Better Generalized Few-Shot Learning Even without Base Data

Seong-Woong Kim, Dong-Wan Choi*

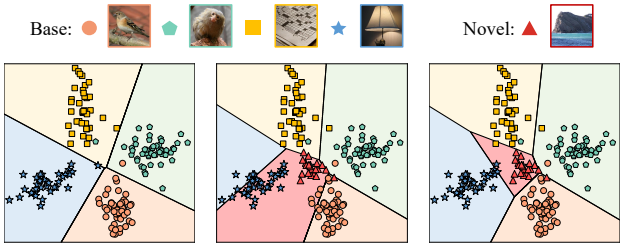
Department of Computer Science and Engineering, Inha University, South Korea
wauri6@gmail.com, dchoi@inha.ac.kr

Abstract

This paper introduces and studies *zero-base generalized few-shot learning* (*zero-base GFSL*), which is an extreme yet practical version of few-shot learning problem. Motivated by the cases where base data is not available due to privacy or ethical issues, the goal of zero-base GFSL is to newly incorporate the knowledge of few samples of novel classes into a pretrained model without any samples of base classes. According to our analysis, we discover the fact that both mean and variance of the weight distribution of novel classes are not properly established, compared to those of base classes. The existing GFSL methods attempt to make the weight norms balanced, which we find helps only the variance part, but discard the importance of mean of weights particularly for novel classes, leading to the limited performance in the GFSL problem even with base data. In this paper, we overcome this limitation by proposing a simple yet effective normalization method that can effectively control both mean and variance of the weight distribution of novel classes without using any base samples and thereby achieve a satisfactory performance on both novel and base classes. Our experimental results somewhat surprisingly show that the proposed zero-base GFSL method that does not utilize any base samples even outperforms the existing GFSL methods that make the best use of base data. Our implementation is available at: https://github.com/bigdata-inha/Zero-Base-GFSL.

Introduction

Few-shot learning (FSL) (Fei-Fei, Fergus, and Perona 2006; Lake et al. 2011; Lake, Salakhutdinov, and Tenenbaum 2015) has become a major problem due to the practical difficulties of data collection. Trying to mimic the human's ability to acquire general knowledge with a few observations, the goal of FSL is to learn new knowledge by training only a few samples on a model, which is often a pretrained model. Many existing works focus on a scenario where the resulting model only discriminates novel classes, referred to as *standard few-shot learning* (Finn, Abbeel, and Levine 2017; Koch et al. 2015; Snell, Swersky, and Zemel 2017; Vinyals et al. 2016). It is even more challenging, yet practical, when the model can infer both existing classes (a.k.a. base classes) already trained in the model and unseen classes (a.k.a. novel classes) newly learned with a few samples. Only some recent works (Gidaris and Komodakis 2018, 2019; Kukleva, Kuehne, and Schiele 2021; Shi et al. 2020) tackle this version of FSL, namely *generalized few-shot learning* (*GFSL*), which not only aims to learn about novel classes but also preserve the existing knowledge for base classes.



(a) Pretrained (b) Biased (c) Desired (ours)

Figure 1: Visualization of the decision boundaries of ResNet-18 trained with an auxiliary 2D projection module before the classifier layer on *tiered-ImageNet*, where red triangles are feature vectors of novel classes and the others are those of base classes. Note that feature vectors themselves remain the same in the frozen feature space. *(a)*: Well-formed decision boundaries of the pretrained model of base classes. *(b)*: Fine-tuned decision boundaries biased toward novel classes. *(c)*: Well-balanced decision boundaries for both base and novel classes via our normalization.

Due to its simplicity and performance, transfer learning becomes a dominant approach in GFSL, where we fine-tune a model already trained over base classes to additionally learn the knowledge of new classes with a few samples. Despite the small volume of new data, the fine-tuned model gets easily biased to novel classes to the point that the model becomes pretty useless for base classes. Hence, most existing works in GFSL focus on how to make the resulting model well balanced over base and novel classes by performing balanced fine-tuning (Kukleva, Kuehne, and Schiele 2021; Li et al. 2019; Qi, Brown, and Lowe 2018) or leveraging additional architectures (Gidaris and Komodakis 2018; Ren et al. 2019; Yoon et al. 2020) and supplement information (Li et al. 2020; Shi et al. 2020).

All the aforementioned approaches assume the presence of *base data* (i.e., samples of base classes) for preserving the knowledge of base classes as well as learning the relationship

*Corresponding Author

between novel and base classes. In reality, however, base data may not be always available due to some privacy or ethical issues. For instance, *Google* releases the highly generic model *BiT* (Kolesnikov et al. 2020) of 18,000 classes, but its training data *JFT* (Sun et al. 2017) consists of millions of private images that should not be exposed to public. Furthermore, retraining the base data has never been a perfect solution for GFSL either. In spite of retraining overhead, the overall performance can be highly dependent on which base samples are selected. To be shown by our experimental results, the state-of-the-art GFSL methods using base data turn out to be even less accurate than our GFSL method without employing any base samples.

Beyond the limitation of the existing GFSL approaches, this paper focuses on an even more challenging scenario of few-shot learning, called *zero-base GFSL*, where we are free to use a pretrained model somehow learned over base classes but cannot retrain any of base samples during GFSL. Obviously, fine-tuning a pretrained model with only novel samples leads to the model highly biased toward novel classes even if we freeze the feature extractor of the model. Thus, even in the frozen feature space built on base classes, novel samples are capable of forming undesirably large decision boundaries, which is hard to be addressed without jointly training all of the base and novel samples. Figure 1(b) shows such an example where some base samples, which used to be well discriminated in their pretrained model (Figure 1(a)), can be misclassified as a novel class with a large decision boundary. This leads to the model pretty inaccurate for base classes without normalization as shown in Figure 2.

Then, how could only a few novel samples make the decision boundaries highly biased toward novel classes in the feature space? To answer this question, we investigate the distribution of weights of classifiers that actually determine decision boundaries, and discover the following undesirable facts in terms of its mean and variance. First, there is an imbalance between the variances of weight distributions of base and novel classes, which we also find is related to the norm imbalance tackled by many existing GFSL methods (Fan et al. 2021; Gidaris and Komodakis 2018; Qi, Brown, and Lowe 2018; Wang et al. 2020). However, these methods do not consider how the means of weight distributions are different between base and novel classes. This paper newly observes that the average weight of the novel classifier is positively shifted from that of the base classifier, which we call *mean shifting phenomenon*. Although this phenomenon is indeed a more crucial reason behind the biased model, existing works rarely try to fix the shifted mean of novel classes.

To remedy these undesirable mean and variance of novel classes, this paper designs a new normalization method that can achieve a centered mean as well as a balanced variance without retraining any base samples. As experimentally shown in Figure 1(c), our normalization method enables the classifier to form well-balanced decision boundaries for both base and novel classes, and thereby it is observed in Figure 2 that the performance degradation of base classes is prevented while keeping the performance of the novel classes as high as possible. With only 5-shot of novel classes, we can keep the performance of both base and novel classes

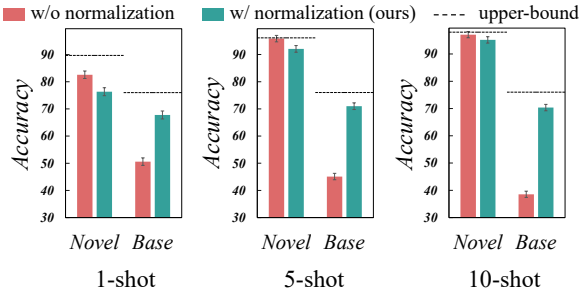


Figure 2: Comparison on the accuracy of zero-base GFSL with ResNet-50 on *ImageNet-800*. The upper bound is the conditional test accuracy given either novel or base classes.

close to their upper bounds that are conditional accuracies given only base and novel classes, respectively. In our experiments, without using any base data, our method even beats the existing state-of-the-art GFSL methods with clear margins, like 4.59% and 3.53% 5-shot accuracy in *mini-* and *tiered-ImageNet*, respectively.

Related Works

Few-shot learning. Few-shot learning (FSL) has been studied mostly for the standard scenario aiming to learn novel classes without having to preserve the existing knowledge. The recent methods of standard FSL can be divided into the following two categories: *meta learning* and *transfer learning*. Meta learning methods commonly train a meta-learner in an episodic way, where the base data is divided into multiples tasks each with a few samples and thereby class-agnostic knowledge is acquired from previous experiences (Finn, Abbeel, and Levine 2017; Jamal and Qi 2019; Lee et al. 2019; Sun et al. 2019; Snell, Swersky, and Zemel 2017; Vinyals et al. 2016; Xu et al. 2021; Zhang et al. 2020).

Unlike meta learning, which requires a complex episodic training, transfer learning (Chen et al. 2019; Dhillon et al. 2020; Tian et al. 2020; Wang et al. 2019) simply re-uses the feature extractor learned from all collected base samples when fine-tuning only the classifier of the model being trained for FSL. These transfer learning approaches often show a better performance than meta learning particularly in deeper neural networks (Dhillon et al. 2020), and therefore are getting more attention from the state-of-the-art methods. Although transfer learning seems to be quite effective in standard FSL, its performance is still unsatisfactory in generalized few-shot learning (GFSL) without leveraging additional techniques. However, this paper claims that our proposed normalization method can enable simple transfer learning to be highly effective in GFSL even without using any base samples.

Generalized few-shot learning. GFSL aims to handle both base and novel classes in a joint space, and hence is regarded to be more difficult than standard FSL. In GFSL, we not only need to learn as much new knowledge as possible but also preserve the pretrained knowledge over base classes. The major approach to this end in GFSL is not to entirely change a pretrained model as we do in transfer learning, but

to train some extra architectures, which inform the original output of the pretrained model to be properly modified for a balanced inference between base and novel classes (Gidaris and Komodakis 2018; Ren et al. 2019; Yoon et al. 2020) occasionally by leveraging supplement information (Li et al. 2020; Shi et al. 2020).

The other approach without introducing any extra architecture is to fine-tune a pretrained model with a balanced dataset of base and novel classes, namely *balanced fine-tuning*. In order to improve the performance, balanced fine-tuning is often performed together with additional techniques like *weighting imprinting* of classifiers (Qi, Brown, and Lowe 2018) and introducing a three-step framework (Kukleva, Kuehne, and Schiele 2021). For a better balanced dataset, *hallucination* approaches (Hariharan and Girshick 2017; Wang et al. 2018) synthesize novel instances based on the base dataset.

Incremental few-shot learning. In continual learning, GFSL is being extended to *incremental few-shot learning* (*IFSL*) (Kukleva, Kuehne, and Schiele 2021; Mazumder, Singh, and Rai 2021; Tao et al. 2020; Zhang et al. 2021), in which the model has to go through a series of tasks of few-shot classes. Similar to the existing GFSL methods, most IFSL methods take the best use of the previously collected dataset to preserve the previous knowledge and learn the relationship among different classes.

All the existing works in GFSL and IFSL somehow need base data for either training extra architectures, balanced fine-tuning, or preserving the previous knowledge. To our best knowledge, this paper is the first study that resolves *zero-base GFSL* via weight normalization without using base data. Although the existing GFSL methods (Fan et al. 2021; Gidaris and Komodakis 2018; Qi, Brown, and Lowe 2018; Wang et al. 2020) also perform basic weight normalization using a cosine classifier, their normalization scheme alone fails to achieve a satisfactory performance without their additional training techniques relying on base data. In this paper, we investigate what occurs to weight distribution during GFSL, and thereby effectively overcome the limitation of the existing basic normalization method.

Preliminary

Basic Framework of GFSL

In generalized few-shot classification, the training dataset is composed of a base dataset, denoted by $\mathcal{D}_{base} = \{(\mathbf{x_i}, \mathbf{y_i})\}_{i=1}^{K_{base}}$, and a novel dataset, denoted by $\mathcal{D}_{novel} = \{(\mathbf{x_j}, \mathbf{y_j})\}_{j=1}^{K_{novel}}$, where $(\mathbf{x}, \mathbf{y})$ is a training instance and $K_{base} \gg K_{novel}$ is usually assumed. We denote C_{base} and C_{novel} be two disjoint sets of base classes and novel classes, respectively, where $|C_{base}| \gg |C_{novel}|$.

Then, the goal of GFSL is to train a model $\Phi = \{\phi, \theta\}$ with $\mathcal{D}_{base} \cup \mathcal{D}_{novel}$ such that Φ can discriminate any classes in $C_{base} \cup C_{novel}$, which is different from standard FSL aiming to classify only C_{novel}. ϕ and θ are parameters of the feature extractor and the *linear* classifier of the model, and we also denote $f_\phi(\mathbf{x})$ be the feature vector returned from ϕ given $\mathbf{x}$, where the dimensionality of each feature vector is denoted as d. We then make inference by taking the softmax of $\theta^\top f_\phi(\mathbf{x})$ for a given input sample $\mathbf{x}$.

The basic framework of GFSL consists of two stages, namely pre-training and fine-tuning. This two-stage transfer learning scheme, which fully utilizes the knowledge of base classes to learn few-shot novel classes, is regarded as one of the leading paradigms for GFSL as well as standard FSL due to its simplicity and effectiveness.

Pre-training. In the pre-training stage, we train a full model $\Phi = \{\phi, \theta_{base}\}$ only with $\mathcal{D}_{base}$ by:

$$\Phi = \arg\min_{\{\phi, \theta_{base}\}} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{base}} -\mathbf{y} \log p(\mathbf{x}) + \mathcal{R}(\phi, \theta_{base}), \quad (1)$$

where $\theta_{base} \in \mathbb{R}^{d \times |C_{base}|}$ is the base classifier, $p(\mathbf{x})$ is the softmax output probability, i.e., $p_i(\mathbf{x}) = \frac{exp(\mathbf{z}_i)}{\sum_{j=1}^{|C_{base}|} exp(\mathbf{z}_j)}$ such that $\mathbf{z} = \theta^\top f_\phi(\mathbf{x})$, and $\mathcal{R}$ is the regularization term for the full model. This stage is equivalent to typical supervised learning as $\mathcal{D}_{base}$ has a sufficient number of samples.

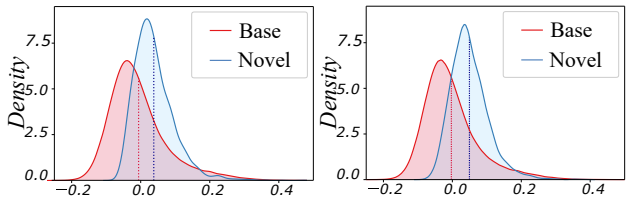
Fine-tuning. Once a model is well-trained over base classes, we then fine-tune the model with respect to $C_{base} \cup C_{novel}$. In order to incorporate novel classes, the model should be extended to $\Phi = \{\phi, \theta\}$ in this fine-tuning stage, where $\theta = \{\theta_{base}, \theta_{novel}\}$, θ_{base} is the trained base classifier and θ_{novel} is the novel classifier randomly initialized. Also, in order to make all the classes well balanced in the fine-tuned model, existing GFSL methods (Kukleva, Kuehne, and Schiele 2021; Qi, Brown, and Lowe 2018) train a balanced dataset, which is either a subset or superset of $\mathcal{D}_{base} \cup \mathcal{D}_{novel}$, constructed by undersampling $\mathcal{D}_{base}$ or oversampling $\mathcal{D}_{novel}$. Given such a balanced dataset, most state-of-the-art methods (Ren et al. 2019; Yoon et al. 2020) fine-tune only the classifier while freezing the feature extractor. This is due to the fact that the feature extractor is already well-trained on a sufficient number of samples and hence better to be frozen not to hurt generalization of the model.

As long as a balanced dataset is ideally constructed, the model can well be trained for both base and novel classes by this basic framework as fine-tuning can properly rearrange all decision boundaries in the joint feature space. Unfortunately, vanilla sampling strategies (Huang et al. 2016; Wang, Ramanan, and Hebert 2017) would not work well in GFSL as undersampling inevitably causes information loss and oversampling inherently suffers from the lack of diversity due to many redundant samples. This is why many existing works propose extra techniques to achieve a better performance, which is not quite satisfactory according to our experiments. More importantly, a balanced dataset can be collected and fine-tuned only if the base dataset is available, which may not always be the case.

Problem Statement of Zero-Base GFSL

In the proposed zero-base GFSL problem, our objective is to train a joint linear classifier θ, which works well for both C_{base} and C_{novel}, with only a few samples of novel classes (i.e., $\mathcal{D}_{novel}$) without using any samples of $\mathcal{D}_{base}$. Following the basic GFSL framework, we only fine-tune the classifier while freezing the feature extractor as:

$$\theta = \arg\min_{\theta} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{novel}} -\mathbf{y} \log p(\mathbf{x}) + \mathcal{R}(\theta). \quad (2)$$



(a) 1-shot (b) 5-shot

Figure 3: Weight distributions of base and novel classifiers where dotted lines represent means, using *tiered-ImageNet* with ResNet-18.

Note that it is even more desired to freeze the feature extractor in zero-base GFSL than in GFSL because the feature extractor can severely be biased toward novel classes after fine-tuning with only novel samples.

Analysis and Methodology on Zero-Base GFSL

This section first investigates why the basic framework of GFSL cannot perform well in zero-base GFSL, and then presents an effective weight normalization method that enables a simple transfer learning scheme to achieve even a better performance than existing GFSL methods assuming the existence of base data.

Hardness of Zero-Base GFSL

As presented in Figure 2, a model trained by the basic framework of GFSL, that is, fine-tuning only the joint classifier θ with $\mathcal{D}_{novel}$, turns out to be extremely inaccurate for base classes yet pretty accurate for novel classes despite freezing the feature extractor. Thus, the fine-tuned classifier gets highly biased toward novel classes due to the absence of base samples.

More specifically, when freezing the feature extractor ϕ, the feature space itself is fixed and so are the feature vectors $f_\phi(\mathbf{x})$ of all the samples $\mathbf{x}$. Consequently, it is decision boundaries in the space that are changed by fine-tuning the classifier and therefore matter to a biased prediction. Without a balanced dataset, it is challenging to properly learn hidden relationships between base and novel classes, leading to *enlarged* (i.e., biased) decision boundaries of novel classes as depicted in Figure 1. In zero-base GFSL, it is our mission to find the best balance between decision boundaries of base and novel classes without retraining any base samples so that we can ideally make the best joint prediction. As decision boundaries between base and novel classes are formed where $\theta_{base}^\top f_\phi(\mathbf{x}) = \theta_{novel}^\top f_\phi(\mathbf{x})$, they are mainly affected by how the weight values of θ_{base} and θ_{novel} are distributed. In the following subsection, we therefore conduct a systematic analysis on the weight distributions of θ_{base} and θ_{novel}.

Analysis on Weight Distributions of Classifiers

We investigate the underlying distributions of weights of base and novel classifiers particularly in terms of their mean μ and variance σ^2. More specifically, given a classifier $\theta = [\theta_1, \theta_2, \ldots, \theta_{|C|}] \in \mathbb{R}^{d\times|C|}$, where $\theta_i \in \mathbb{R}^d$ represents the weight vector corresponding to class i, we first define the vectors of class-wise means and standard deviations of weight values as follows:

$$\mu = [\mu_1, \mu_2, \ldots, \mu_{|C|}] \text{ and } \sigma = [\sigma_1, \sigma_2, \ldots, \sigma_{|C|}],$$

where $\mu_i = \frac{1}{d}\sum_{j=1}^{d}\theta_{i,j}$ and $\sigma_i^2 = \frac{1}{d}\sum_{j=1}^{d}(\theta_{i,j}-\mu_i)^2$. Also, we use μ_{base}, σ_{base}, μ_{novel}, and σ_{novel} to denote the vectors corresponding to the weights of base and novel classifiers (i.e., θ_{base} and θ_{novel}).

In order to examine how the weight values of the novel classifier are different from those of the base classifier, we compute the average of class-wise means and standard deviations as $\bar{\mu} = \frac{1}{|C|}\sum_{i=1}^{|C|}\mu_i$ and $\bar{\sigma} = \frac{1}{|C|}\sum_{i=1}^{|C|}\sigma_i$ and compare $\bar{\mu}_{novel}$ and $\bar{\sigma}_{novel}$ with $\bar{\mu}_{base}$ and $\bar{\sigma}_{base}$. As observed in Figure 3, we discover the following two undesirable facts, both of which should be tackled to obtain a balanced joint classifier.

L2-norms proportional to standard deviations. We first discover that the variance of θ_{base} is greater than that of θ_{novel}, i.e., $\bar{\sigma}_{base} > \bar{\sigma}_{novel}$, as shown in Figure 3. This is somewhat intuitive in that the base classifier is trained on a large number of samples in $\mathcal{D}_{base}$ that can increase the diversity, whereas the novel classifier cannot have such a large diversity due to the lack of training samples in $\mathcal{D}_{novel}$. What is not quite obvious is that these standard deviations are indeed proportional to the L2-norms of weights in neural networks as the following proposition.

Proposition 1. *Consider a parameter θ of N weights in a neural network with the assumption that θ is randomly initialized by a Gaussian distribution with zero mean and trained by a uniformly distributed dataset. Then, it holds that $\sigma = \|\theta\|_2 \cdot \frac{1}{\sqrt{N}}$.*

Proof. Given $E[\theta] = 0$, we have: $\sigma = \sqrt{E[\theta - E(\theta)]^2} = \sqrt{E[\theta^2]} \approx \sqrt{\frac{1}{N}\sum_{i=1}^{N}\theta_i^2} = \frac{1}{\sqrt{N}}\|\theta\|_2$. □

By Proposition 1, $\|\theta_{base}\|_2 > \|\theta_{novel}\|_2$ is implied from the observation of $\bar{\sigma}_{base} > \bar{\sigma}_{novel}$. Somewhat surprisingly, this contradicts to the highly biased results toward novel classes in Figure 2 in the sense that many existing works (Hou et al. 2019; Kang et al. 2020; Zhao et al. 2020) report that larger weight norms of a class tend to produce larger logits leading to more biased predictions toward the corresponding class. Thus, if we follow the existing strategy of equalizing the weight norms (Hou et al. 2019; Kang et al. 2020; Zhao et al. 2020) by increasing the smaller ones yet decreasing the larger ones, the novel-biased classifier in zero-base GFSL would get even more biased toward novel classes, to be confirmed in our experiments.

Then, how could the classifier be highly biased toward novel classes even though the weight norms of novel classes are less than those of base classes? The true answer to this question lies in the average weight of each classifier rather than its weight norm.

Mean shifting phenomenon. In terms of means of weight distributions, we observe the fact that the average weight

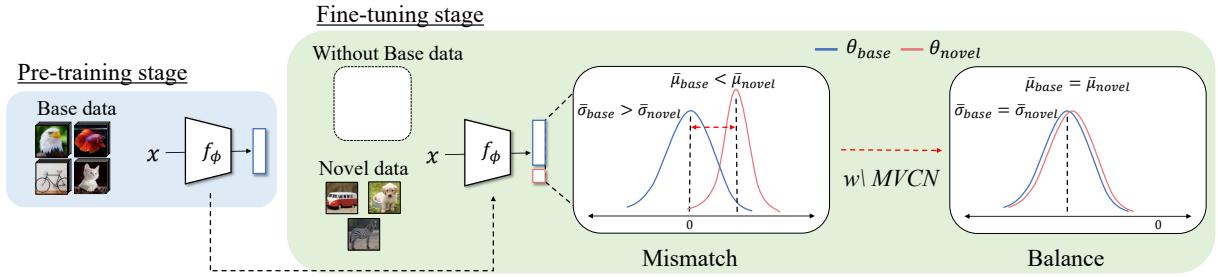


Figure 4: Overview of Mean-Variance Classifier Normalization. In the pre-training stage, we train the entire model on all base classes. In the fine-tuning stage, we normalize the novel classifier by online mean centering in the process of training and adjust the trained weights of the base and novel classifiers by re-scaling them using the standard deviation ratio.

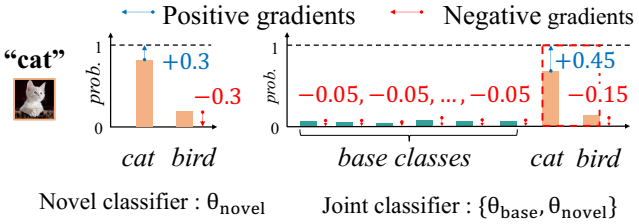


Figure 5: An illustrative example for the positively shifted mean of weights for novel classes, where a cat image gets probabilities $[0.7\ 0.3]$ from θ_{novel} but its probabilities becomes $[0.05\ 0.05\ 0.04\ 0.06\ 0.05\ 0.05 \mid 0.55\ 0.15]$ in $\theta = \{\theta_{base}, \theta_{novel}\}$.

of the novel classifier is positively shifted (i.e., $\bar{\mu}_{novel} > 0$) while that of the base classifier keeps almost zero (i.e., $\bar{\mu}_{base} \approx 0$). We call this situation *mean shifting phenomenon* commonly observed in all the graphs of Figure 3.

Positively shifted $\bar{\mu}_{novel}$ indicates that each weight of θ_{novel} on the average has a more positive value. Recalling that decision boundaries are constructed at $\theta_{base}^{\top} f_\phi(\mathbf{x}) = \theta_{novel}^{\top} f_\phi(\mathbf{x})$, more positive weights of θ_{novel} would increase $\theta_{novel}^{\top} f_\phi(\mathbf{x})$, and therefore enlarge the decision boundaries of novel classes. This implies that mean shifting phenomenon is a more critical reason behind the novel-biased model in zero-base GFSL. As mentioned above, however, the existing works (Hou et al. 2019; Kang et al. 2020; Zhao et al. 2020) only focus on equalizing the weight norms, which turn out to be proportional to the standard deviations of weights, between base and novel classes, and hence their normalization method alone fails to achieve a satisfactory performance.

What makes mean shifted. Let us now analyze why mean shifting phenomenon occurs to the novel classifier in zero-base GFSL. To this end, we first consider how we update the weights of each classifier by gradient descent as $\theta = \theta - \eta \frac{\partial \mathcal{L}}{\partial \theta}(f_\phi(\mathbf{x}), \theta)$, where the gradient is:

$$\frac{\partial}{\partial \theta}\mathcal{L}(f_\phi(\mathbf{x}), \theta) = \frac{\partial \mathcal{L}}{\partial z} \cdot \frac{\partial z}{\partial \theta} = f_\phi(\mathbf{x})(\mathbf{y} - p(\mathbf{x})). \quad (3)$$

Since $f_\phi(\mathbf{x})$ is always non-negative due to the ReLU function and $p(\mathbf{x})$ is also a non-negative probability value, the gradient is positive only if $\mathbf{y}_i = 1$ (i.e., for the probability of the class label of $\mathbf{x}$) and negative for all the other class probabilities.

In training a balanced dataset of all the classes uniformly distributed, the total positive gradient for each class eventually gets similar to the absolute value of its total negative gradient, which is why $\bar{\mu} \approx 0$. Thus, if we train $\mathcal{D}_{novel}$ only to θ_{novel}, μ_{novel} would also be close to zero.

However, when fine-tuning the joint classifier $\theta = \{\theta_{base}, \theta_{novel}\}$ with only $\mathcal{D}_{novel}$, the total positive gradient of θ_{novel} increases whereas the absolute value of its total negative gradient decreases because all the output probabilities of θ_{novel} together become smaller in training θ than they used to be in training only θ_{novel}. To illustrate, consider an example of Figure 5, where output probabilities of θ_{novel} are larger than those of $\theta = \{\theta_{base}, \theta_{novel}\}$ for a given training image of *'cat'*. Thus, even if θ_{novel} outputs $[0.7\ \ 0.3]$ for the cat image, the joint classifier θ would output something like $[0.05\ \ 0.05\ \ 0.04\ \ 0.06\ \ 0.05\ \ 0.05 \mid 0.55\ \ 0.15]$ for 6 base classes followed by 2 novel classes. Consequently, the positive gradient caused by the novel class *'cat'* is increased from 0.3 to 0.45, but the absolute value of the negative gradient for the second novel class *'bird'* is decreased from $|-0.3|$ to $|-0.15|$. Furthermore, θ_{novel} can receive negative gradients only from the instances of the other novel classes due to the absence of base instances. This makes μ_{novel} to be more positively shifted.

As for θ_{base}, only negative gradients will arrive during the entire fine-tuning process, but its total amount would not be that much because the well-trained θ_{base} is not likely to be overconfident to irrelevant novel classes particularly when θ_{novel} gets used to identifying their corresponding novel classes. This is why μ_{base} still keeps to be zero after fine-tuning.

Solution: Mean-Variance Classifier Normalization

To fulfill both the zero-mean and a balanced variance of base and novel classifiers, we propose a simple yet effective normalization method, called *Mean-Variance Classifier Normalization* (*MVCN*). MVCN takes two essential steps, namely *online mean centering* and *offline variance balancing*. We perform online mean centering during fine-tuning the joint classifier, and then proceed to balance variances of weights as a post processing step once the classifier is well trained. The entire process is outlined in Figure 4.

Online mean centering. First, we keep normalizing the weights of the novel classifier to have zero-mean in the pro-

Methods/ Shots	1-shot			5-shot			10-shot		
	Novel	**Base**	**All**	**Novel**	**Base**	**All**	**Novel**	**Base**	**All**
GcGPN (Shi et al. 2020)	39.86	54.65	47.25	56.32	59.30	57.81	-	-	-
IW (Qi, Brown, and Lowe 2018)	41.32	58.04	49.68	59.27	58.68	58.98	45.85	72.53	59.19
DFSL (Gidaris and Komodakis 2018)	31.25	17.72	39.49	46.96	58.92	52.94	66.04	69.87	67.95
AAN (Ren et al. 2019)	45.61	63.92	54.76	60.82	64.14	62.48	66.33	62.49	64.41
LCwoF (Kukleva, Kuehne, and Schiele 2021)	53.78	62.89	57.84	68.58	64.53	66.55	76.71	62.86	69.78
XtarNet (Yoon et al. 2020)	47.04	64.17	55.61	62.46	**70.57**	66.52	70.46	**70.88**	70.67
MVCN (ours)	**51.72**	**65.81**	**58.77**	**75.20**	67.62	**71.22**	**78.70**	68.06	**73.38**

Methods/ Shots	1-shot			5-shot			10-shot		
	Novel	**Base**	**All**	**Novel**	**Base**	**All**	**Novel**	**Base**	**All**
IW (Qi, Brown, and Lowe 2018)	44.95	62.53	53.74	71.85	56.11	63.98	74.01	61.67	65.50
DFSL (Gidaris and Komodakis 2018)	47.32	36.10	41.71	67.94	39.08	53.51	73.97	56.94	65.46
AAN (Ren et al. 2019)	54.39	55.85	55.12	57.76	64.13	60.95	70.88	57.49	64.18
LCwoF (Kukleva, Kuehne, and Schiele 2021)	57.13	60.39	58.76	69.05	63.44	66.25	79.20	61.76	70.57
XtarNet (Yoon et al. 2020)	58.90	**64.02**	61.46	74.49	63.13	68.81	78.36	63.08	70.72
MVCN (ours)	**62.11**	61.23	**61.67**	**79.59**	**66.3**	**72.34**	**83.06**	**67.46**	**75.26**

Table 1: Performance (%) comparison with GFSL methods on *mini-ImageNet* (top) and *tiered-ImageNet* (bottom), where each accuracy value is the average over 600 tasks, and 95% confidence intervals are also shown in Appendix.

cess of training by:

$$\hat{\theta}_{novel} = \theta_{novel} - \mu_{novel}. \tag{4}$$

When fine-tuning the joint classifier, online mean centering is performed only on the novel classifier. This directly reduces the positive weights of the novel classifier and keeps the zero-mean while learning the features of novel classes. Note that we do not give any constraints to the variance of weights during fine-tuning, but rather allow each classifier to learn as many required features as possible.

Offline variance balancing. Once the fine-tuning stage is done, we adjust the weights of base classifier according to the ratio of standard deviations as follows:

$$\hat{\theta}_{base} = \frac{\bar{\sigma}_{novel}}{\sigma_{base}} \cdot \theta_{base}, \tag{5}$$

where $\bar{\sigma}_{novel}$ is the average of class-wise standard deviations for all novel classes. We re-scale each weight of the base classifier by multiplying the ratio of the standard deviation of the novel classifier to that of the base classifier (i.e., $\frac{\bar{\sigma}_{novel}}{\sigma_{base}}$), and thereby the weight variance of the base classifier gets similar to that of the novel classifier. Note that we do not directly normalize θ by their σ, which makes the zero-mean and unit-variance $\frac{\theta-\mu}{\sigma}$, because the standard deviation of weights is much smaller than 1.

Post linear optimization without base data. Finally, we introduce class-wise learnable parameters γ_i and β_i for $i \in C_{base} \cup C_{novel}$, which are intended for further optimizing decision boundaries of novel classes. With freezing $\theta = \{\theta_{base}, \theta_{novel}\}$, γ_i and β_i are similarly trained by Eq. (2) and used at inference time as $\gamma \cdot \theta^\top f_\phi(\mathbf{x}) + \beta$. Through these additional parameters, we can further improve the performance of novel classes particularly in an extreme case of 1-shot learning. Note that this optimization process is performed still without using any base samples.

Experiments

Experimental Settings

Datasets. We compare our method with the state-of-the-art (SOTA) GFSL methods using two datasets, *mini-ImageNet* (Vinyals et al. 2016) and *tiered-ImageNet* (Ren et al. 2018), which are most widely used in the literature of GFSL. The *mini-ImageNet* contains 100 classes and 60,000 sample images from *ImageNet* (Russakovsky et al. 2015), which are then randomly split into 64 training classes, 16 validation classes, and 20 testing classes, proposed by (Ravi and Larochelle 2017). The *tiered-ImageNet* is another subset of *ImageNet*, containing 608 classes, which are then split into 351 training, 97 validation, 160 testing classes. This setting is more challenging since base classes and novel classes come from different super classes. The size of all images in both datasets is 84×84. In addition, we test whether our method works better than the SOTA *GZSL* (*Generalized Zero-Shot Learning*) methods using the three most common datasets, *CUB* , *AWA1* (Lampert, Nickisch, and Harmeling 2009), and *AWA2* (Xian et al. 2019).

Implementation details. We implement all the methods in PyTorch, and train each model on a machine with NVIDIA A100. We use ResNet-12 (He et al. 2016) for *mini-ImageNet* and ResNet-18 for *tiered-ImageNet*, according to (Ren et al. 2019; Yoon et al. 2020). For *CUB*, *AWA1* and *AWA2*, we commonly use ResNet-101. Full details of our settings are covered in Appendix.

Experimental Results

Overall performance. Table 1 summarizes the overall performance of the compared GFSL methods on *ImageNet*. Even without base data, it is clearly observed that our MVCN method outperforms the other GFSL methods exploiting base

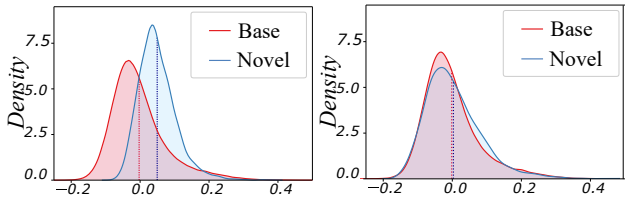
Datasets	CUB				AWA1				AWA2			
Methods/ Shots	**1-shot**	**2-shot**	**5-shot**	**10-shot**	**1-shot**	**2-shot**	**5-shot**	**10-shot**	**1-shot**	**2-shot**	**5-shot**	**10-shot**
ReViSE	36.3	41.1	44.6	50.9	56.1	60.3	64.1	67.8	-	-	-	-
CA-VAE	50.6	54.4	59.6	62.2	64.0	71.3	76.6	79.0	41.8	52.7	66.5	76.7
DA-VAE	49.2	54.6	58.8	60.8	68.0	73.0	75.6	76.8	68.6	77.1	81.8	81.3
CADA-VAE	55.2	59.2	63.0	64.9	69.6	73.7	78.1	80.2	73.6	78.9	81.9	85.0
DRAGON	55.3	59.2	63.5	**67.8**	67.1	69.1	76.7	81.9	-	-	-	-
MVCN (ours)	**57.3**	**61.6**	**65.4**	**67.8**	**69.9**	**76.4**	**81.2**	**82.2**	**77.1**	**83.5**	**87.4**	**87.7**

Table 2: Performance (%) comparison with GZSL methods, ReViSE (Tsai, Huang, and Salakhutdinov 2017), CA-VAE (Schönfeld et al. 2019), DA-VAE (Schönfeld et al. 2019), CADA-VAE (Schönfeld et al. 2019), and DRAGON (Samuel, Atzmon, and Chechik 2021), on *CUB, AWA1*, and *AWA2*.

samples. In *tiered-ImageNet*, which is a more challenging scenario where base and novel classes are quite different, MVCN seems to be even more effective when it is in *mini-ImageNet*. This is probably because a larger number of base classes can be pretrained in *tiered-ImageNet* than in *mini-ImageNet*, and the simple transfer learning scheme of MVCN takes more advantage of this well-trained knowledge of many base classes. Especially when we have more novel samples like 5 or 10-shot, the performance gap between ours and the other methods becomes larger as shown in Table 1. The underlying reason is mean shifting phenomenon gets stronger as we fine-tune more novel classes (see Figure 3). Although XtarNet (Yoon et al. 2020) occasionally performs slightly better on base classes than ours, MVCN still manages to outperform XtraNet on novel classes with clear margins. Furthermore, Table 2 shows that MVCN is superior to all the SOTA GZSL methods on *CUB*, *AWA1*, and *AWA2*. Note that each image in these datasets is augmented with textual attributes that are crucially utilized by most GZSL methods while our method does not exploit any extra information.

Qualitative results. As shown in Figure 6, we can observe that MVCN actually resolves *mean shifting phenomenon* and achieves a balanced variance between base and novel classes as the classifier with MVCN has a $\bar{\mu}_{novel}$ 62 times smaller than it is without normalization. Also, we show that $\bar{\sigma}_{base}$ and $\bar{\sigma}_{novel}$ become almost the same after normalization, implying that there is no bias toward either novel or base classes. The confusion matrices of with and without normalization are presented in our supplement material, where we again confirm that our method is effective for the classifier not to be confused about base classes with novel classes.

Ablation study. To examine the effect of each component of our method, which are online mean centering (MC), offline variance balancing (VB), and post linear optimization (LO), we conduct an ablation study in zero-base GFSL using *mini-ImageNet*. Table 3 shows that applying mean centering solely improves the accuracy of base classes by more than 25%, which shows its effectiveness of mitigating the novel-bias problem. In addition to MC, variance balancing tends to make the model more accurate for novel classes yet a bit less accurate for base classes. Considering the results of Table 1, note that MC together with VB beats the SOTA GFSL method XtarNet by large margins in 5-shot case. Through all the components, we get 12.01% performance gain for 5-shot.



(a) w/o normalization (b) w/ normalization (ours)

Figure 6: Weight distributions of base and novel classifiers of ResNet-18 on *tiered-ImageNet* for 5-shot.

MC	VB	LO	Novel	Base	All
✗	✗	✗	79.97	38.44	59.21
✓	✗	✗	73.13	**68.91**	71.02
✓	✓	✗	73.49	68.71	71.10
✓	✓	✓	**75.20**	67.62	**71.22**

Table 3: Ablation study for 5-shot on *mini-ImageNet* to analyze the effectiveness of three components of our method, which are mean centering (MC), variance balancing (VB), and linear optimization (LO).

Conclusion

In this paper, we conducted the first study on zero-base GFSL, where we need to fine-tune a joint classifier with only a few samples of novel classes. Through a systematic analysis, we discovered that mean shifting phenomenon was the critical reason behind a novel-biased classifier, but the existing GFSL methods have been trying to equalize only the variance of weights. Based on our findings, we proposed a simple yet effective weight normalization method without using any base samples, which can even beat the existing GFSL methods that utilize the base dataset. Even though this work dealt with only linear classifiers, which is our limitation, our next plan is to extend our analysis to cover various types of classifiers.

Acknowledgments

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government(MSIT) (No.2022-0-00448, Deep Total Recall: Continual Learning for Human-Like Recall of Artificial Neural Networks, No.RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)), in part by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No.2021R1F1A1060160, No.2022R1A4A3029480), and in part by INHA UNIVERSITY Research Grant.

References

Chen, W.; Liu, Y.; Kira, Z.; Wang, Y. F.; and Huang, J. 2019. A Closer Look at Few-shot Classification. In *International Conference on Learning Representations, ICLR 2019*.

Dhillon, G. S.; Chaudhari, P.; Ravichandran, A.; and Soatto, S. 2020. A Baseline for Few-Shot Image Classification. In *International Conference on Learning Representations, ICLR 2020*.

Fan, Z.; Ma, Y.; Li, Z.; and Sun, J. 2021. Generalized Few-Shot Object Detection Without Forgetting. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, 4527–4536.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-Shot Learning of Object Categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4): 594–611.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, 1126–1135.

Gidaris, S.; and Komodakis, N. 2018. Dynamic Few-Shot Visual Learning Without Forgetting. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 4367–4375.

Gidaris, S.; and Komodakis, N. 2019. Generating Classification Weights With GNN Denoising Autoencoders for Few-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 21–30.

Hariharan, B.; and Girshick, R. B. 2017. Low-Shot Visual Recognition by Shrinking and Hallucinating Features. In *IEEE International Conference on Computer Vision, ICCV 2017*, 3037–3046.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 770–778.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a Unified Classifier Incrementally via Rebalancing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 831–839.

Huang, C.; Li, Y.; Loy, C. C.; and Tang, X. 2016. Learning Deep Representation for Imbalanced Classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 5375–5384.

Jamal, M. A.; and Qi, G. 2019. Task Agnostic Meta-Learning for Few-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 11719–11727.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling Representation and Classifier for Long-Tailed Recognition. In *International Conference on Learning Representations, ICLR 2020*.

Koch, G.; Zemel, R.; Salakhutdinov, R.; et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 0.

Kolesnikov, A.; Beyer, L.; Zhai, X.; Puigcerver, J.; Yung, J.; Gelly, S.; and Houlsby, N. 2020. Big Transfer (BiT): General Visual Representation Learning. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *ECCV 2020 - 16th European Conference, Proceedings, Part V*, volume 12350, 491–507.

Kukleva, A.; Kuehne, H.; and Schiele, B. 2021. Generalized and Incremental Few-Shot Learning by Explicit Learning and Calibration without Forgetting. In *IEEE/CVF International Conference on Computer Vision, ICCV 2021*, 9000–9009.

Lake, B. M.; Salakhutdinov, R.; Gross, J.; and Tenenbaum, J. B. 2011. One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society, CogSci 2011*.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 951–958.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-Learning With Differentiable Convex Optimization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 10657–10665.

Li, A.; Huang, W.; Lan, X.; Feng, J.; Li, Z.; and Wang, L. 2020. Boosting Few-Shot Learning With Adaptive Margin Loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 12573–12581.

Li, A.; Luo, T.; Xiang, T.; Huang, W.; and Wang, L. 2019. Few-Shot Learning With Global Class Representations. In *IEEE/CVF International Conference on Computer Vision, ICCV 2019*, 9714–9723.

Mazumder, P.; Singh, P.; and Rai, P. 2021. Few-Shot Lifelong Learning. In *AAAI Conference on Artificial Intelligence, AAAI 2021*, 2337–2345.

Qi, H.; Brown, M.; and Lowe, D. G. 2018. Low-Shot Learning With Imprinted Weights. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 5822–5830.

Ravi, S.; and Larochelle, H. 2017. Optimization as a Model for Few-Shot Learning. In *International Conference on Learning Representations, ICLR 2017*.

Ren, M.; Liao, R.; Fetaya, E.; and Zemel, R. S. 2019. Incremental Few-Shot Learning with Attention Attractor Networks. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Annual*

Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 5276–5286.

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 *International Conference on Learning Representations, ICLR 2018*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.*, 115(3): 211–252.

Samuel, D.; Atzmon, Y.; and Chechik, G. 2021. From generalized zero-shot learning to long-tail with class descriptors. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, 286–295.

Schönfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 8247–8255.

Shi, X.; Salewski, L.; Schiegg, M.; and Welling, M. 2020. Relational Generalized Few-Shot Learning. In *British Machine Vision Conference 2020, BMVC 2020*.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Annual Conference on Neural Information Processing Systems 2017*, 4077–4087.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, 843–852.

Sun, Q.; Liu, Y.; Chua, T.; and Schiele, B. 2019. Meta-Transfer Learning for Few-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 403–412.

Tao, X.; Hong, X.; Chang, X.; Dong, S.; Wei, X.; and Gong, Y. 2020. Few-Shot Class-Incremental Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 12180–12189.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking Few-Shot Image Classification: A Good Embedding is All You Need? In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *ECCV 2020 - 16th European Conference, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, 266–282.

Tsai, Y. H.; Huang, L.; and Salakhutdinov, R. 2017. Learning Robust Visual-Semantic Embeddings. In *IEEE International Conference on Computer Vision, ICCV 2017*, 3591–3600.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Annual Conference on Neural Information Processing Systems 2016*, 3630–3638.

Wang, X.; Huang, T. E.; Gonzalez, J.; Darrell, T.; and Yu, F. 2020. Frustratingly Simple Few-Shot Object Detection. In *Proceedings of the International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, 9919–9928. PMLR.

Wang, Y.; Chao, W.-L.; Weinberger, K. Q.; and van der Maaten, L. 2019. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*.

Wang, Y.; Girshick, R. B.; Hebert, M.; and Hariharan, B. 2018. Low-Shot Learning From Imaginary Data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 7278–7286.

Wang, Y.; Ramanan, D.; and Hebert, M. 2017. Learning to Model the Tail. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Annual Conference on Neural Information Processing Systems 2017*, 7029–7039.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2019. Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9): 2251–2265.

Xu, W.; Xu, Y.; Wang, H.; and Tu, Z. 2021. Attentional Constellation Nets for Few-Shot Learning. In *International Conference on Learning Representations, ICLR 2021*.

Yoon, S. W.; Kim, D.; Seo, J.; and Moon, J. 2020. XtarNet: Learning to Extract Task-Adaptive Representation for Incremental Few-Shot Learning. In *Proceedings of the International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, 10852–10860.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 12200–12210.

Zhang, C.; Song, N.; Lin, G.; Zheng, Y.; Pan, P.; and Xu, Y. 2021. Few-Shot Incremental Learning With Continually Evolved Classifiers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, 12455–12464.

Zhao, B.; Xiao, X.; Gan, G.; Zhang, B.; and Xia, S. 2020. Maintaining Discrimination and Fairness in Class Incremental Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 13205–13214.