SSFE-Net: Self-Supervised Feature Enhancement for Ultra-Fine-Grained Few-Shot Class Incremental Learning

Zicheng Pan, Xiaohan Yu, Miaohua Zhang, Yongsheng Gao
School of Engineering and Built Environment
Griffith University, QLD, 4111, Australia
{z.pan; xiaohan.yu; lena.zhang; yongsheng.gao}@griffith.edu.au

Abstract

Ultra-Fine-Grained Visual Categorization (ultra-FGVC) has become a popular problem due to its great real-world potential for classifying the same or closely related species with very similar layouts. However, there present many challenges for the existing ultra-FGVC methods, firstly there are always not enough samples in the existing ultra-FGVC datasets based on which the models can easily get overfitting. Secondly, in practice, we are likely to find new species that we have not seen before and need to add them to existing models, which is known as incremental learning. The existing methods solve these problems by Few-Shot Class Incremental Learning (FSCIL), but the main challenge of the FSCIL models on ultra-FGVC tasks lies in their inferior discrimination detection ability since they usually use low-capacity networks to extract features, which leads to insufficient discriminative details extraction from ultra-fine-grained images. In this paper, a self-supervised feature enhancement for the few-shot incremental learning network (SSFE-Net) is proposed to solve this problem. Specifically, a self-supervised learning (SSL) and knowledge distillation (KD) framework is developed to enhance the feature extraction of the low-capacity backbone network for ultra-FGVC few-shot class incremental learning tasks. Besides, we for the first time create a series of benchmarks for FSCIL tasks on two public ultra-FGVC datasets and three normal fine-grained datasets, which will facilitate the development of the Ultra-FGVC community. Extensive experimental results on public ultra-FGVC datasets and other state-of-the-art benchmarks consistently demonstrate the effectiveness of the proposed method.

1. Introduction

Ultra-FGVC tasks start getting people's attention in recent years and have shown their great potential in many science fields like precision agriculture. Compared with classic classification datasets which are usually designed for distinguishing different types of objects, for example (*e.g.*), birds [29], cars [18], and aircrafts [20], ultra-FGVC mainly focuses on identifying objects from the same or closely related species with large intra-class and small inter-class variances, like different kinds of cotton leaves. As shown in Figure. 1. It's clear that the ultra-FGVC leaf datasets have much higher intra-class and smaller inter-class variances, which is more challenging for the model to identify different species compared with normal FGVC datasets.

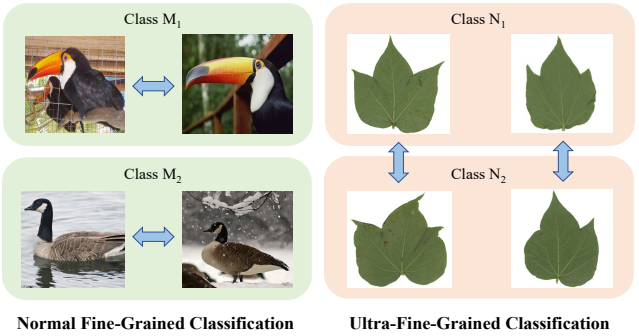

Figure 1: Normal classification tasks (green boxes) vs ultra-FGVC tasks (pink boxes). Images in the same box represent the same species (cultivars). Blue arrows indicate the images which have high similarity.

Despite the high similarity among different classes, another challenge of ultra-FGVC lies in the limitation of training samples. In existing ultra-FGVC datasets, there are usually only a small amount of annotated images available, which matches better with real-world scenarios, especially in scientific fields. For example, there are only six images for each class in the ultra-FGVC dataset CottonCultivar [35]. On the other hand, as the need for in-depth feature extraction increases as well as more and more new classes of objects detected in nature, people find that it's necessary to extend the current ultra-FGVC model without losing too

much performance, which is known as class incremental learning.

The existing methods solve the above few-label and incremental learning problems by few-shot class incremental learning [26, 11, 21, 37, 36, 43, 7] which has received great attention in recent years. However, FSCIL has never been applied to the ultra-FGVC problem which was first proposed by Yu *et.al.* [35] in 2021 with a series of ultra-FGVC datasets and benchmarks. Besides, we find that the state-of-the-art FSCIL methods pay much attention to the catastrophic forgetting problem while ignoring the inferior feature extraction ability of their low-capacity backbones. Learning discriminative representation is vital for extensive vision tasks [40, 39]. For example, most FSCIL methods adopt ResNet18 [26, 21, 7, 43] as their backbone network, leading to insufficient discriminative details extraction from ultra-fine-grained images. The reason why they use low-capacity networks is that training insufficient data samples on complex model will lead to overfitting. To address these problems, we propose a novel SSFE-Net architecture which combines self-supervised learning (SSL) with knowledge distillation (KD). With this design, we can take the advantage of a deeper neural network structure (e.g., ResNet50 [15]) to learn much richer discriminative features under the data-limited circumstances. Besides, the features learned from the self-supervised learning module strengthen the discrimination detection ability of the network from the same sample, which is crucial in ultra-FGVC datasets. Furthermore, the proposed model makes use of the nature of small inter-class variance features on ultra-FGVC datasets by applying a class mean incremental adaptation module. To verify the effectiveness of the SSFE-Net, we conduct experiments on two ultra-FGVC datasets and three commonly used fine-grained datasets. Since this is the first time to apply few-shot class incremental learning on ultra-fine-grained datasets, we create benchmarks on these datasets using different benchmark methods, which will facilitate the development of the ultra-FGVC community. The results indicate that our model has significant improvement compared with other benchmarks and achieves state-of-the-art performance.

The contributions of our work can be summarized as follows:

- A novel self-supervised feature enhancement network (SSFE-Net) is proposed to enhance the feature extraction ability of the low-capacity backbone in ultra-fine-grained FSCIL. An overview of the proposed SSFE-Net is shown in Figure 2.
- A self-supervised learning (SSL) module is developed to extract more high-dimensional feature representations without getting overfitting due to its advantages of robust self-supervised feature augmentation ability. Besides, a knowledge distillation module (KD) is constructed to transfer high-quality features extracted by SSL and augment the low-capacity FSCIL network.
- For the first time, a series of FSCIL benchmarks are created based on two different ultra-FGVC datasets, which will facilitate the development of the ultra-FGVC community.

2. Related Works

2.1. Ultra-Fine-Grained Visual Categorization

Ultra-Fine-Grained Visual Categorization (ultra-FGVC) has gained much attention in recent years [35, 19, 22, 25, 33, 34, 32, 31]. Compared with normal fine-grained visual categorization (FGVC) tasks, ultra-FGVC is more challenging due to the far fewer labelled samples in each class since it is labour-consuming to annotate large-scale datasets. Another challenge is that the ultra-FGVC datasets have small inter-class similarities, even human experts may fail to distinguish different species. Therefore, studying ultra-FGVC will have tremendous development foreground and practical value.

In recent years, many research works are conducted based on ultra-FGVC tasks and make remarkable progress. The first ultra-FGVC tasks was proposed by Larese *et.al* [19] based on a soy leaf dataset. Their work mainly focuses on classifying different soy leaves using traditional machine learning techniques, which demonstrates an effective way of identifying leaves by discriminative vein regions. However, they only made use of vein details of the leaf, which cannot effectively extract and make use of other informative parts of the leaf, e.g., the leaf contour, colour, etc. Besides, their soy leaf dataset only contains three cultivars, which reduces the complexity of the task and model. To better make predictions on ultra-FGVC tasks, Yu *et.al* [35] first proposed several ultra-FGVC datasets and a series of benchmarks for the development of the ultra-FGVC community. These datasets are related to the precision agriculture field and have great diversity and complexity. The authors further investigated these datasets and ultra-FGVC challenges [33], and designed a MaskCOV network architecture to better address the problems. MaskCOV is a feature argumentation method that splits the original images into several equal parts. These patch level covariance features will be masked or randomly combined to form new features. With these modified images, the model not only can focus on general layouts of objects but also can better capture the discriminative areas to make predictions. Based on their studies, we know that the common challenges of ultra-FGVC come from the overfitting problem because of the limitation of data samples as well as their inferior feature extraction ability. Therefore, advancing feature en-

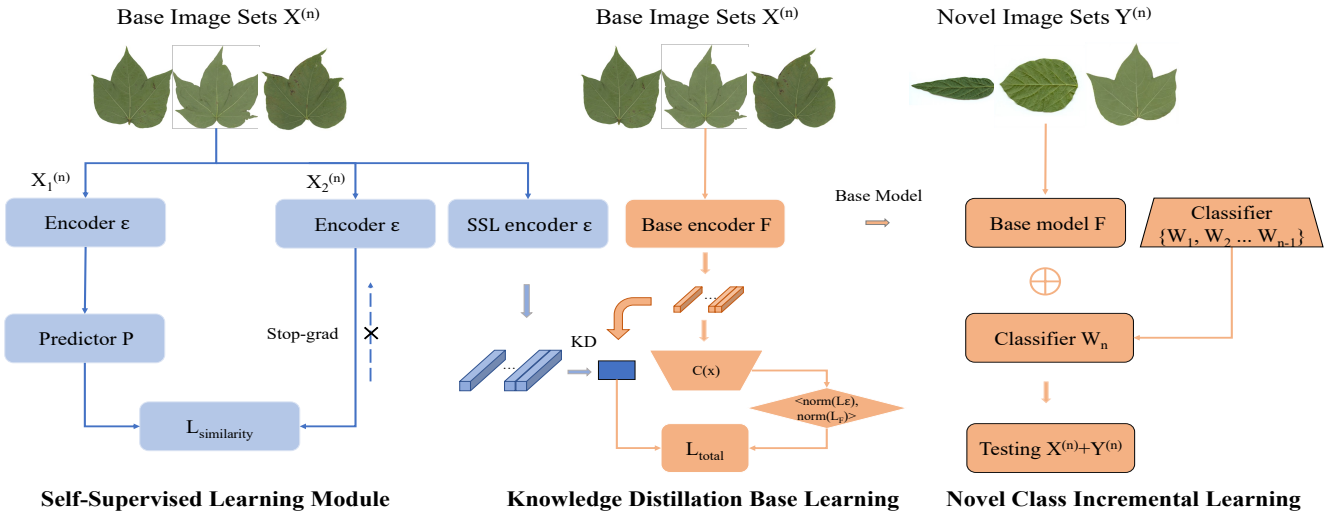

Figure 2: Architecture of SSFE-Net. The SSL model is first trained on the base part of a dataset. The feature representation in SSL will then be combined with normal FSCIL features via KD to produce detailed information. The incremental learning will combine the current session classifier with previous classifiers and take the advantage of the robust base model to produce a better prediction head.

hancement is the most efficient and feasible way to address the ultra-FGVC tasks.

2.2. Few-Shot Class Incremental Learning (FSCIL)

FSCIL aims to explore the new class incrementation ability of normal CNNs with very few training samples. Tao *et.al* [26] first proposed the FSCIL concept in 2020 and named their framework as TOPIC. They introduced a neural gas network to maintain the feature representations space in incremental learning. The most challenging task in FSCIL is the catastrophic forgetting problem during new class incremental stages. Zhu *et.al* [43] introduced a randomly episodic training scheme in which they used random incremental episodes and a self-promoted prototype refinement mechanism to extend the network and maintain the dependencies of old classes. Achituve *et.al* [1] developed a tree-based model using common Gaussian process classification methods with deep kernel learning. Zhao [37] further addressed the catastrophic forgetting problem by balancing the model ability of slow forgetting old knowledge and fast adaptation to novel classes. Mazumder *et.al* [21] proposed an FSLL structure that uses self-supervised learning as an auxiliary loss to train the network. They only updated a small number of parameters of the base training network, which can prevent the network from deviating far away compared with its previous values.

2.3. Self-Supervised Learning (SSL)

It does not require any supervision from human label annotations when performing SSL, which is beneficial for solving real-world problems where we don't always have many labelled samples. Recently, many research works have demonstrated the effectiveness of SSL on both detection and classification tasks [8, 13, 4, 14, 9], and also proven that SSL benefits the deep neural network by learning robust features representations for typical few-shot tasks [12, 24, 6, 10]. Chen *et.al* [6] embedded Augmented Multiscale Deep InfoMax (AMDIM) [3] as their SSL model to few-shot classification tasks. The pre-trained SSL model can maximise the mutual information from different views of an image and significantly improve classification performance. Su *et.al* [24] augmented the unlabelled images by rotation and jigsaw puzzle. The new image representations combined with labelled images are used to improve few-shot learning and increase the network robustness and generalisation ability. Besides the applications of SSL on normal few-shot learning tasks, recently, some studies focus on presenting SSL as auxiliary tasks to support FSCIL problems and achieve great improvement. Particularly, Mazumder *et.al* [21] utilized rotation prediction in SSL to provide auxiliary loss while doing the few-shot class incremental learning. Zhu *et.al* [42] proposed a PASS architecture which also adopted rotational based augmentation SSL to images.

3. Method

In this section, we first give a brief introduction to few-shot class incremental learning and then introduce our method in three parts, including the self-supervised learning module, knowledge distillation module, and the final incre-

mental learning.

3.1. Few-shot Class Incremental Learning (FSCIL) Task Formulation

In FSCIL tasks, the dataset is split into a stream of labelled subsets ($\mathcal{D}^1_{train}$, $\mathcal{D}^2_{train}$, ..., $\mathcal{D}^n_{train}$), where n represents different training sessions. The data stream has no overlapping classes with each other and only $\mathcal{D}^n_{train}$ is available at training session n. On the other hand, for the test samples $\mathcal{D}^n_{test}$ at session n, all testing samples in previous seen classes will be used for evaluation. Normally, the first training session $\mathcal{D}^1_{train}$ is a relatively large dataset for training a base model, all following few-shot training sessions $\mathcal{D}^{n>1}_{train}$ contain N classes and K training samples per class denoted as N-way K-shot in FSCIL.

Inspired by [6] which adopted SSL to train a large embedding network for a two-state paradigm, we use SSL to pre-train the base learning part in the FSCIL task for in-depth features. The SSL model will be trained using SimSiam architecture with a contrastive learning technique to extract semantic information for simplicity. The proposed network has no restriction on SSL methods so other SSL techniques also work. In the FSCIL base training stage, the feature maps and feature vectors extracted by ResNet18 will be combined with SSL mutual information via KD to provide more discriminative features.

In the incremental learning stage, since the main restriction of FSCIL on ultra-FGVC tasks comes from the lack of ability to obtain in-depth discriminative area on different objects at the base training model, improving the model discriminative feature extraction ability is more important at this beginning stage. In the following, we will introduce the proposed self-supervised learning module and few-shot class incremental learning subsequently.

3.2. Self-Supervised Learning (SSL) Module

Many studies have shown that SSL can be a good initialization of a model [6, 27, 2]. We use SSL to maximize the mutual information and provide multiple views for sample images. The SSL model is pre-trained by only using the base part of a dataset, all data in novel sessions will not get involved at this stage. The SSL model we used applies the contrastive learning concept as indicated in the self-supervised learning module of Figure 2. The network randomly augments an image x^n as two different views x^n_1 and x^n_2. The two views will then be further processed by a feature extraction encoder network $\mathcal{E}(x)$ and a prediction head $\mathcal{P}(x)$. $\mathcal{E}(x)$ can share weights between two views during training and the prediction head $\mathcal{P}(x)$ which is responsible for transferring as well as matching one view feature to another. The outputs from different views are simplified as $p_1 \triangleq \mathcal{E}(x_1)$ and $z_2 \triangleq \mathcal{P}(\mathcal{E}(x_2))$. Then the negative cosine similarity between two outputs is defined by:

$$\boldsymbol{\mathcal{G}}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \times \frac{z_2}{\|z_2\|_2}, \tag{1}$$

in which $\|\cdot\|_2$ represents l_2-norm. Note that a stop-gradient operation is applied on z_2 to prevent dimensional collapsing, which means z_2 becomes a constant $stopgrad(z)$. Finally, the symmetrized loss following [13] is defined as:

$$\boldsymbol{\mathcal{L}} = \frac{1}{2}\boldsymbol{\mathcal{G}}(p_1, stopgrad(z_2)) + \frac{1}{2}\boldsymbol{\mathcal{G}}(p_2, stopgrad(z_1)) \tag{2}$$

By studying the images with their own variances, the SSL module enhances the representation learning ability of the model and provides more details of discriminative areas without overfitting problems.

3.3. Knowledge Distillation (KD) Module

The mutual information obtained in the pre-trained SSL module will be further processed and combined with standard few-shot incremental training at the knowledge distillation stage in Figure 2. Same as normal FSCIL settings, the incremental learning in the proposed framework can be divided into base training and multiple novel class training sessions. Inspired by [38], the student network will be trained under the guidance of the teacher network's feature-aligned distillation and feature similarity distillation. Since there is no need to recognize landmark positions on images as in [38], we propose feature vectors aligned distillation instead of feature maps to determine the distillation loss. The vectorized distillation loss can be defined by making use of the $Kullback - Leibler$ (KL) divergence loss.

During the base training stage, the FSCIL backbone $\mathcal{F}_{base}$ generates feature vectors $\mathcal{V}_{base} = \mathcal{F}_{base}(\mathcal{D}^1_{train})$ without classification head. At the same time, the SSL network also processes feature maps and resizes the embedding to match the FSCIL output $\mathcal{V}_{ssl} = Conv(\mathcal{F}_{ssl}(\mathcal{D}^1_{train})$. The model will leverage two feature embeddings and compare their similarity. The loss is generated by integrating the cross entropy loss $\mathcal{L}_{ce}$ of model predictions and the KL divergence loss $\mathcal{L}_{KL}$:

$$\boldsymbol{\mathcal{L}}_{total} = \boldsymbol{\mathcal{L}}_{ce}(\beta\mathcal{V}_{base} + \gamma\mathcal{V}_{ssl}) + \alpha\boldsymbol{\mathcal{L}}_{KL}(\mathcal{V}_{base}, \mathcal{V}_{ssl}), \tag{3}$$

where the parameters β, γ, and α denote the weight contribution hyper-parameters of $\mathcal{L}_{ce}$ and $\mathcal{L}_{KL}$.

3.4. Incremental Learning

In the incremental stage, we use decoupling backbone with classifier strategy to update the classifier without modifying the backbone parameters similar to the continually evolved classifiers [36]. This method uses the mean of each class feature representation and has great benefits on ultra-FGVC incremental tasks since the objects are all similar,

learning the mean of representations can be easily transferred to novel classes without heavy training. For the individual session n, the classifier w_n is generated by replacing the original predicting prototype classifier with the mean of the feature representations from images of same class via

$$\boldsymbol{w}_n^j = \frac{1}{M_j} \sum_{m=1}^{M_j} F(\mathcal{D}_j^n), \tag{4}$$

where M is the number of images of class j. The new incremental classifier w_n will be concatenated with classifiers in the previous sessions to form an evolved classifier $\mathcal{W}_n = \{w_1, w_2, ..., w_n\}$. When making predictions, the network computes the cosine similarity between normalized samples and the classifier by projecting the original representations to the new sample space via inner product calculation:

$$\boldsymbol{P} = \langle norm(\mathcal{D}_j^n), norm(\mathcal{W}_n) \rangle, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product calculation to find the similarity. Each novel session, we update the classifier with the mean of new training sample representations and aggregate it with the original classifier to test over all seen classes.

4. Experiments

To verify the effectiveness of the proposed method, we carry out experiments on two public ultra-fine-grained and three normal fine-grained datasets. Some example images from each dataset are shown in Figure 3. The proposed algorithm is tested and compared with three recently published approaches. Besides, ablation studies are done under different experimental settings to further verify the effectiveness of the proposed method.

4.1. Datasets

CottonCultivar [35]. The CottonCultivar dataset was first proposed in [35] for ultra-FGVC tasks based on different cultivars of cotton leaves. There are 80 classes in total, 40 for base training, and the remaining 40 classes are split into 8 incremental sessions with 5 images for training and 1 for testing per class at the FSCIL stage. The images are resized to 512×512 and cropped to 448×448 at the training stage.
SoyCultivarLocal [35]. SoyCultivarLocal is a larger ultra-FGVC leaf dataset that focuses on cultivars of soybeans. It has 200 classes with 6 images per class. We use the same training and testing split as CottonCultivar and the same base novel splits as CUB200. Each image is resized to 512×512 and then cropped to 448×448 during training.
PlantVillage [17]. PlantVillage is a public dataset designed for plant disease detection systems initially. It consists of 38 different classes of leaf diseases and species. Since the image number is imbalanced among different classes, we randomly choose 100 images, resize them to 256×256, and crop them to 224×224 from each class for training. There are 23 classes in the base training session and the rest 15 classes are further equally split into 3 novel sessions.
Caltech-UCSD Birds-200-2011 (CUB200) [29]. CUB200 consists of 11788 images from 200 different bird categories. Each image is resized to 256×256 and then cropped to 224×224 in the training stage. We use 100 categories for base training and the rest 100 for incremental training following the setting in [26]
Mini-ImageNet [28]. It's a subset of the ImageNet-1k dataset which contains 100 categories with 600 samples of 84×84 color images per class. Same as [26], the first 60 classes are used for base training and the rest is split equally into 8 sessions.

Table 1: Benchmark results on SoyCultivarLocal dataset using 10-way-5-shot setting.

Methods	Sessions (%)										
	1	2	3	4	5	6	7	8	9	10	11
SPPR [43]	6.00	7.27	4.62	5.00	4.67	5.00	4.12	4.12	3.33	2.11	3.00
PASS [42]	8.00	11.81	10.00	6.92	5.71	5.33	4.38	4.12	2.78	3.16	2.50
CEC [36]	26.00	24.56	23.33	**36.33**	19.66	20.42	19.14	20.85	21.91	20.20	18.07
SSFE-Net	**28.73**	**27.27**	**27.50**	26.15	**25.71**	**24.00**	**23.75**	**21.76**	**22.22**	**21.05**	**20.00**

Table 2: Benchmark results on CottonCultivar dataset using 5-way-5-shot setting.

Methods	Sessions (%)								
	1	2	3	4	5	6	7	8	9
SPPR [43]	12.50	6.98	6.52	6.12	5.77	5.45	5.17	4.92	4.69
PASS [42]	10.00	11.11	12.00	3.64	6,67	7.69	7.14	8.00	5.00
CEC [36]	17.50	15.55	14.00	12.73	8.33	4.69	13.02	9.23	7.81
SSFE-Net	**25.00**	**17.78**	**18.00**	**14.55**	**15.00**	**13.85**	**15.71**	**14.67**	**13.75**

Table 3: Benchmark results on PlantVillage dataset using 5-way-5-shot setting. The FSCIL parameter shows different methods' trainable parameter amounts at the FSCIL stage.

Methods	FSCIL Param	Sessions (%)			
		1	2	3	4
SPPR [43]	12.31M	91.13	71.16	57.71	46.62
PASS [42]	**11.35M**	86.85	75.75	64.30	53.79
CEC [36]	12.33M	95.81	88.01	78.90	71.24
SSFE-Net	11.46M	**97.31**	**89.28**	**79.04**	**72.90**

4.2. Implementation Details

All experiments are conducted under the PyTorch framework. As indicated in Eq.3, there are three trade-off hyper-parameters β, γ, and α when calculating the loss, they are set to $8e^{-1}$, $2e^{-1}$, and $9e^{-1}$ respectively. ResNet50 is adopted at the SSL pre-training stage to extract more detailed features, the learning rate is decayed by 0.1 every 20 epochs. We use a common 5-way-5-shot setting for CUB200, Mini-ImageNet, and PlantVillage following the

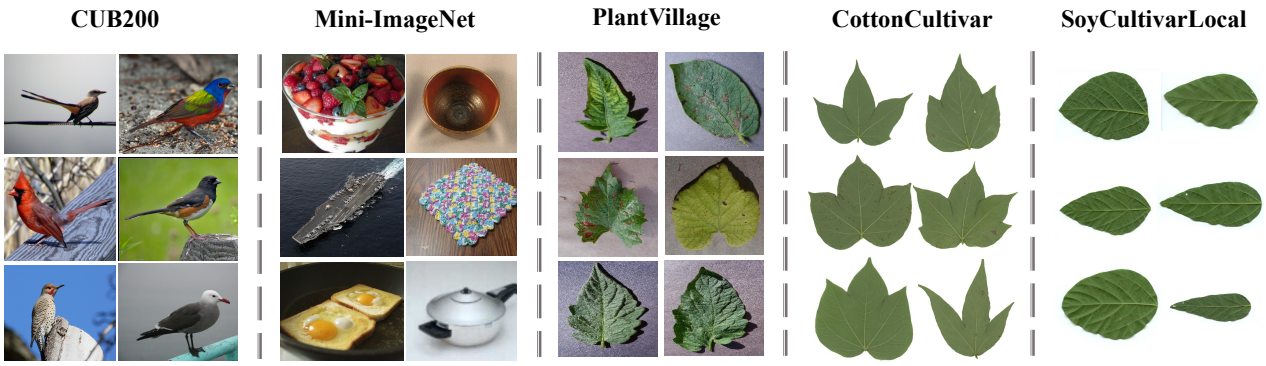

Figure 3: Sample images from five datasets that are used in our study. Each image represents an isolated class of the object.

Table 4: Experimental results on CUB200 dataset using different methods.

Methods	Sessions (%)										
	1	2	3	4	5	6	7	8	9	10	11
iCaRL [23]	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16
EEIL [5]	68.68	53.63	47.91	44.20	36.30	27.46	25.93	24.70	23.95	24.13	22.11
NCM [16]	68.68	57.12	44.21	28.78	26.71	25.66	24.62	21.52	20.12	20.06	19.87
TOPIC [26]	68.68	62.49	54.81	49.99	45.25	41.40	38.35	35.36	32.22	28.31	26.28
SPPR [43]	68.68	61.85	57.43	52.68	50.19	46.88	44.65	43.07	40.17	39.63	37.33
SDC [30]	72.29	68.22	61.94	61.32	59.83	57.30	55.48	54.20	49.99	48.85	42.58
GP-Tree [1]	72.84	67.00	62.98	58.19	54.84	51.77	49.40	47.57	45.47	44.05	42.72
DC [28]	75.52	70.95	66.46	61.20	60.86	56.88	55.40	53.49	51.94	50.93	49.31
FSLL+SS [21]	75.63	71.81	68.16	64.32	62.61	60.10	58.82	58.70	56.45	**56.41**	**55.82**
CEC [36]	75.85	71.94	68.50	63.50	62.43	58.27	57.73	55.81	54.83	53.52	52.28
SSFE-Net	**76.38**	**72.11**	**68.82**	**64.77**	**63.59**	**60.56**	**59.84**	**58.93**	**57.33**	56.23	54.28

work in [2]. Due to the limitation of training samples per class in CottonCultivar and SoyCultivarLocal datasets, the 5-way-3-shot training setting is applied to all the involved comparison methods. In the FSCIL stage, a commonly used ResNet18 structure is adopted as the backbone. We use the 10-way 5-shot setting for the Mini-ImageNet as well as SoyCultivarLocal datasets while the 5-way 5-shot for the remaining datasets, which is similar to the settings in work [26]. For fair comparisons, all experiments we performed on ultra-FGVC datasets do not use the backbone with the ImageNet pre-trained ResNet18. We also conduct ablation studies with the ImageNet pre-trained backbone using the proposed model and CEC benchmark. More details of the hyper-parameter settings and the tuning process can be found in the supplementary materials Section A.

4.3. Compare with the State-Of-The-Art Methods

To verify the effectiveness of our proposed method, we compare our model with other competitive state-of-the-art methods. For the SoyCultivarLocal, CottonCultivar, and PlantVillage datasets, Tables 1 - 3 present the performances of different methods on these benchmark datasets, the best accuracy is highlighted in bold. Since only a small number of the existing models fully release their source code, we only use SPPR [43], PASS [42], as well as CEC [36] for ultra-FGVC benchmark comparison. We apply optimal fine-tuning settings for all benchmark models under the new dataset splittings, which are far better than they claimed in their corresponding papers. The experimental results in these tables consistently verify that the proposed method has superior performance on all ultra-FGVC datasets. Since the SSL model will not get updated at the FSCIL stage, the SSFE-Net has a similar amount of trainable parameters to other benchmarks as presented in Table 3 during the FSCIL stage.

We also conduct experiments on the commonly used fine-grained datasets CUB200 and Mini-ImageNet of traditional FSCIL tasks to further validate the generalisation ability of the proposed SSFE-Net. The comparison methods include iCaRL [23], EEIL [5], NCM [16], TOPIC [26], SPPR [43], SDC [30], GP-Tree [1], Decoupled-Cosine (DC) [28], FSLL+SS [21], and CEC [36]. Tables 4 and 5 show the testing performances on CUB200 and Mini-ImageNet. It's clear that the SSFE-Net can still achieve

competitive performance on normal FSCIL datasets.

Table 5: Experimental results on Mini-ImageNet dataset using different methods.

Methods	Sessions (%)								
	1	2	3	4	5	6	7	8	9
iCaRL [23]	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21
EEIL [5]	61.31	46.58	44.00	37.29	33.14	27.12	24.10	21.57	19.58
NCM [16]	61.31	47.80	39.31	31.91	25.68	21.35	18.67	17.24	14.17
TOPIC [26]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42
SPPR [43]	61.45	63.80	59.53	55.53	52.50	49.60	46.69	43.79	41.92
SDC [30]	64.62	59.63	55.39	50.92	48.30	45.28	42.97	42.51	41.24
GP-Tree [1]	62.32	57.10	52.90	49.36	46.28	43.55	41.13	38.97	37.02
DC [28]	70.37	65.45	61.41	58.00	54.81	51.89	49.10	47.27	45.63
FSLL+SS [21]	68.85	63.14	59.24	55.23	52.24	49.65	47.74	45.23	43.92
CEC [36]	72.00	**66.83**	**62.97**	59.43	**56.70**	53.73	51.19	49.24	47.63
SSFE-Net	**72.06**	66.17	62.25	**59.74**	56.36	**53.85**	**51.96**	**49.55**	**47.73**

5. Ablation Studies

The first ablation study focuses on comparing the contributions of SSL models and KD to the improvement of the FSCIL tasks. We adopt different SSL backbones with/without the presence of KD and ImageNet pre-trained backbone to analyse the effectiveness of different components in the SSFE-Net. On the other hand, we test the ultra-FGVC datasets with other commonly used incremental class split proportions and verify that the dataset splitting does not have a large influence on the model performance.

5.1. The Contributions of SSL Embeddings and KD

To verify whether the SSL feature enhancement network and KD embedding fusion have great improvement on FSCIL tasks, we conduct extensive experiments to analyse the contribution of different SSL backbone components and KD on CUB200, SoyCultivarLocal, and PlantVillage datasets. The performance comparisons under different component combinations in the base training session (session 1) are shown in Table 6, which clearly shows that the self-feature enhancement module (SSL) and KD have great improvements on FSCIL tasks. The SSL module with ResNet50 backbone has the greatest improvement among other backbones since it can generate higher dimensional features compared with ResNet34 and is less prone to overfitting than ResNet101. Besides the SSL training with 5-way 5-shot/3-shot settings, ablation experiments under the 5-way 1-shot (5w1s) SSL setting are also conducted to further investigate the model performance. The results consistently verify that the performance with the 5w1s setting still benefits from the SSL module, and the KD component is capable of properly fusing the SSL features into FSCIL.

The running time of the model is evaluated by running experiments on an NVIDIA A5000 GPU for 150 epochs. Furthermore, we visualise the differences between the normal FSCIL method and the SSFE-Net via class activation maps (CAM) [41]. Please refer to supplementary materials for comparisons. It's clear that the SSFE-Net can better focus on the most discriminative areas of the objects.

5.2. ImageNet Pre-trained FSCIL

The experiments on ultra-FGVC datasets are conducted by training the ResNet18 of FSCIL from scratch. To further explore the potential of the model, we adopt ImageNet pre-trained parameters to SSFE-Net as well as the best benchmark model CEC. The rest of the settings remain the same. The comparison results are shown in Tables 7 and 8 from which we can see that the overall performances of both methods under the above settings are significantly enhanced, and the proposed SSFE-Net still maintains its superiority over CEC.

5.3. Dataset Split Proportion in FSCIL

Since there's no benchmark on the few-shot incremental learning with ultra-FGVC datasets in the literatures, we mainly follow the split proportion for the CUB200 dataset in [26] where half of the data classes are used as base learning. The proposed method is further tested under other commonly used data split settings on PlantVillage and CottonCultivar datasets. For PlantVillage, 20 classes are used for base training and the remaining 18 classes are further separated into 6 sessions for incremental learning following a 3-way-5-shot setting. For CottonCultivar, 60 classes are selected for base training, and the remaining 20 classes are split into 4 incremental sessions following a 5-way-5-shot setting. The performance of CEC and SSFE-Net with different splits on PlantVillage and CottonCultivar are reported in Tables 9 and 10 from which we can see that the proposed method still achieves state-of-the-art performance.

6. Discussions

The experimental results in Sections 4-5 clearly show that the proposed SSFE-Net has superior performance on the ultra-fine-grained few-shot incremental learning tasks. From the CAM graph visual analysis of samples (see supplementary materials for details), we can see that the proposed model has a better ability to locate and focus on the discriminative areas of the images compared with other methods. The experiments on the CottonCultivar dataset further demonstrate its strong feature enhancement ability. The second best method CEC experiences a dramatic drop in classification performance in Table 2 and 10 because it has no mechanism for processing ultra Fine-grained data. The leaf samples from the same class are coming from different parts of the cotton plant so they look very different. CEC lacks detail capture ability and fails to identify different species when new classes come in. With the consideration of the small inter-class similarity and large intra-class

Table 6: Base model training ablation studies on the presenting of different SSL backbones and KD on CUB200, SoyCultivarLocal, and PlantVillage datasets.

Model Components			CUB200		SoyCultivarLocal		PlantVillage	
SSL Backbone	KD	FSCIL	Session 1 Acc. (%)	Training Time (s)	Session 1 Acc. (%)	Training Time (s)	Session 1 Acc. (%)	Training Time (s)
✗	✗	✓	73.88	1145	19.88	818	95.38	1839
ResNet34	✗	✓	73.13	1264	20.92	1053	96.13	2012
ResNet50	✗	✓	73.43	1382	22.26	1271	96.24	2178
ResNet101	✗	✓	73.67	2055	23.83	1485	95.39	2342
ResNet34	✓	✓	74.92	1279	26.48	1055	96.81	2018
ResNet101	✓	✓	74.18	2079	25.66	1476	95.57	2367
ResNet50 (5w1s)	✓	✓	76.05	1385	27.98	1206	97.27	2190
ResNet50	✓	✓	76.58	1379	28.73	1219	97.31	2197

Table 7: Benchmark results on SoyCultivarLocal dataset with ImageNet pre-trained FSCIL model.

Methods	Sessions (%)										
	1	2	3	4	5	6	7	8	9	10	11
CEC [36]	31.00	27.27	28.33	13.67	13.67	19.71	20.31	20.83	**22.11**	**21.37**	19.44
SSFE-Net	**37.24**	**29.27**	**29.10**	**23.85**	**22.86**	**21.67**	**21.00**	**21.17**	20.33	20.37	**20.00**

Table 8: Benchmark results on CottonCultivar dataset with ImageNet pre-trained FSCIL model.

Methods	Sessions (%)								
	1	2	3	4	5	6	7	8	9
CEC [36]	52.50	42.22	36.00	32.77	28.33	30.77	28.57	28.00	27.50
SSFE-Net	**60.00**	**53.33**	**46.00**	**40.00**	**36.67**	**36.92**	**35.71**	**30.67**	**28.74**

Table 9: Experimental results from different methods on PlantVillage dataset with different data splittings under the 3-way-5-shot setting.

Methods	Sessions (%)						
	1	2	3	4	5	6	7
CEC [36]	95.26	90.96	**81.96**	78.50	**73.24**	66.86	63.23
SSFE-Net	**96.45**	**92.27**	81.77	**78.95**	72.56	**68.13**	**63.98**

Table 10: Experimental results from different methods on the CottonCultivar dataset with 5-way 5-shot setting and 60 classes for the base session.

Methods	Sessions (%)				
	1	2	3	4	5
CEC [36]	18.33	13.33	10.00	9.33	6.25
SSFE-Net	**27.50**	**18.46**	**20.00**	**17.33**	**16.25**

variance properties of the ultra-FGVC datasets, SSFE-Net makes use of the strong SSL feature embeddings and can better locate the discriminative areas, which is beneficial for slowing down performance drop in the novel sessions. The detailed capture ability of the proposed method can be verified by the CAM graph in the supplementary materials. Besides, the SSL model helps the model easily transfer the similar feature distribution and knowledge learned from one class to another, which also improves the generalisation ability of the model and benefits the novel session training.

However, compared with prior art FSCIL methods, the SSFE-Net requires a pre-trained SSL model and slightly more time to extract the detailed information from the images since it needs to generate two sets of feature embeddings for the same image. Besides, due to the fact that the SSL model only enhances the features during the base session, the novel incremental prototypes may suffer from insufficient details and a lack of recognizability with different class prototypes. For example, the proposed SSFE-Net method accuracy is slightly lower than the CEC method in sessions 3 and 5, as shown in Table 9. More detailed failure examples are presented in the supplementary materials Section C. Future works will focus on enhancing the discriminative representations of feature prototypes in the novel sessions and adapting them with the base feature space to further reduce the catastrophic forgetting problem.

7. Conclusion

In this paper, a novel SSFE-Net architecture is proposed to improve the feature extraction ability of the low-capacity network backbone in ultra-fine-grained few-shot incremental learning. Specifically, a self-supervised feature enhancement mechanism is developed to extract fine-grained details from the image and achieve great performance on FSCIL tasks. The network utilizes a deep self-supervised learning network to obtain more features from samples and overcomes the bottleneck problem brought by the low capacity network of normal FSCIL. The high-dimensional features from SSL are then used to augment the FSCIL network via knowledge distillation. In addition, the SSFE-Net makes use of the high similarity attribute of different objects in ultra-FGVC tasks and transfers the learning feature from old classes to new classes, which reduces forgetting problems during incremental learning. On top of the proposed model, a series of FSCIL benchmarks are carried out for the first time based on two different ultra-FGVC datasets to facilitate the development of the ultra-FGVC community.

References

[1] Idan Achituve, Aviv Navon, Yochai Yemini, Gal Chechik, and Ethan Fetaya. Gp-tree: A gaussian process classifier for few-shot incremental learning. In *International Conference on Machine Learning (ICML)*, pages 54–65, 2021.

[2] Yuexuan An, Hui Xue, Xingyu Zhao, and Lu Zhang. Conditional self-supervised learning for few-shot classification. In *the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2140–2146, 8 2021.

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9912–9924, 2020.

[5] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *European Conference on Computer Vision (ECCV)*, pages 233–248, 2018.

[6] Da Chen, Yuefeng Chen, Yuhong Li, Feng Mao, Yuan He, and Hui Xue. Self-supervised learning for few-shot image classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1745–1749, 2021.

[7] Kuilin Chen and Chi-Guhn Lee. Incremental few-shot learning via vector quantization in deep embedded space. In *International Conference on Learning Representations (ICLR)*, 2020.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021.

[10] Zhengyu Chen, Jixie Ge, Heshen Zhan, Siteng Huang, and Donglin Wang. Pareto self-supervised training for few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13663–13672, 2021.

[11] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2534–2543, 2021.

[12] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8059–8068, 2019.

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21271–21284, 2020.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 831–839, 2019.

[17] David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015.

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision workshops (ICCVW)*, pages 554–561, 2013.

[19] Mónica G Larese, Ariel E Bayá, Roque M Craviotto, Miriam R Arango, Carina Gallo, and Pablo M Granitto. Multiscale recognition of legume varieties based on leaf venation images. *Expert Systems with Applications*, 41:4638–4647, 2014.

[20] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[21] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 2337–2345, 2021.

[22] Zicheng Pan, Xiaohan Yu, Miaohua Zhang, and Yongsheng Gao. Mask-guided feature extraction and augmentation for ultra-fine-grained visual categorization. In *Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2021.

[23] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2001–2010, 2017.

[24] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *European Conference on Computer Vision (ECCV)*, pages 645–666, 2020.

[25] Yajie Sun, Miaohua Zhang, Xiaohan Yu, Yi Liao, and Yongsheng Gao. A compositional feature embedding and similarity metric for ultra-fine-grained visual categorization. In *Digital Image Computing: Techniques and Applications (DICTA)*, pages 01–08, 2021.

[26] 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)*, pages 12183–12192, 2020.

[27] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, pages 266–282, 2020.

[28] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016.

[29] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[30] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6982–6991, 2020.

[31] Xiaohan Yu, Yongsheng Gao, Mohammed Bennamoun, and Shengwu Xiong. A lie algebra representation for efficient 2d shape classification. *Pattern Recognition*, 134:109078, 2023.

[32] Xiaohan Yu, Yang Zhao, and Yongsheng Gao. Spare: Self-supervised part erasing for ultra-fine-grained visual categorization. *Pattern Recognition*, 128:108691, 2022.

[33] Xiaohan Yu, Yang Zhao, Yongsheng Gao, and Shengwu Xiong. Maskcov: A random mask covariance network for ultra-fine-grained visual categorization. *Pattern Recognition*, 119:108067, 2021.

[34] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan. Patchy image structure classification using multi-orientation region transform. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 12741–12748, 2020.

[35] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Xiaohui Yuan, and Shengwu Xiong. Benchmark platform for ultra-fine-grained visual categorization beyond human performance. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10285–10295, 2021.

[36] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12455–12464, 2021.

[37] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[38] Yang Zhao, Yifan Liu, Chunhua Shen, Yongsheng Gao, and Shengwu Xiong. Mobilefan: Transferring deep hidden representation for face alignment. *Pattern Recognition*, 100:107114, 2020.

[39] Yang Zhao, Chunhua Shen, Xiaohan Yu, Hao Chen, Yongsheng Gao, and Shengwu Xiong. Learning deep part-aware embedding for person retrieval. *Pattern Recognition*, 116:107938, 2021.

[40] Yang Zhao, Xiaohan Yu, Yongsheng Gao, and Chunhua Shen. Learning discriminative region representation for person retrieval. *Pattern Recognition*, 121:108229, 2022.

[41] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[42] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5871–5880, 2021.

[43] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6801–6810, 2021.