A Universal Photography Suggestion System Utilizing Composition Detection, Orientation Detection, and Subject Position Detection

by

Iftikhar Shams Niloy
24241296
Syeda Mahjabin Proma
24241295

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
BRAC University
June 2025

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Iftikhar Shams Niloy	Syeda Mahjabin Proma
24241296	24241295

Approval

The thesis titled "A Universal Photography Suggestion System Utilizing Composition Detection, Orientation Detection, and Subject Position Detection." submitted by

1. Iftikhar Shams Niloy(24241296)
2. Syeda Mahjabin Proma(24241295)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025

Examining Committee:

Supervisor:
(Member)

Mr. Dibyo Fabian Dofadar

Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Mr. Md. Sabbir Ahmed

Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Photography is one of the most popular hobby and images are one of the most important content types on social media, and the impact of a photo often hinges on its composition as much as its subject. In response to this, we proposed a system that classifies the compositional structure, detects orientation and subject of a given photo and suggests improvements based on established photography rules. For the classification of the composition, the photo will be categorized into one of five classes(CC, ROT, LL, FIF, PAT). Then, it will determine the orientation of an image. Lastly, this system uses YOLOv8 object detection model to find the objects of a photograph and through logics and conditions the subject is determined. The proposed system will provide the final suggestion based on the three results of the three proposed models. The main goal of the research is to develop a suggestion system that utilizes the detection models built using Deep Learning(DL) algorithms and find the optimal models that will accurately determine the composition, orientation and subject (if any) of a photograph. We have achieved up to 74.34% accuracy in our composition detection model and a minimum of 0.5870 mean square error (MSE) on our orientation detection model. The subject detection conditions capable of properly detecting the subject of an image most of the cases. Our approach aims to assist users in improving their photography skills and elevating the quality of visual content on any media platforms.

Keywords: Photography, Composition, Orientation, Subject, Deep Learning, CNN.

Acknowledgement

Firstly, all praise goes to the Great Allah for whom our thesis has been completed without any major interruption.
Secondly, to our Supervisor Mr. Dibyo Fabian Dofadar sir and Co-supervisor Mr. Md. Sabbir Ahmed sir for their kind support and advice in our work. They helped us whenever we needed help.
And finally, to our parents without their throughout support, it may not be possible.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In the modern world, visual information and graphics are crucial for drawing the attention of viewers and others on any platform. A strong visual composition is crucial for a photograph to convey the most information about the subject as well as the object. Good photography (appropriate use of light, camera viewpoint and composition) can serve to direct the attention of viewer and elicit emotional reactions, claims Michael Freeman in [3]. Ordinary individuals are not adept at keeping the composition of their photos correct. For capturing a photo and maintaining a good composition, two major factors: zoom, angle of the camera and framing are needed to be considered [1]. It might be challenging for inexperienced photographers to maintain these three factors while capturing a picture. However, the task can be completed swiftly and almost accurately by utilizing the artificial intelligence (AI) which will suggest whether to rotate or apply an alternative framing by moving the camera. Artificial intelligence(AI) is implemented in a variety of software applications and other platforms through leveraging machine learning methods. Moreover, it is widely used to make autonomous devices and learning tools. It was not achieved to give photographers real-time image rating and feedback using multiple deep learning (DL) algorithms [29]. Most of the models built using ML and DL focuses on detecting objects rather than scenarios. On the other hand, there are various kinds of AI models that can identify the objects in scenery photos. However, both the subjects and the objects in the photo need to be considered for the picture to be good and convey the most information. The initial step to create an AI model with ML or DL algorithms is to train the model on a dataset, in every detection-based system, either using images or videos. For model performance evaluation, it needs to be tested using proper evaluation metrics. This proposed model will feed the model with annotated images that will prioritize building a system with rotation sugges-

tion and camera movement suggestion and the composition of the image that has been followed. The photographs by professional photographers may help to train the system to make the right decision about composition and the other features of a good photo. The technology can assist both inexperienced photographers and the general public in taking quality photos by using the composition map viewpoint selection method.

1.2 Problem Statement

Photography is believed to be a form of art. Like other forms of art, everyone has a different idea of what makes a photo good. Therefore, developing a model that decides whether the photo is excellent or not means that not everyone will find the system accurate. This issue can be resolved by using back-propagation, in which user data will be fed into the system to personalize the model to the preferences of the users. However, since now the model is trained using images of someone, privacy issues arise. Secondly, datasets related to the model will be hard to make and authenticate as even the perspectives of experts may differ from each other. Moreover, finding or creating an annotated dataset that focuses on the composition of a photo rather than the subjects and objects of the photos will be a difficult task to complete. A created dataset can not be validated without the authorization of experts as photographic opinion differs from person to person. Additionally, there are many kinds of photographs. Landscape scenery-based photographs do not usually contain a subject and they prioritize the whole scenery as the main focus. On the other hand, portrait photography mainly focuses on the subject and there is a background as the object elements. To properly suggest the composition, the model needs to find what kind of photo the targeted data is, then it has to do the following. The following method includes giving suggestions on zooming, rotating and framing of the photo. This includes another step of artificial intelligence training in the process. For the completion of the proposed model, three kinds and steps of artificial intelligence are required which makes the model complex and requires a lot of computational power to complete.

1.3 Research Importance

Visual materials have a crucial role in practically every aspect of our lives in modern society. It is essential to keep a nice composition while taking pictures for business or to advertise a beautiful location. Moreover, a well composed photo is important for effective marketing and a more accurate portrayal of a scenic location. On the other hand, a well composed portrait photo conveys the story of the subject in a

finer way. Automating a system to take beautiful pictures could be the next stage in expanding the automation development of the world. This paper explores and identifies the composition problem of photos and how it can be solved using artificial intelligence (AI). This proposed model will help novice photographers in building their skills and knowledge about photography. Moreover, it can be used to build artistic photography robots that will be able to take visually attractive photos. In an additional use case, the proposed model of this paper can be used to detect faulty photos on social media and recommend the users a better modified version of the images. This may help provide better user experience (UX) in social media applications. The model consists of 3 layer, so multi layer perceptron (MLP) can be used to build the model. MLP can learn non-linear and complex patterns of a dataset which will help the model to detect multiple type of factors at a time. Moreover, MLP has the capability to transform input of any dimension to the number of dimension we want. This will be helpful for making the detection model to be used in different platforms. Moreover, using backpropagation, the system can be customized, analyzing the photography style of the users.

1.4 Research Objectives

Photography is related to operating a camera. Therefore, the objective prioritize to build a model that can be implemented in a smartphone application or a digital lens-single reflex (DSLR) camera. By implementing deep learning methods and neural networks, the model aims to accomplish the following goals:

- Detection of the compositional features of an image and classify the composition as one of the five selected composition classes.
- Identify the difference between particular object detection and the detection of the composition of an image.
- To develop models that suggest rotating and framing, compositional features utilizing image classification and object detection algorithms, which will be used to give suggestions for the perfecting composition.
- The detection, classification and suggestion system will be built, so that the model can be implemented to build applications to be used as a learning tool for novice photographers to learn photography.

Chapter 2

Literature Review

For literature review, we carefully choose research papers from different reputed websites such as IEEE, ResearchGate, and ScienceDirect. We skimmed through many papers related to Deep Learning, AI, Image Processing, photographic composition and personalized detection techniques to determine the compositional lines. This literature review part helped us to gain knowledge about photographic composition, dissection lines in a photo, formulas to detect dissection lines in a photo and how to use different detection-based algorithms of Deep Learning to work with them.

The paper [5] mentioned that, the interest in building a system that can advise photographers the right composition, has grown over time. The paper was published in 2008 and according to them, no modern digital cameras had not yet achieved the ability to intelligently advise composition to the photographer. However, there are several generally acknowledged essential photography principles that can be used in practical situations, even though no approach can be considered as the perfect one to take a decent picture. Additionally, this paper suggested that, in order to improve the aesthetics of environmental portraits, identification of the dissection line is essential. Rule of Thirds composition was leveraged to ascertain the ideal placement of a subject in a photograph. After that, they employed the Hough Transform to identify any straight lines that are present in the image. Then, any discovered straight lines are disregarded if it does not intersect with the calculated ROI. Moreover, they identified Whole Detection Line (WDL) and Half Detection Line (HDL) differently, in order to make the advising mechanism more accurate.

In the article [34], the algorithm the authors used is to sharpen the image which helps the system to identify the subjects in the picture. After that, the paper covers a summary of common composition patterns and the images are categorized based on their photographic composition. Moreover, this research article offered an approach to categorize the picture with image composition, based on the law

of photographs. The suggested approach by the authors investigates the location information of subjects in the image to take into consideration the aesthetic appeal of the picture. This categorization claimed to be aligned with the attributes of the human vision system and taste in art. Enhancing the accuracy of image retrieval is beneficial, as this makes it easier for users to locate the correct photo among the retrieval results. The findings of this study demonstrate that the categorization results agree with the subjective assessments of the participants. This research paper partially solved one of the challenges, which is making a system preferred by most people by taking feedback from the participants of the study they conducted.

The authors of [33] claim that CLAHE, Hough Transform, and custom algorithm of detection is useful for detecting compositional features from room images. These compositional lines denote corner of boundaries of a room. In this research, out of 36 photos, 26 could have their composition correctly identified by the suggested method. Which means, eighty percent of the photos were successfully detected. For completing their purpose, firstly, the contrast limited adaptive histogram equalization (CLAHE) technique is used in this paper to change the contrast. Secondly, the Hough transform algorithm is used to find the line features. Lastly, the suggested approach chooses a suitable composition that is nearest to the identified composition to help the photograph to be compositionally more accurate. It is possible to incorporate this detection system into an aesthetic composition suggestor system in which it can identify and recommend the ideal composition in an indoor photography setting.

In the research paper [29], Gaussian Blur, Hough Transform and Edge Detection algorithms are used to build a model that can rate the photo compositions in real-time. The primary goal of this research is to create a photo-taking app that uses a rating system to provide real-time recommendations to help users shoot quality pictures. To achieve their goal, the authors of this paper first converted all of the photos to grayscale before applying Gaussian blur. Gaussian blur was applied to get rid of any possible noise in the image. Secondly, they build a custom method to detect the edges of the photo. Then they applied the Hough transform method, adjusting the parameters to enable the combination of two subdivided lines into a single straight line in order to identify every line that could exist.

According to the article [20], the authors focused in the Deep Learning (DL) field and used algorithms like YOLO, PSPNet and RTMPPE to build their desired detection system. This study used deep learning models and techniques involving image retrieval to develop an intelligent framework for portrait composition. Their

method is more modern than the ones that previously mentioned above. They build the model aiming to assist inexperienced or novice photographers by providing insightful criticism. The system recognizes and extracts the components of a scene that represent a correlated semantic model. The suggested method concentrates on the aesthetic elements and scene semantics in portraits. Additionally, via the steps of composition and breakdown, the image features retrieval system looks for pictures with similar semantics alongside comparable images. After giving the photographer suggestions, the camera attempts to create a posture-shot by matching the final stance with one of the suggested ones that was obtained.

The book [13] is on architectural photography. To build a system that can assist a photographer in any kind of composition, architectural photography and composition should also be taken into consideration. Moreover, the book also mentioned about how to set up and use a camera to take a perfect architectural photo that can represent the structure of the house perfectly. This will help us to build the suggestion system of our model.

In the paper [22], it has discussed about building a composition detection model based on DL and particularly focused on the CNN algorithm. In this study, the authors aimed to build a composition classification system utilizing the CNN that can classify an outdoor image into any of the nine aimed classes. The exclusive subject of the article was outside photography, which presents additional difficulties in classifying compositions due to its natural landscapes and textures. The nine target classes are RoT, center, horizontal, symmetric, diagonal, curved, vertical, triangle, and pattern[22]. Each photograph is annotated with one or more composition classes. The authors successfully created composition element detectors for each composition class and also proposed a sky area detector that can differentiate the sky from the land in a scenic photograph. As this paper focuses on developing a classifier that can classify the composition of an image that does not particularly have any subject, this can help in building out model to be more accurate and flexible in detecting and suggesting the composition of a photograph. Moreover, this paper user CNN to build the model, which is similar to our proposed model.

The paper [9] used a custom model to find the spot for human in a photo. It particularly focused on building a system that will suggest a place in the frame for human. This article initially looked into a set of aesthetic composition rules and perception of visual principles with the aim to build a model that can give an aesthetic score for a given image. The authors of the paper suggested that by using geometric detection, compositional characteristics can be extracted from an image.

Furthermore, in order to find the best closure for the person in the picture, the authors employed an effective hierarchical search in the suggested model. This article focuses on composing a photo such that the subject is in the ideal location. Which is quite the opposite, in contrast to the prior analyzed article. The model is trained by providing high-quality portrait photos to the system, then the built model helps the photographer to position the subject appropriately in the frame, following a certain composition. The attributes of this model can aid in improving the accuracy of our suggested model when it comes to offering suggestions for images that contain a subject.

The study [2] states that an algorithm based on anthropometric relations has been employed that specifically concentrates on face size. The study suggests that the variations in the identification of face proportions might lead to varied outcomes. Moreover, the method for automating photographic composition principles for human-populated settings is explained in this study, which might be helpful to the occasional photographers in different ways of imaging and printing applications. It concentrated on the three major composition principles which are rule of thirds, zoom and subject integrity. The authors prioritized zoom the most then the rule of thirds and lastly, the subject integrity principle for building the automation system.

The research [15] aimed to build a system that assists a photographer in photography to take high-quality souvenir images by optimizing the location of human beings in a given background scene. Moreover, the designed system investigates human-scenery positional correlations. For training their model, the authors collected thousands of carefully prepared portrait photos with the intent to make the model understand human-scenery aesthetic composition guidelines. Using the intelligent model they developed an android application that uses an input of a static scene to create a human-shaped symbol at a suggested location on a static scenery frame. This advising system may help novice photographers correctly position their subject in the scene of the shot. This research is quite relevant to ours but does not include all the features of our proposed model. Our model may develop this existing model by detecting different scenes, and more flexibility in suggesting the optimal composition.

In contrast to traditional composition evaluation techniques, which correspond to renowned criteria and principles like the rule of thirds and diagonal rule, the authors of [25] concentrate on Yohaku of photographs to support the creation of an aesthetically appealing composition in more flexible manner. Yohaku is the Japanese term for the perpetual empty space. The basic goal of the Japanese number puzzle game Yohaku is to fill in the empty cells in columns and rows such that the sum of each

row or column equals the required amount. By assessing the aesthetic worth of situations using this system, autonomous photography was accomplished through the use of the Composition Map perspective selection approach. Ultimately, using the perspective selecting technique that the authors have previously shown in their study, autonomous photography was achieved by taking the compositional features of Yohaku into account. This method of automating the photography system is quite unique then the rest of the discussed research. This could be helpful in building a custom model with customized formula for our proposed detection and suggestion model.

In order to understand the positioning of items and how to establish effective photographic composition, the article [20] concentrated on evaluating expertly captured photograph by professionals. To achieve their goals, they followed deep learning techniques to build their model. The aurthors required an algorithm in order to retrive an image overlay from panorama pictures of 360-degree and so they proposed a dual chanel CNN model. The 360-degree panorama images were used to mimic the room-in and room-out procedures while shooting a picture. According to the authors, the majority of the portion of panorama pictures allows users to replicate the depth of focus and shooting angle of real-world photography. It has also been proposed that, they will enhance the functionalities of the system and develop it to include a built-in camera.

According to research [1], a large portion of earlier research has concentrated on geometric and temporal characteristics such as view angle, shot distance and obstruction minimization to determine collision and obstruction-free camera motion pathways. For this reason, objects that are not the subject, get ignored majority of the time. The authors proposed a technique to assist a virtual camera that studies the picture and suggests small adjustments to enhance compositional elements like balance or the Rule of Thirds. Rule of Thirds has been applied to divide the frame into three parts and to align the single limited matter in the right position. The system can apply the RoT to the entire composition by taking all the visible parts into account. This paper is mainly focused on a virtual camera like a video game camera. However, this paper mentioned one of the probable challenge our proposed model might face. Which is differentiating the background and the subject. After the differentiation is done, then the model can perform the feature to suggest movement to fit a composition.

The article [50] presents a framework for automated photo acquisition that combines robotic photography with advanced human language supervision. Their model

uses large language models (LLMs) and visual language models (VLMs) to provide reference pictures based on scene observations and user questions. Users may then mimic positions from these references for the photographic robot to take the photo. The PhotoBot uses a perspective-n-point (PnP) problem formulation to modify the camera view. It does this by utilizing pre–trained features from a vision transformer to capture similarities in the image semantic and compare them between multiple picture. According to the authors, PhotoBot can create visually appealing images, frequently surpassing user-taken images in terms of composition and compliance with user instructions. Additionally, their system has the ability to generalize different reference materials, including paintings. The study advances the area by tackling the social dimension of photography, which was frequently been disregarded in earlier robotic photography studies that have mostly concentrated on technical control and navigation. The use of semantic keypoint correspondences for camera adjustments and the combination of LLMs and VLMs for reference image retrieval are important developments. The efficiency and accuracy of PhotoBot is confirmed by user tests, which also demonstrate how it may improve automated photography user engagement and image quality.

The study by Bo Zhang, Li Niu and Liqing Zhang [35] presents an innovative approach for assessing the quality of picture composition, a crucial but unexplored component of visual aesthetics. In response to the scarcity of trustworthy datasets in this field, the authors provide CADB, the first dataset specifically designed for composition evaluation, It consists of 9,497 photos evaluated by five qualified raters. They present SAMP-Net, a deep learning network with a Saliency-Augmented Multi-pattern Pooling (SAMP) module that integrates saliency information to improve geometric and spatial comprehension while analyzing visual layout from numerous composition patterns. Moreover, to lessen content bias, the model incorporates a weighted Earth Mover's Distance (EMD) loss and makes use of composition-relevant variables. This study also investigates how the model may be applied to other datasets, such PCCD to confirm its efficacy and efficiency. According to experimental data, SAMP-net works better than current aesthetic evaluation techniques and offers comprehensible advice for enhancing image composition.

By combining Convolutional Neural Networks (CNNs) with Transformers, in the research [42] the authors investigate a method of computational aesthetics and enhances picture aesthetics rating. Transformers are excellent at modeling long-range relationships, while traditional CNN-based models are good at capturing local aesthetic characteristics but have trouble with global properties. The authors suggest a two-stream paradigm in which a Vision Transformer (ViT), assisted by super-

pixel segmentation algorithm, considers overall aesthetics while CNN retrieves local information. For the final evaluation, the retrieved image features and characteristics are merged, showing a better comprehension of both local and global aesthetic factor. According to the experimental test results on AVA dataset, the suggested approach performs competitively in regression and distribution tasks and reaches state-of-the-art classification accuracy of 84.5%. According to the authors, this model is more in line with human perception than earlier methods because it successfully strikes a balance between comprehensive and detailed aesthetic qualities. The study emphasizes the need of combining local and global information and makes recommendations for future work, which includes using graph neural networks and improving Transformer-based models for smaller datasets.

The paper [52] addresses the difficulty of amateur photographers obtaining professional-level composition by presenting a unique approach for real-time image enhancement through view adjustment. The authors of the paper suggest a View Adjustment Prediction Network (VAP-Net) that uses a composition feature extraction module to evaluate photos according to symmetry and the rule of thirds, which are the two most important photography guidelines. As a proactive substitute for post-processing methods like cropping, the model forecasts changes in displacement, scaling and rotating to enhance image composition before to capture. According to this paper, the usage of a parameter generation-prediction network, which is modeled after the Hyper-Net architecture, is a significant breakthrough. It continuously creates parameters for view changes according to the content of the images. In order to train and test the authors additionally present a View Adjustment Dataset (VAD) that was created from the GAICD dataset and includes a variety of adjustment types. The model's performance is demonstrated by the experimental findings, which show correct adjustments with low mean absolute error (MAE), which is 0.0392, mean squared error (MSE) of 0.0037 and an IoU score as high as 0.783. In both qualitative and quantitative tests, the model performs better than current approaches, including cutting-edge picture cropping algorithms. When dealing with photographs that include complicated material, the model may not always make precise adjustments. Moreover, the authors claimed that, to further enhance performance, future research will use deep reinforcement learning and add more data in the dataset.

The study [16] suggests a method for enhancing picture aesthetics by integrating photographic composition guideline into saliency detection. The accuracy of traditional saliency detection techniques are limited because they frequently ignore spatial information. The rule of thirds, diagonal dominance and visual balance are examples of photographic composition concepts that the authors use as priors for

saliency detection in order to solve this. Their approach improves on current detection algorithms by employing an online parameter selection strategy to fine-tune salient feature segmentation. Additionally they provide a post-processing framework that uses realistic blurring to apply depth-of-field (DOF) effects. In order to produce a more visually appealing composition, the technique uses a depth estimation method to selectively brighten salient parts while blurring non-salient areas. The beneficial effect of this strategy is demonstrated by experimental findings on benchmark datasets, such as MSRA, which show better saliency detection and segmentation than current techniques. In addition to highlighting the possibilities of incorporating composition-based priors into saliency detection, the authors of this research paper stated that future research might further improve image aesthetics by refining depth estimation algorithms.

The paper [6] suggests a technique for automatically figuring out if a picture follows the rule of thirds, which is a basic compositional element in photography. The authors of this paper acknowledged that comprehending semantic information is necessary for identifying this rule, which is still difficult in computer vision. They used recently developed features in object identification and saliency analysis to solve this by estimating the position of significant things in a picture. To deduce spatial correlations in picture, they extracted a variety of characteristics, such as raw saliency maps, grid-based thirds maps and saliency map centroids. An accuracy of about 80% in rule-of-thirds identification is then attained by combining these characteristics with several machine learning classifiers, such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN) and Adaboost. According to experimental data by the authors of this paper, objectness identification only slightly improves performance compared to raw saliency maps. The study draws attention to the shortcomings of saliency-based methods by showing the difference between high-level content comprehension and low-level stimuli detection. Lastly, the authors proposed that performance might be greatly improved by advances in generic object identification.

The study [8] describes a saliency-based technique for identifying compliance with the rule of thirds in photographic pictures. In order to manage extensive picture archives and enhance image search, sorting and editing, the authors of this paper highlighted the need of automated aesthetic analysis. In contrast to earlier approaches that depend on computationally costly parameter adjustment, their strategy makes use of discriminant analysis techniques to increase accuray and effectiveness. To find salient areas in photos, they employ two very good saliency detection algorithms: Global Contrast (GC) and Regional Contrast (RC) along with

Context-Aware (CA). Following the reduction of feature dimensionally by Principal Component Analysis (PCA), classifiers like Support Vector Machines (SVM), Linear Discriminant Analysis (LDA) and Quadratic Discrimant Analysis (QDA) are employed to classify between Rule of Thirds and None-Rule of Thirds. According to the test results, the RC saliency approach performs better than CA in terms of accurary and computing efficiency, which makes it more appropriate for real-time applications. QDA performed similarly to SVM and had the greatest accuracy of any classifier, which is 77.71%. Lastly, this study emphasized the problem of incorrectly categorized salient areat and proposes that detection accuracy may be increased by advances in semantic content understanding.

The paper [4] examines the color constancy of objects and their backgrounds in order to identify image similarities. The authors provided an approach for evaluating the realism of an image based on object-based color temperature measurements and regional color data. Support Vector Machine (SVM) classifier has been used in order to seperate pictures into object and background areas and extract color consistency signals that differentiate actual photos from composites. In this study, an improved dataset of composite pictures with both realistic and unrealistic compositions is introduced in the study to assess the performance of the model. Under controlled settings, experimental findings show that the suggested technique reaches 70% accuracy, proving to be accurate in identifying proper and altered photos. The study also demonstrates that the optimal results for assessing color consistency are obtained using the Kullback-Leibler (KL) distance measure and the HSV color space. Furthermore, even for complex manipulated photos, the object-based color temperature measure improves tampering detection. As this paper introduced a method to differentiate object and their background, it will also be convenient for features extraction for composition detection benefiting our research.

The study [47] investigates an innovative way of improving image composition on mobile devices by utilizing edge intelligence and deep learning. According to the authors of this paper, conventional picture composition techniques depend on cloud-based processing, which presents deployment issues and latency. So they suggested an Image Composition Guidance System (ICGS), which allows for real-time composition support on cellphones, as a solution to these problems. To guarante effective deployment on mobile devices with limited resources, thesystem makes use of MobileNet-SSD lightweight object recognition model that has been tuned by weight pruning and post training quantization (PTQ). The ICGS provides manual composition guidance (MCG) with target-tracking algorithm for user-defined composition regions and automatic composition guidance (ACG) with a deep learning model for

real-time subject placement. According to the results of experiments, guided visuals receive better aesthetic evaluations that are consistent with societal aesthetic norms. The system is suitable for daily usage due to its reliable stability. By combining deep learning with real-time user mentoring in image design, the research advances the domains of computational photography and mobile artificial intelligence.

The paper [43] presented a multitask Generative Adversarial Network (MT-GAN) with the goal of enhancing image composition by resolving problems with appearance consistency and geometry. The authors stated that, disparities in geometry, lighting and texture make it difficult for traditional picture composition techniques to harmonize items from various sources. In order to address this a multitask GAN architecture has been suggested, that combines a Spatial Feature Extraction Network (SFEN) for maximizing spatial feature representation and a Geometric Consistency Network (GCN) for aligning object placements. To produce realistic and well-balanced composite pictures, the model employs three learning objectives: compositing loss, adversarial loss and Region of Interest (RoI) consistency loss. In order to compare the performance the study presents a fresh dataset of 7,756 photos with ground-truth annotations. Test results state that, MT-GAN performs better than other approaches, such as ST-GAN and DoveNet and achieves great Mean Square Error (MSE), Peak Signal-to Noise Ratio (PSNR) and realism ratings. The model is a flexible and reliable option for picture composition as it successfully suggests realistic shadows, reflections and appropriate object placements.

The paper [45] stated an image resizing method that combines composition detection with computational aesthetics in order to improve the visual quality of scaled photos. Typical method of picture scaling frequently result in distortion or the loss of significan visual components due to image warping and seam slicing. The authors suggest a composition-aware scaling approach to address these problems while maintaining the visual balance and importance substance of photos. The technique starts with a CNN-based composition detection module that categorizes photos into four basic composition types: symmetric, horizontal, center and rule of thirds. Then the algorithm directs the scaling procedure to preserve the structural integrity and aesthetic appeal of the image based on the detected categorization of the composition. Using quantitative measure of picture quality, the study assess the stratgy and shows that the suggested method ourperforms conventional resizing procedures in terms of aesthetics. According to the experimental findings, the authors stated that the content-aware scaling that incorporates composition principles greatly improves visual harmony and viewer perception.

Chapter 3

Work Plan

The work plan includes all the predicted tasks that need to be completed for the thesis to be successful. Firstly, a topic has been established based on our areas of interest for the thesis. Subsequently, the subject is discussed in groups to create a list of difficulties we might encounter when working on the thesis. Consequently, we examine scholarly articles that are relevant to our subject in an effort to solve the challenges. We initially require a dataset in order to train any machine learning or deep learning model. However, previously built datasets may not have considered compositional lines during annotation. Therefore, the gathered datasets will undergo analysis before being labeled in accordance with the compositional features of a photo. To avoid complexity we have decided to build the initial composition detection model using Image Classification. For image classification, we have classified each image as one of the five composition methods that has been selected for our model. After that, we have researched about machine learning (ML) and deep learning (DL) algorithms to gain a clear understanding of what is and is not possible. Doing the research we have found Convolutional Neural Network as the best suited algorithm for our initial classification model.As our suggested model incorporates several detecting sub-models and the outcomes will be combined to provide the desired result. It has been decided to build the composition detection model then using the prediction of that model we will build a composition suggestion system that will suggest how to improve the composition of that specific image. We are planning to use saliency mapping along with leading lines of each image to train the suggestion system. So both model will be trained using the same images but different annotation and image features. Through a trial and error process, we will find the best suitable algorithm for our detection model along with the optimal pipeline. Then we will combine both models to build our hybrid model that can detect the nearest composition of an image then suggest methods (move left or right, zoom in or zoom out) to improve the composition. Lastly, the result will be written as a research paper and submitted to the authority.

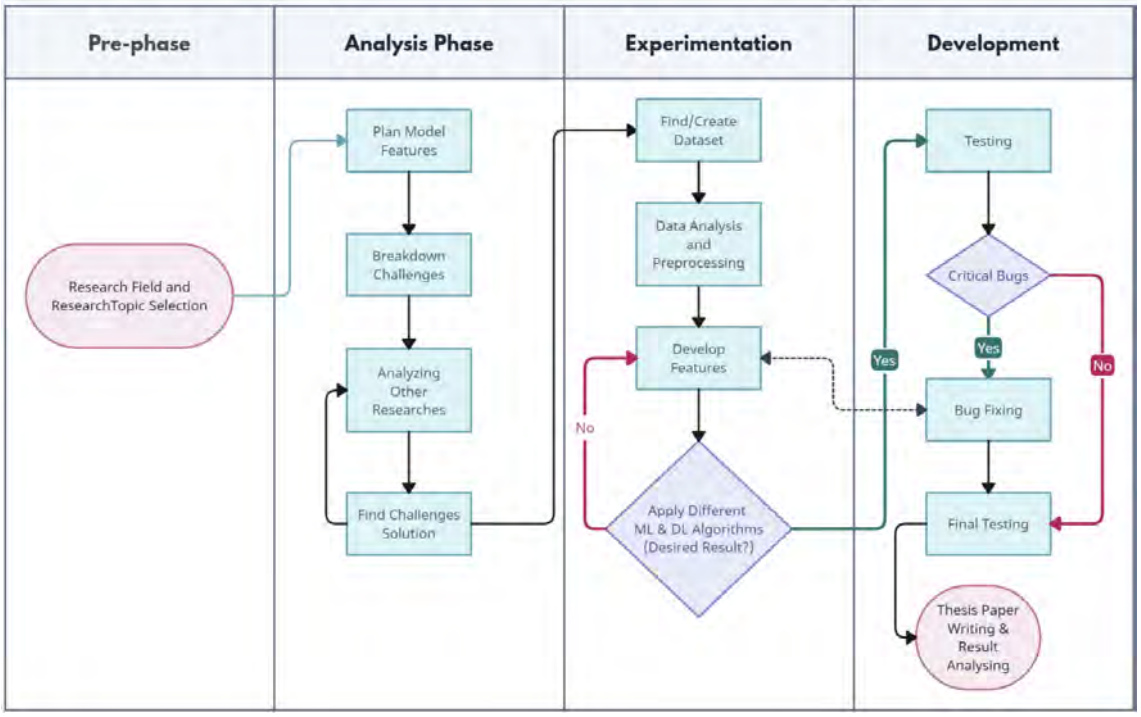

Figure 3.1: Work Plan Flowchart

Chapter 4

Dataset Description and Pre-processing

For building our model, we have used three different dataset. One for subject detection, one for rotation detection and another for photo composition detection.

4.1 CCIP Dataset

Common Composition in Photography (CCIP) is our custom dataset which is used to detect the composition of the photography. This composition detection dataset contains 1515 photos from five distinct photographic composition classes: Leading Lines, Rule of Thirds, Centered Composition and Symmetry, Frame in Frame and Patterns. To guarantee that all selected photos are unrestricted by copyright, each image is gathered from several copyright-free websites such as Freepik [56], Pexels [57], Unsplash [58], Image Composition Assessment DataBase (CADB) [35] and some pictures from Aesthetics and Attributes DataBase (AADB) [17]. Furthermore, we took some of the photos in the dataset by ourselves. The dataset is useful for various picture classification tasks and photographic study since it is concentrated on the main composition criterias of photography which are comparatively easier to distinguish. An annotation file that associates picture names with their corresponding composition classes and photos.

Composition Name	Image
Rule of Third	332
Centered Composition and Symmetry	332
Leading Lines	330
Frame in Frame	300
Patterns	221

Table 4.1: CCIP Dataset

4.1.1 Resizing images

The dataset contains images varying in resolution. For better performance, we have resized all images into 224×224 as this is the standard input size for different CNN architectures.

4.1.2 Splitting the dataset

We have decided to use image annotation on our dataset as we needed to consider the whole image to determine the composition. That is why we annotated each image as one of the 5 target classes and used a .csv file to keep the information recorded.

4.1.3 Normalizing the values

As neural networks work better with a smaller range of input data (0 to 1), we normalized the pixel value of all images by dividing it by 255.

4.1.4 Annotation

The dataset was annotated using a CSV file that maps each image filename to its corresponding composition category. This form of annotation contains metadata for example: class name with the images for supervised training. The dataset itself consists of a single image folder and a CSV file with two columns named as: image_name and composition_name.

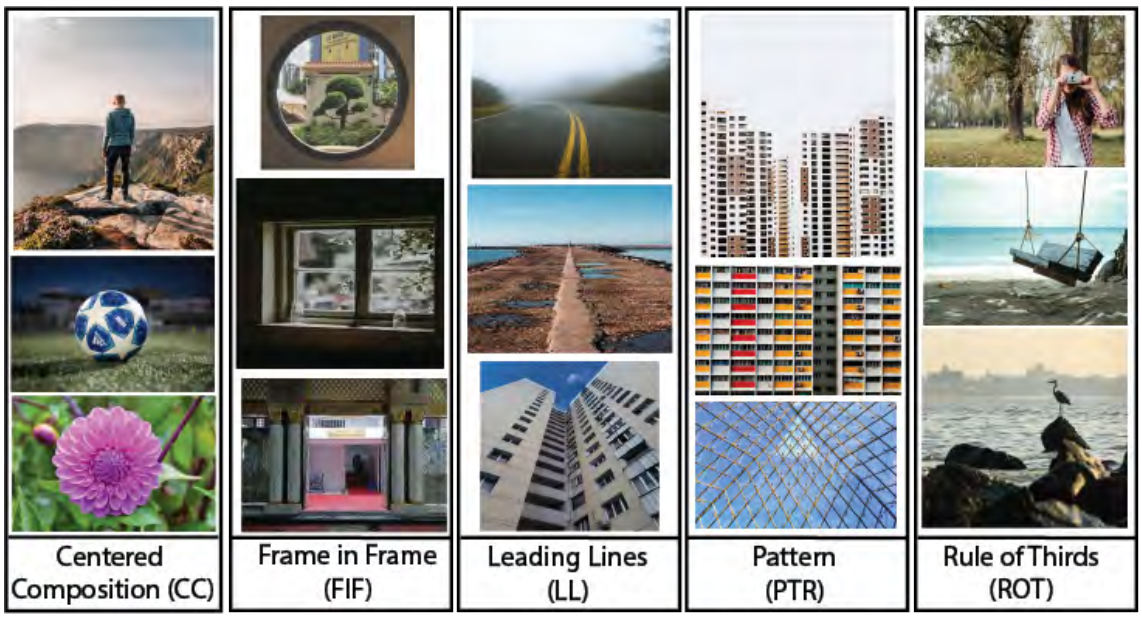

Figure 4.1: CCIP datase

4.2 DRC-D

We have used the Rotation Correction dataset from Deep Rotation Correction without Angle Prior [44] paper to train our model for rotation detection and correction. This dataset consists of images that are randomly rotated between -10° and +10°, making it ideal for learning how to identify misaligned or tilted images. For our orientation correction model, we used the DRC-D to detect whether an image is in correct orientation or not and suggest the camera angle adjustment required to bring it to a 0° upright position. Instead of generating corrected images like the one mentioned in the paper[44], our model was trained to predict the necessary camera angle rotation for example: 4.0° clockwise or 5.0° counter-clockwise to help the photographer correct the orientation of camera properly.

4.2.1 Dataset Modification

In the paper[44], the dataset is used to create an image to image correction model. But in our case, we want to build a regression model which can predict the orientation of an image. To support this, we customized the dataset by including a ground truth (GT) image at 0-degrees for each type of image. Although the model does not use GT images as direct input, they were included to validate angle prediction consistency and thus increasing the size of the dataset. This approach ensured that, while the model outputs only a rotation angle, we could qualitatively assess its accuracy by comparing the predictions against the test dataset.

4.2.2 Image Resizing & Preprocessing

Each image in the dataset has the same resolution of 512×384. As we are trying to build an orientation detection model, which prefers to have more spacial information. For this reason, the images are kept in 512×384 resolution to hold more spatial information. However, as most of neural network model prefer to have a value ranged from 0 to 1, we have normalized the dataset by dividing the pixel values by 255.

4.2.3 Splitting the Dataset

The dataset comes splitted into train and test images.After modification, the dataset consists of 7019 images for training and 1216 images for testing.

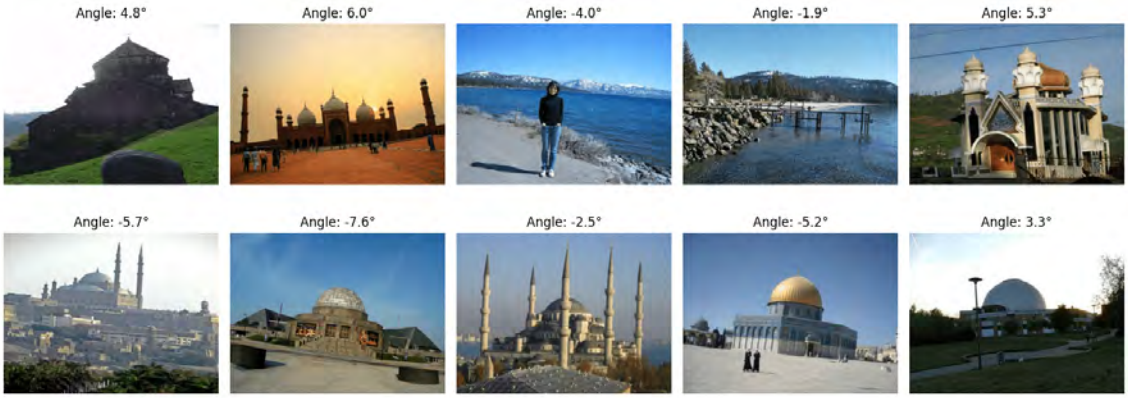

Figure 4.2: DRC-D

4.3 COCO Dataset

The COCO (Common Objects in Context) dataset [10] is a large-scale dataset made by Microsoft, comes with multi-type annotation for object detection. It contains over 330000 images and over 200000 labeled images with more than 80 object categories. In our work, the COCO dataset is utilized to detect and identify various objects within a photograph. Among the detected objects, the one with the largest bounding box area is considered the main Subject of the image.

Figure 4.3: COCO Dataset

Chapter 5

Architecture Analysis

5.1 Convolutional Neural Networks (CNN)

As our dataset is imaged-based data, deep learning models will be more appropriate for us. As we are working with a photo composition detection system, we need to extract features like lines, shapes, texture, and different patterns of an image. Learning with data that has already been labeled are then taken as inputs that serve as targets is known as supervised learning [11]. A collection of input values as vectors and multiple correspondingly defined output values will be included in every training sample [11]. The issue of feature selection is resolved by deep learning approaches, which identify the important features automatically from raw input for a given problem rather than requiring pre-selected features [21]. So we found the Convolutional Neural Network (CNN) model to be more appropriate for our work.

CNN is a deep learning model designed to process images by extracting and analyzing features. It begins by taking an image as input and then applies convolutional layers to detect different patterns like edges and texture of the image. The layers are followed by an activation function to introduce non linearity and that makes the model capable of learning complexity representations. Layers according to the need of feature extraction of the image can be added after the input layer to increase the usefulness of this algorithm suggested in [19]. Different filters can be linked to each layer. Thus, we are able to extract many elements from the provided image using CNN [19].

After that the pooling layers reduce the image's dimensions while retaining essential features, improving computational efficiency. Finally, the extracted features pass through fully connected layers where the network interprets them and produces an output such as classifying an image and detecting objects within it.

RGB images have three colour channels red, green and blue on the other hand, grayscale i,ages have only one colour channel. CNN processes all of these images by analyzing pixel values. In convolutional, a small matrix or filter moves over the image to extract the feature edge or texture. Early layer detects the simple patterns, the deeper layer detects complex objects and the final layer generates an activation map by assigning confidence scores to classify the image based on learned features. It was determined that it is better to search for specific regions inside the image rather than a complete connection [19].

5.1.1 CNNs Layers

CNN is mostly used for object detection of an image by creating a bunting box with class labels and confidence scores. There are two stages to detect the object. The initial half of the network focuses to create feature extractor consisting of many convolution and pooling layered alternately [55]. This allows the system for processing of preprocesses input data and convert it into more general feature representation [55]. The second part consists of fully-connected layers, in which every neuron inside a layer is linked to every alternate neuron from its preceding layer [27] and classify the image into one of the selected classes. The descriptions of different layers in CNN as follows:

Convolution Layer

The convolution layer in CNN is mainly used for fetching the features of an image by sliding a window of filters (also known as kernels) over the image pixels inputted as matrix. This operation is known as convolution. Two of the most important parameters are strides and padding, which controls the step size of filters and amount of adding extra pixels around the corners, respectively.

Pooling Layer

The function of the pooling layers is to decrease the size of features retrived from convolution. One of the pooling method is max pooling, which selects the maximum value from a certain amount of pixels and replaces all the pixel with one pixel with maximum value. Another method is average pooling which works the same but takes the average value of certain amount of pixels to complete the pooling.

Activation Function Layer

Some common activation layers are ReLU, Sigmoid and Tanh. This introduces non linearity into the dataset allowing the network to model complex patterns. Activation functions mimic the function that only neurals that surpass a specific threshold can only be transferred to the subsequent neuron [38]. CNN kernels symbolize numerous receptors that can react to diverse aspects [38].

Batch Normalize Layer

This layer works to normalize the inputs to each neuron, helping with faster training, regularization and reduce overfitting. It also stabilizes learning by reducing internal covariate shift and helps to improve the generalization performance.

Dropout Layer

The dropout layer prevents overfitting by randomly setting the proportion of units of zero during training.

Classification Layer

This is the final layer. This layer has two sub layers. These are fully connected layers (FC) and global average pooling layers (GAP). The FC layer connects all the neurons from the previous layers to the current layer and the GAP layer reduces feature maps by averaging values, preventing overfitting with fewer parameters but limiting complex decision boundaries.

All these layers work together to extract features, prevent overfitting and generate accurate predictions for object detection.

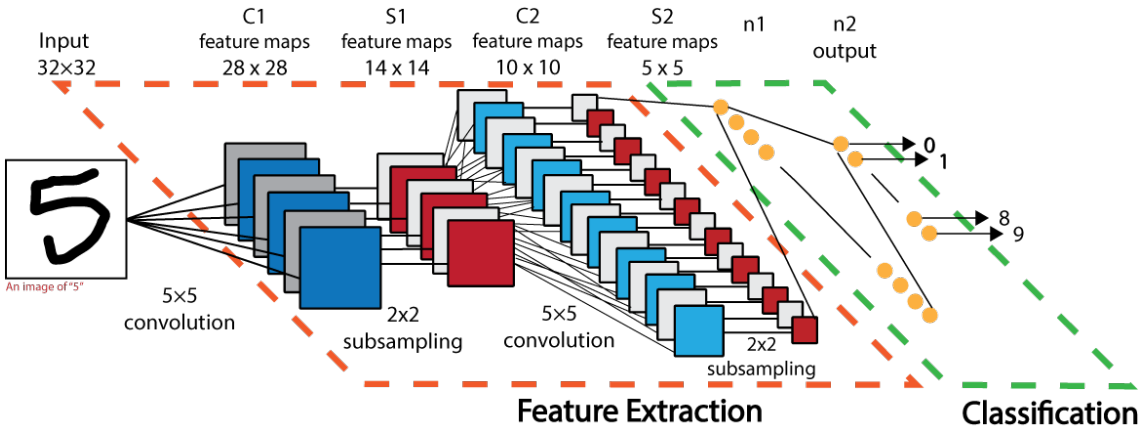

Figure 5.1: Internal Layers of CNNs [24]

5.1.2 Different CNN Architectures

Different CNN architectures are also designed to automatically extract spatial features from images using conventional layers and this make them very reliable for doing image classification, object detection and segmentation etc. Different architectures such as VGG, ResNet, MobileNet, EfficientNet, Inception, LeeNet, Alexnet etc optimize the performance through depth, skip connections and efficient scaling.

LeeNet

LeeNet was first proposed by Yann LeCun in 1998 for handwritten digit recognition. This is the first successful CNN architecture and also the foundation of many modern CNN architecture. It consists of two convolutional layers. These layers contain 6 and 16 filters of size 5×5. Each layer is followed by a 2×2 max pooling layer for feature extraction and two fully connected layers with 120 and 80 neurons for classification and 10 neurons for output layer. Though it's simple, it is a very useful architecture as a starting point of an image classification task.

AlexNet

AlexNet got into spotlight by outperforming previous models in ImageNet Large Scale Visual Recognition Challenge known as ILSVRC in short. The architecture of AlexNet have a total of 5 layers for convolution: 96 filters of 11×11, 256 filters of 5×5 and 384 filters of 3×3. Each convolution layer has a max pooling layer of 2×2 subsequently and three fully connected layers (two with 4096 neurons and an output layer consists of a thousand neurons for ImageNet classification). AlexNet introduced key innovations like ReLU activation, data augmentation, and dropout regularization, making it a breakthrough in deep learning and inspiring many modern architectures.

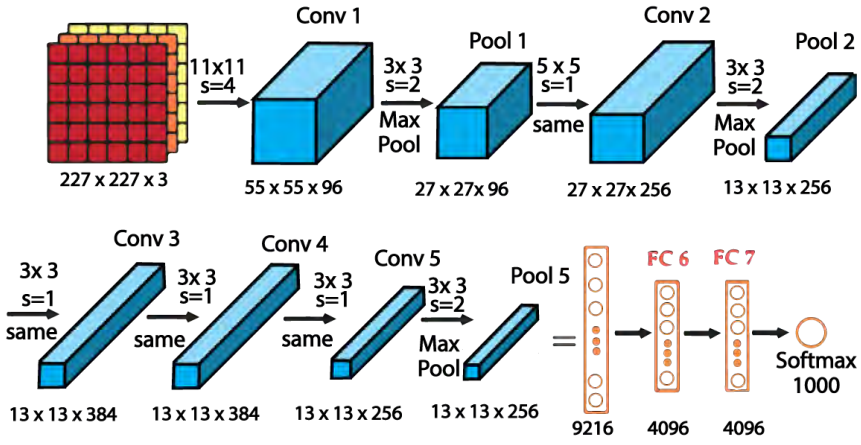

Figure 5.2: AlexNet Architecture [7]

ResNet

ResNet architecture addresses the problem related to vanishing gradient, utilizing residual connections which bypass one or more layers and allow gradients to flow directly to earlier layers. ResNet enables the training of very deep networks like: ResNet-50, ResNet-101 and ResNet-152 without degradation in performance. The unique artchitecture of ResNet has made is good for a wide range of computer vision tasks.

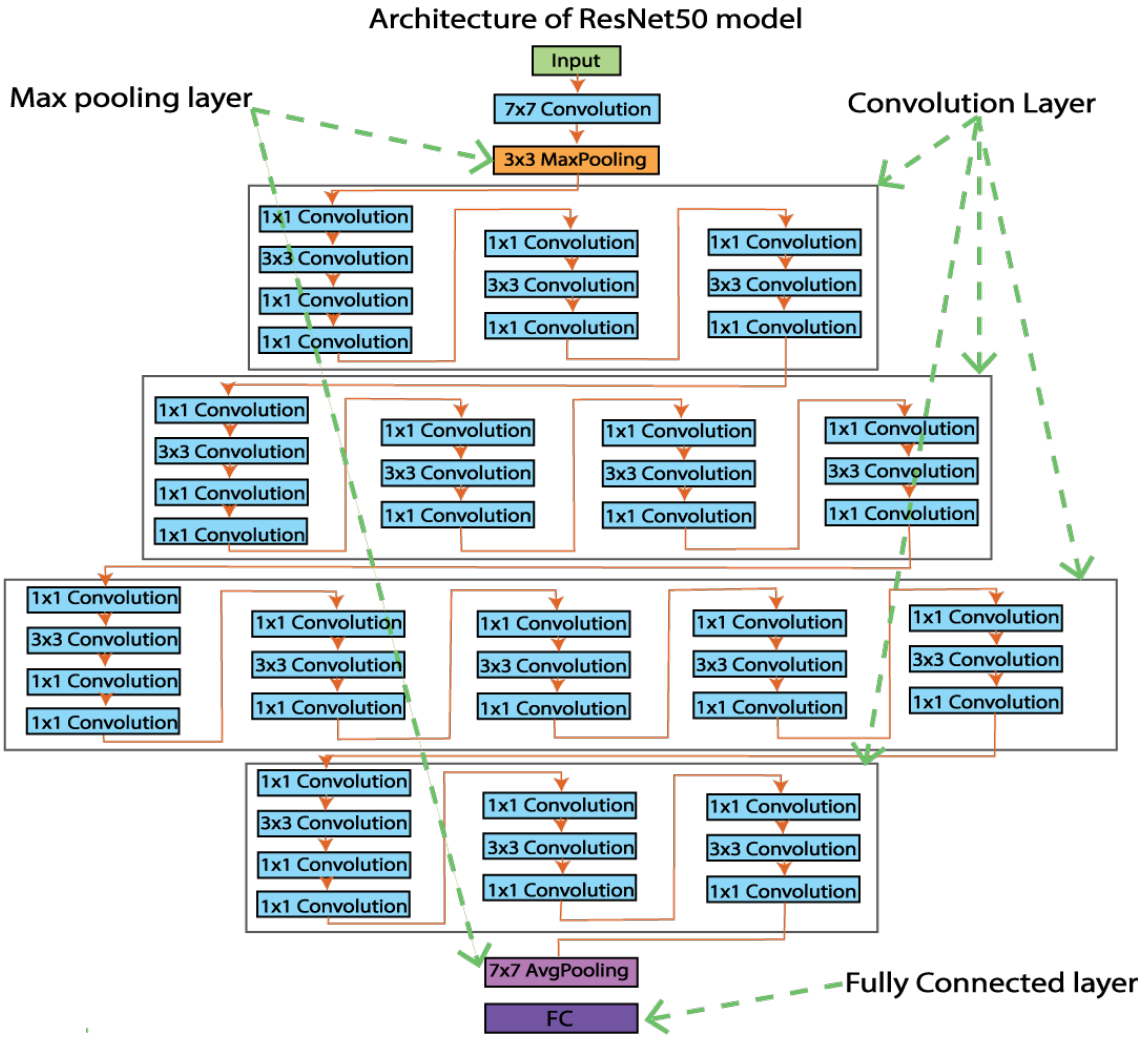

Figure 5.3: Resnet-50 Architecture [31]

MobileNet

MobileNet was developed by Google in 2017. MobileNet is a lightweight CNN architecture. It is good for embedding AI technology in devices with less computational power like mobile devices. This architecture retains good accuracy while drastically reducing the computational cost and model size by using depthwise separable

convolutions and this make it optimal for real-time applications on devices with limited resources [14]. MobileNet comes in multiple versions like: V1, V2 and V3. MobileNet V3 offers the best balance between accuracy and efficiency. MobileNet is widely used in applications like classification of an image, detection of objects in an image, face recognition and real-time vision tasks on mobile and edge devices for the architecture's compact and simple design.

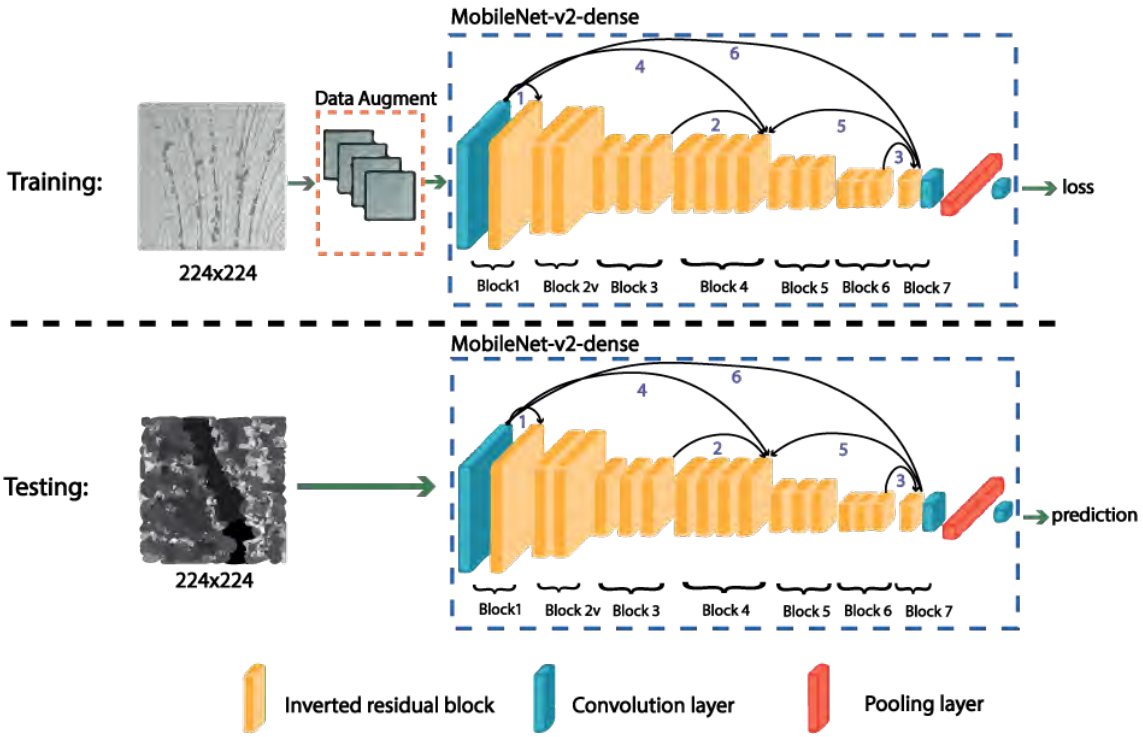

Figure 5.4: MobileNetV2 Architecture [30]

InceptionNet

Another CNN architecture is InceptionNet which was introduced by Google in 2014, designed to enhance computational efficiency and precision in activities involving image recognition. A key feature of InceptionNet is the "Inception module", which processes input data through multiple convolutional filters of varying sizes like 1×1, 3×3, 5×5 simultaneously. This CNN architecture has the capacity to identify intricate patterns, which allows it to record characteristics at various sizes inside the same layer [14]. The original version was known as GoogLeNet or Inception v1 which consists of 22 layers and was notable for its depth and performance. Then Inception v2 and v3 was introduced which featured enhancements like batch normalization and factorized convolutions to further improve traning efficiency. InceptionNet architecture allows for increased width and depth of the network while maintaining a

managable computational budget and this makes it suitable for various applications, including classification of images and detection of objects [14].

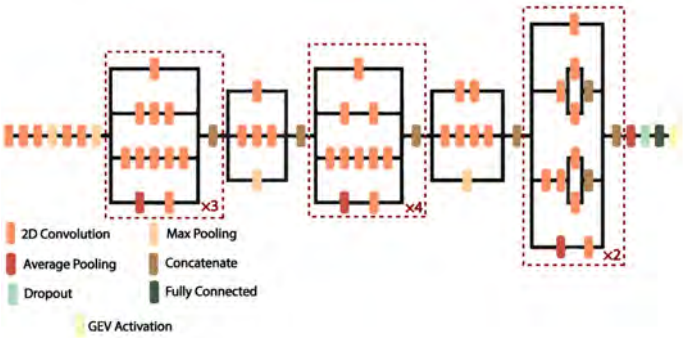

Figure 5.5: InceptionNet-V3 Architecture [36]

VGGNet

The speciality of this deep CNN acrchitecture is the simplicity and better performance in image classification. A small 3×3 convolutional filters and a deep architecture of up to 19 layers make it more efficient at capturing complex patterns. VGGNet's structured design and consistent layer size made it a foundational model for wide range of computer vision applications, including object detection and image segmentation.

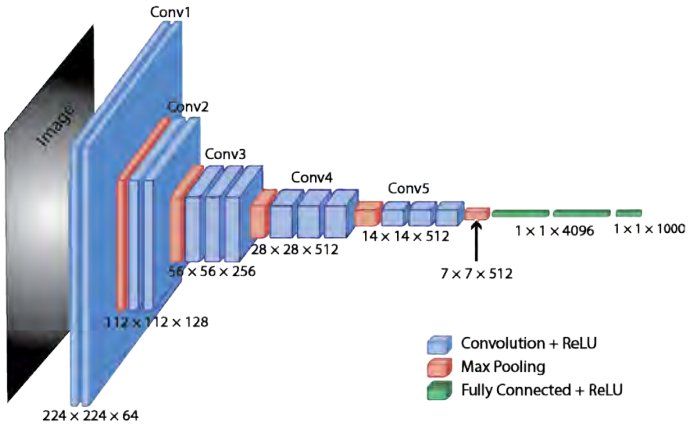

Figure 5.6: VGG-16 Architecture [30]

ConvNeXt

Influenced by the concept of Vision Tranformers, ConvNeXt was first proposed in 2020 in the event "A ConvNet for the 2020s"[39]. The goal of this architecture is to reduce the performance gaps between CNNs and ViTs. One of the key innovative design of ConvNeXt is that it substitute the 7×7 depthwise convolutions for smaller

convolution 3×3. Like the MLP blocks of transformers, this architecture uses expansion ratio in hidden layers which enhances representation capacity. Moreover, with ViT procedures for stability, Layer Normalization (LN) is used in replace of Batch Normalization (BN). For all these reasons, ConvNeXt performs better than classical ResNet architectures.

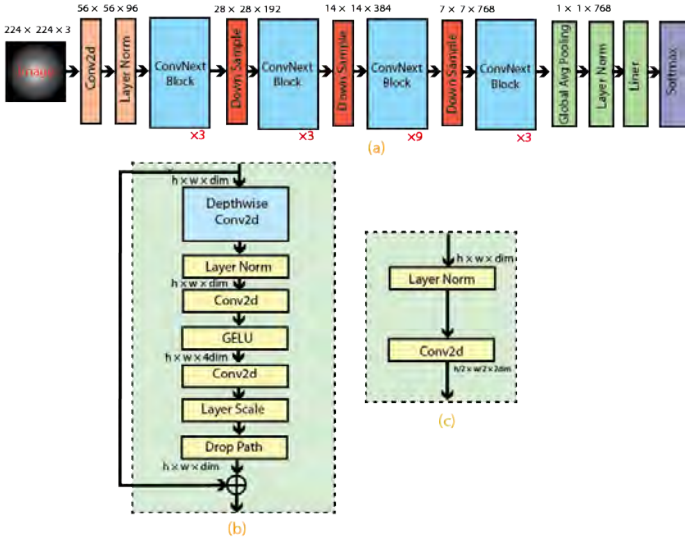

Figure 5.7: ConvNext Architecture [46]

EfficientNet

Google Research introduced EfficientNet first in 2019 [32]. In ResNet and MobileNet there is an issue of arbitrary scaling. This issue is solved in Efficient by adapting a systematic technique to evenly scale network depth, breadth and also the resolution. This architecture also uses MBConv block which is designed by AutoML MNAS and it is mobile-friendly. Instead of using common activation function ReLu, it used Swish activation and uses stochastic depth for regularization. Smooth scalability without redesign and quicker inference compared to larger models are two of the main advantages of EfficientNet.

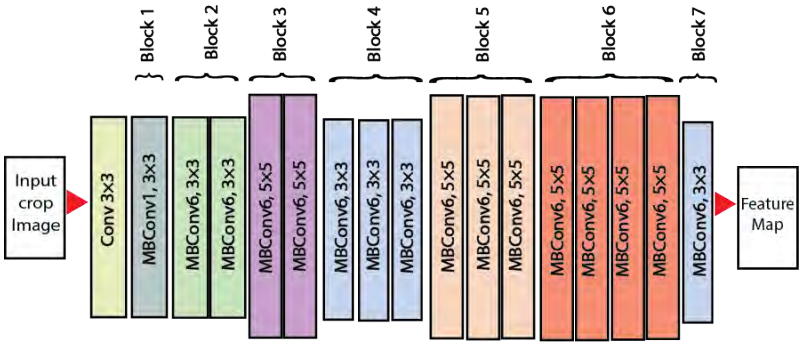

Figure 5.8: EfficientNet Architecture [51]

Convolutional Neural Networks (CNNs) have revolutionized computer vision by enabling systems like image classification, object detection and segmentation with high accuracy. Their robustness to translation and rotation, ability to handle large datasets and support for transfer learning make them highly effective for various applications. However, CNNs come with challenges such as high computational costs, susceptibility to overfitting with limited data and lack of interpretability often making them black box models. Despite these limitations, CNN remains a fundamental building block in deep learning, continuously evolving to become more efficient and accessible for real world applications.

5.2 You Only Look Once (YOLO)

For real time detections of objects, Yolo (You Only Look Once) is a modern architecture for deep learning [12]. Once the input picture is divided into a fixed grid, each grid cell forecasts the odds of a bounding box, confidence value and class simultaneously. This extreamly fast unified methodology allows YOLO to be recognized by a single forward network path. For distinctive extraction, the architecture is usually composed of folding CNN backbone [12].

5.2.1 Different version of YOLO

YOLOv1 introduced a unified detection framework with a custom CNN backbone but it struggled to localize small objects [12]. YOLOv2 made improvements by adding anchor boxes, batch normalization and multi scale training. It used a Darknet-19 backbone for better feature extraction [18]. YOLOv3 improved performance further with a deeper Darknet-53 backbone and multi-scale predictions. This enabled more accurate detection of objects of different sizes [23]. YOLOv4 included techniques like the CSPDarknet53 backbone, Mish activation, spatial pyramid pooling and path aggregation networks to effectively balance speed and accuracy [26]. YOLOv5 was created by Ultralytics to focus on user friendliness and scalability with a PyTorch implementation which offered various model sizes for different performance needs [28]. Later versions like YOLOv6 and YOLOv7 introduced more efficient modules and training methods, pushing the limits of detection speed and accuracy [37], [40]. YOLOv8 introduced a modular and versatile architecture that supports multiple vision tasks beyond detection. YOLOv9 introduced Programmable Gradient Information (PGI) and the GELAN backbone to improve efficiency and reduce information bottlenecks in lightweight models [54]. YOLOv10 eliminated the need for NMS by using a dual-assignment, end-to-end training approach, achieving

faster and more accurate real-time detection [53]. YOLOv11 enhanced multi-task performance with modules like C3k2, SPPF, and C2PSA, supporting detection, segmentation, and pose estimation in a unified architecture [49].

5.2.2 YOLO version 8

Ultralytics published YOLOv8 in 2023. We have used this version of YOLO because it is the latest version offered by Ultralytics that can be imported and can be used for object detection directly in Google Colab. YOLOv8 offers a completely modular design integrated into PyTorch, it marks a significant advancement in the YOLO series. Withing one cohesive framework, the version 8 of YOLO can handle a variety of vision tasks, such as object identification, instance segmentation and classification[41]. This version incorporates advanced backbone and neck designs that optimize feature extraction and fusion, alongside novel training strategies that improve convergence and accuracy. YOLOv8 is optimized for real time deployment as it offers improved inference speed without compromising precision. Its flexible design simplifies customization and integration across diverse applications making it a state of the art model that balances efficiency and versatility in computer vision [41].

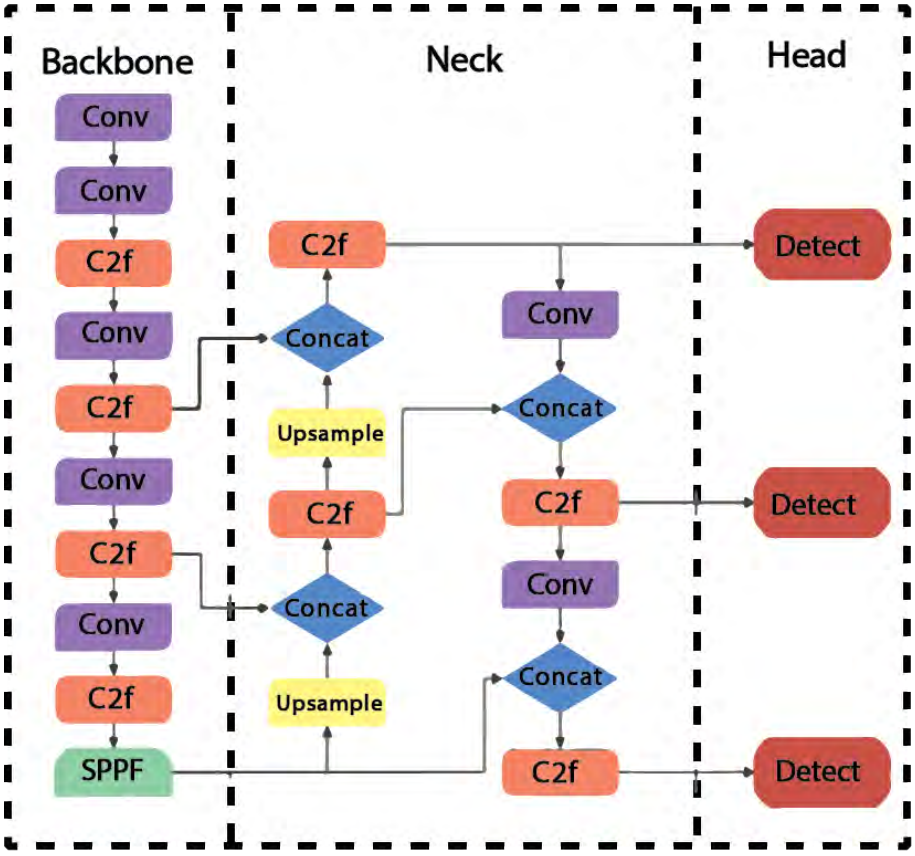

Figure 5.9: YOLOv8 Architecture [48]

Chapter 6

Model Building

6.1 Composition Detection Model

In the first stage of the system, the goal is to determine the composition of a photo by identifying whether it follows any of the predefined photographic rules such as Center Composition (CC), Rule of Thirds (ROT), Leading Lines (LL), Frame in Frame (FIF) or Patterns (PAT). The dataset is organized with a set of images placed in a directory and a corresponding CSV annotation file that contains image filenames and their respective composition class labels. The model is expected to classify each input image into one of these five composition categories.

6.1.1 Importing Necessary Libraries

To develop and train the composition detection model, several essential Python libraries were imported. TensorFlow and Keras were used for constructing and training the deep learning model, while NumPy provided support for numerical computations. Image loading and preprocessing were handled using PIL.Image and cv2. For reading CSV annotations and managing the dataset, pandas and os were utilized. Visualization of training performance was done using matplotlib.pyplot, and tools like sklearn.metrics were used for evaluating classification accuracy and generating confusion matrices. These libraries formed the foundational building blocks of the entire pipeline.

6.1.2 Model Architecture and Training

For composition classification we have used 3 different transfer learning models (EfficientNetB0, InceptionV3 and VGG-19) with fine-tuning for a result comparison and analysis.

EfficientNet B0

This transfer learning model was imported from tensorflow.keras library and built by removing the initial classification head and adding custom classification weight. The custom head involves a global average pooling followed by two dense layers with softmax output and 256-unit ReLu. For regularization we applied dropout (0.5). By following this method, pre-trained heirarchical characteristics of the model are preserved while the model is being adjusted according our goal. The model was trained in 2 stages. For initial training we used frozen base and train the model for 15 epochs using batch size of 8. For stage-2, we fine-tuned the mode by unfreezing last 100 layers and freezing the rest. Then the model was trained for 50 epochs. To calculate the loss, sparse categorical cross-entropy has been used.

VGG-19

A VGG-19 pre-trained model is used with custom classification head that consists of three dense 256-units ReLu layers each and dropout of 0.5 with L2 regularization. Then a softmax output layer is added for multi-class classification along with the other layers of custom classification head. The model is compiled with initial learning rate of 0.001 and Adam optimizer. However, we used decaying learning rate monitoring the minimum validation loss with a patience value of 3 epochs. The model was initially frozen to maintain pre-learned weight and trained for 10 epochs. Then the model was fine-tuned by releasing the last 10 layers and trained for another 10 epochs.

Inception V3

The pre-trained ImageNet InceptionV3 model is modified by subsititing classification head with a custom head of one 256-unit ReLu dense layer, dropout layer of 0.5 and softmax output for our classification model. Three stages are used in the training strategy. In stage-1, the model is trained for 2 epochs with batch size of 4, fine-tuned by releasing layer 150 onwards and learning rate set to 0.001. Then, in stage-2, the model is fine-tuned by releasing from layers 200 onwards and trained for 2 epochs with batch size of 8 and a learning rate of 0.0005. Lastly, in stage-3, the model is trained unfreezing layers from 200 and onwards and trained for 2 epochs with a batchsize of 16 and learning rate of 0.0001. This training method is know as progressive unfreezing. Additionally, the model is trained with Adam optimizer, sparse categorical crossentropy as loss and callbacks for early stopping, learning rate reductions and checkpoints.

6.1.3 Evaluation and Visualization

After training, the performance of model was evaluated on the test dataset using classification reports and confusion matrices. Visualization tools such as ConfusionMatrixDisplay from scikit-learn and training history plots (loss vs. accuracy) from Matplotlib were used to assess model convergence and class-wise performance. These insights helped verify whether the model could reliably classify different composition styles.

6.2 Orientation Detection Model

In the second stage of our system, the goal was to determine whether the input image was properly oriented which means if the angle is at 0 degree angle or tilted. Instead of correcting the image directly, our model predicts the degree and direction like clockwise or counterclockwise in which the camera should be adjusted to achieve a proper upright orientation. We trained our model using the customized DRC-D dataset, which contains natural images with rotation variations.

6.2.1 Importing Necessary Libraries

To implement and train the orientation model, we imported various essential Python libraries such as: cv2 for image manipulation, tensorflow and keras for model development, numpy for numerical operations and matplotlib for result visualization. Other helpful modules like os, random and sklearn were also used to manage data splitting, loading and evaluation.

6.2.2 Model Architecture and Training Framework

We have constructed four distinct models for orientation detection and compared the outcomes for evaluation. The four model includes one custom model, a Resnet50 V2 model, a ConvNeXt model and an EfficientNet B2 model.

Custom Model

We built a custom CNN-based regression model using TensorFlow and Keras. Instead of classifying images into fixed rotation classes, our model was trained to regress the exact angle by which the image is rotated. This makes the model more flexible for real-world camera orientation guidance. The model architecture includes 4-layer CNN with increasing filter from 32 to 256. The model uses gobal average pooling, dense layers, batch normalization, max pooling and L2 regularization. Swish activation has been used in the last dense layer for gradient flow. To avoid

overfitting dropout layer from 0.2 to 0.5 has been used along with weight decay. Adam optimizer is utilized with weight decay of 0.005. The model was trained for 15 epochs.

ResNet50 V2

The ResNet50 V2 is modified with convolutional attention mechanism which is activated by 1×1 sigmoid, applied to feature maps prior to pooling. Only top layers are frozen while the last 143 layers are fine-tuned. Moreover, to highlight imporant regions, spatial attention weights used to increase the feature maps. 128-unit ReLu dense layer is added as custom regression head along with global pooling. The model is trained using 15 epochs.

ConvNeXt

ConvNeXt base model is modified using custom regression head of 512-unit ReLU dense layer and linear output for orientation prediction. For speedier computation withou sacrificn accuracy, mixed-precision training using datatype of float16 and float32 is used. As the optimizer, Adam optimizer is used with a learning rate of 0.0001 and weight decay of 0.005. Fine-tuning is acheived by releasing the last 30 layers of the architecture. Lastly, L2 weight decay and dropout of 0.3 is added to reduce overfitting in dense layer. The model is trained using 8 epochs.

EfficientNet B2

Utilizing pre-trained EfficientNet B2 backbone with ImageNet weights, the model is enhanced with a custom regression head consisting of linear output lear and 256-unit dense layer with ReLU activation followed by droupout rate of 0.3. Fine-tuning is achieved by making the first 50 and last 50 layers trainable. For optimization AdamW is used with a learning rate of 0.00005. The evaluation metrics while training were mean squared error (MSE) and mean absolute error (MAE). When handling the high-resolution 384 by 512 pixel input, regularization strategies like weight decay and dropout help avoid overfitting. The model was trained for 20 epochs with a batch size of 16.

6.2.3 Loss Function and Optimization

Since this was a regression task, we used Mean Absolute Error (MAE) as our primary loss function. The model was compiled using the Adam optimizer with a learning rate scheduler and early stopping to prevent overfitting. We also tracked metrics like Mean Squared Error (MSE) during training for deeper insight.

6.2.4 Prediction and Angle Suggestion

During inference, the model receives an image and predicts a float value indicating how many degrees the camera should rotate in clockwise or counterclockwise to align the image upright. For example, if the prediction is -3.4°, the suggestion would be to rotate the camera 3.4° clockwise.

6.2.5 Visual Feedback and Testing

To visualize predictions, we overlaid the predicted rotation angle on sample test images. We also plotted training and validation MAE loss curves using Matplotlib to evaluate model convergence. For additional visual cues, tilted images and their predicted correction angles were displayed side-by-side.

6.2.6 Evaluation and Model Utilities

We used scikit-learn's mean_absolute_error function to validate performance on the test set. The model achieved a low average MAE on unseen data, demonstrating strong generalization. TensorBoard callbacks were optionally used for logging training performance over time.

6.3 Subject Detection Model

In the final stage of our system, the primary objective was to detect all objects within an input image and determine the main subject among them. This model leverages the YOLOv8 object detection architecture, pre-trained on the COCO dataset, to identify objects. The main subject is selected based on object area and spatial relations like overlap. This serves as the foundational block for further composition suggestions in later models.

6.3.1 Model Setup and Initialization

We imported different Python libraries such as Ultralytics' YOLO for object detection, OpenCV for image processing, NumPy for numerical operations, and PyTorch as the backend framework. PIL was used for image loading, and Matplotlib for visualization. Additionally, io and math supported data handling and calculations. We used the YOLOv8s model with pre-trained weights (yolov8s.pt) from the COCO dataset, enabling detection of 80+ object classes without additional training.

6.3.2 Image Inference and Object Detection

An input image is loaded using PIL, and inference is performed using the YOLO model. The detected objects are returned along with their bounding boxes, class labels and confidence scores. These detections are visualized on the image for interpretability. For each detection, we extract bounding box coordinates, confidence scores, class labels (e.g., person, car, cat), and compute the bounding box area. All object information is stored in a list for further analysis.

6.3.3 Main Subject Identification

To determine the main subject, we apply a multi-step strategy based on bounding box area and spatial overlap. The strategies includes the following:

- If no objects overlaps:
  - The largest object is selected as the subject.
- If only one object overlaps the largest object:
  - The object infront (smaller) is selected as the subject.
- If more than one object overlaps the largest object:
  - The object infront (smaller than the largest) and with highest confidence is selected as subject.
  - If no such multi-overlapping object exists, rely on the rule no.2.

These rules follow the assumption that prominent subjects either dominate the frame or appear in the foreground. Overlap is assessed using Intersection over Union (IoU) and relative area metrics. The smaller object must overlap at least 50% on the largest object for it to be considered as an overlap.

6.3.4 Visualization and Composition Guidance

All detected objects are annotated using OpenCV with distinct colors—green for general objects, orange for overlapping ones, red for the main subject, and magenta for the final subject box. To support composition-based suggestions, the subject's center is computed. Two guidance functions are implemented: the Center Composition (CC) suggestion, which compares the subject's center to the image center, and the Rule of Thirds (ROT) suggestion, which evaluates its proximity to the nearest intersection point in a rule-of-thirds grid. The suggestions are visually represented and printed as feedback to guide better framing.

6.4 Final Suggestion System

All three detection models are used to generate the final suggestion system. The logic for building the suggestion follows:

- If composition detection predicts ROT or CC:
  - → Use the orientation detection model to predict the orientation of the image and then suggest how to make it ground level (0°)
  - → If subject found, generate camera movement suggestion according to center point postion of the subject (from subject detection model) according to the composition followed.
    - ⇒ If ROT, find the nearest Rule of Thirds intersection points to generate camera movement suggestion.
    - ⇒ If CC, use the center of the image to generate camera movement suggestion.
  - → If no subject found, do not generate any camera movement suggestion.
- If composition detection predicts LL or FIF:
  - → Use the orientation detection model to predict the orientation of the image and then suggest how to make it ground level (0°)
  - → If subject found, generate camera movement suggestion accordin to the minimum distance between centerpoint or Rule of Thirds intersection points.
  - → If no subject found, do not generate any camera movement suggestion.
- If composition detection predicts PAT:
  - → Do not generate any orientation suggestion.
  - → Do not generate any camera movement suggestion.

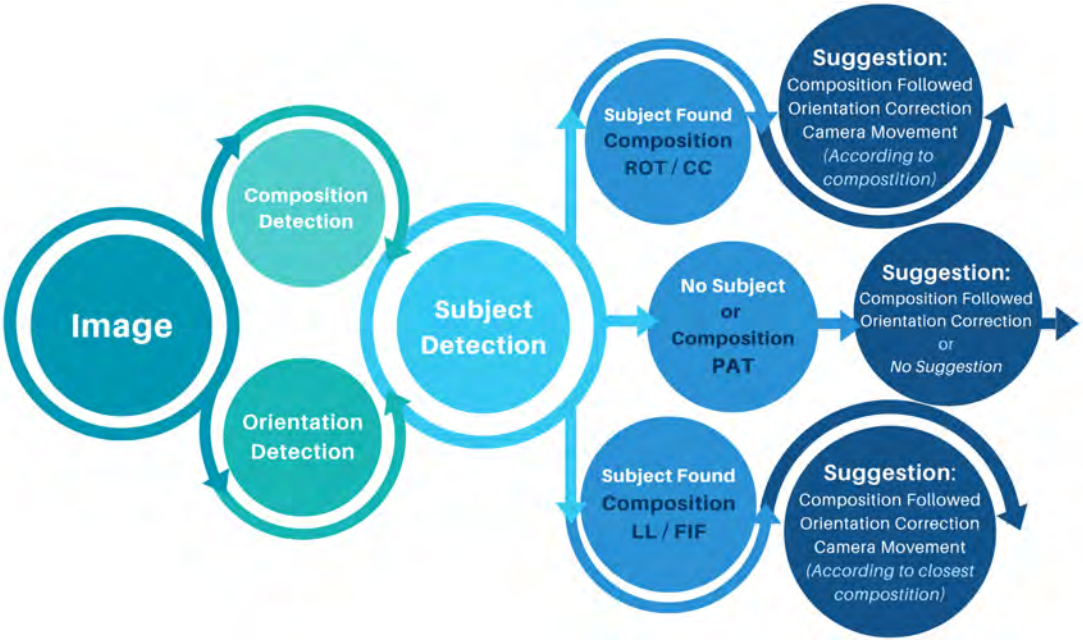

Figure 6.1: Visualization of Final System

Chapter 7

Result Analysis

7.1 Composition Detection Model

7.1.1 Efficientnet B0

In our model, the accuracy for the Efficientnet B0 model is 58.55% and the F1 scores claim that this model performed best for Patterns class and worst for the Leading Lines class.

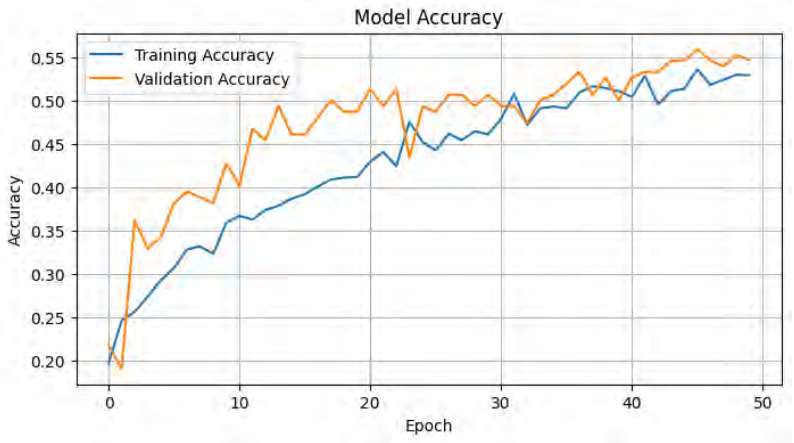

Figure 7.1: Efficientnet B0 Model Accuracy

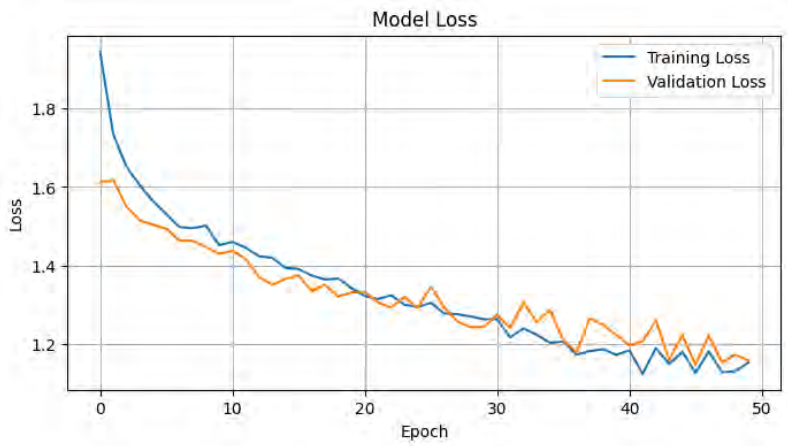

Figure 7.2: Efficientnet B0 Model Loss

The accuracy, training and validation performance of EfficientNet B0 over 50 epochs is displayed in the two graphs above. The validation accuracy peaked at 58.55 and train accuracy peaked at 52.5% which is when the model is saved according to call-back conditions. The model seems not to overfit as the train and validation loss almost merges in the end.

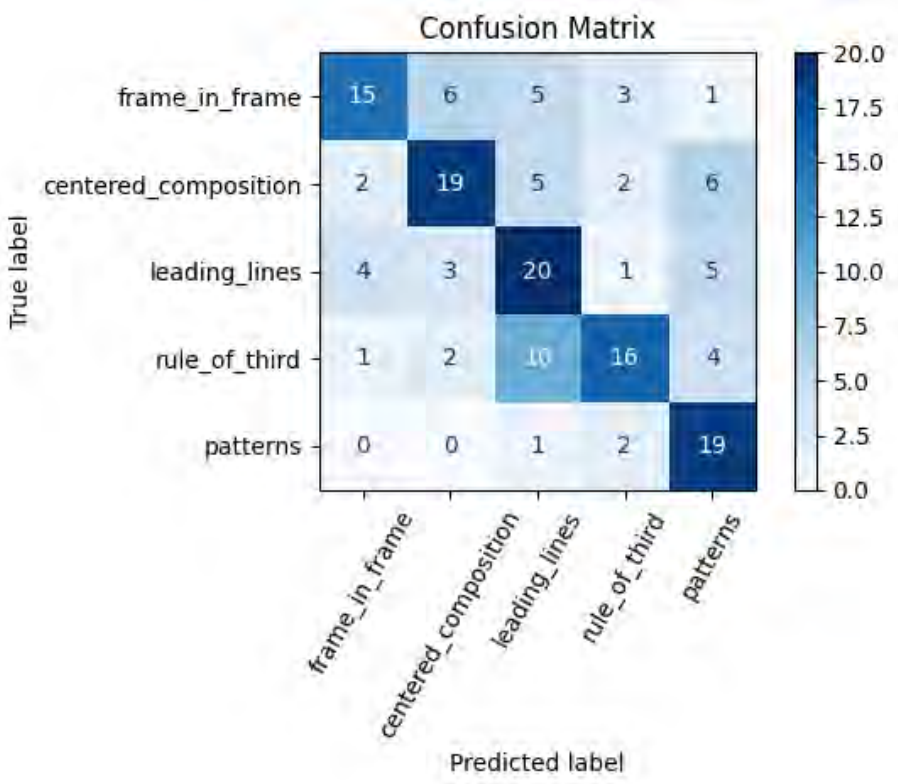

Figure 7.3: Efficientnet B0 Model Confusion Matrix

The classification into 5 different class on several images by the model can be seen in the confusion matrix. The diagonal (15, 19, 20, 16, 19) displays accurate predictions, with ”patterns” struggling the most.

Class Name	Precision	Recall	F1-Score	Support
Centered Composition and Symmetry	0.63	0.56	0.59	34
Frame in Frame	0.68	0.50	0.58	30
Pattern	0.54	0.86	0.67	22
Rule of Third	0.67	0.48	0.56	33
Leading Lines	0.49	0.61	0.54	33

Table 7.1: Efficientnet B0: Precision, Recall, F1-Score, and Support for Each Class

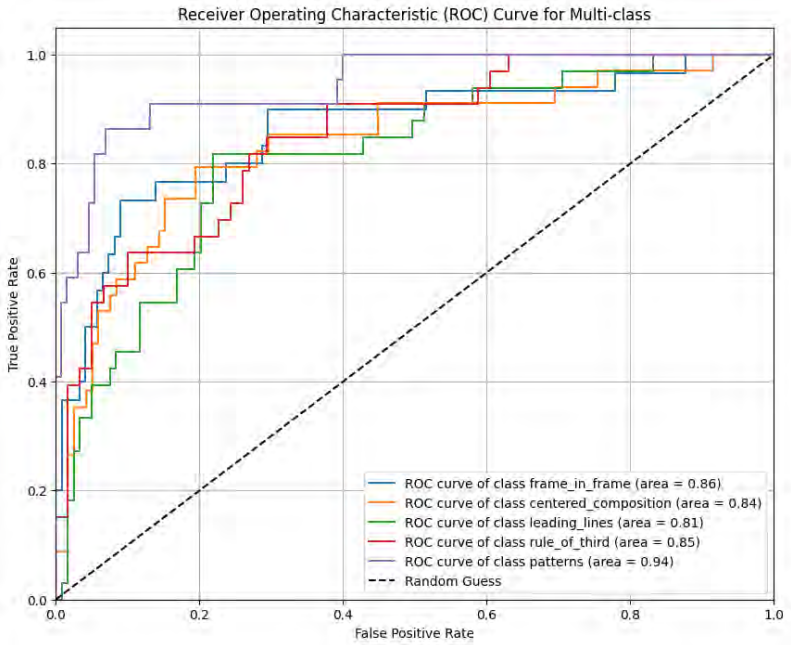

Figure 7.4: Efficientnet B0 Model ROC

In EfficientNetB0, the validation macro F1-score of 0.59 and accuracy of 59% indicate moderate performance in photo composition detection. The model performed best on the patterns class (F1: 0.67, recall: 0.86), but struggled with leading lines (precision: 0.49) and showed low recall for frame in frame and rule of thirds. The accuracy and loss graphs indicate stable training without significant overfitting. Additionally, ROC AUC scores from 0.81 to 0.94 show that the model can effectively separate the classes.

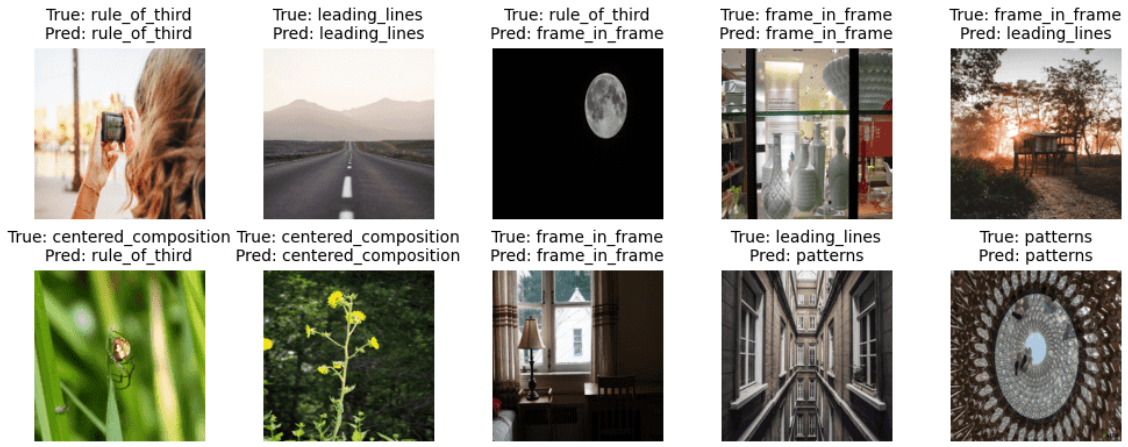

Figure 7.5: Efficientnet B0 Test Result

7.1.2 InceptionV3

The accuracy of InceptionV3 is 74.34% and it shows a balanced performance for all classes.

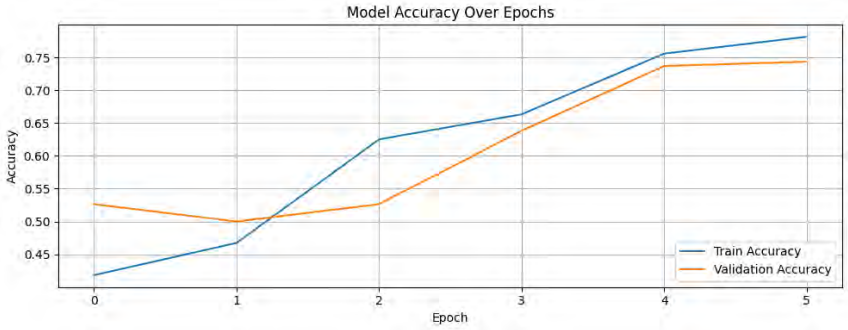

Figure 7.6: InceptionV3 Model Accuracy

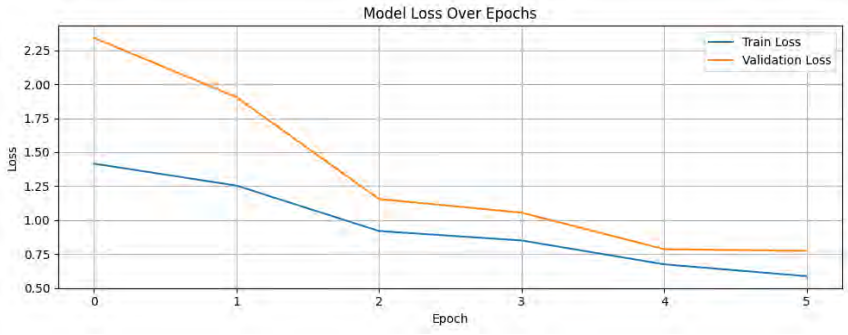

Figure 7.7: InceptionV3 Model Loss

The Inception V3 model shows to increase training and validation accuracy rapidly through all 5 epochs. The model ended training at 0.5598 training loss and 0.7737 validation loss. Which indicates that the model is not overfitting but would overfit if it was trained with higher epochs.

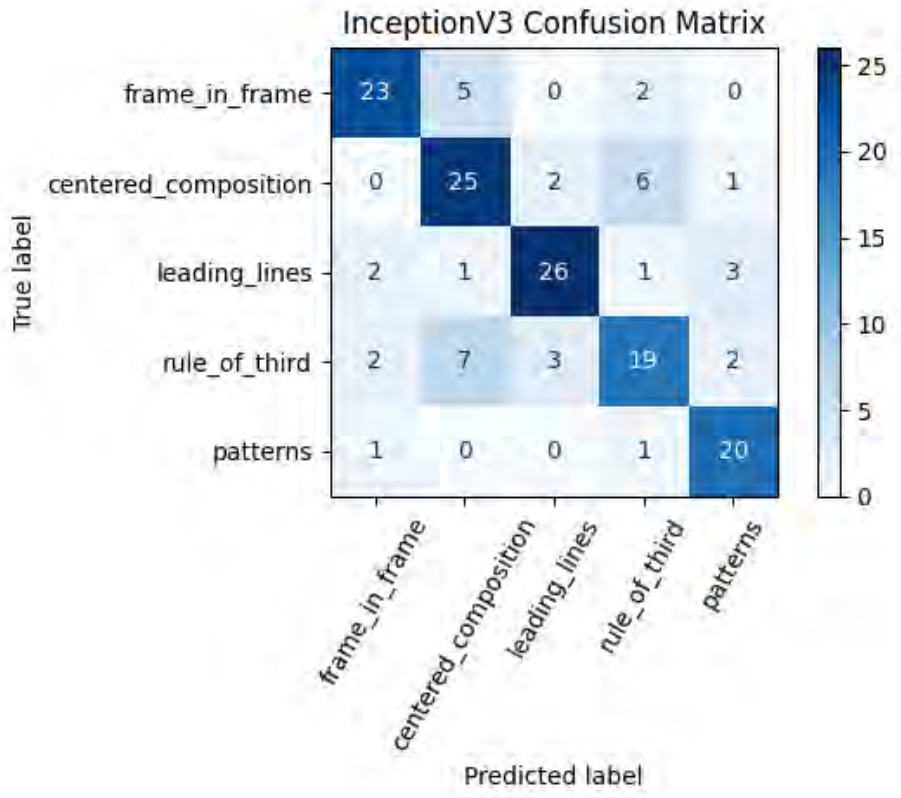

Figure 7.8: InceptionV3 Model Confusion Matrix

The confusion matrix of Inception V3 shows that the classification performance across various compositional styles is very good. The model confuses a bit between Rule of Thirds and Centered Compostion. However, it will not affect significantly in the final suggestion system.

Class Name	Precision	Recall	F1-Score	Support
Centered Composition and Symmetry	0.66	0.74	0.69	34
Frame in Frame	0.82	0.77	0.79	30
Pattern	0.77	0.91	0.83	22
Rule of Third	0.66	0.58	0.61	33
Leading Lines	0.84	0.79	0.81	33

Table 7.2: InceptionV3: Precision, Recall, F1-Score, and Support for Each Class

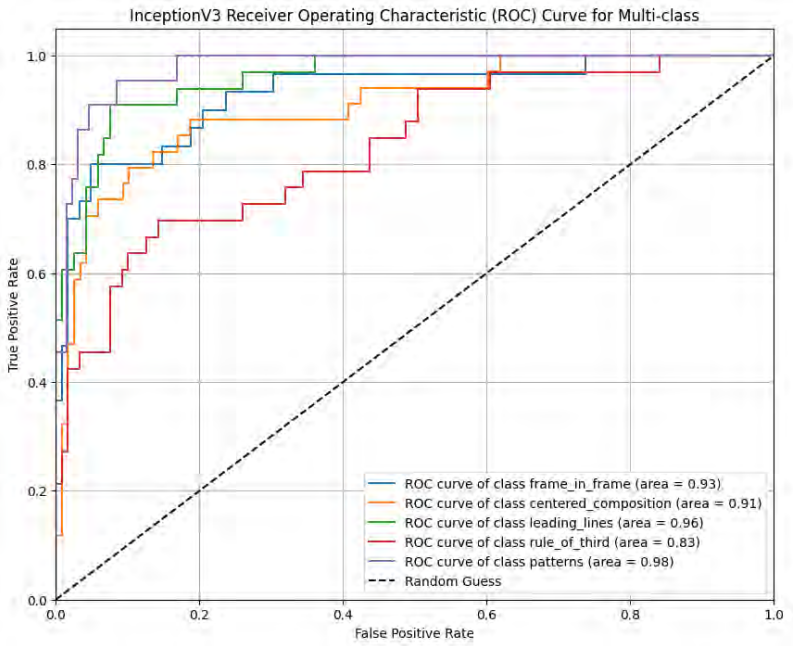

Figure 7.9: InceptionV3 Model ROC

The InceptionV3 model performed strongly in classifying photographic composition styles with a validation accuracy of 74% and a macro F1-score of 0.75. It performed best on the Patterns class (F1: 0.83, AUC: 0.98) and struggled slightly with detecting rule_of_third (F1: 0.61) due to visual similarity with other classes. The high ROC AUC scores across all classes (0.83–0.98) indicate strong class separability.

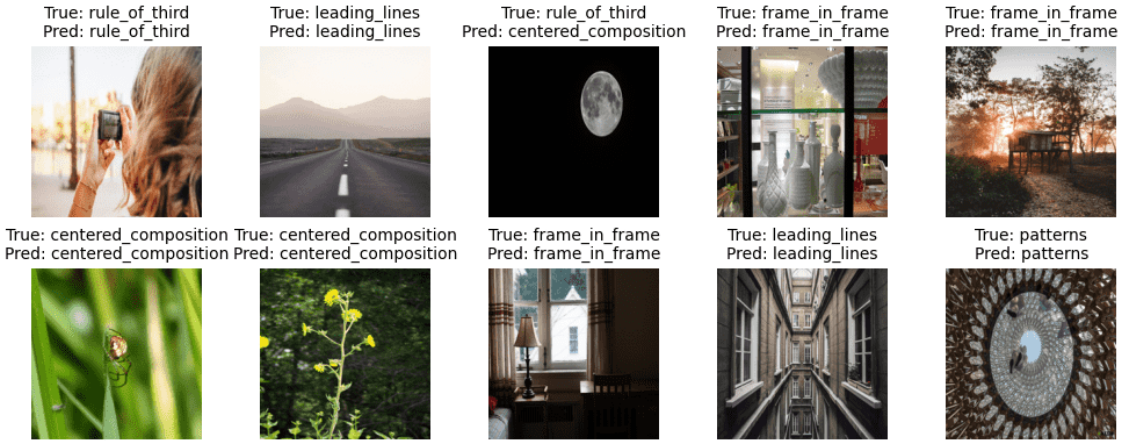

Figure 7.10: InceptionV3 Model Test Result

7.1.3 VGG19

The overall accuracy of VGG19 is 68.42% and it performed well in classifying Frame in Frame and Patterns.

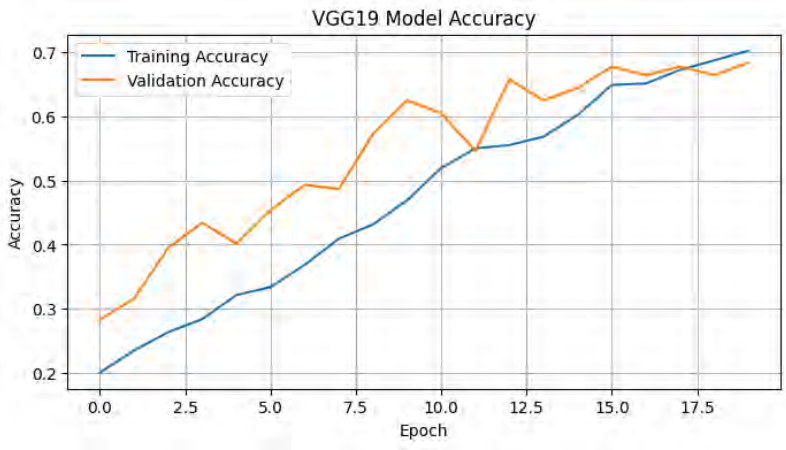

Figure 7.11: InceptionV3 Model Accuracy

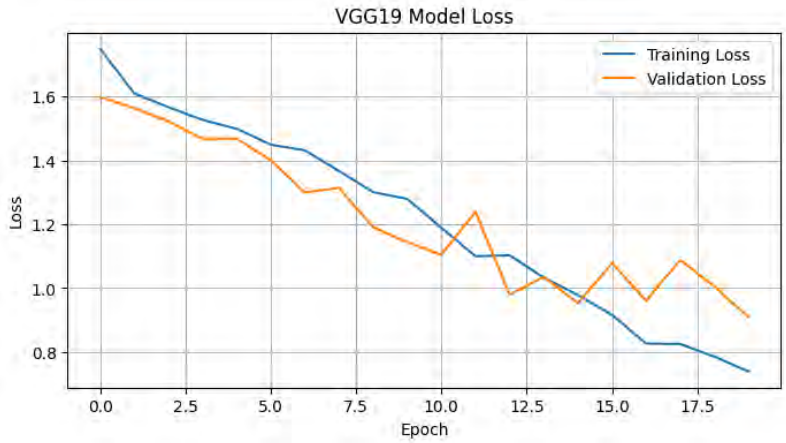

Figure 7.12: VGG19 Model Loss

The accuracy curve of VGG-19 shows that, as the epoch increases train and validation accuracy also increase to roughly 70% and 60% accordingly. Similar to this, the loss curves show a steady convergence pattern, with training and validation losses declining simultaneously before leveling off at around 0.8 and 1.0 accordingly. This indicated the model is appropiate for the given task complexity.

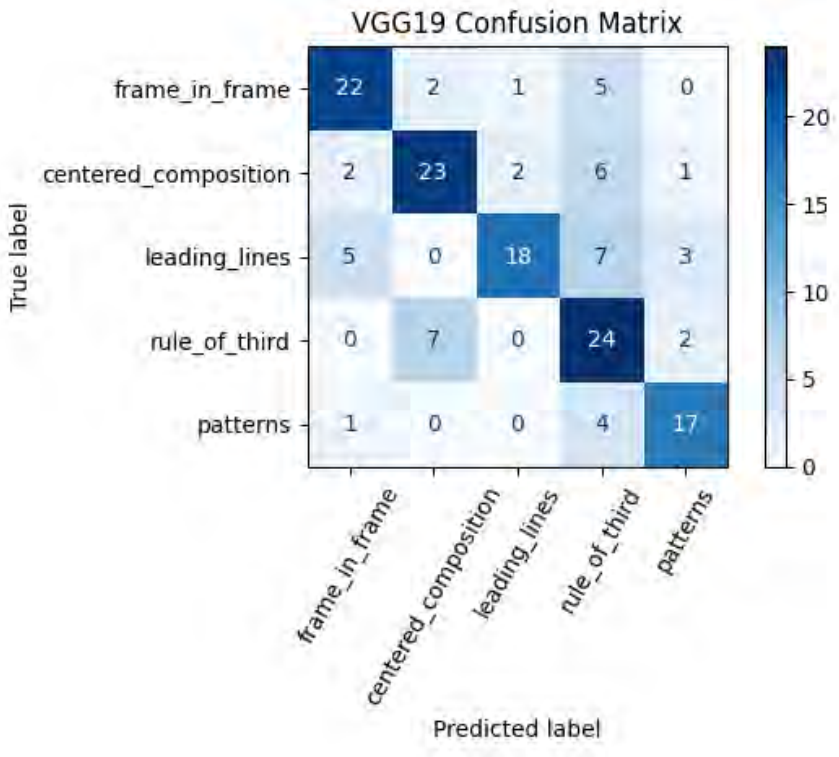

Figure 7.13: VGG19 Model Confusion Matrix

Analyzing the confusion matrix we can see that, it occasionally fails to predict similar styles between Centered Composition and Leading Lines (6-7 errors), the VGG-19 performs good is other cases. Specially it almost flawless in categorizing images of Pattern composition(17/18).

Class Name	Precision	Recall	F1-Score	Support
Centered Composition and Symmetry	0.72	0.68	0.70	34
Frame in Frame	0.73	0.73	0.73	30
Pattern	0.74	0.97	0.76	22
Rule of Third	0.52	0.73	0.61	33
Leading Lines	0.86	0.55	0.67	33

Table 7.3: VGG19: Precision, Recall, F1-Score, and Support for Each Class

The validation accuracy of VGG19 is 68% and macro F1 score is close to 0.70. So the performance of this algorithm was moderate to strong in classifying photographic composition. It performed best on patterns (recall: 0.77, AUC: 0.95) and leading_lines (precision: 0.86). Some misclassifications occurred in visually similar classes. The accuracy and loss curves showed stable training with no significant overfitting. ROC AUC scores ranging from 0.85 to 0.95 confirmed strong class separability.

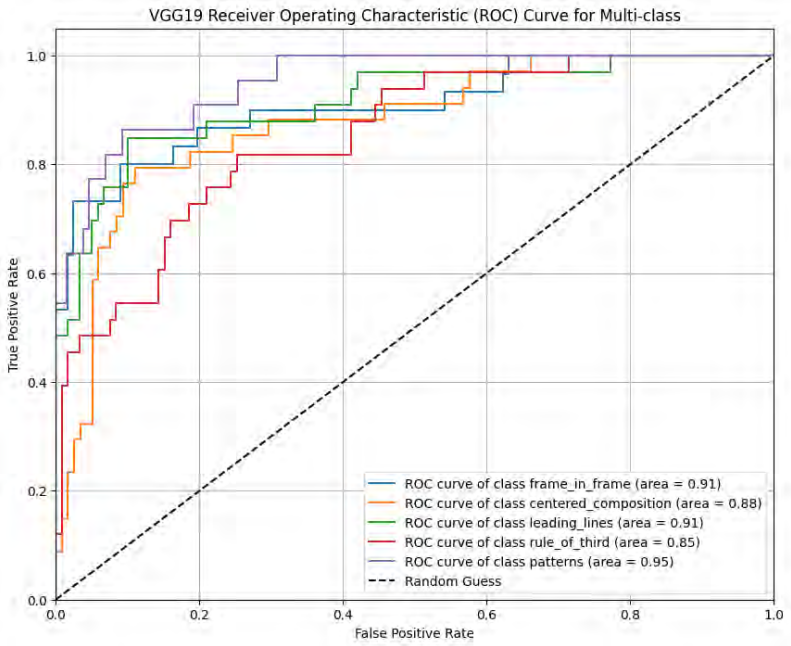

Figure 7.14: VGG19 Model ROC

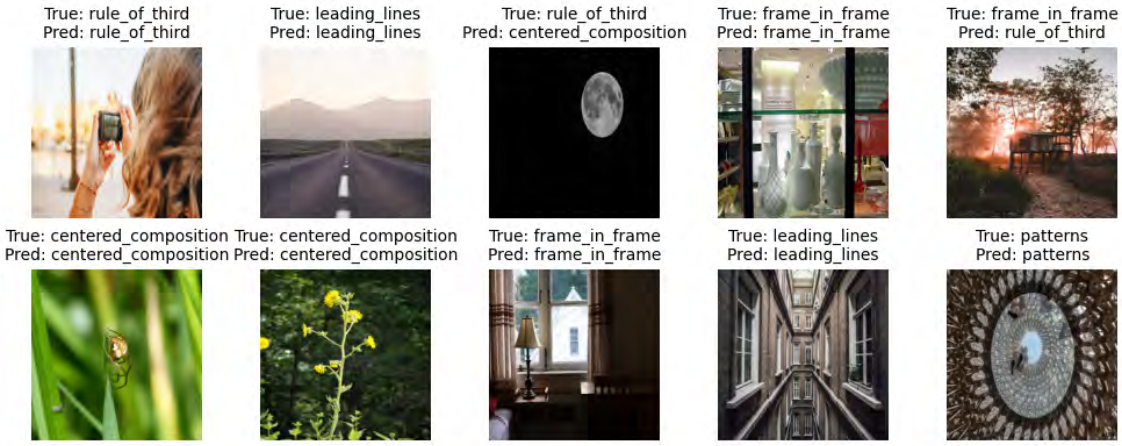

Figure 7.15: VGG19 Model Test Result

7.1.4 Composition Detection Models Comparison

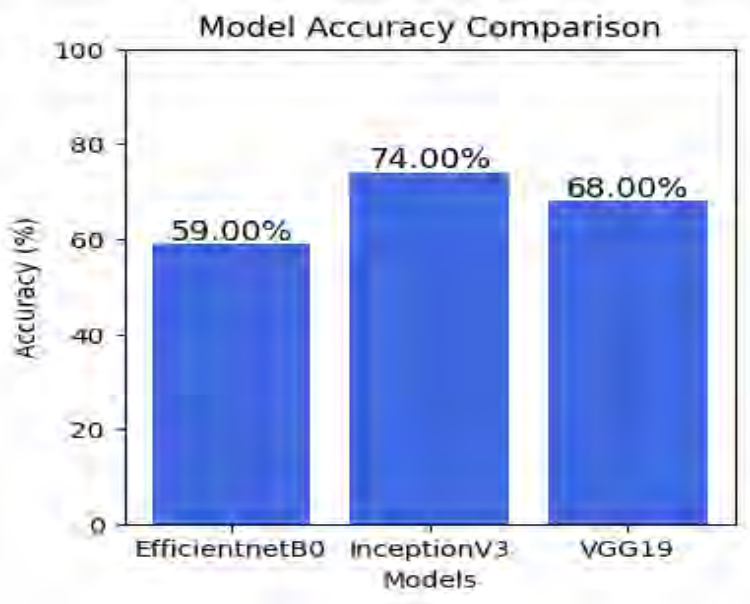

Figure 7.16: Accuracy of EfficientNetB0, InceptionV3 and VGG19

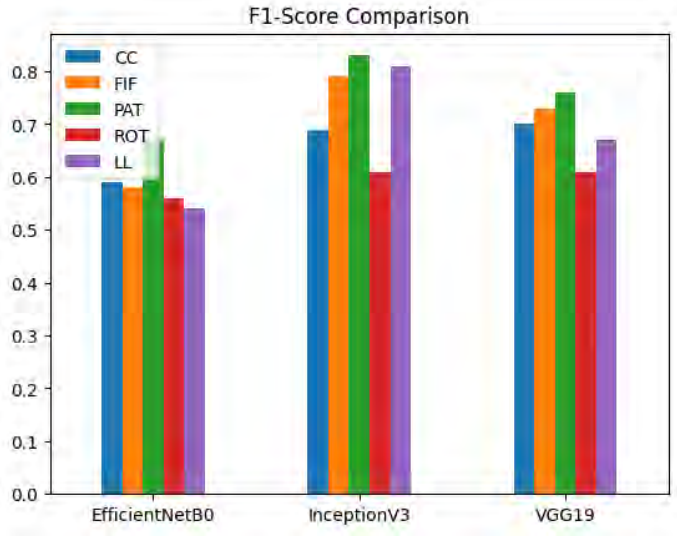

Figure 7.17: F1 Score of Implemented CNN Architectures

It is clear that Inception V3 performs the best at classifying the images based on their compostion when comparing the evaluation metrics of all the composition detection models. Overall, the model shows no indications of overfitting and the accuracy of the model along with F1-scores through all classes are higher compared to other models that have been trained. Furthermore, the InceptoinV3 model mostly confuses between Rule of Thirds and Centered Composition. Since these two classes are quite comparable and the final suggestion system indicates that the logic of both classes is similar and it will not cause any major error.

7.2 Orientation Detection Model

7.2.1 Custom Model

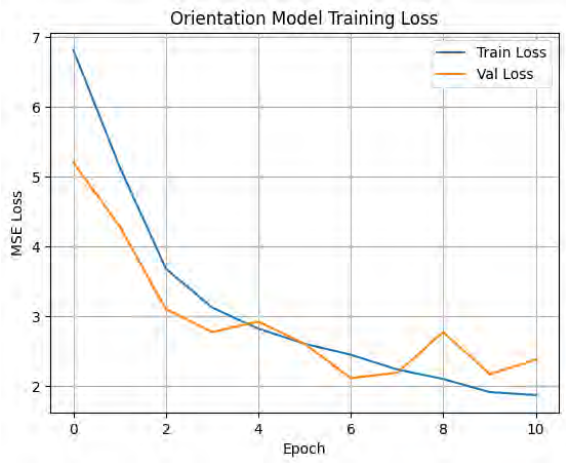

Figure 7.18: Train VS Validation loss

From figure-7.18 we can see that the validation loss and train loss decreases gradually over epoch. However, signs of starting to overfit can be noticable as the validation loss is increase in the end while the train loss is decreasing.

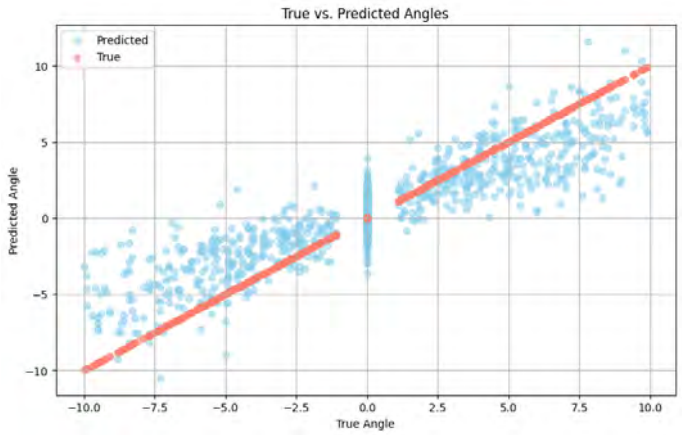

Figure 7.19: Truth VS Predicted Angle

Our custom model does not perform very well with the test dataset as we can see from figure:7.17. The model even fails to predict orientation of images that are labeled as 0°.

Figure 7.20: Test Result of Custom Model

7.2.2 ResNet50 V2

Figure 7.21: Train VS Validation loss

The train vs validation loss graph (figure-7.19) shows that ResNet50 is not doing a good job in training. While the train loss is gradually decreasing, the validation loss does not seem to decrease much the begining till the end.

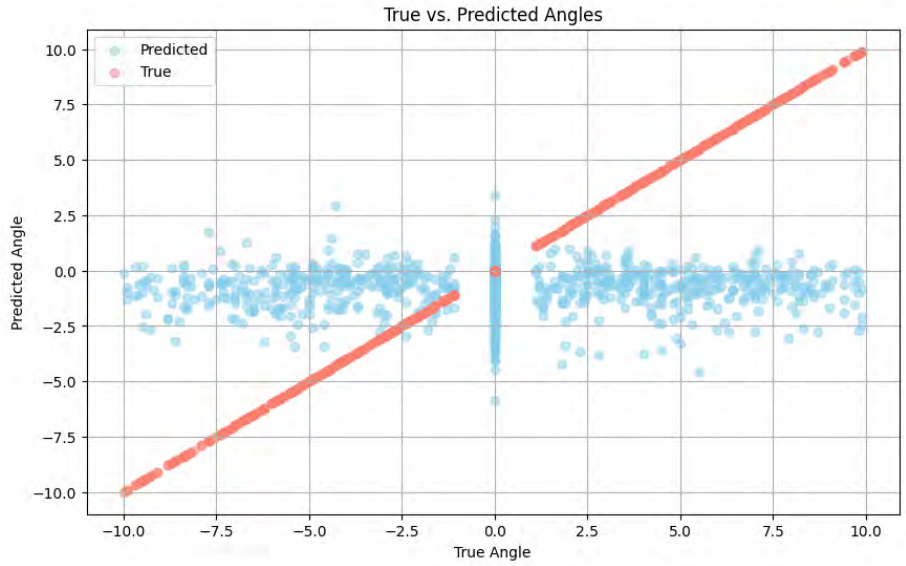

Figure 7.22: Truth VS Predicted Angle

As expected from the training process, ResNet50 V2 does not perform well on the test data which can be seen in figure:7.20. The difference between the predicted and true is huge in many cases.

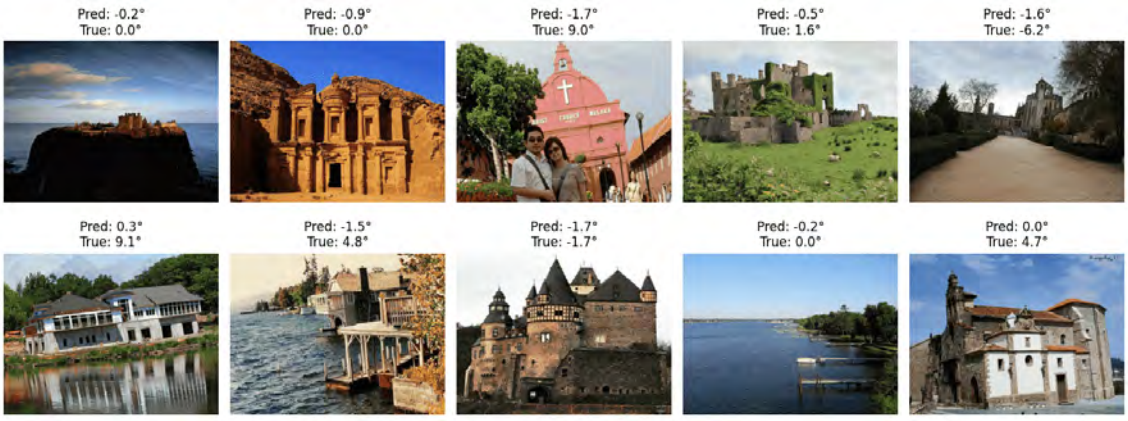

Figure 7.23: Test Result of Resnet50V2 Model

7.2.3 ConvNeXt

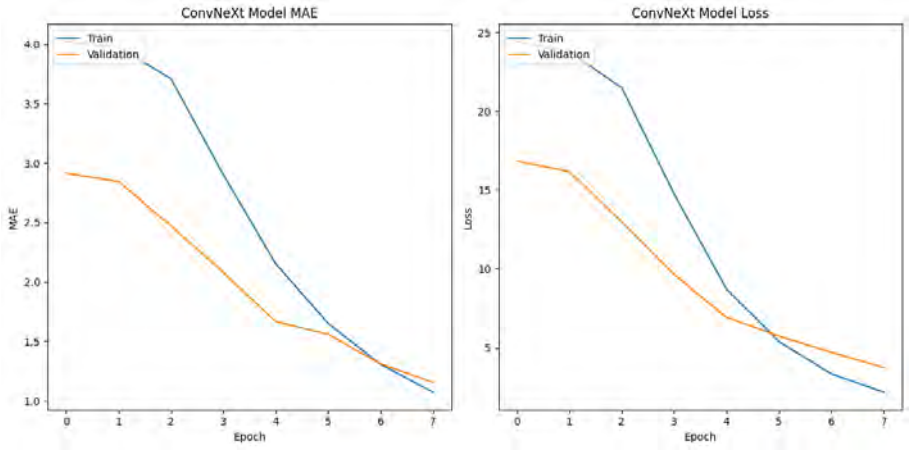

Figure 7.24: Train VS Validation loss

The stable alignment between trainig and validation, which can be seen in the figure:7.22, indicated balanced model capacity and effective regularization. However, the graph also shows that the model is staring to show signs of overfitting after epoch 6.

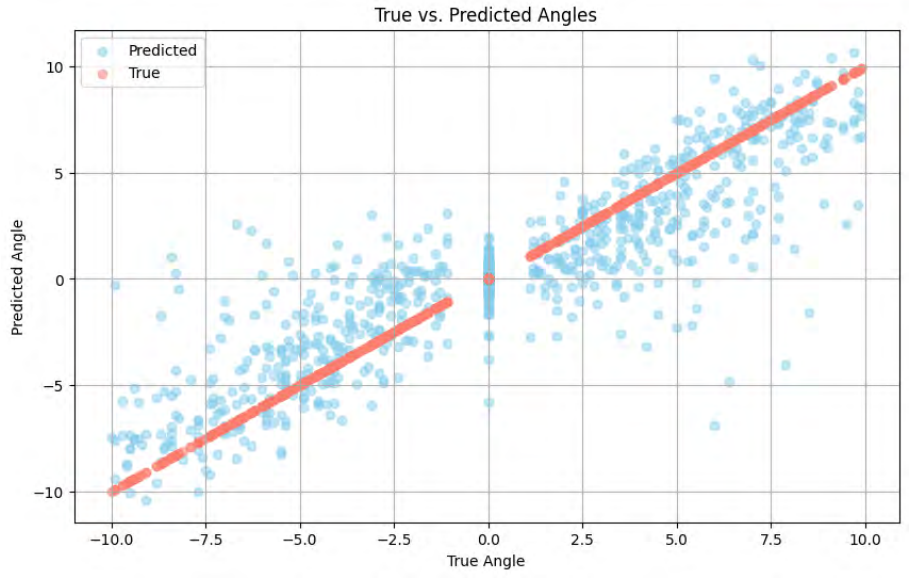

Figure 7.25: Truth VS Predicted Angle

Even after having larger validation loss value than ResNet50 V2, this model seems to perform better. The prediction vs truth graph shows that the predictions made by the model is not perfect but better than the two models mentioned above

Figure 7.26: Sample Test Result of ConvNeXt Model

7.2.4 EfficientNetB2

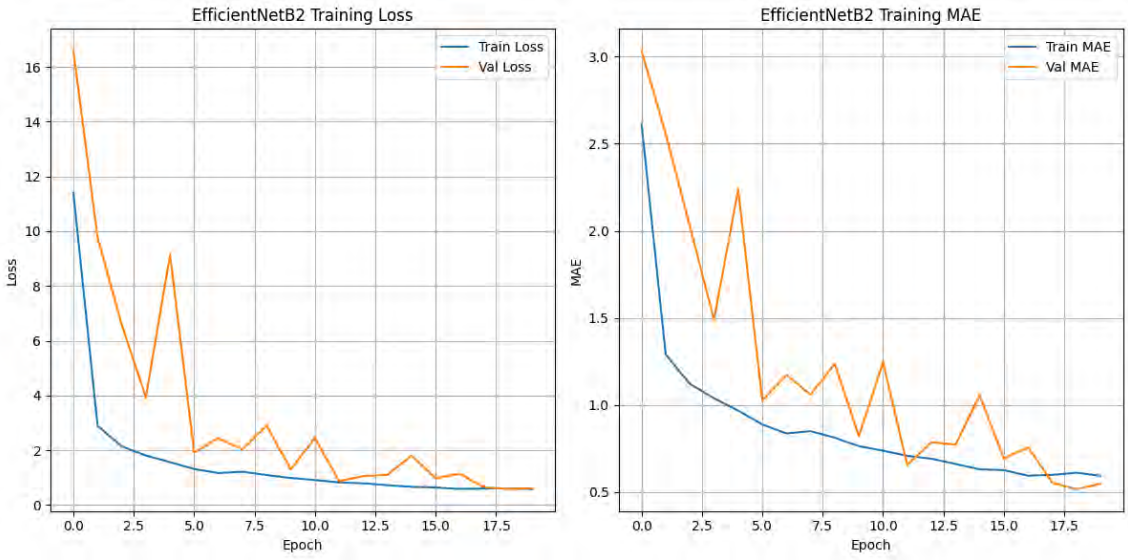

Figure 7.27: Train VS Validation loss

The model was traind for 20 epochs and it shows a gradual decrease of train and validation loss still epoch 3. However, the validation loss seem to increase dramatically in epoch 4 and had a major fall in epoch 5. From this point onwards the validation loss decreased alongside with training loss with subtle spikes in between epochs. At the end of the training process, the loss seems to decrease significanlty and the model seems not to overfit at all as the validation and training loss is nearly the same.

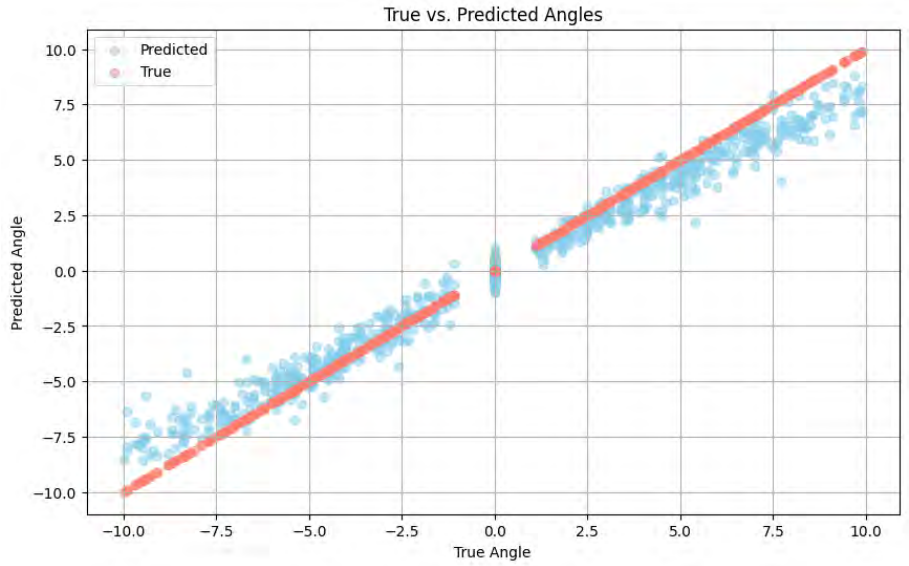

Figure 7.28: Truth VS Predicted Angle

The model perform pretty good as it seems to predict image orientation almost perfectly. The above figure 7.28 also shows that, the model can even predict corner cases with minimum level of error.

Figure 7.29: Test Result of EfficientNetB2

7.2.5 Orientation Detection Models Comparison

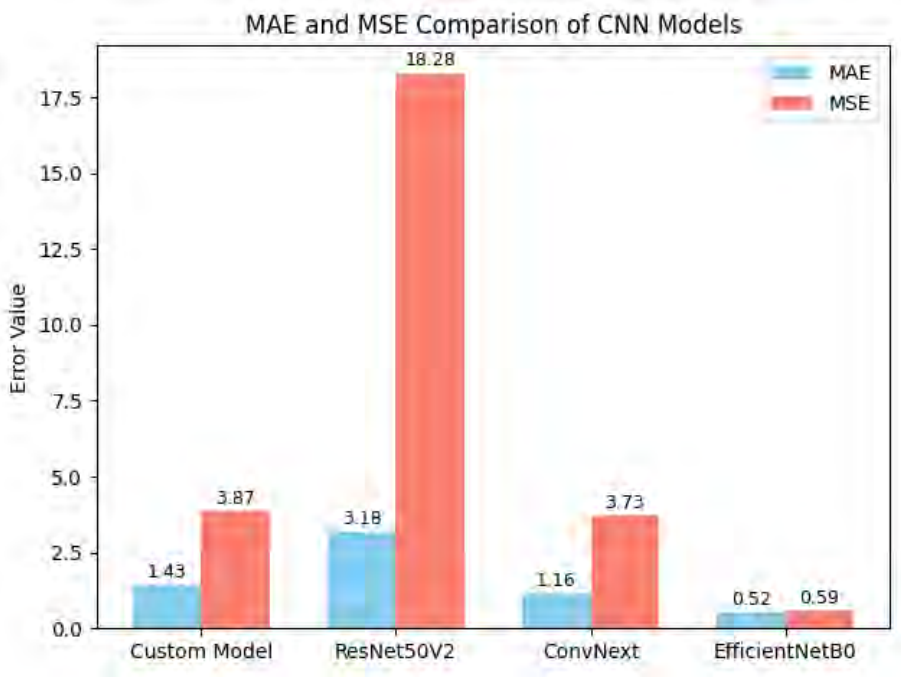

Figure 7.30: MAE and MSE comparison

Different performance characteristics across the analyzed models can be observed by the comparative study of MAE and MSE value across CNN architectures in Figure 7.30. The custom model shows competitive error metric result (MAE: 1.43, MSE:3.87), indicating good architectural choices for the target task. ConvNeXt performs little bit better than out custom model and ResNet50 V2 performs the worst. The bad performance of ResNet50 V2 can be of uneven distribution of training data and the dataset was too small for the model to be trained on. The EfficientNet B0 performs the best, having a negligible amount of MAE and MSE value. The capacity of the model to be trained on a limited dataset explains its good performance.

7.3 Object Subject Detection Model

First we detect all the objects using the YOLOv8 pre-trained model by Ultralytics. Then utilizing the information of object inferences we find the subject using certain logics and conditions (mentioned in chapter-6).

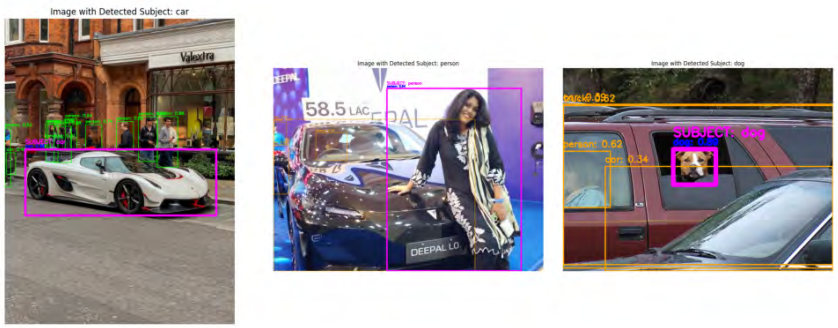

Figure 7.31: Subject Detection

In Figure 7.31 it can be seen that the applied logics work well even in different overlapping conditions. In the leftmost image, the bounding box of the car overlaps with the person behind but does not crosses the threshold of 50% overlap. Which is why the largest object is determined as the subject. However, in both the center and rightmost images, the smaller object which is overlapping the larger one is determined as the subject. However, these conditions and logics does not always

Figure 7.32: Subject Detection

determine the subject properly. In the left image of Figure 7.32, it is evident that the person is the subject, but the system determines the ”handbag” as the subject. In case of the image in the right, it follows Pattern(PAT) composition and it is better not trying to find the subject as the result is quite unpredictable in case of images with Pattern composition.

7.4 Final Suggestion System

The final suggestion system utilizes the best models, one from each of the Composition Detection and Orientation Detection models along with pre-trained YOLOv8 for subject detection. Since multiple models and logics have been used to reduce the overall error, the system works quite efficiently. As confirmed by enthusiastic photographers, the suggestion(s) by the final system is fairly acceptable.

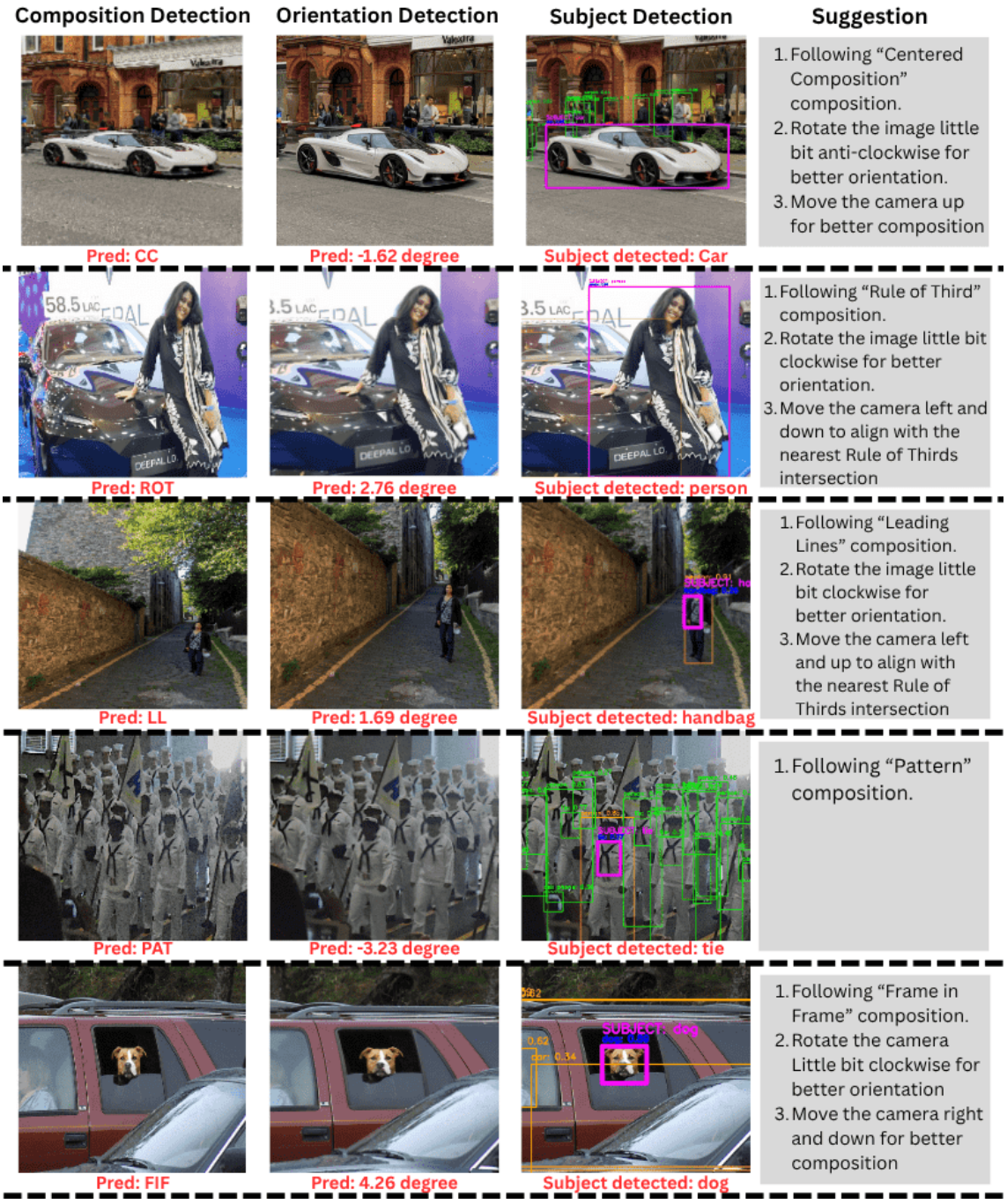

Figure 7.33: Suggestion System

Chapter 8

Conclusion and Future Works

The research aimed to build a model that firstly can detect which composition the photographer is trying to maintain while taking photographs and then suggest zoom, angle or framing to make the composition perfect. Attempts have been made to build the compostion detection model using CNN which shows 74.34% accuracy with the test data at best. Analyzing the dataset we found out some of the images overlap with multiple classes. For better result in image classification, firstly we need to exapand the dataset. Then we need to clean and preprocess the dataset properly for better composition detection accuracy. Secondly, for orientation detection EfficientNet B2 shows significantly better result with a MSE of 0.59. Lastly, YOLOv8 is used from the Ultralytics library for objects detection. From the detected objects logics and conditions have been used to find the image subject thus making it a subject detection model. For the suggestion system we used the result of composition detection, orientation detection and subject detection to determine which composition to focus to make suggestion accordingly. There suggestions are of 2 type, which are Orientation Correction & Camera Movement. For orientation correction, the suggestions are like rotate camera clockwise or anti-clockwise and for camera movement the suggestions look like move camera up or move the camera left or move the camera down the right. This suggestion system will be able to help novice photographers to learn photography in an effective yet easy way.

The suggestion system is initially designed to create a photographic assistant that can assist the photographer in selecting and suggesting according to the composition of a shot given different circumstances. This model has promising features that can be embedded into e-commerce platforms, where it can improve the aesthetics of product photographs and influence sales. Moreover, it can be implemented in automated composition correction and aesthetic editing system.

Bibliography

[1] W. Bares, "A photographic composition assistant for intelligent virtual 3d camera systems," in *International Symposium on Smart Graphics*, Springer, 2006, pp. 172–183.

[2] C. Cavalcanti, H. Gomes, R. Meireles, and W. Guerra, "Towards automating photographic composition of people," in *Proceedings of the IASTED International Conference on Visualization, Imaging, and Image Processing*, Citeseer, 2006, pp. 25–30.

[3] M. Freeman, *The Photographer's Eye: Composition and Design for Better Digital Photos*. Focal Press, 2007, ISBN: 9780240809342. [Online]. Available: https://books.google.com.bd/books?id=3d_lwAEACAAJ.

[4] G. Cao, Y. Zhao, and R. Ni, "Image composition detection using object-based color consistency," in *2008 9th International Conference on Signal Processing*, 2008, pp. 1186–1189. DOI: 10.1109/ICOSP.2008.4697342.

[5] C. Shen, J. Liu, S. Shih, and J. Hong, "Towards intelligent photo composition-automatic detection of unintentional dissection lines in environmental portrait photos," *Expert Systems with Applications*, vol. 36, no. 5, pp. 9024–9030, 2009, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.12.041. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417408009214.

[6] L. Mai, H. Le, Y. Niu, and F. Liu, "Rule of thirds detection from photograph," *2011 IEEE International Symposium on Multimedia*, pp. 91–96, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:6378697.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[8] M. Maleš, A. Heđi, and M. Grgić, "Compositional rule of thirds detection," in *Proceedings ELMAR-2012*, IEEE, 2012, pp. 41–44.

[9] S. Ma, Y. Fan, and C. W. Chen, "Finding your spot: A photography suggestion system for placing human in the scene," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 556–560. DOI: 10.1109/ICIP.2014.7025111.

[10] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, *Microsoft coco: Common objects in context*, 2015. arXiv: 1405.0312 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1405.0312.

[11] K. O'Shea, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Jun. 2015. DOI: 10.48550/arXiv.1506.02640.

[13] A. Schulz, *Architectural Photography, 3rd Edition: Composition, Capture, and Digital Image Processing*. Rocky Nook, 2015, ISBN: 9781681980010. [Online]. Available: https://books.google.com.bd/books?id=eQsUDgAAQBAJ.

[14] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[15] Y. Wang, M. Song, D. Tao, *et al.*, "Where2stand: A human position recommendation system for souvenir photography," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 1, Oct. 2015, ISSN: 2157-6904. DOI: 10.1145/2770879. [Online]. Available: https://doi.org/10.1145/2770879.

[16] H. Zhao, J. Chen, Y. Han, and X. Cao, "Image aesthetics enhancement using composition-based saliency detection," *Multimedia Systems*, vol. 21, pp. 159–168, 2015.

[17] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, *Photo aesthetics ranking network with attributes and content adaptation*, 2016. arXiv: 1606.01621 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1606.01621.

[18] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," Dec. 2016. DOI: 10.48550/arXiv.1612.08242.

[19] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[20] F. Farhat, M. M. Kamani, S. Mishra, and J. Z. Wang, "Intelligent portrait composition assistance: Integrating deep-learned models and photography idea retrieval," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, Mountain View, California, USA: Association for Computing Machinery, 2017, pp. 17–25, ISBN: 9781450354165. DOI: 10.1145/3126686.3126710. [Online]. Available: https://doi.org/10.1145/3126686.3126710.

[21] S. Indolia, A. K. Goswami, S. P. Mishra, and P. Asopa, "Conceptual understanding of convolutional neural network-a deep learning approach," *Procedia computer science*, vol. 132, pp. 679–688, 2018.

[22] J.-T. Lee, H.-U. Kim, C. Lee, and C.-S. Kim, "Photographic composition classification and dominant geometric element detection for outdoor scenes," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 91–105, 2018, ISSN: 1047-3203. DOI: https://doi.org/10.1016/j.jvcir.2018.05.018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320318301147.

[23] J. Redmon and A. Farhadi, *Yolov3: An incremental improvement*, 2018. arXiv: 1804.02767 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1804.02767.

[24] N. Sharma, V. Jain, and A. Mishra, "An analysis of convolutional neural networks for image classification," *Procedia computer science*, vol. 132, pp. 377–384, 2018.

[25] K. Lan and K. Sekiyama, "Autonomous robot photographer system based on aesthetic composition evaluation using yohaku," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 101–106. DOI: 10.1109/SMC.2019.8914467.

[26] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, *Yolov4: Optimal speed and accuracy of object detection*, 2020. arXiv: 2004.10934 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2004.10934.

[27] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, "Fundamental concepts of convolutional neural network," *Recent trends and advances in artificial intelligence and Internet of Things*, pp. 519–567, 2020.

[28] G. Jocher, *Yolov5 by ultralytics*, version 7.0, 2020. DOI: 10.5281/zenodo.3908559. [Online]. Available: https://github.com/ultralytics/yolov5.

[29] Y.-F. Li, C.-K. Yang, and Y.-Z. Chang, "Photo composition with real-time rating," *Sensors*, vol. 20, no. 3, 2020, ISSN: 1424-8220. DOI: 10.3390/s20030582. [Online]. Available: https://www.mdpi.com/1424-8220/20/3/582.

[30] Z. Lin, H. Ye, B. Zhan, and X. Huang, "An efficient network for surface defect detection," *Applied Sciences*, vol. 10, no. 17, p. 6085, 2020.

[31] J. Peng, S. Kang, Z. Ning, *et al.*, "Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from ct imaging," *European radiology*, vol. 30, pp. 413–424, 2020.

[32] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1905.11946.

[33] T. Akama, M. Suzuki, H. Kimura, Y. Asada, K. Suzuki, and H. Takahashi, "Composition presentation based on compositional features for photography assistance," in *International Workshop on Advanced Imaging Technology (IWAIT) 2021*, M. Nakajima, J.-G. Kim, W.-N. Lie, and Q. Kemao, Eds., International Society for Optics and Photonics, vol. 11766, SPIE, 2021, p. 117661I. DOI: 10.1117/12.2590758. [Online]. Available: https://doi.org/10.1117/12.2590758.

[34] Y. Li, "A new method of image classification with photography composition," in *2nd International Conference on Computer Vision, Image, and Deep Learning*, SPIE, vol. 11911, 2021, pp. 110–114.

[35] B. Zhang, L. Niu, and L. Zhang, "Image composition assessment with saliency-augmented multi-pattern pooling," *ArXiv*, vol. abs/2104.03133, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233168684.

[36] J. Bridge, L. Fu, W. Lin, Y. Xue, G. Y. Lip, and Y. Zheng, "Artificial intelligence to detect abnormal heart rhythm from scanned electrocardiogram tracings," *Journal of Arrhythmia*, vol. 38, no. 3, pp. 425–431, 2022.

[37] C. Li, L. Li, H. Jiang, *et al.*, *Yolov6: A single-stage object detection framework for industrial applications*, 2022. arXiv: 2209.02976 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2209.02976.

[38] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022. DOI: 10.1109/TNNLS.2021.3084827.

[39] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, *A convnet for the 2020s*, 2022. arXiv: 2201.03545 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2201.03545.

[40] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, *Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors*, 2022. arXiv: 2207.02696 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2207.02696.

[41] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics yolov8*, version 8.0.0, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[42] Y. Ke, Y. Wang, K. Wang, F. Qin, J. Guo, and S. Yang, "Image aesthetics assessment using composite features from transformer and cnn," *Multimedia Systems*, vol. 29, no. 5, pp. 2483–2494, 2023.

[43] X. Li, G. Teng, P. An, and H.-y. Yao, "Mt-gan: Toward realistic image composition based on spatial features," *EURASIP Journal on Advances in Signal Processing*, vol. 2023, no. 1, p. 46, 2023.

[44] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Deep rotation correction without angle prior," *IEEE Transactions on Image Processing*, vol. 32, pp. 2879–2888, 2023, ISSN: 1941-0042. DOI: 10.1109/tip.2023.3275869. [Online]. Available: http://dx.doi.org/10.1109/TIP.2023.3275869.

[45] B. Wang, H. Si, H. Fu, *et al.*, "Content-aware image resizing technology based on composition detection and composition rules," *Electronics*, vol. 12, no. 14, p. 3096, 2023.

[46] D. Yu, H. Fu, Y. Song, W. Xie, and X. Zhijie, "Deep transfer learning rolling bearing fault diagnosis method based on convolutional neural network feature fusion," *Measurement Science and Technology*, vol. 35, Oct. 2023. DOI: 10.1088/1361-6501/acfe31.

[47] P. Yuan, Z. Han, and X. Zhao, "Integrating the edge intelligence technology into image composition: A case study," *Peer-to-Peer Networking and Applications*, vol. 16, no. 4, pp. 1641–1651, 2023.

[48] A. Elhanashi, P. Dini, S. Saponara, and Z. Qinghe, "Telestroke: Real-time stroke detection with federated learning and yolov8 on edge devices," *Journal of Real-Time Image Processing*, vol. 21, Jun. 2024. DOI: 10.1007/s11554-024-01500-1.

[49] R. Khanam and M. Hussain, *Yolov11: An overview of the key architectural enhancements*, 2024. arXiv: 2410.17725 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2410.17725.

[50] O. Limoyo, J. Li, D. Rivkin, J. Kelly, and G. Dudek, "Photobot: Reference-guided interactive photography via natural language," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 2479–2486. DOI: 10.1109/IROS58592.2024.10801790.

[51] J. Riza, S. Barman, S. Ridita, Z. Mahmud, and A. Bhattacharya, "Phytocare : A hybrid approach for identifying rice, potato and corn diseases," Ph.D. dissertation, Feb. 2024.

[52] N. Sheng, Y. Ke, S. Yang, Y. Yang, and L. Chen, "View adjustment: Helping users improve photographic composition," *Multimedia Systems*, vol. 30, no. 5, p. 293, 2024.

[53] A. Wang, H. Chen, L. Liu, *et al.*, *Yolov10: Real-time end-to-end object detection*, 2024. arXiv: 2405.14458 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2405.14458.

[54] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, *Yolov9: Learning what you want to learn using programmable gradient information*, 2024. arXiv: 2402.13616 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2402.13616.

[55] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, p. 99, 2024.

[56] Freepik, *Stock photos to download / freepik*, en. [Online]. Available: https://www.freepik.com.

[57] Pexels, *Free stock photos, royalty free stock images copyright free pictures · pexels*. [Online]. Available: https://www.pexels.com/.

[58] Unsplash, *Beautiful free images pictures / unsplash*. [Online]. Available: https://unsplash.com/.