CLE-ViT: Contrastive Learning Encoded Transformer for Ultra-Fine-Grained Visual Categorization

Xiaohan Yu[1] , **Jun Wang**[2] and **Yongsheng Gao**[1]

[1]School of Engineering and Built Environment, Griffith University, Australia
[2]Department of Computer Science, University of Warwick, UK
{xiaohan.yu, yongsheng.gao}@griffith.edu.au, jun.wang.3@warwick.ac.uk

Abstract

Ultra-fine-grained visual classification (ultra-FGVC) targets at classifying sub-grained categories of fine-grained objects. This inevitably requires discriminative representation learning within a limited training set. Exploring intrinsic features from the object itself, *e.g.*, predicting the rotation of a given image, has demonstrated great progress towards learning discriminative representation. Yet none of these works consider explicit supervision for learning mutual information at instance level. To this end, this paper introduces CLE-ViT, a novel contrastive learning encoded transformer, to address the fundamental problem in ultra-FGVC. The core design is a self-supervised module that performs self-shuffling and masking and then distinguishes these altered images from other images. This drives the model to learn an optimized feature space that has a large inter-class distance while remaining tolerant to intra-class variations. By incorporating this self-supervised module, the network acquires more knowledge from the intrinsic structure of the input data, which improves the generalization ability without requiring extra manual annotations. CLE-ViT demonstrates strong performance on 7 publicly available datasets, demonstrating its effectiveness in the ultra-FGVC task. The code is available at https://github.com/Markin-Wang/CLEViT

1 Introduction

Ultra-fine-grained visual categorization (ultra-FGVC) distinguishes a sub-category of images from a single fine-grained category. As an emerging topic, ultra-FGVC demonstrates potential in artificial intelligence agriculture and smart farming *e.g.*, automatic crop cultivar classification and plant disease classification [Yu *et al.*, 2020; Larese *et al.*, 2014; Yu *et al.*, 2023]. The intrinsic challenge of ultra-FGVC lies in that very limited samples are provided due to the granularity further moving down to a sub-category level [Huang and Li, 2020]. Another key observation is that the visual variances are difficult to distinguish among different classes, while intra-class variances can be very large (see Figure 2 as an example). Thus how to learn discriminative representation within limited training samples becomes a core question in ultra-FGVC.

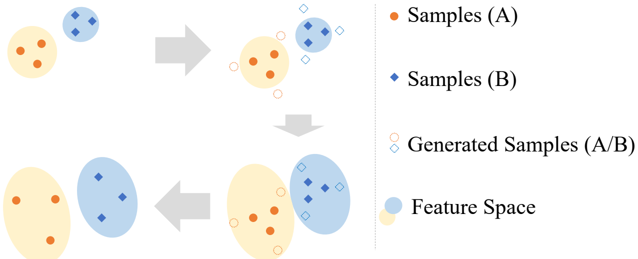


Figure 1: An example of illustrating the process of learning more generalized feature space via incorporating intra-class variations. By generating more diversified samples (from left to right in the first row), the model is required to reshape the feature space to be more tolerant to larger intra-class variations (top to bottom in the right column). This enables a more generalized feature space to adapt to new (testing) samples.

Recent progress has been made by incorporating self-supervised learning to jointly optimize the objectives of representation learning. This is achieved by training with some predefined pretext tasks to drive the model to better understand the intrinsic feature of the data itself. For instance, learning a representation by training a model to predict the rotation [Gidaris *et al.*, 2018] or spatial context [Doersch *et al.*, 2015] of input images. Several works [Yu *et al.*, 2023; Yu *et al.*, 2022] have demonstrated such a strategy can lead to significant performance gain. However, none of these works consider explicit supervision for learning mutual information at **instance** level. This motivates us to introduce self-supervised instance-level contrastive learning to gain a more discriminative representation via understanding mutual information between augmented views of a single image.

In this paper, we introduce CLE-ViT, a novel contrastive learning encoded transformer, to address the intrinsic challenges of ultra-FGVC. The core design is a self-supervised module that performs self-shuffling and masking and then distinguishes these altered image views from other images. This drives the model to learn an optimized feature space that has a large inter-class distance while remaining tolerant to intra-class variations. By incorporating this self-supervised module, the network acquires more knowledge from the in-

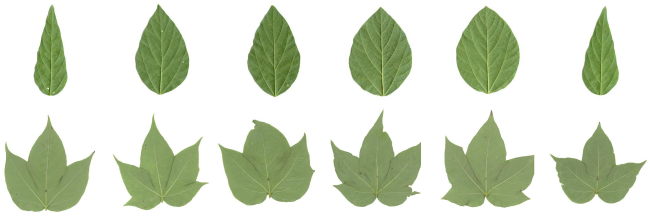


Figure 2: An example of illustrating large intra-class variations in ultra-fine-grained image datasets. The top row shows six images from the same category in the SoyLocal dataset. The bottom row shows six images from the same category in the Cotton80 dataset.

trinsic structure of the input data, thus improving the generalization ability with limited training samples.

CLE-ViT achieves superior (or comparable) performance on five ultra-fine-grained datasets, one plant disease dataset, and the CUB-200-2011 dataset. With promising performance on crop cultivar classification and plant disease classification, the proposed CLE-ViT may provide a promising solution to pushing forward the progress in smart farming. The main contributions of this paper are summarized as follows:

1) We introduce CLE-ViT, a novel vision transformer model that performs instance-level contrastive learning towards discriminative classification with limited training data.

2) CLE-ViT learns an optimized feature space that has a large inter-class distance while remaining tolerant to intra-class variance.

3) CLE-ViT achieves strong performance on five ultra-fine-grained datasets and two fine-grained datasets, demonstrating its effectiveness for ultra-FGVC.

2 Related Work

Ultra-fine-grained visual categorization (ultra-FGVC) identifies objects at a very fine granularity where even humans feel difficult to accurately describe the visual difference. In comparison with fine-grained visual categorization (FGVC), ultra-FGVC has two unique properties/challenges: 1) the annotations are not labeled by human experts or volunteers but obtained from genetic source bank [Yu *et al.*, 2021b]; 2) the classification granularity has moved from species level (FGVC) to a subordinate level, *i.e.*, cultivar level [Yu *et al.*, 2020]. An example of comparing ultra-FGVC and FGVC is shown in Figure 2. FGVC aims to distinguish between images from the top row and those from the bottom row, while ultra-FGVC classifies different images from a single row. This challenging research topic is attracting increasing attention for its significant potential in artificial intelligence agriculture and smart farming.

Earlier exploration started with a small ultra-fine-grained image dataset, which contains 600 images of 100 soybean cultivars [Yu *et al.*, 2020]. Despite encouraging performance on this challenging dataset, their proposed feature modeling method requires a manually segmented vein structure that is inherently difficult to use for practical applications. Recently, [Yu *et al.*, 2021b] released a benchmark platform with baseline performances of 13 state-of-the-art CNN methods on a large-scale ultra-fine-grained cultivar leaf dataset including in total of 47,114 leaf images from two plant species and 3,526 different cultivars. [Yu *et al.*, 2021a] proposed a random mask covariance network (MaskCOV) to learn discriminative representation for ultra-FGVC. The MaskCOV randomly shuffles and masks out image patches, and then predicts the original position of each patch via a self-learning module. Despite its state-of-the-art performance on the ultra-FGVC tasks, the MaskCOV together with all the baseline methods in [Yu *et al.*, 2021b] are all CNNs.

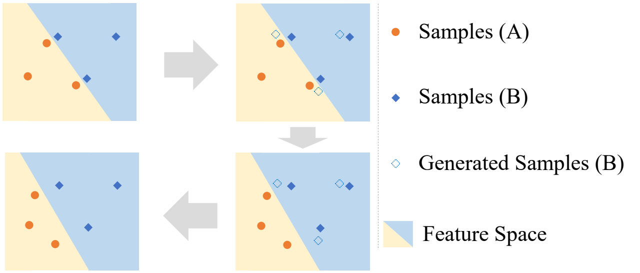


Figure 3: An example of enhancing the generalization capability by enlarging inter-class distance. By generating more diversified samples (from left to right in the first row), the model reshapes the feature space to ensure these samples are aligned to their ground-truth category (top to bottom in the right column). Thus a more separated feature space is formed which better adapts to new (testing) samples.

Transformer [Vaswani *et al.*, 2017] was first developed on natural language processing and is now gaining increasing attention due to its effectiveness in extensive computer-vision tasks. Especially the Vision transformer (ViT) [Dosovitskiy *et al.*, 2020] which adopted a pure transformer directly to deal with sequences of image patches, has now demonstrated very competitive performance in image classification. Yet ViT-based methods require a large-scale dataset for model pretraining. To that end, [Touvron *et al.*, 2021] introduced DeiT, that employed a teacher-student strategy to speed up ViT training. Transformer models were further applied to other popular computer vision tasks. TransFG and FFVT [He *et al.*, 2022; Wang *et al.*, 2021a] proposed to use a few important tokens for final classification and explored ViT in the context of fine-grained visual classification. More recently, Mix-ViT [Yu *et al.*, 2023] and SPARE [Yu *et al.*, 2022] both introduce predefined pretext tasks as a supervision signal for implicit self-supervised learning in ultra-FGVC. Mix-ViT predicts the position of mixed tokens. SPARE classifies masked semantic part regions. None of them consider explicit supervision for learning mutual information at the stance level. In contrast, we develop explicit supervision for learning mutual information at the instance level. Our proposed self-supervised instance-level contrastive learning enables a desirable feature space that has a large inter-class distance while remaining tolerant to intra-class variance.

3 Methods

3.1 Overview & Motivation

Overview. Figure 4 illustrates the overall framework of the proposed CLE-ViT. A given image is first projected into two different views via the following operations: 1) standard aug-

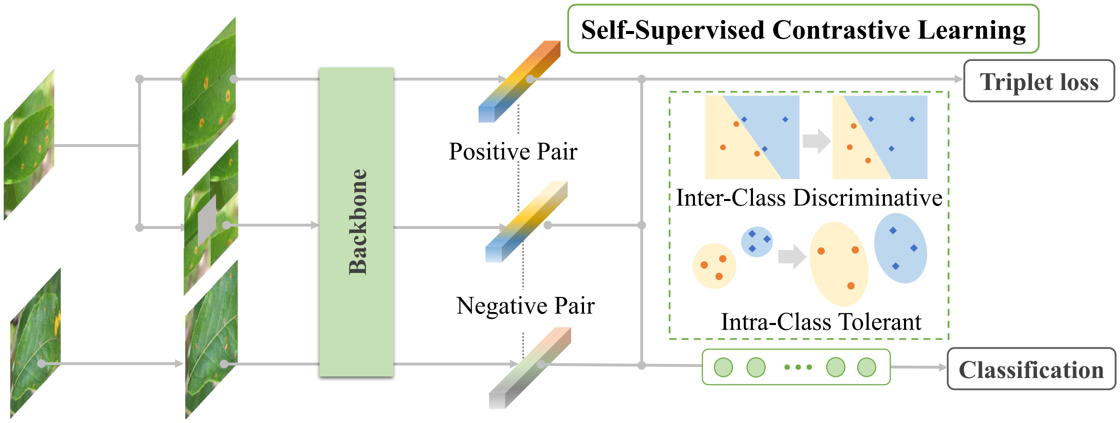


Figure 4: Overview of the proposed method. A given image is first projected into two different views via the following operations: 1) standard augmentation and 2) random shuffling and masking. Then the two views are sent to the backbone network for feature extraction. In addition to the standard classification optimization, their output features form a positive pair, while remaining images from the same batch form negative samples for self-supervised contrastive learning.

mentation and 2) random shuffling and masking. Then the two views are sent to the backbone network for feature extraction. In addition to the standard classification optimization, their output features form a positive pair, while remaining images from the same batch form negative samples for self-supervised contrastive learning. The whole network is trained in an end-to-end manner. The self-supervised module is detached in the inference stage.

Motivation. A desirable learned feature space should be tolerant to intra-class variations. Once the model is overfitting the training samples from the same category, the associated feature space of this class may be closely clustered (as demonstrated in Figure 1). This may hurt the generalization capability as the new (testing) samples may not fall into such a concentrated cluster due to the intrinsic intra-class variations in ultra-fine-grained samples (see Figure 2). To that end, we propose to conduct the contrastive learning on instance-level, instead of the class level in previous works [He *et al.*, 2022; Wang *et al.*, 2021b]. On one hand, this creates more diversified training samples to enable a more intra-class variance-tolerant feature space such that the training process becomes less likely to overfit. On the other hand, the diversified samples may also enlarge the feature space distance between different categories (see Figure 3).

3.2 Image Classification

Given an input image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$ and its associated one-hot category label $\boldsymbol{y} \in \mathbb{R}^{1\times N_c}$, we firstly employ a feature extractor, *e.g.*, ResNet50 [Simonyan and Zisserman, 2015] and Swin Transformer [Liu *et al.*, 2021], to obtain its patch features $\boldsymbol{V} \in \mathbb{R}^{N_p\times D}$ and global feature representation $\boldsymbol{u} \in \mathbb{R}^{1\times D}$ which is used to undertake the category categorization via the classification head. Note that the H, W, N_c, N_p, D are the height, and width of the image, the number of total classes in the datasets, the number of patches in the final stage, and the feature dimension of the final global features respectively. This process can be expressed as:

$$\{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_i, ..., \boldsymbol{v}_{N_{p-1}}, \boldsymbol{v}_{N_p}\} = f_{ife}(\boldsymbol{I}), \tag{1}$$

$$\boldsymbol{p} = \sigma(Head(\boldsymbol{v}_g)), \boldsymbol{u} = \frac{1}{N_p}\sum_{i=0}^{1} \boldsymbol{v}_i \tag{2}$$

Where σ denotes the softmax function and $\boldsymbol{p}$ is the probability distribution given by the backbone model. f_{ife} refers to the image feature extractor. After obtaining the category prediction, the model is normally optimized by a Cross-Entropy loss in an end-to-end manner:

$$L_{cls} = -\frac{1}{N_c}\sum_{i=1}^{N_c} \boldsymbol{y}_i \cdot log(\boldsymbol{p}_i). \tag{3}$$

3.3 Instance-level Contrastive Learning

The proposed instance-level contrastive learning module is trained in a self-supervised manner as it is formed without the need for any extra manual label information. When there are only limited training samples, incorporating such self-supervised tasks can improve the representation learning without requiring extra annotations.

Positive Pair Construction

The anchor image is obtained by applying the standard data augmentation methods, *e.g.*, Random Horizontal Flip, and Random Rotation to the input image. To form the positive sample for the anchor image, we first follow the same standard data augmentation methods as the anchor image to generate the base image, and perform a strong data augmentation, i.e., randomly mask out a proportion of pixels. Specifically, given the anchor image $\boldsymbol{I}_a \in \boldsymbol{R}^{H\times W\times 3}$, the process of obtaining the positive sample $\boldsymbol{I}_{p^*} \in \mathbb{R}^{H\times W\times 3}$ can be formulated by:

$$\boldsymbol{I}_{p^*}(i,j) = \begin{cases} \boldsymbol{I}_a(i,j) & if(i,j) \in \boldsymbol{H} \\ 0 & if(i,j) \notin \boldsymbol{H} \end{cases}, \tag{4}$$

$$\boldsymbol{H} = \{\ (i,j) \mid l < i <= l+k, m < j <= m+t\}, \tag{5}$$

, where (i,j) denotes the position index in the images and $\boldsymbol{H}$, is the set containing all the pixel positions to be masked.

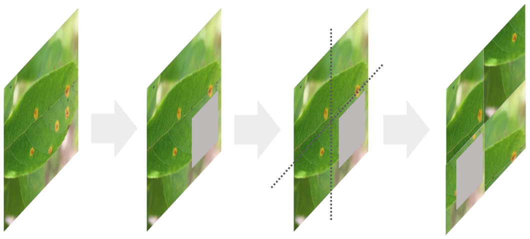


Figure 5: An example of illustrating masked and shuffled samples in positive image pair generation.

k and t determine the masked size and are randomly selected. These parameters are controlled by a term $\alpha = \frac{k*t}{H*W}$, the proportion of masked region.

However, the model can easily infer the original information of masked pixels from the anchor image containing the masked regions within minor transformation in standard augmentation. To this end, we propose to randomly shuffle the positive image to enhance the difficulty in not only acquiring the masked information from the anchor image but also recognizing the category of the image. In particular, we evenly split the image into $n = s * s$ part where s is an integer larger than 1, then the masked image can be reformulated as:

$$\boldsymbol{I}_{p^*} = \begin{bmatrix} P_{(1,1)} & P_{(1,2)} & \cdots & P_{(1,s)} \\ P_{(2,1)} & P_{(2,2)} & \cdots & P_{(2,s)} \\ \vdots & \vdots & \ddots & \vdots \\ P_{(s,1)} & P_{(s,2)} & \cdots & P_{(s,s)} \end{bmatrix}, \tag{6}$$

where we represent the $P(i, j)$ as the image patch in i^{th} row and j^{th} column. The final positive sample $\boldsymbol{I}_p$ is obtained by randomly shuffling the n parts in $\boldsymbol{I}_{p^*}$. An example of generating $\boldsymbol{I}_p$ is shown in Figure 5.

Negative Pair Construction

Since we aim to establish instance-level contrastive learning, the negative sample can be any other sample in the same batch. In detail, for each sample, we randomly pick up one **anchor** image (excluding itself) from the batch as its negative sample $\boldsymbol{I_n}$ to form the negative pair $< \boldsymbol{I}_a, \boldsymbol{I}_n >$.

Instance-level Contrastive Learning

After obtaining the positive pair $< \boldsymbol{I}_a, \boldsymbol{I}_p >$ and negative pair $< \boldsymbol{I}_a, \boldsymbol{I}_n >$, the next problem is how to conduct the contrastive learning. Here, we employ the triplet loss as our contrastive learning metric. Specifically, let's denote the global features of the anchor image, its positive sample and negative sample as $\boldsymbol{u}_a, \boldsymbol{u}_p$ and $\boldsymbol{u}_n$ respectively, our instance-level triplet loss is formulated as:

$$\boldsymbol{L}_{icl} = \frac{1}{B}\sum_{i=1}^{B} \max(\sigma(\boldsymbol{u}_a^i - \boldsymbol{u}_p^j) - \sigma(\boldsymbol{u}_a^i - \boldsymbol{u}_n^j) + \beta, 0), \tag{7}$$

where B is the number of samples in one batch and i denotes the i^{th} sample in the batch. β refers to the margin to control the difficulty in contrastive learning. σ is the L_2 normalization operation. Note that our proposed self-supervised approach can be added to any triple losses.

Granularity	Dataset	#Class	#Train	#Test
Ultra-fine-grained	Cotton80	80	240	240
	SoyLocal	200	600	600
	SoyGene	1,110	12,763	11,143
	SoyAgeing	198	4,950	4,950
	SoyGlobal	1,938	5,814	5,814
Fine-grained	CUB-200-2011	200	5,994	5,794
	Apple Foliar disease	4	1,366	455

Table 1: Statistics of the benchmark datasets.

3.4 Objective Function

The model is jointly optimized by the Cross-Entropy Loss (Equation 3) and the proposed instance-level contrastive loss (Equation 7). We denote the Cross-Entropy Loss on anchor images as L_{clsa} and its positive images as L_{clsp}. Then, the final objective function can be expressed as:

$$L_{fnl} = L_{clsa} + \lambda L_{clsp} + \gamma L_{icl}, \tag{8}$$

where λ and γ are the hyper-parameters to control the contribution among the classification loss on anchor and positive images, and the self-supervised loss.

3.5 Discussion

Why instance-level contrastive learning? To enable instance-level contrastive learning, the first step is to perform in-image augmentation to generate instance-level positive pairs. Such in-image augmentation creates more diversified samples such that the training process is less likely to overfit. More importantly, distinguishing different views of the same sample from other samples can reshape the learned feature space to have a larger distance between different categories and within the same category. As such, the learned feature space is more generalized to adapt to new samples. Moreover, we also perform class-level contrastive learning as an ablation study for a comprehensive evaluation (will be discussed in Section 4.4).

Why triplet loss? Triplet loss and infoNCE loss [Oord *et al.*, 2018] share a similar spirit of separating positive samples from negative samples. Yet infoNCE loss treats each negative sample as a unique category which is less applicable in instance-level setting given a number of more than 10K instances (categories) for some benchmarks. In addition, applying triplet loss focuses on optimizing the distance between positive samples and negative samples rather than urging the model to map the positive pairs to the same points in InfoNCE, thus can avoid overfitting and enhance generalization capability.

4 Experiments

4.1 Datasets & Benchemark Methods

Following [Yu *et al.*, 2023], five ultra-fine-grained image datasets are adopted for evaluation including Cotton80, SoyLocal, SoyGene, SoyAgeing and SoyGlobal. Moreover, two fine-grained datasets, Apple Foliar disease dataset [Thapa *et al.*, 2020] and CUB-200-2011 (CUB) [Wah *et al.*, 2011] are also used to further verify the effectiveness of the proposed

Method	Backbone	Top 1 Accuracy (%)					
		Cotton	S.Loc	S.Gene	S.Age	S.Glo	A.F.
Alexnet [Krizhevsky *et al.*, 2012]	Alexnet	22.92	19.50	13.12	44.93	13.21	95.16
VGG-16 [Simonyan and Zisserman, 2015]	VGG-16	39.33	39.33	63.54	70.44	45.17	95.60
ResNet-50 [He *et al.*, 2016]	ResNet-50	52.50	38.83	70.21	67.15	25.59	94.73
SimCLR (FT) [Chen *et al.*, 2020a]	ResNet-50	51.67	37.33	62.68	64.73	42.54	93.63
SimCLR (L) [Chen *et al.*, 2020a]	ResNet-50	41.25	29.17	29.62	46.18	13.48	82.86
MoCo v2 (FT) [Chen *et al.*, 2020b]	ResNet-50	45.00	32.67	56.49	59.13	29.26	96.04
MoCo v2 (L) [Chen *et al.*, 2020b]	ResNet-50	30.42	27.67	26.58	38.26	12.99	85.49
BYOL (FT) [Grill *et al.*, 2020]	ResNet-50	52.92	33.17	60.65	64.75	41.35	96.04
BYOL (L) [Grill *et al.*, 2020]	ResNet-50	47.92	25.50	35.13	49.53	18.44	87.03
Cutout (8) [DeVries and Taylor, 2017]	ResNet-50	55.83	37.67	61.12	65.70	47.06	94.95
Cutout (16) [DeVries and Taylor, 2017]	ResNet-50	54.58	31.67	62.46	63.68	44.65	94.95
Hide and Seek [Singh and Lee, 2017]	ResNet-50	48.33	28.00	61.27	60.48	23.74	96.26
ADL (0.5) [Choe and Shim, 2019]	ResNet-50	43.75	34.67	55.19	61.70	39.35	96.04
ADL (0.25) [Choe and Shim, 2019]	ResNet-50	40.83	28.00	52.18	51.56	29.50	94.51
Cutmix [Yun *et al.*, 2019]	ResNet-50	45.00	26.33	66.39	62.68	30.31	93.19
DCL [Chen *et al.*, 2019]	ResNet-50	53.75	45.33	71.41	73.19	42.21	94.73
MaskCOV [Yu *et al.*, 2021a]	ResNet-50	58.75	46.17	73.57	75.86	50.28	95.82
SPARE [Yu *et al.*, 2022]	ResNet-50	60.42	44.67	79.41	75.72	56.45	96.70
ViT [Dosovitskiy *et al.*, 2020]	Transformer	52.50	38.83	53.63	66.95	40.57	96.48
DeiT [Touvron *et al.*, 2021]	Transformer	54.17	38.67	66.80	69.54	45.34	96.26
TransFG [He *et al.*, 2022]	Transformer	54.58	40.67	22.38	72.16	21.24	97.14
Hybrid ViT [Dosovitskiy *et al.*, 2020]	Transformer&ResNet	50.83	37.00	71.74	73.56	18.82	96.48
Mix-ViT [Yu *et al.*, 2023]	Transformer&ResNet	60.42	**56.17**	**79.94**	76.30	51.00	97.36
Proposed Method	Transformer	**63.33**	47.17	78.50	**82.14**	**75.21**	**97.58**

Table 2: The classification accuracies on the benchmark datasets. The results of the best-performing method are in boldface, while the second-best performances are underlined. Here Cotton represents Cotton80, S.Local represents SoyLocal, S.Gene represents SoyGene, S.Age represents SoyAgeing, S.Glo represents SoyGlobal and A.F. represents Apple Foliar disease. L and FT indicates linear and fine-tuning evaluation, respectively.

method. Table 1 summarizes the statistics of benchmark datasets, *i.e.*, the numbers of classes, training images, and testing images. For fair comparisons, the proposed CLE-ViT is compared with the same 17 benchmark methods as adopted in [Yu *et al.*, 2023] for comprehensive evaluations.

4.2 Implementation

Model Details. We adopt the Swin Transformer Base (Swin-B) [Liu *et al.*, 2021] as our backbone model by taking both precision and efficiency into consideration. The same as the most transformer-based works [Yu *et al.*, 2023; Touvron *et al.*, 2021], our backbone is initialized by the ImageNet21K [Deng *et al.*, 2009] pre-trained model. The proportion of the masked region and the number of parts n and are set to $[0.15, 0.45]$ and 4 respectively. The margin β in Equation 7 is 1. λ and γ are both set to 1 for all datasets except 0.3 and 0.5 for CUB dataset.

Training and inference. Following the [He *et al.*, 2022; Touvron *et al.*, 2021; Wang *et al.*, 2021a], input images are first resized to 600×600 for all datasets. Random (Center) cropping is then applied to crop the images into 448×448 during the training (inference) phase. After that, we adopt random horizontal flipping, color jitter, and random rotation during the training. The standard augmentation consists of the aforementioned transformations. The whole architecture is optimized by AdamW optimizer. In our experiment settings, the batch size and the learning rate are set to 12 and 1e-3 for all the datasets.

4.3 Comparison to The State-of-The-Arts

Evaluation on ultra-fine-grained image datasets. Table 2 lists the classification accuracy of all the competing methods and their backbone networks on five ultra-fine-grained datasets. The proposed CLE-ViT achieves 75.21% classification accuracy on the SoyGlobal dataset, outperforming other competing methods with a significant margin (more than 19%). We also observe a similar trend in SoyAgeing, where the proposed method surpasses other competing methods with a margin of more than 5% of classification accuracy. Among the ultra-fine-grained image datasets, the SoyAgeing dataset covers five subsets and each subset contains images collected from a specific cultivating stage. The comparison results of all competing methods on the five subsets are summarized in Table 3. The proposed method achieves strong performance compared with other competing methods, demonstrating its effectiveness in ultra-FGVC tasks.

Evaluation on fine-grained datasets. we present evaluation results in Table 4 of the proposed CLE-ViT on the widely used fine-grained image dataset, CUB-200-2011 [Wah *et al.*, 2011]. CLE-ViT achieves competitive performance (ranked 2nd) among the state-of-the-art methods on CUB-200-2011, demonstrating its effectiveness and generalization capability in fine-grained visual classification. To further verify the effectiveness of the proposed method, we evaluate a plant disease classification image dataset, the Apple Foliar disease dataset. The comparison results are summarized in Table 2. The proposed method achieves the best classification accuracy of 97.58% on the Apple Foliar disease dataset.

Method	Backbone	Top 1 Accuracy (%)					
		R1	R3	R4	R5	R6	Avg
Alexnet [Krizhevsky *et al.*, 2012]	Alexnet	49.90	44.65	45.15	47.47	37.47	44.93
VGG-16 [Simonyan and Zisserman, 2015]	VGG-16	72.32	72.53	74.95	71.11	61.31	70.44
ResNet-50 [He *et al.*, 2016]	ResNet-50	70.00	64.24	74.04	72.63	54.85	67.15
SimCLR (L) [Chen *et al.*, 2020a]	ResNet-50	53.64	45.66	45.35	50.40	35.86	46.18
SimCLR (FT) [Chen *et al.*, 2020a]	ResNet-50	70.00	66.57	64.24	68.38	54.44	64.73
MoCo v2 (L) [Chen *et al.*, 2020b]	ResNet-50	42.93	38.59	38.99	38.99	31.82	38.26
MoCo v2 (FT) [Chen *et al.*, 2020b]	ResNet-50	62.73	56.16	61.31	65.96	49.49	59.13
BYOL (L) [Grill *et al.*, 2020]	ResNet-50	55.35	48.38	50.40	49.60	43.94	49.53
BYOL (FT) [Grill *et al.*, 2020]	ResNet-50	71.11	66.16	65.76	64.65	56.06	64.75
Cutout (16) [DeVries and Taylor, 2017]	ResNet-50	70.20	61.92	62.32	69.70	54.24	63.68
Cutout (8) [DeVries and Taylor, 2017]	ResNet-50	66.87	64.04	67.78	73.43	56.36	65.70
Hide and Seek [Singh and Lee, 2017]	ResNet-50	64.04	58.99	61.31	64.75	53.33	60.48
ADL (0.25) [Choe and Shim, 2019]	ResNet-50	53.54	54.34	55.15	52.83	41.92	51.56
ADL (0.5) [Choe and Shim, 2019]	ResNet-50	66.67	58.89	64.75	68.48	49.70	61.70
Cutmix [Yun *et al.*, 2019]	ResNet-50	65.56	59.19	64.24	68.79	53.64	62.28
DCL [Chen *et al.*, 2019]	ResNet-50	76.87	73.84	76.16	76.16	62.93	73.19
MaskCOV [Yu *et al.*, 2021a]	ResNet-50	79.80	74.65	79.60	78.28	66.97	75.86
SPARE [Yu *et al.*, 2022]	ResNet-50	78.28	79.90	78.69	77.27	64.44	75.72
ViT [Dosovitskiy *et al.*, 2020]	Transformer	69.29	64.55	70.40	71.01	59.49	66.95
DeiT [Touvron *et al.*, 2021]	Transformer	73.03	70.40	69.09	74.65	60.51	69.54
TransFG [He *et al.*, 2022]	Transformer	74.95	74.55	74.24	76.26	60.81	72.16
Hybrid ViT [Dosovitskiy *et al.*, 2020]	Transformer&ResNet	77.17	76.97	74.75	76.36	62.53	73.56
Mix-ViT [Yu *et al.*, 2023]	Transformer&ResNet	79.29	77.17	77.98	79.19	67.88	76.30
Proposed Method	Transformer	**80.81**	**83.33**	**84.24**	**86.36**	**75.96**	**82.14**

Table 3: The classification accuracies of the competing methods on the five subsets of the SoyAgeing dataset. "Avg" denotes the average classification accuracy of the five subsets. The results of the best-performing method are in boldface, while the second-best performances are underlined. L indicates linear evaluation. FT denotes fine-tuning evaluation.

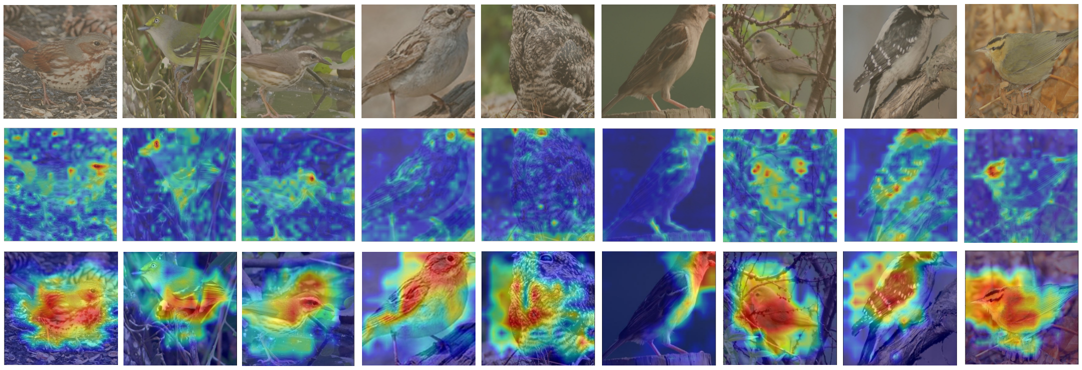
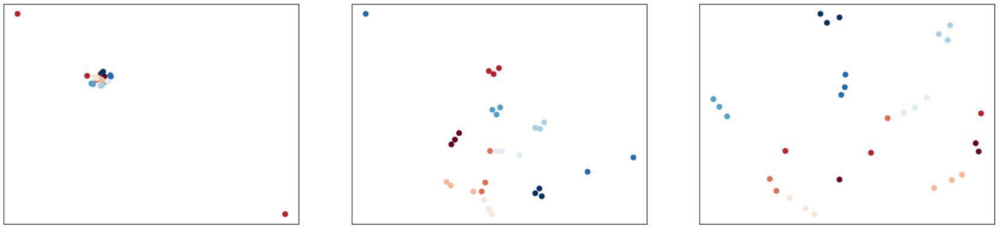


Figure 6: Visual comparison of Mix-ViT (middle) and the proposed CLE-ViT (bottom) with original images (top) on CUB dataset.



(a) Baseline

(b) Baseline+Class Contrastive

(c) Ours

Figure 7: tSNE visualization of learned features from (a) Baseline and (b) Baseline + Class Contrastive and (c) Ours on the SoyGlobal dataset. Each color indicates a unique category (10 categories in total). In comparison with Baseline and Baseline+Class Contrastive, the proposed method learns a more optimized feature space that has a large inter-class distance while remaining tolerant to intra-class variance.

Method	Backbone	Accuracy (%)
PMG [Du *et al.*, 2020]	ResNet50	89.6
DCL [Chen *et al.*, 2019]	ResNet-50	87.8
MaskCOV [Yu *et al.*, 2021a]	ResNet-50	86.6
SPARE [Yu *et al.*, 2022]	ResNet-50	86.8
API-Net [Zhuang *et al.*, 2020]	DenseNet161	90.0
StackedLSTM [Ge *et al.*, 2019]	GoogleNet	90.4
TransFG [He *et al.*, 2022]	Transformer	91.7
DeiT [Touvron *et al.*, 2021]	Transformer	90.0
ViT [Dosovitskiy *et al.*, 2020]	Transformer	90.6
Mix-ViT [Yu *et al.*, 2023]	Transformer	91.0
Proposed Method	Transformer	91.2

Table 4: The classification accuracies on the CUB-200-2011.

Visualization. We visualize the attention maps of Mix-ViT [Yu *et al.*, 2023] and the proposed CLE-ViT on the CUB dataset [Wah *et al.*, 2011] in Figure 6, where highlighted regions contribute significantly to the visual categorization. We observe that the attentive regions of birds consistently cover unique patterns, verifying the reliability of the effectiveness. This is consistent with human observations and common sense that visual features from those unique patterns are vital for the determination of bird species.

4.4 Ablation Study & Analysis

Role of the self-supervised contrastive learning module. To further verify the contribution of the proposed self-supervised contrastive learning module, we present a comprehensive ablation on 4 benchmark datasets. The baseline removes the self-supervised module from the CLE-ViT. In addition, contrastive learning can also be used at class level, *i.e.*, all images from the same category are used to form positive pairs while remaining images from other categories are negative samples. For a comprehensive evaluation, we replace the instance-level contrastive learning with a standard class-level contrastive learning, denoted as Baseline+Class Contrastive (Baseline+CC). The comparison results of baseline, baseline+CC, and CLE-ViT are shown in Figure 8. We observe that CLE-ViT consistently improves the performance over the baseline method and Baseline+CC, verifying the effectiveness of the proposed self-supervised learning module.

Feature space. Figure 7 shows tSNE visualization of learned features from baseline, baseline+class contrastive and ours. Here the samples are randomly selected from 10 categories (testing set). We observe that the clusters from the baseline are not well separable. The tSNE result from Baseline+Class Contrastive shows a larger inter-class distance compared with baseline while the intra-class distance remains small. This may hurt the generalization ability given that samples in ultra-fine-grained image datasets often have large intra-class variance, *e.g.*, the point in the top left corner 7 (b). In contrast, the proposed CLE-ViT shows both large inter-class distance and intra-class distance, indicating a better generalization capability.

Ablation study on L_{clsp}. Table 5 shows an ablation study of L_{clsp} on Soy.Loc dataset by varying the weight from 0 to 1.5 at a step size of 0.5. Here λ partially balances the contribution from instance-level contrastive learning and category-level learning. $\lambda = 0$ means positive samples are unable to receive category-level supervision. Without category-level supervision, contrastive learning can still optimize the distance at the instance-level but might be unable to predict the category label of positive samples. As λ increases, the contribution from category-level supervision becomes larger and achieves the best performance when $\lambda = 1$. But when λ becomes too large, the category-level supervision will completely dominate the training while ignoring the contribution from contrastive learning, thus may lead to overfitting and performance drop.

λ	0.0	0.5	1.0	1.5
Accuracy(%)	43.67	44.00	**47.17**	43.00

Table 5: The ablation study of the weight λ of CE loss for positive samples on Soy.Loc dataset.

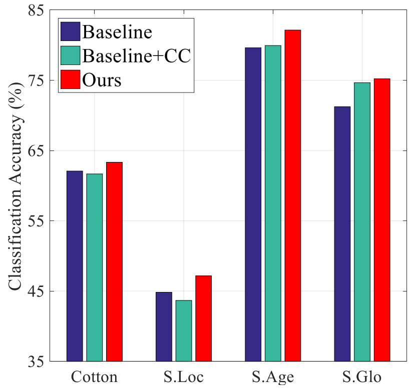


Figure 8: Ablation study on ultra-fine-grained image datasets.

5 Conclusion

This paper introduced a novel contrastive learning encoded vision transformer, CLE-ViT, to address intrinsic challenges in ultra-fine-grained visual categorization. A new self-supervised learning module has been proposed, which drives the model to learn an optimized feature space that has a large inter-class distance while remaining tolerant to intra-class variations. By incorporating this self-supervised module, the network acquires more knowledge from the intrinsic structure of the input data, which improves the generalization ability of limited training samples. CLE-ViT has achieved competitive performance on seven public datasets, demonstrating its effectiveness in ultra-FGVC, birds, and plant disease classification tasks.

We also observe that negative samples are randomly sampled in batches for better efficiency, thus potentially introducing easy pairs. A promising direction in future work is to explore efficient hard sample mining approaches when forming negative pairs.

Contribution Statement

Xiaohan Yu and Jun Wang contributed equally to this paper.

References

[Chen *et al.*, 2019] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *CVPR*, pages 5157–5166, 2019.

[Chen *et al.*, 2020a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[Chen *et al.*, 2020b] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[Choe and Shim, 2019] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, pages 2219–2228, 2019.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[DeVries and Taylor, 2017] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[Doersch *et al.*, 2015] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[Du *et al.*, 2020] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *ECCV*, pages 153–168. Springer, 2020.

[Ge *et al.*, 2019] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *CVPR*, pages 3034–3043, 2019.

[Gidaris *et al.*, 2018] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[Grill *et al.*, 2020] 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. In *NeurIPS*, volume 33, pages 21271–21284, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2022] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *AAAI*, volume 36, pages 852–860, 2022.

[Huang and Li, 2020] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *CVPR*, pages 8662–8672, 2020.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[Larese *et al.*, 2014] Mónica G Larese, Rafael Namías, Roque M Craviotto, Miriam R Arango, Carina Gallo, and Pablo M Granitto. Automatic classification of legumes using leaf vein image features. *Pattern Recognition*, 47(1):158–168, 2014.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Singh and Lee, 2017] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, pages 3544–3553, 2017.

[Thapa *et al.*, 2020] Ranjita Thapa, Noah Snavely, Serge Belongie, and Awais Khan. The plant pathology 2020 challenge dataset to classify foliar disease of apples. *arXiv preprint arXiv:2004.11958*, 2020.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICLR*, pages 10347–10357, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report CNS-TR-2011-001*, 2011.

[Wang *et al.*, 2021a] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization. In *BMVC*, 2021.

[Wang *et al.*, 2021b] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 943–952, 2021.

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

[Yu *et al.*, 2021a] 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.

[Yu *et al.*, 2021b] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Xiaohui Yuan, and Shengwu Xiong. Benchmark platform for ultra-fine-grained visual categorization beyond human performance. In *ICCV*, pages 10285–10295, 2021.

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

[Yu *et al.*, 2023] Xiaohan Yu, Jun Wang, Yang Zhao, and Yongsheng Gao. Mix-vit: Mixing attentive vision transformer for ultra-fine-grained visual categorization. *Pattern Recognition*, 135:109131, 2023.

[Yun *et al.*, 2019] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019.

[Zhuang *et al.*, 2020] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. In *AAAI*, volume 34, pages 13130–13137, 2020.