5	**0.900** (0.010)	**0.942** (0.001)	**0.941** (0.001)	**0.856** (0.009)
mini-ImageNet	Spectral	Classic	-	**0.903** (0.003)	**0.941** (0.001)	**0.941** (0.001)	**0.832** (0.006)
mini-ImageNet	Spectral	Episodic	-	0.893 (0.002)	0.937 (0.000)	0.936 (0.000)	0.813 (0.001)
tired-ImageNet	K-means	Classic	-	0.975 (0.024)	0.984 (0.007)	0.984 (0.007)	0.965 (0.028)
tired-ImageNet	K-means	Episodic	0.5	**0.984** (0.021)	**0.986** (0.007)	**0.986** (0.007)	**0.977** (0.023)
tired-ImageNet	Spectral	Classic	-	0.978 (0.018)	0.982 (0.007)	0.982 (0.007)	0.966 (0.023)
tired-ImageNet	Spectral	Episodic	0.2	**0.982** (0.019)	**0.984** (0.007)	**0.984** (0.007)	**0.971** (0.024)

100 classes, where each class has 600 images of size 224x224 (Vinyals et al., 2016). In alignment with established practices (Ravi and Larochelle, 2017; Wang et al., 2019), we adopt a partitioning scheme consisting of 60 training classes, complemented by 20 validation classes and 20 test classes.

Tiered-ImageNet (Ren et al., 2018) is a larger subset of the larger ILSVRC-12 dataset (Russakovsky et al., 2015). It contains 608 classes with 779,165 images partitioned into disjoint sets for training (351 classes), validation and testing. Images for evaluation are split into 97 classes for validation and 160 classes for testing.

Caltech Birds-200-2011 (CUB) (Wah et al., 2011) is a fine-grained image classification dataset. We adopt Chen et al. (Chen et al., 2019) for few-shot classification on CUB, which splits into 120 training, 40 validation, and 40 test classes with a total of 11,788 images for the experiments. To maintain consistency and comparability with mini-ImageNet, the images from CUB are uniformly resized to 224x224 pixels.

The EU Moths dataset, as introduced by Böhlke et al. (Böhlke et al., 2021), encapsulates a collection of 200 moth species prevalent in Central Europe. Notably, each species is delineated by few samples, comprising a mere 11 images, resulting in a total of 2,205 images within the dataset. This dataset was chosen for our evaluation of the multivariate scatter loss because it presents the most challenging for clustering, with only a few samples per class and species with highly similar appearances. The images have high resolution, even surpassing 1000x1000 pixels, and have been resized to 224x224 pixels. The dataset is split into 100 training, 50 validation, and 50 test classes for experiments.

5 RESULTS AND DISCUSSION

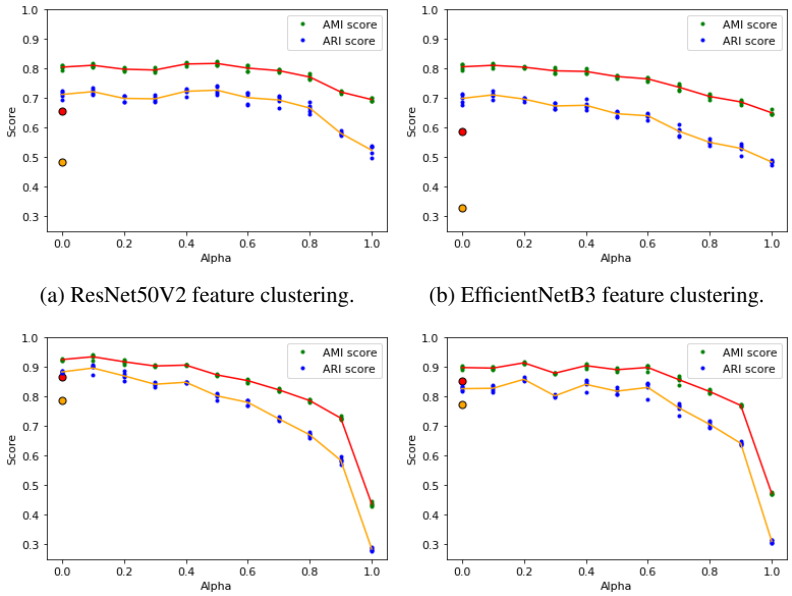
Several experiments were performed on the datasets to evaluate clustering and episodic training with the multivariate scatter loss function. Here, results were obtained with pre-trained models on ImageNet and fine-tuning with classic and episodic training on the four datasets. Since mini-ImageNet and tiered-ImageNet are a subset of ImageNet we only expect to see minor improvements when fine-tuning the models with classic or episodic training.

Detailed results as shown in Table 1 for training models with ResNet50v2, EfficientNetB3, ConvNeXt-B and ViT-B/16 can be found on Github[1] with the source code for the experimental results.

5.1 Evaluation of the Multivariate Scatter Loss

Spectral clustering of features from fine-tuned models, using both classic and episodic training on the EU moth dataset, is shown in Figure 3. The scores represent the average of five runs. It is important to emphasize that the optimal model, along with the corre-

[1] https://github.com/kimbjerge/few-shot-clustering



(a) ResNet50V2 feature clustering.

(b) EfficientNetB3 feature clustering.

(c) ConvNeXt-B feature clustering.

(d) ViT-B/16 feature clustering.

Figure 3: Shows the ARI and AMI metrics with Spectral clustering of features from fine-tuned models on the EU moth dataset with different α values. Each green (AMI) and blue (ARI) dot represents a clustering result out of five random runs for each metric. The red (AMI) and yellow (ARI) circles at Alpha=0 are the metrics for the classic trained models.

sponding α values, was determined through episodic fine-tuning using the validation sets.

All four models achieved higher AMI and ARI scores with episodic training compared to fine-tuning with classic training. The highest relative increase in performance with episodic training is observed for ResNet50v2 and EfficientNetB3. However, ConvNeXt-B and ViT-B/16 achieved the highest AMI and ARI scores overall. As the value of α increases, both AMI and ARI scores tend to decrease. Nonetheless, the highest ARI scores were observed with non-zero α values at: $\alpha = 0.5$ (ResNet50v2), $\alpha = 0.1$ (EfficientNetB3), $\alpha = 0.1$ (ConvNeXt-B) and $\alpha = 0.2$ (ViT-B/16). This indicates that the multivariate scatter loss positively impacts episodic training, leading to improved Spectral clustering performance. However, there is only very little improvement compared to $\alpha = 0$ for all networks which limits the impact of the scatter loss.

In Table 1 and on Github[1], we present the CA, NMI, AMI, and ARI metrics for both classic and episodic fine-tuning of models, with clustering performed using K-means and Spectral clustering. For K-means clustering with $\alpha = 0.1$, the best results were achieved with episodic fine-tuning of the ResNet50v2, ConvNeXt-B, and ViT-B/16 models. However, for EfficientNetB3, the multivariate scatter loss did not result in performance improvement for K-means clustering.

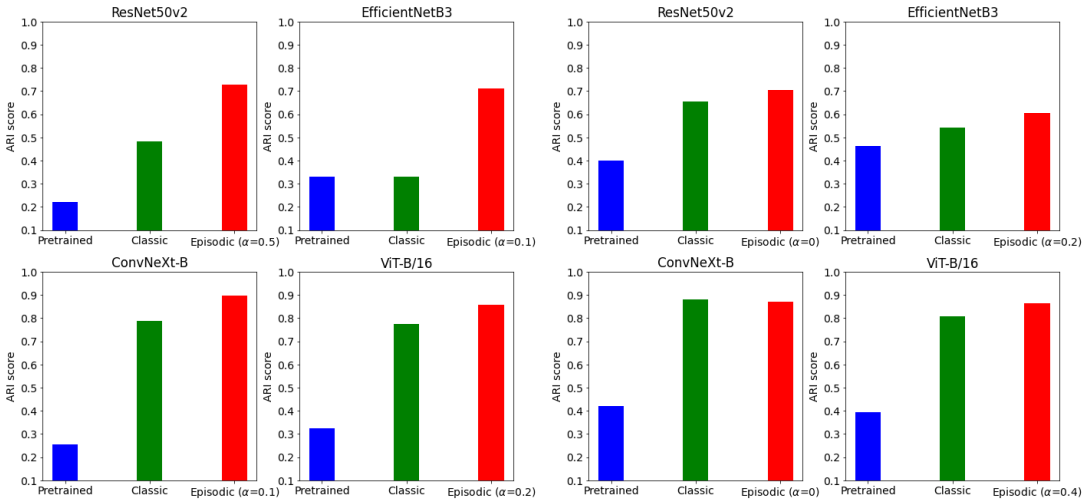
Fine-tuning models on the mini-ImageNet dataset demonstrates that the multivariate scatter loss indicates minor improved model performance, with the episodic fine-tuned models achieving the best average ARI of 0.977 and AMI of 0.984 with $\alpha = 0.5$ and ConvNeXt-B. In two cases, using ResNet50v2 and ViT-B/16, the highest scores were obtained with Spectral clustering, solely by applying multivariate scatter loss ($\alpha = 1.0$).

5.2 Clustering with the EU moths and CUB datasets

Figure 4 summarizes the ARI scores for Spectral clustering of features from models pre-trained on ImageNet and fine-tuned using either classic or episodic training on the EU moths and CUB datasets.

As expected, the fine-tuned models outperform the pre-trained models, as the fine-tuning process adapts the models to datasets within the same domain, in our case different species of moths or birds.

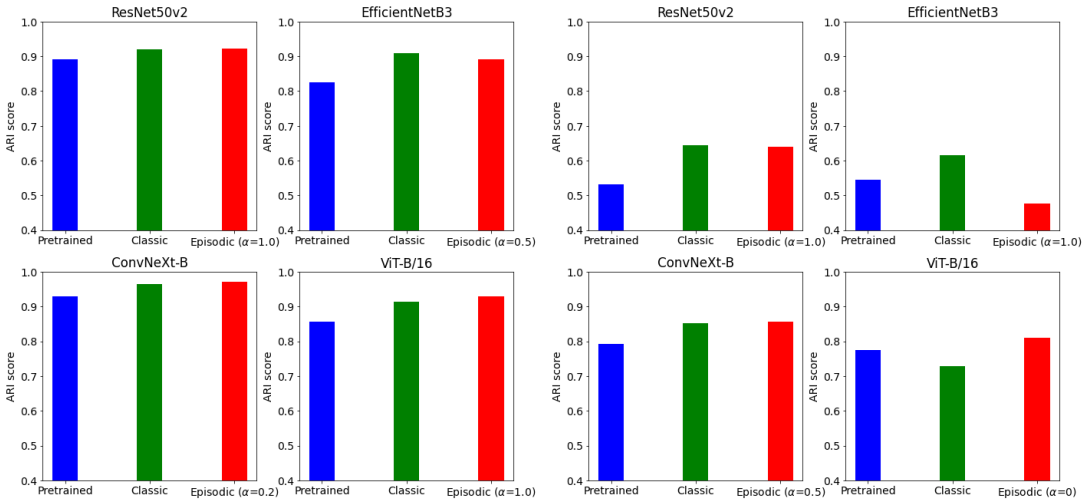
Notably, models fine-tuned with episodic training consistently outperform those trained with clas-



(a) EU moths with Spectral clustering.

(b) CUB dataset with Spectral clustering.

Figure 4: Shows the ARI score with Spectral clustering of features from pre-trained and fine-tuned models on the EU moth and CUB dataset. The blue bars presents clustering of features from classic pre-trained models on ImageNet. The green bars presents clustering of features from classic fine-tuned models on the respective dataset. The red bars presents clustering of features from the best episodic fine-tuned models.



(a) Mini-ImageNet - Spectral clustering.

(b) TieredImageNet - K-means clustering.

Figure 5: Shows the ARI score with Spectral and K-means clustering of features from pre-trained and fine-tuned models on the mini-ImageNet and tiered-ImagneNet dataset. The blue bars presents clustering of features from classic pre-trained models on ImageNet. The green bars presents clustering of features from classic fine-tuned models on the respective dataset. The red bars presents clustering of features from the best episodic fine-tuned models.

sic fine-tuning. This suggests that episodic training enhances the models' ability to generalize and adapt to new domains, such as switching between different species in the moth and bird datasets. For the CUB dataset, the performance of ConvNeXt-B is the same for both classic and episodic training. However, this is not the case for the EU moths dataset, which has fewer samples per class. The stronger performance of episodic training on the EU moths dataset highlights its effectiveness, particularly when working with datasets that have limited samples per class.

Table 1 presents the CA, NMI, AMI, and ARI metrics for K-means and Spectral clustering of features extracted from fine-tuned ConvNeXt-B models using both classic and episodic training of the EU moths, CUB, mini-ImageNet and tiered-ImageNet datasets. Metrics in all tables are computed across 5 random runs of clustering feature vectors from fine-tuned DL models, with average (AVG) and standard deviations (SD). The best results were obtained using Spectral clustering on features from a ConvNeXt-B model trained with episodic learning. For both datasets, Spectral clustering outperforms K-means clustering with an increase of 8% - 14% on all clustering metrics.

On the EU moths dataset, the best model achieved NMI=0.962, AMI=0.935, and ARI=0.897 with multivariate scatter loss ($\alpha = 0.1$). For the CUB dataset, ConvNeXt-B model reached NMI=0.930, AMI=0.917, and ARI=0.869 with $\alpha = 0.0$. However, classic fine-tuning yielded slightly better results, with an ARI of 0.881. All scores are reported as the average of 5 runs. The CUB dataset contains more samples compared to the EU moths dataset, suggesting that the multivariate scatter loss has a greater impact with episodic training on datasets with fewer data samples per class.

5.3 Clustering with the mini-ImageNet and tiered-ImageNet datasets

Figure 5 summarizes the ARI scores for Spectral and K-means clustering of features extracted from models pre-trained on ImageNet and fine-tuned using either classic or episodic training on the mini-ImageNet and tiered-ImageNet datasets. As expected, both classic and episodic fine-tuning outperform the pre-trained models. While episodic fine-tuning generally yields slightly better results than classic fine-tuning, EfficientNetB3 is an exception, where classic fine-tuning is the best.

The multivariate scatter loss enhances episodic training even without the use of cross-entropy loss ($\alpha = 1.0$). Given that the models were pre-trained on ImageNet, the multivariate scatter loss appears to improve the distribution of features into well-defined clusters, particularly when the model has already been pre-trained on the same classes.

The test dataset for tiered-ImageNet contains 160 classes, and Spectral clustering requires several hours (10-20) to process. As a result, only the performance of ConvNeXt-B is included in the detailed results on Github[1]. Interestingly, K-means clustering outperforms Spectral clustering, achieving an ARI of 0.856 compared to 0.813 with Spectral clustering. This suggests that while Spectral clustering excels with test datasets containing 20-50 classes, K-means performs better when the number of classes increases above 150, it might be the preferred method for larger datasets.

6 CONCLUSION

In this study, we present a novel method for clustering images using Model-Agnostic Meta-Learning within the context of few-shot learning and episodic training with multivariate scatter loss. The proposed method was evaluated on four commonly used few-shot learning datasets, employing four state-of-the-art DL models for feature extraction. ConvNeXt-B outperformed the other networks, achieving ARI scores of 0.897 and 0.869 on the EU moths and Caltech Birds datasets, respectively. On the mini-ImageNet and tiered-ImageNet datasets, episodic fine-tuned models with multivariate scatter loss further improved clustering performance, with ARI scores of 0.977 and 0.856 ($\alpha = 0.5$). The multivariate scatter loss consistently enhanced clustering performance during episodic fine-tuning across most experiments, particularly on the EU moths dataset, where its effectiveness was demonstrated with only 11 samples per class. This highlights the method's potential in handling datasets with limited samples.

We explored two commonly used clustering algorithms: K-means and Spectral clustering. On the EU moths and Caltech Birds datasets, Spectral clustering outperformed K-means, with an 8% to 14% improvement across all clustering metrics. However, on the tiered-ImageNet dataset, which contains a large number of classes (160), K-means clustering delivered the best results. Future experiments could explore additional clustering algorithms, especially when applied to datasets with few samples in each class.

Our solution represents an advancement in tackling the challenge of image clustering, especially in real-world scenarios where class samples are sparse. The demonstrated accuracy and effectiveness of our

proposed method highlight its potential as a valuable tool in settings with limited labeled data, offering promising applications across a wide range of practical domains such as clustering of images of animal species in biology.

Acknowledgement

During the preparation of this work, the first author utilized ChatGPT (OpenAI, 2023) to enhance the clarity and formulation for parts of the written text. However, the authors takes full responsibility for the content of the publication.

REFERENCES

Ankerst, M., Breunig, M. M., Kriegel, H. P., and Sander, J. (1999). OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 28(2).

Binta Islam, S., Valles, D., Hibbitts, T. J., Ryberg, W. A., Walkup, D. K., and Forstner, M. R. (2023). Animal Species Recognition with Deep Convolutional Neural Networks from Ecological Camera Trap Images. *Animals*, 13(9).

Bjerge, K., Bodesheim, P., and Karstoft, H. (2024). Few-shot learning with novelty detection. In Fred, A., Hadjali, A., Gusikhin, O., and Sansone, C., editors, *Deep Learning Theory and Applications*, pages 340–363, Cham. Springer Nature Switzerland.

Bjerge, K., Geissmann, Q., Alison, J., Mann, H. M., Høye, T. T., Dyrmann, M., and Karstoft, H. (2023). Hierarchical classification of insects with multitask learning and anomaly detection. *Ecological Informatics*, 77:102278.

Bo, D., Wang, X., Shi, C., Zhu, M., Lu, E., and Cui, P. (2020). Structural Deep Clustering Network. In *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*.

Bodesheim, P., Freytag, A., Rodner, E., Kemmler, M., and Denzler, J. (2013). Kernel null space methods for novelty detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Böhlke, J., Korsch, D., Bodesheim, P., and Denzler, J. (2021). Exploiting Web Images for Moth Species Classification. In *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI)*, volume P-314.

Boudiaf, M., Masud, Z. I., Rony, J., Dolz, J., Piantanida, P., and Ayed, I. B. (2020). Transductive information maximization for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 2020-December.

Cai, J., Hao, J., Yang, H., Zhao, X., and Yang, Y. (2023). A review on semi-supervised clustering. *Information Sciences*, 632.

Chen, W. Y., Wang, Y. C. F., Liu, Y. C., Kira, Z., and Huang, J. B. (2019). A closer look at few-shot classification. In *7th International Conference on Learning Representations, ICLR 2019*.

Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, volume I.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm . *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(1).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16X16 words: Transformers for image recognition at scale. In *ICLR 2021 - 9th International Conference on Learning Representations*.

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings - 2nd International Conference on Knowledge Discovery and Data Mining, KDD 1996*.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., and Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3).

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *34th International Conference on Machine Learning, ICML 2017*, volume 3.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. (2022). Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9).

Huang, H., Wang, C., Wei, X., and Zhou, Y. (2024). Deep image clustering: A survey. *Neurocomputing*, 599:128101.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1).

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8).

Jiang, H., Huang, Y., and Li, Q. (2022). Spectral clustering of single cells using Siamese nerual network combined with improved affinity matrix. *Briefings in Bioinformatics*, 23(3).

Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y. Z., and Hospedales, T. (2019). Episodic training for domain generalization. In *Proceedings of the IEEE Interna-*

tional Conference on Computer Vision, volume 2019-October.

Liu, J., Song, L., and Qin, Y. (2020). Prototype Rectification for Few-Shot Learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12346 LNCS:741–756.

Lu, Y., Li, H., Li, Y., Lin, Y., and Peng, X. (2024). A survey on deep clustering: from the prior perspective. *Vicinagearth*, 1(1):1–17.

Macqueen, J. (1967). Some methods for classification and analysis of multivarite observation. Preceeding of the 5th Berkeley symposium on mathematical statistics and probability, Berkeley. *University of california press*, 281.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11).

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., and Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biology*, 9(8).

OpenAI (2023). ChatGPT 3.5.

Qin, Y., Ding, S., Wang, L., and Wang, Y. (2019). Research Progress on Semi-Supervised Clustering.

Ravi, S. and Larochelle, H. (2017). Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*.

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H., and Zemel, R. S. (2018). Meta-learning for semi-supervised few-shot classification. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Ren, Y., Hu, K., Dai, X., Pan, L., Hoi, S. C., and Xu, Z. (2019). Semi-supervised deep embedded clustering. *Neurocomputing*, 325.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., and Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLoS ONE*, 14(1).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3).

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 2017-December.

Song, Y., Wang, T., Cai, P., Mondal, S. K., and Sahoo, J. P. (2023). A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities. *ACM Computing Surveys*, 55(13s).

Strehl, A. and Ghosh, J. (2003). Cluster ensembles - A knowledge reuse framework for combining multiple partitions. In *Journal of Machine Learning Research*, volume 3.

Sun, B., Zhou, P., Du, L., and Li, X. (2022). Active deep image clustering. *Knowledge-Based Systems*, 252.

Tan, M. and Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *36th International Conference on Machine Learning, ICML 2019*, volume 97, pages 6105–6114.

Todi, A., Narula, N., Sharma, M., and Gupta, U. (2023). ConvNext: A Contemporary Architecture for Convolutional Neural Networks for Image Classification. In *Proceedings - 2023 3rd International Conference on Innovative Sustainable Computational Technologies, CISCT 2023*.

Vilhagra, L. A., Fernandes, E. R., and Nogueira, B. M. (2020). TextCSN: A semi-supervised approach for text clustering using pairwise constraints and convolutional siamese network. In *Proceedings of the ACM Symposium on Applied Computing*.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4).

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). Caltech-UCSD Birds-200-2011 (CUB-200-2011). *California Institute of Technology*, CNS-TR-2011-001.

Wang, Y., Chao, W.-L., Weinberger, K. Q., and van der Maaten, L. (2019). SimpleShot: Revisiting Nearest-Neighbor Classification for Few-Shot Learning. *arXiv*.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, 53(3).

Wu, J., Long, K., Wang, F., Qian, C., Li, C., Lin, Z., and Zha, H. (2019). Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October.

Xie, J., Girshick, R., and Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In *33rd International Conference on Machine Learning, ICML 2016*, volume 1.

Yang, J., Parikh, D., and Batra, D. (2016). Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December.