Local Foreground Selection Aware Attentive Feature Reconstruction for Few-Shot Fine-Grained Plant Species Classification

Aisha Zulfiqar[a] and Ebroul Izquierdo

School of Electronics Engineering and Computer Science, Queen Mary University of London, U.K.

fi

Keywords: Few-Shot Learning, Plant Species Classification, Fine-Grained Classification, Attention Mechanism.

Abstract: Plant species exhibit subtle distinctions, requiring a reduction in intra-class variation and an increase in inter-class differences to improve accuracy. This paper addresses plant species classification using a limited number of labelled samples and introduces a novel Local Foreground Selection(LFS) attention mechanism. Based on the proposed attention Local Foreground Selection Module(LFSM) is a straightforward module designed to generate discriminative support and query feature maps. It operates by integrating two types of attention: local attention, which captures local spatial details to enhance feature discrimination and increase inter-class differentiation, and foreground selection attention, which emphasizes the foreground plant object while mitigating background interference. By focusing on the foreground, the query and support features selectively highlight relevant feature sequences and disregard less significant background sequences, thereby reducing intra-class differences. Experimental results from three plant species datasets demonstrate the effectiveness of the proposed LFS attention and its complementary advantages over previous feature reconstruction methods.

1 INTRODUCTION

Automatic plant species classification is essential for ecology and biodiversity conservation. It is a fine-grained image classification problem, characterized by subtle differences between species. Large intra-class variations arise from differences in background, illumination and pose, while inter-class variations tend to be minimal due to similar morphology. Consequently, effective training must focus on learning discriminative features to enhance classification performance. Few-shot fine-grained image classification is particularly challenging due to limited training data, necessitating the optimization of both inter-class and intra-class variations. Conventional few-shot learning (Snell et al., 2017) (Sung et al., 2018) typically addresses only class-level differences, which limits its effectiveness for fine-grained classification tasks. Recent few-shot fine-grained classification methods use feature reconstruction (Wertheimer et al., 2021) (Doersch et al., 2020), where support features reconstruct query features to enhance class separation to optimize inter-class and intra-class variations.

In this study, we present a novel Local Foreground Selection (LFS) attention mechanism, designed to localize discriminative regions through a dual-action attention approach that generates distinct features. It effectively reduces background effects while emphasizing the foreground, along with capturing essential local discriminative details. This is achieved by combining local attention, which extracts local spatial context to enhance inter-class variation, and foreground selection attention, which minimizes intra-class variation by highlighting foreground object and reducing background effects. The novelty of the LFS attention lies in its integration of local and foreground selection attention, which together outperforms their individual results. This attention mechanism produces highly discriminative features, facilitating the reconstruction of support and query features in existing few-shot fine-grained classification methods (Wertheimer et al., 2021) (Wu et al., 2023). Our main contributions can be summarized as follows:

- In this study, we propose a novel Local Foreground Selection attention designed to optimize both inter-class and intra-class variations.
- The proposed attention is incorporated into a vision transformer encoder, creating the Local Foreground Selection Module (LFSM), which produces discriminative feature maps.
- The LFSM module significantly improves perfor-

[a] https://orcid.org/0000-0001-6647-0134

Zulfiqar, A. and Izquierdo, E.
Local Foreground Selection Aware Attentive Feature Reconstruction for Few-Shot Fine-Grained Plant Species Classification.
DOI: 10.5220/0013184000003912
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 3: VISAPP*, pages 532-539
ISBN: 978-989-758-728-3; ISSN: 2184-4321
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

mance for plant images with natural background, when combined with state of the art feature reconstruction approaches.

2 RELATED WORKS

Metric Based Few-Shot Learning. Metric learning is a widely used approach for few-shot learning. Prominent metric-based methods include Matching Networks (Vinyals et al., 2016), Prototypical Networks (Snell et al., 2017), and Relation Networks (Sung et al., 2018). These methods transform training data into feature vectors within a shared feature space, where the similarity between two feature vectors is measured by their distance. Metric learning approaches for fine-grained image classification leverage euclidean or cosine distances as metric. Recent approaches include low-rank pairwise bilinear pooling to learn effective distance metrics (Huang et al., 2021b) and focus area location mechanisms for identifying similar regions among objects (Sun et al., 2021). Techniques like multi-attention meta-learning (MattML) (Zhu et al., 2020) and local descriptor-based image-to-class measures (Li et al., 2019) enhance feature metrics for better classification. Non-linear data projection networks (NDPNet) (Zhang et al., 2021) improve similarity measures through advanced projections, while bi-similarity network (Li et al., 2020) employ dual similarity checks for more discriminative features. Additionally, target-oriented alignment networks (TOAN) (Huang et al., 2021a) align support and query images to minimize intra-class variance and maximize inter-class variance, contributing to more robust classification performance.

Feature Reconstruction Method. While metric-based methods necessitate the creation of a single vector that preserves spatial locations, feature reconstruction approaches address this limitation. DeepEMD (Zhang et al., 2020) does not perform matching at the image level like traditional metric learning methods; instead, it partitions the image into a set of local representations. Optimal matching is conducted on these representations from two images to assess similarity, with Earth Mover's distance. In the Feature Reconstruction Network (FRN) (Wertheimer et al., 2021), feature maps are reconstructed by pooling support set feature maps into a matrix, where each column represents the concatenated feature maps of a channel. For classification, each location in the query image's feature map is reconstructed using a weighted sum of the support features from the corresponding class. The LCCRN (Li et al., 2023) network enhances local information extraction by introducing a local content extraction module, while a separate embedding module preserves appearance details. Bi-FRN (Wu et al., 2023) adopts a feature reconstruction strategy which reconstructs query images from support images and vice versa. This results in a feature pool derived from both support and query images, with pooled features mutually reconstructed from one another.

Attention Mechanism. The transformer self-attention (Vaswani et al., 2017) has been incorporated into several few-shot learning methods. The Few-shot Embedding Adaptation with Transformer (FEAT) (Ye et al., 2020) utilizes it to perform task-specific embedding adaptation. This approach employs a set-to-set function to derive more discriminative instance representations and models interactions among images within each set. The CTX model (Doersch et al., 2020) proposes a self-supervised learning framework combined with a Cross Transformer, representing images as spatial tensors and generating query-aligned prototypes. CTX uses self-attention to identify the spatial attention weights of support and query images, facilitating the learning of query-aligned class prototypes. The Few-shot Cosine Transformer (FS-CT) (Nguyen et al., 2023) introduces cosine attention, which produces an effective correlation map between support and query images, outperforming softmax attention. Additionally, the Bi-directional Feature Reconstruction Network (Bi-FRN) (Wu et al., 2023) utilizes self-attention to generate discriminative query and support features.

3 METHOD

Classification of plant species represents a fine-grained visual classification (FGVC) task, characterized by subtle differences among various species. This challenge is compounded by large intra-class variations and minimal inter-class differences that must be addressed. Additionally, plant classification is conducted on images within their natural environments, necessitating the removal of background effects. We propose a novel attention-based module designed to generate discriminative feature maps for use with existing few-shot fine-grained approaches using feature reconstruction approaches. This novel module can enhance the performance of prior works (Wu et al., 2023) (Wertheimer et al., 2021). The overall framework of our method is illustrated in Figure 1. For an episode, the support and query samples are passed through feature embedding where feature maps are obtained. To obtain discriminative feature maps suitable for fine grained classification we refine them with our proposed attention mechanism and pass them through the Local Foreground Selection module (LFSM). Subsequently, the output of the LFSM is fed into the feature reconstruction networks (Wu et al., 2023) (Wertheimer et al., 2021), which reconstruct query and support features. Finally, the similarity metric computes the euclidean distance between the original and reconstructed feature maps.

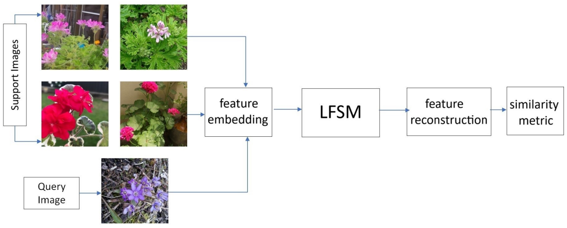

Figure 1: Local Foreground Selection aware Attentive Feature Reconstruction Network.

Problem Formulation. For a given dataset D, it is divided into D_{train}, D_{val} and D_{test} such that the categories in these sets are disjoint. Few-shot classification performs the task of C-way and K-shot classification on D_{test} by learning the knowledge from D_{train} and D_{val}. The task is performed such that C classes from the test set are selected. From each of these classes only K labelled samples are selected which serve as the support samples(S) and M unlabelled samples are selected which form part of the query samples(Q). In this few-shot task the classification accuracy is determined on D_{test}.

3.1 Feature Embedding

The first step is to extract feature maps of support and query samples. Following literature we use the Conv-4 and ResNet-12 networks as backbones to obtain feature maps. The architectures of these networks are the same as in (Wertheimer et al., 2021). For a C-way and K-shot task $C \times (K+M)$ images are input to the embedding module where features are extracted.

3.2 Local Foreground Selection Module

The LFSM module is responsible for generating discriminative features for support and query pools. For a C-way, K-shot task, the extracted features from the embedding module are represented as $u_i = f_\theta(u_i) \in \mathbb{R}^{d\times h\times w}$, where d denotes the number of channels, h is the height of the feature maps, and w represents the width. The input features to the LFSM are denoted as u_i, while the output from this module is $y_i \in \mathbb{R}^{d\times r}$. Within the LFSM module, the input features u_i are transformed into r local features $[u_i^1, u_i^2, u_i^3, \ldots, u_i^r] + E_{pos}$, which are then fed into a vision transformer encoder.

The LFSM constructs discriminative feature maps utilizing the proposed Local Foreground Selection (LFS) attention embedded within a vision transformer encoder. Originally, vision transformer employs self-attention but instead we introduce LFS attention. It performs two tasks: the local attention task and foreground selective attention task. The local attention is to emphasize the local details of the plant object which increases inter-class differences. The presence of natural background in plant images increases intra-class variations, we aim to mitigate its impact by using a foreground selection attention. The foreground selection attention filters out sequences associated with the background of the target plant object, ensuring that only the important tokens relevant to the foreground are retained. The two attentions—local and foreground selection—are aggregated and generate distinct features for fine grained classification. The architecture of the LFS module is presented in the Figure 2 and a comparison of the heatmap resulting from the LFS attention with the output of the feature embedding is visualized in Figure 3.

Local Attention. The local attention mechanism integrates convolutions into the vision transformer block to effectively model local spatial context. Instead of position-wise linear projections for Multi-Head Self-Attention we use convolutional projections for the attention operation. Specifically, depth-wise separable convolutions are utilized for these projections, which are applied to derive the queries (Q), keys (K), and values (V). The feature maps, reshaped earlier in the LFSM module, serve as input for the depth-wise convolutional projections. This function returns the attention scores to the vision transformer encoder.

Foreground Selection Attention. The purpose of foreground selection attention is to enable a focus on the plant object rather than the background, as it does not contribute to fine-grained classification. In transformer self-attention, all tokens are assigned relevance weights that reduce the influence of less important tokens. Tokens not associated with the plant object receive lower attention scores; however, these background tokens still exert some influence. To further diminish their impact, these tokens are discarded, allowing the feature maps to concentrate on the plant object, thereby aiding in the classification of subtle class differences. The foreground selection attention generates relevance scores by selecting tokens based on a defined threshold, discarding those below it. The

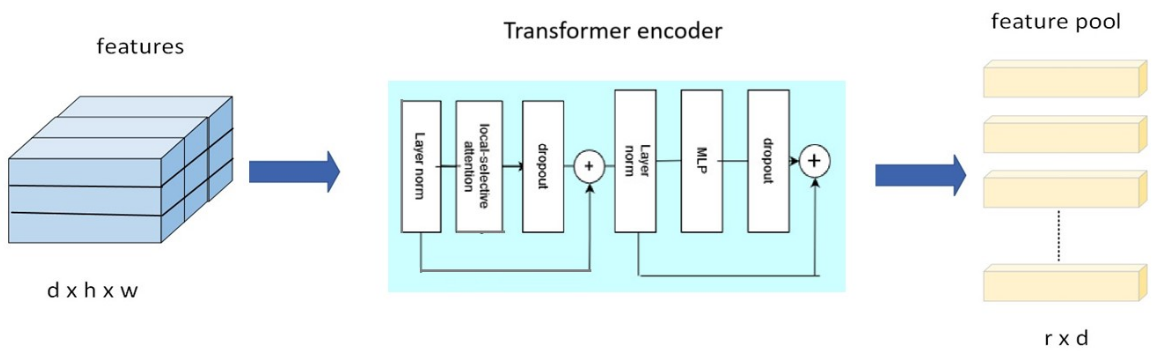

Figure 2: Local Foreground Selection Module where features are fed into the vision transformer and outputs feature pool.

dot product of Q and K^T creates an attention score matrix with dimensions $m \times m$. The relevance of all tokens is then evaluated relative to one another and stored in the relevance matrix, where the relevance of the i^{th} token with respect to the j^{th} token is denoted by the matrix element $relevance_{i,j}$. Relevance scores are calculated for each row of the attention matrix. Tokens representing the plant object yield high relevance scores, while those associated with the background have low scores. Consequently, the relevance scores higher than the threshold are preserved, while lower relevance scores are set to zero. To implement this, we arrange the rows in descending order. After sorting the i^{th} row of the relevance matrix, we identify the $(m \times \text{FS-ratio})^{th}$ index, where the FS-ratio determines the number of tokens selected and ranges from 0.1 to 1.0. The relevance score at this selected index is recorded as the threshold for comparison. A comparison is then made between $relevance_{i,j}$ and the threshold; if $relevance_{i,j}$ exceeds the threshold, it is retained, otherwise set to zero. The relevance matrix is then normalized by softmax function. The matrix multiplication of the attention score and V is calculated to get the weighted sum over V as indicated by the following equation. This process is mathematically illustrated in the following equations:

$$\text{relevance} = \frac{QK^T}{\sqrt{d_k}} \tag{1}$$

$$\text{index} = \text{argsort}\left(-\text{relevance}_i\right)\left[\text{FS-ratio} \times m\right] \tag{2}$$

$$\text{relevance}_i = \begin{cases} \text{relevance}_i, & \text{if } > \text{relevance}_i[\text{index}] \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$\text{FS-Attention} = \text{softmax}\left(\text{relevance}_i\right) V \tag{4}$$

After getting the Foreground Selection attention(FS-attention) we replace all the non-zero weights to one. In this way the elements of the matrix are either zero or one. The background related tokens are made zero and plant object tokens are made one. This takes the form of a binary matrix in which the relevance score positions that contribute to the object are represented by ones and the non-contributing scores that represent the background are zero.

$$\text{FS-Attention} = \begin{cases} 1 & \text{if FS-Attention} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Local Foreground Selection Attention. The local foreground selection attention is a bi-functional attention mechanism designed to capture local details of the plant object, as reflected in the local attention matrix. The foreground selection attention indicates which locations within the attention matrix have scores originating from the plant object versus the background. In the vision transformer encoder, the attention scores derived from local attention are element-wise multiplied with the foreground selection attention matrix. This ensures that regions with lower relevance score corresponding to the background are set to zero, effectively discarding them. Consequently, the resulting attention matrix contains values that specifically pertain to the plant object. Utilizing attention scores from local attention, rather than self-attention, enhances the representation of local details of the plant object. This targeted attention improves focus on the plant object while minimizing background interference, ultimately leading to increased accuracy. After getting the attention weights the generated features are obtained by iterative Layer Normalization and Multi Layer Perceptron producing feature pools as output.

$$\hat{y}_i = \text{Attention}\left(y_i W_\alpha^Q, y_i W_\alpha^K, y_i W_\alpha^V\right), \quad \hat{y}_i \in \mathbb{R}^{r \times d} \tag{5}$$

Where W_α^Q ,W_α^K and W_α^V are learnable weight parameters having size $d \times d$. The $\hat{y}_i$ are obtained using the following equation from Multi Layer Perceptron (MLP) and Layer Normalization (LN).

$$\hat{y}_i = MLP\left(LN\left(y_i + \hat{y}_i\right)\right) \tag{6}$$

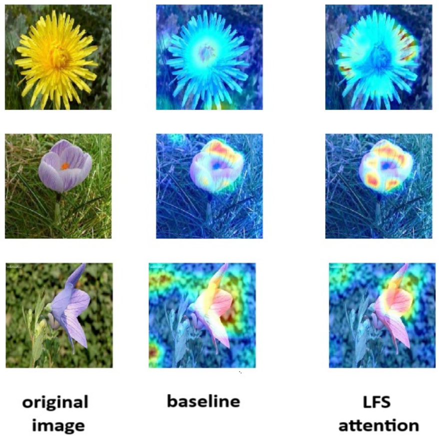

Figure 3: Heat map comparing the original feature maps to our proposed Local Foreground Selection attention on Oxford Flower-102 dataset.

4 EXPERIMENTAL RESULTS AND ANALYSIS

Datasets. The plant species classification problem is considered with natural background images from three datasets. The Oxford Flower-102 dataset (Nilsback and Zisserman, 2008) is a benchmark for few-shot fine-grained classification, featuring 102 flower categories. The iNaturalist 2019 dataset (Van Horn et al., 2018), known for its long-tailed distribution, focuses on a subset of 348 plant species from the Plantae category, each containing 50–1000 images. Similarly, the Plant Net 300-K dataset (Garcin et al., 2021), originally comprising 1081 plant species, uses a subset of 252 species having 50–1000 images for few-shot classification. All images across the datasets are resized to 84×84 without any bounding box cropping.

Implementation Details. The experiments are conducted using GeForce 3090 GPU, implemented with Pytorch on Conv-4 and ResNet-12 backbones. All methods including the baseline, state-of-the art and ours are trained from scratch. The training of our method is performed for 1200 epochs using SGD with Nesterov momentum of 0.9. The initial learning rate is taken as 0.1 and weight decay is 5e-4. Learning rate decreases by a scaling factor of 10 every 400 epochs. For Conv-4 models, we train using 30-way 5-shot episodes, query images per class are 15. For ResNet-12 models the training episodes are 15-way 5-shot. Testing for both backbones is performed for 1-shot and 5-shot. Data augmentation methods like centre crop, random horizontal flip and colour jitter are used. The best-performing model are selected based on the validation set every 20 epochs. For all experiments, we report the mean accuracy of 10,000 randomly generated tasks on D_test with 95% confidence intervals on the standard 1-shot and 5-shot settings.

4.1 Comparison with State-of-the-Art

The accuracy of our method is tested for plant species classification via few-shot learning. Experiments are conducted on the above mentioned three plant species datasets. We obtain results on the state-of-the art methods ourselves using their official codes. The state-of-the-art methods that we compare our method to include ProtoNet (Snell et al., 2017), DN4 (Li et al., 2019), CTX (Doersch et al., 2020), FRN (Wertheimer et al., 2021), LCCRN (Li et al., 2023), BiFRN (Wu et al., 2023), BSFA (Zha et al., 2023) and FSCT (Nguyen et al., 2023). A comparison of the performance of our method with the state-of-the-art methods is reflected in Table 1 for Conv-4 backbone and Table 2 for ResNet-12 backbone. For the Conv-4 backbone our method outperforms all other state of the art methods for all three plant species dataset. If we carefully observe we can see that the performance of our method with FRN (Wertheimer et al., 2021) significantly improves from 2-7 %. The 1-shot performance improves significantly for all three datasets. The accuracy also increases appreciably for 5-shot as well. When our method is used with BiFRN (Wu et al., 2023) our local foreground selection attention works well for both 1-shot and 5-shot as it removes the effect of background and focus on local details. For the ResNet-12 backbone also the performance of our method is better for 5-way 5-shot and 5-way 1-shot for all three datasets. The accuracy improves by 1-6 % when our attention mechanism is introduced with FRN and BiFRN. The LFS attention used with FRN outperforms all state of the arts as well as our LFS with BiFRN. This show that our attention mechanism gives highly discriminative feature maps that are suitable for feature reconstruction. Even though FRN performs unidirectional feature reconstruction the LFS attention make its performance comparable to the bi-directional feature reconstruction in BiFRN.

4.2 Ablation Study

The Effectiveness of Local Foreground Selection Attention. Our proposed LFS attention improves the overall accuracy when used with both FRN and BiFRN. So, we test its effectiveness of as compared to self attention, local attention and foreground selection attention(FS-attention) . A comparison is presented with the baseline i.e ProtoNet (Snell et al., 2017) and other aforementioned attentions. Our lo-

Table 1: Comparison with state-of the-art for 1-shot and 5-shot classification of plant species on Conv-4 backbone.

Method	**Oxford Flower-102**		**iNaturalist19**		**Plant Net 300-K**	
	1-shot	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Proto(Snell et al., 2017)	59.19±.22	83.10±.14	54.40±.23	76.21±.16	54.83±.22	77.60±.15
DN4(Li et al., 2019)	51.86±.21	69.13±.21	38.88±.15	49.23±.13	40.45±.16	55.41±.23
CTX(Doersch et al., 2020)	70.52±.91	80.82±.71	46.96±.94	63.18±.79	49.69±.88	65.67±.76
FRN(Wertheimer et al., 2021)	69.40±.20	87.45±.12	62.08±.22	80.19±.15	62.01±.22	80.81±.14
BiFRN(Wu et al., 2023)	75.11±.17	89.55±.15	64.54±.20	81.89±.12	62.65±.23	81.56±.17
LCCRN(Li et al., 2023)	71.38±.21	85.70±.14	64.47±.23	80.36±.16	63.50±.22	80.55±.15
BSFA(Zha et al., 2023)	71.42±.47	83.85±.33	60.51±.53	75.85±.40	62.57±.50	77.16±.36
FSCT-Cosine(Nguyen et al., 2023)	66.30±.87	82.04±.68	51.83±.96	64.46±.83	52.07±.96	66.03±.75
FSCT-Softmax(Nguyen et al., 2023)	62.96±.91	76.22±.70	50.09±.95	62.04±.74	49.84±.95	62.12±.82
Ours: LFS + BiFRN	75.71±.10	**90.30±.13**	65.36±.15	82.59±.11	63.51±.20	82.19±.14
Ours: LFS + FRN	**76.46±.20**	89.94±.13	**66.62±.22**	**82.76±.15**	**64.67±.20**	**82.31±.14**

Table 2: Comparison with state-of-the-art for 1-shot and 5-shot classification of plant species on ResNet-12 backbone.

Method	**Oxford Flower-102**		**iNaturalist19**		**Plant Net 300-K**	
	1-shot	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Proto(Snell et al., 2017)	75.23±.20	88.43±.13	72.44±.29	87.37±.21	69.96±.21	86.34±.13
DN4(Li et al., 2019)	68.62±.21	74.19±.21	58.08±.15	59.37±.13	56.45±.16	65.41±.13
CTX(Doersch et al., 2020)	77.12±.91	83.62±.71	55.36±.94	60.58±.79	59.61±.88	70.47±.76
FRN(Wertheimer et al., 2021)	79.20±.20	90.10±.11	74.62±.23	87.52±.14	72.52±.21	86.33±.14
BiFRN(Wu et al., 2023)	77.75±.17	89.94±.15	72.24±.16	87.20±.12	70.84±.15	86.52±.12
LCCRN(Li et al., 2023)	78.95±.19	92.60±.10	74.96±0.21	87.65±.12	72.63±.22	86.50±.15
BSFA(Zha et al., 2023)	77.02±.44	89.06±.26	76.23±.48	87.31±.30	74.21±.49	86.63±.29
FSCT-Cosine(Nguyen et al., 2023)	63.10±.87	85.16±.68	58.85±.96	61.86±.83	62.14±.96	71.83±.75
FSCT-Softmax(Nguyen et al., 2023)	59.29±.91	79.52±.70	59.79±.95	58.84±.74	60.01±.95	67.12±.82
Ours: LFS + BiFRN	79.77±.12	90.87±.10	72.95±.15	87.67±.11	71.04±.15	86.70±.14
Ours: LFS + FRN	**79.85±.20**	**93.50±.11**	**79.61±.11**	**93.33±.15**	**77.42±.15**	**88.71±.14**

cal foreground selection attention(LFS-attention) performs better with the combination of both local and foreground selection attention. They do not give the best results when used alone which is the evidence that LFS-attention is more effective. The results are shown in Table 3 and comparison is made for two datasets. The results reflect that our proposed LFS-attention is a superior approach for few-shot fine-grained classification.

The Effect of FS-Ratio. FS-ratio defined in the foreground selection attention part defines the number of tokens that will be accepted. By incorporating FS-ratio we can see the impact of number of tokens that will be accepted to provide the best results. The accuracy of classification based on different FS-ratio varies according to the backbone and the dataset. This is reflected in the Table 4 and Table 5. This presents evidence for the effect of number of background tokens accepted in the local foreground selection attention. The results in the tables present the performance on different FS-ratio ranging from 0.1 to 1.0. The FS-ratio that gives the best performance is highlighted with bold letters, for each method and the respective dataset. For Oxford Flower-102 the best results are given by FS-ratio 0.3 for FRN(Conv-4) and 0.5 for BiFRN(Conv-4). For ResNet-12 backbone the FS-ratio 0.1 gives best results for both FRN and BiFRN methods. For the iNaturalist 2019 the Conv-4 gives best results with FS-ratio 0.3 for FRN and 0.5 for BiFRN. While for ResNet-12, 0.3 FS-ratio works

Table 3: Ablation study for effectiveness of LFS attention on Oxford Flowers102 and iNaturalist 2019 datasets.

Backbone	**Method**	**Oxford Flower-102**		**iNaturalist19**	
		1-shot	**5-shot**	**1-shot**	**5-shot**
Conv-4	Baseline	59.19±.22	83.10±.14	54.40±.22	76.21±.17
FRN (Conv-4)	Self-attention	74.20±.20	87.96±.13	64.21±.21	81.21±.14
	Local attention	74.89±.18	88.20±.13	64.68±.25	81.51±.11
	Selective attention	75.15±.17	88.65±.14	64.92±.20	81.70±.12
	LFS-attention	**76.46±.20**	**89.94±.13**	**66.62±.22**	**82.76±.15**
BiFRN (Conv-4)	Self-attention	75.11±.17	89.55±.15	64.54±.20	81.89±.12
	Local attention	74.91±.17	89.60±.15	65.03±.20	82.28±.12
	Selective attention	74.84±.17	89.79±.15	64.91±.20	82.36±.12
	LFS-attention	**75.71±.10**	**90.30±.13**	**65.36±.15**	**82.59±.11**
ResNet-12	Baseline	70.99±.20	86.99±.13	64.59±.29	81.65±.21
FRN (ResNet-12)	Self-attention	79.01±.20	91.50±.11	77.69±.11	90.52±.16
	Local attention	78.88±.16	91.89±.12	78.12±0.15	91.26±.13
	Selective attention	78.95±.21	92.26±.15	78.54±.12	91.98±.17
	LFS-attention	**79.85±.20**	**93.50±.11**	**79.61±.11**	**93.33±.15**
BiFRN (ResNet-12)	Self-attention	77.75±.17	89.94±.15	72.24±.16	87.20±.12
	Local attention	78.31±.17	90.08±.15	71.97±.16	87.12±.12
	Selective attention	77.75±.28	88.79±.15	72.44±.16	87.27±.12
	LFS-attention	**79.77±.12**	**90.87±.10**	**72.95±.15**	**87.67±.11**

Table 4: Performance with different select ratios on Oxford flowers-102 dataset for 5-shot and 1-shot.

FRN	**FS-ratio/ Oxford flower-102**				
	0.1	**0.3**	**0.5**	**0.7**	**0.9**
Conv					
5-shot	89.42	**89.94**	88.83	89.30	88.70
1-shot	76.03	**76.46**	75.26	75.16	74.20
ResNet					
5-shot	**93.50**	93.37	93.04	92.88	93.12
1-shot	**79.85**	79.73	79.43	79.46	79.45
BiFRN	**FS-ratio/ Oxford flower-102**				
Conv					
5-shot	89.69	89.32	**90.30**	89.44	89.91
1-shot	74.96	74.64	**75.71**	74.51	74.91
ResNet					
5-shot	**90.87**	90.40	89.90	89.81	90.24
1-shot	**79.77**	78.97	77.83	78.55	78.51

Table 5: Performance with different select ratios on iNaturalist 19 dataset for 5-shot and 1-shot.

FRN	**FS-ratio/ iNaturalist 2019**				
	0.1	**0.3**	**0.5**	**0.7**	**0.9**
Conv					
5-shot	82.69	**82.76**	82.48	82.52	82.40
1-shot	66.15	**66.62**	66.14	65.83	66.01
ResNet					
5-shot	93.21	**93.33**	92.82	93.06	92.82
1-shot	79.44	**79.61**	78.81	79.12	78.69
BiFRN	**FS-ratio/ iNaturalist 2019**				
Conv					
5-shot	82.39	82.05	**82.59**	81.82	82.50
1-shot	65.00	64.43	**65.36**	64.63	65.18
ResNet					
5-shot	86.68	**87.67**	86.59	87.37	85.92
1-shot	70.99	**72.95**	70.98	72.19	70.25

best for both FRN and BiFRN. In this way different datasets have different suitable FS-ratio for each method and the relevant backbone.

5 CONCLUSION

In this paper we propose a novel Local Foreground Selection(LFS) attention based module called Local Foreground Selection Module (LFSM), tailored to generate discriminative query and support features suitable for few-shot fine grained classification. The proposed attention optimizes large intra-class and small inter-class variations for the plant species classification task. The Foreground Selection attention works by highlighting the foreground and reducing the effect of background which contributes to alleviate the high intra-class variations. Secondly, we extract local spatial details to focus on the fine-grained details of the plant object with the local attention. Combining the local and foreground selection attentions enhances accuracy. The proposed LFSM module complements feature reconstruction methods and improve performance evident on plant species datasets and the ablation studies also validate the effectiveness of LFS attention.

ACKNOWLEDGMENTS

This research utilized Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045.

The research leading to this paper was partially supported by the EU Horizon Europe research and innovation Programme under Grant Agreement No. 101086461.

REFERENCES

Doersch, C., Gupta, A., and Zisserman, A. (2020). Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems*, 33:21981–21993.

Garcin, C., Joly, A., Bonnet, P., Lombardo, J.-C., Affouard, A., Chouet, M., Servajean, M., Lorieul, T., and Salmon, J. (2021). Pl@ ntnet-300k: a plant image dataset with high label ambiguity and a long-tailed distribution. In *NeurIPS 2021-35th Conference on Neural Information Processing Systems*.

Huang, H., Zhang, J., Yu, L., Zhang, J., Wu, Q., and Xu, C. (2021a). Toan: Target-oriented alignment network for fine-grained image categorization with few labeled samples. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):853–866.

Huang, H., Zhang, J., Zhang, J., Xu, J., and Wu, Q. (2021b). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. *IEEE Transactions on Multimedia*, 23:1666–1680.

Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., and Luo, J. (2019). Revisiting local descriptor based image-to-class measure for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7260–7268.

Li, X., Song, Q., Wu, J., Zhu, R., Ma, Z., and Xue, J.-H. (2023). Locally-enriched cross-reconstruction for few-shot fine-grained image classification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, X., Wu, J., Sun, Z., Ma, Z., Cao, J., and Xue, J.-H. (2020). Bsnet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Transactions on Image Processing*, 30:1318–1331.

Nguyen, H. Q., Nguyen, C. Q., Le, D. D., and Pham, H. H. (2023). Enhancing few-shot image classification with cosine transformer. *IEEE Access*.

Nilsback, M.-E. and Zisserman, A. (2008). Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sun, X., Xv, H., Dong, J., Zhou, H., Chen, C., and Li, Q. (2021). Few-shot learning for domain-specific fine-grained image classification. *IEEE Transactions on Industrial Electronics*, 68(4):3588–3598.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2018). The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wertheimer, D., Tang, L., and Hariharan, B. (2021). Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8012–8021.

Wu, J., Chang, D., Sain, A., Li, X., Ma, Z., Cao, J., Guo, J., and Song, Y.-Z. (2023). Bi-directional feature reconstruction network for fine-grained few-shot image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2821–2829.

Ye, H.-J., Hu, H., Zhan, D.-C., and Sha, F. (2020). Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8808–8817.

Zha, Z., Tang, H., Sun, Y., and Tang, J. (2023). Boosting few-shot fine-grained recognition with background suppression and foreground alignment. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, C., Cai, Y., Lin, G., and Shen, C. (2020). Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12203–12213.

Zhang, W., Liu, X., Xue, Z., Gao, Y., and Sun, C. (2021). Ndpnet: A novel non-linear data projection network for few-shot fine-grained image classification. *arXiv preprint arXiv:2106.06988*.

Zhu, Y., Liu, C., and Jiang, S. (2020). Multi-attention meta learning for few-shot fine-grained image recognition. In *IJCAI*, pages 1090–1096.