Down-Sampling Inter-Layer Adapter with Self-Supervised Adaptation Warmup for Parameter and Computation Efficient Ultra-Fine-Grained Image Recognition

EDWIN ARKEL RIOS*, Institute of Electronics, National Yang Ming Chiao Tung University, Taiwan

FEMILOYE OYERINDE*, Cohere for AI Community, Nigeria

FERNANDO MIKAEL, Computer Science Department, National Tsing Hua University, Taiwan

OSWIN GOSAL, Computer Science Department, National Tsing Hua University, Taiwan

MIN-CHUN HU, Computer Science Department, National Tsing Hua University, Taiwan

BO-CHENG LAI, Institute of Electronics, National Yang Ming Chiao Tung University, Taiwan

Ultra-fine-grained image recognition (UFGIR) involves classifying objects with extremely subtle inter-class differences, such as distinguishing between cultivars within the same species under conditions of limited per-class samples. In this work, we extend our previous Inter-Layer Adapter (ILA) framework by integrating a Self-Supervised Adaptation Warmup (SAW) stage that leverages a Dual-Attention-driven Mix Augmentation SupCon (DAMAS) strategy. Instead of fine-tuning the entire Vision Transformer (ViT) backbone, our approach trains only lightweight adapter modules, thereby achieving significant reductions in both the number of trainable parameters and FLOPs. Comprehensive experiments conducted on 5 diverse UFGIR datasets demonstrate that our extended method substantially improves recognition accuracy—yielding average gains of up to 11% at higher resolutions—while maintaining a highly efficient computational profile. These results underscore the effectiveness of our self-supervised warmup in bridging the modality gap between generic pretraining and the fine-grained domain, making our approach particularly suitable for deployment in resource-constrained environments.

CCS Concepts: • **Computing methodologies** → **Object recognition**.

*Both authors contributed equally to this research.

Authors' addresses: Edwin Arkel Rios, edwinarkelrios.ee08@nycu.edu.tw, Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan; Femiloye Oyerinde, oyerindefemiloye@gmail.com, Cohere for AI Community, Abuja, Nigeria; Fernando Mikael, fm113065427@gapp.nthu.edu.tw, Computer Science Department, National Tsing Hua University, Hsinchu, Taiwan, 300044; Oswin Gosal, oswingosal02@gapp.nthu.edu.tw, Computer Science Department, National Tsing Hua University, Hsinchu, Taiwan, 300044; Min-Chun Hu, anitahu@cs.nthu.edu.tw, Computer Science Department, National Tsing Hua University, Hsinchu, Taiwan, 300044; Bo-Cheng Lai, bclai@nycu.edu.tw, Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan.

ACM XXXX-XXXX/2018/3-ART
https://doi.org/XXXXXXX.XXXXXXX

Additional Key Words and Phrases: Vision transformer, parameter-efficient transfer learning, fine-tuning, fine grained visual analysis, object categorization, image classification, transformer, efficient

ACM Reference Format:

Edwin Arkel Rios, Femiloye Oyerinde, Fernando Mikael, Oswin Gosal, Min-Chun Hu, and Bo-Cheng Lai. 2018. Down-Sampling Inter-Layer Adapter with Self-Supervised Adaptation Warmup for Parameter and Computation Efficient Ultra-Fine-Grained Image Recognition. 1, 1 (March 2018), 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Ultra-fine-grained image recognition (UFGIR) addresses classification at an exceptionally granular level, such as distinguishing between cultivars of a plant species. While fine grained recognition tackles sub-categories (e.g., bird species), UFGIR operates within macro- categories where inter-class variations are subtle and intra-class diversity is pronounced [45] as seen from Fig. 1. Applications span agriculture [27, 32], medicine [34], and industry [28], but challenges persist due to limited labeled data and the inadequacy of conventional vision backbones to capture discriminative features [51].

Fig. 1. Examples of ultra-fine-grained image classes, illustrating the subtle inter-class variations and pronounced intra-class diversity characteristic of UFGIR tasks.

Prior ultra-fine-grained recognition approaches typically augment coarse-grained backbones [43, 46–50] with task-specific modules [14, 43] or metric-learning losses [12, 49] to amplify discriminative features. While these strategies improve sensitivity to inter-class variations in standard FGIR, they struggle with UFGIR's extreme granularity due to their reliance on local receptive fields and handcrafted feature interactions.

Vision Transformers (ViTs) [11] have emerged as promising backbones for UFGIR, leveraging global self-attention [42] to aggregate nuanced features [20, 36, 40]. However, full fine-tuning of large ViTs is impractical for multi-task deployment. Parameter- efficient transfer learning (PETL) methods [4, 17, 23, 41] mitigate this by freezing most parameters, yet our analysis reveals a critical limitation in UFGIR: frozen ViTs suffer from attention collapse [9, 44, 54], where deeper layers produce highly similar features as observed from Fig. 4. This undermines their ability to resolve ultra-fine distinctions.

Previous PETL approaches for generic FGIR underperform in UFGIR settings, even when specialized FGIR modules are selectively tuned (PEFGIR), and the attention collapse issue still persists as seen in Fig. 4. In our prior work, we introduced the Inter-Layer Adapter (ILA) [38] to address this issue by inserting lightweight adapters between frozen ViT layers. The ILA module uses dual spatial downsampling to enforce a hierarchical feature extraction that alleviates attention collapse and reduces background overfitting.

However, while ILA significantly improves performance, a modality gap remains: the frozen encoder is pretrained on generic tasks (e.g., ImageNet classification), which do not fully capture the localized feature demands of UFGIR. Previous methods have attempted to bridge this gap via self-supervised pretraining [12, 46, 47], yet they often overlook the importance of cross-dataset diversity during pretraining and data-aware augmentations to resolve ultra subtle visual distinctions. [39] employed scaled-rank adapters with an intermediate pretraining phase, where adapters are first optimized to align representations with the target domain.

Our attention rollout analysis (see Fig. 2) further reveals that vanilla ViTs and previous adapter methods, including ILA tend to overfit to background regions, thereby failing to focus on discriminative parts of the image. To address these challenges, we propose an extension to ILA by integrating a Self-Supervised Adaptation Warmup (SAW) stage. SAW pretrains the adapters using a supervised contrastive objective, combined with a dual-attention-driven mix augmentation strategy on a diverse corpus of ultra-fine-grained data, and subsequently fine-tunes on the target dataset. This two-stage adaptation not only further mitigates attention collapse but also aligns adapter features with the pretrained backbone, enhancing the extraction of subtle, discriminative features essential for UFGIR.

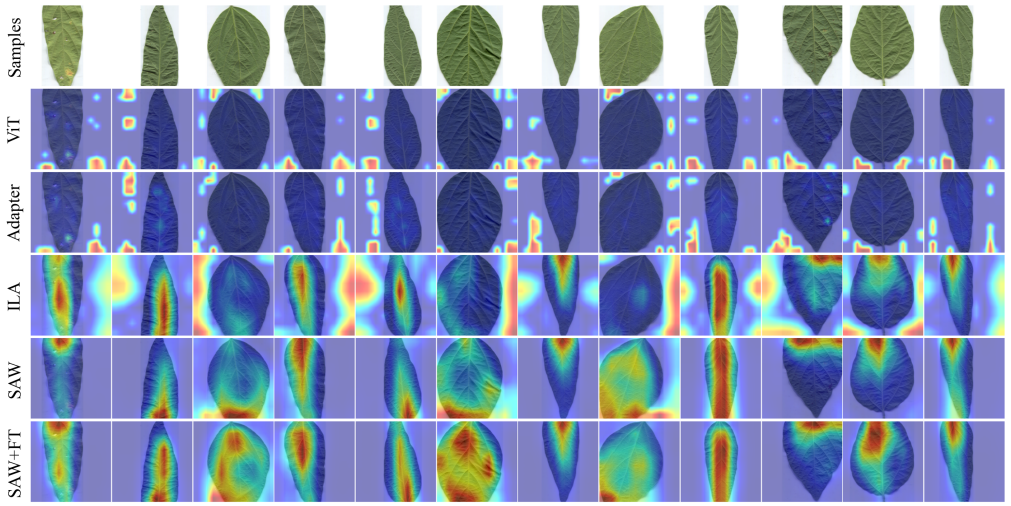


Fig. 2. Attention rollouts from layers 8 and 12 for the SoyAgeing dataset. The figure demonstrates that vanilla ViTs and conventional adapter methods tend to overfit to background regions, whereas our SAW extension yields more focused attention maps that emphasize discriminative image regions.

The contributions of our previous work are as follows:

(1) We proposed a novel ILA module to address the attention collapse problem in frozen ViTs for UFGIR tasks. The ILA employs dual spatial downsampling branches to aggregate discriminative features while reducing computational cost.
(2) We conducted comprehensive experiments across ten UFGIR datasets, comparing over 15 state-of-the-art methods. Our ILA framework achieved outstanding classification performance and computational efficiency, increasing average accuracy by at least 6.8% compared to other parameter-efficient methods, while requiring 123× fewer trainable parameters and 30% fewer FLOPs than current SOTA UFGIR methods.

And we susbtantially extend ILA with the following contributions:

(1) We propose Dual-Attention-driven Mix Augmentation SupCon (DAMAS), which leverages early-layer attention maps to guide region mixing in contrastive learning. This dual-attention strategy improves efficiency while providing robust guidance for data-aware augmentation
(2) Our extensive ablation studies across diverse backbones demonstrate the effectiveness of the proposed components and reveal the accuracy–cost trade-off. By adjusting the adapter's downsampling rate and insertion location, we provide valuable insights for optimizing performance in ultra-fine-grained recognition tasks.

There should be two points regarding experiment contributions. One with more general results highlighting how the method with SAW pushes accuracy even further (by how much compared to our previous SotA) and then a separate one for the additional ablations we did (backbone, design of ILA which show the versatility and how to obtain a better accuracy vs cost trade-off for our method))

2 RELATED WORK

2.1 Ultra Fine-Grained Image Recognition

Ultra fine-grained image recognition (UFGIR) remains a challenging task due to subtle inter-class differences and limited labeled data. To address these challenges, many methods build on generic recognition backbones and augment them with modules or loss functions that effectively select and aggregate discriminative features [14, 43, 48, 49]. For instance, approaches such as FFVT [43] leverage Vision Transformer (ViT) attention scores to extract and aggregate intermediate low-, medium-, and high-level features that capture minute inter-class variations and are aggregated through the last transformer encoder block.

Given the data constraints inherent in UFGIR, recent research has increasingly explored self-supervised techniques and tailored data augmentation strategies [5, 50, 51]. Methods such as MaskCOV, SPARE, and Mix-ViT utilize these strategies to learn intrinsic, fine-grained details from scarce data. In parallel, contrastive learning approaches exemplified by CLE-ViT [46] and CSDNet [12], which also incorporates self-knowledge distillation [7, 16, 53] have been employed to further enhance the extraction of discriminative features. However, most of these methods employ ViT backbones with large number of parameters that all need to be stored for deployment and also spend significant resources during training.

2.2 Parameter-Efficient Transfer Learning

Parameter-efficient transfer learning (PETL) techniques seek to fine-tune only a small subset of a model's parameters while keeping the majority of the pretrained backbone fixed. These methods generally fall into two categories: prompt-tuning and adapter-based approaches.

Prompt-tuning [22, 29] augments the input sequence with additional task-specific, learnable tokens appended at various stages of the transformer. For instance, VQT [41] employs these tokens as queries to aggregate layer-wise information, which is then integrated into the classification head. However, the incorporation of extra tokens not only increases the computational cost during the forward pass but can also lead to a rapid escalation in the number of parameters within the classification head.

In contrast, adapter-based methods insert lightweight, non-linear modules directly within transformer layers. Originally proposed by Houlsby et al. [17] in the natural language processing

domain, adapters have been successfully adapted for vision transformers. For example, ConvPass [23] extends this concept by integrating 2D convolutions to inject spatial biases, thereby enhancing the extraction of fine-grained features. However, these existing methods PETL methods do not incorporate inductive bias such as features downsampling, which impacts performance under the data scarcity of ultra-fine-grained image tasks.

2.3 Self-Supervised Learning

Self-supervised learning (SSL) is widely used for representation learning and can be broadly categorized into generative and discriminative approaches. The foundational contrastive loss [33] pulls positive pairs closer and pushes negative pairs apart in the embedding space. Early approaches like SimCLR [8] generate positive pairs via generic data augmentations and requires a large batch size. SupCon [24] instead redefines positive pairs by using class labels: for an anchor image $\mathbf{x}_i$ with label y_i, all instances $\{\mathbf{x}_j \mid y_j = y_i\}$ in the batch are treated as positives. This label-aware approach reduces dependence on large batch sizes and improves representation quality. Other methods, such as PID [6] treats each image as a unique class and uses augmentations like CutMix [52] and MultiCrop, and SelfCon [3], which generates multiple views via auxiliary sub-networks to contrast with the backbone features, both pose a challenge as they introduce more parameters to the network. Notably, SupCon is the most efficient of these methods due to its label awareness which makes it possible to leverage data-aware augmentations, and most importantly, it does not introduce any extra parameters.

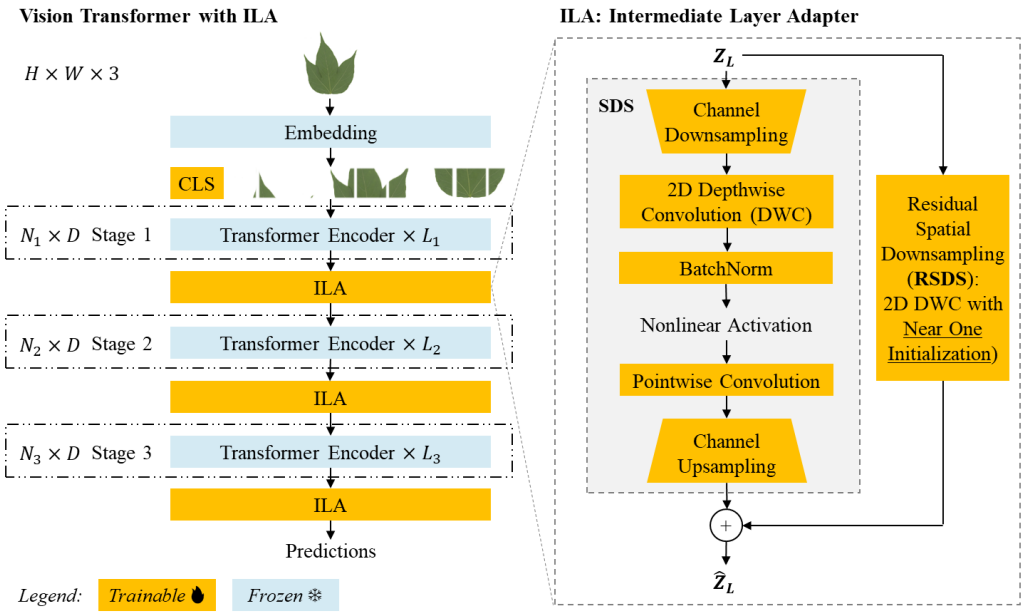
```
This section should tell me briefly what is SSL and what is the problem
with current SSL and ideally how our method addreses this (DADA)
  Generic data augmentation, fine-grained details -> DADA
```

2.4 Data-Aware Data Augmentation

In fine-grained image recognition (FGIR) where limited data and subtle inter-class variations pose significant challenges, it is non-trivial to employ data augmentation methods to improve performance. Early methods relied on basic transformations such as cropping or rotating the image [cites], which increase data diversity but often fail to capture fine-grained details. However, recent approaches have shifted toward data-specific techniques. For example, part-based augmentation [cite] and SnapMix [21] exploit the inherent structure of the data to guide augmentations, while WSDAN [19] employs attention from all layers of a Vision Transformer (ViT) for attention-guided cropping and dropping, albeit at high computational cost. To mitigate this, our method uses attention rollout from only the early layers of the ViT, demonstrating that these layers already attend to the most discriminative regions. This strategy achieves effective data-aware augmentation with significantly reduced computational overhead.

3 METHOD

Figure 3 illustrates our proposed method. We build on our Inter-Layer Adapter (ILA) framework, which mitigates attention collapse by enforcing hierarchical feature extraction via lightweight adapters inserted between frozen ViT layers. In this work, we extend ILA with a Self-Supervised Adaptation Warmup (SAW) stage. During SAW, the adapter modules are pretrained using a self-supervised contrastive objective on a diverse set of ultra-fine-grained datasets. This warmup aligns the adapter features with the frozen backbone's representations and enhances the extraction of subtle, discriminative cues. After this adaptation phase, the model is fine-tuned on the target dataset,



Mention 3 stage spatial hierarchy and shorten caption. Add description

Fig. 3. Overview of ViT with our proposed Intermediate Layer Adapter (ILA). Trainable modules are shown in orange while frozen ones are shown in blue. An image is embedded into tokens and forwarded through a series of transformer encoder blocks, which we divide into three groups. After the first two encoder groups the sequence is passed through the ILA. After passing through all the encoder blocks the CLS token is forwarded through a classification head to obtain predictions. In the ILA tokens are forwarded through two spatial downsampling (SDS) branches. In the main SDS branch (highlighted as a grey box) tokens are first downsampled channel-wise and then spatially downsampled through the usage of a 2D depth-wise convolution. The sequence is then forwarded through a BatchNorm layer, a non-linear activation, and a point-wise convolution, before being up-sampled channel-wise. To allow for residual gradient flow we also forward the tokens through a Residual Spatial Downsampling (RSDS) branch implemented as a 2D depth-wise convolution initialized with values near one. Initializing the kernel to values near one allows the RSDS to behave as a learnable identity or pooling function. Then, the outputs of the dual SDS branches are added together and forwarded to the next encoder group.

and the refined CLS token is used for final classification. This integrated approach effectively bridges the modality gap while maintaining computational efficiency.

3.1 Vision Transformer Encoder

Rename sequence length N_0 to N_1 and redefine it as N_1 = F_h_1 ^^2 where F_h_1 = H / P (assuming H=W) and reference it later in the ILA section when describing N_i+1 = F_h_i+1 ^ 2 = F h i - (KS - 1)

The Vision Transformer (ViT) encoder processes input images by first patchifying them into fixed-size patches via a convolution with kernel size P. These patches are then flattened and linearly embedded into D-dimensional tokens, resulting in a sequence of length $N_0 = \frac{h}{P} \times \frac{w}{P}$, where h and w denote the image height and width, respectively. A learnable [CLS] token [10] is appended to the sequence, and positional embeddings are added to encode spatial relationships. The resulting token sequence is then processed through a series of transformer encoder blocks, each comprising multi-head self-attention (MHSA) and position-wise feed-forward networks (PWFFN) [42]. The output of

each block is represented as $\mathbf{z}_l \in \mathbb{R}^{N_l \times D}$. Finally, a LayerNorm [2] and a linear classification layer are applied to produce the final predictions.

3.2 ViT Attention Collapse

Recent studies have revealed that deep Vision Transformers (ViTs) often experience attention collapse [44, 54], where self-attention maps become increasingly uniform across layers. This over-smoothing effect, as described in [13], results in the loss of high-frequency information and limits the diversity of the learned representations. Such a collapse is especially detrimental for ultra-fine-grained image recognition (UFGIR), where capturing subtle inter-class differences is crucial. In parameter-efficient transfer learning (PETL) settings, where only the adapter weights are fine-tuned and the ViT backbone remains frozen, we observe that attention collapse is further exacerbated, as illustrated in Fig. 4. The resulting uniformity in attention maps in deeper layers restricts the extraction of discriminative features, highlighting the need for specialized adaptation strategies.

3.3 Inter-Layer Adapter

You need to describe that we are inspired by previous methods (cite Swin, PVT) to incorporate hierarchy (suggest you to copy the section from the ECCV version, page 5 on ILA

To alleviate the attention collapse issue, we employ lightweight downsampling adapters inserted between the layers of the vision transformer to enforce the heirachical extraction of ultra-fine grained features. The Inter-Layer-Adapter module consists of two branches: the Spatial Downsampling Branch (SDS) and the Residual Spatial Downsampling (RSDS) branch.

3.3.1 ***Spatial Downsampling Branch (SDS).*** In the SDS branch, the input feature map $\mathbf{z}_l$ at layer l undergoes channel downsampling (CDS), followed by a depthwise separable convolution (DWConv) for spatial downsampling, and finally channel upsampling (CUS). This design, inspired by [23], incorporates two key modifications: (i) the use of depthwise separable convolution [18] to improve efficiency, and (ii) the omission of padding, so that the spatial resolution is reduced during convolution. These changes not only lower the computational cost but also enhance attention diversity by enforcing a strict hierarchy in the feature maps. The overall operation of the SDS branch is summarized by:

$$\mathbf{z}_1 = W_{\text{CDS}}^T \mathbf{z}_l + b_{\text{CDS}}, \tag{1}$$

where $\mathbf{z}_l$ represents the input feature map at layer l. As in our previous work, the ILA module is interleaved between intermediate layers of the Vision Transformer encoder [11], with only the adapters being trained while the encoder remains frozen.

For each channel c and spatial location (i, j), the depthwise convolution is defined as:

$$\mathbf{z}_2(c, i, j) = \sum_{p=0}^{K_h-1} \sum_{q=0}^{K_w-1} \mathbf{z}_1(c, i+p, j+q)\, k_c(p, q), \tag{2}$$

where $k_c(p, q)$ denotes the kernel weight for channel c. This output is then normalized, activated by the GELU function, and processed by a pointwise convolution (PWConv) to mix channel information. We denote this combined operation as:

$$\mathbf{z}_2 = \text{PWConv}\Big(\text{GELU}\big(\text{BN}(\text{DWConv}(\mathbf{z}_1))\big)\Big). \tag{3}$$

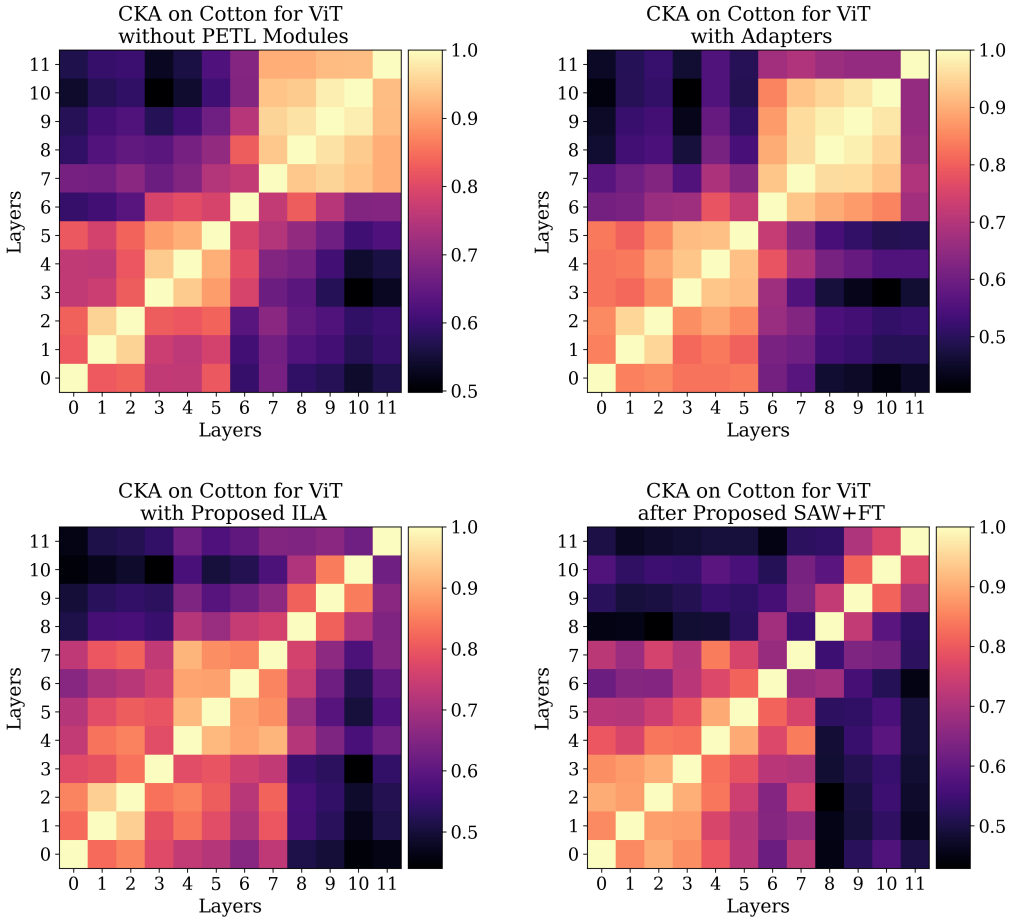


Fig. 4. Centered Kernel Alignment (CKA) similarity [26] between attention layers of a ViT for the vanilla ViT (left) and ours (right). Lighter colors indicate higher similarity.

and finally, the channel upsampling (CUS) takes $\mathbf{z}_2$ as input to give $\mathbf{m}$

$$\mathbf{m} = W_{\text{CUS}}^{T}\,\mathbf{z}_2 + b_{\text{CUS}}, \tag{4}$$

3.3.2 ***Residual Spatial Downsampling (RSDS) branch.*** To compensate for the reduction in spatial dimensions introduced by the SDS branch, we integrate a learnable residual connection via the RSDS branch, which facilitates the smooth flow of information between layers and mitigate the risk of vanishing gradients within the network [15]. Instead of fixed pooling or interpolation—which may discard crucial local details—we utilize a 1D depthwise convolution (DWC) with kernel weights initialized near one. For each channel d and spatial position n, the residual output is computed as:

$$r_{d,n} = \sum_{k=0}^{K-1} z_l^{d,n+k} \cdot W_{d,k}, \qquad d = 0, 1, \ldots, D; \quad n = 0, 1, \ldots, N, \tag{5}$$

which, for a kernel size $K = 1$ and near-one initialization, approximates the identity mapping:

$$r_{d,n} \approx z_l^{d,n}, \qquad d = 0, 1, \ldots, D; \quad n = 0, 1, \ldots, N. \tag{6}$$

This learnable residual connection acts as an adaptive gate, allowing the network to modulate the contribution of the original features, and for larger kernels, it functions similarly to a sum-pooling operation.

3.4 Self-Supervised Adaptation Warmup (SAW)

From Figure 2 with Dual-Attention guided Mix Augmentation Supcon (DAMAS)

From Sec. 1 (see Fig. 2), we show that previous PETL methods, including ILA fail to focus on the discriminative parts of the image, but overfit to the background regions. This motivates us to propose a Self-Supervised Adaptation Warmup (SAW) stage that integrates into the ILA architecture to bridge the data modality gap [39] and mitigate attention collapse in frozen Vision Transformers for ultra-fine-grained tasks [44, 54]. Unlike the original ILA, which only employs a dual spatial downsampling strategy to enforce hierarchical feature extraction [23], our extended method first trains the adapters discriminatively on a diverse collection of 5 ultra-fine-grained datasets using a supervised contrastive objective [24], and then fine-tunes on the target dataset. As shown in Fig. 4, incorporating this adaptation warmup stage results in deeper transformer layers that capture more distinct and discriminative features, in contrast to the collapsed representations observed with the frozen backbone alone.

Remind us of the issue with why did we incorporate the data-aware mix augmentation

1) Overview of the method: mix augmentation + supcon
2) Why did we incorporate mix? issues with generic augs for UFGIR
3) Explain how does the mix aug work (fig 5)
4) Why do we incorporate the early dual-attention selection: single-attention or full attention rollout
5) Propose a new method based on dual-attention selection
6) Highlight the advantages: lower cost, similar visualization (Fig 6)

3.4.1 ***Supervised Contrastive Learning.*** We use the supervised contrastive loss [24] with a data-aware augmentation strategy that cuts and swaps discriminative regions between image pairs to generate label-aware positive samples. In the stage of pretraining, we pool across 5 ultra-fine-grained task datasets to increase the data diversity.

$$\mathcal{L}_{\text{sup}} = \sum_{i \in I} -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}, \tag{7}$$

where $P(i)$ denotes indices of positives distinct from i, $A(i)$ includes all samples except i, and τ controls similarity sharpness

3.4.2 ***Early Attention Data Aware Augmentation.*** In Vision Transformers (ViTs), self-attention captures relationships among image patches, and attention rollout aggregates these weights across layers to yield a global view of the model's focus. We leverage attention rollout to identify the most discriminative regions of an image. These regions are then swapped between image pairs to create positive samples for data-aware augmentation.

Performing full attention rollout over all L layers has high computational complexity, typically $O(L \cdot n^2 \cdot d)$ or equivalently $O(L \cdot S^3)$ when expressed in terms of the attention matrix $S \in \mathbb{R}^{n \times n}$, where n is the number of patches and d is the hidden dimension. To mitigate this cost, we propose a dual rollout approach that computes global attention flow using only two layers. Specifically, we extract the first row of S (which corresponds to the [CLS] token) as it effectively summarizes the entire image context. This strategy reduces the complexity by a factor of $L \times S$ while still guiding the augmentation process efficiently.

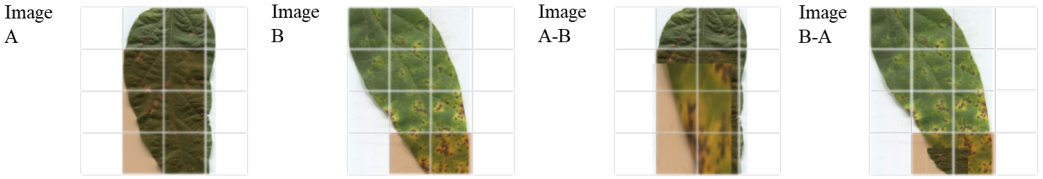


Fig. 5. Illustration of the self-aware data augmentation process. We identify discriminative regions in the original leaf images (left) and swap these regions between pairs (right), creating more challenging positive samples for ultra fine-grained recognition tasks.

```
Cost: O(LN^3)
```

3.4.3 ***Dual-Attention Layers Selection.*** We study which layers attentions lead to higher accuracy and cost effective. Instead of using attention from a single layer, which often results in noisy or inconsistent attention maps, or more than two layers, which significantly increases computational cost, we find that using exactly two attention layers strikes the best balance. This choice reduces FLOPs by a factor of N, where N is the sequence length, making it much more efficient than deeper attention-based strategies. In Fig. 6 we show the results of selecting different attention layers for augmentation. A more computationally expensive variant of our method applies attention rollout from the first four layers, increasing cost by a factor of N compared to the dual-attention approach. While this method provides more fine-grained information for augmentation selection, the trade-off in efficiency is non-negligible.

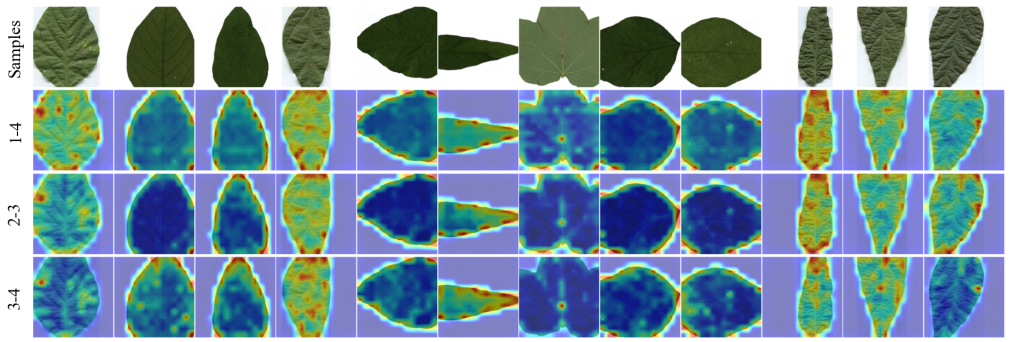


Fig. 6. Results of selecting different attention layers for data-aware augmentation. Using exactly two attention layers strikes an optimal balance between capturing discriminative features and computational efficiency.

3.4.4 ***Data Diversity.*** Ultra fine-grained image recognition (UFGIR) tasks often suffer from limited data per class and a high number of classes, which can hinder a model's ability to capture subtle differences. To address this, we pool all 10 ultra-fine-grained datasets used in our experiments for the adaptation warmup stage. This strategy increases the diversity of training data

4 EXPERIMENTAL METHODOLOGY

In Tab. 1 we describe the datasets used for our experiments. These are ultra-fine-grained leaves datasets collected by Yu et al. [51] where each category represents a confirmed cultivar name attached to the seed obtained from the genetic resource bank.

We conduct our experiments in two stages. First, we perform a learning rate search over the set $\{0.3, 0.1, 0.03, 0.01, 0.003\}$ using a subset of the training data, selecting the learning rate that yields the highest validation accuracy. In the second stage, we train the model on the full training set using the selected learning rate and evaluate performance on the test set over three independent runs. All models are optimized with Stochastic Gradient Descent (SGD) (momentum 0.9), a batch size of 8, and a cosine learning rate scheduler with a 500-step warmup. Training is performed for 50 epochs with automatic mixed precision.

Table 1. Dataset statistics for the ultra-fine-grained leaves recognition experiments. The table lists the number of classes and the distribution of training and testing images for each dataset, including the Cotton dataset, SoyAgeing (and its subsets), SoyGene, SoyGlobal, and SoyLocal. These diverse datasets underscore the varying scales and complexities encountered in UFGIR tasks.

Datasets	Classes	Train Images	Test Images
Cotton	80	240	240
SoyAgeing	198	4950	4950
SoyAgeingR1	198	990	990
SoyAgeingR3	198	990	990
SoyAgeingR4	198	990	990
SoyAgeingR5	198	990	990
SoyAgeingR6	198	990	990
SoyGene	1110	12763	11143
SoyGlobal	1938	5814	5814
SoyLocal	200	600	600

For data preprocessing, images are first resized to 300×300 or 600×600 and then randomly cropped (or center-cropped during inference) to 224×224 or 448×448. All images are horizontally flipped and normalized using the standard ImageNet mean and standard deviation.

We evaluate our methods on five ultra-fine-grained leaves datasets [51], where each category represents a distinct cultivar. Results are reported in terms of top-1 accuracy (percentage) along with standard deviations over three runs. In addition, we assess computational efficiency by measuring total trainable task parameters (TTTP)—the subset of parameters fine-tuned for the task—as well as floating-point operations (FLOPs), throughput (images/second), and the overall number of trainable parameters. This comprehensive evaluation enables us to compare both the performance and resource efficiency of each method.

All experiments employ the ViT-B/16 [11] backbone pretrained on ImageNet-21k, with a patch size of 16, 12 transformer layers, and a hidden dimension of 768. We compare our proposed Self-Supervised Adaptation Warmup (SAW) framework with three variants of our original ILA:

- ILA: Intermediate layer adapter modules with downsampling are inserted after layers 4 and 8.
- ILA^{+}: In addition to ILA, non-downsampling adapters are inserted in all layers except 4 and 8.
- ILA^{++}: Extends ILA by also incorporating traditional intra-layer adapters [17] in every layer.

- SAW: Extends ILA^{++} by integrating a self-supervised adaptation warmup stage, wherein our proposed Dual-Attention-driven Mix Augmentation SupCon (DAMAS) strategy pretrains the adapters on a diverse corpus of ultra-fine-grained data before fine-tuning on the target dataset.

When applicable, the best performance values are highlighted in **bold** and the second-best values are underlined.

Similarly, we compare our SAW framework against 15 state-of-the-art models in the parameter-efficient setting. These methods are organized into three families based on their fine-tuning strategies and feature aggregation mechanisms:

- Linear Classifier (Baseline): the most simple PE method which keeps all backbone parameters frozen and only fine-tunes the classification head.
- Low-Rank Bilinear Pooling (LR-BLP) [25]: employs a low-rank projection to reduce the dimensionality of the bilinearly pooled features.
- Matrix Power Normalized Covariance (MPN-Cov) [30, 31]: applies covariance pooling of high-level features to lessen instabilities of bilinear pooling.
- Intermediate Features Aggregation (IFA) [36]: selects the CLS tokens from intermediate layers and forwards them through a small MLP to first aggregate cross-layer features before outputting classification predictions.

The second category includes FGIR methods where a module is designed to explicitly select features based on some criteria, along with a possible module to aggregate these selected discriminative features. We evaluate these models in the parameter-efficient setting (PEFGIR) where only a small percentage of modules are fine-tuned. It is composed of:

- CAL [35]: employs counter-factuality to train a bilinear attention module which is used for both pooling features and to generate augmented versions of the input images (crops and masked). The fine-tuned components include the attention module and the bilinear attention pooling classification head.
- TransFG [14]: selects features from the previous-to-last layer based on head-wise attention rollout [1], a matrix-multiplication based aggregation of attention scores across layers. We fine-tune the last transformer encoder block where the feature aggregation happens and the linear classification head.
- FFVT [43]: selects and aggregates intermediate features based on layer-wise attention. Same as the previous one: the last transformer encoder block and the classification head.
- RAMS-T [20]: crops the image for data augmentation based on attention rollout [20]. Fine-tunes only the classification head.
- GLSim [37]: computes the similarity between global and local representations of an image to select crops. Fine-tunes an aggregator transformer encoder block and the classification head.

The third category includes dedicated PETL methods as follows:

- VPT-Shallow (VPT-Sh) [22]: appends learnable prompts to the sequence at the start of the transformer.

- VPT-Deep [22]: appends learnable prompts to the sequence before each transformer block and then removes them after each block.
- Visual Query Tuning (VQT) [41]: appends learnable prompts to be used as queries only (not keys or values) to the sequence prior to the MHSA module of each layer. These prompts are expedited towards the classification head where they are concatenated into a single large dimensional linear layer.
- Adapter [17]: incorporates a small MLP inserted after the MHSA and PWFFN of each transformer encoder block.
- ConvPass [23]: similar to the previous it incorporates a small MLP inserted inside the transformer block, but this MLP incorporates a 3×3 convolution in between the channel downsampling and upsampling of the adapter.

5 RESULTS AND DISCUSSION

5.1 Top-1 Accuracy Comparison with State-of-the-Art PETL Methods

5.1.1 ***224×224 Training Results.*** We evaluate our proposed Self-Supervised Adaptation Warmup (SAW) method on 5 ultra-fine-grained datasets using images of size 224×224, employing the standard ViT-B/16 pretrained on ImageNet-21K. As shown in Table 2, SAW consistently outperforms the best ILA variant (ILA++) and other conventional PETL approaches across the evaluated datasets. In particular, SAW achieves an average accuracy improvement of approximately 12% over ILA++, and about 25% higher than leading conventional methods. These improvements underscore the effectiveness of our adaptation warmup stage in aligning adapter features with the frozen backbone, thereby enhancing the extraction of fine-grained, discriminative features while maintaining high parameter efficiency.

Furthermore, Table 3 indicates that on the SoyAgeing R1–R6 subset, SAW yields substantial gains. These findings highlight the robustness and superior performance of SAW at 224×224 resolution.

5.1.2 ***448×448 Training Results.*** We further evaluate our methods at a higher resolution of 448×448. As shown in Table 4, increasing the resolution leads to a substantial improvement in performance: SAW achieves an average top-1 accuracy improvement of approximately 11% across the five datasets against ILA++. In contrast, while methods such as CSDNet [12] that use a self-supervised module to [what does CSDNet do] sample tokens achieve competitive results, CSDNet fine-tunes the entire model, whereas our method employs parameter-efficient transfer learning by training only the adapter. This makes our approach significantly more cost-efficient and reusable.

From Table 5 we show the top-1 accuracy results on the SoyAge R1–R6 datasets at 448×448 resolution. Our proposed SAW method achieves striking performance gains over the baseline and existing PETL methods. For instance, the baseline ViT B-16 model attains accuracies in the range of approximately 55–60%, whereas SAW reaches around 83–84% on most subsets. In particular, SAW outperforms the best ILA variant (ILA++) by roughly 10% on SoyAgeR1.

These results confirm that leveraging higher-resolution inputs and our SAW extension enables the network to capture finer, more discriminative features, yielding a compelling balance between performance and efficiency for ultra-fine-grained recognition.

Table 2. Top-1 accuracy comparison of state-of-the-art PETL methods on five ultra-fine-grained datasets with an image resolution of 224 × 224. Our proposed SAW method achieves the highest accuracy across all datasets, significantly outperforming previous PETL approaches, including ILA++.

method	Cotton	SoyAgeing	SoyGene	SoyGlobal	SoyLocal
ViT B-16	35.42 ± 1.1	44.91 ± 0.07	19.45 ± 0.16	17.88 ± 0.4	28.83 ± 1.04
LR-BLP	30.56 ± 1.73	30.38 ± 0.49	10.73 ± 0.13	5.54 ± 0.21	17.39 ± 1.14
MPNC [30]	39.72 ± 1.34	38.24 ± 0.78	18.26 ± 0.31	17.86 ± 0.19	26.0 ± 1.09
IFA [36]	37.92 ± 2.73	49.84 ± 0.22	25.08 ± 2.19	27.88 ± 0.18	32.67 ± 3.63
TransFG [14]	44.86 ± 0.86	53.52 ± 0.15	32.51 ± 0.15	34.93 ± 0.55	35.83 ± 1.42
FFVT[43]	40.28 ± 2.84	63.88 ± 0.29	40.56 ± 0.17	38.11 ± 0.83	40.78 ± 0.48
CAL CAL[35]	40.0 ± 0.72	51.38 ± 0.61	19.54 ± 0.22	24.47 ± 0.65	31.83 ± 0.84
RAMS[20]	34.44 ± 0.24	47.32 ± 0.91	20.62 ± 0.38	18.63 ± 0.31	27.28 ± 0.19
GLSim[37]	38.89 ± 0.86	53.73 ± 0.2	27.58 ± 0.32	31.42 ± 0.25	28.72 ± 0.69
VQT [41]	39.44 ± 1.58	57.95 ± 0.55	31.14 ± 0.56	24.91 ± 0.18	34.45 ± 0.75
VPT-S [22]	32.5 ± 1.91	46.49 ± 0.62	23.45 ± 0.4	18.63 ± 0.93	26.78 ± 0.79
VPT-D [4]	24.3 ± 8.67	56.12 ± 1.04	35.16 ± 0.73	30.35 ± 1.35	19.06 ± 0.59
ConvP [23]	40.28 ± 2.05	56.01 ± 0.38	43.72 ± 0.72	31.41 ± 0.65	33.72 ± 1.55
Adapter [17]	41.11 ± 1.47	64.95 ± 0.72	44.92 ± 0.37	35.04 ± 1.17	33.33 ± 1.26
ILA[38]	47.78 ± 1.27	56.88 ± 0.52	36.95 ± 0.25	34.37 ± 0.6	38.0 ± 0.88
ILA+[38]	49.72 ± 0.64	60.98 ± 0.29	44.18 ± 0.53	35.21 ± 0.4	37.61 ± 0.35
ILA++[38]	48.33 ± 2.5	67.39 ± 0.86	52.1 ± 0.41	43.48 ± 0.21	41.28 ± 0.98
SAW	**50.28 ± 1.05**	**70.18 ± 0.44**	**59.85 ± 0.17**	**49.62 ± 0.41**	**53.33 ± 0.29**

5.2 Accuracy Vs FLOPs and Throughput Comparison

In Table 6, we report the top-1 accuracy, total trainable task parameters (TTTP), FLOPs, task-deployable parameters (TP), and memory usage for various PETL methods evaluated at 448×448 resolution. Notably, while prior methods can achieve competitive accuracy, they do so at significantly higher computational costs. For example, CSDNet attains a top-1 accuracy of 68.4% but requires 123× more parameters than our SAW, which achieves the highest accuracy of 73.5%. Moreover, by extending ILA with our Self-Supervised Adaptation Warmup (SAW) stage, we improve its accuracy by over 7% without incurring additional TTTP, FLOPs, or memory overhead. From Fig. 7 and Fig. 8 we visually observe how our method improves performance over ILA++ and other existing methods with no increasing cost. These findings confirm that our approach effectively leverages higher-resolution inputs to capture finer details while maintaining a low computational footprint, thereby offering a cost-efficient and reusable solution for ultra-fine-grained image recognition.

Table 3. Top-1 accuracy comparison of state-of-the-art PETL methods on the SoyAgeing dataset across five rounds (R1–R6) with an image resolution of 224 × 224. Our proposed SAW method significantly outperforms previous approaches, including ILA++, demonstrating superior adaptation and accuracy.

method	SoyAgeingR1	SoyAgeingR3	SoyAgeingR4	SoyAgeingR5	SoyAgeingR6
ViT B-16	54.92 ± 0.87	50.74 ± 0.62	52.79 ± 1.09	53.43 ± 0.1	43.47 ± 1.03
LR-BLP	45.19 ± 0.42	38.42 ± 0.31	39.96 ± 1.71	42.29 ± 1.02	33.1 ± 0.38
MPNC [30]	55.32 ± 1.02	50.13 ± 1.61	52.19 ± 0.66	52.02 ± 0.36	39.86 ± 1.54
IFA[36]	60.47 ± 0.92	56.09 ± 0.47	57.37 ± 2.02	61.08 ± 1.04	50.98 ± 1.22
TransFG[14]	65.22 ± 0.61	61.98 ± 0.58	63.87 ± 0.6	64.75 ± 0.56	53.2 ± 0.57
FFVT [43]	65.69 ± 1.96	66.97 ± 0.61	68.69 ± 1.72	69.06 ± 0.15	54.48 ± 0.84
CAL[35]	59.86 ± 0.48	56.7 ± 0.33	53.13 ± 0.46	57.58 ± 0.53	45.39 ± 0.94
RAMS[20]	57.27 ± 0.88	53.64 ± 0.71	55.39 ± 0.46	56.33 ± 0.59	44.75 ± 0.1
GLSim [37]	61.04 ± 0.21	58.96 ± 1.02	60.91 ± 0.44	59.66 ± 0.56	47.24 ± 0.31
VQT[41]	64.01 ± 0.92	59.22 ± 0.25	60.07 ± 1.01	63.54 ± 1.47	51.28 ± 0.26
VPT-S [22]	53.44 ± 1.46	49.03 ± 1.08	49.33 ± 0.36	52.79 ± 1.34	43.33 ± 1.15
VPT-D [22]	59.19 ± 2.54	36.2 ± 2.68	53.36 ± 0.06	55.66 ± 1.39	46.53 ± 1.74
ConvP [23]	64.04 ± 0.4	61.31 ± 0.71	64.01 ± 0.56	67.34 ± 0.65	53.84 ± 0.36
Adapter [17]	66.4 ± 1.26	64.82 ± 1.61	64.75 ± 0.56	67.54 ± 0.32	54.01 ± 1.55
ILA [38]	65.72 ± 0.76	63.0 ± 0.62	65.08 ± 0.23	66.67 ± 0.63	56.43 ± 0.74
ILA+ [38]	69.7 ± 1.05	66.67 ± 0.83	68.69 ± 0.83	69.46 ± 0.76	60.57 ± 0.21
ILA++ [38]	70.81 ± 0.2	69.09 ± 1.21	68.28 ± 0.88	69.7 ± 0.3	61.78 ± 1.11
SAW	**77.41 ± 0.35**	**75.35 ± 0.71**	**78.01 ± 0.85**	**78.18 ± 0.44**	**66.77 ± 1.05**

5.3 Ablations

5.3.1 ***Importance of Downsampling.*** Our ablation studies reveal that enforcing hierarchical feature extraction through spatial downsampling is critical for learning deeper, more discriminative representations. As shown in Table 7, omitting spatial downsampling results in an accuracy drop of approximately 9.2% on SoyGlobal and 8.2% on SoyLocal, underscoring its role in guiding the network to focus on salient, task-relevant features. However, because spatial downsampling reduces the spatial dimensions of the feature maps, it creates a mismatch that prevents the direct application of standard skip connections [15]. To address this, we explored several options and ultimately incorporated a residual downsampling branch—a convolutional layer with near-one initialization designed to approximate an identity function while remaining learnable. This design ensures smooth feature propagation by effectively bridging the shape mismatch. Together, these results demonstrate that both spatial downsampling and an appropriately designed residual skip connection are essential for maximizing performance in ultra-fine-grained recognition.

5.3.2 ***Deciding Downsampling ILA Position.*** The position at which the adapter is inserted into the transformer layers plays a crucial role in performance, as observed in Tab. 8. We evaluate different insertion locations and compare them against ILA++, which places the adapter between layers 3 to 7 of the transformer model. Notably, Loc 4-9 achieves the highest accuracy. However,

Table 4. Top-1 accuracy (%) on ultra-fine-grained datasets at 448×448 resolution. Results are reported as average ± standard deviation. Increasing the resolution from 224x224 to 448x448 leads to significant performance gains for SAW, demonstrating the effectiveness of our method in leveraging higher-resolution inputs for improved fine-grained recognition.

method	Cotton	SoyAgeing	SoyGene	SoyGlobal	SoyLocal
SIM-Trans*	54.58	34.76	15.46	**70.69**	25
CSDNet* [12]	**57.92**	75.39	**70.82**	56.3	46.17
ViT B-16	39.03 ± 0.48	48.61 ± 0.58	21.31 ± 0.45	24.97 ± 0.8	28.72 ± 1.67
MPNC [30]	43.89 ± 1.05	40.43 ± 2.61	20.22 ± 0.43	25.79 ± 1.07	27.94 ± 1.92
IFA [36]	44.58 ± 3.0	56.01 ± 2.14	33.09 ± 0.87	35.82 ± 1.02	29.22 ± 1.99
TransFG [14]	51.67 ± 2.73	57.54 ± 0.82	38.79 ± 0.18	45.35 ± 0.15	38.06 ± 0.79
FFVT [43]	51.94 ± 2.06	69.93 ± 1.03	49.97 ± 0.46	47.7 ± 0.28	42.22 ± 1.3
CAL [35]	44.03 ± 2.92	51.16 ± 0.27	23.03 ± 0.32	34.98 ± 1.39	31.61 ± 1.29
RAMS [20]	38.47 ± 1.47	50.73 ± 0.45	25.83 ± 0.44	25.17 ± 0.87	29.67 ± 0.84
GLSim [37]	45.7 ± 1.69	56.58 ± 0.46	33.04 ± 0.32	39.85 ± 0.31	29.95 ± 0.63
VQT [41]	50.97 ± 2.1	65.65 ± 0.68	36.8 ± 0.17	33.99 ± 1.53	36.33 ± 1.17
VPT-S [22]	38.19 ± 2.51	49.12 ± 0.28	27.43 ± 0.59	27.37 ± 3.77	28.39 ± 0.95
VPT-D [22]	43.19 ± 1.73	60.76 ± 1.03	38.28 ± 0.83	36.83 ± 1.55	25.28 ± 0.69
ConvP [23]	48.33 ± 2.5	60.18 ± 1.64	53.43 ± 1.43	45.13 ± 1.25	34.22 ± 1.3
Adapter[17]	48.19 ± 1.88	73.31 ± 0.81	57.04 ± 0.98	47.27 ± 0.41	36.83 ± 1.09
ILA[38]	51.94 ± 1.68	66.32 ± 0.73	44.65 ± 0.81	46.9 ± 0.24	43.56 ± 0.42
ILA+ [38]	53.33 ± 1.1	68.79 ± 0.63	52.65 ± 0.89	48.29 ± 0.38	46.56 ± 0.63
ILA++ [38]	55.42 ± 1.67	75.0 ± 0.36	62.19 ± 0.41	58.14 ± 0.24	50.83 ± 0.6
SAW	55.42 ± 1.1	**79.65 ± 0.11**	70.53 ± 0.22	64.02 ± 0.56	**58.33 ± 0.67**

since the performance difference between Loc 4-9 and Loc 3-7 is minimal, and Loc 3-7 requires fewer FLOPs, we consider layers 3-7 to be the optimal choice for inserting inter-layer adapters.

5.3.3 ***Backbone SAW ablation.*** We compare the performance of our proposed SAW method across different backbone configurations: ViT, DeiT, and DeiT3 on the SoyGlobal and SoyLocal datasets. The results in Table 9 indicate that SAW consistently outperforms both the baseline and the corresponding adapter and ILA++ variants for each backbone. For example, on ViT, SAW improves the top-1 accuracy from 17.9% (Baseline) to 49.6% on SoyGlobal and from 28.8% to 53.3% on SoyLocal. Similar trends are observed for DeiT and DeiT3, where SAW (DeiT) and SAW (DeiT3) achieve the highest accuracies, demonstrating the robustness of our method regardless of the underlying backbone. These findings underscore that incorporating a self-supervised adaptation warmup leads to significant gains in discriminative feature learning while remaining parameter-efficient.

5.3.4 ***Choosing Convolution Kernel Size.*** In our ILA framework, a depthwise separable convolution is employed to maintain parameter efficiency. We conduct an ablation study on the convolution kernel size to examine its trade-offs with accuracy, FLOPs, and total trainable parameters (TTP). As shown in Table 10, increasing the kernel size tends to boost accuracy on both the SoyGlobal and SoyLocal datasets while simultaneously reducing FLOPs. However, larger kernel

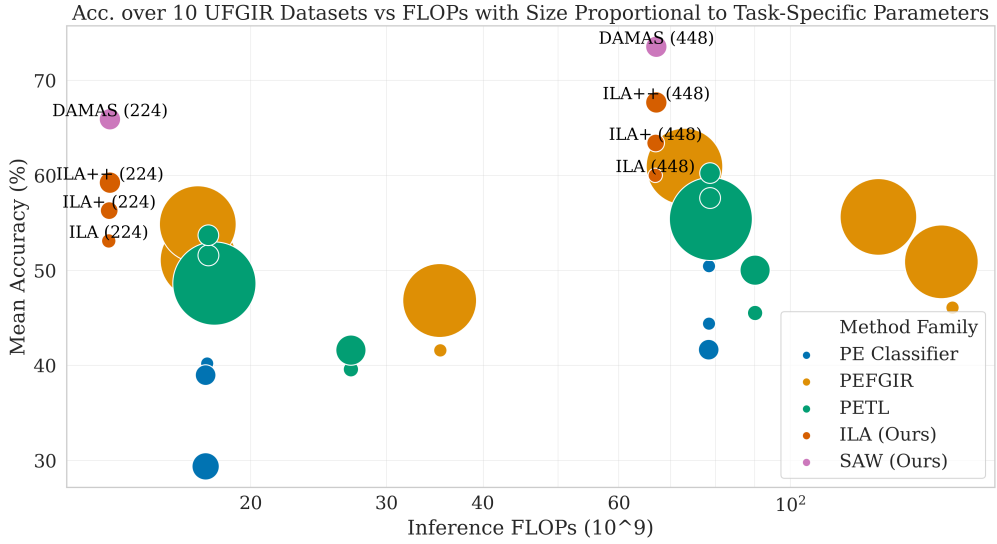


Fig. 7. Performance trade-offs for various methods at 224×224 and 448×448 resolutions. Our proposed DAMAS achieves the highest mean accuracy, nearly 70% at 224×224, and significantly outperforms ILA++ while using the same number of parameters. At higher resolution, DAMAS further improves accuracy with only a moderate increase in FLOPs.

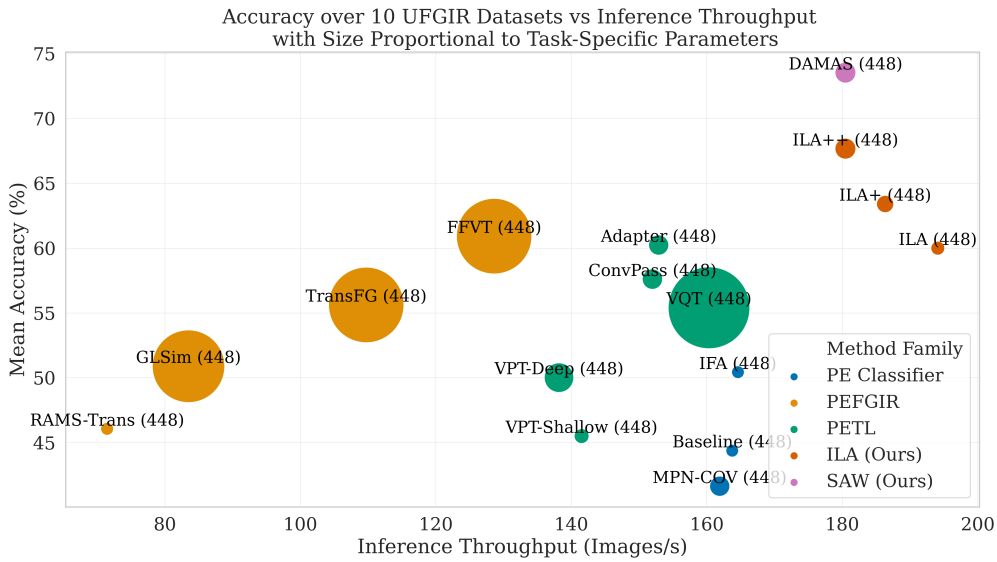


Fig. 8. Performance trade-offs at 448×448 resolution. The plot illustrates the relationship between accuracy and throughput (images/second) for various methods. Our DAMAS and ILA-based approaches achieve the highest throughput along with superior mean accuracy, while methods such as Adapter, despite high throughput, exhibit substantially lower accuracy. This demonstrates that our method is both computationally efficient and outperforms state-of-the-art techniques.

Table 5. Top-1 accuracy (%) on Soyaging R1-R6 series datasets at 448×448 resolution. Results are reported as average ± standard deviation over multiple runs. Our proposed SAW methods achieve the highest performance, highlighting the benefits of high-resolution inputs and the self-supervised adaptation warmup stage for fine-grained recognition.

method	SoyAgeR1	SoyAgeR3	SoyAgeR4	SoyAgeR5	SoyAgeR6
SIM-Trans*	69.9	73.23	73.13	73.94	63.23
CSDNet* [12]	75.15	76.57	77.27	78.18	69.8
ViT B-16	60.47 ± 0.21	55.66 ± 0.91	58.25 ± 0.56	59.6 ± 0.1	47.17 ± 0.52
MPNC [30]	54.88 ± 1.37	51.01 ± 2.8	54.95 ± 0.53	53.87 ± 1.74	43.4 ± 1.49
IFA[36]	63.77 ± 2.89	60.07 ± 2.53	62.53 ± 3.85	64.21 ± 1.81	55.02 ± 4.74
TransFG [14]	68.42 ± 0.21	62.73 ± 0.44	68.86 ± 1.02	68.99 ± 0.0	55.76 ± 0.88
FFVT[43]	72.39 ± 0.73	70.34 ± 0.84	72.8 ± 2.56	72.49 ± 0.47	59.39 ± 1.03
CAL[35]	58.62 ± 0.31	60.4 ± 1.37	56.36 ± 1.24	62.22 ± 2.52	48.79 ± 1.84
RAMS[20]	60.67 ± 1.29	58.72 ± 0.29	59.23 ± 0.59	62.39 ± 1.12	49.76 ± 0.71
GLSim[37]	61.65 ± 0.77	62.36 ± 0.36	63.97 ± 0.56	62.19 ± 0.32	53.6 ± 0.38
VQT [41]	69.29 ± 1.82	66.53 ± 1.13	67.14 ± 0.85	69.83 ± 1.37	57.44 ± 2.13
VPT-S [22]	59.66 ± 0.93	54.88 ± 0.31	59.53 ± 2.97	61.15 ± 0.46	49.29 ± 0.44
VPT-D [22]	66.1 ± 3.95	51.82 ± 3.75	60.34 ± 2.35	63.84 ± 1.05	53.67 ± 1.76
ConvP [23]	68.45 ± 1.76	67.88 ± 0.66	68.28 ± 0.63	71.62 ± 0.93	58.52 ± 0.91
Adapter [17]	69.02 ± 0.31	67.84 ± 1.39	71.45 ± 0.67	72.26 ± 1.32	59.09 ± 2.93
ILA [38]	69.49 ± 0.56	70.17 ± 0.56	70.74 ± 0.38	73.33 ± 0.53	62.83 ± 0.56
ILA+[38]	73.1 ± 1.0	73.74 ± 1.06	75.45 ± 0.86	75.52 ± 1.31	66.53 ± 0.65
ILA++[38]	75.86 ± 0.1	74.65 ± 0.71	76.8 ± 0.86	77.88 ± 0.44	69.9 ± 0.66
SAW	**83.27 ± 0.25**	**81.85 ± 0.29**	**83.7 ± 0.92**	**84.31 ± 0.31**	**74.21 ± 1.23**

sizes also incur a higher TTP, and the incremental accuracy gains are not substantial enough to justify the additional cost. Our experiments indicate that a kernel size of 3 represents the optimal balance among accuracy, FLOPs, and TTP.

Table 6. Comparison of top-1 accuracy, total trainable parameters (TTTP), FLOPs, TP, and memory usage (in GB) for various PETL methods evaluated at 448×448 resolution. This table highlights the trade-offs between accuracy and computational efficiency, demonstrating that our proposed approaches achieve competitive or superior performance while reducing resource requirements.

Method	Acc.	TTTP	FLOPs (10^9)	TP	Memory (GB)
ViT B-16*	56.6 ± 18.1	866	78.5	163.8	3
MPNC* [30]	53.3 ± 18.4	872.4	78.4	161.9	5.1
TransFG* [14]	65.9 ± 13	866	130.1	109.7	20.2
FFVT* [43]	66.5 ± 12.1	866	73	128.6	20.2
CAL*[35]	59.7 ± 20.9	973.7	627.4	21.5	11
RAMS* [20]	63.5 ± 13.5	866	162.4	71.5	22.1
GLSim* [37]	65.6 ± 14.8	936.9	157	83.5	4.6
SIM-Trans*	55.4 ± 22.2	996.6	81.8	153.9	14.7
CSDNet* [12]	68.4 ± 11	866	78.4	173.3	2.3
ViT B-16	44.4 ± 14.5	3.5	78.5	163.8	3
MPNC [30]	41.6 ± 12.5	9.9	78.4	161.9	5.1
IFA [36]	50.4 ± 13.3	3.5	78.5	164.6	12.6
TransFG [14]	55.6 ± 11.6	74.4	130.1	109.7	20.2
FFVT [43]	60.9 ± 11.7	74.4	73	128.6	20.2
CAL [35]	47.1 ± 13	111.2	627.4	21.5	11
RAMS[20]	46.1 ± 14.5	3.5	162.4	71.5	22.1
GLSim [37]	50.9 ± 12.4	74.4	157	83.5	4.6
VQT [41]	55.4 ± 14.3	212	78.9	160.3	4.7
VPT-S [22]	45.5 ± 13.6	4.3	90.1	141.5	5.4
VPT-D [22]	50 ± 13.2	12.7	90.1	138.2	5.6
ConvP [23]	57.6 ± 11.9	6.8	78.8	152	4.5
Adapter [17]	60.2 ± 12.3	6.6	78.8	152.9	4.5
ILA [38]	60 ± 11.5	3.9	66.9	194.1	4.3
ILA+ [38]	63.4 ± 11.4	5.2	67	186.3	5.3
ILA++ [38]	67.7 ± 9.8	7	67.1	180.5	4.3
SAW	**73.5 ± 10.6**	7	67.1	180.5	4.3

5.3.5 ***Data-Aware Data Augmentation Impact.*** Our Dual-Attention-driven Mix Augmentation SupCon (DAMAS) is a key component of the Self-Supervised Adaptation Warmup (SAW) framework. Table 11 summarizes our ablation study on different warmup loss functions and data diversity settings. When using the SupCon loss with our dual-attention-driven mix augmentation (denoted as DAMAS), we observe the highest accuracy; approximately 49.6% on SoyGlobal and 53.3% on SoyLocal. In contrast, omitting the mix augmentation and using only the SupCon loss causes a significant drop in performance (to 33.4% on SoyGlobal and 33.7% on SoyLocal). We also evaluate alternative losses: DAMASelf, which employs a self-contrastive loss, yields competitive results but incurs additional parameters due to auxiliary sub-networks, and the parametric instance discrimination (PID) loss introduces more parameters which grows with the size of the datasets. Moreover, our experiments show that pooling data from all five datasets during the adaptation

Different experiments such as combining different datasets should have their own small sub-sections or at least a different paragraph so it's easy for the reader to identify the key components from these experiments
Group them at least into the following in their respective paragraphs and also add horizontals bars into the table to distinguish between them if possible
Loss function (PID, DAMAS, DAMASelf)
Data-augmentation type (Crop, mask, mix)
Data-aware source: dual-attention vs full rollout
Data-diversity (single dataset vs leaf pooled)

Table 7. Impact of Spatial Downsampling and Residual Connections on UFGIR Accuracy. The table reports the top-1 accuracy (average ± standard deviation) on the SoyGlobal and SoyLocal datasets for various ablation configurations of our ILA module

Model	SoyGlobal	SoyLocal
ViT B-16	17.9 ± 0.4	28.8 ± 1
No Downsampling	34.3 ± 0.6	33.1 ± 1
No Residual	2.2 ± 0.9	6.6 ± 1.4
AvgPool	29.1 ± 0.4	31.2 ± 1.3
Conv	34.9 ± 1.4	27.8 ± 5.6
DWC (Normal Init)	29.2 ± 1.8	25.2 ± 2
DWC (Near Ones Init, Def.)	**43.5 ± 0.2**	**41.3 ± 1**

Table 8. Comparison of ILA downsampling positions in transformer layers, showing accuracy on SoyGlobal and SoyLocal datasets along with FLOPs. Loc. 4-8 achieves the best balance between accuracy and computational cost.

Model	SoyGlobal Acc.	SoyLocal Acc.	FLOPs
ViT B-16	17.9 ± 0.4	28.8 ± 1	17.6
Loc. 1-2	33.9 ± 0.2	32.1 ± 0.9	**10.0**
Loc. 2-4	41.5 ± 0.4	38.2 ± 1.1	11.0
Loc. 3-6	41.7 ± 0.4	38.3 ± 1.2	12.1
Loc. 4-8 (Def.)	**43.5 ± 0.2**	41.3 ± 1	13.1
Loc. 5-10	42.9 ± 0.2	**43.6 ± 1.1**	14.2
Loc. 6-12	42.4 ± 0.4	41.1 ± 1	15.3

Table 9. Top-1 Accuracy (%) on SoyGlobal and SoyLocal for different backbone configurations. Results are reported as average ± standard deviation. For each backbone, SAW outperforms the baseline, Adapter, and ILA++ variants, highlighting its effectiveness across different transformer architectures.

Model	SoyGlobal	SoyLocal
ViT B-16	17.9 ± 0.4	28.8 ± 1
Adapter (ViT)	35 ± 1.2	33.3 ± 1.3
ILA++ (ViT)	44 ± 0.3	41.2 ± 0.8
SAW (ViT)	**49.6 ± 0.4**	**53.3 ± 0.3**
Baseline (DeiT)	24.8 ± 0.1	37.4 ± 0.5
Adapter (DeiT)	39.3 ± 1.1	43.8 ± 1.1
ILA++ (DeiT)	43.7 ± 0.4	46.4 ± 0.8
SAW (DeiT)	**50.5 ± 0.1**	**53.1 ± 1.8**
Baseline (DeiT3)	15.8 ± 0.2	21.7 ± 0.6
Adapter (DeiT3)	20.8 ± 1.8	28.1 ± 1.1
ILA++ (DeiT3)	36.3 ± 1.2	36.8 ± 2.1
SAW (DeiT3)	**47.3 ± 0.3**	**42.8 ± 0.8**

Table 10. Trade-off analysis of convolution kernel size in the ILA module. Top-1 accuracy (avg ± std), FLOPs (G), and TTP (M) are reported for SoyGlobal and SoyLocal; a kernel size of 3 provides an optimal balance.

Model	SoyGlobal	SoyLocal	FLOPs	TTP
Baseline	17.9 ± 0.4	28.8 ± 1	17.57	**0.15**
KS=1	34.4 ± 0.4	33.6 ± 0.5	17.63	0.5
KS=3 (Def.)	**43.5 ± 0.2**	41.3 ± 1	13.14	0.51
KS=5	42.7 ± 0.2	**43.2 ± 0.3**	9.95	0.53
KS=7		42.7 ± 0.3	**7.97**	0.57

warmup is crucial; pretraining on a single dataset (DAMAS-SD) results in a 3% and 8% accuracy decline on SoyGlobal and SoyLocal, respectively. Overall, our analysis confirms that the combination of dual-attention-driven mix augmentation and increased data diversity substantially improves top-1 accuracy over ILA++, by over 12% on SoyGlobal and 30% on SoyLocal.

Table 11. Ablation study of warmup loss functions and data diversity for self-aware data augmentation. Results are reported as top-1 accuracy (average ± standard deviation) on SoyGlobal and SoyLocal.

Model	SoyGlobal	SoyLocal
ViT B-16	17.9 ± 0.4	28.8 ± 1
ILA++ (ConvCLSAd)	43.5 ± 0.2	41.3 ± 1
ILA++ (DWCLSAd)	44 ± 0.3	41.2 ± 0.8
SupCon	33.4 ± 0.3	33.7 ± 0.2
PID	42.1 ± 0.2	43.4 ± 1.2
DACAS	42.9 ± 0.1	42.3 ± 0.2
DARAS	39.9 ± 1.1	44.4 ± 0.4
DAMAS-SD	47.9 ± 0.6	48.4 ± 1
DAMAS	49.6 ± 0.4	**53.3 ± 0.3**
ROMAS	49.6 ± 0.1	52.8 ± 1.3
DAMASelf	**49.9 ± 0.2**	53.2 ± 0.5

6 CONCLUSION

In this work, we extended our prior Inter-Layer Adapter (ILA) framework with a Self-Supervised Adaptation Warmup (SAW) stage to address the unique challenges of ultra-fine-grained image recognition (UFGIR). By integrating a Dual-Attention-driven Mix Augmentation SupCon (DAMAS) strategy and leveraging diverse datasets during adaptation warmup, our approach effectively bridges the modality gap between generic pretraining and fine-grained tasks. Extensive experiments across multiple datasets demonstrate that SAW not only mitigates the attention collapse in frozen Vision Transformers but also achieves significant accuracy gains while drastically reducing computational cost and parameter overhead. These results highlight the potential of self-supervised adaptation for achieving cost-efficient, high-performance fine-grained recognition, and they open avenues for future research in adaptive and lightweight transfer learning strategies.

ACKNOWLEDGMENTS

This work was supported by the National Science and Technology Council, Taiwan, under grant NSTC 111-2221-E-A49-092-MY3, NSTC 113-2640-E-A49-005, NSTC 112-2221-E-007-079-MY3 and NSTC 113-2218-E-007-020. We also thank the National Center for High-performance Computing (NCHC) and NYCU HPC for providing computational and storage resources.

REFERENCES

[1] Samira Abnar and Willem Zuidema. 2020. Quantifying Attention Flow in Transformers. https://doi.org/10.48550/arXiv.2005.00928 arXiv:2005.00928 [cs].

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. https://doi.org/10.48550/arXiv.1607.06450 arXiv:1607.06450 [cs, stat].

[3] Sangmin Bae, Sungnyun Kim, Jongwoo Ko, Gihun Lee, Seungjong Noh, and Se-Young Yun. 2023. Self-contrastive learning: single-viewed supervised contrastive framework using sub-network. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23, Vol. 37)*. AAAI Press, 197–205. https://doi.org/10.1609/aaai.v37i1.25091

[4] Jiawang Bai, Li Yuan, Shu-Tao Xia, Shuicheng Yan, Zhifeng Li, and Wei Liu. 2022. Improving Vision Transformers by Revisiting High-Frequency Components. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Vol. 13684. Springer Nature Switzerland, Cham, 1–18. https://doi.org/10.1007/978-3-031-20053-3_1 Series Title: Lecture Notes in Computer Science.

[5] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. 2023. A Cookbook of Self-Supervised Learning. https://doi.org/10.48550/arXiv.2304.12210 arXiv:2304.12210 [cs].

[6] Yun-Hao Cao, Hao Yu, and Jianxin Wu. 2022. Training Vision Transformers with only 2040 Images. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 220–237. https://doi.org/10.1007/978-3-031-19806-9_13

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9630–9640. https://doi.org/10.1109/ICCV48922.2021.00951 ISSN: 2380-7504.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 1597–1607. ISSN: 2640-3498.

[9] Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. 2022. The Principle of Diversity: Training Stronger Vision Transformers Calls for Reducing All Levels of Redundancy. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 12010–12020. https://doi.org/10.1109/CVPR52688.2022.01171

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). arXiv: 1810.04805.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929 [cs]* (Oct. 2020). arXiv: 2010.11929.

[12] Ziye Fang, Xin Jiang, Hao Tang, and Zechao Li. 2024. Learning Contrastive Self-Distillation for Ultra-Fine-Grained Visual Categorization Targeting Limited Samples. *IEEE Transactions on Circuits and Systems for Video Technology* (2024), 1–1. https://doi.org/10.1109/TCSVT.2024.3370731 Conference Name: IEEE Transactions on Circuits and Systems for Video Technology.

[13] Junhyeong Go and Jongbin Ryu. 2025. Channel Propagation Networks for Refreshable Vision Transformer. 1353–1362.

[14] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. 2022. TransFG: A Transformer Architecture for Fine-Grained Recognition. In *Proceedings of the First MiniCon Conference*.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]* (Dec. 2015). arXiv: 1512.03385.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. https://doi.org/10.48550/arXiv.1503.02531 arXiv:1503.02531 [cs, stat].

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2790–2799. ISSN: 2640-3498.

[18] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. https://doi.org/10.48550/arXiv.1704.04861 arXiv:1704.04861 [cs].

[19] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. 2019. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification. https://doi.org/10.48550/arXiv.1901.09891 arXiv:1901.09891 [cs].

[20] Yunqing Hu, Xuan Jin, Yin Zhang, Haiwen Hong, Jingfeng Zhang, Yuan He, and Hui Xue. 2021. RAMS-Trans: Recurrent Attention Multi-scale Transformer for Fine-grained Image Recognition. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*. Association for Computing Machinery, New York, NY, USA, 4239–4248. https://doi.org/10.1145/3474085.3475561

[21] Shaoli Huang, Xinchao Wang, and Dacheng Tao. 2021. SnapMix: Semantically Proportional Mixing for Augmenting Fine-grained Data. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 2 (May 2021), 1628–1636. https://doi.org/10.1609/aaai.v35i2.16255

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual Prompt Tuning. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Vol. 13693. Springer Nature Switzerland, Cham, 709–727. https://doi.org/10.1007/978-3-031-19827-4_41 Series Title: Lecture Notes in Computer Science.

[23] Shibo Jie and Zhi-Hong Deng. 2022. Convolutional Bypasses Are Better Vision Transformer Adapters. https://doi.org/10.48550/arXiv.2207.07039 arXiv:2207.07039 [cs].

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, 18661–18673.

[25] Shu Kong and Charless Fowlkes. 2017. Low-Rank Bilinear Pooling for Fine-Grained Classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, HI, 7025–7034. https://doi.org/10.1109/CVPR.2017.743

[26] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 3519–3529. ISSN: 2640-3498.

[27] Mónica G. Larese, Ariel E. Bayá, Roque M. Craviotto, Miriam R. Arango, Carina Gallo, and Pablo M. Granitto. 2014. Multiscale recognition of legume varieties based on leaf venation images. *Expert Systems with Applications* 41, 10 (Aug. 2014), 4638–4647. https://doi.org/10.1016/j.eswa.2014.01.029

[28] Jan Lehr, Alik Sargsyan, Martin Pape, Jan Philipps, and Jörg Krüger. 2020. Automated Optical Inspection Using Anomaly Detection and Unsupervised Defect Clustering. In *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Vol. 1. 1235–1238. https://doi.org/10.1109/ETFA46521.2020.9212172 ISSN: 1946-0759.

[29] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3045–3059. https://doi.org/10.18653/v1/2021.emnlp-main.243

[30] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. 2018. Towards Faster Training of Global Covariance Pooling Networks by Iterative Matrix Square Root Normalization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, 947–955. https://doi.org/10.1109/CVPR.2018.00105

[31] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. 2017. Is Second-Order Information Helpful for Large-Scale Visual Recognition?. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, 2089–2097. https://doi.org/10.1109/ICCV.2017.228

[32] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. 2016. Using Deep Learning for Image-Based Plant Disease Detection. *Frontiers in Plant Science* 7 (Sept. 2016). https://doi.org/10.3389/fpls.2016.01419 Publisher: Frontiers.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. https://doi.org/10.48550/arXiv.1807.03748 arXiv:1807.03748 [cs].

[34] Wongi Park and Jongbin Ryu. 2023. Fine-Grained Self-Supervised Learning with Jigsaw Puzzles for Medical Image Classification. https://doi.org/10.48550/arXiv.2308.05770 arXiv:2308.05770 [cs].

[35] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. 2021. Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-Identification. 1025–1034.

[36] Edwin Arkel Rios, Min-Chun Hu, and Bo-Cheng Lai. 2022. Anime Character Recognition using Intermediate Features Aggregation. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. 424–428. https://doi.org/10.1109/ISCAS48785.2022.9937519 ISSN: 2158-1525.

[37] Edwin Arkel Rios, Min-Chun Hu, and Bo-Cheng Lai. 2024. Global-Local Similarity for Efficient Fine-Grained Image Recognition with Vision Transformers. arXiv:2407.12891 [cs].

[38] Edwin Arkel Rios, Femiloye Oyerinde, Min-Chun Hu, and Bo-Cheng Lai. 2024. Down-Sampling Inter-Layer Adapter for Parameter and Computation Efficient Ultra-Fine-Grained Image Recognition. In *Computer Vision – ECCV 2024 Workshops*. Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2409.11051 arXiv:2409.11051 [cs].

[39] Linus Scheibenreif, Michael Mommert, and Damian Borth. 2024. Parameter Efficient Self-Supervised Geospatial Domain Adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 27841–27851. https://doi.org/10.1109/CVPR52733.2024.02630

[40] Hongbo Sun, Xiangteng He, and Yuxin Peng. 2022. SIM-Trans: Structure Information Modeling Transformer for Fine-grained Visual Categorization. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. Association for Computing Machinery, New York, NY, USA, 5853–5861. https://doi.org/10.1145/3503161.3548308

[41] Cheng-Hao Tu, Zheda Mai, and Wei-Lun Chao. 2023. Visual Query Tuning: Towards Effective Usage of Intermediate Representations for Parameter and Memory Efficient Transfer Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Vancouver, BC, Canada, 7725–7735. https://doi.org/10.1109/CVPR52729.2023.00746

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[43] Jun Wang, Xiaohan Yu, and Yongsheng Gao. 2021. Feature Fusion Vision Transformer for Fine-Grained Visual Categorization. In *British Machine Vision Conference (BMVC)*. arXiv: 2107.02341.

[44] Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. 2022. Anti-Oversmoothing in Deep Vision Transformers via the Fourier Domain Analysis: From Theory to Practice. https://doi.org/10.48550/arXiv.2203.05962 arXiv:2203.05962 [cs].

[45] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. 2021. Fine-Grained Image Analysis with Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. https://doi.org/10.1109/TPAMI.2021.3126648 Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[46] Xiaohan Yu, Jun Wang, and Yongsheng Gao. 2023. CLE-ViT: Contrastive Learning Encoded Transformer for Ultra-Fine-Grained Visual Categorization. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China, 4531–4539. https://doi.org/10.24963/ijcai.2023/504

[47] Xiaohan Yu, Jun Wang, Yang Zhao, and Yongsheng Gao. 2023. Mix-ViT: Mixing attentive vision transformer for ultra-fine-grained visual categorization. *Pattern Recognition* 135 (March 2023), 109131. https://doi.org/10.1016/j.patcog.2022.109131

[48] Xiaohan Yu, Yang Zhao, and Yongsheng Gao. 2022. SPARE: Self-supervised part erasing for ultra-fine-grained visual categorization. *Pattern Recognition* 128 (Aug. 2022), 108691. https://doi.org/10.1016/j.patcog.2022.108691

[49] Xiaohan Yu, Yang Zhao, Yongsheng Gao, and Shengwu Xiong. 2021. MaskCOV: A random mask covariance network for ultra-fine-grained visual categorization. *Pattern Recognition* 119 (Nov. 2021), 108067. https://doi.org/10.1016/j.patcog.2021.108067

[50] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan. 2020. Patchy Image Structure Classification Using Multi-Orientation Region Transform. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07 (April 2020), 12741–12748. https://doi.org/10.1609/aaai.v34i07.6968 Number: 07.

[51] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Xiaohui Yuan, and Shengwu Xiong. 2021. Benchmark Platform for Ultra-Fine-Grained Visual Categorization Beyond Human Performance. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 10265–10275. https://doi.org/10.1109/ICCV48922.2021.01012 ISSN: 2380-7504.

[52] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 6022–6031. https://doi.org/10.1109/ICCV.2019.00612

[53] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. 2022. Self-Distillation: Towards Efficient and Compact Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 08 (Aug. 2022), 4388–4403. https://doi.org/10.1109/TPAMI.2021.3067100 Publisher: IEEE Computer Society.

[54] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. 2021. DeepViT: Towards Deeper Vision Transformer. https://doi.org/10.48550/arXiv.2103.11886 arXiv:2103.11886 [cs].

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009