CHAPTER 1
INTRODUCTION

1.1 Background

Image classification represents one of the challenges in today's era of technological advancement. By enabling the grouping or categorization of data based on prior training, image classification simplifies complex tasks. As a result, there has been substantial progress in research within this field. However, a new challenge has emerged due to the inherent limitations of traditional image classification, which is primarily effective for images with clear or pronounced differences. It struggles when dealing with subtle distinctions, such as differentiating between various bird species or dog breeds. This limitation has given rise to a specialized research area called Fine-Grained Visual Classification (FGVC). The complexity of FGVC is illustrated in Fig. 1.1, where three visually similar dog breeds (Malamute, Eskimo, and Siberian Husky) pose a significant challenge for classification, as they belong to distinct classes despite their resemblance. These intricacies make FGVC a demanding task that often requires expert knowledge and involves considerable costs in its implementation. The main difficulty in visual recognition lies in accurately identifying objects that exhibit significant similarities or complex details. FGVC tackles this issue by facilitating the classification of objects with fine-grained distinctions, ensuring that even those with minor differences can be precisely categorized. Numerous deep learning methods [1] have been developed to tackle these challenges, broadly categorized into two approaches: part locating and feature encoding. The first category, part locating relies on bounding boxes around distinctive parts to identify subtle differences across input images [2]. Initially, these methods used annotations, such as bounding boxes or part labels, to locate discriminative regions. However, this approach faced limitations due to its reliance on manual annotations, which restricted its practicality [2]-[3].

To overcome these drawbacks, weakly supervised part locating methods were introduced, leveraging object detection to generate the information of part bounding box without manual intervention [4]. Despite this advancement, the method still has shortcomings, as it requires a bounding box generation module, which adds complexity to the network. To enable the automatic identification of multiple object parts, attention mechanisms have been introduced. These mechanisms help detect

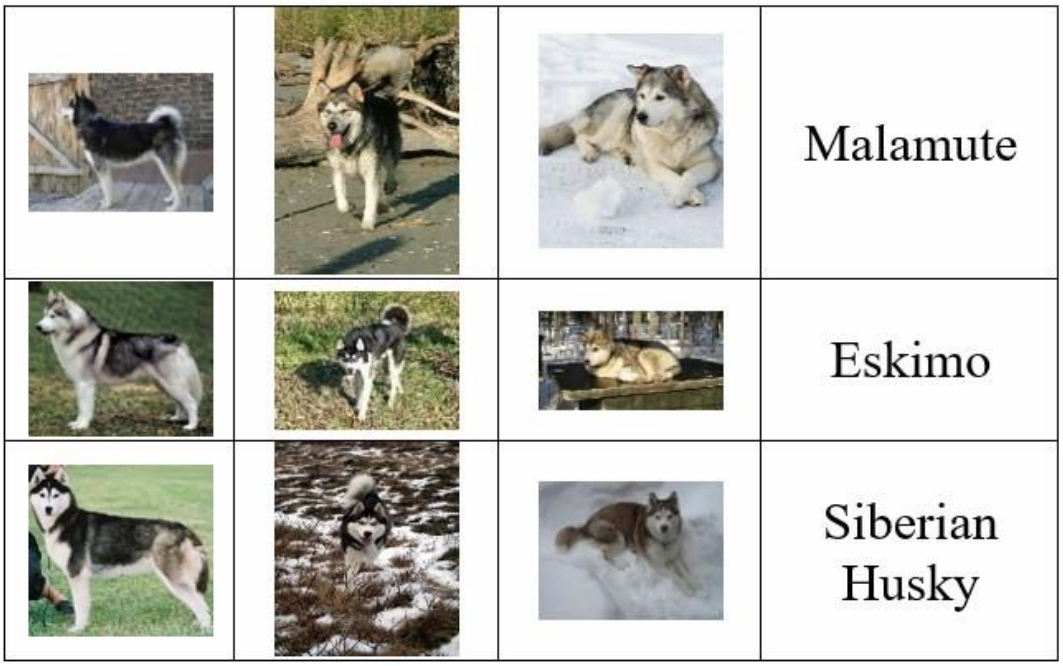

Fig. 1.1 The image above is an example of several dog breeds that are nearly similar. This presents a problem and a significant challenge for FGVC to perform image classification

discriminative or specific regions without the need for additional annotations [5] [6]. For instance, in [7], the approach involves combining discriminative regions from samples of different categories with existing labels, enabling the network to identify both common and specific features across categories as required by the initial problem. Beyond this, the latest advancement is the introduction of vision transformers (ViTs)[8] . The core of ViTs [8] lies in the Multi Head Self-Attention (MHSA Mechanism), it calculates relationships between input patches and captures them as a global feature representation. However, the capability to recognize distinct regions differs among various layers and attention heads within the MHSA in ViTs [9], [10].

1.2 Problem Identification

The previously proposed methods The previously proposed methods [11], [12], [13] have reported success in combining cross-layer features to enhance feature representation. However, direct fusion of cross-layer features may involve excessive information from unreliable regions, impacting the final classification performance. To better exploit cross layer features and reduce noise, [14] suggests the use of a Cross-Layer Refinement (CLR) module. In this module, the inputs are tokens selected from all previous layers, considered as cross-layer tokens. Based on these cross-layer tokens, several tokens are chosen using the MHV module, referred to as

refined tokens. Cross-layer features and refined features are extracted from cross-layer tokens and refined tokens, respectively, by the transformer layers. To avoid losing details, the assist logits operation is designed, as recommended in [15]. The refined feature and cross-layer feature are operated by assist logits to generate the final prediction results.

1.3 Objectives

The objective of this research is to enhance the performance of final prediction in Fine-Grained Visual Classification (FGVC) tasks through the integration of ensemble learning methods and the development of specialized modules to 8 address issues with the multi-head self-attention's ability to comprehend local features. This research is expected to provide significant benefits, including improved accuracy in fine-grained recognition and optimization of ViT usage in fine-grained recognition tasks. Additionally, this study has the potential to contribute to the development of more reliable and efficient deep learning techniques to overcome the challenges of high-level fine-grained recognition.

1.4 Scope of Work

To keep the experiment from being too long, this thesis limits the works as follows:

1. The Backbone Model to be utilized is a Vit-B-16 [8], which has been pre-trained on the ImageNet21K dataset.

2. During training, techniques such as random cropping, horizontal flipping, and colour augmentation were applied. However, during testing, center cropping was used.

3. The model training process utilized only classification labels without additional annotations.

4. In this study, the images of Oxford IIIT-Pet[16] were resized to a resolution of 448×448 pixels. The dataset comprises 3,680 training samples and 3,669 testing samples, containing two animal categories: dogs and cats.

1.5 Expected Result

The proposed final prediction is expected to address the limitations of direct fusion of cross-layer features by effectively extracting refined tokens and refining cross-layer features. By utilizing the MHV module to select tokens and incorporating the assist logits operation, the proposed methods aim to enhance feature representation while reducing noise. It is anticipated that the integration of these techniques will lead to improved classification performance compared to previous methods. Through performance parameter and comparative analysis, the thesis expects to improving final prediction of the output.

1.6 Research Methodology

This thesis used fundamental study and experiment based on work packages (WP). These are the WP that is used in this thesis:

- WP 1: Literature review is conducted to comprehend the fundamental concepts related to fine-grained recognition, Vision Transformer, and other relevant techniques.
- WP 2: Model Selection and Configuration is choosing an appropriate Vision Transformer (ViT) model for experiments. Model configurations, including hyperparameters and training settings, will be set.
- WP 3: Implementation of Improvements Developing modifications and enhancements to the final prediction to improve its ability to improve the results.
- WP 4: Performance analysis, conducting a series of experiments on various fine-grained recognition datasets using suitable performance metrics. The results will be used to measure the effectiveness of the proposed improvements.
- WP 5: Analysing experimental data to evaluate the extent to which the proposed improvements successfully address the issue of comprehending local features in input images.