Analysis on Fine-Grained Image Recognition

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY (M.Tech)
IN
ARTIFICIAL INTELLIGENCE

Submitted by

Utkarsh Singh (2K21/AFI/29)

Under the supervision of

Prof. (Dr.) Shailender Kumar

**DEPARTMENT OF COMPUTER SCIENCE &
ENGINEERING**
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2023

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Utkarsh Singh**, Roll No. 2K21/AFI/29 students of M.Tech (ARTIFICIAL INTELLIGENCE),hereby declare that the project Dissertation titled "**Analysis on Fine-Grained Image Recognition**" which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi — Utkarsh Singh

Date: 29 May 2023 — (2K21/AFI/29)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "**Analysis on Fine-Grained Image Recognition**" which is submitted by **Utkarsh Singh**, Roll No. **2K21/AFI/29**, **Department of Computer Science & Engineering** ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi **Prof. (Dr.) Shailender Kumar**

Date: 29 May 2023 **Professor, Ph.D. Coordinator**

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to **Dr. Shailender Kumar**, Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi for his continuous guidance and mentorship that he provided us during the project. He showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. He was always ready to help us and clear our doubts regarding any hurdles in this project. Without hir constant support and motivation, this project would not have been successful.

Place: Delhi Utkarsh Singh

Date: 29 May 2023 2K21/AFI/29

Abstract

In addition to being able to distinguish between a dog and any bird, humans are amazingly capable of visualising or picturing fine-grained detail, we can also discern an American Bulldog from an English Bulldog To educate or train machines to apprehended in a fine-grained manner, fine-grained image recognition was brought to the academic community. FGIR is used in different academic and business settings. applications in both industry and academia. These applications have had a positive impact on fields like conservation and commerce. The objective of FGIR has been a long-term goal of the deep learning community. It is to retrieve and identify images from several intra-class categories of a super-class category, such as various dogs, aircraft, and plant species, retail product types, etc. Determining how to distinguish between items that are remarkably similar in appearance but different in fine-grained attributes is, consequently, the main difficulty.

Intricate nuances and subtle visual signals that might not be clearly seen in low-resolution photos are frequently used in fine-grained image identification. We have put efforts into an approach to this problem using super-resolution to recover finer details, such as textures, patterns, or minor differentiating traits. These details are essential for the categorization or localization of fine-grained categories. Usually, accurate localization of objects or particular regions of interest within an image is necessary, which is performed by specialised feature detection. We also performed a survey on recent trends in the field, and we tried to identify and classify the generalised techniques used in FGIR into groups based on the types of techniques followed and the results that we were trying to achieve.

Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

O	Output Set, m
P	Probaility
Wi	Weight of ith feature
FGIR	Fine-Grained Image Recognition
CNN	Convolutional Neural Network
ViT	Vision Transformer

Chapter 1

INTRODUCTION

We understand an image as a 2D geometric function in spatial coordinates. The pixels of an image can be represented as a matrix in this space. This function can be manipulated in order to extract intrinsic patterns. Several techniques, like filtering, can be used to modify or even enhance an image. Filters in image processing are used for extracting image segments of interest, highlighting them, or even removing some other features like the median filter, the Gaussian filter, the Laplacian filter, the Sobel filter, and image scaling.

1.1 Object Detection

Identifying and detecting items of interest within digital photographs or video frames is the problem of object detection, which falls under the umbrella of computer vision. It is a key issue in the study of computer vision and has a wide range of uses in robotics, autonomous vehicles, surveillance, and picture retrieval, among other fields. Identifying items in an image and properly determining their spatial position by tracing bounding boxes around them are the two objectives of object detection. This makes it feasible to create intelligent systems that can communicate with the outside world by enabling robots to comprehend and interpret visual data.

Traditionally, handmade characteristics and machine learning classifiers were used by object identification algorithms to identify items. However, the area has made great strides thanks to recent developments in deep learning, notably with convolutional neural networks (CNNs). Deep learning-based algorithms have shown to be significantly more accurate and efficient than traditional methods in object identification tasks. Modern object detection systems frequently employ a one- or two-stage methodology. In the two-stage method, the computer system first creates a list of region recommendations that represent probable places for objects in the image. These ideas are then categorised in order to establish whether or not they include an object. The system may lessen the computing load by concentrating solely on pertinent regions thanks to this two-step method.

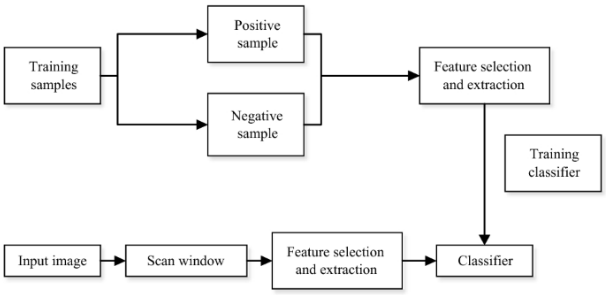


Figure 1.1: Flow chart representing Object detection

Without the requirement for explicit region proposal construction, object identification is carried out immediately in the one-stage technique. Each grid cell in the full image forecasts a set of bounding boxes together with the class probabilities. Comparing this procedure to the two-stage approaches, it is typically faster but may result in some accuracy loss.

The creation of object detection algorithms mainly relies on labelled datasets that offer instances of things and their related bounding boxes. The effectiveness of object identification algorithms is trained on and assessed against these datasets. The COCO (Common Objects in Context) dataset [12] and ImageNet [13] are a few well-liked object detection datasets. With the introduction of cutting-edge designs like Faster R-CNN [14], SSD (Single Shot MultiBox Detector) [6], YOLO (You Only Look Once) [15] and, EfficientDet [16], object detection has recently experienced fast advancement. These models use deep learning methods to deliver cutting-edge performance in terms of accuracy and speed, including convolutional neural networks and numerous optimisations. Research on object identification is still strong, and it is anticipated that further developments in deep learning, hardware acceleration, and dataset accessibility will further improve its capabilities. A vast range of applications across several sectors and areas are made possible by the capability to precisely and effectively recognise objects in photos and videos.

1.2 Image Super-Resolution

The technique of increasing a picture's resolution and quality to create a higher-resolution version with more detailed information is known as image super-resolution. It is a significant problem in image processing and computer vision to reconstruct lost high-frequency information and increase visual clarity. The need for image super-resolution arises in a variety of situations, including enhancing the visual quality of outdated or compressed images, enhancing the resolution of medical images for precise diagnosis, and enhancing the quality of images in video applications. The difficulty with picture super-resolution is precisely predicting the high-frequency information that is absent. It is typically challenging to immediately recover the lost high-frequency features since the low-resolution image only offers a small amount of information. Researchers have created a number of strategies based on various concepts and methods to overcome this.

Interpolation and filtering techniques were frequently used in traditional super-resolution methods to boost resolution, but the results were frequently fuzzy or unrealistic. Convolutional neural networks (CNNs) have demonstrated astounding performance in picture super-resolution since the development of deep learning. These deep learning-based methods pick up complicated visual patterns and produce more precise high-resolution images by learning from huge datasets. The exploration of innovative network designs, loss functions, and training techniques has been the main focus of recent developments in picture super-resolution. To produce more visually realistic and detailed high-resolution photographs, generative adversarial networks (GANs) [17] have been used in super-resolution. In order to increase the preservation of small features and improve reconstruction quality, attention methods and residual connections have also been added into network topologies.

The creation of single-image super-resolution techniques, which attempt to improve the resolution of a single picture without depending on numerous input images or prior knowledge, is another field of research in image super-resolution. These techniques use advanced deep learning models to extrapolate high-frequency information from input photos with poor resolution. In order to solve numerous difficulties with picture quality concurrently, researchers are also investigating the coupling of super-resolution with other tasks, such as image denoising and inpainting. This encourages the creation of more robust and thorough picture enhancing techniques. Medical imaging, remote sensing, surveillance, digital photography, and video processing are just a few of the fields where image super-resolution finds important use. It improves picture analysis, provides better visualisation, and improves the user experience overall by increasing image resolution and quality.

1.3 Fine-grained Image Recognition

The goal of FGIR is to handle objects that fall under different subcategories of the same meta category. We have seen different techniques of image processing used in order to extract the most meaningful parts of the image or the areas of interest to the problem in this domain. Different image processing techniques work to extract the meaningful parts of the image and also work to eliminate errors or filter out non-required parts of the image. Along with this, we have also seen different techniques used with computer vision models.

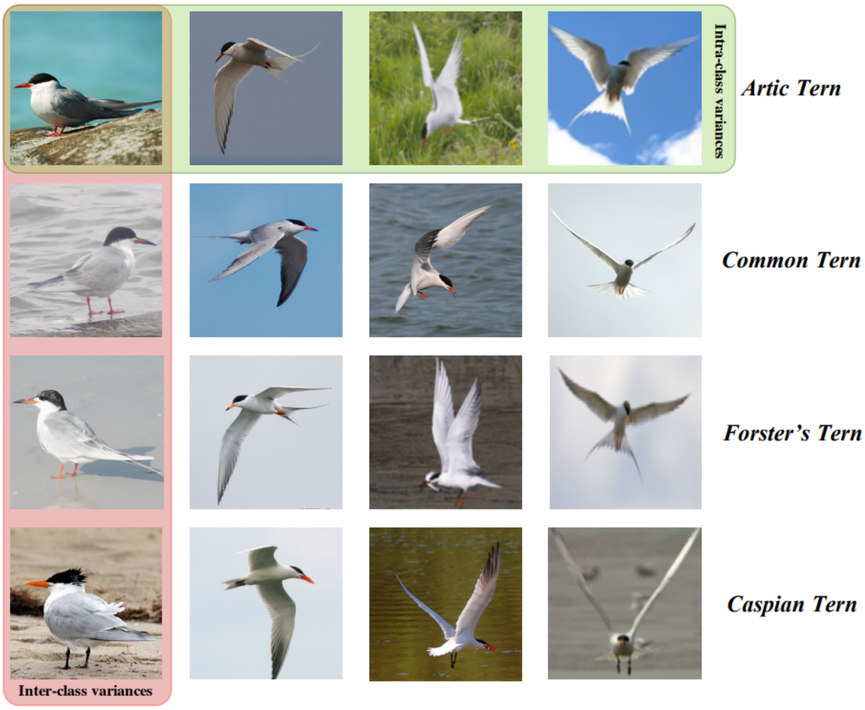


Figure 1.2: Fine-grained grained image intricacies [9]

The considerable homogeneity and minute variations across item categories make fine-grained image identification jobs difficult. We try to enhance the features using super-resolution by boosting the discriminating ability of the learned features. The model may be more resistant to intra-class fluctuations and enable better distinction between items with similar appearances. The thought behind this is that fine-grained image collection in certain situations could be dominated by low-resolution pictures. These photos may be up-scaled to a greater quality, which effectively increases the quantity of training data available. This can help models generalise while reducing the drawbacks of having less training data. Several approaches have been used in the past to perform feature localization. including bounding-box annotations [18], region proposal methods [19], part-based approaches [20], and attention mechanisms. Boundingbox annotations can help localise items in an image by anticipating bounding boxes. coordinates with object identification algorithms such as Faster R-CNN, YOLO [15], or SSD. Normally, these algorithms are trained on large object identification datasets, but they may also be adjusted or customised for a particular fine-grained image recognition application. This study mostly performs the above procedure on the CUB200-2011 dataset [21].

The CUB-200 dataset, which includes bird photos with class labels, was the first thing we acquired. We divided the dataset into subgroups for training and testing. We annotated the CUB200 dataset with bounding boxes in a YOLO-compatible style.

1.4 Overview

The objective of fine-grained image identification is to identify and categorise objects that fall into similar categories but have distinct qualities that may be difficult for humans or traditional computer vision systems to discern. Tasks requiring fine-grained identification include recognising several bird species, dog types, flower types, or automobile models.

Because the visual distinctions between subcategories may be fairly subtle and frequently include minute local characteristics or particular areas of an item, fine-grained image identification is a challenging task. A model must have the ability to concentrate on these minute details and recognise the small visual signals that distinguish one subcategory from another. Large labelled datasets with fine-grained annotations, where each item is accurately marked with its associated subcategory name, are often needed for accurate fine-grained identification. These datasets are used to train convolutional neural networks (CNNs), which are particularly good at extracting hierarchical features from visual input, and other deep learning models.

Several methods have been developed to address the difficulties involved with fine-grained picture recognition. One such method involves combining localization information with classification in order to not only identify the item but also to draw attention to the distinguishing features or areas that influenced the categorization choice. Understanding these crucial hints makes it easier to distinguish between related subcategories.

Another essential method for fine-grained recognition is transfer learning, which involves fine-tuning or using feature extractors for pre-trained models built on massive datasets (like ImageNet). This makes use of the generic characteristics that were learned during the pre-training phase and enhances the models' ability to identify fine-grained categories with less training data.

Furthermore, attention methods enable models to focus on useful regions while disregarding irrelevant background regions, which has proven successful in fine-grained image identification. The model's capacity to pay attention to minute details is improved by attention mechanisms, which also increase accuracy.

Numerous sectors, including biodiversity monitoring, species conservation, product recognition, and fashion analysis, use fine-grained picture recognition. It provides more accurate item recognition and classification, enhancing comprehension, analysis, and decision-making based on visual information.

The creation of cutting-edge architectures, network regularisation strategies, and bigger fine-grained datasets will continue to drive advancements in fine-grained image identification as deep learning and computer vision research advances. These developments will increase the capacity of robots to comprehend and interpret minute changes in visual appearance, making them useful tools in a variety of real-world applications.

Improvements in accuracy, efficiency, and resilience have been prioritised as recent research trends in object identification have centred on tackling these issues. The availability of large datasets, unique network topologies, and improvements in deep learning have all led to a number of important innovations and trends. Here is a summary of several significant object detection research trends: One-stage detectors have drawn a lot of interest because of how straightforward and effective they are. Popular examples are the YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector) models. Without explicitly generating region suggestions, these models carry out item detection across the board. One-stage detectors are currently being improved in order to increase accuracy while preserving real-time performance.

FGIR is utilised in the business world to classify intricate, specific classifications that are essentially difficult to classify using traditional image recognition algorithms. FGIR-related techniques have been created and have excelled in well-known computer vision benchmarks. The FGIR benchmarks, however, were created in a static, controlled environment. Another significant drawback of the current FGIR datasets is that much of the image frame is usually taken up by the object of interest, which makes them unrepresentative of real-world applications.

We classified the FGIR techniques into three categories:

1) Localization and Classification
2) External Information
3) End-to-End Encoding

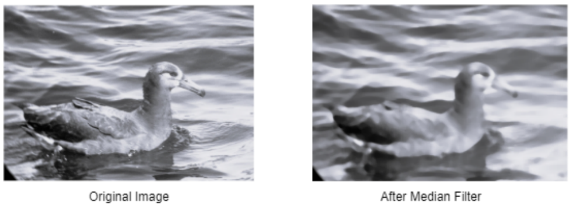
With regard to the most recent advancements and the direction that the field is heading, we have carefully summarised the papers on this topic. The scarcity of labelled photos, particularly those belonging to a particular meta-category and its corresponding sub-categories, severely restricts the use of sample data in research at the moment. One of the main obstacles preventing this discipline from progressing at the moment is the absence of a common dataset. Some academics have attempted to solve this issue by directly getting labelled data from the web and using the ones that can be applied in this field. Since unlabeled photos are widely available on the internet and within the deep learning community, several researchers have used these unlabeled images and carried out image captioning as a challenging technique. The outcome is a labelled picture that may be used directly in this issue.

Given that a lot of work on picture captioning is being done on a larger scale than that of this domain, we can consider doing FGIR using this technique. However, the paper claims that as a model grows more exact, the necessity for it to carry out fine-grained analysis increases. It qualifies as being incredibly accurate in terms of minute details.

Filters in image processing are used for extracting image segments of interest or highlighting or even removing some other features like:

1.4.1 Median Filter

Remove noise from images and use a simple technique of replacing the central pixel with the median pixel value of the surrounding pixels. This takes linear time complexity and is efficient.



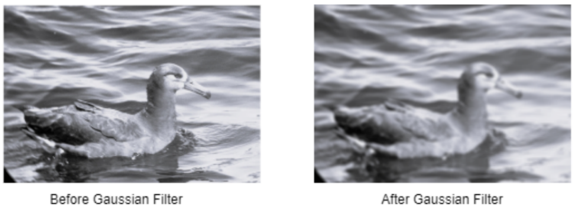
Original Image

After Median Filter

Figure 1.3: Median filter to a black and white image

1.4.2 Gaussian Filter

Also removes noise from images but mainly blurs them. It uses a kernel that follows a Gaussian distribution so that the image is filtered In a way, it focuses on normal distribution.



Before Gaussian Filter

After Gaussian Filter

Figure 1.4: Gaussian filter to a black and white image

1.4.3 Sobel Filter

Sobel filters use kernels that maximise the vertical and horizontal lines in the image. The vertical and horizontal edges pop out of the image and are shown combined together, which works as extracting the edges in the image. It is a computationally cheaper technique for identifying strong edges in an image where we do not need to spend much computation and time on training a machine learning model to identify image edges.

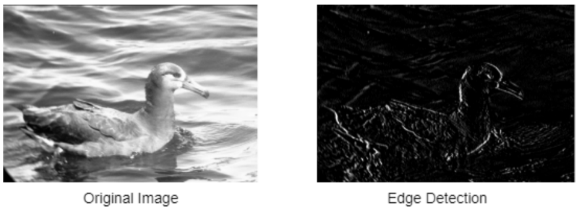


Figure 1.5: Sobel filter to a black and white image

1.4.4 Image Scaling

Image downscaling and upscaling refers to reducing and increasing the resolution of the image respectively.

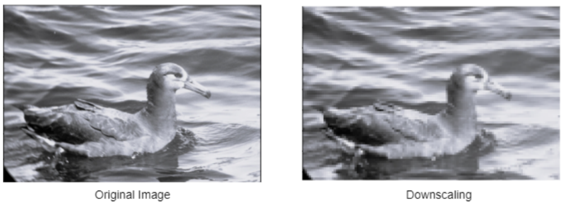


Figure 1.6: Image Downscaling

Figure 1.7: Image Upscaling

1.5 Problem statement

The following is a description of the research questions we set in order to choose the research articles for evaluation that were relevant to our field:
1) What are the available literature sources for FGIR
2) What are the different techniques that can be used to perform FGIR
3) Which datasets are used for FGIR.
4) Also to come up with an approach to perform the task.

1.6 Identified Problem

The current issue is the identification and categorization of minute elements in photographs. The process of finding and classifying objects or features that display minute variances and subtleties within a certain category is known as fine-grained image detection. The difficulty comes from the richness and intricateness of these minute features, which frequently call for a high level of visual discernment and specialised expertise. Examples of such situations include recognising various automobile models, differentiating between similar bird species, and identifying certain varieties of flowers.

Since they mostly depend on generic visual cues that might not capture the small distinctions required for exact classification, current image detection and classification algorithms frequently struggle with reliably discriminating fine-grained details. Therefore, sophisticated methods that can accurately collect and analyse these complex aspects are required. The objective is to create a reliable and accurate fine-grained image detection system that can categorise objects with minute variances in an accurate manner. To enhance classification performance, this system should make use of cutting-edge computer vision algorithms and deep learning approaches and maybe include domain-specific information. Numerous fields, such as botany, the motor industry, art, and animal protection, will benefit from the answer to this issue. The suggested technology would improve our capacity to comprehend and engage with the visual world at a very granular level by precisely recognising and categorising fine-grained elements in images, opening up a wide range of useful applications and insights.

Chapter 2

LITERATURE REVIEW

Object detection has frequently employed two-stage detectors, such as Faster R-CNN (Region-based Convolutional Neural Networks) [14]. They produce prospective object areas using a region proposal network (RPN), categorise, and refine those suggestions. Increasing speed and efficiency has been the focus of recent research on two-stage detectors, frequently by incorporating innovative network components or optimisation approaches.

Traditional object detectors use known anchor boxes to compare bounding boxes to ground truth boxes. Anchor-free detectors, on the other hand, do away with the need for anchor boxes and directly predict the locations and sizes of objects. With the growing demand for real-time applications and resource-constrained situations, researchers have concentrated on creating efficient object detection algorithms. This covers network architecture design optimisation, knowledge distillation, quantization, pruning, and other approaches for reducing computational complexity and memory footprint while maintaining considerable accuracy. Transformer [4] designs, popularised by their effectiveness in natural language processing, have also found their way into object identification. Models like DETR (DEtection TRansformer) [22] use techniques of self-attention to identify objects. Transformer-based detectors offer a versatile framework for managing object detection tasks and have demonstrated promising accuracy and speed results.

Self-supervised learning has gained popularity as a method of using vast volumes of unlabeled data for pretraining object identification models. Models can capture richer representations and generalise to downstream tasks more effectively by learning from unlabeled input. Pretext tasks and contrastive learning have both demonstrated significant promise for enhancing object identification abilities during self-supervised pretraining.

It is frequently difficult for object identification algorithms trained on one domain to generalise successfully to other domains or undiscovered classes. Domain adaptation methods for transferring knowledge from labelled source domains to unlabeled target domains have been the subject of recent study. Furthermore, few-shot object identification tries to identify things with just a few labelled instances, necessitating models to generalise from sparse training data. Small disruptions can cause object detectors to be misclassified or have their bounding boxes altered in adversarial attacks. To make object detectors more resistant to adversarial assaults, research has examined strategies including adversarial training, defensive distillation, and robust loss functions. The continuous attempts to improve object identification algorithms' precision, effectiveness, generalizability, and resilience are reflected in these research themes. It is anticipated that ongoing research into innovative network topologies, training methods, and dataset developments will lead to additional breakthroughs and push the limits of object identification in a variety of practical applications.

The goal of fine-grained image recognition, often referred to as fine-grained visual categorization, is to discriminate between visually identical item categories with minute changes. Fine-grained image recognition digs deeper to detect certain subcategories or species within a given class, in contrast to classical object identification, which seeks to categorise things into broad categories like dogs, cats, or vehicles.

The CUB-200 dataset, which includes bird photos with class labels, was the first thing we acquired. We divided the dataset into subgroups for training and testing. We annotated the CUB200 dataset with bounding boxes in YOLO. Since the programme's beginning approximately 20 years ago, significant progress has been made. Particularly, deep learning has been a potent technique for discriminative feature learning, which has produced amazing advances in the field of fine-grained image recognition. The practical application of these techniques in a variety of application scenarios has been significantly improved by deep learning-enabled FGIR. Due to the minor inter-class changes brought on by extremely comparable sub-categories and the significant intra-class variations in poses, sizes, and rotations, the problem's fine-grained nature presents researchers with a different and difficult challenge. As such, it is the antithesis of general image analysis. Since the beginning of the use of neural networks in the area of computation with the introduction of convolutional neural networks by Y. Lecun et al. [10], researchers have started the classification and recognition of images. Further advancements were made, including One major improvement was made by Alex Krizhevsky et al.'s [23] accommodating style.

We have looked at a number of recent publications on fine-grained image recognition from prestigious journals. Because of the significant intra-class variance and little inter-class variation in fine-grained picture recognition, it is a challenging challenge. Deep learning-driven technologies are profiting from deep learning advances.

Over the past four years, many developments have occurred in the area. of fine-grained image identification have been made. Objects belonging to a certain category, such as several bird species or Dog breeds, which frequently have identical external characteristics, are distinguished using fine-grained image recognition. A list of some significant advancements in this area is shown below:

2.1 Deep Neural Networks

A family of machine learning models known as deep neural networks (DNNs) are modelled after the structure and operation of the human brain. They are made to learn and represent intricate correlations and patterns in data, which enables them to complete a variety of tasks requiring advanced comprehension and judgement. A layer-based design made up of linked nodes, commonly referred to as artificial neurons or units, is the basis of a deep neural network. These units are arranged in layers, with an input layer, one or more hidden layers, and an output layer being the standard layer configuration. DNNs get their name from the hidden layers, which can be numerous and extensive.

Weights, which signify the significance or strength of the connection, are attached to the links forming the network's nodes' connections. In order to properly tune the model to provide correct predictions or classifications, the network learns to modify these weights based on input data and intended output throughout the training phase. Deep neural networks' main advantage is its capacity to autonomously learn hierarchical data representations. By drawing on the representations that the network's earlier levels learnt, each layer learns ever-more sophisticated characteristics. Intricate patterns and relationships in the data may be captured by DNNs thanks to this hierarchical learning, which makes them very good at tasks like speech recognition, picture identification, and natural language processing.

Modern convolutional neural networks and other deep neural network architectures continue to be used in fine-grained image recognition models. These designs have shown considerable gains in accuracy and generalisation power when used in conjunction with methodologies like transfer learning and network ensembles. We have seen CNN being used in some form or another, with different loss functions and regularizations.

2.1.1 Convolutional neural networks

Convolutional neural networks (CNN) are inspired by the visual cortex of human and animal brains. It extracts information from the images using hidden layers, consisting mainly of the convolution layer and the fully connected layer. Images on CNN are seen as a matrix of pixels.

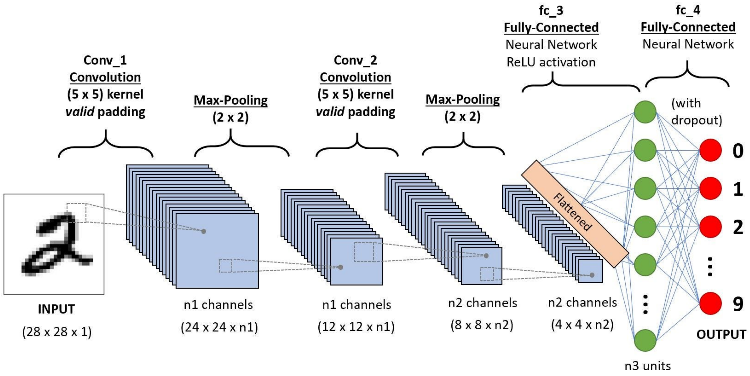


Figure 2.1: General CNN architecture [10]

Convolutional Neural Networks (CNNs) are a particular kind of deep neural network created for the processing and analysis of organised grid-like input, such as photos and movies. The foundation of computer vision tasks including image classification, object identification, picture segmentation, and others is now provided by CNNs. The capacity of CNNs to automatically learn hierarchical representations of visual data is its distinguishing feature. Convolutional layers, pooling layers, and completely linked layers, among others, are used to do this.

The foundational units of CNNs are convolutional layers. They use a series of trainable filters or kernels to extract regional information and record spatial correlations from the input data. The network learns to recognise edges, textures, and more complex patterns at various sizes and orientations by swiping these filters across the input. Convolutional layer results are referred to as feature maps.

The feature maps are downsampled using pooling layers to reduce their spatial dimensions while preserving the most important data. Max pooling, which chooses the highest value within an area, and average pooling, which calculates the average value, are common pooling processes. Translation invariance can be attained and computational complexity is decreased by pooling.

Fully linked layers, often referred to as dense layers, are frequently positioned towards the network's edge. They translate the high-level characteristics acquired by the convolutional layers to the required prediction or output classes. The network may make final judgements based on the retrieved features thanks to fully linked layers. The process of training a CNN entails providing it with labelled training data made up of pairs of input samples and the intended outputs they should produce. Backpropagation is a technique used by the network to learn from these instances where mistakes in the output predictions are transmitted backward through the network to update the weights and enhance the performance of the model.

2.1.2 Region-based Convolutional Neural Network(R-CNN) [1]

CNN uses a region proposal generator, which uses a selective search algorithm. The selective search algorithm performs segmentation and is used as an object detection algorithm. It uses support vector machines to perform object detection. It calculates the colour, texture, size, and fill similarity of images and feeds them to SVM. The output of the selective search algorithm is a feature-extracted image, which is fed to a backbone CNN.

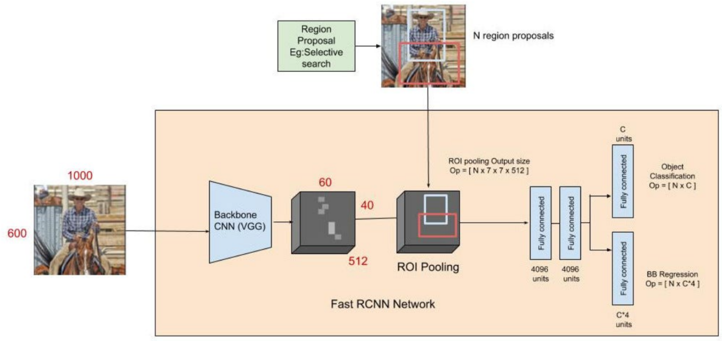


Figure 2.2: R-CNN architecture [1]

R-CNN, a ground-breaking object detection framework, transformed the industry by fusing deep learning with region proposal methods. A group of researchers under the direction of Ross Girshick first presented it in 2014. The challenge of object detection, which entails locating and categorising things inside an image, is addressed by the R-CNN method. Traditional approaches, which were inaccurate and computationally costly, depended on handmade features and sliding window techniques.

Convolutional neural networks (CNNs) were used by R-CNN, a pioneering technology, to detect objects. The three primary components of the framework are object categorization, feature extraction, and region proposal. In the first stage, probable item bounding boxes inside the picture are generated using a region proposal technique, such as Selective Search. The CNN then processes these areas, which are assumed to be home to objects. The next step is to shrink each area proposal to a defined size before feeding it into a CNN that has already been trained using an architecture like AlexNet or VGGNet. From the area of interest, the CNN derives high-dimensional feature representations that encode both spatial and semantic data.

2.1.3 Faster R-CNN

The extracted features are then fed into a set of support vector machines (SVMs) [24] for classification. The SVMs learn to classify the presence or absence of specific object categories within each region proposal. Additionally, a bounding box regression is performed to refine the location of the proposed bounding boxes.

During the training phase, R-CNN employs a multi-stage process. First, the CNN is pre-trained on a large-scale image classification dataset, such as ImageNet, to learn generic visual features. Then, the CNN is fine-tuned using the region proposals and their associated ground-truth labels.

R-CNN demonstrated significant improvements in object detection accuracy compared to previous methods. However, it suffered from slow inference speed due to the need for individual CNN forward passes for each region proposal.

Faster R-CNN is a speed improvement on R-CNN. It performs a region of interest pooling layer (RoI), which focuses on the part of the image that we are interested in. This RoI layer is generally smaller and removes the need for the selective search algorithm. which is the main reason R-CNN was slower. The output of the pooled image contains the extracted features of the image, which contain the area of interest of the image. These extracted features are then input to the faster R-CNN network for classification, and the image's belongingness to the output class is calculated and represented in the form of a bounding box. Except for the additional layer used to predict segmentation, it is quite similar to Faster R-CNN. The second step, which operates concurrently, predicts class, generates bounding boxes, and produces a binary mask for each RoI. This stage of region proposal production is the same in both architectures.

Later efforts expanded on R-CNN to solve its drawbacks. Fast R-CNN is one notable modification that shared the convolutional features across all area suggestions, greatly accelerating the inference procedure. Faster R-CNN is another development that included a region proposal network (RPN) incorporated into the CNN architecture, allowing end-to-end training and substantially boosting speed and accuracy.

R-CNN and its offspring have established frameworks for object detection that have influenced a large number of further models. Numerous applications, including as autonomous driving, video monitoring, and image analysis, have widely embraced them.

Overall, the combination of deep learning and region recommendations was pioneered by R-CNN, opening the door for major improvements in object identification and laying the groundwork for further state-of-the-art methods.

2.1.4 Mask R-CNN [2]

Faster R-CNN, which integrates region proposal generation and object categorization into a single network architecture, serves as the foundation for Mask R-CNN. Mask R-CNN extends this architecture to predict a binary mask for each instance of an object in addition to object recognition, offering precise pixel-level segmentation.A backbone network, a region proposal network (RPN), and a mask prediction network make up the architecture of Mask R-CNN. The backbone network analyses the input picture and extracts a feature map that encodes rich semantic and spatial information. It is commonly built on a deep convolutional neural network (CNN) architecture like ResNet or VGGNet.

Mask R-CNN is similar to Faster R-CNN with an added layer for segment detection. The problem with Faster R-CNN was that it did not perform segmentation or was not able to identify the shape of the object that had been detected in the image. Mask R-CNN contains a backbone network, which is a faster R-CNN, and a region proposal network (RPN), which is a mechanism to determine an objectness score. The output is then passed through a procedure that performs masking. Particular masks are used so that, by using these masks, it differentiates the segments that are more discriminative. These masks are learned using a RoIAlign mechanism. This mechanism performs RoI pooling and generates fixed-sized regions of interest.

The mask prediction network classifies objects and segments instances at the pixel level using the region suggestions from the RPN. It anticipates the class label of each suggestion for object classification. For each area suggestion, it simultaneously creates a binary mask containing the precise pixels that correspond to the object instance.

The multi-task loss function used by Mask R-CNN during training consists of three parts: the classification loss, the bounding box regression loss, and the mask loss. The mask loss gauges how closely the expected and actual masks resemble one another, promoting precise segmentation at the pixel level.

Mask R-CNN has developed into a commonly used framework in computer vision applications due to its impressive performance in segmentation tasks. Tasks requiring accurate object knowledge, such as splitting overlapping instances, segmenting fine-grained objects, and tracking objects between frames, have greatly benefited from the use of Mask R-CNN. Following models built on its basis and explored modifications to increase speed, accuracy, and efficiency It provides exact localization. classifying and segmenting several object instances within an image at the pixel level.

Advancements in a variety of fields, including robotics, autonomous driving, and medical imaging, have all been made possible by the capacity to produce high-quality masks.

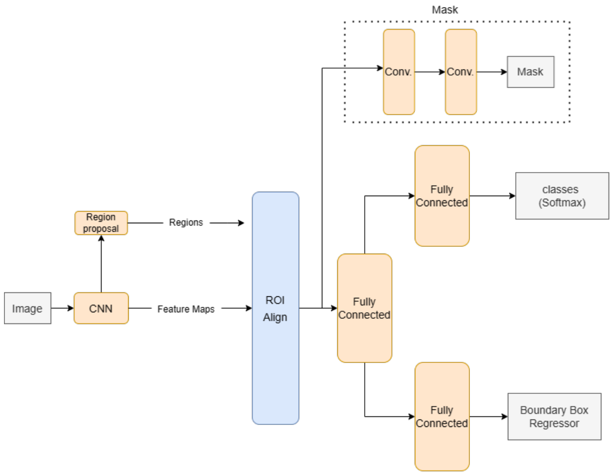


Figure 2.3: Mask R-CNN architecture [2]

2.2 Part Based Approach

In order to manage minuscule distinctions among comparable categories, part-based techniques have been widely employed in fine-grained image identification. These techniques locate and identify regions or elements of an item that are specific to a particular category and may be used to identify that thing. To concentrate on these relevant areas, attention mechanisms and spatial transforms are frequently used. Part information can also be used to represent features.

2.2.1 ResNet-50 [3]

The well-known vanishing/exploding gradient is one issue ResNets aims to address. This is due to the fact that when the network is too deep, the gradients required to compute the loss function simply decrease to zero after numerous chain rule operations. As a result, learning is not taking place since the weights' values are constant.

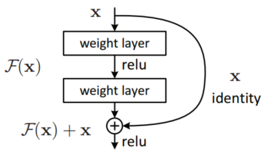

x
weight layer
relu
$\mathcal{F}(\mathbf{x})$
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2.4: ResNet residual block [3]

Instead of learning the intended output directly, this skip link allows the network to learn residual mappings by adding the original input to the output of the stacked layers.

In order to overcome the difficulty of training very deep neural networks, the ResNet architecture was created. This problem is addressed by ResNet-50, which introduces the idea of residual learning. The network can skip one or more layers by using a ”skip connection” or ”shortcut connection,” which makes the gradient flow more readily during training. The issue of vanishing gradients affects traditional deep neural networks, where the gradients get smaller exponentially as they backpropagate through several layers. This hampers the performance of deep networks and makes it harder to train them successfully.

ResNet is a deep neural network that was inspired by VGG-16. ResNet solved the problem of vanishing or exploding gradients, which happened with networks that were very deep (nearly more than 20 layers deep),This is when the gradient becomes too small or too large for the computation to be performed. This was an early problem with deep neural networks. However, ResNet tried to solve this problem to a certain extent. It uses skip connections.which allows a neuron in a particular layer to connect to another neuron a fewlayers ahead. This reduced the high connectivity of the neurons among layers. This also helped in skipping the layer of neurons that were affecting the performance.of the network.
Increasing the number of channels while progressively decreasing the spatial dimensions After that, fully linked layers get the final output for categorization. Convolutional layers, pooling layers, fully linked layers, and residual blocks are among the 50 layers of the ResNet-50 architecture.

The essential element of ResNet that enables the learning of residual mappings is the residual blocks. Along with the skip connection, each residual block normally has two or three convolutional layers. ResNet-50 was created especially for jobs involving picture categorization. An input picture is processed through a number of convolutional layers, pooling layers, and residual blocks during training. ResNet can train very deep networks efficiently and accurately by using residual blocks.

These architectures helped deep learning in computer vision evolve and provided a foundation for several later models. ResNet-50’s architecture was carefully planned, striking a balance between computing complexity and depth. Deeper variations of the ResNet family, such as ResNet-101 and ResNet-152, which have even greater accuracy but cost more to compute, have been added by researchers. On benchmark datasets, it performs superbly and outperforms earlier models in terms of accuracy. Object recognition, picture segmentation, and image classification are just a few of the computer vision applications that make use of the architecture as a backbone network.

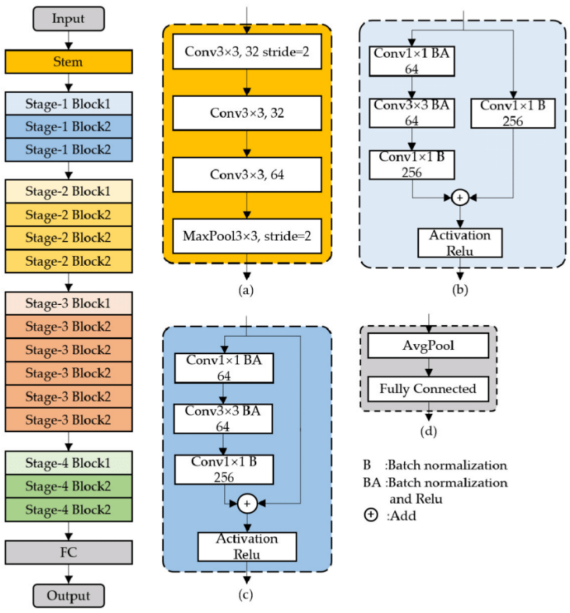


Figure 2.5: ResNet-50 architecture [3]

2.3 Attention Mechanism

Deep learning approaches known as attention mechanisms allow models to concentrate on particular areas of input data. The outstanding performance of Mask R-CNN in segmentation tasks has led to its widespread adoption in computer vision applications. The use of Mask R-CNN has substantially benefited tasks requiring precise object knowledge, such as separating overlapping instances, segmenting fine-grained objects, and tracking objects across frames. The ability to create high-quality masks has enabled advancements in a number of industries, such as robotics, autonomous driving, and medical imaging. On its foundation, models were developed that examined adjustments to boost productivity. It offers precise localization. categorising and dividing up several item instances in a picture.

A lot of research has been done on attention processes in fine-grained picture recognition. With the aid of these techniques, models may concentrate on distinguishable areas or facets of an item that are essential for precise categorization. By focusing on the image's informative areas, they aid in the acquisition of small details and enhance performance.

In several applications, attention mechanisms have shown substantial advancements. By revealing which elements of the input data are critical for the model's decision-making process, attention mechanisms aid interpretability. They make it possible for models to focus on specific words or phrases in natural language processing, which helps with language comprehension and translation jobs. To improve model performance and interpretability across many domains and activities, researchers continue to investigate and develop novel variants and uses of attention processes. Attention mechanisms aid models in computer vision by helping them to concentrate on important visual areas and objects, improving image categorization, object recognition, and image captioning.

2.3.1 Transformer [4]

Transformer Neural Network Architecture solved the problem of sequential data. being computed serially. Transformer can process data parallelly, i.e., for a paragraph of words, each word would not be processed serially, but the entire paragraph would act as a single input, and the output would be a fixed-length vector called word embeddings. Transformer can compute the embeddings simultaneously.

The initial notion of "transformer, which was afterwards referred to as text transformer, is expanded upon by the concept of "vision transformer" (ViT). It is only the implementation of Transformer in the domain, with a small tweak to accommodate the various data modalities. A ViT particularly employs several tokenization and embedding techniques. The general architecture is the same, though.

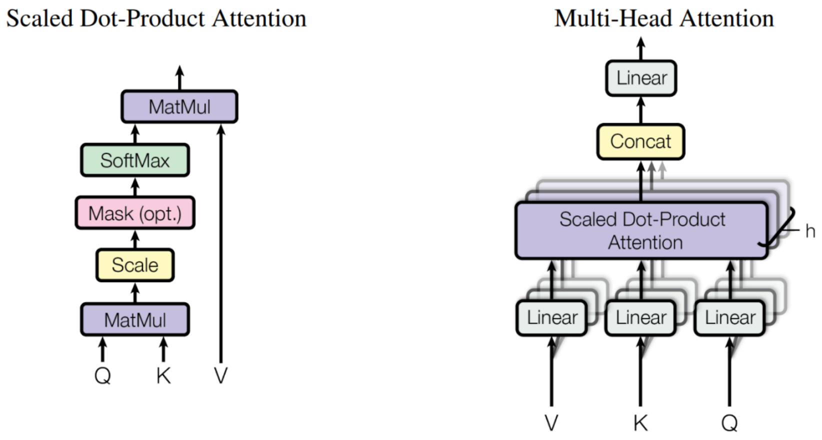


Figure 2.6: Scaled Dot-Product Attention and Multi-Head Attention [4]

Transformer has an encoder part and a decoder part and uses an attention mechanism. At input, positional encoding and input embeddings create word embeddings. The output is an embedding that maps every word to a point in space where similar words are closer to each other. For every word, there is an attention vector, which represents every word's contextual relationship with another word in the same sentence. The embeddings are passed on to a fully connected feed-forward network.

The Transformer does not rely on sequential processing or convolutional processes, in

contrast to earlier sequence models like recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Instead, it makes use of a self-attention mechanism that permits the processing of all input locations in parallel, enabling it to effectively capture dependencies between any two places in a sequence.
Scaled dot-product attention, commonly known as the self-attention notion, is the fundamental principle underlying the Transformer design. Each point in the input sequence may attend to all other locations and give them weights depending on their significance, thanks to self-attention. This enables the model to flexibly concentrate during computation on various elements of the input sequence.

Self-attention layers are present in the decoder as well, but encoder-decoder attention represents a new attention mechanism. The encoder and the decoder are the two fundamental parts of the transformer design. While the decoder creates the output sequence, the encoder analyses the input sequence. A feed-forward neural network is followed by a self-attention mechanism in each of the encoder's levels, which are made up of several layers. Each layer's self-attention technique enables the encoder to concurrently collect dependencies between all points in the input sequence. By using this attention technique, the decoder may focus on the necessary portions of the input sequence as it creates the output sequence. The self-attention outputs are subjected to non-linear changes by the feed-forward network, increasing the model's capacity for representation.

2.3.2 Vision Transformer [5]

ViT, on the other hand, is a transformer that is targeted at vision processing tasks. Vision first transforms and splits an image into fixed-size pixels and flattens them. From these flattened patches, lower-dimensional linear embeddings are then created. Positional embeddings are included and fed to a transformer encoder. The ViT model is pre-trained with image labels and then fully supervised on a big dataset. Fine-tuning is then done on the dataset for image.

This is especially useful for NLP tasks like machine translation or document interpretation, which depend on extracting contextual information from far-off words or phrases. In many NLP tasks, the Transformer design has demonstrated impressive performance, outperforming earlier models and producing new state-of-the-art outcomes.

In order to prevent the model from attending to future places in the input sequence during training, the Transformer architecture includes a method known as ”masked self-attention” in the decoder. This approach makes sure that the model only considers past

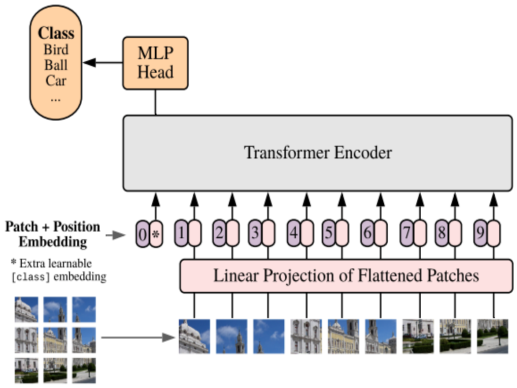

Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure 2.7: Vision Transformer Architecture [5]

and present data for predictions.In order to further illustrate its adaptability and potency, it has also been used for tasks other than NLP, such as picture creation and audio processing. The Transformer design has the capacity to more successfully model long-range relationships than conventional sequential models, which is a noteworthy advantage.

A collection of fixed-size patches from an input image are linearly projected and flattened into a list of tokens via ViT. These tokens serve as the input for the Transformer model, in which each token has the ability to attend to all other tokens and gain context awareness owing to mechanisms for self-attention.

2.4 Survey

Below are some of the implementational details of a survey.

Table 2.1: Survey Report

Method	Categorization	Methodology
ResNet-50	Localization and Classification - Attention Mechanism	CNN used for extracting features. Attention mechanism used for generating attention maps. Attention based sampling performed for extracting details. [25]

ResNet-50	Localization and Classification - Segmentation and Sampling	Convolutional layers used with learned Sparse Attention to extract dominating features. Selective sparse sampling to assign more importance to dominating regions. [26]
ResNet-50	External Information - Adversarial Learning	Web data and training data is input to different CNN and specific prediction loss is calculated for both of them. Joint Loss is calculated by 2-way classification where CNN determines which dataset the image belongs to Joint optimization by reducing the Joint Loss by backpropagation. [27]
ResNet-50	Localization and Classification - Attention Mechanism	Branch Routing module created using two CNN modules resulting in two output feature maps classifying two discriminative regions. Attention Module uses an attention mechanism which outputs feature maps. The Label Prediction module uses convolution and pooling operations for prediction on feature maps. A Binary tree formed with Branch Routing Modules and Attention Modules as nodes and Label Prediction Module as leaf. [28]

ResNet-50	Localization and Classification - Segmentation and Sampling	Feature map from CNN is passed to a Graph Propagation sub-networks which generates a discriminative score for features.Discriminative Features are selected using the discriminative score through graph-propagation and the resultant features are concatenated. Resultant features are passed through CNN. [29]
ResNet-50	Localization and Classification - Segmentation and Sampling	Region Proposal Network(RPN) used for extracting discriminative regions. Another CNN used as a feature extractor. Feature extractor uses discriminative regions as input followed by a fully connected layer for Region-based feature learning. Another feature extractor uses an input image followed by a fully connected layer for Object-based feature learning. [30]
ResNet-50	Localization and Classification - Attention Mechanism	Context-Aware Attention Pooling (CAP) is applied to the features extracted from the image using CNN. CAP reduces dimensions using Bilinear Pooling and the resultant region based features uses attention mechanism to extract contextual information. A learnable pooling approach is used for classification. [31]

ResNet-101	Localization and Classification - Attention Mechanism	CNN for generating feature space along with Attention Maps used as input for expert networks. Multiple experts in a cascaded manner. Semantic Grouping of discriminating feature space obtained from CNN. Deep Bilinear transformation (DBT) of grouped features. Concatenation of Group Bilinear forms a feature vector. [19]
ResNet-101	Localization and Classification - Segmentation and Sampling	CNN used for feature extraction which is then used to extract a Part Dictionary based on segmentation. An Attention Map is created using the Decision Function which groups pixels into meaningful regions. Attention based classification performed on Attention Map. [32]

ResNet-101	End-to-End Encoding - Loss Function	For a single image, features are extracted using CNN and are used for calculating Channel Interaction weights for further modifying the feature space. For a single image, features are extracted using CNN and are used for calculating Channel Interaction. For two images, the modified feature space of both is used for calculating Channel Interaction weights again. Finally a contrastive loss is calculated between the two images for classification. weights for further modifying the feature space. cite
ResNet-101	End-to-End Encoding - Loss Function	Features from a pair of images extracted using CNN called as Initial Feature are concatenated to form Mutual Vectors. Mutual vectors are thencompared with the initial vectors pairwise. A score-ranking regularization along with softmax based feature priority is used as loss function. [33]

ViT-B-16	Localization and Classification - Attention Mechanism	Image segments are flattened and peak suppression is performed to reduce noise and then the resultant images are input to the Vision Transformer. A classification is performed on the output representation called Knowledge Learning. This outputs knowledge based representation of images. The Vision Transformer image representation and knowledge based representations are fused using Fusion representation prediction mechanism. [5]

Chapter 3

Background

The correct classification and identification of objects or elements with subtle visual differences within a given category is the focus of the computer vision subfield known as "fine-grained image recognition". It involves recognising variations of an object that resemble one another but differ in some respects, such as distinct bird species, flower varieties, or car models.

Due to the large variety of practical applications, fine-grained image identification has attracted a lot of interest. Identification of species, population monitoring, and a grasp of ecological dynamics are crucial for animal conservation. The accurate classification of plant species and the early identification of disease are made possible by fine-grained recognition in botany and agriculture. In order to recognise fine-grained pictures, it is challenging to capture and analyse minute visual differences. These visual variations might be a result of modifications to colour, texture, shape, or specific geographical patterns. Conventional computer vision methods relying on global features or shallow classifiers usually struggle to adequately capture these fine-grained qualities.

Advanced techniques, such as attention mechanisms and deep learning architectures like convolutional neural networks (CNNs), have been developed to address these issues. These models are able to automatically learn hierarchical representations from images, gathering both low-level visual details and high-level semantic data. Using transfer learning to gain refinement from pre-trained models on specific fine-grained datasets has also been effective in boosting classification performance.With continuous research, innovative network designs, attention processes, and data augmentation strategies are being incorporated into fine-grained image recognition. The goal is to create dependable and scalable techniques that can correctly categorise and comprehend the subtle variances within fine-grained categories, enabling breakthroughs in a variety of sectors and enhancing our capacity to comprehend and interact with the visual environment.

3.1 Collecting Dataset

Carefully selected datasets are specialized datasets that concentrate on identifying and differentiating visually comparable categories within a particular field. These datasets are intended to test models to detect slight discrepancies and subtleties among closely related categories, necessitating meticulous differentiation. Here is an overview of a few renowned finely detailed datasets:

3.1.1 CUB200-2011

The given data set comprises 200 kinds of birds, with 30 images for both training and testing per species. It comes with bounding box annotations and attribute labels, thus being apt for fine-grained classification and attribute forecasting tasks.

3.1.2 Oxford-IIIT Pets [6]

The collection of data comprises 37 classes of domesticated animals, encompassing more than 7,000 pictures. It presents difficulties in identifying subtle distinctions between various types of felines and canines.

3.1.3 Stanford Dogs

Over 20,000 pictures of 120 different dog breeds may be found in this collection. It is frequently used to investigate breed recognition methods and perform fine-grained classification jobs.

3.1.4 FGVC-Aircraft [7]

There are around 100 photos per category in the Fine-Grained Visual Categorization Aircraft dataset, which includes photographs of 100 different aircraft types. It focuses on identifying different aircraft models by their minute visual variations.

3.1.5 Stanford Cars

This collection includes more than 16,000 photos of 196 distinct automobile models from a variety of angles. It is frequently used for jobs requiring fine-grained classifying of automobile models.

3.1.6 Food-101

With almost 1,000 photos per category, this dataset comprises of 101 different food categories. It is a difficulty for models to distinguish minute variations among various food products.

3.1.7 iNaturalist [8]

A sizable dataset called iNaturalist contains a wide variety of plant and animal species. It is appropriate for fine-grained categorization and species recognition tasks because it spans over 8,000 categories and contains millions of photos.

A common benchmark for assessing the effectiveness of fine-grained image classification methods is the CUB200-2011 dataset. It is frequently used by researchers to create and contrast algorithms for tasks such as feature extraction, attribute prediction, and bird species detection. In the context of fine-grained picture categorization, it has also been used as a testbed for investigating methodologies including transfer learning, deep learning, and attention processes.

The CUB200-2011 dataset is useful for researching item localization, attribute-based recognition, and other related tasks in addition to fine-grained classification since it includes bounding box annotations and attribute labels. Data preparation is necessary for fine-grained picture recognition in order to improve the data quality and model performance. Here are a few examples of standard data preparation techniques for accurate image recognition:

3.2 Data Pre-processing

3.2.1 Data Augmentation

Limited training data are frequently needed for fine-grained recognition, although data augmentation approaches assist grow the dataset and enhance model generalisation. Random rotations, translations, flips, scaling, and brightness/contrast modifications are typical augmentation techniques. These methods make the model more adaptable to various situations by simulating variances that could appear in real-world events.

A computer vision approach called data augmentation uses modifications of existing pictures to artificially increase the size of a training dataset. By applying a series of predetermined alterations to the original photos, enhanced copies of the images are created that have distinct aesthetic traits but the same semantic content. By adding more variation to the training data, data augmentation is frequently used to increase the generalisation and resilience of machine learning models.

Data augmentation entails subjecting the original photos to a number of modifications. Among these changes are the following. Images can be translated, rotated, scaled, flipped, and sheared using geometric transformations to imitate various views or locations. Colour transformations: changing the hue, saturation, brightness, contrast, or colour channel to produce different lighting effects Adding random noise to pictures, such as dropout or Gaussian noise, can imitate real-world changes and improve model resilience. Occlusion and cutout: encouraging the model to focus on various areas of the picture by partially or completely occluding segments of the image or randomly chopping out square patches.

Data augmentation approaches frequently use randomization to broaden their diversity. When performing transformations, random parameters or probabilities are applied, allowing for a stochastic variance in the enhanced pictures. The model cannot rely on certain patterns or artefacts found in the source photos because of this unpredictability.

On-the-Fly Augmentation: During the training phase, data augmentation can be carried out immediately. Each training image is enhanced in real-time before being input into the model, as opposed to producing and storing augmented images offline. This method conserves disc space and permits almost infinite numbers of enhanced samples.

Data augmentation is carried out using a specific set of transformations, which are specified by augmentation policies.

Application Domains: Semantic segmentation, object identification, picture creation, image classification, and other computer vision tasks all make extensive use of data augmentation. By learning robust and invariant features, lowering overfitting, and enhancing performance on novel or difficult cases, it aids models in generalising more effectively.

Importance of the Validation Set: When using data augmentation, it's important to make sure that the enhanced photos don't add biases or skew the labels that represent the real world. Therefore, it is customary to use the same transformations during training for both the labels and the accompanying pictures. Additionally, the model's performance on the original, unaltered photos should be assessed using a different validation set.

3.2.2 Image Resizing

It is crucial to resize photographs to a constant resolution so that they all have the same dimensions and can be processed quickly. Resizing aids in lowering memory needs and computational complexity. Images are frequently cropped and scaled to a specific size, or they are resized while keeping the original aspect ratio.

Changing an image's size or resolution is a fundamental computer vision technique known as image resizing. It is a typical preprocessing procedure used to make sure that pictures have uniform dimensions, satisfy particular specifications, or promote effective processing by machine learning models. In order to resize a picture, its width and height must be changed while the aspect ratio is preserved.

The aspect ratio of a picture must be preserved when it is resized in order to avoid distortion. The proportion between the image's width and height is known as the aspect ratio. The content and aesthetic elements of a picture are not stretched or compressed in an unnatural way when it is resized while maintaining its aspect ratio.

Sizing a picture by multiplying its dimensions by a scale factor is the process of scaling. One can scale up (increasing size) or scale down (decreasing size). When working with huge datasets or high-resolution pictures, scaling down is frequently employed to minimise computational complexity and memory needs during training and inference.

Interpolation is the process of estimating the pixel values for the new picture dimensions when scaling an image. Typical interpolation strategies include:

The simplest interpolation technique, nearest neighbour, assigns the value of the closest pixel to the new pixel. Although it is quick, the output might be grainy or blocky. Bilinear: A more complex technique that determines pixel values by averaging the weights of the four pixels that are closest to it. Compared to closest neighbour, bilinear interpolation yields results that are more consistent. Bicubic: A higher-order interpolation technique that accounts for more nearby pixels and yields results that are smoother. Although more computationally costly, bicubic interpolation can produce results of higher quality.

3.2.3 Preprocessing Tools and Libraries

Data preparation for fine-grained picture recognition is facilitated by a number of libraries and tools. Image resizing, cropping, and enhancement operations are offered by well-known libraries as OpenCV, PIL (Python Imaging Library), and scikit-image. Additionally, built-in features for data normalisation and augmentation are available in deep learning frameworks like TensorFlow and PyTorch.

In computer vision jobs, preprocessing is essential because it improves the quality of input data, extracts useful features, and gets the data ready for additional analysis or machine learning techniques. In computer vision, a variety of preprocessing methods and tools are used to modify and convert pictures. In computer vision, the following preprocessing methods are frequently used:

Image resizing: One frequent preprocessing procedure is resizing images to a standard size. It guarantees consistent picture dimensions, which are frequently required for effective processing and compliance with model structures. Cropping a photograph includes eliminating distracting elements to concentrate on the area of interest. Cropping can help the model focus on key characteristics by removing undesired or unnecessary background elements. Image normalisation: Normalisation is the process of bringing an image's pixel values into a uniform range. It entails adjusting the mean and standard deviation of pixel values to zero and one, respectively, or scaling the pixel intensities to a given range (for example, [0, 1]).

Image Grayscale Conversion: When colour information is not required for the work at hand, converting pictures to grayscale (single-channel) is helpful. Grayscale pictures maintain crucial structural information while reducing computer complexity and memory needs. Denoising techniques for photos are designed to get rid of or cut down on noise. Environmental variables, sensor limits, compression artefacts, and other factors can all contribute to noise. Denoising techniques assist to enhance the quality of photographs by removing undesirable noise elements. Image enhancement methods are used to increase the aesthetic appeal and clarity of photographs. The overall look, details, and contrast of photos are improved using methods including contrast enhancement, histogram equalisation, and adaptive filtering.

Image Augmentation: To increase the size of the training dataset, other versions of already-existing photos are created. It entails actions like translation, scaling, flipping, rotation, and adding random noise. The diversification of training data is increased, generalisation is improved, and overfitting is decreased by augmentation. Edge detection techniques locate the borders separating items in a picture. Edges exhibit sharp variations in colour or intensity and can reveal crucial structural details. For processes like object identification, segmentation, and feature extraction, edge detection is frequently employed as a preprocessing step. Filtering: Certain elements or components of photographs can be enhanced or suppressed using filtering techniques.

3.2.4 Normalization

To ensure that deep learning models train and converge effectively, the picture data must be normalised. Scaling the pixel values to a common range, such as [0, 1] or [-1, 1], is known as normalisation. Pixel values can be normalised by dividing them by 255, by removing the mean and dividing by the standard deviation, or by adopting a particular normalisation.

When working with photos that have various lighting, colour distributions, or dynamic ranges, normalisation is very helpful. Here are some essential features of normalisation in computer vision .In computer vision, normalisation is a typical preprocessing method that includes converting an image's pixel values to a standardised range or distribution. It seeks to boost model performance, increase convergence of machine learning algorithms, and lessen the influence of fluctuations in picture properties.

Scaling the values of the pixels to a certain range, such as [0, 1] or [-1, 1], is a typical method of normalisation. To do this, multiply the pixel values by either the range of pixel values or the highest feasible value (255 in the case of 8-bit grayscale photos). Scaling to a certain range makes ensuring that all pixel values fall inside a predetermined range, which is frequently advantageous for calculation speed and numerical stability. The mean value of the pixel intensities throughout the whole dataset or each picture separately is subtracted as another normalisation procedure. This basically eliminates the mean intensity by centering the pixel values around zero.

3.2.5 Feature Extraction

Extraction of characteristics from the pictures before feeding them into the model may be required for fine-grained recognition. In this stage, low-level features like colour histograms, texture descriptors, or local binary patterns may be extracted. More sophisticated methods, such as feature pyramids or convolutional neural networks (CNNs), may also be used to capture hierarchical representations.

Standardisation, sometimes referred to as z-score normalisation, entails dividing the pixel intensities' standard deviation by their mean value. The pixel values produced by this normalisation method have a zero mean and unit variance. Standardisation can aid machine learning algorithms that presume data is normally distributed by helping to make the pixel distributions more Gaussian-like.

Normalisation can be carried out channel-wise in colour photographs with several channels (such as RGB photos). This entails doing normalisation on each colour channel separately. For applications where colour information is crucial, channel-wise normalisation is helpful in maintaining the relative connections between several colour channels.

Benefits and Points to Keep in Mind: Normalisation helps models concentrate on crucial visual elements rather than pointless changes by reducing the influence of shifting lighting circumstances, contrast levels, or colour biases. It can increase generalisation, model convergence, and fair comparisons between various datasets or picture inputs. To maintain consistency and prevent biases from being introduced, normalisation should be used consistently throughout the training, validation, and testing sets.

Deep learning normalisation techniques: In deep learning, normalisation is frequently included as a layer inside the neural network design. Deep neural networks' intermediate layers' activations are frequently normalised using methods like batch normalisation (BN) and instance normalisation (IN).

3.2.6 Annotation and Labeling

To capture the minute changes among categories in fine-grained datasets, annotations must be exact and comprehensive. For the data to be accurately labelled with the distinguishing qualities, domain-specific specialists may need to annotate it.

Annotation and labelling, which entail marking or labelling certain objects, areas, or properties within an image or a collection of pictures, are crucial computer vision operations. Annotations offer ground truth data that helps with object identification, semantic segmentation, object tracking, and several other computer vision tasks, as well as training and testing machine learning models. Annotation and labelling in computer vision are broken down as follows:

Marking the bounding boxes around items of interest in a picture is known as object localisation. The model can locate and identify objects inside an image by using bounding box annotation, which provides the coordinates for the object's rectangular area.

Semantic Segmentation: Semantic segmentation is the process of assigning a class or category to each pixel in a picture. It entails dividing the picture into sections that represent several item types or groupings. By providing pixel-level comprehension of the image, this annotation approach enables models to distinguish between and recognise various items or areas.

3.2.7 Data Balancing

In fine-grained datasets, there may be a class imbalance where some categories contain much fewer samples than others. Unbalanced class representation may result in biassed model training and subpar results for classes that are underrepresented. To balance the dataset and guarantee equal representation of all categories, many strategies may be utilised, including oversampling, undersampling, and class-weighted loss functions.

3.2.8 Patch Extraction

To collect local information, images are routinely divided into smaller patches for fine-grained identification. Sliding panes or preconfigured grids can be used to obtain these patches. By considering the extracted patches as separate samples during training, the model may then focus on local properties inside the image.

Data balancing is a crucial computer vision approach that seeks to solve the problem of class imbalance in a dataset. Class imbalance happens when there is a large disparity in the number of instances in various classes, which might result in biassed model training and perhaps subpar results for underrepresented classes. By changing the dataset's class distribution, data balancing techniques assist to solve this issue.

Oversampling: By copying or creating additional samples, oversampling entails increasing the number of examples in the minority class or classes. This can be done at random or with the help of certain algorithms like SMOTE (Synthetic Minority Oversampling Technique), ADASYN (Adaptive Synthetic Sampling), or strategies for data augmentation. By ensuring that the model contains enough samples from the underrepresented classes, oversampling promotes greater generalisation and learning.

Undersampling: Undersampling is the process of arbitrarily eliminating samples to lower the number of occurrences in the majority class or classes. Undersampling is a simple procedure, but it may result in the loss of information that may be useful. While reducing its dominance, care must be made to preserve a representative distribution of the majority class.

Class Weighting: During model training, distinct classes are given varying weights. The model assigns more weight to the instances of the minority class or classes during optimisation by allocating higher weights to them. In algorithms like support vector machines (SVMs) or loss functions like cross-entropy loss, class weighting is frequently utilised. It aids in balancing the effect of several classes on learning.

Stratified sampling: Stratified sampling makes sure that the training, validation, and testing sets all preserve the same level of similarity in the class distribution as the original dataset. This method preserves the proportional proportions of various classes while randomly selecting samples. By using stratified sampling, all classes are represented in each subset and bias is prevented from being introduced when the dataset is divided.

3.3 Feature Selection

To find the most discriminative and informative features that are necessary for precise classification, feature selection is a crucial step in fine-grained image recognition. Several crucial elements and methods for feature selection in fine-grained picture recognition are listed below:

3.3.1 Local Feature Extraction

Local characteristics that identify certain aspects in the image are frequently used for fine-grained identification. Local features can be retrieved using methods like Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), Histogram of Oriented Gradients (HOG), or Convolutional Neural Networks (CNNs). These regional characteristics aid in identifying the minute variations that separate fine-grained groupings.

3.3.2 Fine-Grained Descriptors

Designing specialised descriptors suited to the particular topic or issue may be required for fine-grained recognition. These descriptions are meant to represent the distinctive qualities of the fine-grained categories. For instance, descriptors can concentrate on capturing the form and colour patterns of bird plumage in order to identify different kinds of birds.

3.3.3 Dimensionality Reduction

Techniques for dimensionality reduction try to save the most important data while reducing the dimensionality of the feature space. By using some techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) techniques, the original feature space is converted into a lower-dimensional space while maintaining the highest level of discrimination.

3.3.4 Joint Feature Learning

From basic visual characteristics to complex semantic representations, joint feature learning algorithms strive to concurrently learn many layers of features. This method aids in capturing both local information and a wider context, enhancing the models' capacity to discriminate. End-to-end joint feature learning is made possible by CNNs and Transformers.

3.3.5 Domain-Specific Knowledge

It is possible to improve feature selection in fine-grained recognition by using domain-specific information. Specific bird characteristics, such as beak or wing patterns, may be essential for categorization, for instance, in botany. The accuracy and interpretability of the models can be improved by incorporating this knowledge into the feature selection process.

3.3.6 Sequential Feature Selection

Sequential feature selection strategies iteratively choose subsets of features depending on their ability to discriminate. These techniques assess several feature subsets and select the most insightful characteristics that enhance classification performance. The sequential feature selection methods Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) are two examples.

3.3.7 Evaluation Metrics

For evaluating feature selection algorithms' efficacy in fine-grained recognition, appropriate evaluation criteria are crucial. Metrics like classification accuracy, precision, recall, F1-score, or Area Under the Curve (AUC) are frequently used to assess how well feature selection algorithms perform when properly differentiating between fine-grained categories.

In fine-grained image recognition, effective feature selection reduces dimensionality, gets rid of extraneous data, and concentrates on discriminative characteristics that are essential for precise classification. Feature selection improves the performance and interpretability of fine-grained recognition models by choosing the most useful features.

Chapter 4

METHODOLOGY

We highlight three traits that predominate in bird images, i.e., head, wings, and bottom/claws. These three characteristics are the most prevalent. localised segments that are taken into account in the case of a bird. This is useful for both object recognition and subsequent feature extraction. To extract features from these annotated regions and concentrate on the discriminating information, there are region-based techniques that can be used. such as region-based CNNs or attention mechanisms. Darknet format, which specifies the object's class label and the bounding box's normalised coordinates in relation to the picture size, is what Yolo commonly employs. The YOLOV5 network design and settings were setup. The number of classes (bird species) in the dataset, the anchor boxes for each grid cell, the network topology (e.g., number of convolutional layers, filters, etc.), learning rate, and other hyperparameters are all defined. We used pretrained weights to start the YOLOV5 model (on ImageNet). The network gains the ability to forecast bounds, box coordinates, and class probabilities for each grid cell during training.

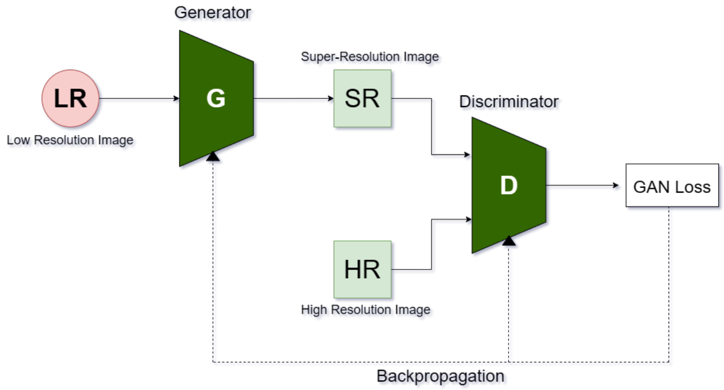


Figure 4.1: Image Super-Resolution Generative Adversarial Network (SRGAN) Architecture [11]

In the form of clipped picture features, the output of the preceding operation is gathered. The dataset contains low-resolution images and the associated high-resolution versions of them. By applying methods like bicubic interpolation, the high-resolution images may be downscaled to create the low-resolution images. A generator network and a discriminator network make up the two primary parts of SRGAN. While the discriminator network seeks to discriminate between genuine high-resolution photos and created high-resolution images, the generator network is in charge of converting low-resolution images into their equivalents in high-resolution.

We train the generator network to produce high-resolution pictures from low-resolution input after initialising it with random weights. To learn the mapping from low-resolution to high-resolution space, the generator network uses methods like deep convolutional neural networks. The ground-truth high-resolution photos and the produced high-resolution images are compared while the generator is being trained using a loss function. To train the networks efficiently, SRGAN uses a variety of loss functions. Usually, they involve adversarial loss, which motivates the generator to create high-resolution pictures that deceive the discriminator, and content loss, which contrasts the created high-resolution image with the original high-resolution image. To retain picture features and textures, reconstruction loss or perceptual loss is also used. The loss function uses the standard GAN discriminator loss to train the discriminator. The generator network is used to upgrade fresh, low-quality photos to higher resolution after it has been trained. To further improve their visual quality, the resulting high-resolution photographs may go through further post-processing operations like denoising or sharpening.

4.1 Dataset

A well-known benchmark dataset for fine-grained image recognition that is primarily targeted at bird species categorization is the CUB-200-2011 (Caltech-UCSD Birds-200-2011) dataset. 200 species of birds that are often found in North America are included in the CUB-200-2011 dataset. 11,788 different bird photographs make up the whole collection, with 30 to 60 images per species. The dataset is split into two primary subsets: a training set and a testing set.

4.1.1 Dataset Characteristics

The images in the CUB-200-2011 collection show birds in a variety of positions, looks, and surroundings. The photos come in a variety of resolutions, sizes, and aspect ratios to mimic the difficulties that come up in real-world circumstances with fine-grained identification. The dataset offers a wide variety of bird species, both common and uncommon.

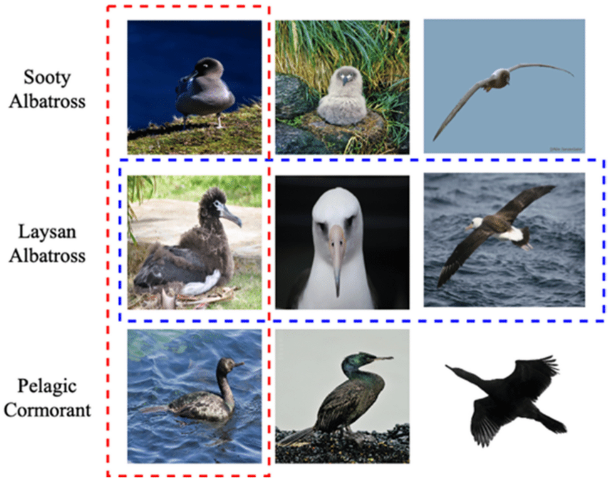


Figure 4.2: View of CUB200-2011 dataset

4.1.2 Annotations

Fine-grained class labels identifying the bird species are tagged on each image in the CUB-200-2011 collection. Additionally, component annotations that identify the positions of various body parts, including the head, torso, beak, wings, and legs, are accessible for certain photos. More in-depth analysis and precise localization tasks are made possible by these annotations.

4.1.3 Training and Testing Splits

There are training and testing sets created from the dataset. Approximately 5,994 photos make up the training set, whereas 5,794 images make up the testing set. Splits prevent the same bird species from showing up in both the training and testing sets, offering a trustworthy assessment framework for fine-grained recognition models.

4.2 Approach

In order to detect the head, wings, and bottom of birds, we prepare the dataset for the object detection job. We utilise a YOLOV5 model that has been pretrained on Imagenet and annotate the data for these 3 characteristics in order to recognise parts. The dataset includes annotations for a bird's head, wings, and bottom segments since these traits are the most important for differentiating one bird from another. We train the model after annotating a small sample of the data (around 40%). We take the input RGB image for resolution 500x500 and then normalize it. For this dataset, we divide it into training and test sets in an 8:2 ratio. The learning rate is kept at 0.01, momentum at 0.937, and weight decay at 0.0005 for 100 epochs. The model begins accurately annotating the bounding boxes of about 97%, and we use the model to make predictions about the remaining photos.

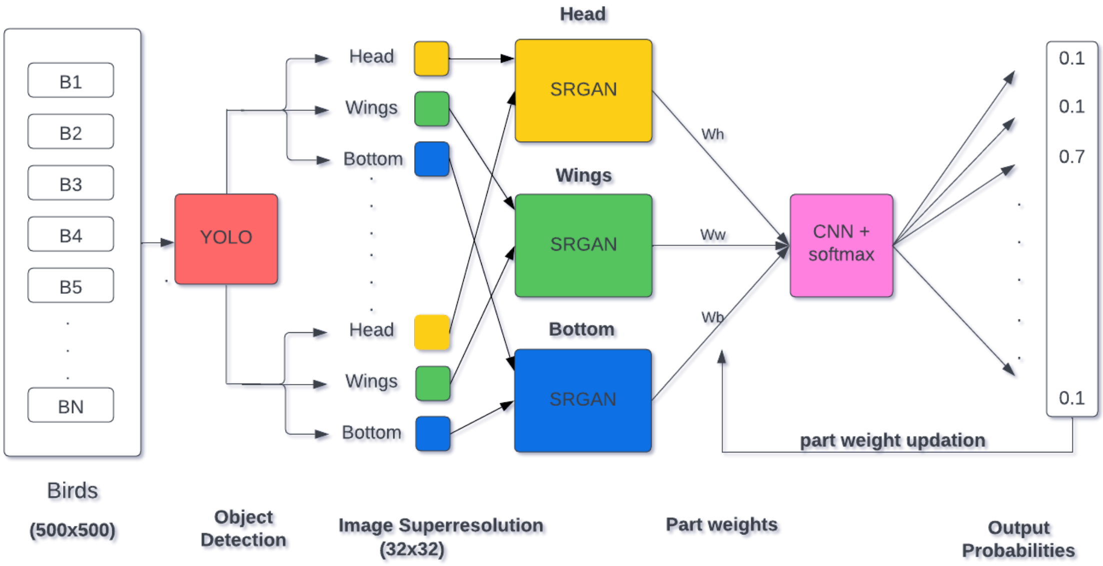


Figure 4.3: Proposed model architecture

The resulting images are cropped in order to acquire the features identified by the prior approach from their bounding boxes. The photos obtained are retained at a maximum resolution of 32 x 32. The model's output is then used in a superresolution process using SRGAN. Using the ready dataset, we train the SRGAN model. The following steps are part of the training process: A generator network in SRGAN receives training on how to upgrade low-resolution pictures to high-resolution ones. It should be trained with a loss function thatfavours perceptual similarity between the produced pictures and the high-resolution ground truth images. The discriminator network separates produced high-resolution pictures from actual high-resolution images. It is taught to distinguish between actual and produced pictures using a binary classification loss. We take the input RGB image of resolution 32 X 32 and normalise the dataset with a learning rate of 0.00008, first-order momentum for Adam decay as 0.5, and second-order momentum as 0.999. We run the model for 250 epochs.

4.2.1 Feature Weightage

We give each of the three feature categories a varied weight because various features or sections may offer varying levels of information that may be used to categorise the image feature element of a class. As the bird's head offers the most data to distinguish it from other birds, it should be evident that the head characteristics should be given the greatest weight. The wings follow next. Given that the interpretation of wings can vary greatly depending on the bird's posture, such as whether it is sitting or flying, less importance is

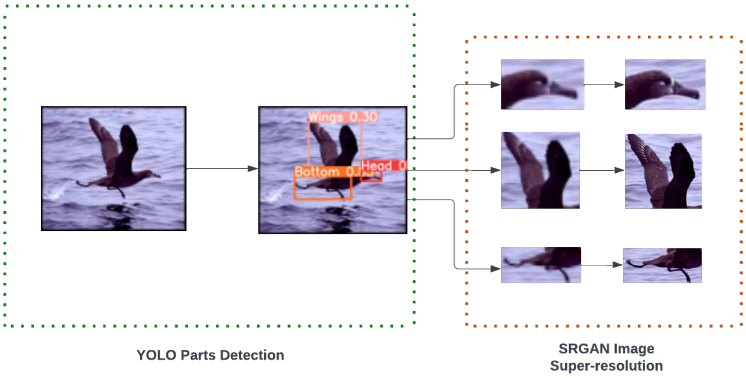


Figure 4.4: Division of task in proposed model

placed on them, resulting in lesser weights. In the previous model, the bottom feature of the image was the least accurately recognised. A bird's bottom is also quite subjective to the bird photograph that was captured. The bird's claws will be folded while viewed in flight, but they will be splayed out when viewed in a sitting position, giving the bird a distinct appearance. The bottom image has been given the least weight in terms of bird recognition.

4.3 Analysis

For different images, there was a high difference in the part images. Techniques like data augmentation, posture normalisation, and multi-view learning can be used to overcome these difficulties. To aid the model's generalisation, data augmentation techniques may be used to intentionally incorporate differences in bird positions during training. The following points were observed as reasons for this issue:

4.3.1 Differences in Appearance

Depending on their positions, birds can display major variances in appearance. It can be challenging to extract consistent visual data from a bird due to variations in the orientation, size, and shape of the bird caused by different stances. For instance, a bird viewed from the front may seem different than when viewed from the side or from above.

4.3.2 Perspective Dependence

It is necessary for fine-grained recognition models to be resistant to different points of view. Bird stances, however, might have elements that depend on the viewpoint. The form and placement of the feathers, for example, may be distinguishing visual traits of the same bird seen from several perspectives and are difficult to generalise to other situations.

4.3.3 Occlusion

Occlusion, or when a bird's body or characteristics are obscured or blocked, can happen in certain positions. Finding important distinguishing characteristics and differentiating between different bird species might be difficult when there is occlusion. For instance, a bird's extended wings may obscure its body or head, making it more difficult to identify.

Chapter 5

RESULTS

In terms of fine-grained detection, it would appear that YOLOv5 is not the most specialised architecture. Other models, such as faster R-CNN or models made especially for fine-grained detection, such as B-CNN or TASN, could do better. However, YOLOv5 may still produce better results on applications requiring fine-grained image recognition, as in the case of the CUB200-2011 dataset, when given the right tweaks and training.

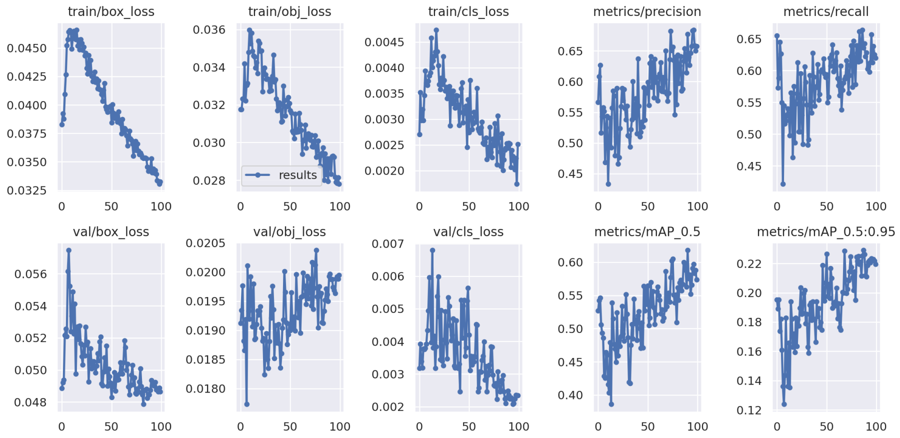


Figure 5.1: YOLOV5 test loss, train loss and other metric result

As for SRGAN, the total number of training images for both the generator network and discriminator network is around 1,00,000 samples for the head and bottom and 80,000 samples for the wings. Some of the future improvements in the method depend on the fact that fine-grained pictures are greatly influenced by the object's posture. In our example, the bird was occasionally sitting or swimming on water, which obscured the claws. Any non-flying image of the bird also showed distinct wing structures from the image of a flying bird.

Modern models' classification accuracy on the CUB-200-2011 dataset was remarkable. Top-1 accuracy ratings vary from about 80% to over 90%, showing that models can correctly identify different bird species based on minute visual variations.

Small visual differences between different bird species may be captured and distinguished using fine-grained image recognition algorithms. They can recognise differentiating traits, including colour patterns, form details, and texture changes, in order to discriminate between superficially similar species. The model architecture, training methods, hyperparameters, and precise experimental settings employed in various research projects may all have an impact on the precise results and accuracy. However, the overall development in fine-grained picture recognition on the CUB-200-2011 dataset demonstrates how deep learning models are successful at identifying minute visual distinctions and correctly categorising bird species.

5.0.1 Output

A fine-grained image recognition model's projected class label for the input picture is its main output. It reflects the acknowledged bird species or precise category that the model thinks the picture falls within. The projected class label frequently takes the form of a verbal description or a numeric code that designates a certain group.

Table 5.1: Accuracy of some models on different datasets

Method	Dataset(%)		
	CUB200-2011	Stanford Dogs [34]	FGVC-Aircraft [7]
ResNet-50 [32]	87.5	94.1	92.6
ResNet-101 [32]	88.1	**94.5**	92.8
ViT-B-16 [35]	90.4	91.4	**93.6**
Our Method	**91.3**	90.6	86.7

The model may offer confidence ratings or probability values for each class label in addition to the projected class label. These ratings show how confident the model is in each category's forecast. Higher scores imply a greater level of assurance in the anticipated label, while lower levels might signify doubt or ambiguity.

Below table can be referred for the accuracy of the approach

5.0.2 Feature Detection Results

While YOLOv5 is a strong framework for generic object detection, it may not be as competitive or as easy to get performance results on fine-grained picture recognition tasks using YOLOv5. To get the best performance in this particular area, it is crucial to take into account specialised architectures and methods that have been designed and tested

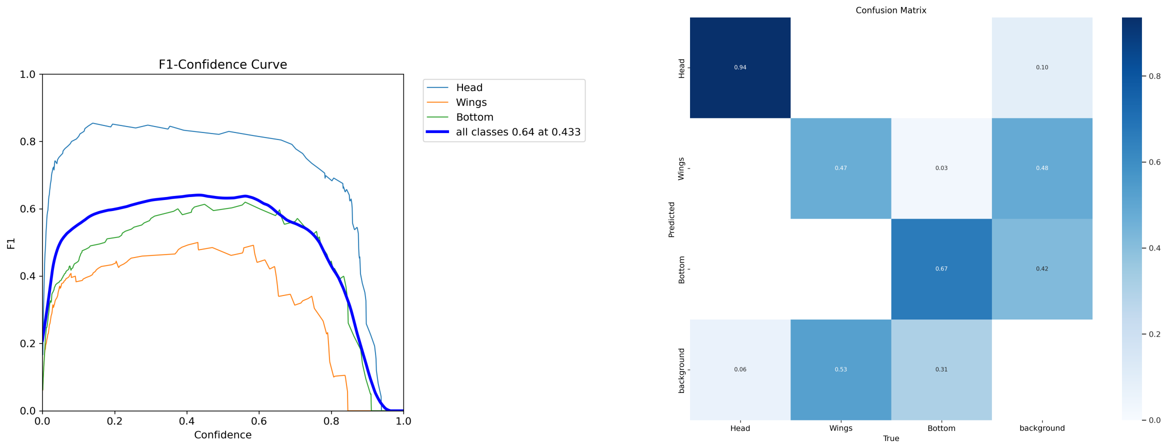


Figure 5.2: YOLOV5 F1-Curve(left) and confusion matrix(right)

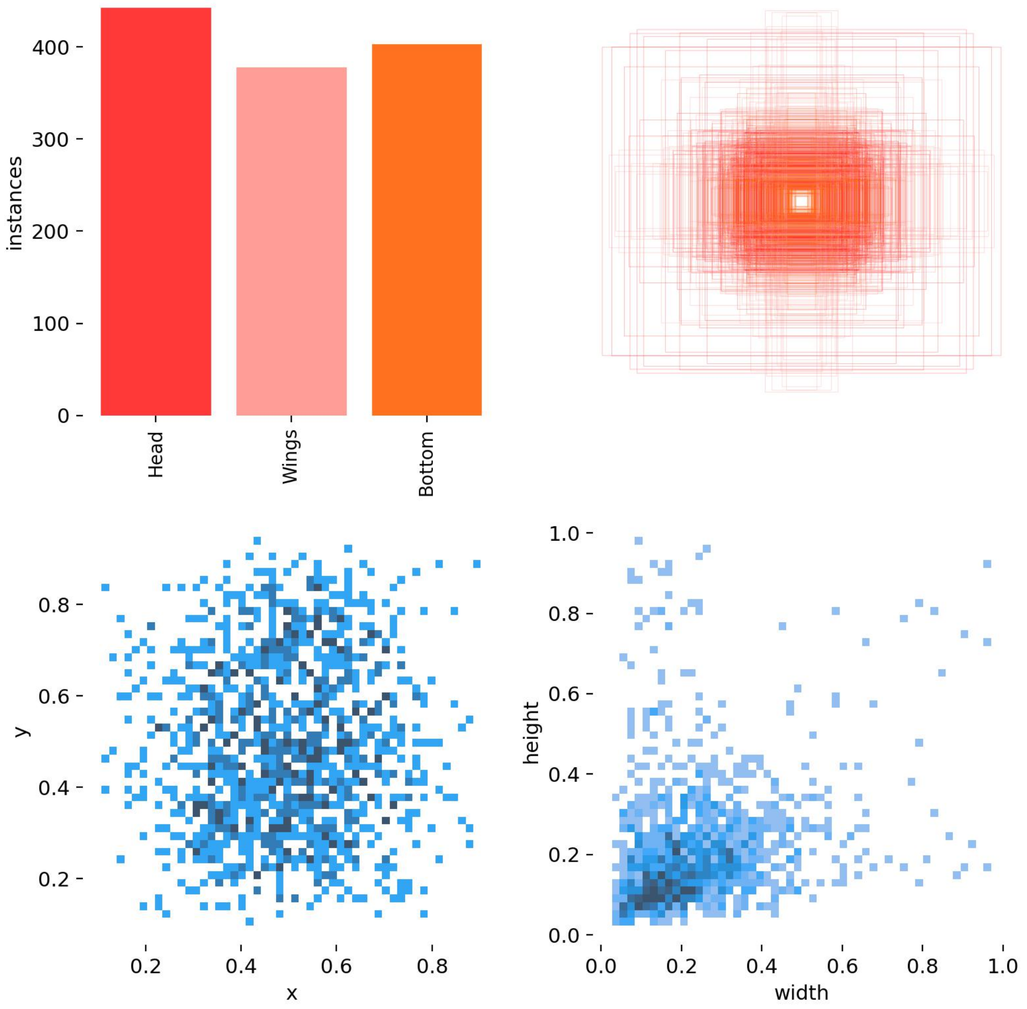


Figure 5.3: YOLOV5 output labels

for fine-grained picture recognition.

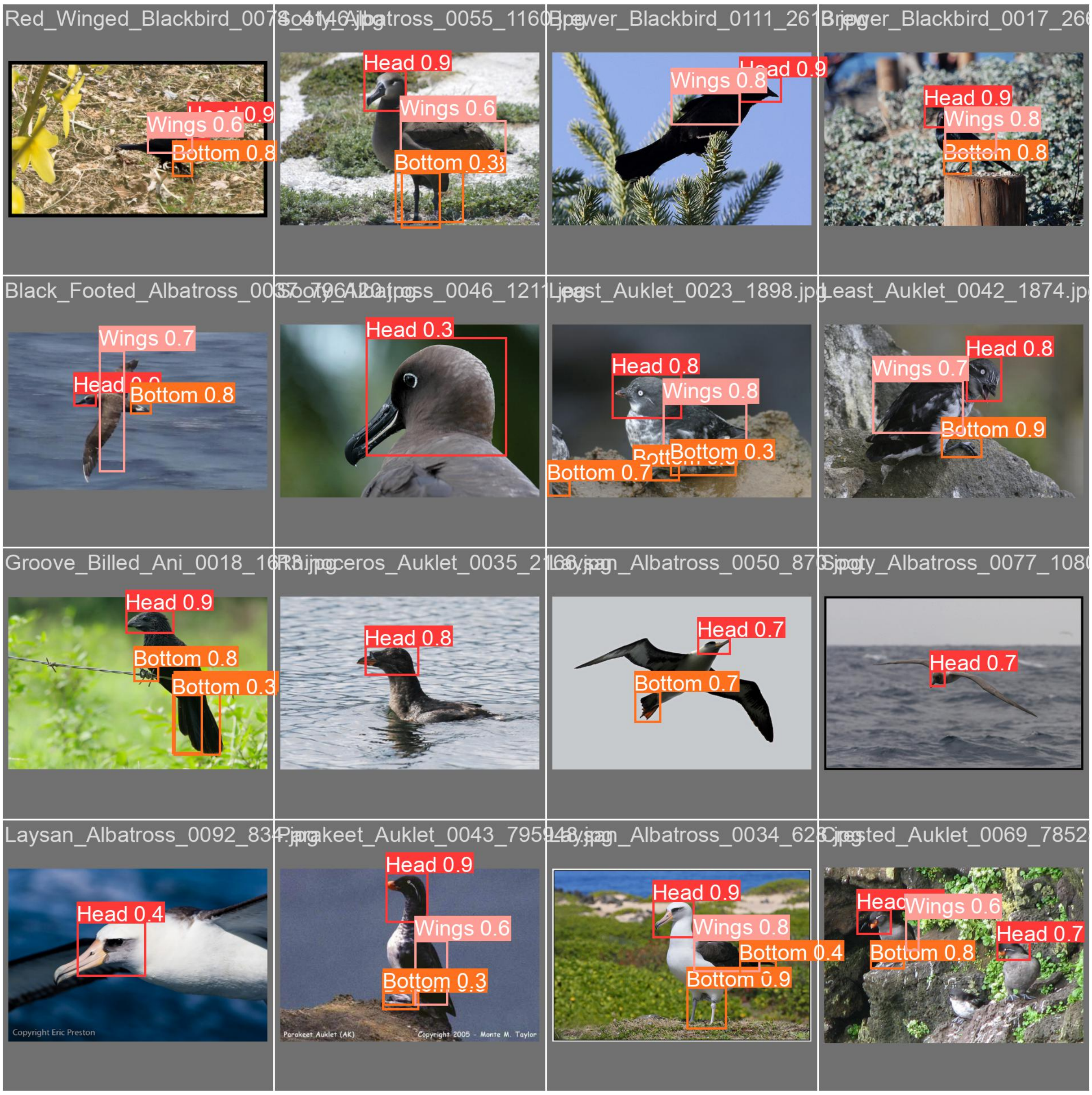

Red_Winged_Blackbird_0078_4146.jpg
Head 0.9
Wings 0.6
Bottom 0.8
Sooty_Albatross_0055_1160.jpg
Head 0.9
Wings 0.6
Bottom 0.3
Brewer_Blackbird_0111_2618.jpg
Head 0.9
Wings 0.8
Brewer_Blackbird_0017_266
Head 0.9
Wings 0.8
Bottom 0.8
Black_Footed_Albatross_0037_796120.jpg
Wings 0.7
Head 0.9
Bottom 0.8
Sooty_Albatross_0046_1211.jpg
Head 0.3
Least_Auklet_0023_1898.jpg
Head 0.8
Wings 0.8
Bottom 0.3
Bottom 0.7
Least_Auklet_0042_1874.jpg
Head 0.8
Wings 0.7
Bottom 0.9
Groove_Billed_Ani_0018_1680.jpg
Head 0.9
Bottom 0.8
Bottom 0.3
Rhinoceros_Auklet_0035_2166.jpg
Head 0.8
Laysan_Albatross_0050_870.jpg
Head 0.7
Bottom 0.7
Sooty_Albatross_0077_1080
Head 0.7
Laysan_Albatross_0092_834.jpg
Head 0.4
Copyright Eric Preston
Parakeet_Auklet_0043_795948.jpg
Head 0.9
Wings 0.6
Bottom 0.3
Parakeet Auklet (AK)
Copyright 2005 - Monte M. Taylor
Laysan_Albatross_0034_628.jpg
Head 0.9
Wings 0.8
Bottom 0.4
Bottom 0.9
Crested_Auklet_0069_7852
Head
Wings 0.6
Head 0.7
Bottom 0.8

Figure 5.4: YOLOV5 output summarized through an image

5.0.3 SRGAN Results

SRGAN can enhance the aesthetic appeal of low-resolution images, but the outcomes of fine-grained image recognition that it directly affects will depend on the particular dataset and job. Metrics like classification accuracy, precision, recall, and F1 score are frequently used to gauge how well fine-grained image recognition models perform. Researchers often use specialised architectures and approaches that concentrate on collecting fine-grained features, modelling small visual changes, and adding domain-specific information to obtain state-of-the-art outcomes in fine-grained image recognition.

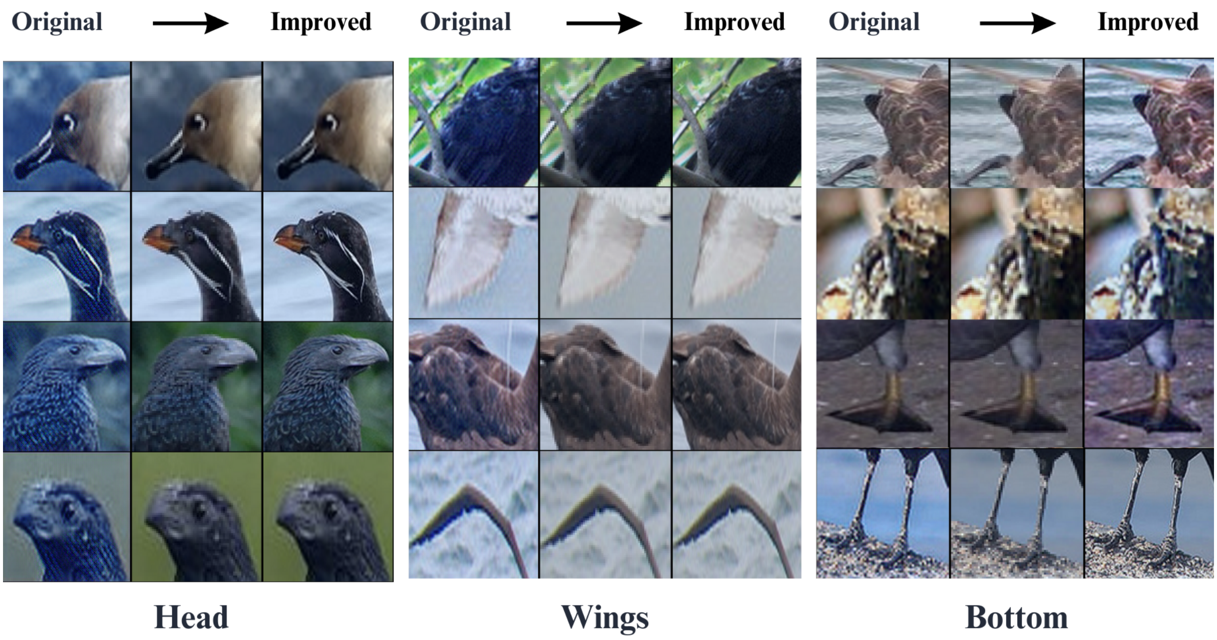


Figure 5.5: YOLOV5 output summarized through an image

Chapter 6

CONCLUSION AND FUTURE SCOPE

On benchmark datasets including the Oxford-IIIT Pet Dataset, Stanford Dogs Dataset, and Caltech-USD Birds-200-2011 (CUB200-2011), notable improvements in fine-grained picture identification have been made. Modern models that make use of deep learning architectures, attention mechanisms, part-based strategies, and self-supervised learning techniques constantly increase the accuracy of these datasets.

The top-1 accuracies of the top-performing models, which range from about 80% to over 90%, show that they can successfully differentiate between visually identical categories. To capture fine-grained subtleties and fluctuations within the data, these models frequently use complicated architectures that make use of hierarchical features, part-based representations, and attention processes.

The development of fine-grained image identification has not only enhanced computer vision but also found uses in a number of other fields. It has been used in fields where accurate distinction between visually similar classes is crucial, including species recognition, medical imaging, product classification, and more.

Fine-grained image identification is still a topic of current study despite major advancements since there are still difficulties to be solved. Limited labelled data, similarities across classes, differences within classes, and generalisation to unknown categories are some of these difficulties. To further enhance the functionality and reliability of fine-grained image recognition algorithms, researchers are continuously investigating novel approaches and datasets.

There have been a number of recent improvements made to fine-grained image detection, which is an important field of study. Recently, attention processes, part-based models, hierarchical representations, and the utilisation of massive datasets have been the main topics of study in fine-grained image identification. In order to address issues like inadequately labelled data and exact annotations, strategies including domain adaptation, metric learning, and weakly supervised Learning and self-supervised learning have been investigated. The performance of fine-grained image detection is being improved by all of these developments.

6.1 Future Scope

Although recent years have seen a substantial breakthrough in fine-grained image identification, there are still many promising areas for future development. The following are some crucial topics for fine-grained image recognition's future that show promise:

6.1.1 Improved Models and Architectures

For fine-grained image recognition, researchers can keep creating more sophisticated and specialised models and architectures. To better capture minute features and small variations between visually identical categories, this involves investigating innovative deep learning architectures, attention processes, and incorporating strategies like graph neural networks or capsule networks.

6.1.2 Data Augmentation and Synthesis

Additionally, artificially fine-grained pictures may be produced to supplement the training set using data synthesis techniques like ge Due to the difficulty of assembling substantial fine-grained datasets, fine-grained image identification frequently suffers from a lack of labelled data. Future studies can concentrate on creating efficient data augmentation methods that maintain minute details while producing more training examples. To add to the training set, artificially fine-grained pictures can be produced using data synthesis techniques like generative adversarial networks (GANs).nerative adversarial networks (GANs).

6.1.3 Transfer Learning and Domain Adaptation

Further research into transfer learning approaches can be used to optimise pre-trained models for fine-grained image identification tasks using large-scale datasets like ImageNet. The effectiveness of fine-grained recognition in certain domains or subcategories can be improved by using domain adaptation methods to transfer knowledge from a source domain with a wealth of data to a target domain with sparse data.

There is an immense amount of applications that FGIR can have. Mostly the applications that are there for object detection are also followed by FGIR.

In this survey we have observed that the transformer based models and techniques are performing way better than other deep learning models. Vision Transformers require an extreme amount of computation but perform better than earlier models. Further we can see a significant amount of work on Vision Transformer related models which can provide a better insight and performance into this domain. The image processing techniques which currently exist are performing better for some datasets, for others there is an open area of research. The domain specific image processing is needed for specialized datasets in order to attain the State of the Art accuracy.

A newer generation of Fine-Grained image dataset is needed for some domains in order to accelerate the advancement in this field. Also with better computation and parallel processing 3D image recognition also provides a wide area of research.

6.2 Applications

There is an immense amount of applications that FGIR can have. Mostly the applications that are there for object detection are also followed by FGIR.

6.2.1 Biological Research

FGIR systems can help biologists in recognizing the different species and breed of animals. Some animals due to their habitat or living conditions are harder to picture in a constructive way that can be useful in determining the species of the animal. FGIR systems can be used in order to clarify the results from a biologist.

6.2.2 Agriculture Disease Detection

Images from leaves and crops that are infected or not can be harder to classify for even humans. The fungal, bacterial and viral diseases need to be detected and clearly classified as such. They are much more active and can infect plants more frequently in warm, humid settings. New pests and illnesses have emerged as a result of climate change and cropping patterns, causing an annual loss in Indian agriculture. In this situation, numerous studies have been conducted using technologies like computer vision. As a result, we're attempting to compile all recent works on machine learning and deep learning. so that it can aid in advancing the work.

6.2.3 Medical Feature Detection

Medical images are one of the most difficult to classify even for a doctor. The minute details of the medical features align better with the scope of the FGIR. The medical images are both black and white and colored, and sometimes 3D too. Current scenario of FGIR can help in detecting or classifying the black and white and colored images but there needs to be much more research that needs to be done in order to extend the scope of FGIR to 3D imagery.

6.3 Conclusion

We have provided an extensive study on fine-grained image recognition in the form of a survey in which we classify the overall development in this field in recent time into different groups and subgroups. The development in the area of Natural language processing can be used to solve vision tasks. Keeping in In light of the recent developments in deep learning, there is an increasing requirement in the computation power
Future trends in the discipline include more comprehensive and dynamic methods for FGIR. For improved detection rates, these models should uncover and extract new characteristics and evaluate massive volumes of data. In addition, semisupervised learning approaches should aid in coping with the rapid expansion of fine-grained generators and the propagation of their material via the internet. Future research will need a more dynamic method or even a mix of supervised and unsupervised learning to immediately discover and actively follow trends associated with contemporary and complicated algorithms for the development of false content without requiring enormous quantities of data. Research in the future could focus on incorporating audio content into forgery detection and pipeline optimisation.

Bibliography

[1] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013. [Online]. Available: http://arxiv.org/abs/1311.2524

[2] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

[7] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," Tech. Rep., 2013.

[8] G. V. Horn, O. M. Aodha, Y. Song, A. Shepard, H. Adam, P. Perona, and S. J. Belongie, "The inaturalist challenge 2017 dataset," *CoRR*, vol. abs/1707.06642, 2017. [Online]. Available: http://arxiv.org/abs/1707.06642

[9] X. Wei, Y. Song, O. M. Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. J. Belongie, "Fine-grained image analysis with deep learning: A survey," *CoRR*, vol. abs/2111.06119, 2021. [Online]. Available: https://arxiv.org/abs/2111.06119

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016. [Online]. Available: http://arxiv.org/abs/1609.04802

[12] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[15] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios," *CoRR*, vol. abs/2108.11539, 2021. [Online]. Available: https://arxiv.org/abs/2108.11539

[16] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," *CoRR*, vol. abs/1911.09070, 2019. [Online]. Available: http://arxiv.org/abs/1911.09070

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[18] X. Ke, Y. Cai, B. Chen, H. Liu, and W. Guo, "Granularity-aware distillation and structure modeling region proposal network for fine-grained image classification," *Pattern Recognition*, p. 109305, 2023.

[19] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, and T. Mei, "Learning rich part hierarchies with progressive attention networks for fine-grained image recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 476–488, 2019.

[20] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 130–13 137.

[21] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *CoRR*, vol. abs/2005.12872, 2020. [Online]. Available: https://arxiv.org/abs/2005.12872

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[24] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[25] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5012–5021.

[26] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8662–8672.

[27] C. Zhang, Y. Yao, H. Liu, G.-S. Xie, X. Shu, T. Zhou, Z. Zhang, F. Shen, and Z. Tang, "Web-supervised network with softly update-drop training for fine-grained visual classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 781–12 788.

[28] R. Ji, L. Wen, L. Zhang, D. Du, Y. Wu, C. Zhao, X. Liu, and F. Huang, "Attention convolutional binary neural tree for fine-grained visual categorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 468–10 477.

[29] Z. Wang, S. Wang, H. Li, Z. Dou, and J. Li, "Graph-propagation based correlation learning for weakly supervised fine-grained image classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 289–12 296.

[30] C. Liu, H. Xie, Z.-J. Zha, L. Ma, L. Yu, and Y. Zhang, "Filtration and distillation: Enhancing region attention for fine-grained visual categorization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 555–11 562.

[31] A. Behera, Z. Wharton, P. R. Hewage, and A. Bera, "Context-aware attentional pooling (cap) for fine-grained visual classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 929–937.

[32] Y. Gao, X. Han, X. Wang, W. Huang, and M. Scott, "Channel interaction networks for fine-grained image categorization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 10 818–10 825.

[33] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y.-Z. Song, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 29, pp. 4683–4695, 2020.

[34] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[35] K. Liu, K. Chen, and K. Jia, "Convolutional fine-grained classification with self-supervised target relation regularization," *IEEE Transactions on Image Processing*, vol. 31, pp. 5570–5584, 2022.

PAPER NAME

Utkarsh_2K21_AFI_29_Major_Thesis_PLAG (2).pdf

WORD COUNT	CHARACTER COUNT
17039 Words	**99751 Characters**
PAGE COUNT	FILE SIZE
67 Pages	**9.7MB**
SUBMISSION DATE	REPORT DATE
May 30, 2023 8:46 AM GMT+5:30	**May 30, 2023 8:47 AM GMT+5:30**

● 13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 8% Internet database
- Crossref database
- 10% Submitted Works database
- 5% Publications database
- Crossref Posted Content database

● Excluded from Similarity Report

- Bibliographic material
- Cited material
- Quoted material
- Small Matches (Less then 8 words)