IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 9, 2019, accepted December 20, 2019, date of publication December 25, 2019, date of current version January 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962258

Semi-Supervised Learning for Fine-Grained Classification With Self-Training

OBED TETTEY NARTEY[1], GUOWU YANG[1,3], JINZHAO WU[3,4], AND SARPONG KWADWO ASARE[2]

[1]Big Data Research Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]School of Electronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[3]Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Guangxi University for Nationalities, Nanning 530006, China
[4]School of Computer Science and Electronic Information, Guangxi University, Nanning 530004, China

Corresponding author: Obed Tettey Nartey (ashong.nartey@std.uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572109 and Grant 61772006, in part by the Science and Technology Program of Guangxi under Grant AB17129012, in part by the Science and Technology Major Project of Guangxi under Grant AA17204096, in part by the special Fund for Scientific and Technological Bases and Talents of Gungxi under Grant 2016AD05050, and in part by the Special Fund for Bagui Scholars of Guangxi.

ABSTRACT Semi-supervised learning is a machine learning approach that tackles the challenge of having a large set of unlabeled data and few labeled ones. In this paper we adopt a semi-supervised self-training method to increase the amount of training data, prevent overfitting and improve the performance of deep models by proposing a novel selection algorithm that prevents mistake reinforcement which is a common thing in conventional self-training models. The model leverages, unlabeled data and specifically, after each training, we first generate pseudo-labels on the unlabeled set to be added to the labeled training samples. Next, we select the top-k most-confident pseudo-labeled images from each unlabeled class with their pseudo-labels and update the training data, and retrain the network on the updated training data. The method improves the accuracy in two-fold; bridging the gap in the appearance of visual objects, and enlarging the training set to meet the demands of deep models. We demonstrated the effectiveness of the model by conducting experiments on four state-of-the-art fine-grained datasets, which include Stanford Dogs, Stanford Cars, 102-Oxford flowers, and CUB-200-2011. We further evaluated the model on some coarse-grain data. Experimental results clearly show that our proposed framework has better performance than some previous works on the same data; the model obtained higher classification accuracy than most of the supervised learning models.

INDEX TERMS Fine-grained classification, pseudo-labels, self-training, semi-supervised learning.

I. INTRODUCTION

An exponential growth of image classification has been witnessed over the few two decades and its semantic organization has become exceedingly expensive and difficult to manually categorize. These large datasets are mostly organized in a hierarchical order, where the deeper one goes in the hierarchy, the finer the categories and the rare annotated training data becomes, therefore obtaining training data for fine-grained images becomes expensive as most of the time expert knowledge is needed [1]. Distinguishing between a dog and a car is easy because there are plenty of helpful visual cues. In comparison, the difference between fine-grained classes can be very subtle, and only a few key features matter. In feature space, these subcategories are fundamentally distinct from one another. Another difficult task that we face is labeling all categories of fine-grained images. A large number of images in the ever-changing world makes it overly expensive to gain labeled examples for every category. The fine-grained classification task emphasizes differentiating between hard-to-distinguish image classes such as species of birds, dogs, flowers or even models of automobiles. Most previous works have focused on tackling the intra-class variations in perspective, illumination and pose using some localizing techniques and augmenting the training data with additional ones from the web with a little focus on the intra-class similarities that tend to make powerful models not able to generalize well on fine-grained classification tasks. It was observed that these fine-grained classification tasks tackle the problem as Large Scale Visual Classification tasks with much focus on the

The associate editor coordinating the review of this manuscript and approving it for publication was Fatih Emre Boran.

inter-class variations as compared to the intra-class variations present in the former. Also, fine-grained datasets are a combination of small, non-uniform, and minute inter-class differences, that make the classification tasks challenging even for powerful state-of-the-art learning models. In view of the issues above, we propose a learning method focusing on the problem of semi-supervised learning for fine-grained visual classification where we try to train a model that generalizes well on target samples, given the condition that there is a provision of both well-labeled source samples and labeled together with unlabeled target samples at training time. We propose to assign pseudo labels to target samples, and via self-training, train the target-specific model with the pseudo-labels as if they were true labels.

Conventional Self-training is a semi-supervised learning method that can learn decision boundaries from samples. It is a commonly used method in domains, such as Natural Language Processing [2]–[4] and object detection and recognition [5]. Traditionally, it can learn a better decision boundary for the source and target samples with hand-crafted features with less consideration on features distribution matching. Thereby combining Convolutional Neural Network (CNN) with self-training becomes a powerful method that can learn a far better decision boundary and find a more advanced feature space that matches source and target samples distribution. This is somewhat similar to adversarial training based methods; however, it uses a simpler approach where feature learning is guided by a cross-entropy loss that enhances the closeness of the source and target features as well aligning the class-wise features. Self-training is carried out by occurring in turns; a generation of a set of pseudo-labels corresponding to a large selected metric probability score, and fine-tuning a network using the training set together with the generated pseudo-labels based on an assumption that the test or target samples with the highest prediction probability are selected to be added to the labeled training set. Sometimes the visual domain gap between training and test domains is usually different between classes. This can result in different degrees of difficulty for the network to learn transferable features for each class. In addition, there is an imbalance in the distribution of the various classes, and this causes prediction confidence problems for various classes, but this proposed work, focuses on the best way of selecting the highest predicted probability confidence pseudo-labels. In all, we propose a typical CNN-based self-training method for Fine-grained objects recognition with a focus on:

Improving deep neural networks for fine-grained classification by combining self-training and Convolutional Neural Network for fine-grained objects classification. We formulate a loss minimization scheme, solving it by using an end-to-end approach. Both domain-invariant features and a classifier are expected to be learned. Therefore, aiming at how to learn the discriminatory features by building a target specific network and feed it with artificially labeled samples. Self-training with a standard network architecture as a base-network, we leverage a classifier to artificially label unlabeled samples and retrain the classifier. However, this method does not assume that the labeled samples are drawn from different domains. We employ a k-fold cross-validation method to solve the class imbalance problem. In self-training for fine-grained classification, the implementation has no added overhead in training or prediction time and provides performance improvements both in fine-grained classification tasks and coarse-grain tasks that involve transfer learning with small amounts of training data. We obtained a great deal of performance on four of the most widely-used fine-grained recognition datasets, improving over some previous-best published methods.

II. RELATED WORK

Studies on classifying images are very important in the field of computer vision. Over the years, many image classifications have been done, most of them being done by the use of deep neural networks. Deep neural networks have over the years achieved tremendous performances on different datasets [6]–[9]. Generally, deep learning models are built to learn hierarchical feature representation directly from categorical data, where they have been very successful. However, in the face of sub-categorical classification, deep learning algorithms have struggled and obtained poor performance [10], due to less availability of well-annotated data, occlusion, high intra-class variance and a vast set of similarities present in these sets. In this work, we investigate a semi-supervised self-training technique for fine-grained recognition. The concept of fine-grained here can be seen as objects with similar properties or attributes. By associating attributes to subcategories, it becomes easier to estimate these attributes by conducting a classification task on fine-grained image samples. CNNs, since their inception from the early days of the 1990s, have been very consistent and competitive with other machine learning techniques for classifying images. Tasks such as character recognition, image classification and object recognition in videos have been successfully implemented with high detection and classification accuracies. Notably of these tasks where the best results have been achieved are MNIST [6], CIFAR [11] and the almighty ImageNet classification challenge [8], which has become the standard for evaluating for what is the current state-of-the-art in image classification and recognition. Several large and deep CNN algorithms [7]–[9], [12]–[14] have been proposed which have achieved different accuracies and losses that are considerably better than using conventional machine learning techniques.

A. FINE-GRAINED VISUAL CLASSIFICATION

This task has been well studied particularly since the inception of Deep Neural Networks (DNNs) [1], [15]–[22]. However, its applicability, in reality, is hindered by the limited amount of available well-annotated data. Some works such as [18] have used large available datasets and adapted them to a small set, whereas in some settings large scale noisy data

were used in training the models [23]. Such approach makes it difficult for models to easily generalize well to real-world tasks, given that these training images may have been derived from field guides. And because of the significant variations in the appearance of objects, between the real-world images and the training sets, model generalization is hampered.

B. SEMI-SUPERVISED SELF-TRAINING

The availability of large unlabeled data and less labeled ones is a common challenge in the application of computer vision models. A better way to solve this challenging task is by using a semi-supervised learning approach that will make use of both labeled and unlabeled data. Supervised learning relies on labeled samples to train a good classifier. However, the time-consuming nature, along with expert guidance of data labeling makes it difficult to acquire enough labeled data. These tend to hinder the applicability of supervised learning models in real-world scenarios [24]. A way to deal with this is using a semi-supervised learning paradigm that uses unlabeled data [25]–[28]. This methodology necessarily does not need all samples to be labeled. It uses an amount of unlabeled data, together with the labeled data, to build better models that require less human effort and yields high performance [26]. Self-training is a semi-supervised learning scheme that iteratively, enlarges the labeled training sample set [29]. Initially, a model is trained with a set of labeled data samples, followed by prediction on the unlabeled data and then a selection of the unlabeled data with high confidence to be incrementally appended to the labeled training data with their predicted labels. It is a technique that leverages a supervised model to generate pseudo-labels for unlabeled data samples and add the samples that are selected with the highest confidence to the training data together with their generated pseudo-labels, thereby enlarging the training data size. This procedure is repeated until the model converges. And it is a method that has been implemented and used in several [3]–[5], [29]–[32] studies and applications. The classifier uses its predictions, which are the pseudo-labels generated for unlabeled data to teach itself. Typically, the highest-confident unlabeled points, together with their predicted labels are the ones selected. In [32], a method was proposed to use a self-paced learning model that learns to detect objects from images and adapt those objects to videos by learning labeled source samples and target data with pseudo-labels in an easy-to-hard manner. Xuanyi Dong et al. also proposed few-example object detection [33] scheme. With a limited number of annotated data they trained models that would go on to iteratively learn and detect objects in images by utilizing the self-paced learning technique to solve object detection problem. Their model was able to learn the discriminative features and reliably select samples from the large pool of unlabeled samples. The self-paced learning algorithm proposed by Xuanyi was further improved in [34] to detect facial landmarks by learning from partially labeled samples. In that work, the authors improved the self-paced learning with meta-learning. A detector was trained on a set of labeled images to generate new training samples using this detector's prediction as pseudo-labels of unlabeled images and retrain the detector on the labeled samples and partial pseudo-labeled samples. An easy-to-train interaction mechanism between teacher and students to provide more reliable pseudo-labeled samples was proposed. The teacher network judges the quality of pseudo-labels generated from students network and give a feedback to the students by selecting a qualified pseudo-labeled samples to retrain the students network for them to become more robust. Another study [35] proposed the use of constraints to label a pool of unlabeled data and then use that newly generated labeled data to enlarge and update the model to bridge the gap between source and target domains, by slowly adding them to the training set from both the target features and instances in which the model is highly confident. The method was modified to suit a sentimental classification task in a different research [36]. In that work, the authors proposed a technique that uses a high precision classifier that is made up of linguistic rules to influence the selection of training candidates, which are artificially labeled by the base learner in an iterative self-training process. The linguistic knowledge encoded in the classifier is highly precise, that, it does not just select the high-confidence training samples but also in the pre-processing stage, corrects the high-confidence errors made by the base-classifier based on the work proposed in [35]. We propose a variant self-training that differs from the conventional self-training method in the way the selection of examples are appended in a single iteration with the learner not constrained. Self-training has been very efficient, therefore in this study; we adopt a CNN learner as the base-learner and a novel selection algorithm that prevents mistake reinforcement; reinforcing wrong predictions to enlarge set along the training process, which is a common thing in conventional self-training models or Expectation-Maximization algorithms.

III. SELF-TRAINING FOR FINE-GRAINED CLASSIFICATION

Fine-grained datasets in computer vision are way smaller than general-object large visual classification datasets and possess a great deal of imbalance across classes. Moreover, the samples of a class are not the accurate representation of the characteristic differences in the visual class itself. Also, there is over-fitting when deep neural networks are trained with a smaller set of data even when either trained with huge parameters or preliminary layers frozen. Besides, the training samples are mostly not the complete representation of the real-world data, with some classes having abundant samples than others. Challenges such as these impede the application of fine-grained classification tasks in reality. Classes with fewer samples may not be well represented during training, sometimes forcing the neural network to latch onto sample-specific artifacts in the image, instead of learning the versatile artistic features for the target image. A pleasing strategy is to follow

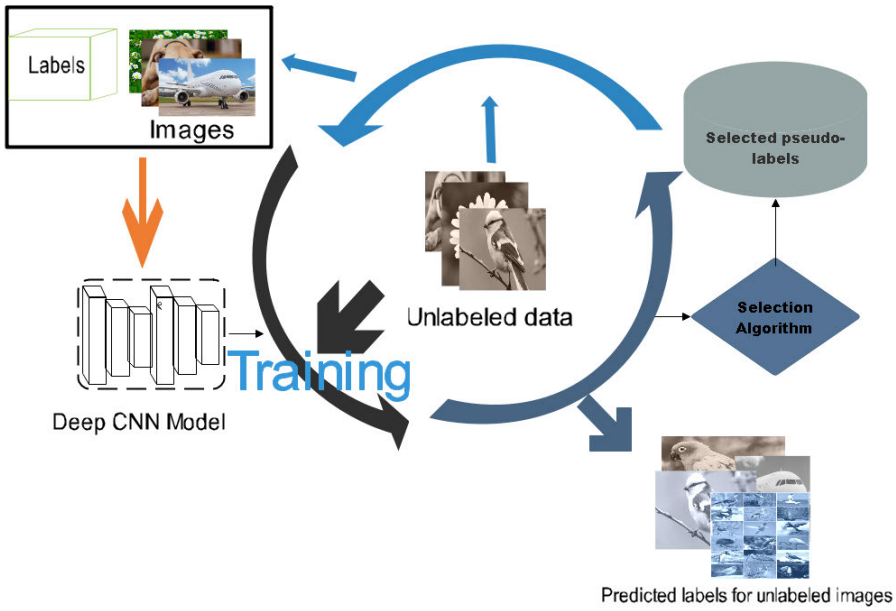

FIGURE 1. Illustration of semi-supervised self-training framework algorithm workflow: A deep-CNN classifier is initially trained on a set of labeled images, generate new pseudo-labeled training samples from unlabeled images, select most confident pseudo-labeled samples via selection algorithm and retrain the classifier. The premium pseudo - labeled data along with real - labeled data are used for the retraining process.

an 'easy-to-hard' design via self-paced curriculum learning, generating pseudo-labels from the most confident predictions with a hope that they are mostly correct, ensuring that, the model is updated and better adapted to the test domain, with a further examination of the less-confident predictions. A new pointillistic method, called Semi-Supervised Learning for Fine-Grained Classification (SSLGFC) via self-training, which integrates Semi-Supervised Learning(SSL) and Fine-Tuning CNNs, is proposed considering the three characteristics of fine-grained visual classification, i.e., the high intra-class similarities, the small sample size, and the existence of unlabeled data. Looking at the workflow of the proposed system illustrated in Fig. 1, a DNN is first trained with images fed into it and then a prediction is done on unlabeled data by the model after training, generating approximate labels known as pseudo-labels for the unlabeled samples. A selection algorithm is then run to select the unlabeled samples that have the highest-confidence probability prediction together with their approximated labels to be added to the training set, this cycle is iterated over for a number of times. Compared to other visual classification approaches, the main merit of SSLGFC is in two aspects: (1) Self-training is introduced to use a self-paced ''easy-to-hard'' way to avoid reinforcing wrong predictions to enlarge the training set during the training process. In short, this is to avoid mistake reinforcement, and (2) an enhanced semi-supervised self-training classification algorithm that can effectively classify fine-grained samples by using a small labeled sample size and utilize its effectiveness for both labeled and unlabeled data. To our best of knowledge, this work is the first to study this classification approach in the fine-grained scenario.

A. PRELIMINARIES

Given the labels for few images in the same task for both source and target, the most direct way to improve classification and generalization is supervised fine-tuning models on both domains. Following a setting with n classes, the desired classification objective is defined as a standard softmax loss on the labeled source data as inputs x_s, y_s and the target data x_t, y_t.

$$L_c(\chi, y : \theta_c)_W = -\sum_k 1[y = k] \log P_k. \tag{1}$$

In Eq. 1, the goal is to produce a classifier θ_c that can correctly classify target samples at the time of testing. However, based on the assumption that, access to a limited amount of labeled target data, potentially from only a subset of the categories-of-interest, the transfer of representations using Fine-Tuning becomes inefficient. Therefore, we propose to use a semi-supervised model with softmax output by formulating the problem as minimizing the loss function:

$$\begin{aligned} minL_{st}(W)_W = &-\sum_{s=1}^{S}\sum_{n=1}^{N} \mathcal{Y}_{s,n}^T \log\left(P_n(W, I_s)\right) \\ &-\sum_{s=1}^{T}\sum_{n=1}^{N} \mathcal{Y}_{t,n}^T \log\left(P_n(W, I_t)\right). \end{aligned} \tag{2}$$

where $\boldsymbol{I_s}$ denotes the image in source domain indexed by $\boldsymbol{s} = 1, 2, \ldots, \boldsymbol{S}$. $\mathcal{Y}_{s,n}$ the true labels for the $\boldsymbol{nth}$ image ($n = 1, 2, \ldots, N$) for $\boldsymbol{I_s}$, and $\boldsymbol{W}$ contains the network weights. $\boldsymbol{P_n(w, I_s)}$ is the softmax output containing the class probabilities. Such similar definitions goes for $I_t, \mathcal{Y}_{t,n}$ and $p_n(w, I_t)$

at evaluation time.

$$minL_{st}\left(W,\hat{\mathcal{Y}}\right)_{W,\hat{\mathcal{Y}}} = -\sum_{s=1}^{S}\sum_{n=1}^{N}\mathcal{Y}_{s,n}^{T}\log\left(P_n\left(W,I_s\right)\right) - \sum_{s=1}^{T}\sum_{n=1}^{N}\hat{\mathcal{Y}}_{t,n}^{T}\log\left(P_n(W,I_t)\right). \quad (3)$$

The problem can further be formulated to minimize the loss function in Eq. 3. Given a situation where some of the target labels are unavailable, the model considers the labels to be hidden and learns from an approximate target labels $\hat{\mathcal{Y}}$ for $\hat{\mathcal{C}}$ representing the number of classes. We refer to $\hat{\mathcal{Y}}$ as pseudo-labels which shall be used to train the model again iteratively and such training strategy is known as Self-training.

B. SELF-TRAINING FOR CLASSIFICATION WITH SELF-PACED LEARNING

A paradigm of self-training that jointly learns a model and optimizes pseudo-labels on a set of unlabeled data is difficult as there is a possibility to incorrectly generate pseudo-labels which approximates the true ground labels. In order to avoid reinforcing wrong predictions into the training set, an 'easy-to-hard' self-training approach would be used. An 'easy-to-hard' self-training approach generates pseudo-labels from the most confident and correct predictions, updates the model and better generalize on classifying unlabeled target data. The approach then later explores the less confident pseudo-labels that are remaining, and including this in the scheme, we modify Eq. 3 to:

$$minL_{st}\left(W,\hat{\mathcal{Y}}\right)_{W,\hat{\mathcal{Y}}} = -\sum_{s=1}^{S}\sum_{n=1}^{N}\mathcal{Y}_{s,n}^{T}\log\left(P_n\left(W,I_s\right)\right) - \sum_{s=1}^{T}\sum_{n=1}^{N}\left[\int_{1}^{2}\hat{\mathcal{Y}}_{t,n}^{T}\log\left(P_n\left(W,I_t\right)\right) + k\left|\hat{\mathcal{Y}}_{t,n}^{T}\right|_{1}\right]. \quad (4)$$

In the modified formulated loss in Eq. 4, $\mathcal{Y}$ is assigned 0 when the pseudo-label $\hat{\mathcal{Y}}$ is ignored during the model training phase. L_1 regularizer is added to the loss function to serve as a negation to prevent the case of ignoring a large number of pseudo-labels. The factor $k > 0$ ensures that more pseudo-labels are selected for training the model. Meaning a larger k ensures that a large number of pseudo-labels are selected. Minimizing the Loss in Eq. 4 uses the two coordinating steps below:

a) Initialize W and minimize the loss in Eq. 4 with respect to $\hat{\mathcal{Y}}_{t,n}$.
b) Set $\hat{\mathcal{Y}}_{t,n}$ and optimize the objective function in Eq. 4 with respect to W.

Executing step a) followed by step b) is considered to be a single **iteration**. As mentioned early on, we propose a semi-supervised method, a self-training algorithm where step a) and step b) are executed in succession and repeated for multiple rounds. Meaning, in step a), a portion of the most confident pseudo-labels are selected while step b) does the training of the network model when the pseudo-labels have been selected in step a). The algorithm flow in the self-training for fine-grained classification framework is depicted in Fig. 1. Step b) results in network learning with an optimizer. But step a), given the optimization over discrete variables, needs a nonlinear function. So step a) can be reformulated to Eq. 5, given $k > 0$.

$$min_{\hat{\mathcal{Y}}} -\sum_{t=1}^{T}\sum_{n=1}^{N}\left[\sum\hat{\mathcal{Y}}_{t,y}^{(c)}\log\left(p_n(c|w,I_t)\right) + k|\hat{\mathcal{Y}}_{t,n}|_1\right]. \quad s.t.\ k > 0 \quad (5)$$

Because $\hat{\mathcal{Y}}_{t,n}$ is either needed to be a discrete one-hot vector or a vector with a null magnitude, the pseudo-label framework is optimized by way of using the solver in Eq. 6.

$$\hat{\mathcal{Y}}_{t,y}^{(c*)} = \begin{cases} 1, & \text{if } c = \arg\max p_n\left(c|w,I_t\right), \\ & p_n\left(c|w,I_t\right) > \exp(-k). \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Compared with conventional self-training methods that are able to learn domain-invariant classifiers, CNN based self-training methods can learn both domain-invariant classifiers and domain-invariant features. Intuitively, CNN based self-training can better learn and alleviate the intra-class variations challenge that faces deep learning models in classifying fine-grained images. Therefore, the softmax loss in Eq. 6 tries to reduce the domain variations in feature space. It can also make the models learn features and weights with no prior observation of unlabeled samples to solve the missing value (pseudo-label) problem that befalls both traditional self-training and Expectation-Maximization (EM) methods. From Eq. 6, it can be seen that the generation of pseudo-labels in Eq. 5 hinges on the output $(p_n(c|w,I_t))$. Assignment of pseudo-labels is done using the output. Because, self-training generates pseudo-labels that corresponds to large confidence, a challenge that prevails most often is, models tend to be biased toward classes with large-sized samples that are, initially transferred well and do away with less-sized classes during the training process. Thus self-training algorithms do not perform well in multi-class classification problems. To gain mastery over this issue, $k|\hat{\mathcal{Y}}_{t,n}|$ is introduced in Eq. 5, to determine the proportion of pseudo-labels $\hat{\mathcal{Y}}_{t,n}$ that would be selected from each class and to assign a pseudo-label to a sample and one-hot encode, the output probability $(p_n(c|w,I_t))$ in the solver Eq. 6 doesn't have to be less than $\exp(-k)$ otherwise it is assigned a zero-vector and ignored. It may be noted that, the proposed configuration of the semi-supervised self-training algorithm is similar to [31], [32] and many other related works. However, the suggested framework presents a generalized model for self-training, with much focus on the generation and selection of pseudo-labels with self-paced learning using a curriculum learning method. The generation of pseudo-labels is coupled with curriculum learning under a single unified learning

Algorithm 1 Algorithm for Determining k In

input : Deep Learning Network $P(w)$, unlabeled Images I_t, selected pseudo-labels p
output: k
for $t \leftarrow 1$ **to** T **do**
  $P_{I_t} = P(w, I_t)$;
  $MP_{I_t} = \max(P_{I_t}, axis = 0)$;
  $M = \left[M, from - matrix - to - vector(MP_{I_t})\right]$
end
$M = sort(M, order = descending)$ $L = length(M) \times p$
$k = -\log(M[L])$;
$return(k)$

framework. More significantly, in terms of the specific application of classifying fine-grained images, the above self-training structure throws more light on a relatively new direction for classification models. The proposed model highlights on a different way of classifying unlabeled images from a small amount of well-labeled images in the fine-grained domain with an extension to coarse-grained domain samples.

C. FIXING AND FINDING (k) IN SELF TRAINING ALGORITHM

k is a crucial determinant that decides the number of pseudo-labels to be added to the training set after each iteration phase. It filters out the pseudo-labels with their probabilities less than k. We set k by taking the maximum probability on each sample, sort these probabilities across all samples and all classes in a descending order. k is then set, so that $\exp(-k)$ would equal the ranked probability round at $(p * T * N)$, where p is a portion number between [0, 1]. Optimizing the pseudo-labels in this case produces $p \times 100\%$ confident pseudo-labels for training. Algorithm 1 gives the concise determination of k in the proposed framework. It is designed to allow more pseudo-labels to be added to the training set for each additional round. To be specific, p starts from 10% of the most confident predictions, and at each additional round, the top 5% is included in the next iteration of the pseudo-label generation process. The maximum limit of p is set to 50%. From the algorithm, M is the maximum probability output on each sample, sort such probabilities across samples and classes.

IV. EXPERIMENTS

In this section, a comprehensive evaluation of the proposed method by performing experiments on benchmark datasets, are provided. We explore each unit of SSLGFC on four publicly available fine-grained visual classification (FGVC) datasets. We conduct additional experiments on some selected coarse-grained classification datasets (Natural Images [37], caltech-101 [38] and food101 [39]) to demonstrate the effectiveness of the semi-supervised self-training method in a limited labeled data setting.

TABLE 1. A summary information of four state-of-the-art fine-grained datasets.

Data Set	Description	Classes #	Training #	Testing #
Stanford Cars	Cars	196	6,667	3,333
CUB-200-2011	Birds	200	5,994	5,794
Flower-102	Flower species	200	5,994	5,794
Stanford Dogs	Dogs	120	12,000	8,580

A. DATASETS AND EXPERIMENTAL SETUP

1) DATASETS

We experimented the method on four state-of-the-art FGVC, which are CUB-200-2011 [40], Stanford Dogs [41], Stanford cars [42], and 102 oxford flower [43]. The complexity of these datasets is high due to the uncountable similarities existing among classes, occlusion, high intra-class variations, imbalance set, just to mention a few. A summary of the specifics on these datasets have been provided in Table 1.

2) STANFORD DOGS

A set of images containing breeds of dogs across the globe. It is a subset of the Imagenet [44] annotations that are prepared for the task of fine-grained visual classification. The annotations have class-labels as well as bounding-boxes. However, for the sake of semi-supervised classification purposes the bounding boxes are dropped. The test set consisting of 8580 samples was used as the unlabeled data in this experiment.

3) STANFORD CARS

The Cars dataset contains 16,185 images of 196 classes of cars. The data is split into 8,144 training images and 8,041 testing images, where each class at the level of the model, make, year, etc. has been split approximately 50-50 percentage-wise. Following the supervised training procedure, the training set with its annotation was used to train the base-learner. It was split into a ratio of 70% for training and 30% for validation. The test set was used for the semi-supervised learning phase where the annotations were dropped thereby using the test as unlabeled data.

4) CUB-200-2011

A challenging set of 200 bird species put together for fine-grained classification purpose. It contains 11,788 images of birds species with 5,994 training samples and 5,794 images in the test set. The experiment had two phases, with the first being the supervised learning phase where the base-learner is trained with the training set, including their true labels, and the second phase being the semi-supervised learning stage. In the second phase, the test set without its annotation was used to train the classifier.

5) FLOWER-102

A set of 102 category dataset, consisting of 102 flower categories which are chosen to be flowers frequently

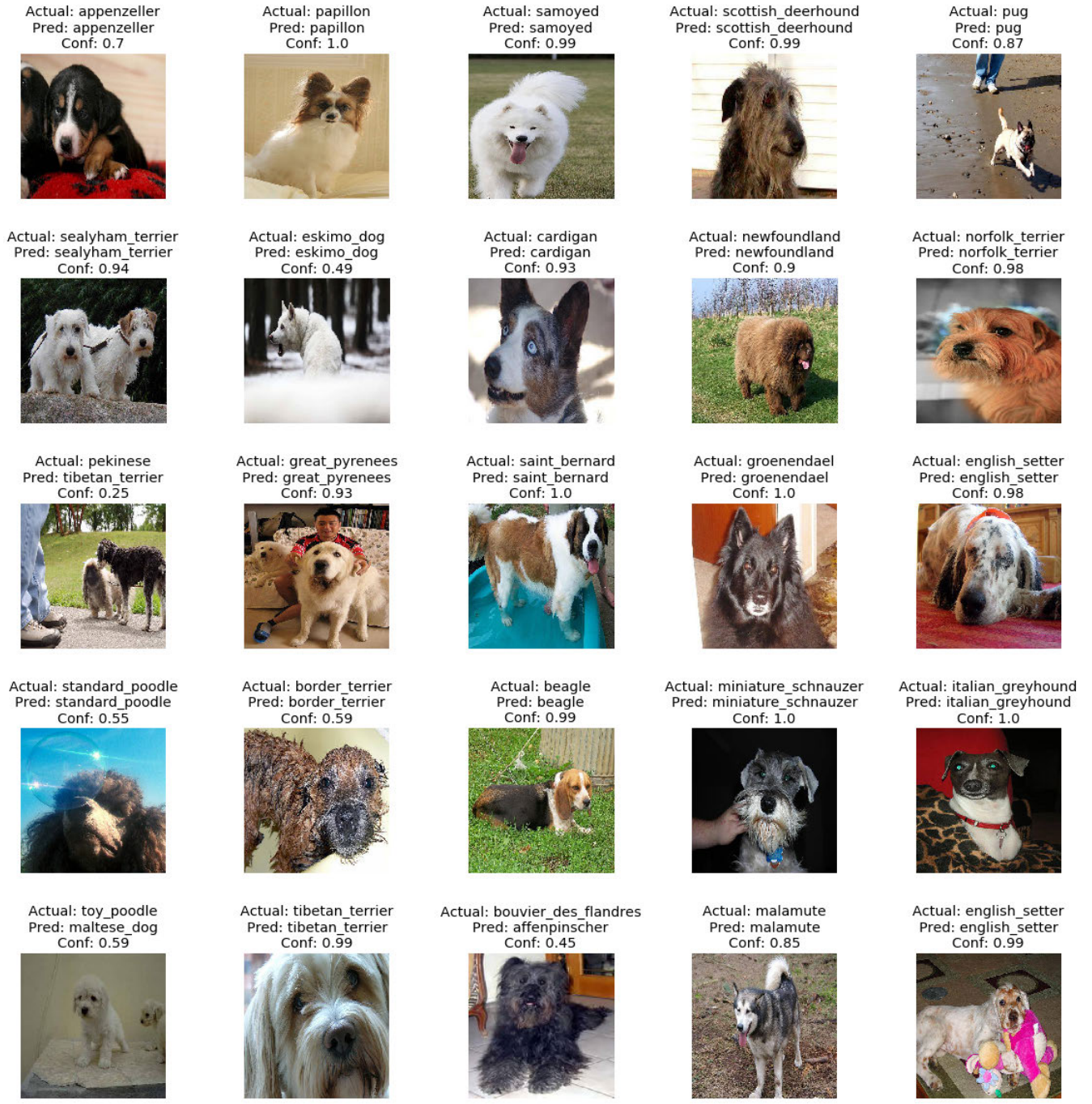

FIGURE 2. **Prediction on Stanford Dogs using Top-10% ST approach: In the figure, the true class and the predicted class together with their probabilities have been given. We observed that most of the images were predicted correctly with just three out of 25 wrongly predicted. And the prediction probabilities as depicted in figure, are high.**

occuring in the United Kingdom. Each class consists of between 40 and 258 images. It is a challenging set, due to both small inter-class and large intra-class variances and the images have large scale, pose and light variations. It consists of 8,189 samples that have been divided into a training set, validation set and a test set. The number of samples for the training set is 1,020, and the validation set just as the training set contains 1,020. The test set has 6,149 samples. And in this experiment setup, the test set is used as the unlabeled data for the self-training part.

Further experiments were run on coarse-grain datasets, to evaluate the method. The summary of which has been divulged in Table 2. Due to hardware limitations, only 21,000 samples of the food101 dataset were experimented on, with 93 categories having 208 and the remaining 8 being made of 207 samples. Also, the coarse-grained datasets had

TABLE 2. A summary information of some selected coarse-grain datasets.

Data Set	Description	Classes #	Samples #
Natural-Images	natural objects	10	6,899
caltech-101	pictures of objects	101	8,677
food101	food categories	101	21,000

TABLE 3. Self-training(ST) approach performance on Stanford Dogs. Comparisons of accuracies on the Stanford dogs dataset by the various self training approaches and our proposed method. Baseline (fine-tuning) indicates that, the pre-trained model was fine-tuned with only labeled samples. The ratio number in the brackets represents the portion of the most confident pseudo-labels that we use. Compared to supervised algorithm which use 100% labels, ST obtained higher accuracy and improved significantly when a portion of the generated pseudo-labeled samples was used.

ST-approach	Accuracy(%)
Baseline (fine-tuning)	85.70
All pseudo labels	88.43
K(top 5) pseudo-labels	90.74
K(top 10) pseudo-labels	**91.67**
K(top 20) pseudo-labels	89.89
model-weight	90.28

no test-set which would go on to be used as unlabeled set to suit the scenario, so we divided the entire set into approximately two equal ratios with one part used for supervised training and the remaining for the semi-supervised setting.

B. IMPLEMENTATION DETAILS

In the experiments, the pre-trained Inception_ResNetV2 [45] was chosen as the baseline model of the proposed protocol. Inception_ResNetV2 is a variation of Inception_V3 [46] model which borrows some ideas from Microsoft's ResNet [47]. It is able to significantly improve recognition performance of objects at a relatively low computational cost. Fine-tuning of pre-trained networks using Imagenet has been evaluated by various previous studies and it has been shown to be among the best techniques for deep CNNs to gain improved performance when applied to small scale data. And in this scenario, where the problem is a fine-grained classification task with limited amount of data, fine-tuning technique is implemented at the fully supervised learning phase to initialize the model weights and also reduce the variance. Data augmentation of random rotation, vertical flips, and zooming techniques were performed to regularize our model during training. We trained the model using an Adam optimizer [48] with $\beta_1 = 0.9\beta_2 = 0.999$, 50 number of epochs, a mini-batch of 32, and an initial learning rate of 1e-3 that decays upon countering a plateau in the learning process. A single NVIDIA GTX1080Ti GPU was used to run the experiments. We retrain our SSLFGC model with hyper-parameters for top k using 5%, 10% and 20% pseudo-labeled samples of the unlabeled data, in all our experiments for simplicity. According to Eq. 4, the factor k ensures more higher-confident pseudo-labeled samples are selected for the retraining process. It is a critical and

TABLE 4. Self-training(ST) approach accuracy on Fine-grained datasets: The ratio of the most confident pseudo-labels that we used are put in brackets. Compared to the supervised algorithms which use 100% labels, the conventional way of self training by using 100% of the pseudo-labeled samples and the model weight, the accuracy was bettered in cases where the top k pseudo-labeled samples was used.

ST-approach	Birds(%)	Cars(%)	flowers(%)	Dogs(%)
Baseline (fine-tuning)	76.73	84.44	93	85.70
All pseudo labels	78.06	88.52	93.83	88.43
K(top 5) pseudo-labels	**83.31**	91.27	94.83	90.74
K(top 10) pseudo-labels	82.72	**93.71**	**96.72**	**91.67**
K(top 20) pseudo-labels	81.20	93.11	95.38	89.89
model-weight	80.65	93.65	95.69	90.28

TABLE 5. Comparison with related works on Stanford Dogs: Comparisons of accuracies with related works and our proposed method. Baseline (fine-tuning) indicates that, the pre-trained model was fine-tuned with only labeled samples. The ratio number in the brackets represents the portion k = 5%, 10 *and* 20% respectively of the most confident pseudo-labels that were used. Compared to the supervised algorithms which use 100% labels, ST obtained higher accuracy and improved significantly when a portion of the generated pseudo-labeled samples was used. We observed The top 20% did not obtained a high accuracy and was just a percent more than FCAN.

Method	Accuracy(%)
PC [49]	83.75
FCAN [20]	88.9
MAMC [50]	85.2
Baseline (fine-tuning)	85.70
SSLFGC(top 5)	90.74
SSLFGC(top 10)	**91.67**
SSLFGC(top 20)	89.89

a crucial factor so by setting k to 5%, 10% and 20% means the k-maximum probabilities across all samples and classes that has been regularized with the L_1 regularizer are used in the *Tth* iteration. In the implementation of our experiments, suppose that the number of selected images for the *cth* class is p then $k = -\log(M[L])$ as detailed in Algorithm 1.

C. ACCURACY CONTRIBUTION AND COMPARISON

Our Self-training method was deployed using three different approaches. 1. Using all the generated pseudo-labels for the unlabeled set; 2. Setting a threshold (K) to use the top-5%, top-10% and top-20% confident pseudo-labels; 3. And using the weight of the trained model. However, as described above, the second approach was investigated deeply, although accuracies are reported on each aprroach. We experimented on the fine-grained dataset and demonstrated that each approach uniquely improves the classification accuracy as shown in Table 4. There was a significant improved performance by the SSLFGC model on all the datasets. A further fully-supervised experiment was conducted using a fine-tuning approach and the performance was compared with the proposed model. It can be seen from Table 5, Table 6, Table 7, Table 8, and Table 9 that, the SSLFGC method performed better with regards to obtaining a higher accuracy than the supervised learning approach of using and fine-tuning the Imagenet pre-trained model.

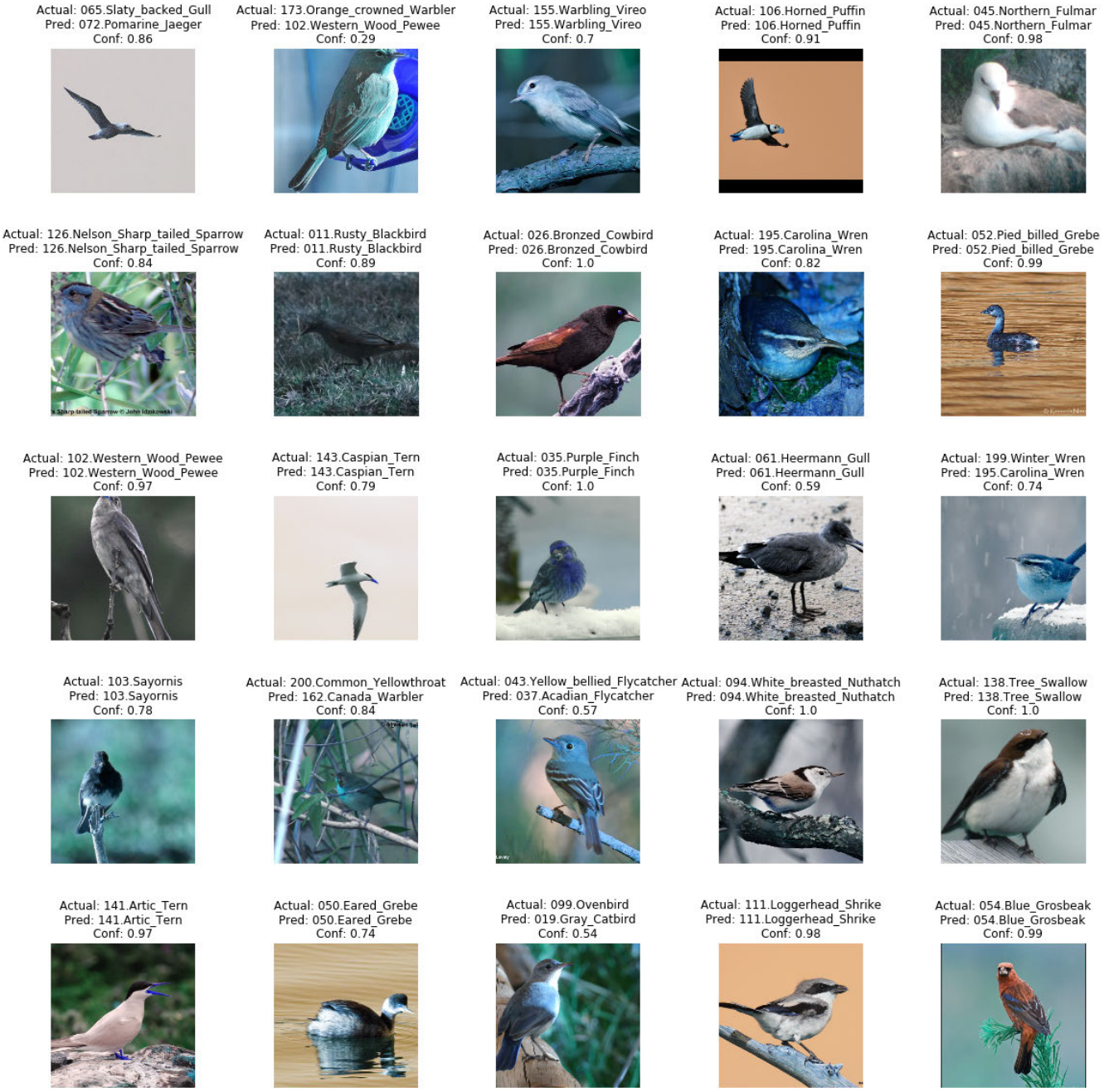

FIGURE 3. SSLFGC Prediction on CUB-200-2011 using Top-10%: The model struggled to obtain the same similar performance it had achieved on the other datasets. Yet the model predicted the images with very high confidence although 6 samples were wrongly classified.

D. COMPARISON WITH RELATED WORKS

1) FINE-GRAINED RESULTS

A comparison of our semi-supervised method with some related works on fine-grained classification tasks. We experimented on Stanford-dogs and demonstrated how each approach of the model impacts the classification accuracy which is provided in Table 3. It can be seen that, in the first approach where all the generated pseudo-labels for the unlabeled set were used in the self-training process, although it obtained a performance higher than the supervised learning approach, the model could not achieve the accuracy that was achieved by the other two self-training approaches. Both the top-k and the weight usage approaches significantly obtained higher accuracies on all the datasets. For instance, on the CUB-200-2011 dataset, almost all the approaches had a setback to obtain a higher accuracy but such was not the case for the top-5%. A decent accuracy was obtained and it goes to confirm the point that, given the most-confident generated pseudo-labels, the model will achieve a significantly improved accuracy. A visualization of predictions

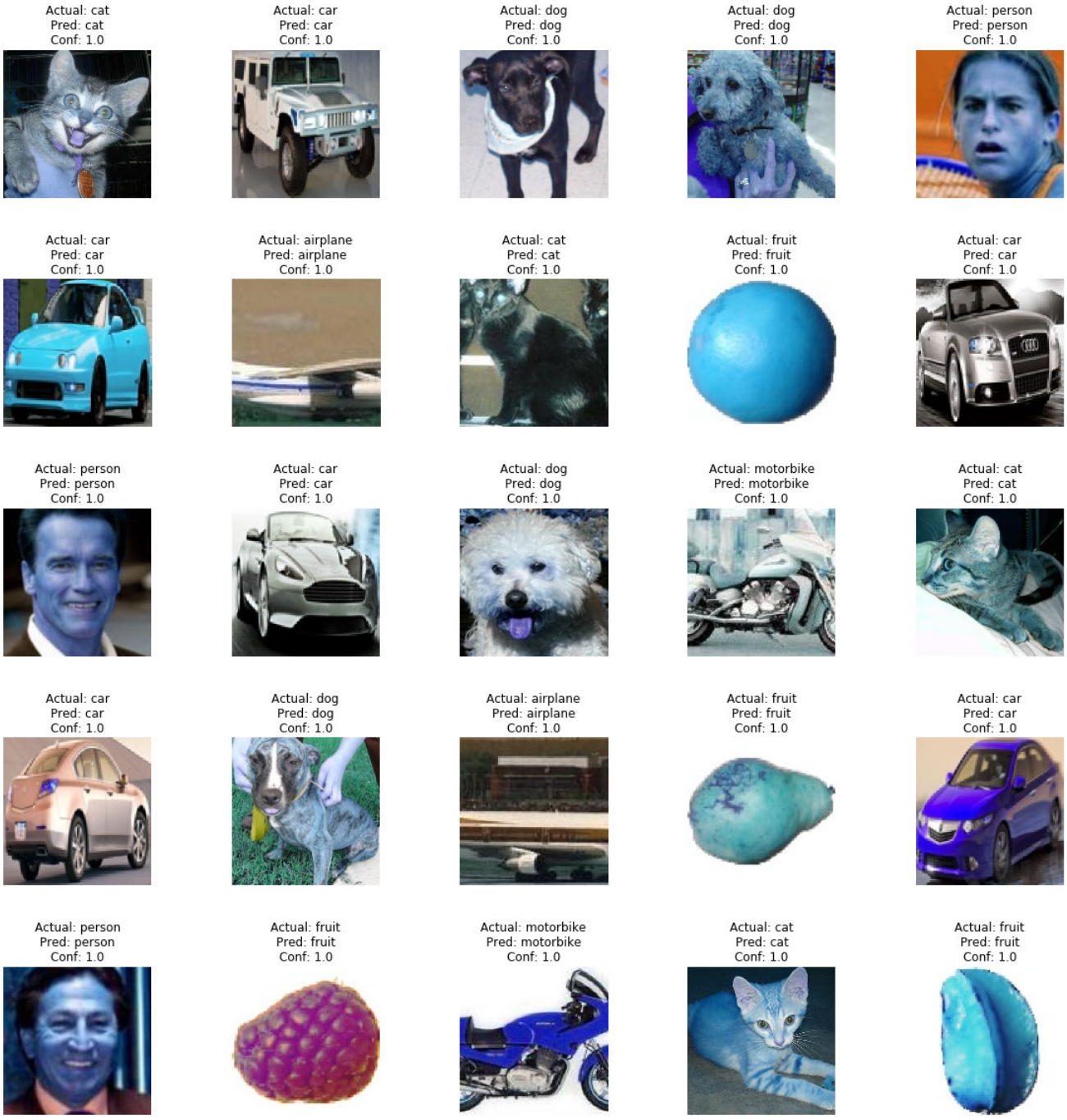

FIGURE 4. **Prediction on natural images using Top-10% approach: The model achieved a perfect performance of being able to classify samples drawn from the natural dataset. Given that there were so many discriminative features existing among the samples, the classifier had a 100% prediction accuracy.**

on the validation set proved the capability of the top-10% approach as provided in Fig. 2. In Fig. 2, the *top* − 10% approach had three predictions out of twenty-five wrongly classified, and not only did the model predict the correct dog in most cases but did that with very high confidence. However, we realized the high accuracy obtained by the model on the Stanford dogs, Stanford cars, and 102 Oxford-flowers was not the case when the model was experimented on the CUB-200-2011 birds dataset as provided in Table 7. The model managed to obtain a decent performance accuracy when compared with the other datasets although the model choked. We visualized to investigate the prediction confidence and further observed that, although the performance accuracy was not so high as compared

TABLE 6. Comparison with related works on Stanford Cars: Comparisons of accuracies on the Stanford cars by some supervised approaches and our proposed method. In the supervised algorithm which use 100% labels, the SSLGFC algorithm with top-10% bettered the accuracy with a slight margin when compared to PC. The top 5% however did not do so well and we observed that, utilizing just 5% of the most confident pseudo-labeled samples was not enough to meet the hunger of deep CNNs.

Method	Accuracy(%)
PC [49]	93.43
FCAN [20]	91.5
MAMC [50]	93.0
Baseline (fine-tuning)	89.0
SSLFGC(top 5)	91.27
SSLFGC(top 10)	**93.71**
SSLFGC(top 20)	93.11

to the other datasets, most of the predictions were right, and that was done with high confidence too. The various results have been provided in Table 5, Table 6, Table 7, and Table 8 respectively on the various fine-grained datasets. From the tables, we give the performance comparisons between SSLFGC and some works on the various datasets. Pairwise confusion(PC) [49], Fully Convolutional Attention Localization Networks(FCAN) [20] and Multi-attention multi-class constraint(MAMC) [50] have been recognized as powerful supervised classification methods for dealing with small sample size and fine-grained classification tasks. However, we found that the semi-supervised learning method proposed, i.e., SSLFGC performed significantly better than the supervised algorithms with improved accuracy of 91.67% for the dogs, 92.30% for cars, 96.72% for flowers and 82.72% for CUB-200-2011 respectively. When compared, the performance was not so good on the CUB-200-2011 birds' dataset, and a possible reason is that the several similarities present, the background, and the small sample size for the various classes affected the model's performance. Aside from that, SSLFGC effectively utilized the information buried in unlabeled samples and thus achieved a better classification performance than other models. One intriguing thing is the consistency at which the top-K approach obtained improved accuracies than other methods and approaches. It goes on to say that the proposed selection algorithm which was integrated as part of the self-training process was helpful and even in a case where the model struggled, it could still manage a decent result.

2) COARSE-GRAIN RESULTS

We evaluated the method on coarse-grain datasets and compared the results (see Table 9). among the three ways through which self-training is implemented. The method obtained high performance in terms of accuracy, largely; however, the performance in the approach, where the determinant k; here, the top-5%, top10% and top-20% were used achieved higher accuracy than the rest of the approaches. There was not so much a big difference between the accuracy obtained by the top-5% and the top-10% when compared to the other

TABLE 7. Comparison with related works on CUB-200-2011: The supervised related works chalked better performance than our SSLFGC algorithm. Although utilizing the top 5% pseudo-labeled samples for the semi-supervised learning phase, the top 5% did not achieve the accuracy it obtain on CUB-200-2011 on the other three datasets. It obtained a decent accuracy and fall short by a margin of almost 4% when compared to PC which obtained 86.87%. We observed that when it comes to classifying birds, it is not only about quantity but rather the discriminative features can help greatly.

Method	Accuracy(%)
PC [49]	**86.87**
FCAN [20]	84.3
MAMC [50]	86.5
Baseline (fine-tuning)	76.63
SSLFGC(top 5)	83.31
SSLFGC(top 10)	82.72
SSLFGC(top 20)	81.20

TABLE 8. Comparison with related works on Oxford 102-flowers: 100% labeled samples was used for the supervised fine-tuning,and even with that SSLGFC algorithm with top-10% bettered the accuracy with a big margin when compared to related works.

8 Method	Accuracy(%)
PC [49]	93.65
Efficient object detection and segmentation [51]	80.66
Flower species recognition system [52]	93.41
Baseline (fine-tuning)	94.3
SSLFGC(top 5)	94.83
SSLFGC(top 10)	**96.72**
SSLFGC(top 20)	95.38

TABLE 9. Results of the approach on coarse-grain datasets: Performance of the model using the three different self-training approaches on coarse-grained. The supervised learning method is seen back to its best by achieving a high classification accuracy. In the supervised algorithm which use 100% labels, the SSLGFC algorithm with top-10% bettered the accuracies obtained by the supervised learning. On the Caltech-101, top 5% obtained the same accuracy just as the top 10%. On the food dataset, the performance is decent but not great.

Method	Caltech-101	Food-101	Natural Images
Baseline (fine-tuning)	94.33	87.92	95.26
All pseudo labels	95.14	81.41	97.28
K(top 5) pseudo-labels	**97.68**	83.11	99.32
K(top 10) pseudo-labels	**97.68**	**84.0**	**99.74**
K(top 20) pseudo-labels	96.33	81.87	99.01
model-weight	95.94	82.57	99.72

results. Although, there were some misclassified samples in the fine-grained datasets, such was not the case in the coarse-grained sets as depicted Fig. 4, and we can attribute it to the fact that the vast variance present in coarse-grained sets hugely contributed to their 100% correct predictions. Also the impressive results obtained for the coarse-grained datasets go to confirm how challenging and difficult fine-grained classification tasks are.

V. CONCLUSION

The high intra-class similarities, small sample size, and enrichment of unlabeled samples of fine-grained data represent significant obstacles for learning approaches to

Fine-grained visual classifications. To overcome these difficulties, we proposed a novel semi-supervised self-training method that uses self-paced learning, called SSLFGC, for visual classification using fine-grained data. Self-paced learning is introduced to alleviate reinforcing wrongly generated pseudo-labels for unlabeled samples to enlarge the training set. To utilize the information from the unlabeled data, the self-training technique was applied in SSLFGC. However, the traditional self-training method is prone to reinforcing model mistakes, termed as mistake-reinforcement. In light of this, a new and efficient sample selection procedure was developed to alleviate the mistake-reinforcement problem of conventional self-training methods. The results of experiments have demonstrated the capability of utilizing unlabeled data. The performances of the SSLFGC are further compared with some state-of-the-art classification methods on four separate benchmark fine-grained datasets and three other selected coarse-grained data, and the SSLFGC obtained higher performances when compared with other techniques. Not only did the SSLFGC performed so well, but could predict classes of samples with very confidence. For instance, the birds dataset that SSLFGC struggled to obtain a high accuracy on was classified with high confidence, which goes to say the proposed selection scheme is highly efficient and able to mitigate mistake reinforcement. Despite the encouraging performance of SSLFGC, as our future work, we are investigating the benefits of neural network architecture search. We will also come out with an efficient semi-supervised algorithm that will be able to balance class sample size to further improve the classification accuracy on tiny images.

ACKNOWLEDGMENT

(Obed Tettey Nartey and Sarpong Kwadwo Asare contributed equally to this work.) The authors would like to thank the anonymous reviewers for their careful reading of this article and for their helpful and constructive comments.

REFERENCES

[1] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 580–587.

[2] E. Riloff, J. Wiebe, and W. Phillips, "Exploiting subjectivity classification to improve information extraction," in *Proc. 20th Conf. Artif. Intell. (AAAI)*, 2005.

[3] B. Wang, B. Spencer, C. X. Ling, and H. Zhang, "Semi-supervised self-training for sentence subjectivity classification," in *Advances in Artificial Intelligence*, S. Bergler, Ed. Berlin, Germany: Springer, 2008, pp. 344–355.

[4] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. 33rd Ann. Meeting Assoc. Comput. Linguistics*, 1995, pp. 189–196.

[5] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Proc. 7th IEEE Workshop Appl. Comput. Vis.*, Jan. 2005, pp. 29–36.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR) (Oral)*, 2015.

[10] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Proc. IEEE Winter Conf. Workshop Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 924–933.

[11] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *CoRR*, May 2012.

[12] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, vol. 1, no. 2, p. 3.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR) (Banff)*, Dec. 2013.

[14] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012. [Online]. Available: https://arxiv.org/abs/1207.0580

[15] Y. Lin, S. Zhu, and A. Angelova, "Image segmentation for large-scale subcategory flower recognition," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Washington, DC, USA, Jan. 2013, pp. 39–45, doi: 10.1109/WACV.2013.6474997.

[16] E. Gavves, B. Fernando, C. Snoek, A. Smeulders, and T. Tuytelaars, "Fine-grained categorization by alignments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2013, pp. 1713–1720.

[17] S. Branson, G. V. Horn, S. J. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Nottingham, U.K., Sep. 2014. [Online]. Available: http://vision.cornell.edu/se3/wp-content/uploads/2015/02/BMVC14.pdf

[18] T. Gebru, J. Hoffman, and L. Fei-Fei, "Fine-grained recognition in the wild: A multi-task domain adaptation approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1349–1358.

[19] Y. Chai, E. Rahtu, V. Lempitsky, L. Van Gool, and A. Zisserman, "TriCoS: A Tri-level class-discriminative co-segmentation method for image classification," in *Computer Vision—ECCV*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Germany: Springer, 2012, pp. 794–807.

[20] X. Liu, T. Xia, J. Wang, and Y. Lin, "Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition," *CoRR*, vol. abs/1603.06765, Mar. 2016.

[21] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, Z. Luo, V.-A. Nguyen, and M. Do, "Weakly supervised fine-grained image categorization," *CoRR*, vol. abs/1504.04943, pp. 4321–4329, Apr. 2015.

[22] C. Göring, A. Freytag, E. Rodner, and J. Denzler, "Fine-grained categorization—Short summary of our entry for the ImageNet challenge 2012," *CoRR*, vol. abs/1310.4759, pp. 1–7, Oct. 2013.

[23] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei, "The unreasonable effectiveness of noisy data for fine-grained recognition," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016, pp. 301–320.

[24] J. Han, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann, 2005.

[25] F. Schwenker and E. Trentin, "Pattern classification and clustering: A review of partially supervised learning approaches," *Pattern Recognit. Lett.*, vol. 37, pp. 4–14, Feb. 2014.

[26] X. Zhu, "Semi-supervised learning literature survey," Ph.D. dissertation, Dept. Comput. Sci., Univ. Wisconsin, Madison, WI, USA, Tech. Rep. 07, 2008.

[27] I. Triguero, S. García, and F. Herrera, "Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study," *Knowl. Inf. Syst.*, vol. 42, no. 2, pp. 245–284, Feb. 2015, doi: 10.1007/s10115-013-0706-y.

[28] S. S. Azab, M. F. A. Hady, and H. A. Hefny, "Semi-supervised classification: Cluster and label approach using particle swarm optimization," *Int. J. Comput. Appl.*, vol. 160, no. 3, pp. 39–44, Feb. 2017. [Online]. Available: http://www.ijcaonline.org/archives/volume160/number3/27056-2017913013

[29] M. Li and Z. H. Zhou, "Setred: Self-training with editing," in *Advances in Knowledge Discovery and Data Mining (PAKDD)* (Lecture Notes in Computer Science), vol. 3518. Springer, 2005, pp. 611–621.
[30] D. Wu, M. S. Shang, X. Luo, J. Xu, H.-Y. Yan, W.-H. Deng, and G.-Y. Wang, "Self-training semi-supervised classification based on density peaks of data," *Neurocomputing*, vol. 275, pp. 180–191, May 2017.
[31] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018.
[32] K. D. Tang, V. Ramanathan, F. Li, and D. Koller, "Shifting weights: Adapting object detectors from image to video," in *Proc. Adv. Neural Inf. Process. Syst. 26th Annu. Conf. Neural Inf. Process. Syst. Meeting Held*, Lake Tahoe, NV, USA, 2012, pp. 647–655. [Online]. Available: http://papers.nips.cc/paper/4691-shifting-weights-adapting-object-detectors-from-image-to-video
[33] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, "Few-example object detection with model communication," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1641–1654, Jul. 2019.
[34] X. Dong and Y. Yang, "Teacher supervises students how to learn from partially labeled images for facial landmark detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 783–792.
[35] M. wei Chang, L. Ratinov, and D. Roth, "Guiding semi-supervision with constraint-driven learning," in *Proc. Annu. Meeting ACL*, 2007.
[36] B. Drury, L. Torgo, and J. J. Almeida, "Guided self training for sentiment classification," in *Proc. Workshop Robust Unsupervised Semisupervised Methods Natural Lang. Process.*, Hissar, Bulgaria, Sep. 2011, pp. 9–16. [Online]. Available: https://www.aclweb.org/anthology/W11-3902
[37] P. Roy, S. Ghosh, S. Bhattacharya, and U. Pal, "Effects of degradations on deep neural network architectures," *CoRR*, vol. abs/1807.10108, 2018.
[38] R. F. L. Fei-Fei and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Comput. Vis. Pattern Recognit., Workshop Generative-Model Based Vis.*, 2004.
[39] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101—Mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.*, 2014.
[40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
[41] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *Proc. 1st Workshop Fine-Grained Vis. Categorization, IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 1–2.
[42] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. 4th Int. IEEE Workshop 3D Represent. Recognit. (3dRR-13)*, Sydney, NSW, Australia, Dec. 2013, pp. 554–561.
[43] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf Comput. Vis., Graph. Image Process. (ICVGIP)*, 2008, pp. 722–729.
[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.
[45] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, 2016.
[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[48] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2014.
[49] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 71–88.
[50] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2018, pp. 834–850.
[51] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 811–818.
[52] I. Gogul and V. S. Kumar, "Flower species recognition system using convolution neural networks and transfer learning," in *Proc. 4th Int. Conf. Signal Process., Commun. Netw. (ICSCN)*, 2017, pp. 1–6.

OBED TETTEY NARTEY received the B.Sc. degree from All Nations University College, Ghana, in 2011, and the M.Sc. degree from the University of Electronic Science and Technology of China, in 2015, where he is currently pursuing the Ph.D. degree. His research interests include artificial intelligence, computer vision, machine learning and pattern recognition, data mining, and deep learning.

GUOWU YANG received the B.S. degree from the University of Science and Technology of China, in 1989, the M.S. degree from the Wuhan University of Technology, in 1994, and the Ph.D. degree in electrical and computer engineering from Portland State University, in 2005. He is currently a Full Professor with the University of Electronic Science and Technology of China. He has published more than 100 journal articles and conference papers. His research interests include machine learning, logic synthesis, and quantum computing.

JINZHAO WU was born in 1965. He received the Ph.D. degree in science from the Institute of Systems Science, Chinese Academy of Sciences. He is currently a Professor with the Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Guangxi University. His research interests include the fields of formal methods, symbolic computation, and automated reasoning.

SARPONG KWADWO ASARE received the B.Sc. degree from the University of Ghana, in 2012, and the M.Sc. degree from the University of Electronic Science and Technology of China, in 2015, where he is currently pursuing the Ph.D. degree. His research interests include artificial intelligence, machine learning, statistical learning, and deep learning methods and applications.