

Semantic-aware Knowledge Distillation for Few-Shot Class-Incremental Learning

Ali Cheraghian^{*1,2}, Shafin Rahman^{*3}, Pengfei Fang^{1,2}, Soumava Kumar Roy^{1,2}

Lars Petersson^{1,2}, Mehrtash Harandi^{2,4}

¹Australian National University, ²Data61-CSIRO, Australia

³ North South University, Bangladesh, ⁴Monash University, Australia

{Ali.Cheraghian, Pengfei.Fang, Soumava.KumarRoy}@anu.edu.au,

shafin.rahman@northsouth.edu, Lars.Petersson@data61.csiro.au, mehrtash.harandi@monash.edu

Abstract

Few-shot class incremental learning (FSCIL) portrays the problem of learning new concepts gradually, where only a few examples per concept are available to the learner. Due to the limited number of examples for training, the techniques developed for standard incremental learning cannot be applied verbatim to FSCIL. In this work, we introduce a distillation algorithm to address the problem of FSCIL and propose to make use of semantic information during training. To this end, we make use of word embeddings as semantic information which is cheap to obtain and which facilitate the distillation process. Furthermore, we propose a method based on an attention mechanism on multiple parallel embeddings of visual data to align visual and semantic vectors, which reduces issues related to catastrophic forgetting. Via experiments on MiniImageNet, CUB200, and CIFAR100 dataset, we establish new state-of-the-art results by outperforming existing approaches.

1. Introduction

In a real world scenario, we may not get access to information about all possible classes when the system is first trained. It is more realistic to assume that we will obtain class-specific data incrementally as time goes by. Therefore, in such a scenario, we require that our model can be adapted with new information made available without hampering the performance on what has been learnt so far. Although a natural task for human beings, it is a difficult task for an intelligent machine due to the possibility of catastrophic forgetting [17]. A trained model tends to forget old tasks when learning new information. There are three different streams of work in the literature addressing such an

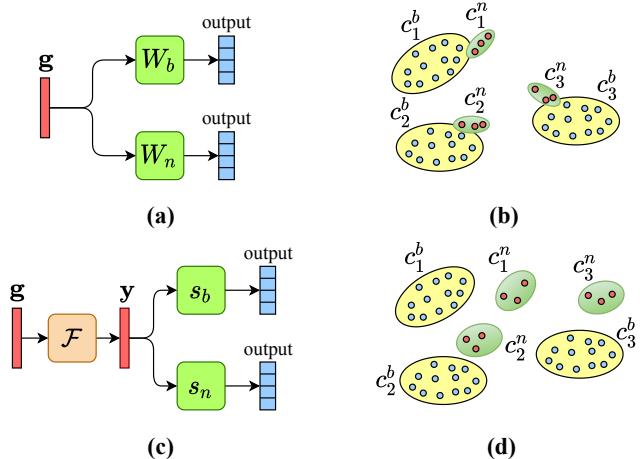


Figure 1: (a) Knowledge distillation as described in [16] does not work on few-shot class-incremental learning [32] since adding new tasks appends new trainable weights (W_n) to the network in addition to base weights (W_b). (b) The impact of using only a few instances of novel classes. As few samples are not sufficient to learn new parameters, the network gets biased towards base classes, overfitted on few examples of novel classes, and not well-separated from base classes. (c) Our semantically guided network does not add new parameters while adding new classes incrementally. We only include word vectors of new tasks (s_n) in addition to the base classes (s_b) and keep fine-tuning the base network (\mathcal{F}). (d) As a result, the knowledge distillation process can help the network, remembering base training, generalizing to novel classes, and finding well-separated representation of classes.

incremental or continual learning paradigm [33]. Firstly, *task-incremental learning* divides all classes into different tasks, where each task contains a few classes, and then learns each task individually. The task labels of the test instances are made available during testing which means the model does not need to predict the correct label between all classes but only between classes that are defined for a specific task. Secondly, *domain-incremental learning* does

^{*}denotes equal contribution.

not reveal the task label at test time, but the model always solves the current task at hand without inferring the true class label. Thirdly, *class-incremental learning* predicts the class label between all classes during test time as the output of all tasks are merged into one unified classifier without having access to the task label. Being the most realistic of the three, in this paper, we are interested in this third setting. Furthermore, in many applications, new tasks (a set of novel classes) come with only a few examples per class, making the class-incremental learning even more challenging. This setting is called *few-shot class-incremental learning* (FSCIL) [32]. The main challenges in FSCIL are catastrophic forgetting of already acquired knowledge and overfitting the network to novel classes. Challenges of that nature are addressed by the work on *knowledge distillation* in [11]. However, [32] showed that knowledge distillation is not the preferred approach for FSCIL due to class imbalance in the few-shot scenario and the performance trade-off between novel classes and base classes. In this paper, we propose an augmented knowledge distillation approach suitable for the case of few-shot incremental learning.

In order to apply knowledge distillation to novel tasks, scores of the previously trained model are needed as well as many instances of the new classes to be learned. Those new instances help to learn the new trainable weights that are added while learning novel tasks. For incremental learning with few-shot data, we can preserve previous scores but cannot provide enough samples to learn the corresponding weights for novel classes. For this reason, knowledge distillation [32] becomes a difficult problem in our case. Addressing this issue, we take advantage of a semantic word vector (word2vec [18] or GloVe [22]) which provides a semantic representation for each class as auxiliary knowledge. Being inspired by the literature on zero-shot learning [7, 6, 5, 20, 37, 24, 25], given an image as input, we estimate the semantic word vectors for the input instead of directly predicting its class label. Then, we measure the similarity of the predicted word vectors with the word vectors from the set of possible class labels, followed by a softmax layer applied to the similarity values to get the final score of the classes. One key benefit of this approach is that adding new classes while training on novel tasks does not come with new weights to train because the model attempts to predict fixed-length word vectors as an intermediate representation. No matter how many classes are present during fine-tuning, the network continues with its previous task of estimating the word vectors. In this new set-up, the distillation loss can easily accommodate new classes. One challenge of this approach is to obtain a good alignment of visual and semantic word vectors of few-shot instances. To address this issue, we employ automatically assigned superclass information of classes to train multiple embedding modules in parallel after the backbone network. The set

of superclasses is attained from the semantic word vector space representations of the base task, and are then held fixed for the novel classes that follow. We determine an embedding for each superclass during training such that each embedding sees only the superclass set of classes. Hence, given a novel class, there is a selection of embeddings that each may be more or less suited. We employ an attention module [8] to merge multiple embedding outputs and a loss to train the alignment appropriately with few-shot instances. It helps the network not to overfit on the few-shot instances as well as not becoming biased to the base classes. Figure 1 describes the key differences between conventional works and our method. With our proposed approach, we successfully beat the current state-of-the-art [32] on MiniImageNet, CUB200, and the CIFAR100 datasets thanks to the combined effect of using the auxiliary semantic information from word vectors and knowledge distillation in concert.

In summary, the contributions of this paper are: **(1)** A semantically-guided knowledge distillation approach for few-shot class-incremental learning using semantic word vectors, **(2)** A new visual-semantic alignment strategy for few-shot class-incremental learning using automatically assigned superclass annotations, **(3)** Extensive experiments validating the approach on MiniImageNet, CUB200, and CIFAR100 while achieving new state-of-the-art results.

2. Related work

Incremental learning: Incremental learning means learning from a sequence of data which appear over time. In the literature [33], incremental learning techniques are categorized into three groups, task-incremental learning [4, 28, 21], domain-incremental learning [39, 29], and class-incremental learning [26, 1, 13, 36, 31]. In this paper, we are only concerned with the third category, class-incremental learning, as we consider a unified output where the task label is not available during test time. Rebuffi *et al.* [26] keeps an “episodic memory” of the samples and incrementally adapts the nearest-neighbor classifier for the novel tasks. Castro *et al.* [1] proposed an end-to-end incremental learning method. In this method, a knowledge distillation loss is used to keep information about previously seen classes, and a classification loss is employed to learn the new classes. Hou *et al.* [13] introduced a novel approach for incrementally learning a unified classifier that reduces the imbalance between old and new classes by cosine similarity, which excludes the bias in the classifier. Wu *et al.* [36] proposed a method for large scale incremental learning, where they correct the bias in the output of the model with the help of a linear model. Simon *et al.* [31] propose a novel approach to the arsenal of distillation techniques. They construct low-dimensional manifolds for previous and current responses and minimize the dissimilarity between the geodesic responses connecting the manifolds.

Few-shot incremental learning: There are not many works addressing the FSCIL setting. Tao *et al.* [32] proposed this setting for the first time. They utilize a neural gas (NG) network to learn and maintain the topology of the feature manifold produced by various classes. Specifically, they introduce a method that alleviates the forgetting of the old classes by stabilizing the topology of the NG and enhancing the representation learning for few-shot novel classes by expanding and modifying NG to novel training samples. There is another category, called dynamic few-shot learning (DFSL) [9, 27, 38, 12], which is similar to FSCIL. Some works call this setting incremental few-shot learning but for clarity, in this work, we call this setting DFSL. The only difference is that in DFSL, there are only two sequences of tasks, in comparison to FSCIL, which contains several tasks. The first task containing many training samples is called base task, and the second task containing only a few training samples is called a novel task. Gidaris *et al.* [9] proposed an attention-based method that generates a classifier for the novel task from the classifier of the base task. Mengye *et al.* [27] notes that the method of recurrent back-propagation can back-propagate within the optimization process and helps the learning of novel tasks. Yoon *et al.* [38] proposed a method which obtains a task-adaptive representation for novel tasks based on the information provided from the base task by an attention module.

Knowledge distillation: Knowledge distillation is a well-known procedure that is employed in incremental learning to address catastrophic forgetting. Distillation loss was initially introduced to convey knowledge between separate neural networks [11]. Later, Li *et al.* [16] used a distillation loss to preserve the knowledge of the old tasks while learning the new ones using a classification loss. Shmelkov *et al.* [30] proposed a method where the embedding and the classifier are trained together without the need for keeping samples of the training data. Castro *et al.* [2] introduced an end-to-end method which consists of a classification loss for learning novel tasks and a distillation loss to retain information of the old task. Zhang *et al.* [40] introduced an approach to train an individual network for the novel classes, and then merging this new network with the network based on previous classes using a double distillation objective. Zhao *et al.* [41] employed knowledge distillation, at first, to keep the discrimination of old classes. Next, to further keep the balance between old and new classes, they introduced a method to refine the bias weights in the FC layer following the regular training. While it has been shown in [32] that regular knowledge distillation is not working well in the FSCIL setting, we offer, in this work, a method that enables the use of knowledge distillation for this purpose.

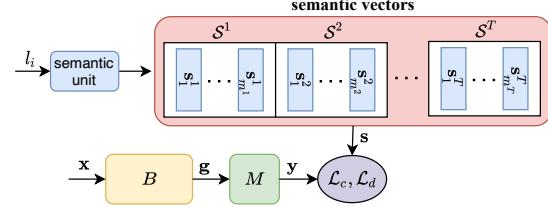


Figure 2: Our simplified proposed architecture for knowledge distillation. In this design, the input image \mathbf{x} is forwarded into the backbone B to extract a feature representation $\mathbf{g} \in \mathbb{R}^u$. Then, the extracted feature \mathbf{g} is mapped into the semantic domain via a mapping module M to form the estimation of the semantic vector $\mathbf{y} \in \mathbb{R}^d$.

3. Method

3.1. Problem Formulation

Suppose, there is a sequence of disjoint tasks $\mathcal{D} = \{\mathcal{D}^1, \dots, \mathcal{D}^T\}$, where $\mathcal{C}^t = \{c_1^t, \dots, c_{m_t}^t\}$ is the set of classes in the task \mathcal{D}^t . Additionally, a set of d -dimensional semantic class embeddings for each class label in the task \mathcal{D}^t defined as S^t is available during training. To be more specific, in the task $\mathcal{D}^t = \{(\mathbf{x}_i^t, l_i^t, \mathbf{s}_i^t)\}_{i=1}^{|\mathcal{D}^t|}$, \mathbf{x}_i^t is the i -th sample, l_i^t is its associated ground truth, and \mathbf{s}_i^t is its associated semantic representation. There are many training samples in the first task \mathcal{D}^1 , termed the base task. However, in the following tasks $t > 1$, termed the novel task, there are only a few training samples (5-shots per class) for each class. It is essential to mention that the classes between all tasks are disjoint, *i.e.*, $\mathcal{C}^i \cap \mathcal{C}^j = \emptyset, \forall i, j \in \{1, \dots, T\}$, where $i \neq j$. The objective of our work is to incrementally train a model with a unified output, while only training samples of the t -th task is available at the t -th session. At test time, we expect the trained model on task \mathcal{D}^t to predict the output for the current task and all the previous tasks $\{\mathcal{D}^1, \dots, \mathcal{D}^{t-1}\}$.

3.2. Knowledge Distillation

Knowledge distillation is a common approach [26, 1, 13] for incremental learning to address the catastrophic forgetting. Even though it has shown promising results on incremental learning, it cannot be employed directly to few-shot class incremental learning (FSCIL) due to the imbalanced data and trade-off issues [32]. In this paper (see Figure 2), we will show how to successfully take advantage of knowledge distillation in the FSCIL setting. To illustrate our proposed method for knowledge distillation, we use a simplified version of our proposed architecture. In this scheme, the input image \mathbf{x} goes into the backbone B , which is trained on only the first task \mathcal{D}^1 , as we have many training samples in this task, and is kept frozen on other tasks. The output of the backbone B is a feature representation $\mathbf{g} \in \mathbb{R}^u$. Next, the mapping network M is used to

project the feature representation \mathbf{g} into the semantic domain, where the projected feature $\mathbf{y} \in \mathbb{R}^d$ is aligned with its associated semantic representation $\mathbf{s} \in \mathbb{R}^d$. Lets assume we want to train the model with the task \mathcal{D}^t , which has m or $|\mathcal{C}^t|$ classes, and the number of classes of the previous tasks are n . Also, we choose one representation for each class of the former tasks saved in a small memory \mathcal{M} , where each prototype $\mathbf{g}^{\mathcal{M}} \in \mathbb{R}^u$ is the average of all training samples from each class. The aim is to map the input image \mathbf{x}_i into the semantic domain \mathbf{y}_i by the function $\mathbf{y}_i = \kappa_{\theta}(\mathbf{x}_i)$, which consists all training parameters of the backbone B and the mapping module M . We consider the cosine distance as the similarity between the projected feature \mathbf{y}_i and the semantic representation \mathbf{s}_k , $d(\mathbf{s}_k, \mathbf{y}_i) = \cos(\mathbf{s}_k, \mathbf{y}_i)$. If the output of the classifier before adding the novel task \mathcal{D}^t is $\mathbf{d}' = [d'(\mathbf{s}_1, \mathbf{y}_i), \dots, d'(\mathbf{s}_n, \mathbf{y}_i)]$, and if the output of the classifier after adding the novel task is $\mathbf{d} = [d(\mathbf{s}_1, \mathbf{y}_i), \dots, d(\mathbf{s}_n, \mathbf{y}_i), d(\mathbf{s}_{n+1}, \mathbf{y}_i), \dots, d(\mathbf{s}_{n+m}, \mathbf{y}_i)]$, the distillation loss is defined as,

$$\mathcal{L}_d = -\frac{1}{N_c} \sum_{i=1}^{N_c} \sum_{k=1}^n p_k \log(q_k), \quad (1)$$

$$p_k = \frac{e^{-d'(\mathbf{s}_k, \mathbf{y}_i)/\tau}}{\sum_{j=1}^n e^{-d'(\mathbf{s}_j, \mathbf{y}_i)/\tau}}, \quad q_k = \frac{e^{-d(\mathbf{s}_k, \mathbf{y}_i)/\tau}}{\sum_{j=1}^n e^{-d(\mathbf{s}_j, \mathbf{y}_i)/\tau}},$$

where $\mathbf{y}_i = M(\mathbf{g}_i)$, and $\mathbf{g}_i \in B(\mathbf{x}_i^t) \cup \mathcal{M}$. Here, τ is the temperature scalar. The τ is set to 2 for all experiments. Also, N_c is the number of samples in the task \mathcal{D}^t and memory \mathcal{M} , i.e., $N_c = |\mathcal{D}^t| + |\mathcal{M}|$.

Remark. In this paper, we show that knowledge distillation can be used for few-shot class incremental learning. To this end, we use a semantic word vector in our pipeline as additional information.

Additionally, we employ a cross-entropy loss as a classification loss,

$$\mathcal{L}_c = -\frac{1}{N_c} \sum_{i=1}^{N_c} \sum_{k=1}^{n+m} \mathbf{1}[l_i == k] \log\left(\frac{e^{-d(\mathbf{s}_k, \mathbf{y}_i)}}{\sum_{j=1}^{n+m} e^{-d(\mathbf{s}_j, \mathbf{y}_i)}}\right), \quad (2)$$

where $\mathbf{1}[\cdot]$ is the indicator function. Then, the total loss is defined as,

$$\mathcal{L} = \lambda_1 \mathcal{L}_d + \lambda_2 \mathcal{L}_c,$$

where λ_1 and λ_2 are used to control the effect of each loss in the final loss \mathcal{L} .

This approach does not add new parameters during the incremental learning stage. Instead, it only fine-tunes M with newly available data incrementally. It helps the distillation process and learning without forgetting. Also, this



Figure 3: For a novel class, Hyena, our model finds a set of base class (e.g., Dalmatian, Saluki, Meerkat) instances, which reside in close proximity in the semantic embedding space. The presence of shared semantics (e.g., face, body-shape, 4-feet, short-tail) between novel and base classes help to understand Hyena as a novel class and not to forget base classes.

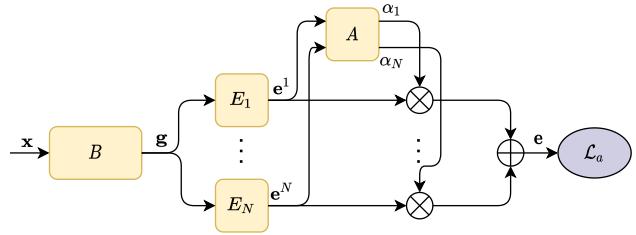


Figure 4: Our proposed architecture is designed specifically for tasks with only a few training samples. The input image \mathbf{x} goes into the backbone B , which generates a global representation $\mathbf{g} \in \mathbb{R}^u$. Then, the feature \mathbf{g} separates into several embedding modules E . The attention module A is used to merge all of them to generate the final representation $\mathbf{e} \in \mathbb{R}^u$.

approach reduces the data imbalance problem of few-shot learning. The network may not have had the opportunity to see the novel objects, however, newly discovered novel objects may very well share semantic properties with the objects it has already seen among the base classes. For example, if hyena is a novel class, the network did not have any opportunity to see any hyena. However, many typical hyena attributes like ‘face’, ‘body’, etc., are seen by the network from other base class animals (e.g., Dalmatian, Saluki, Meerkat). It helps the network to understand hyena by reducing the data dependency. An intuitive illustration is shown in Figure 3.

3.3. Multiple embeddings for few shot tasks

The main challenge with the tasks containing a few training samples is the overfitting issue, as it is difficult to learn the distribution of a task sufficiently with only a few training data. To this end, in our proposed method (see Figure 4), we generate multiple embeddings, where each is designed specifically for a group of classes. We use the word vector semantic to separate classes into several groups. We design these groups based on the classes that we have in the first task \mathcal{D}^1 because the first task contains more classes with many training samples. After that, we train each embedding E_i using a cross-entropy loss based on the superclass

labels obtained in the previous stage. Each embedding observes only part of the entire set of classes, which results in embeddings that are experts on a particular superclass. For the novel tasks, we assign a superclass label to each class based on the cluster that we obtained in the previous stage.

Superclass: The number of the embeddings is defined by the superclass knowledge obtained from the semantic word vector space. In the semantic space, there is a class embedding \mathbf{s} for each class. We apply k -means clustering of the semantic representations of the classes of the first task \mathcal{D}^1 . In this way, similar classes fall into the same category. After applying k -means clustering, we assign a superclass label to each class $\mathcal{R} = \{1, \dots, N\}$. For the other tasks, which have only a few training samples, we use the cluster centers (obtained for the first task \mathcal{D}^1) to assign superclass labels to the classes in these tasks. To assign a superclass label to novel classes, we simply calculate the minimum Euclidean distance between the semantic vector of novel class and cluster centers.

Attention: For the i -th sample in the task \mathcal{D}^t , the embedding representation $\mathbf{e}_i \in \mathbb{R}^u$ is the weighted average of all private modules $\mathbf{e}_i^k \in \mathbb{R}^u$, $k \in \mathcal{R}$, where a neural network determines weights, which is an attention A module. Furthermore, the weights must sum to 1 to stay invariant to the number of private modules. Then, the final embedding representation is $\mathbf{e}_i = \sum_{k=1}^N \alpha^k \mathbf{e}_i^k$, where α^k is, where

$$\alpha^k = \frac{e^{\mathbf{w}^T \tanh(\mathbf{V}(\mathbf{e}_i^k)^T)}}{\sum_{j=1}^N e^{\mathbf{w}^T \tanh(\mathbf{V}(\mathbf{e}_i^j)^T)}}. \quad (3)$$

Here, $\mathbf{w} \in \mathbb{R}^{L \times 1}$ and $\mathbf{V} \in \mathbb{R}^{L \times d}$ are trainable parameters.

Training: We train the attention modules A so that the final representation \mathbf{e} becomes similar to the corresponding embedding module given the associated superclass label. To this end, we use the following loss function,

$$\mathcal{L}_a = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{e^{-d(\mathbf{e}_i, \mathbf{e}_i^k)}}{\sum_{j=1}^N e^{-d(\mathbf{e}_i, \mathbf{e}_i^j)}} \quad (4)$$

where the superclass label for i -th sample is $k \in \mathcal{R}$, and N_t is the number of samples in the task \mathcal{D}^t . This approach helps to not overfit the network on only a few novel class data. Multiple embeddings specialized on related classes, belonging to the same superclass, describe the novel instances. Combining multiple embedding features and the global feature enables a strong generalization when classifying both base and novel classes.

3.4. Model Overview

The final proposed architecture is shown in Figure 5. The input image \mathbf{x} goes into the backbone B , which is pretrained on the first task D^1 with a cross-entropy loss, to generate

Algorithm 1 Training procedure of the proposed method

```

1: function TRAIN( $\mathcal{D}$ )
2:   Hyperparameters:  $\lambda_1, \lambda_2, \lambda_3, N$ 
3:    $\mathcal{M} \leftarrow \{\}$ 
4:   Train the backbone  $B$  on the first task  $\mathcal{D}^1$ 
5:   Apply  $k$ -means clustering, where  $k = N$ , on base
   semantic vectors and assign a superclass label
6:   Train  $N$  embedding  $E_i$  (Fig. 4) on the base task
    $\mathcal{D}^1$  using superclass labels as cluster identity
      ▷ Train the final architecture (Fig. 5)
7:   for  $h = 1$  to  $epochs$  do
8:     Calculate the classification loss using Eq 2
9:     Calculate the attention loss using Eq 4
10:     $\mathcal{L} = \lambda_1 \mathcal{L}_c + \lambda_3 \mathcal{L}_a$ 
11:    Backpropagate and update  $A$ , and  $M$ 
12:   end for
13:    $\mathcal{M} \leftarrow \text{UPDATEMEMORY}(\mathcal{D}^1, \mathcal{M}, \mathcal{C}^1)$ 
14:   for  $t = 2$  to  $T$  do
15:     Assign superclass labels to task  $\mathcal{D}^t$  using
       cluster centers obtained for the base task
16:     for  $h = 1$  to  $epochs$  do
17:       Calculate the classification loss using Eq 2
18:       Calculate the distillation loss using Eq 1
19:       Calculate the attention loss using Eq 4
20:        $\mathcal{L} = \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_d + \lambda_3 \mathcal{L}_a$ 
21:       Backpropagate and update  $A$ ,  $M$ , and  $E$ 
22:     end for
23:      $\mathcal{M} \leftarrow \text{UPDATEMEMORY}(\mathcal{D}^t, \mathcal{M}, \mathcal{C}_{m^t})$ 
24:   end for
25: end function
26: function UPDATEMEMORY( $\mathcal{D}^t, \mathcal{M}, \mathcal{C}_{m^t}$ )
27:   for  $c = 1$  to  $|\mathcal{C}_{m^t}|$  do
28:     Calculate a prototype  $\mathbf{g}_c^{\mathcal{M}}$  for each class by
       averaging of all training samples from each class
29:      $\mathcal{M} \leftarrow \mathcal{M} \cup (\mathbf{g}_c^{\mathcal{M}}, \mathbf{l}_c^t)$ 
30:   end for
31:   return  $\mathcal{M}$ 
32: end function

```

the global feature representation $\mathbf{g} \in \mathbb{R}^u$. When training the other tasks $\mathcal{D}^t, t > 1$, the backbone network B is kept frozen to prevent overfitting to classes with only a few training samples. Then, the extracted feature \mathbf{g} is fed into N embedding modules E_i , $i = 1, \dots, N$, where they are trained on the first task based on the superclass information and are updated for the novel tasks. To this end, the output of all embedding $\mathbf{e}^k \in \mathbb{R}^u$ are fused based on the weights generated by an attention module A . Then, the global feature \mathbf{g} is concatenated with the feature generated from the embedding modules \mathbf{e} to form the feature $\mathbf{f} \in \mathbb{R}^{2u}$. In the next stage, the feature \mathbf{f} is projected from the visual domain to the semantic domain by the mapping module M to form

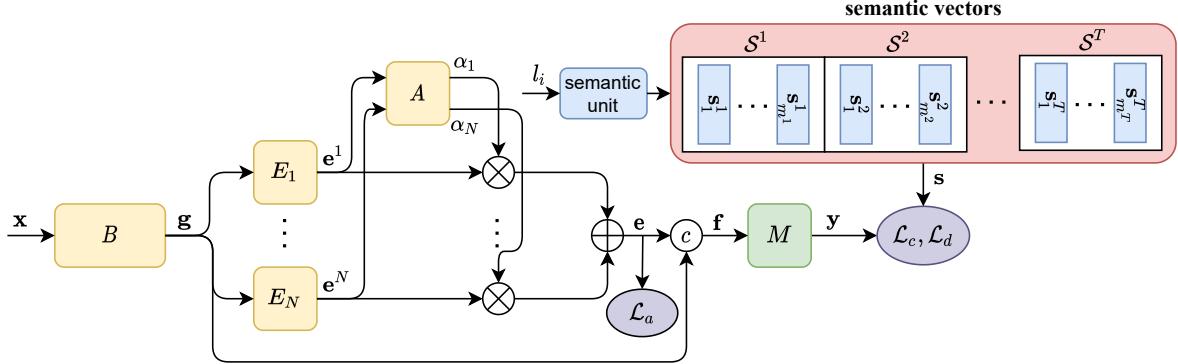


Figure 5: The proposed architecture. The image \mathbf{x} is forwarded to the Backbone B to obtain the global feature representation $\mathbf{g} \in \mathbb{R}^u$. The feature \mathbf{g} is fed into several embedding modules to get a representation $\mathbf{e} \in \mathbb{R}^u$ based on the superclass information obtained from the word vector semantic space. After this stage, the feature vector $\mathbf{f} \in \mathbb{R}^{2u}$ is obtained by concatenating the global and embedding representations. In the next stage, the \mathbf{f} is projected via a mapping module M from the visual space into the semantic space $\mathbf{y} \in \mathbb{R}^d$ to align the visual information with their associated semantic representation.

feature $\mathbf{y} \in \mathbb{R}^d$, where it is aligned with its associated semantic representation $\mathbf{s} \in \mathbb{R}^d$. In order to train our proposed architecture, we use the following loss function,

$$\mathcal{L} = \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_d + \lambda_3 \mathcal{L}_a, \quad (5)$$

where λ_1 , λ_2 , and λ_3 are used to control the effect of each term in the final loss function. The pseudo code of the training procedure is shown in Algorithm 1.

4. Experiments

This section contains two parts. In the first part, we evaluate our method on FSCIL [32], and we conduct a set of ablation studies to investigate the recommended approach. Next, we investigate the dynamic few-shot learning [9] (DFSL) setting to demonstrate the capability of our proposed method in a different setting.

4.1. Experiments on FSCIL

Datasets: We evaluate our proposed method on three well-known datasets, MiniImageNet [34], CUB200 [35], and CIFAR100 [15]. MiniImageNet contains 100 classes, where each class include 500 training samples and 100 testing samples. The size of each image is 84×84 . CUB200 contains 200 fine-grained classes, separated into 6000 training images, and 6000 testing images. The image size in this dataset is 224×224 . CIFAR100 includes 100 classes, where each class has 600 images, separated into 500 training images and 100 test images. Each image has a size of 32×32 . In this paper, we follow the setting proposed by [32]. In this setting, for MiniImageNet and CIFAR100, 60 and 40 are selected as the number of base and novel classes. For novel classes, a 5-way 5-shot setting is considered. There are nine sessions for MiniImageNet and CIFAR100 datasets (1 base

session + 8 novel sessions). For the CUB200 dataset, a 10-way 5-shot setting is considered, where 100 classes are selected as base classes, and the remaining 100 classes are split into 10 sessions.

Semantic Features: We use unsupervised word vectors trained on an unannotated text corpus as a class semantic embedding. For MiniImageNet, CUB200, and CIFAR100, we employ 1000, 400, and 300 dimensional word2vec [18], respectively. For the ablation study, we also use the 300 dimensional GloVe [22] for the CUB200 dataset.

Validation: To find hyperparameters, we conducted a grid search. The ranges we consider, $\lambda_1, \lambda_2, \lambda_3 \in [0, 2]$, # of embedding modules $E \in [1, 10]$, and the temperature $\tau \in [0, 5]$ value. We split the training set into two sets: base set, which consists 60% of the training classes, and validation set which consists rest of the classes added incrementally.

Implementation details¹: We employ ResNet18 [10] for the backbone B , where features of the input image are derived from the final pooling layer with 512 dimensions. The backbone B is trained on the first/base task, and kept fixed for the following tasks. For the embedding modules E_i , we use a fully connected layer with 512 dimensions. Moreover, we utilize a few fully connected layers for the attention module A . Ultimately, for the mapping module M , we used three fully-connected layers with 512, 728, and d , which is the dimension of the semantic word vector, hidden units, where all layers have a ReLU function. In all experiments, we utilize the Adam optimizer [14], where the learning rate and batch size were set to 0.001 and 128, respectively. The number embedding modules for MiniImageNet, CUB200, and CIFAR100 are selected as 3, 5, and 3 respectively. Also, the value of λ_1 , λ_2 , and λ_3 are 0.7, 1.1, and 0.6 respectively

¹Code is available at: <https://github.com/ali-chr/Semantic-aware-Knowledge-Distillation-for-Few-ShotClass-Incremental-Learning>

Table 1: CUB200 results with ResNet18 based on the 10-way 5-shot setting.

Method	Sessions										
	1	2	3	4	5	6	7	8	9	10	11
iCaRL [26]	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16
EEIL [1]	68.68	53.63	47.91	44.20	36.30	27.46	25.93	24.70	23.95	24.13	22.11
NCM [13]	68.68	57.12	44.21	28.78	26.71	25.66	24.62	21.52	20.12	20.06	19.87
AL-MML [32]	68.68	62.49	54.81	49.99	45.25	41.40	38.35	35.36	32.22	28.31	26.28
Ours	68.23	60.45	55.70	50.45	45.72	42.90	40.89	38.77	36.51	34.87	32.96

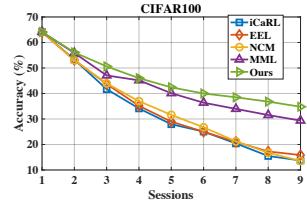
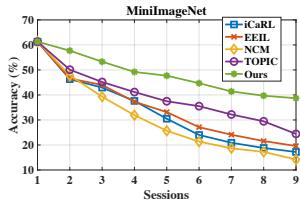


Figure 6: Results on MiniImageNet and CIFAR100 based on the 5-way 5-shot FSCIL setting.

for all datasets. Since we use semantic word vectors in our pipeline, we get slightly different result on the first task in comparison to the setting proposed by [32].

4.1.1 Results

In this part, we compare our proposed method with state-of-the-art [26, 1, 13, 32] on the 5/10-way and 5-shot setting. Figure 6 presents our results on MiniImageNet and CIFAR100 using the 5-way 5-shot setting. Moreover, we show the result on the CUB200 dataset in Table 1. Overall, on these three datasets, our method beats all state-of-the-art methods. As additional sequences of new tasks arrive, our approach shows its advantages to the other methods. To be more specific, in MiniImageNet, in the last session, we get 39.04% accuracy, while the second-best one (TOPIC) achieves 24.42% accuracy which demonstrates that our method surpasses the state-of-the-art by a large margin (more than 14%). On CIFAR100, our method reaches the absolute accuracy of 34.80%, while the second-best (TOPIC) one accuracy is 29.37%. Also, on CUB200, our approach achieves 32.96% in the last session, which is superior to the other approaches.

4.1.2 Ablation study

Impact of loss function and embedding module: In this part, we report the individual effect of the distillation loss \mathcal{L}_d and attention loss \mathcal{L}_a functions in the total loss. As can be seen in Figure 7, where only the accuracy of the last session is shown, \mathcal{L}_d is more effective than \mathcal{L}_a , as \mathcal{L}_d helps the model to remember the previously seen tasks, while \mathcal{L}_a helps the model to generate a richer feature representation for the novel tasks, which have only a few training samples.

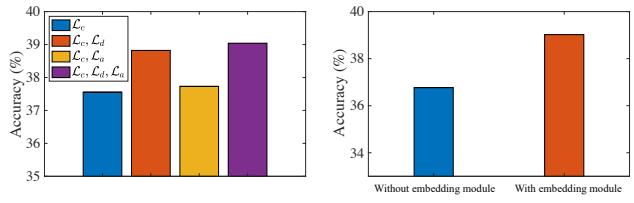


Figure 7: (Left) The influence of \mathcal{L}_d and \mathcal{L}_a losses, and (Right) the impact of using multiple embedding

We also evaluate the effect of multiple embedding modules in our proposed algorithm. To understand the influence of these in our design, we utilise a baseline architecture which does not include the multiple embedding module and has only the backbone B . As presented in Figure 7, our proposed method exceeds the baseline, which indicates the value of the embedding module inside our architecture.

Impact of different backbones and semantic information: In this ablation study, we evaluate the influence of the alternative semantic word vector, GloVe. We also conduct this analysis using two backbones, ResNet18 and ResNet101, to understand the effect of different backbones. The accuracy of our approach is evaluated on test samples of the base task and the novel tasks. The accuracy of the base task Acc_b is the performance of our model on the first/base task \mathcal{D}^1 , and the accuracy of the novel task Acc_n is considered the performance on the test samples of the current task and all previous novel tasks $\{\mathcal{D}^2, \dots, \mathcal{D}^t\}$. To evaluate the contribution of the base and novel instances in the final accuracy, we also report the Harmonic Mean (HM) [3] of the accuracy of the base and novel classes.

The values for Acc_b , Acc_n , and HM are shown in Figure 8. As can be seen, the combination of GloVe and ResNet101 forget less of the base classes, which is followed by better learning on novel classes. As a result, it provides a greater HM value. It reveals that having a different type of semantic vector is valuable in our suggested pipeline. Additionally, a deeper backbone (ResNet101) is useful in FS-CIL as it generates more valuable feature representations. We present our results starting from session 2 to 11 because there is no novel task in session 1.

Impact of Temperature, τ : We notice that the method is robust to τ in a wide range. Figure 9 (left) shows the ef-

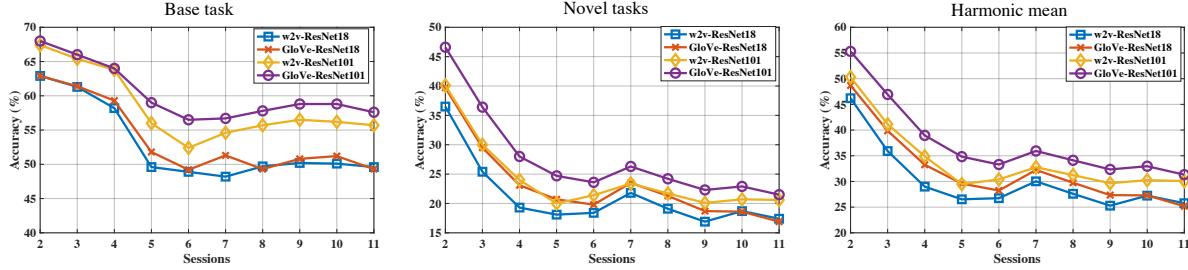


Figure 8: The results of our method on base (left), novel tasks (middle), and their harmonic mean (right) in different incremental sessions.

Table 2: miniImageNet 64+5-way results

Method	1-shot		5-shot	
	accuracy	Δ	accuracy	Δ
Imprint [23]	$41.34 \pm 0.54\%$	-23.79%	$46.34 \pm 0.54\%$	-25.25%
LwoF [9]	$49.65 \pm 0.64\%$	-14.47%	$59.66 \pm 0.55\%$	-12.35%
AA [27]	$54.95 \pm 0.30\%$	-11.84%	$63.04 \pm 0.30\%$	-10.66%
XtarNet [38]	$55.28 \pm 0.33\%$	-13.13%	$66.86 \pm 0.31\%$	-10.34%
Ours	$58.07 \pm 0.27\%$	-10.83%	$68.03 \pm 0.44\%$	-8.25%

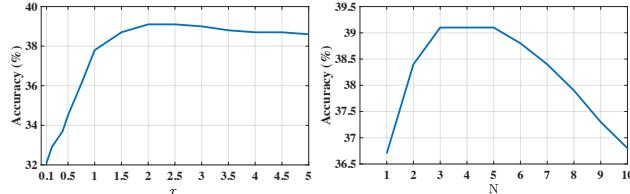


Figure 9: Impact of (left) Temperature, τ and (right) number of Supercluster (N) on the last task of MiniImageNet.

fect of τ on MiniImageNet on the last task. Increasing τ is helpful for the forgetting issue until $\tau = 2$ by balancing the contribution of p_k and q_k , after that the accuracy decreases.

Impact of number of Supercluster (N): In the Figure 9 (right), we present the performances on MiniImageNet on the last task when N varies. When N is very small (e.g., 1 or 2), it seems that our method only extracts global information, merging semantics of different super-categories, thus failing to encode local information and subtle differences within each super-category. Conversely, when N is very high (e.g., ≥ 7), fewer classes fall into the same super-category, which leads to indistinguishable information. As we concatenate both global and local information, with high N , local information does not help. Thus, using $N = 10$, we got a similar performance to only global, $N = 1$. Empirically, $3 \leq N \leq 5$ both global and local information work together for MiniImageNet.

4.2. Experiments on DFSL

To further evaluate our proposed method, we apply our design to the DFSL setting. As discussed in the related work section, DFSL is similar to FSCIL, though the only difference is that we have only two sequences of tasks in DFSL whereas there are various sequences of tasks in FSCIL. In

this section, we use the setting introduced by [9]. In this experiment, we use miniImageNet, which is split into two tasks, a base task and a novel task. The base task consists of 64 classes, while the novel task has five classes which are randomly selected in an episode way from 20 classes. It is necessary to mention that the classes in the base and novel tasks are disjoint. In each episode, the training (support) set is built by selecting 1 or 5 examples of the novel classes, representing a 1-shot or 5-shot scenario. The test (query) set consists of examples of the base and novel tasks. In this experiment, we use ResNet12 as adopted in [19] as the backbone. The forgetting in the DFSL setting is calculated as the gap between joint and individual performances for base and novel classes. Individual performances for the base/novel task is calculated when only the base/novel classifier is used. Δ_b and Δ_n are used to indicate the gap between the base and novel tasks, respectively. The average of these gaps is represented as $\Delta = (\Delta_b + \Delta_n)/2$, which show the amount of forgetting,

We compare our method with state-of-the-art methods in Table 2. As can be seen, our method outperforms all other methods on joint accuracy and forgetting Δ on both the 1-shot and 5-shot settings.

5. Conclusion

We proposed a semantic-aware knowledge distillation method for few-shot class incremental learning (FSCIL). Due to a limited amount of training data for the novel classes, the knowledge distillation technique as previously used did not work well in this problem domain. In this paper, using auxiliary information from class semantics (word vectors), we propose a new FSCIL method where knowledge distillation can indeed perform learning without forgetting. Moreover, we offer an attention mechanism based on multiple embedding representations of visual data to describe the novel classes that also demonstrates better generalization. Three well-known datasets, MiniImageNet, CUB200, and CIFAR100, are used to show that class semantics can be a useful source of information for knowledge distillation. We outperform state-of-the-art methods of FSCIL by a large margin.

References

- [1] Francisco M. Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-End Incremental Learning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV 2018 - European Conference on Computer Vision*, Munich, Germany, Sept. 2018.
- [2] Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Cordelia Schmid, and Karteek Alahari. End-to-End Incremental Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [3] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 52–68, Cham, 2016. Springer International Publishing.
- [4] Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [5] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Mitigating the hubness problem for zero-shot learning of 3d objects. In *British Machine Vision Conference (BMVC’19)*, 2019.
- [6] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Transductive zero-shot learning for 3d point cloud classification. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 912–922, 2020.
- [7] Ali Cheraghian, Shafin Rahman, and Lars Petersson. Zero-shot learning of 3d point cloud objects. In *International Conference on Machine Vision Applications (MVA)*, 2019.
- [8] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear Attention Networks for Person Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [9] S. Gidaris and N. Komodakis. Dynamic Few-Shot Visual Learning Without Forgetting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [11] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.
- [12] Jie Hong, Pengfei Fang, Weihao Li, Zhang Tong, Christian Simon, Lars Petersson, and Mehrtash Harandi. Reinforced Attention for Few-Shot Learning and Beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.
- [13] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a Unified Classifier Incrementally via Rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [14] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [15] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research).
- [16] Z. Li and D. Hoiem. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.
- [17] Michael McCloskey and Neal J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. volume 24 of *Psychology of Learning and Motivation*, pages 109 – 165. Academic Press, 1989.
- [18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.
- [19] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A Simple Neural Attentive Meta-Learner. In *International Conference on Learning Representations*, 2018.
- [20] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees G. M. Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *European Conference on Computer Vision*, August 2020.
- [21] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational Continual Learning. In *International Conference on Learning Representations*, 2018.
- [22] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [23] H. Qi, M. Brown, and D. G. Lowe. Low-Shot Learning with Imprinted Weights. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5822–5830, 2018.
- [24] Shafin Rahman, Salman Khan, and Nick Barnes. Transductive learning for zero-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [25] Shafin Rahman, Salman Khan, Nick Barnes, and Fahad Shahbaz Khan. Any-shot object detection. In *Computer Vision – ACCV 2020*, pages 89–106, Cham, 2021. Springer International Publishing.
- [26] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental Classifier and Representation Learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, 2017.
- [27] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard S. Zemel. Incremental Few-Shot Learning with Attention Attractor Networks, booktitle= Advances in Neural Information Processing Systems (NeurIPS), year = 2019.
- [28] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, , and Gerald Tesauro. Learning to

- Learn Without Forgetting By Maximizing Transfer and Minimizing Interference. In *International Conference on Learning Representations*, 2019.
- [29] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual Learning with Deep Generative Replay. In *Advances in Neural Information Processing Systems 30*, pages 2990–2999, 2017.
- [30] Konstantin Shmelkov, Cordelia Schmid, and Kartek Alahari. Incremental Learning of Object Detectors Without Catastrophic Forgetting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [31] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. On Learning the Geodesic Path for Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [32] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-Shot Class-Incremental Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [33] Gido M. van de Ven and Andreas S. Tolias. Three Scenarios for Continual Learning. *CoRR*, abs/1904.07734, 2019.
- [34] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3630–3638. Curran Associates, Inc., 2016.
- [35] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- [36] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large Scale Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [37] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [38] Sung Whan Yoon, Do-Yeon Kim, Jun Seo, and Jaekyun Moon. Xtarnet: Learning to Extract Task-Adaptive Representation for Incremental Few-Shot Learning. In *In Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.
- [39] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual Learning Through Synaptic Intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995, 06–11 Aug 2017.
- [40] J. Zhang, J. Zhang, S. Ghosh, D. Li, S. Tasci, L. Heck, H. Zhang, and C. . Jay Kuo. Class-incremental Learning via Deep Model Consolidation. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1120–1129, 2020.
- [41] B. Zhao, X. Xiao, G. Gan, B. Zhang, and S. T. Xia. Maintaining Discrimination and Fairness in Class Incremental Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13205–13214, 2020.