

Understanding the Geometry of Photographic images using Deep Learning

Sarvani Chakrabarty

A Dissertation

Presented to the University of Dublin, Trinity College
in partial fulfilment of the requirements for the degree of

**Master of Science in Computer Science
(Intelligent Systems)**

Supervisor: Aljosa Smolic, Koustav Ghosal

September 2020

Declaration

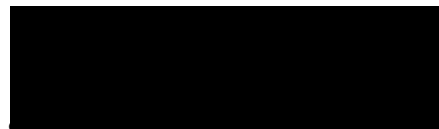
I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.


Sarvani Chakrabarty

September 6, 2020

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

A black rectangular box redacting the signature of the undersigned.

Sarvam Chakrabarty

September 6, 2020

Acknowledgments

I would like to thank my supervisor, Dr. Aljosa Smolic, for his support, feedback and empathy throughout the project. I would also like to thank Koustav Ghosal for his advice on research directions and weekly meetings throughout the duration of the project. Their guidance leads to the final success of this thesis.

I would also like to thank my parents, Dr. Bandana Chakrabarty and Mr. Raj Chakrabarty, for giving valuable suggestions on improving the thesis in terms of literature review conducted and the process of carrying out a research. Their immovable faith in me encourages and motivates me to not give up.

I would like to specially thank Gail Weadick for being so efficient in preparing the letter for the Embassy which made it possible for me to come back to Ireland from India during the pandemic lockdown and also to Michael and Niall from the SCSS helpdesk for putting up with my incessant questions about getting things to run on the remote desktop.

SARVANI CHAKRABARTY

*University of Dublin, Trinity College
September 2020*

Understanding the Geometry of Photographic images using Deep Learning

Sarvani Chakrabarty, Master of Science in Computer Science
University of Dublin, Trinity College, 2020

Supervisor: Aljosa Smolic, Koustav Ghosal

Photographs are composed of different attributes and styles which play a significant role in the way they are viewed. These attributes contribute towards the overall aesthetics of photographs. While a considerable amount of research has been carried out in understanding the appearances and texture of photographs using deep learning, the work done for identifying the geometry of photographs using deep learning becomes extremely limited. This dissertation is a step towards exploring different neural network architectures and determining their performances in detection of geometric attributes in photographic images. The dissertation presents a novel dataset, Geo-Style, which has 12,000 images taken from Flickr and annotated with 7 geometric style labels: Architecture, Lines, Frames, Repetitions, Rule of Thirds, Silhouettes and Symmetry. Neural network architectures are used for style classification and the results achieved show commendable understanding of geometric style by the selected networks. Additionally, the dissertation demonstrates explanations with the help of confusion matrix and feature map visualizations to justify the results achieved by the neural networks. The dissertation finally concludes by conducting a user study to provide a human baseline for the evaluation of Geo-Style dataset.

Contents

Acknowledgments	iii
Abstract	iv
List of Tables	vii
List of Figures	viii
Chapter 1 Introduction	1
1.1 Overview	1
1.2 Motivation	2
1.3 Dissertation Structure	4
Chapter 2 Background Research	5
2.1 Aesthetics in Photography	5
2.2 Datasets for style detection in images	7
2.2.1 AVA Dataset	7
2.2.2 Flickr-Style Dataset	8
2.3 Data Augmentation techniques	9
2.4 Approaches for style and object detection in images	10
2.4.1 Saliency based approach	10
2.4.2 Bounding box approach	12
2.4.3 Geometric approach	15
2.5 Evaluation/Visualization of algorithm performances in Style Detection	16
2.5.1 Evaluation metrics	16
2.5.2 Visualization techniques	17

Chapter 3	Methodology	20
3.1	Geometric properties of Images	20
3.1.1	Architecture	20
3.1.2	Frames	21
3.1.3	Lines	21
3.1.4	Repetitions	22
3.1.5	Rule of Thirds	22
3.1.6	Silhouettes	23
3.1.7	Symmetry	23
3.2	Geometry-Style Dataset	24
3.3	Implementation	25
3.3.1	Data Augmentation	25
3.3.2	Classification models	27
3.3.3	Experiments	28
Chapter 4	Results	30
4.1	Evaluation Metrics	30
4.2	Results	33
4.2.1	Mean Average Precision	33
4.2.2	Per-class precision scores	33
4.2.3	Confusion Matrix	35
4.2.4	Visualization of feature maps	36
4.2.5	GUI results	38
4.2.6	User Study	38
4.3	Discussion	39
Chapter 5	Conclusion	42
5.1	Main contributions	43
5.2	Future work	43
Bibliography		45

List of Tables

3.1	ImageNet 1-crop error rates (224x224) of selected networks. Source: https://pytorch.org/docs/stable/torchvision/models.html . . .	27
4.1	Category wise performance of models with center crop	34
4.2	Category wise performance of models with Random Crop	34

List of Figures

1.1	A sense of Symmetry: M. Gustave’s speech in jail, The Grand Budapest Hotel	3
2.1	Examples of the AVA dataset	8
2.2	Examples of Flickr-Style Dataset	9
2.3	Traditional Augmentation techniques	10
2.4	Example of the Rule of Thirds	11
2.5	Double columned CNN architecture, A Geometry-Sensitive Approach for Photographic Style Classification	12
2.6	The efficiency of ConvNets for detection, OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks	13
2.7	Examples of bounding boxes produced by the regression network, OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks	14
2.8	The YOLO Model: The YOLO model treats detection as a regression problem. It divides the image into a grid of $A \times A$. It predicts N bounding boxes, their confidence and class probabilities C for each grid cell. These predictions are encoded as an $A \times A \times (N \times 5 + C)$ tensor.	14
2.9	Pipeline of Deep Hough Transform	15
2.10	Grad-CAM overview	18
3.1	Examples of Architecture. Source: https://www.flickr.com/	21
3.2	Examples of Frames. Source: https://www.flickr.com/	21
3.3	Examples of Lines. Source: https://www.flickr.com/	22
3.4	Examples of Repetitions. Source: https://www.flickr.com/	22

3.5	Examples of Rule of Thirds. Source: https://www.flickr.com/	23
3.6	Examples of Silhouettes. Source: https://www.flickr.com/	23
3.7	Examples of Symmetry. Source: https://www.flickr.com/	24
3.8	Examples of Geo-Style Dataset : Row 1: Architecture, Row 2: Frames, Row 3: Lines, Row 4: Repetitions, Row 5: Rule of Thirds, Row 6: Silhouettes, Row 7: Symmetry	26
3.9	High level visual representation of the pipeline	28
4.1	Screenshots of the user study. The user can select multiple options of geometric styles that they think are most relevant to the given image. The geometric style with the highest number of votes for each image is considered for evaluation.	32
4.2	Distribution of subjects in user study with respect to their familiarity with geometry of photographs (Figure a) and age (Figure b).	32
4.3	Mean Average Precision: ResNet152 with Center Crop gives the highest MAP followed by DenseNet161 Center Crop	33
4.4	Confusion Matrices showing performances of each neural network . . .	35
4.5	Coarse localization feature maps of the regions in the image that are ‘important’ for predictions from the models	37
4.6	GUI snapshots of images. ResNet152 center crop is used to evaluate these images since it has achieved the highest MAP. If we see the probability scores, the network quite correctly detects the major geometric style present in the image	38
4.7	User study: Confusion Matrices of top three performing models’ predictions with human baseLines.	39

Chapter 1

Introduction

1.1 Overview

Photography is a visual art form and aesthetics are one of its essential principles which encompass visual beauty. Aesthetics has a major impact on an image's character and determines the response that it will get from viewers. Every photographic image is characterized by one or more unique photographic styles or aesthetics. These styles can be broadly classified into appearance-based and geometry-based [1]. Appearance-based photographic style helps in determining the texture and appearance of subjects in an image such as grainy texture, motion blur, complementary colours and soft focus to name a few. Geometry based photographic style, on the other hand, are focused more on the arrangement and shape of subjects in an image. A few examples of geometry-based styles are the rule of thirds, symmetry and frames.

This dissertation focuses on exploring the different geometric properties which are present in photographic images and creating a novel dataset which incorporates these properties. The dissertation further involves exploring the ability of CNNs to detect these properties and comparing the performances of different CNN architectures in determining the same. There is also some research carried out on data augmentation techniques in this dissertation, as they have a significant amount of impact on the performances of CNNs [2].

Seven different types of geometric styles are defined in this dissertation and a new dataset of photographs selected from "Flickr" and annotated with these style labels

is created. The dataset is split in an 80-20% fashion for training and validation combined and testing respectively. Pretrained models of DenseNet161, ResNet152, VGG16, WideResNet-101 and GoogLeNet are loaded and fine tuned by training them on this specific data. They are then further applied to perform the style detection task. The performances of these neural networks are evaluated using Mean Average Precision. Feature maps are also visualized for the last layer of every selected network to have a sense of explainability as to which features of an image are being captured by the neural networks before detecting the geometric property of the image. The dissertation also conducts a user study in order to prove a human baseline for the evaluation of the novel dataset. Since the interpretation of photographic styles varies from person to person, a comparison is carried out between the predictions of the neural networks with human observers.

1.2 Motivation

The style of an image plays a vital role in the way it is viewed, and its popularity in the field of computer vision is steadily increasing. Texture and Appearances are fundamental characteristics of images. Texture representation is one of the challenging and important problems in pattern recognition and Computer Vision which has attracted extensive research attention over the years.

Although very recognizable to human observers, aesthetic and visual style is a difficult concept to define rigorously. Ever since AlexNet [3] won the first the ILSVRC Competition in 2012 [4], Deep Convolutional Neural Networks (CNNs) have shown a remarkable evolution with impressive performances in computer vision tasks. Datasets such as AVA [5] and Flickr-Style [6] as well as ImageNet [7] and COCO [8] are especially curated for aesthetics and object detection respectively in images. However, while CNNs do a commendable job in learning texture and appearance-based features in an image, its understanding of the geometric aspects of images become limited. Geometric features are the image features which are constructed by a set of geometric elements like points, lines or curves. They also include several photographic composition dependencies such as spiral or circular framing and The Rule of Thirds. The performance of CNNs gets restricted while detecting these features are due to two factors. First, the data augmentation techniques such as cropping, warping etc. distort

the photograph’s composition, thereby affecting the performance [1]. Second, CNN features are translation-invariant in principle whereas some geometric attributes such as the Rule of Thirds are position-dependent and appearance-invariant [1].

There is a general rise in the interest in photography due to the growth of camera technology, whether it is a digital one or smartphone. Proper use of geometry in cinematography can also create power impacts on viewer. A notable geometric shot is of that in movie ”The Grand Budapest Hotel” when M. Gustave recites a letter to his colleagues in jail. In this cutaway, M. Gustave is framed at the centre of the shot, with prisoners and guards arrayed evenly on either side, all facing the camera (Figure 1.1). This shot provokes laughter because of the absurdity between seriousness and awkward composition within the exaggeratedly poetic context [9].



Figure 1.1: A sense of Symmetry: M. Gustave’s speech in jail, The Grand Budapest Hotel

A few simple adjustments to the set and the camera’s position can make a dramatic difference to the perceived depth and complexity of the scene and hence the problem is worth exploring. The dissertation’s goal is to examine and evaluate performances of different CNNs in detecting geometric styles of images by curating a geometry specific dataset and experimenting with different data augmentation techniques. A comparative analysis is carried out between these networks by calculating their Mean Average Precision as the evaluation metric to validate the best performing model.

1.3 Dissertation Structure

The thesis is organised into five chapters, starting with an introduction to this research and its motivation in Chapter 1. Then it is followed by Chapter 2 which is a literature review on the background of the performances of different deep neural networks, development of aesthetics related datasets, different types of geometric styles present in an image and the art of data augmentation strategies. The chapter also also talks about the evaluation metrics used to evaluate the networks and ways to visualize their feature maps. In Chapter 3, the design of the proposed pipeline is explained along with the reasons for choosing the selected neural networks. It also presents the process of curating a geometry-specific dataset, the experimentation details and different models which were trained in this project. Chapter 4 gives a summary of the evaluation results from multiple perspectives - Mean Average Precision and per class precision scores to evaluate the overall performance of networks and their performances in each geometric style respectively. The dissertation finally concludes with Chapter 5 summarizing the main contributions and providing possible avenues for future work.

Chapter 2

Background Research

This chapter mainly presents a literature review on the development of photographic style and object detection techniques. It deep dives into the different aesthetic properties of images. It also introduces the different datasets which are present online curated specifically for object detection tasks. The chapter covers different ways in which evaluations can be carried out and the effectiveness of the state of the art data augmentation strategies used in image classification tasks during the past few years. Finally, we look into the techniques of feature map visualizations to help provide explainability.

2.1 Aesthetics in Photography

The aesthetic quality of a photographic image is highly influenced by its composition. Composition is nothing but the amalgam of photographic styles and attributes that a viewer perceives towards the essence of an image. The paper [1] has done an objective analysis to establish a broad categorization of photographic styles. These styles can be classified into local or appearance-based (focus, image grain, motion blur, etc.) and geometry-based (aspect ratio, rule of thirds, framing, etc.) While appearance-based styles refer to the texture of images, geometry-based style include arrangements and shapes of subjects in an image. There has been some research done in the area of appearance-based photographic styles. The datasets of AVA [5] and Flickr-Style [6] have introduced several photographic styles that are present in an image. They have collected pictures which are annotated with appearance-based photographic styles such

as Duotones, HDR, Long Exposure etc.

A relation exists between the composition of a photograph and its subject. Similar subjects are typically photographed in a similar style. The paper [10] investigates the use of photographic style for category-level image classification on the assumption that images within a category share a similar style defined by several attributes. The paper experiments with various photographic styles such as colorfulness, lighting, depth of field, viewpoint, rule of thirds and saliency. The experiments suggest that grouping salient points and global composition benefit the image classification tasks and improve the existing state-of-the-art.

In [11], the author talks about the different image principles that modern images follow such as high contrast (in use of tones, colors, font sizes, size and types of shapes, etc.), limited color palettes and simple geometric shapes. Geometric shapes might include thick lines as design elements, large empty white spaces, asymmetrical as well as symmetrical composition, strong visual rhythm created via repeating elements, parallel lines, etc. or parallel projection that results in parallel rather than converging lines (and repeating parallel lines create visual rhythm.)

The effect that pictures have on viewers is highly subjective. Every picture is judged differently by each person. There is no unanimously agreed standard for measuring aesthetic value of a photo. Datta et al in [12] have attempted to explore the relationship between emotions which pictures arouse in people, and their low-level content. Certain visual features are identified and extracted based on the intuition that they can discriminate between aesthetically pleasing and displeasing images. Automated classifiers are then built using support vector machines and classification trees to rate an image highly or lowly. The work done in this paper identifies certain features which are relevant to photographic quality while building the classifiers. The candidate features are exposure of light and colorfulness, saturation, hue, the rule of thirds, wavelet-based texture (grainy or smooth), size and aspect ratio, depth of field and shape convexity. This work is a significant step towards the highly challenging task of understanding the correlation of human emotions and images they see by a computational approach.

2.2 Datasets for style detection in images

Images which are created deliberately convey meanings and visual style often holds a significant amount of gravity for the image meanings. Realistic, diverse and challenging datasets have been curated over the years which help train convolutional neural networks to correctly identify the meaning that the image tries to convey in terms of aesthetics and objects. This section explores the existing datasets which have inspired the development of a novel dataset in this dissertation which is specific to geometric properties of photographic images.

2.2.1 AVA Dataset

With the motivation of keeping the process of organizing and navigating through the ever-expanding volume of visual content available online by aesthetic preferences, Murray et al in [5] has introduced a large-scale database in order to conduct Aesthetic Visual Analysis (AVA). The AVA database consists of over 250,000 images which were obtained from www.dpchallenge.com. The AVA dataset also consists of a rich variety of meta-data including a large number of aesthetic scores for each image, semantic labels for over 60 categories as well as labels related to photographic style. In the Dpchallenge forum, photographers while critiquing an image not only comments on how much they like it, but also explains why they like or dislike it. The AVA database consists of style annotations and using that, it tries to replicate the mentioned qualitative assessment of aesthetic properties in images. The authors use 14 different style annotations (derived from the titles and descriptions of the photographic challenges to which photos were submitted) which mount upto 14,079 images that they use for this qualitative assessment of classifying the photographic style of an image and to judge if the images are consistent with their respective styles. Those styles are: Complementary colors, Duotones, HRD, Image Grain, Light on White, Long Exposure, Macro, Motion Blur, Negative Image, Rule of Thirds, Shallow DOF, Silhouettes, Soft Focus and Vanishing Point. A few snapshots of the dataset is given in the Figure 2.1. For the style labels, the publishers of the dataset provide a train/test split, where training images have only one label, but test images may be multi-labeled. The AVA database design has helped in providing a large-scale benchmark and training resource. It also provided insights into aesthetic preferences and proved how richer and especially larger

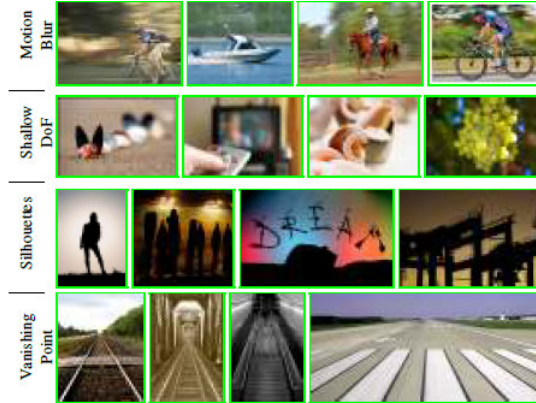


Figure 2.1: Examples of the AVA dataset

datasets could help to improve existing applications and enable new ones. A lot of experiments has been carried out, since, on this rich dataset to understand the different aesthetic properties of an image.

2.2.2 Flickr-Style Dataset

Style plays a critical role in the manmade imagery that we experience daily and another dataset which has made significant progress in defining the problem of understanding photographic style is the Flickr-Style dataset [6]. Although Flickr users often provide free-form tags for their uploaded images, the tags tend to be quite unreliable. The paper does a research on how Flickr groups, which are community curated collections of visual concepts, give more precise results. For example, the Flickr Group “Geometry Beauty” is described, in part, as “Circles, triangles, rectangles, symmetric objects, repeated patterns”. Using images of these communities, 20 visual styles were identifies for the Flickr-Style dataset such as Macro, Bokeh, Depth-of-field, Long Exposure, Hazy, Sunny, Pastel, Bright etc. A few examples of the dataset are shown in Figure 2.2 Working with a large database definitely paves the way of encountering images which are not labeled. Mechanical Turk Evaluation is used to tackle such problems. The paper further performs a thorough evaluation of the dataset by using performance metrics such as Average Precision and per-class accuracy on subsets of data. Confusion Matrices are also plotted to understand which classes have more confusions than the others such as Depth of Field vs Macro and Vintage vs Melancholy. This paper helped

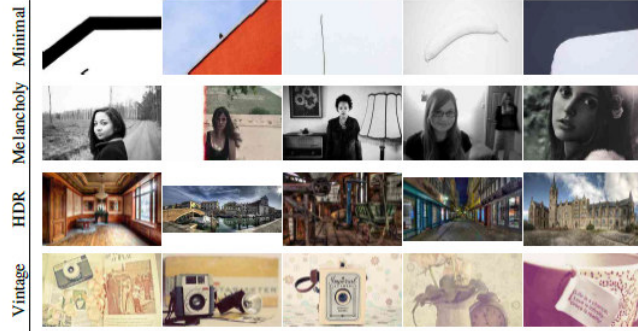


Figure 2.2: Examples of Flickr-Style Dataset

in giving insights into the possible evaluation metrics that can be used for the dataset developed in this dissertation.

2.3 Data Augmentation techniques

Deep learning requires large volumes of training data to train models. Often, there is insufficient training data available and augmentation is used to expand the dataset. This section introduces a few data augmentation strategies for image classifications that have been used in recent years.

Historically, only simple forms of augmentation, such as cropping and horizontal flips, were used. In [2], the authors investigate the effectiveness of different data augmentation strategies such as rotation, skew tilt, shear, random erase, random distortion and Gaussian distortion. It was found that combination of the original dataset with multiple single augmentations increases accuracy by +2.36%. Random distortion and Gaussian distortion are found to be the worst augmentation techniques whereas random erase performs the best.

The images of ImageNet [3] were subjected to different transformation strategies to avoid overfitting such as flipping, cropping and rescaling. In this paper, the authors also apply two kinds of augmentations to enhance the model performance: horizontal reflection and application of PCA on the RGB values of image pixels. Using these augmentation strategies have proven to improve performances of models, thereby demonstrating their importance in training a neural network. Figure 2.3 shows how applying different augmentation techniques on an image change the subject matter of the image

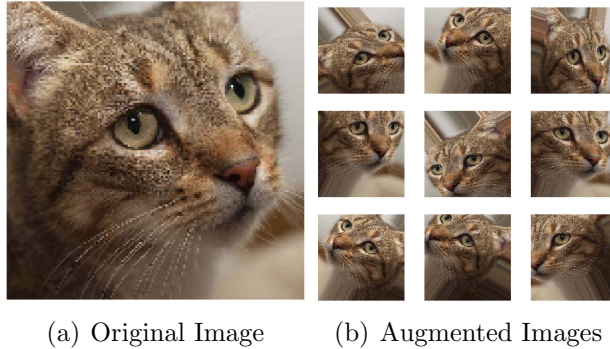


Figure 2.3: Traditional Augmentation techniques

(in this case, the cat).

2.4 Approaches for style and object detection in images

Convolutional Neural Networks have shown a commendable evolution with impressive performances in the area of Computer vision. However, since very limited work has been carried out in the field of studying geometric properties images, we aim to seek inspiration from the existing different style and object detection techniques. This section explores a variety of architectures/approaches that has been implemented and holds a record of performing well in their intended tasks.

2.4.1 Saliency based approach

Rule of Thirds is one of most popular geometric feature of a photographic image. According to the rule of thirds, important objects are placed along either the imagery thirds lines or around their intersections of the image and that often leads to highly aesthetic photos [13] as can be seen in Figure 2.4. Long Mai, et al in [13] put forward a methodology to detect the rule of thirds from a photo using Saliency based feature design. Since rule of thirds require the knowledge of important content location, this paper explores methods in generic objectness and saliency analysis in order to infer crucial object locations and designed features accordingly. Their method utilizes three algorithms for the estimation of saliency: GBVS, FT and GC and OBJ.



Figure 2.4: Example of the Rule of Thirds

Each of these methods take an image as the input and provide as an output a map indicating the saliency value at each pixel. The paper experiments with Saliency Map Centroid, Saliency around third lines and intersection and a simple war saliency map. A photo that respects the rule of thirds usually has more saliency around the thirds lines and their intersections. The method then applies a range of classic machine learning techniques for the rule of thirds detection, including the Naive Bayesian Classifier, Support Vector Machine, Adaboost, and K-Nearest Neighbor method. It is observed that raw saliency map gives the best performance in detecting the rule of thirds from a photo. The experiments conducted in the paper show a promising result although future improvement is certainly needed.

Another saliency based approach is presented in [1] where a double columned CNN is introduced with one column having general RGB features as input and the other column having saliency maps as the input (Figure 2.5). This novel input representation is geometry-sensitive, position-cognizant and appearance-invariant. The overall architecture of the network used in this paper consists of three main blocks - the saliency detector, the double columned feature extractor and the classifier. The authors use CNN-LSTM (long and short term memory) framework for saliency detection. The feature extractor consists of two parallel and independent columns. One column is used for the saliency map. The features from the RGB channel are fed to the other column. These features are computed using pre-trained CNN model DenseNet161 from PyTorch which is fine tuned on the datasets used for this research. Feature maps from the two columns are then concatenated and fused together using a fully connected layer which replaces the last layer of DenseNet. The authors use both AVA and Flickr-Style datasets and compare the performances of the network using Mean Average Precision

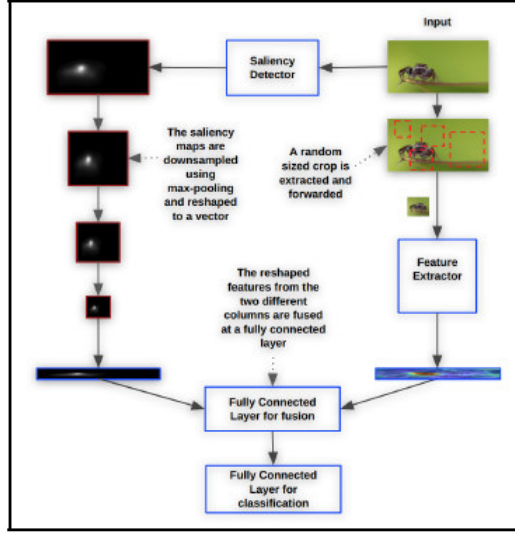


Figure 2.5: Double columned CNN architecture, A Geometry-Sensitive Approach for Photographic Style Classification

(MAP). The authors also calculate per class precision (PCP) scores to evaluate the performance of the network for each class of attributes of the dataset. It was observed that this architecture with two columns elevate the state-of-art and perform better than just passing RGB features through a CNN such as DenseNet161 and ResNet152.

2.4.2 Bounding box approach

Sermanet et al in [14] has presented a multi-scale, sliding window approach within a ConvNet that can be used for three computer vision tasks: classification, localization and detection, wherein each task is a sub-task of the next. The paper states that while most sliding window approaches compute a complete pipeline for each window of the input image one at a time, ConvNets prove to be efficient when applied in a sliding fashion since they share computations common to overlapping regions. The complete network is diagrammed in Figure 2.6 where the top part shows ConvNet producing only a single spatial output during training. But, when applied at a test time over a larger image, the authors simply extend the output of each layer to enclose the new image size by applying each convolution over the extent of the complete image. Thus, a spatial output map (eg. 2x2) is produced as shown in the bottom of the figure. In the

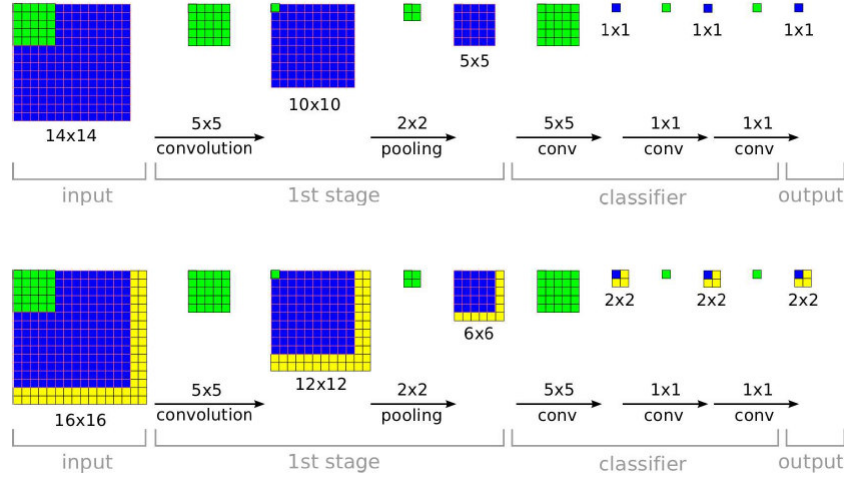


Figure 2.6: The efficiency of ConvNets for detection, OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks

classification task of this challenge, each image is single labeled in correspondence to the main object in the image. Since the images might have multiple unlabeled objects, a maximum of 5 guesses per image are granted. In the localization task, in addition to the guesses, a bounding box for the predicted object is returned with each guess. To evaluate the correctness, the predicted box must match the ground truth by at least 50% (using the PASCAL criterion of union over intersection), as well as be labeled with the correct class (i.e. each prediction is a label and bounding box that are associated together). The authors replace the classifier layer of the classification-trained network by a regression network which is trained to predict the object bounding boxes at each spatial location and scale. To generate object bounding box predictions, the classifier and regressor networks are simultaneously run across all locations and scales. Due to these sharing the same feature extraction layers, only the final regression layers need to be recomputed after computing the classification network. The output of the final softmax layer for a class c at each location provides a score of confidence that an object of class c is present (though not necessarily fully contained) in the corresponding field of view. Thus we can assign a confidence to each bounding box (Figure 2.7). This approach is applied to the ILSVRC 2013 datasets and as of that time, ranked the 4th in classification and 1st in both localization and detection.

While OverFeat [14] performs window detection efficiently, it still is a disjoint sys-



Figure 2.7: Examples of bounding boxes produced by the regression network, OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks

tem. Localization is optimized, however, that doesn't hold true for detection performance. These limitations are overcome in the paper [15] where the authors introduce a more agile approach called "You Only Look Once" (YOLO). Here, object detection is considered to be a regression problem and thereby the bounding boxes are spatially separated from the associated class probabilities.

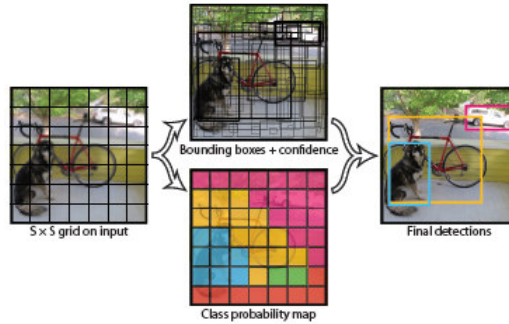


Figure 2.8: The YOLO Model: The YOLO model treats detection as a regression problem. It divides the image into a grid of $A \times A$. It predicts N bounding boxes, their confidence and class probabilities C for each grid cell. These predictions are encoded as an $A \times A \times (N \times 5 + C)$ tensor.

YOLO uses only a single neural network that predicts class probabilities as well as bounding boxes in one evaluation from full images directly (Figure 2.8). It detects objects in an image with a single pass through a neural network. To summarize, in YOLO, the image is first divided into a grid and for each grid cells, two bounding

boxes are predicted. The bounding box which performs the best across the complete image is then identified. YOLO is super fast and even works on videos since it only requires a single pass through a network. The limitation of this process is that the accuracy of the model is a little compromised when it comes to generalizing objects in new or unusual aspect ratios due to its nature of learning to predict bounding boxes from data.

2.4.3 Geometric approach

Kai Zhao et al in [16] proposes a simple, yet effective method to detect meaningful straight lines (semantic lines) in given images. Prior detection methods take line detections as a special case of object detection. They neglect the inherent characteristics of lines, thereby leading to less efficient and suboptimal results. To tackle this, the classic method of Hough Transform is incorporated into CNNs for straight line detection in natural images. The pipeline consists of the following four principle components: 1) a CNN en-coder that extracts pixel-wise deep representations; 2) the deep Hough transform (DHT) that converts the spatial representations to a parametric space; 3) a line detector that is responsible to detect lines in the parametric space, and 4) a reverse Hough transform (RHT) that converts the detected lines back to image space. This is represented diagrammatically in the Figure 2.9. All these components are unified

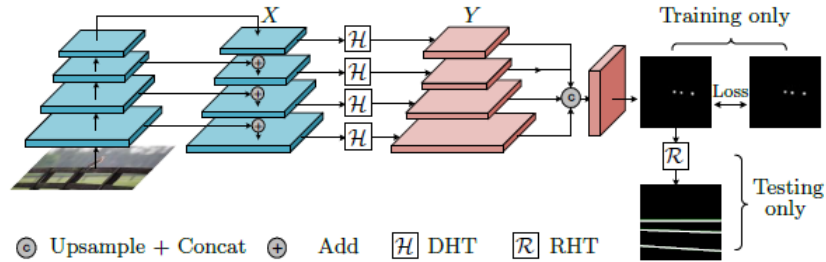


Figure 2.9: Pipeline of Deep Hough Transform

in a framework that performs forward inference and backward training in an end-to-end manner. PyTorch models of ResNet50 and VGGNet16 are further implemented as backbones. A new evaluation metric called EA-Score is proposed for line structures, which considers both Euclidean distance and angular distance between a pair

of lines. Both quantitative and qualitative results revealed that the proposed method outperforms previous arts in terms of both detection quality and speed significantly.

Another paper which researches on object detection from a geometric aspect is the work done by Xie et al in [17] which develops a new edge detection algorithm called holistically-nested edge detection (HED). HED automatically grasps valuable hierarchical portrayals that are important in order to resolve the challenging ambiguity in edge and object boundary detection. This proposed algorithm addresses two significant long standing issues in vision problems: (1) holistic training as well as prediction of images and (2) multi-scale as well as multi-level learning of features. The HED model structure and parameters are initialized by loading a pre-trained trimmed VGGNet. The HED model is evaluated using Average Precision and F- Scores on the BSD500 and NYU Depth datasets. It is observed that the proposed method significantly improves the state-of-the-art and shows promising results in performing image-to-image learning by combining multi scale and level visual responses, even though clear-cut contextual and significant information have not been enforced.

2.5 Evaluation/Visualization of algorithm performances in Style Detection

It is an essential phase of research to evaluate the findings of the process to judge if it improves the state-of-the-art. Visualizing feature maps and confusion matrices give more clarity and explainability on the working of the algorithms since the working of these networks are completely black-box. Several metrics exist for the same and this section takes a look at the metrics used for measuring the performances of image detection systems.

2.5.1 Evaluation metrics

Mean Average Precision has been widely used to measure the performance of models responsible for image classification or object detection tasks. Koustav et al in [1] has used mAP to check the compare the performances of their double columned CNN architecture with pretrained models such as DenseNet161 and ResNet152. Flickr-Style Dataset [6] has also computed mAP to evaluate the performance of classifiers trained on

the dataset in correctly determining the photographic style of images. For evaluating the correctness of the AVA dataset [5] for the 14 style annotations, mAP has been used to evaluate the classification model. Average Precision is basically the area under a precision-recall curve. Long Mai et al in [13] uses this approach to evaluate the performance of saliency maps for rule of thirds detection.

Precision is the ratio of true positives to the sum of true and false positives and it helps in evaluating the ability of a classifier not to label as positive a sample that is negative. In a classification task, precision score of 1.0 for a class means every item labeled as belonging to that class is indeed true. For style detection methods, this metric gives the evaluation of different classifying models' performances at a more granular level. In [1] per class precision scores are calculated to gain deeper insights on the performances of models on individual photographic attributes. The COCO dataset [8] was introduced for detection and segmentation of object found in everyday life. The dataset consists of a vast collection of object instances which are annotated and with the aim to advance research in object detection and segmentation algorithms. The paper uses precision to verify the image annotations which were carried out by using Amazon's Mechanical Turk.

2.5.2 Visualization techniques

In the field of machine learning and specifically the problem of classification, a confusion matrix is a table which allows visualization of the performance of algorithms. It is a great tool to evaluate the behavior and understand the effectiveness of a binary or categorical classifier. A confusion matrix heatmap visualization makes it easy to understand which classes are being confused more often in a classification problem. [6] has used Confusion Matrices to understand the behavior of the classifiers in detecting the different annotated styles of the dataset. This helps the authors to detect the existence of anomalies and surprising errors and give scope to ponder and compute the reasons of the errors. For example, in Flickr-Style dataset [6], there were understandable confusions of Romantic with Pastel and Vintage with Melancholy. However, there also existed surprising sources of mistakes such as Macro with Bright/Energetic. To explain this particular confusion, it was observed that lots of Macro photos contain bright flowers, insects, or birds, often against vibrant greenery. In [1], a double

columned CNN model was built and implemented on the AVA dataset of 14 style annotations. The plotting of a confusion matrix heatmap helped the authors to understand the limitations of their algorithm such as confusion between Long Exposure and Motion Blur, which makes sense, since the capture of both attributes require using a slow shutter speed and occur mostly at night.

Visualizing filters and feature maps of input images to a particular model help explaining the functionality and demonstrated behavior of the neural networks. Grad-CAM (Gradient-weighted Class Activation Mapping) [18] proposes a technique for making Convolutional Neural Network (CNN)-based models more transparent by visualizing the regions of input that are ‘important’ for predictions for these models via visual explanations. It uses the class-specific gradient information flowing into the final convolutional layer of a CNN to produce a coarse localization map of the important regions in the image and is a strict generalization of the Class Activation Mapping (CAM). In CAM, the feature maps of a layer of a CNN are spatially pooled using Global Average Pooling (GAP) and linearly transformed to produce a score, say y , for each class. In Grad-CAM, the gradient of this score, y , is computed with respect to the feature maps. These gradients flowing back are global-average-pooled to obtain weights followed by a Rectified Linear unit and captures the ‘importance’ of feature map for a target class (Figure 2.10) . Grad-CAM heat map is a weighted combination

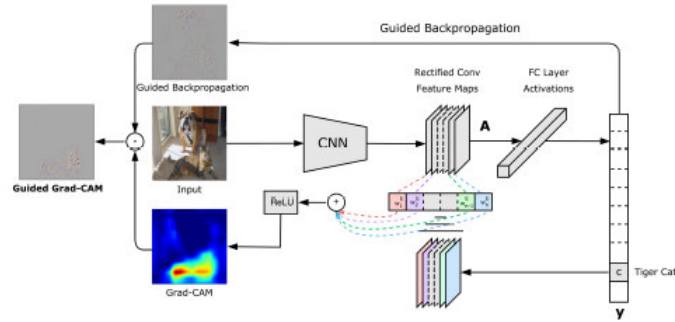


Figure 2.10: Grad-CAM overview

of feature maps, but this is followed by a Rectified Linear unit to prevent localization maps from highlighting more than just the desired class and achieve lower localization performance since we are only interested in the features that have a positive influence on the class of interest, i.e. pixels whose intensity should be increased in order to

increase the score y . This enables us to capture the ‘importance’ of feature map for a target class.

In principle, the pipeline of the dissertation is similar to [1] in the sense that there is an attempt to analyze the performances of CNNs in aesthetic style detection in photographs. However, this work concentrates specifically on the geometric aspects of images. Also, a novel dataset is created in a structure similar to the AVA dataset [5] and using Flickr as the source of images [6]. Traditional data augmentation techniques inspired from [2] are used. One way this dissertation differs from the existing research is, in addition to developing a new dataset for geometric features and evaluating the performances of CNNs in detecting these features, the work also involves exploring the reasons as to why a specific model behaves in a particular way in prediction by using feature map visualizations using [18] as the reference.

Chapter 3

Methodology

This chapter describes the high level design of detecting geometric properties of photographic images. The structure of the chapter starts with introducing the geometric styles that would be evaluated in this dissertation followed by the preparation of a novel dataset pertaining to the selected styles. The chapter further describes the different image classification models used for style detection tasks.

3.1 Geometric properties of Images

The geometry of an image highly influences on how the image is perceived by viewers. Images which display similar subjects are typically photographed in a similar style [10]. Geometry of an image refers to the arrangement and shape of objects in an image. Taking this definition into consideration, seven geometric features have been identified which contribute towards the aesthetics of an image. These features are described in the subsequent sections.

3.1.1 Architecture

Architecture is considered to be the most popular geometric subject in the world of photography [19]. Humans have sought shelter in fabricated structure since ancient history. These structures, over the years, became a testament to ingenuity as well as art. In photography, architecture speaks about to humans' desire for order [19]. Figure 3.1 gives a few examples of Architecture in photography.



Figure 3.1: Examples of Architecture. Source: <https://www.flickr.com/>

3.1.2 Frames

Framing photos is surrounding a subject with different elements to make it stand out and thereby help in grabbing the viewer's attention to the main subject [20]. In addition to directing attention, the use of framing gives a picture context. Foreground elements around the subject, for example, add to the story told by an image. Figure 3.2 gives a few examples of Frames in photography.



Figure 3.2: Examples of Frames. Source: <https://www.flickr.com/>

3.1.3 Lines

Geometric photography and lines are almost inseparable. It's the most critical element of visual art. Lines can be bold, wavy, thin, sharp; there are countless variations [19]. The purpose of lines is to lead the viewer's eyes to the subject of the image [21]. The geometry of line appeals to the way we visualise our surroundings. They define space, momentum and emphasis. Lines can be subject in themselves as well, conveying emotion through form and shape [19]. Figure 3.3 gives a few examples of Architecture in photography.



Figure 3.3: Examples of Lines. Source: <https://www.flickr.com/>

3.1.4 Repetitions

Repetitions help in reinforcing the significance of subjects in images by having multiple presence of them in the same frame. This type of geometric style have greater impact on the viewers as with each repetition, the subject is reinforced [19]. Geometric features of an image often revolve around repetitions since they create a strong foundation for the image to operate on. Figure 3.4 gives a few examples of Repetitions in photography.



Figure 3.4: Examples of Repetitions. Source: <https://www.flickr.com/>

3.1.5 Rule of Thirds

One of the most popular composition rules used by photographers is the rule of thirds. The rule of thirds places important objects along the imagery thirds lines or around their intersections [13]. Among all geometric properties, rule of thirds is the most extensively researched one. Figure 3.5 gives a few examples of Rule of Thirds in photography.



Figure 3.5: Examples of Rule of Thirds. Source: <https://www.flickr.com/>

3.1.6 Silhouettes

Silhouettes is one photographic style which might be considered both as appearance-based and geometry-based. While there has been evaluations on silhouettes in the AVA dataset [5], this photographic attribute could also be included in a dataset dedicated to geometric attributes because in silhouettes, the arrangement is such that the object only has an outline and a featureless interior. Basically, silhouettes show the shape of the subject without any detail and give information about the global shape of scene object, which makes it an interesting geometry style for photographs. Figure 3.6 gives a few examples of Silhouettes in photography.



Figure 3.6: Examples of Silhouettes. Source: <https://www.flickr.com/>

3.1.7 Symmetry

Symmetry can be defined as the geometric property which gives a sense of visual balance to the viewers [19]. Symmetric photographs are visually pleasant to look at as the finer geometric details are evenly distributed throughout the image. It adds an even flow to the image. Figure 3.7 gives a few examples of Symmetry in photography.



Figure 3.7: Examples of Symmetry. Source: <https://www.flickr.com/>

3.2 Geometry-Style Dataset

Building a model such that it carries out classification tasks effectively requires annotated training data. As mentioned in Chapter 2 under Background Research, there are two datasets which are dedicated towards appearance-based photographic style detection in images - AVA and Flickr-Style [5] [6]. Since this dissertation revolves around studying the field of geometry-based photographic style, a novel dataset is created keeping in mind the geometric attributes identified in the section above. The larger the dataset, the more effective the results will be. Hence, a web crawler is built to get images from Flickr, as it has a rich set of photographs annotated by the users.

Geo-Style. Using a web crawler and Flickr API, a database consisting of 12000 annotated images is built. Since the number of geometric properties shortlisted is 7, the dataset of 12000 images is divided into 7 classes, each class corresponding to one geometric property or style. The classes are named as:

- architecture
- frames
- lines
- repetitions
- rule of thirds
- silhouettes
- symmetry

Each class consists of approximately 1700 images. 10,000 of these images are used for training and validation purposes and 2000 images are used for testing.

Unlike the AVA dataset [5], where the testing dataset has multiple labels, the Geo-Style dataset has both training and testing data single labeled to keep things

simple. The possibility of overlap in image classes, for example, lines overlapping with architecture and symmetry, might exist, however, such instances are considered as an unfortunate, albeit, acceptable reality of working with large-scale data. Since the dataset is novel, and the effect that pictures have on viewers is highly subjective, a user study is conducted to cross verify if the images annotated with certain classes make sense or not, the results of which will be covered in Chapter 4. This was inspired by the Mechanical Turk Evaluation used to provide a human baseline for evaluating the Flickr-Style dataset [6]. Example images of the Geo-Style dataset are shown in Figure 3.8.

3.3 Implementation

This section gives an overview of the pipeline followed including specific implementation details of data augmentation techniques and style detection models. The first subsection gives a brief idea on the art of data augmentation exploited for this dissertation. This is followed by a discussion of the classification networks used. Finally, the transformation network and classification networks are combined to form the overall pipeline of the system.

3.3.1 Data Augmentation

The traditional approach of using CNNs for the purpose of natural image classification is to forward a "transformed" version of the input through a series of convolutional, pooling and fully connected layers before finally generating the classification score [1]. Considering the data augmentation techniques mentioned in [2] and [3], the following transformation strategies are applied to and experimented with the images:

- Resize: To resize the input PIL images to a specific size. Here, we have
- Center Crop: To crop the given PIL image at the center.
- Random Crop: To crop the PIL image at a random location.
- Random Horizontal Flip: To horizontally flip the given PIL image randomly with a given probability, in this case we use the default probability parameter which is 0.5.

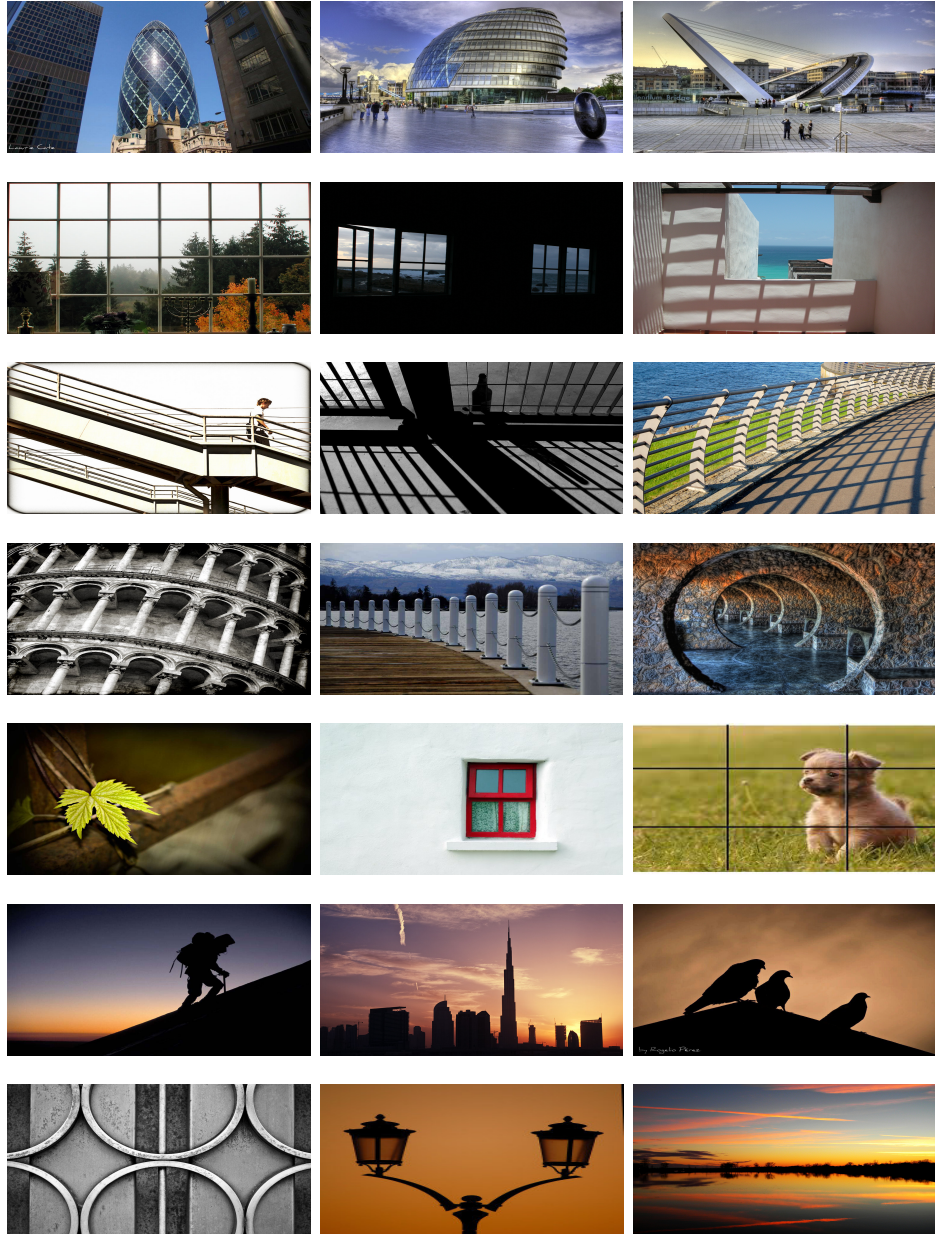


Figure 3.8: Examples of **Geo-Style Dataset**: Row 1: Architecture, Row 2: Frames, Row 3: Lines, Row 4: Repetitions, Row 5: Rule of Thirds, Row 6: Silhouettes, Row 7: Symmetry

- Conversion to Tensor: To convert the PIL image to a tensor, Since the PIL images belong to mode RGB, the tensor is of shape (C x H x W) in the range [0.0, 1.0] [22].
- Normalize: To normalize the tensor image with mean and standard deviation for each channel.

The "Transforms" module of PyTorch [22] is utilized to apply these data augmentations. All the images are scaled to the same size. All images are experimented with Center crop and Random Crop separately along with the other transformations. The performances of neural networks are then evaluated with Center Crop and then with Random Crop, the results of which will be explained in the chapter 4.

3.3.2 Classification models

Five different PyTorch classification models are selected to detect the geoemtric styles of images. They are as follows:

- DenseNet161
- ResNet152
- GoogLeNet
- WideResNet101
- VGG16

These networks were selected due to their superior performances in the ImageNet challenge. A comparative analysis of ImageNet 1-crop error rates (224x224) for these models are shown in Table 3.1

Network	Top-5 error
DenseNet161	6.20
ResNet152	5.94
GoogLeNet	6.67
WideResNet101	5.72
VGG16	9.62

Table 3.1: ImageNet 1-crop error rates (224x224) of selected networks. Source: <https://pytorch.org/docs/stable/torchvision/models.html>

One important thing to consider DenseNet is due to the fact that it differs from the traditional CNNs in the manner that each layer of DenseNet receives as input the concatenated outputs from all previous layers. Wide residual networks are an extension of deep residual networks with decreased depth and increased width [23]. Due to the problem of diminishing feature reuse which often makes training slower in ResNet, WideResNet is claimed to be far superior over their commonly used thin and very deep counterpart [23], which is why the scope of the dissertation also compares the performance of WideResNet101 in addition to ResNet152. The mentioned networks are pre-trained models that are loaded from PyTorch and are finetuned by training them on Geo-Style dataset.

3.3.3 Experiments

Pipeline

The main goal of this dissertation is to compare the performances of the neural networks in detection of geometric properties that exist in a photographic images. To understand the effectiveness of the model's style detection ability, two sets of experiments are conducted: one with center crop and another with random crop in transformations.

For each input image, a set of duplicate images that are resized, cropped (center or random) and horizontally flipped is generated. The output images from the transformation net are fed to a second network called the RGB feature extractor which accepts raw RGB input. The output from the feature extractor is finally fed to the neural networks for style detection purposes. A visual representation of the pipeline using Center crop and ResNet152 is shown in Figure 3.9

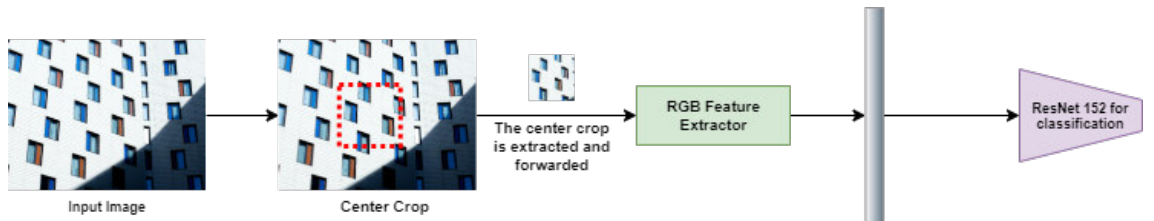


Figure 3.9: High level visual representation of the pipeline

Training

Since the dataset is not so large, the pre-trained models are fine-tuned by training them on the Geo-Style dataset. All the models are coded with PyTorch and Python with Ubuntu 16.04 as the platform. The images in the Geo-Style dataset are scaled to 224x224 and normalized with mean of (0.485, 0.456, 0.406) and standard deviation of (0.229, 0.224, 0.225) for respective channels.

For the code part, the batch size is selected to be 8 with 25 epochs over the training and validation images. The optimizer used is Stochastic Gradient Descent (SGD) and the learning rate is set at 0.001. This dissertation work applies the research done by Koustav et al [1] as the backbone structure and hence the values of these parameters are kept same during the training process. It takes around 5-6 hours on the given machine to complete the training of a network.

Chapter 4

Results

This chapter presents the evaluation metrics that have been employed to compare the performances of the neural networks in detecting the geometric style of images. The results achieved from each of these evaluation metrics are also presented in this chapter along with the inferences that are achieved from the results in the Discussion section.

4.1 Evaluation Metrics

Several performance metrics are considered to evaluate the networks' performances. Confusion Matrices and feature maps are also visualized to get a fair idea about the working of the neural networks in detecting the geometric styles. A graphical user interface (GUI) is built which provides an interactive platform for users to give any image of their choice for style detection. The results of the GUI are explained in more details in section 4.2.5.

Mean Average Precision

To evaluate the performances of the pre-trained and then fine tuned neural networks on Geo-Style dataset, the models are deployed on the test data for style detection. The detection scores are reported in terms of Mean Average Precision (MAP). MAP is the average of per class precision. MAP shows which model has performed the best in detection of geometric styles in photographs.

Per-class precision

Per-class precision scores are further calculated for evaluation of each model’s performance for each of the geometric classes in the testing dataset of Geo-Style.

Confusion Matrix

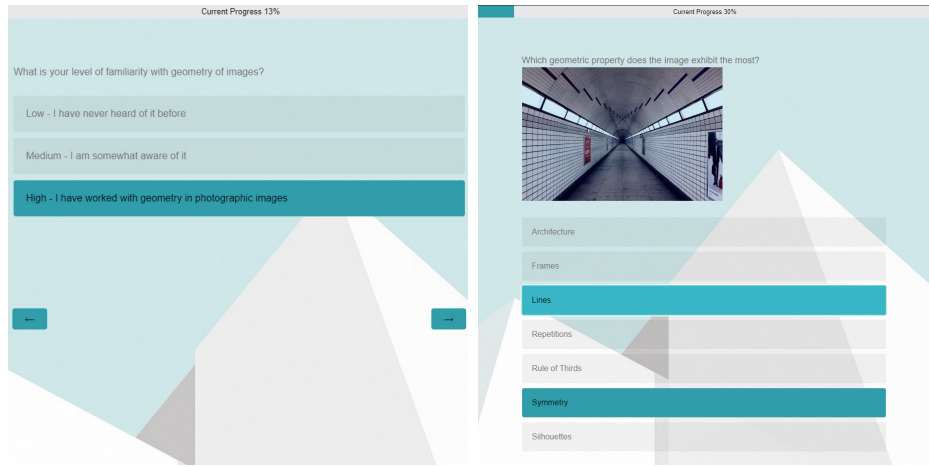
Additionally, confusion matrices are plotted to inspect confusions which occur between certain geometric classes (for example, Lines might often get confused with Architecture). Since assessing photographic styles is a highly subjective task, confusion matrices help provide deeper insights into the working of the neural networks.

Feature Map visualization using Grad-CAM localization technique

Since deep learning models are mostly a black-box and geometric style detection is a fairly new area which has tremendous scope of research, feature maps are visualized to understand which pixels of the image are the neural networks capturing before detecting their geometric style. Gradient-weighted Class Activation Mapping (Grad-CAM) [18] is used to visualize the class-specific gradient information flowing into the final convolutional layer of a CNN and produce a coarse localization map of the important regions in the image. This helps in explaining and advocating the pattern of how the models are classifying images the way they are.

User Study

Since the Geo-Style dataset is novel and interpretation of photographic styles are highly subjective, a comparison is carried out between the predictions of the neural networks with human observers. In order to provide a human baseline for evaluation of the Geo-Style dataset, a user study is conducted. A total of 22 photos from the dataset were selected to get the user reviews on and 43 responses were received. For each photo, seven options corresponding to the seven identified geometric style were shown to the subjects and they evaluated by selecting the styles they think the image is exhibiting the most. A few screenshots of the study are in figure 4.1. A short description of the geometric styles were also provided separately to the subjects for their easy reference. To keep the study diverse, people from all age groups were asked to participate.

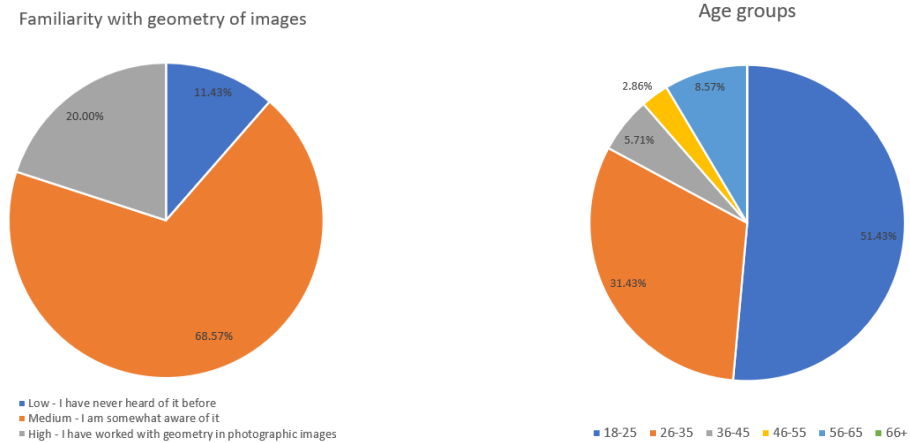


(a) Figure a

(b) Figure b

Figure 4.1: Screenshots of the user study. The user can select multiple options of geometric styles that they think are most relevant to the given image. The geometric style with the highest number of votes for each image is considered for evaluation.

The subjects were required to provide their familiarity with geometry of images, the distribution of which is shown in Figure 4.2.



(a) Figure a

(b) Figure b

Figure 4.2: Distribution of subjects in user study with respect to their familiarity with geometry of photographs (Figure a) and age (Figure b).

4.2 Results

4.2.1 Mean Average Precision

Figure 4.3 gives a comparative analysis of the overall performance of the neural networks with center crop and random crop transformations in geometric style detection in terms of their Mean Average Precision (MAP) score.

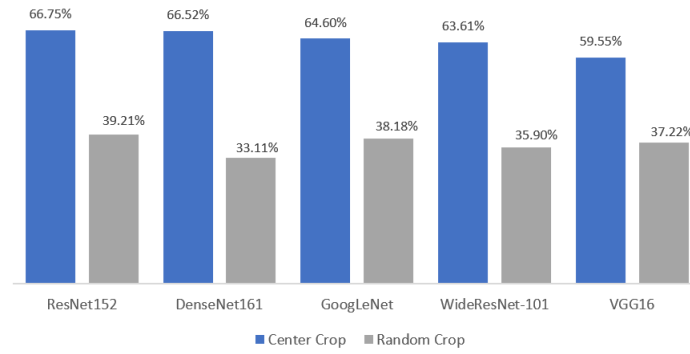


Figure 4.3: Mean Average Precision: ResNet152 with Center Crop gives the highest MAP followed by DenseNet161 Center Crop

As can be observed from the bar graph (Figure 4.3), center crop with all the selected models give better performances than random crop with all models. ResNet152 with center crop gives the best performance with MAP score of 66.75% followed closely by DenseNet161 which has MAP score of 66.57%.

4.2.2 Per-class precision scores

Tables 4.1 and 4.2 demonstrate the Per-class Precision Scores to show the performance for the different neural networks in detecting the geometric styles of photographs with center crop and random crop transformations respectively for each geometric style.

Geometric Styles	DenseNet- 161	ResNet- 152	GoogLeNet	Wide ResNet- 101	VGG16
Architecture	61.82	66.73	59.18	61.74	43.20
Frames	62.58	58.14	63.20	58.4	62.54
Lines	82.05	84.17	82.16	80.59	82.2
Repetitions	45.47	48.88	45.35	46.93	44
RoT	60.91	66.60	51.86	58.65	49.30
Silhouettes	90.49	83.85	85.67	84.62	85.60
Symmetry	62.50	58.97	64.93	61.49	50.01

Table 4.1: Category wise performance of models with center crop

Geometric Styles	DenseNet- 161	ResNet- 152	GoogLeNet	Wide ResNet- 101	VGG16
Architecture	22.89	30.95	33.74	27.57	41.95
Frames	39.02	41.43	39.85	39.30	32.02
Lines	41.74	54.95	48.40	44.53	52.3
Repetitions	31.07	29.80	26.33	23.60	30.22
RoT	25.32	28.74	30.55	26.81	26.14
Silhouettes	46.00	59.54	58.70	61.76	52.80
Symmetry	25.85	29.17	29.84	27.98	24.66

Table 4.2: Category wise performance of models with Random Crop

From tables 4.1 and 4.2, it can be seen that per-class precision scores for each class with center crop as transformation technique are better than the same with random crop as transformation technique. The scores of best performing models for each of the geometric style is highlighted in bold. ResNet152 with center crop performs the best in detecting Architecture, Lines, Repetitions and Rule of Thirds. GoogLeNet with center crop most precisely detects Frames and Symmetry. DenseNet161 with center crop gives the highest precision score for Silhouettes.

4.2.3 Confusion Matrix

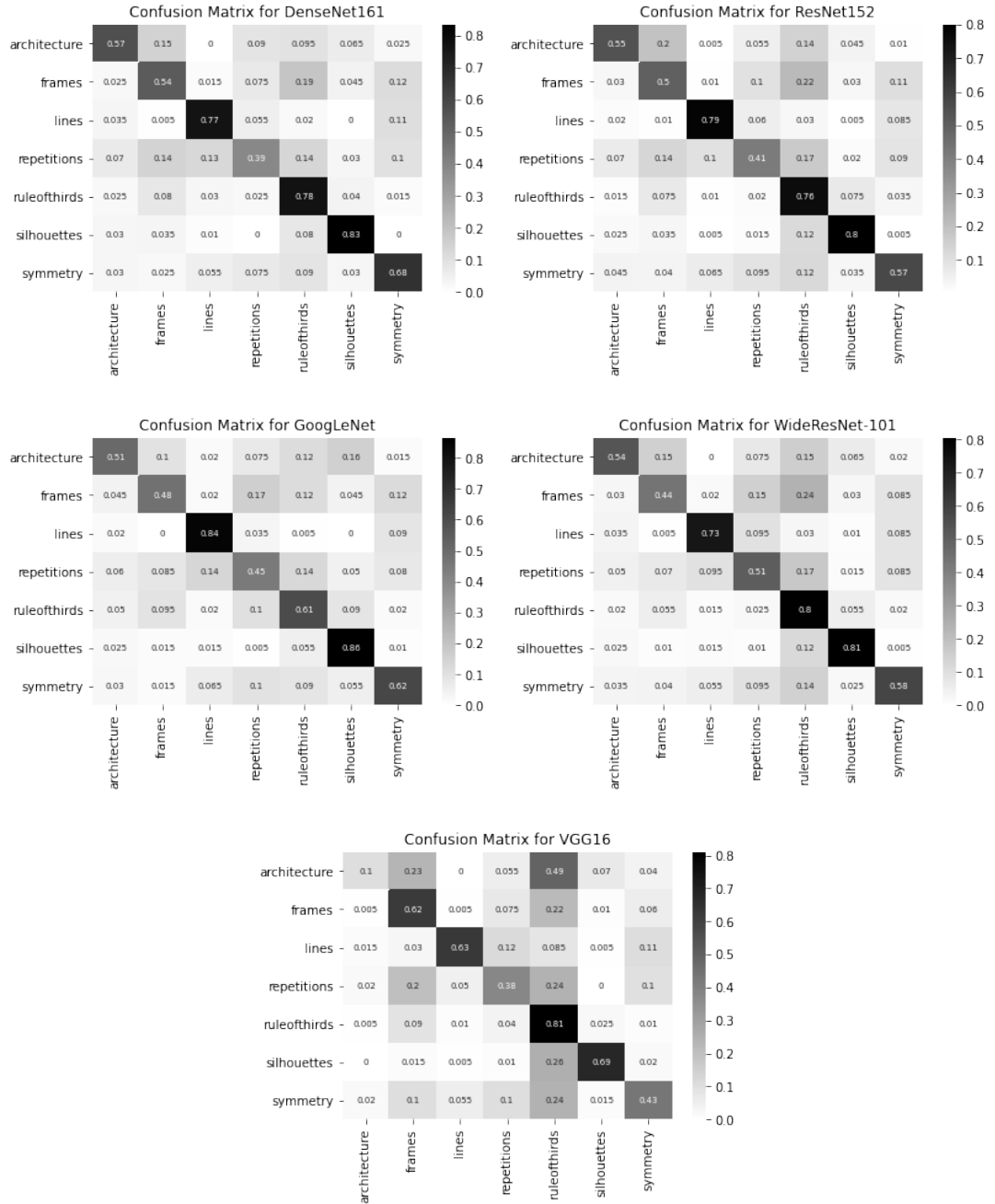


Figure 4.4: Confusion Matrices showing performances of each neural network

Figure 4.4 shows the confusion matrices for each network with center crop. This helps us in the identification of classes which the networks are confusing with each other.

The plotted confusion matrices for each network help in giving deeper insights into the confusions which are occurring between geometric styles. DenseNet161 and ResNet152 have similar instances, where all the geometric features are mostly correctly predicted with some understandable confusions such as that between Architecture and Frames or Frames and Rule of Thirds. Lines, Rule of Thirds, Silhouettes and Symmetry are the strongest classes, whereas Architecture, Frames and Repetitions are the weaker ones. Similar trends are observed for DenseNet161, ResNet152, GoogleNet and WideResNet-101. From an overall perspective, there can be seen a lot of confusions in the confusion matrix of VGG16 which justifies its low performance in detection of geometric features in images.

4.2.4 Visualization of feature maps

Figure 4.5 plots the feature maps of the five selected models with center crop for one image predicted to be belonging to each of the seven geometric styles. From Figure 4.5, it is observed that the superior performances of ResNet152 and DenseNet161 are justified as they are able to detect most significant details of the images which make them belong to a specific geometric style. On the other hand, the feature maps of VGG16 support its poor performance since the model misses out on capturing the critical details of the images and hence is unable to correctly detect the geometric styles.


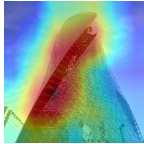
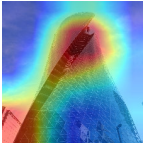
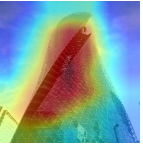
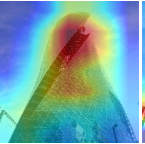
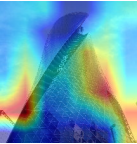

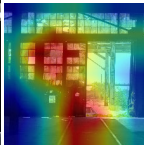
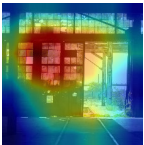
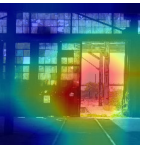
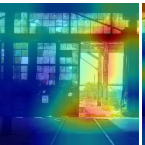
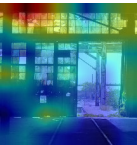
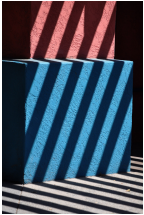
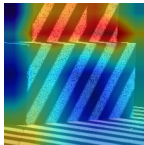
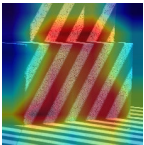
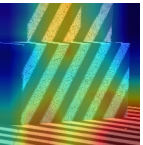
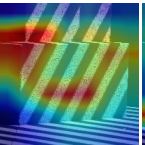
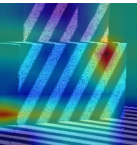

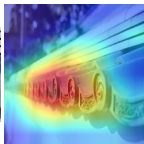
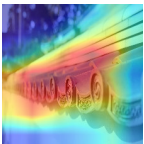
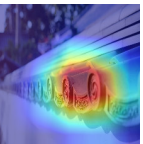
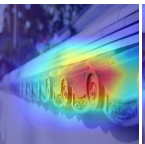
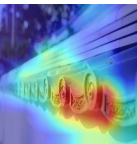



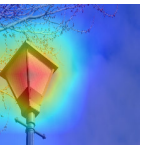
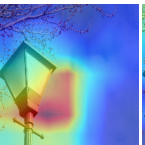
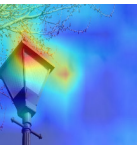







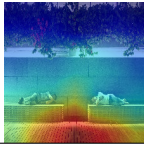
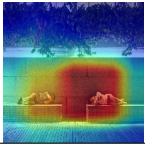
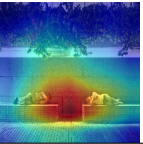
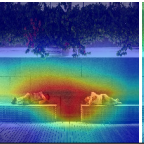
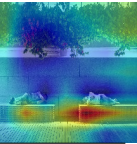
Style	Image	ResNet-152	DenseNet161	GoogLeNet	Wide ResNet-101	VGG16
Architecture						
Frames						
Lines						
Repetitions						
Rule of Thirds						
Silhouettes						
Symmetry						

Figure 4.5: Coarse localization feature maps of the regions in the image that are ‘important’ for predictions from the models

4.2.5 GUI results

To test the performance of the best performing model which is ResNet152 with Center Crop with any random image as an input, a graphical user interface (GUI) is built so that a user can have an interactive experience in automatically evaluating an image of his or her choice with respect to its geometric style. To keep things simple, softmax activation function is used to return the result in terms of the geoemtric style with the highest probability score amongst all as evaluated by ResNet152.

A few screenshots of the GUI are shown in Figure 4.6.

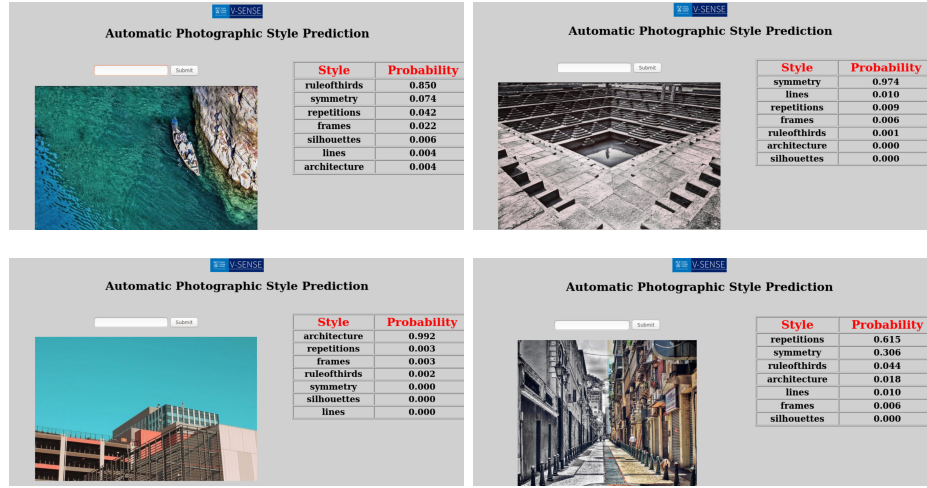


Figure 4.6: GUI snapshots of images. ResNet152 center crop is used to evaluate these images since it has achieved the highest MAP. If we see the probability scores, the network quite correctly detects the major geometric style present in the image

4.2.6 User Study

Style detection of a photograph being a highly subjective task, human views were decided to be incorporated in order to cross verify the predictions by the top three performing models - ResNet152, DenseNet161 and GoogLeNet - with center crop. 22 images from the test dataset were selected for this purpose. Figure 4.7 shows the Confusion matrices that were plotted to visualize the performance of the networks.

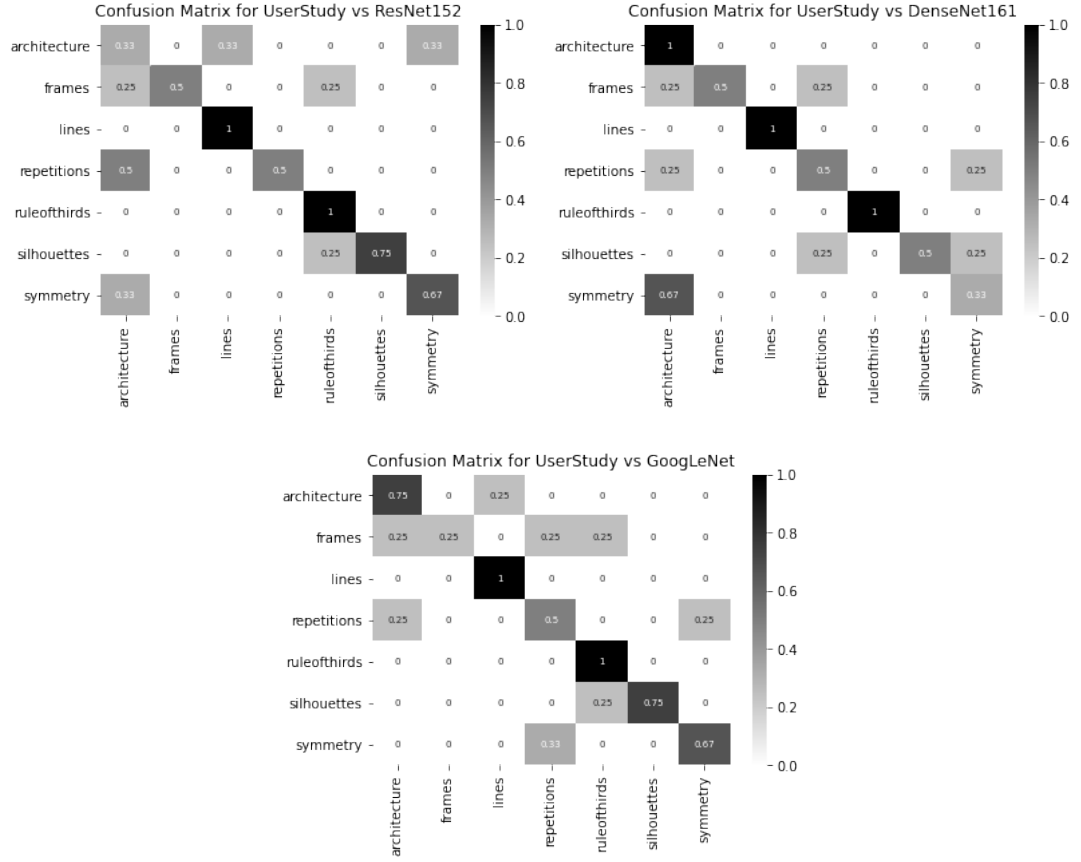


Figure 4.7: User study: Confusion Matrices of top three performing models' predictions with human baseLines.

It is observed that almost all the geometric styles are correctly detected. Instances of understandable confusions between certain styles are also observed, such as between Architecture and Frames, Repetitions and Symmetry.

4.3 Discussion

As observed in the Results section (Section 4.2), ResNet152 with center crop gives the best performance in detection of geometric style in photographic images. This section aims to achieve perspectives from the various evaluation metrics used and whether or not the results achieved from each of these metrics complement each other.

From the MAP and per-class precision scores, it is very evident that center crop with

the other data augmentation techniques performs in a much better fashion than random crop with the same set of data augmentation techniques in detection of geometric features. This might be because geometric style is highly dependent on the arrangement of subjects in images, and hence center crop can probably correctly capture the arrangement upto some extent, especially in geometric features such as Symmetry. On the other hand, if the cropped part of an image in random crop fails to be a part where the subject repeats (case of repetition) or exists in the edge of the image (case of Rule of Thirds), then the neural network will not have the correct set of input to be trained on, thereby leading to incorrect style detection.

In order to deep dive into the number of correct predictions made by the networks for each geometric style, confusion matrices are plotted for each of the networks. Upon inspecting the confusion matrices (Figure 4.4), cases of understandable confusions are observed, for example, Architecture vs Frames which make sense because Frames could be a part of Architecture. Frames and Rule of Thirds too get confused with each other which are plausible if the frame is located on the third line of the image. It is also observed that the confusions increase as we move from the model with higher MAP score to the model with lower MAP score. For example, examining the confusion matrix of VGG16, it is seen that Repetitions and Architecture are two of the weakest classes in VGG16 which further supports the per class precision scores achieved by VGG16 in Repetitions and Architecture of 44% and 43.20% respectively. Thus the confusion matrices help in cross verification of the results achieved by using MAP and per class precision.

The feature maps plotted by using Gradient-weighted Class Activation Mapping give further insights and transparency into the working of CNN based models. The main purpose of plotting these maps (Figure 4.5) is to provide answers to the "Why" by explaining the performances of the models in geometric style detection. The high MAP scores of the ResNet152 and DenseNet161 can be explained by the feature maps which highlight the parts of images that are captured by the models. For example, in Architecture, ResNet152 detects the lines as well as the curve of the structure. It also identifies the repeated subjects in Repetitions. DenseNet161 does a fantastic job in Repetitions where it is able to identify each repeated subject. It also detects the boy's figure in Silhouettes very accurately which bolsters the high precision score of 90.49% by DenseNet. GoogleNet gives the best performance of per-class precision

score of 64.93% in detection of Symmetry in photographic images. This is justified by the feature map of GoogleNet in Symmetry where the localized map is symmetric and captures the even aspect of the image. The feature maps of VGG16 show that the model is unable to detect information in the images which is crucial to classifying them into the correct geometric styles. Thus, these visualizations lend insights into the failure of a model in detecting geometric styles, showing that seemingly unreasonable predictions might have reasonable explanations.

The user study was conducted to have a human baseline and shed some light further into the evaluation of the predictions. The selected images to be evaluated by users were only 22 and confusion matrices were plotted to get a visualization of the strongest and weakest geometric styles. The top three performing models are used for this purpose (Figure 4.7). It is observed that while Lines, Rule of Thirds, Silhouettes and Symmetry work well for ResNet152 and GoogleNet, Symmetry do not work so well for DenseNet161. For DenseNet161, Symmetry is getting confused with Architecture which is logical because Symmetry can be a part of Architecture. A structure can be built in such a way that it is symmetric at the same time. For ResNet152, Architecture is getting confused with Lines and Symmetry. To explain this particular confusion, it is observed that a lot of images which exhibit Architecture as the primary geometric feature also have lines and symmetry as features in them. The user study reaffirms the theory that the manner in which photographic styles are perceived varies from person to person. There is no single definition of style and it is a highly subjective matter. Images often possess more than one photographic style and one way this problem can probably be addressed is by including multilabeled images for the dataset. More description is provided in the next chapter regarding the different research that can be carried out to achieve better performances from CNNs in detection of geometry-based photographic style.

Chapter 5

Conclusion

In this dissertation, a significant chunk of progress is made in defining the problem of understanding geometric styles of photographic images by using deep learning. Seven geometric styles - Architecture, Frames, Lines, Repetitions, Rule of Thirds, Silhouettes and Symmetry - in photography are identified and a novel dataset, Geo-Style, is created that exhibits these types of geometric that have not been previously considered in the literature.

Five different neural networks - ResNet152, DenseNet161, GoogLeNet, WideResNet101 and VGG16 - have been used and fine tuned by training on Geo-Style dataset, the performances of which are compared using several evaluation metrics. The experiments are conducted from two perspectives, first, with center crop as data augmentation technique and second, with random crop as data augmentation technique. ResNet152 with center crop gives the best performance with an MAP of 66.73%.

Confusion Matrices and feature maps are also plotted to demonstrate explainability of the neural networks. It is observed that the confusions made by the networks between classes are understandable as there are bound to be overlaps between photographic styles and the views are highly subjective. This is why a user study is conducted too to have a human baseline to compare the predictions with. Similar patterns are observed where geometric styles like Architecture and Frames get confused.

Feature maps using localization by Grad-CAM have helped in identifying the features of images that have been deemed as 'important' by the networks in the last later of the networks. The heatmap produced complement the performances of the models.

ResNet152 and DenseNet161, being the best performing models, quite accurately capture the important features in a better fashion than VGG16. Thus, this dissertation demonstrates not only the state-of-the-art results in the detection of geometric photography styles, but also provides explanation to justify the detection by the neural networks.

5.1 Main contributions

The methodology of the this dissertation proposes a pipeline to evaluate how well neural networks can detect geometric features in the photographic image. Apart from Rule of Thirds, there has been very few to no research done from a geometry-centric perspective in the field of computer vision. This required to first identify geometric styles which are most prevalent in photography. Seven relevant geometric features are identified in this dissertation which are: Architecture, Frames, Lines, Repetitions, Rule of Thirds, Symmetry and Silhouettes.

Since a dataset concentrating on only geometric properties of images was not found in previously conducted research, a novel dataset is created in this dissertation to accommodate the geometric styles identified in images. The dataset consists of 12,000 annotated single labeled images chosen from Flickr.

Furthermore, this dissertation demonstrates the state-of-the-art results in detection of the geometric styles in images by several neural networks. The performances of five neural networks are compared by various evaluation metrics.

Additionally, the dissertation also provides explanations on the behavior of the networks by visualizing features in images which are considered to be 'important' by the networks. Since interpretation of style is a highly subjective task, a user study is conducted too to get a human baseline to evaluate the networks' predictions.

5.2 Future work

Since the area of research into evaluating geometric styles of photography is still at its early stage and has not been much explored, the scope of it is huge. The dissertation sets the groundwork for further work to be incorporated into it.

The dissertation considers only seven most relevant geometric styles present in photography. This can be extended to accommodate more number of styles and evaluate the performances of CNNs in detection of these additional geometric styles.

The dataset can be further expanded. The current dataset, Geo-Style, consists of 12,000 images taken from Flickr. Other photography sites such as Pinterest could be used too to get more variety. Also, the images in the current dataset for both training and testing are single labeled. The dataset can be modified to include multi-labeled images as well. Since the entire study of understanding photographic styles is highly subjective and a photograph might exhibit more than one styles, a multi-labeled dataset might produce interesting and more accurate results.

Only RGB features are extracted in the dissertation. Just like saliency maps were fed to a second column of CNN architecture along with RGB in the work done in [1], a similar approach might be adopted by using features such as lines and edges along with the general RGB feature extraction for geometric style detection. Since lines and edges are highly present in an image consisting of geometric elements, this might improve the performances of CNNs in detection of geometric styles present in an image.

HoughNet [24] and Holistically-Nested Edge Detection [17] are effective methods to detect meaningful semantic lines and edges in images respectively. Since both these architectures are built to detect lines and edges which happen to be crucial aspects of an image possessing geometry as its photographic style, they might exhibit better performances in detection of geometry-based photographic styles.

The system can also be extended to the domain of video and 360 images. As mentioned in Chapter 1, proper use of geometry in cinematography has the scope of creating powerful impacts on the viewers. Just like Wes Anderson movies that are usually aesthetically very pleasing to look at, understanding geometry of images could contribute towards producing dynamic and compelling scenes in videos.

With the ever expanding volume of visual content, there is a lot that can be done with aesthetics of images, such as navigating and organizing through aesthetic preferences and applying them to convey a story in cinematography. The study in this dissertation serves as a benchmark for understanding geometric-based photography style in images using deep learning and even though it is still in its nascent stages, there is a substantial amount of research that can be conducted into the development of computational models of style detection in the field of computer vision.

Bibliography

- [1] K. Ghosal, M. Prasad, and A. Smolic, “A geometry-sensitive approach for photographic style classification,” *arXiv preprint arXiv:1909.01040*, 2019.
- [2] S. O’Gara and K. McGuinness, “Comparing data augmentation strategies for deep image classification,” 2019.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [5] N. Murray, L. Marchesotti, and F. Perronnin, “Ava: A large-scale database for aesthetic visual analysis,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415, IEEE, 2012.
- [6] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, “Recognizing image style,” *arXiv preprint arXiv:1311.3715*, 2013.
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

- [8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, pp. 740–755, Springer, 2014.
- [9] S. Lee, “Wes anderson’s ambivalent film style: the relation between mise-en-scène and emotion,” *New Review of Film and Television Studies*, vol. 14, no. 4, pp. 409–439, 2016.
- [10] J. C. Van Gemert, “Exploiting photographic style for category-level image classification by generalizing the spatial pyramid,” in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pp. 1–8, 2011.
- [11] L. Manovich, “Subjects and styles in instagram photography (part 1),” *Instagram Book*, 2016.
- [12] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Studying aesthetics in photographic images using a computational approach,” in *European conference on computer vision*, pp. 288–301, Springer, 2006.
- [13] L. Mai, H. Le, Y. Niu, and F. Liu, “Rule of thirds detection from photograph,” in *2011 IEEE International Symposium on Multimedia*, pp. 91–96, IEEE, 2011.
- [14] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” *arXiv preprint arXiv:1312.6229*, 2013.
- [15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.
- [16] Q. Han, K. Zhao, J. Xu, and M.-M. Cheng, “Deep hough transform for semantic line detection,” *arXiv preprint arXiv:2003.04676*, 2020.
- [17] S. Xie and Z. Tu, “Holistically-nested edge detection,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.

- [18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.
- [19] M. Kennedy, *Understand Geometric Photography*, accessed September 3, 2020.
- [20] T. Ivanova, *Why Framing Is Important in Photography*, accessed September 3, 2020.
- [21] T. Ivanova, *How To Understand and Use Aesthetics in Photography*, accessed September 3, 2020.
- [22] *Torchvision Transforms*, accessed September 3, 2020.
- [23] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016.
- [24] N. Samet, S. Hicsonmez, and E. Akbas, “Houghnet: Integrating near and long-range evidence for bottom-up object detection,” *arXiv preprint arXiv:2007.02355*, 2020.