AESTHETIC QUALITY ASSESSMENT FOR REAL ESTATE IMAGES THROUGH DEEP LEARNING METHODS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

NAZLI ÖZGE UÇAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

DECEMBER 2022

Approval of the thesis:

AESTHETIC QUALITY ASSESSMENT FOR REAL ESTATE IMAGES THROUGH DEEP LEARNING METHODS

submitted by **NAZLI ÖZGE UÇAN** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. Ahmet Oğuz Akyüz
Supervisor, **Computer Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Tolga Kurtuluş Çapın
Computer Engineering, TED University _____

Prof. Dr. Ahmet Oğuz Akyüz
Computer Engineering, METU _____

Assist. Prof. Dr. Ramazan Gökberk Cinbiş
Computer Engineering, METU _____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Nazlı Özge Uçan

Signature :

ABSTRACT

AESTHETIC QUALITY ASSESSMENT FOR REAL ESTATE IMAGES THROUGH DEEP LEARNING METHODS

Uçan, Nazlı Özge
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Ahmet Oğuz Akyüz

December 2022, 64 pages

In this thesis, we aim to find the aesthetic quality of real estate images. Although aesthetic assessment is a subjective terminology, it is highly correlated with photographic rules. The aesthetic quality of images in real estate affects the decision of potential people of interest. The aesthetic evaluation of images is established via the Aesthetic Visual Assessment (AVA) dataset benchmark. Although AVA is a publicly available and diverse image dataset, it cannot be adapted to the real estate domain. Therefore, we constructed the Real-Estate Aesthetics Assessment Dataset (RAAD), which consists of real and synthetic real estate images. In order to gather subjective user data on the RAAD, a user study is conducted on a custom web-based scoring platform, serving RAAD image data. We analyzed several different methods involving classical vision classifiers and deep image classification models in order to assign an aesthetic quality score to the given real estate image. The results of those different approaches are presented comparatively on the RAAD data.

Keywords: aesthetic quality assessment, real-estate image quality classification

ÖZ

DERİN ÖĞRENME YÖNTEMLERİ İLE EMLAK GÖRÜNTÜLERİNDE ESTETİK KALİTE DEĞERLENDİRMESİ

Uçan, Nazlı Özge
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Ahmet Oğuz Akyüz

Aralık 2022 , 64 sayfa

Bu tezde emlak görüntülerinin estetik kalitesini bulmayı amaçlıyoruz. Estetik değerlendirme öznel bir konu olmasına rağmen, fotoğrafçılık kurallarıyla yüksek oranda ilişkilidir. Emlak görüntülerin estetik kalitesi, potansiyel ilgililerin kararını etkiler. Estetik Görsel Değerlendirme (AVA), görüntülerin değerlendirilmesi için kullanılan halka açık ve çeşitli bir görüntü veri seti olmasına rağmen, emlak görüntüleri içermez. Bu nedenle, gerçek ve sentetik emlak görüntülerinden oluşan Gayrimenkul Estetik Değerlendirme Veri Kümesi (RAAD) veri tabanını oluşturuldu. RAAD üzerinde öznel kullanıcı verilerini toplamak için, özel olarak geliştirilen web tabanlı puanlama platformunda kullanıcı çalışması yapıldı. Verilen emlak görüntüsüne estetik kalite puanı atamak için klasik görüntü ve derin görüntü sınıflandırma modellerini içeren farklı yöntemleri analiz ettik. Sonuçlar, RAAD üzerinde karşılaştırmalı olarak sunulmuştur.

Anahtar Kelimeler: estetik kalite değerlendirmesi, gayrimenkul resim kalite sınıflandırması

ACKNOWLEDGMENTS

I would like to express my deep gratitude to my supervisor, Prof. Dr. Ahmet Oğuz Akyüz, without his guidance I would have not been able to fulfill my masters. I feel so lucky to have him as my supervisor.

I want to thank my family and friends, for always being there for me in my time of need.

I also acknowledge that this thesis work is supported by TÜBİTAK BİDEB 2210-A National Scholarship Programme for MSc Students through my masters education.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

RAAD	Real-Estate Aesthetics Assessment Dataset
PR-AUC	Precision-Recall Area Under Curve

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

In this research, we focused on the aesthetic quality of images in the real estate domain. While hunting for a new apartment we all browsed through several online real estate sites and talked with real estate agencies. Before going to the house, the photos are all we have for us to make our decisions. Those photos must have high quality, good angles, and lightning and they should represent the place as best as they can. Otherwise, we would not give the site even a chance to visit in person. Many real estate agencies hire photographers for this job or they might take the photos themselves. During the selection of photos among many that have been taken that will go online, some sort of selection must be done. So far, agencies do this process manually. By using the outcome of the image aesthetics quality assessment score, they can easily sort the images from highly desirable ones to bad quality ones and save resources.

1.2 Proposed Methods and Models

In this thesis, handcrafted feature analysis is done on the RAAD dataset and several classification methods are adapted. Comparisons of different approaches are done based on the classification results for real and synthetic data and user given scores.

Figure 1.1: Aesthetic image samples from [1]. 1st row has generally low aesthetics scores whereas 2nd row has high aesthetic scores

1.3 Contributions and Novelties

During the preparations of this thesis, several contributions are done. They can be summarized as follows:

- A novel real estate dataset is crawled from the internet with annotations gathered from users via custom online user study platform.
- Transformer architecture is experimented in the image aesthetic quality assessment domain for the first time.
- Real estate image aesthetic assessment methods are developed.
- Real estate image aesthetics assessment is investigated under the domain of general image aesthetics assessment.

1.4 The Outline of the Thesis

In the following sections, the literature survey, dataset, methodology, experimental results, and discussions are given. In chapter 2, a literature survey on image aesthetics quality assessment is investigated for classical image vision approaches and deep learning approaches. The research done on the real estate domain is also given here. In chapter 3, the different dataset used for image aesthetic quality assessment is compared and the real estate dataset is analyzed. Also, the user study is explained in detail on how to gather user data. In chapter 4, methodologies are given for the suggested problem. In chapter 5, experimental results are given. Chapter 6 has in depth analysis of the outcomes. In chapter 7, the conclusion is given with limitations during the experiments and possible future work.

CHAPTER 2

LITERATURE SURVEY

This chapter aims to give background information on image aesthetics assessment from the classical vision area through more recent deep learning methods. It explains how image aesthetic models are developed through the lens of neuroaesthetics and computational metrics. Later, the research conducted on the real-estate domain is included. The usage areas of the intersection between computational metrics and the real-estate are explained.

2.1 Image Aesthetics Quality Assessment

Aesthetics assessment at its core is a subjective topic. However, there has been much research in the neuroaesthetics domain, where neuroscientists, artists, and psychologists work together to understand how the brain works and responds to aesthetics. They observe the responses of the brain to seeing a conventionally beautiful face or an artwork. They find that under aesthetically pleasing sights, the core emotional areas and reward-related places in the brain brighten up [2], which means that seeing a beautiful face triggers the same area on the brain as winning a scratchcard [3] or it is the reason we feel like crying over a beautiful view. In [4] Ramachandran and Hirstein came up with 'eight laws of artistic experience', some of the main laws of what makes something aesthetically pleasing. Some of those laws come from the reward mechanism of the brain, some from the evolutionary survival skills that we develop, and some from the emotional background. If we were to analyze those eight laws, we can see that there is a high correlation with how computer scientists approach this subject. In the next chapters, we explained the classical vision approaches that

use the computational metrics of those laws and deep learning approaches to further understand aesthetic quality assessment.

2.1.1 Classical Vision Approaches

Assessment of the aesthetic quality of a given image can be accomplished by many different computational methods. Before deep learning methods come into play, all the methods rely upon handcrafted features and classical vision algorithms such as Support Vector Machines and Linear Regression. In [5], classical approaches to image aesthetics quality assessment are divided into 4 categories, based on the type of handcrafted features that each use. In the first category, there are methods using high or low-level simple image features, such as image contrast, sharpness, saturation, exposure, texture, depth, clarity, and so on. The second category consists of methods using image composition features. Image composition features consist of the features representing the salient object, such as the rule of thirds, sky illumination, low depth of field, etc. These methods generally combine simple features as well. The third category uses general-purpose features. Bag-of-Visual words, SIFT descriptors, color descriptors, and Fisher vectors are the general features that are used under this category. Also, in the last category, the task-specific feature extraction methods are explained. They are focused on the human face, landscape, or typography contexts and extract features only for their context.

Most of these methods, make use of the basic photographic principles, such as the rule of thirds, golden ratio, and color harmonies as it is believed to yield the best results.

2.1.2 Deep Learning Approaches

With the increasing use of deep learning methods, classical approaches becomes not adequate for the image aesthetic quality assessment task. In the early stages, the deep classifiers are used only to extract deep features of the image. Those features are then fed into an SVM classifier [6], [7]. Later on, the adoption of deep learning approaches to the aesthetics assessment improves as a classical deep image classi-

fication task. The DMA-Net extracts multiple random fixed-size patches from the image and feeds them into Alex-Net-based CNN and orderless aggregation to combine the patch results [8]. The MNA-CNN explains that while CNN architectures try to learn an image, the usage of random cropping and scaling would affect the aesthetic properties of the image [9]. Hence, they proposed a composition-preserving VGG16-based architecture, where they also used adaptive-spatial pooling to handle varied size inputs. The A-LAMP [10] took the idea of incorporating image patches but they used an adaptive patch-selection method to focus on patches that carry important aspects of the image. They also incorporated a layout-aware subnet to add to information coming from the salient object in the image. APM [11] focuses on the aesthetic score distribution of the image. It uses a ResNet-based CNN to extract the features and use them for regression to obtain the score distribution. NIMA [12] also focuses on the score distribution as APM. It takes out the last layer of the Inception-v2 network and adds a fully-connected layer and a softmax layer to obtain the prediction distributions. MPada [13] approaches the task using multiple image patches and increases or decreases the weight of each patch during training. Hosu et. al [14] use the images in their original size by extracting features with multi-level spatial pooling. In this way, they try to achieve the information loss of image resizing. RGNet [15] emphasizes the composition of the elements in an image. They extract the important regions using semantic segmentation, represent each segment as a graph node, and train a graph convolutional network on those features. Zeng et. al [16] proposed a unified approach to the different ways of image aesthetic assessment; binary classification, score regression, and score distribution. AFDC [17] proposed a kernel using adaptive dilated convolution. This can be plugged into existing CNN architectures to extract information when random cropping or resizing occurs. PA-IAA [18] focuses on the personality trait of the aesthetics assessment. They trained a multi-task network and a Siamese network to find a correlation between personality and the image aesthetic scoring and increase the performance for each. As a more recent approach, HLA-GCN [19] used graph convolutional networks using layout data as graph nodes.

Some of those methods provide an adjustment to the classification model for them to better represent aesthetic assessment. Some approached this task as a binary classification, some as regression, and some as score distribution. Some use the layout infor-

mation in the image, some used the whole image and some used the image patches.

The application areas of image aesthetic assessment are vast. We can use the aesthetic value of the image for photo album summarization [20], [21], automated photo editing and enhancement, multi-shot photo selection, etc.

2.2 Real Estate Assessment

The early implications of deep learning methods in the real estate domain are on the property risk assessment. Ju et. al [22] used simple back propagation neural net and Zhao et. al [23] used a back propagation neural net with genetic algorithms for risk assessment. Along with risk assessment, price estimation models arose. You et. al [24] use RNN on the exterior of the house images. Naumzik and Feuerriegel [25] used VGG-16-based CNN with statistical inference. Wang et. al [26] used deep learning approaches with time series forecasting. Nadai and Lepri [27] used neighborhood information for the house value estimation. Kucklick and Müller [28] used satellite images and real estate text data for multi-kernel learning. Bin et. al [29] used street map information of houses and extract those features with a CNN. They incorporated an attention mechanism to combine the features with house listing data. Law et. al [30] also used satellite images. They trained a CNN and a hedonic price regression model for real estate value estimation.

In the real estate domain, both the people searching for landed properties and the real estate agents attach importance to the images they put online. While people are looking for a property they look at real estate listings online and only when the pictures are parallel with what they are looking for and are of good aesthetic quality, then they go to the showings of the place. Hence, real estate agents or property owners need to put favorable images. With the increasing size of taken photographs, the selection process could be compelling. That is why an automated aesthetics score rating will be beneficial. Based on our research, there has not been any work done on this problem and this work is the very first one.

CHAPTER 3

DATASET

There are many image-related datasets, however, only a few of them are about aesthetics assessment. In this chapter, the details of the dataset used in the literature for aesthetics quality assessment are given. In this research, the aim is to develop a model for visual aesthetic quality assessment of real estate data and the AVA dataset lacks the necessary image contents for this purpose. Hence, a novel dataset; the Real-Estate Aesthetics Assessment Dataset, RAAD is constructed. In the second part, the custom RAAD is defined and the user study is explained in detail on how the scores are gathered.

3.1 Aesthetic Assessment Datasets

As there are many image datasets, there are a few image datasets for aesthetic quality assessment in the literature. PhotoNet [31] and DPChallenge [32] are online photography challenges. In [31], photographers upload their photos and rate others' on a scale of 1 to 7. In [32], amateur or professional photographers can upload their photos. DPChallenge constructs the basis for many other aesthetic datasets in terms of image content. Some of the main datasets used in the literature are listed below:

- Image Aesthetics Dataset, IAD [33] is a large-scale dataset. It has images gathered from DPChallenge [32] and Photo.Net [31]. IAD assigns binary image scores to each image as low or high.
- Aesthetics with Attributes Database, AADB [34] is gathered to give meaning to image aesthetic scores. They gather user score data along with the 11 selected

attributes for each image. In AADB, there are 10000 photos with 5 scores each.

- FLICKR-AES [35] dataset is especially gathered for a personal assessment of image scores. The dataset aims to develop aesthetic assessment not as a generalized concept but rather for per-user-specific use cases. It contains 40000 images, with each image scored by 5 people with a score from 1 to 5.

- Aesthetic Visual Analysis, AVA [1] is the most comprehensive aesthetic assessment dataset in the literature and is used as the main dataset for benchmarking by many researchers. AVA has nearly 255,000 images. It provides 3 different annotations for each image for different image aesthetic analysis areas. Each image is associated with 1 or 2 semantic tags from nature to family. There are 66 tags and they contain the semantic information of the image content. The second annotation is for the photographic styles. They combine different photographic style challenges and came up with a total of 14 photographic styles, including composition, light, HDR, motion blur, etc. Aesthetic annotations are the third annotation associated with each image that we are focusing on in this research. Each image is scored by an average of 210 users based on the perceived aesthetic quality. Since the images are gathered mainly from photographic challenges, the user profile is mainly professional or amateur photographers. They believe that, since the user demographic has a trained eye for aesthetics, the scores given by them contain a high value. Considering that AVA has an average of 210 scores for each image, the researchers can use those scores for different types of image aesthetic quality assessment; image score distribution, image score prediction, and image score classification.

The comparison of image aesthetic datasets is given in 3.1. As we can see from the table, only AVA and IAD have a high number of images. However, IAD does not have the score distribution, instead only the binary aesthetic label of an image. With this, we cannot know whether an aesthetically pleasing image has a score distribution inclined to the highest score or just above the average. This information is useful to make an aesthetic judgment on an image. Taking every aspect into account, we decided to use AVA [1] as the baseline dataset and extract image aesthetic scores as binary labels, i.e aesthetically pleasing and unpleasing.

Table 3.1: Aesthetic Assessment Dataset Comparison

	IAD	AADB	FLICKR-AES	AVA
Number of images	1.5 million	10 000	40 000	255 000
Personalized Scores	✗	✓	✓	✗
Semantic Labels	✗	✓	✗	✓
Score Distribution	✗	✓[a]	✓	✓

[a] Only five people scored the whole dataset, so the distribution cannot be generalized.

3.2 Real Estate Aesthetics Assessment Dataset: RAAD

Toward evaluating real estate image aesthetic quality, there is no dedicated dataset. Although AVA has interior and architecture categories, images in those categories do not represent the real estate domain completely and the number of images in those categories is insufficient. Hence, we constructed the Real Estate Aesthetics Assessment Dataset, RAAD. There are two different image categories in the dataset; real and synthetic images. Real images are the existing real estate property photographs and synthetic images are generated from hyper-realistic 3D reconstructions by [36]. The scoring system is constructed by focusing on the aesthetic annotation structure of AVA. Each image is associated with a set of scores given by users. For score assignment, a user study is developed. Details of each image category and how to conduct the user study are explained in the following sections.

3.2.1 Real Images

Real images are gathered from online real estate websites [37], [38], [39], [40], [41]. We had a few considerations while gathering the images. First, we wanted to make sure that images do not contain a watermark or realtors' logo on them, since they may affect the judgment on the image. Then, we tried to include diverse images. We tried not to focus on one geographic place but rather gather images from different countries. This way, the architectural style would not be biased towards a certain type. Then, for diversity, we included apartments from different price ranges, so that

Figure 3.1: Real estate images gathered from web

the real estate property would contain various interior designs. Lastly, we tried to include both furnished and non-furnished places, since while people are looking for real estate, they would come across both of those kinds of properties. Our purpose with these criteria is that we would construct a diverse dataset representing what a user would experience. In 3.1 there are real estate image selections of the real part of RAAD.

3.2.2 Synthetic Images

Considering the insufficient data from real images, we decided to incorporate synthetic data into our dataset. There are many indoor reconstruction methods are available. Some only use synthetic data [42], [43] and some use scanned real indoor scenes to generate the 3D data [44], [36]. Among them, [36] gives the most realistic rendering of an indoor place. The mirror reflections, texture information, and finer details are eminently realistic. So, for the synthetic images of RAAD, we gathered them through Replica-Dataset [36]. By using the ReplicaSDK they provide, we place a camera object and generate shots of the scenes with configurable pitch, yaw angles, and different saturation and exposure values.

In the Replica Dataset, there are 18 indoor scenes. Out of the available 18 scenes,

we selected the most realistic 5 of them. Other scenes contain rendering errors and they were far from being realistic. For each scene, we place a camera in the middle of the scene, at the eye-sight height. We selected 3 different pitch angles with 40 degrees between them and 6 different non-overlapping yaw angles with 60 degrees. Then, for each angle position, we selected 2 different saturation values; slightly high saturation and slightly low saturation, and 2 different exposure values; slightly high exposure and slightly low exposure. Also, for each setting, we modified the camera zoom option to zoom in, normal zoom, and zoomed out. So, for each scene, we have 18 different angles, 3 different zoom levels, and 4 different saturation-exposure options.

In 3.2, a room scene is shown with different pitch and yaw angles. The first row has the default camera yaw angle and in each row, we increment the yaw angle by $60°$. Each column shows the normal, low, and high pitch values of the current position. We can see that a normal pitch angle has mostly the most flattering images, whereas some angles are not favorable at all.

In 3.3, a hotel scene is shown with different saturation and exposure values. We can observe that the high saturation value yields more color-rich images but they seem more processed and low saturation yield dull images. We can also observe that the low exposure value produces darker images and high exposure produces brighter images but the details are lost. In 3.4, 2 different scenes are given with different zoom levels for the camera. We can observe each of the generated zoom levels in real-life photographs.

3.3 User Study

To gather user data, we conducted a user study. A custom web application is developed for this purpose. When a user opens the user study, the application first selects whether to show synthetic or real image data. Then, it randomly selects N real estate property images. During the initial phase of the user study, we select N as 50, to ensure that users would not get bored and finish the study. With 50 images, it would take 3-5 minutes to finish the study. The user is asked to give a score to all images

Figure 3.2: Synthetic room images. 1st column shows the pitch down, 2nd shows the normal pitch and 3rd shows the pitch up with $40°$ in between. In each row camera yaw angle is rotated by $60°$

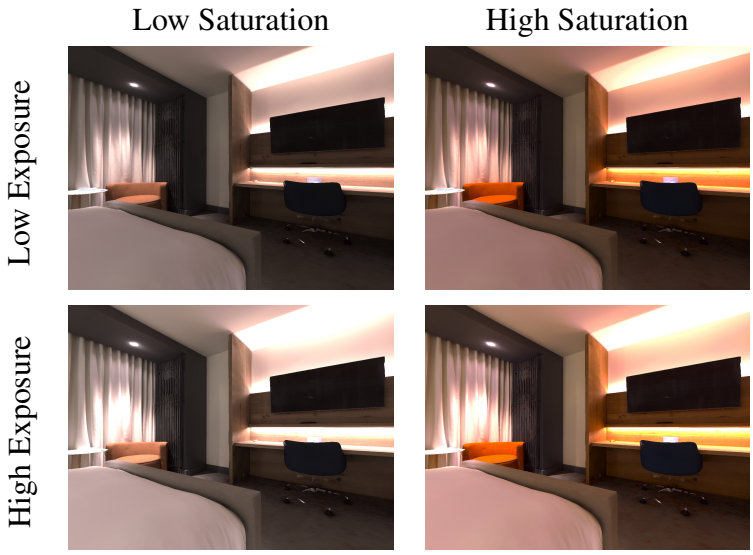


Figure 3.3: Synthetic room images. Bottom row images have slightly higher exposure value than the top row and the second column images have slightly higher saturation value than the first column images.

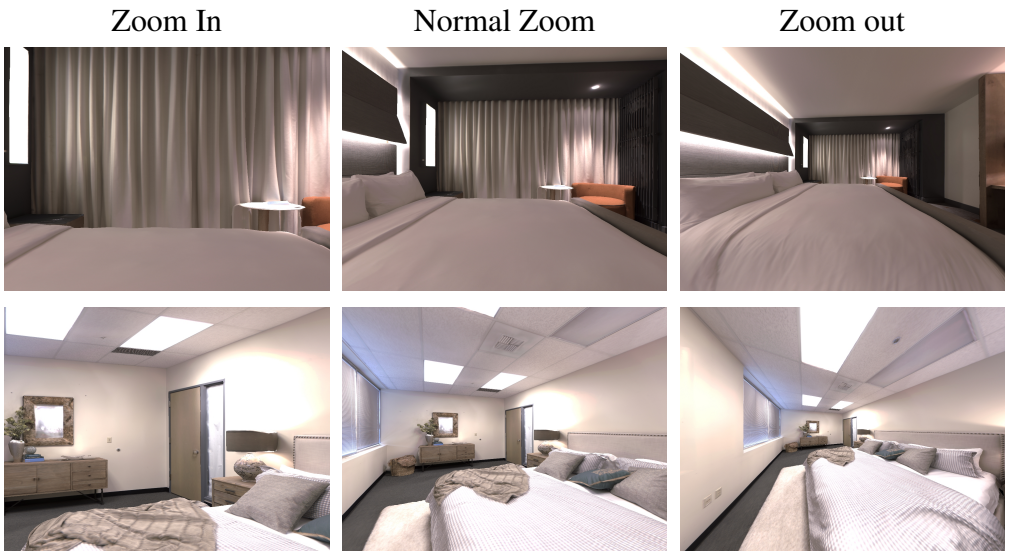


Figure 3.4: 2 different set of synthetic room images. 1st column shows the zoomed in images, 2nd shows the default zoom level and 3rd shows the zoomed out images.

based on how aesthetically pleasing the image is to that user:

> In this experiment, you are asked to indicate the visual appeal of various real-estate photographs from a score of 1 (very unappealing) to 10 (very appealing). In total, you will make a judgment for 50 photographs. There is no time limit but at a normal pace the experiment is expected to finish in about 10 minutes. The definition of visual appeal is a personal one, but try to picture yourself as a potential buyer reviewing real-estates by looking at their online photographs.

To maintain an even distribution of the scores on each image, we keep track of how many scores are given to each image. Upon requesting a new user study, the web application selects images amongst the least selected ones. Also, to preserve the credibility of the scores, we randomly select 3 images among all and present the user with those images twice. If the user gives different scores of margin more than 2, then that user is assigned as untrustworthy, and his/her data would not be used. For instance, if the user gives the first copy of the image a score of 5 and the second copy a score of 8, we do not count that user's result, even if the user gives other double image credible scores. However, if the user gives the first copy of the image a score of 5 and the second copy a score of 7 or 3, we take this as a dependable score and for the end score of that image, we take the average of both of them.

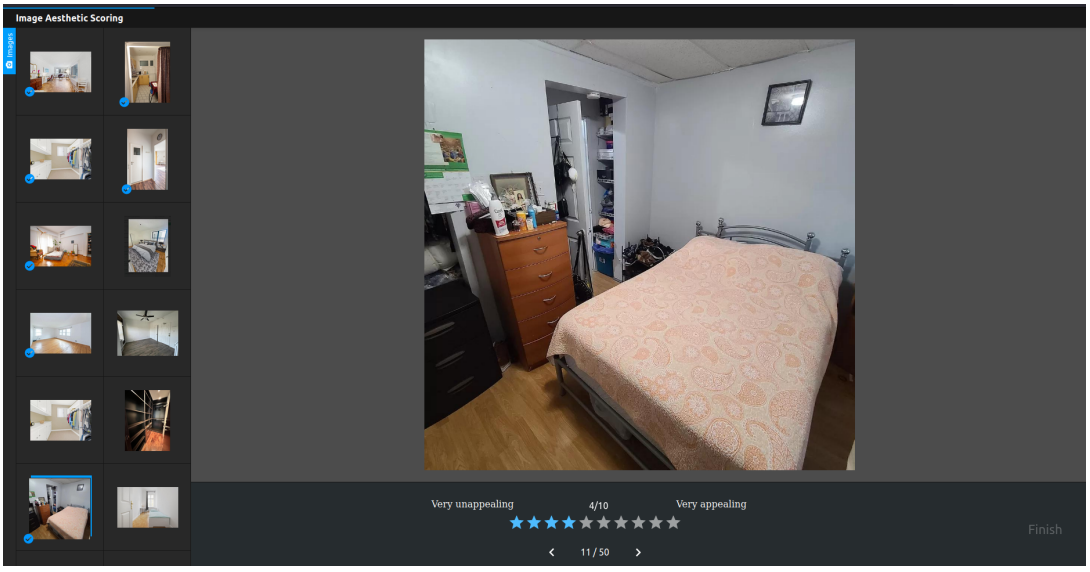


Figure 3.5: Screenshot image of the user study

In 3.5, the screenshot of the user study is shown. The image list, on the left panel, is given to the user in random order. When the user selects an image, it is displayed

in the center as a full size. Below the image, there is a scoring field from 1 very unappealing to 10 very appealing. When the user gives a score to the image, the star indicator changes its color and the score is displayed on top. At the bottom, the user can see the image number and how many images there are in total. When the scoring of an image is finished, its thumbnail on the left panel is updated with a blue mark. After finishing scoring all the images, the "Finish" button on the bottom right becomes clickable. Upon finishing the user study, the user is displayed a thank you note and if the scores are deemed trustworthy, they are saved on the remote server. If the user refreshes the page, he/she can re-do the user study with a different set of images, chosen randomly.

CHAPTER 4

METHODOLOGY

In this chapter, we present the methods for real estate image aesthetic assessment. Through a user study, we obtained the aesthetic scores of real estate images. We treat those scores as the ground truth and find a method to obtain a mapping with a computational metric.

In the literature, much research is done on image aesthetic assessment with many different application areas. In chapter 2.1, details of those methods are given. Some approach aesthetic assessment as a classification problem, some as a regression problem, or even a score distribution prediction problem. In this research, we approach this problem as a binary classification problem. Initially, we treat this as a 10- class classification problem by assigning the average scores given to each image as its ground truth label. Then, we proceed with binary classification by assigning based on the average score; 0 if it's less than 6, 1 otherwise, with 0 being aesthetically unpleasing and 1 being aesthetically pleasing. Also, due to the nature of image differences, we treat synthetic data and real data separately for all of the classifiers. The RAAD data contains 2125 images as mentioned in 3.2. The training, validation, and test splits for both real and synthetic images are given in 4.1. For all of the classification experiments, we use this split ratio.

Table 4.1: Train - Validation - Test Splits of RAAD

	Train	Validation	Test
RAAD Real Images	594	198	199
RAAD Synthetic Images	680	226	228

4.1 Classical Vision Approaches on Real Estate Aesthetics Classification

Initially, the task of the aesthetics assessment is done using classical vision classifiers. For this task, we gathered several handcrafted features and trained classical image classifiers with them.

4.1.1 Handcrafted Features

Several image aesthetics assessment approaches make use of handcrafted image features. To adapt the same principles to the real estate domain, we as well used handcrafted features. For this purpose, we used image mean, standard deviation, energy, hue, saturation, brightness, and texture features. For texture features, we used GLCM [45]. Then to choose which features contain useful information for aesthetic scoring, we used Spearman's and Pearson's correlation analysis. Spearman's correlation analysis gives the correlation value between the values based on the monotonic relationship whereas Pearson's correlation analysis gives the correlation based on their linear dependency of them [46] [47]. We analyze the handcrafted features for both of these analyses to find a high correlation coefficient.

4.1.2 Classical Classifiers

Before deep learning methods, we trained our data for the aesthetics assessment on the classical vision classifiers. For this purpose, we used Support Vector Machine (SVM) [48], Linear Regression (LinReg) [49], Logistic Regression (LogReg) [50], and Ordinal Logistic Regression (OrdLogReg) [51]. For both real and synthetic parts of the RAAD, we trained the classifiers separately using the handcrafted features. Initially, we used 10-class classifiers based on the user scores. Then we moved to binary classification by assigning scores equal to or lower than 5 as 0 and others as 1. During the training, we used PCA analysis to select the best fit 150 features [52] and used 3-fold cross-validation. For each classifier, we calculated the confusion matrix. The confusion matrix represents the class predictions of each class. A sample confusion matrix layout for binary classification is given in 4.2. From the confusion matrix,

Table 4.2: Confusion matrix for binary classification where p denotes positive and n denotes negative samples

		Prediction outcome	
		p′	**n′**
Actual value	**p**	True Positive	False Negative
	n	False Positive	True Negative

true positive (TP), false positive (FP), false negative (FN) and true negative (TN) scores can be obtained. TP represents the correctly classified positive instances, TN represents correctly classified negative instances, FN represents incorrectly classified positive instances, and FP represents incorrectly classified negative instances.

Using the confusion matrix, we calculated accuracy, precision, and recall by using the equations 4.1, 4.3, and 4.2, respectively. Both the precision and recall metrics give information on positive values. So, none of them can be used on their own for the assessment task. Of the simplicity of its nature, the accuracy metric is the most commonly used classifier assessment metric. However, when the class distribution is not even between classes, the accuracy metric would give incorrectly high results.

$$Accuracy = (FP + TN)/(FP + TP + TN + FN) \tag{4.1}$$

$$Recall = TP/(TP + FN) \tag{4.2}$$

$$Precision = TP/(TP + FP) \tag{4.3}$$

Due to the imbalanced nature of our data, we decided to use metrics that are robust to class imbalance. The most popular ones are the receiver operator characteristics (ROC) curve and the precision-recall (PR) curve. ROC curve gives the false positive rate (FPR) 4.4 vs true positive rate (TPR), which is the same as recall, 4.2 for each given class separation threshold value. Each point on the ROC curve represents correctly classified positive instance value for the falsely classified negative instance value.

$$FPR = FP/(FP + TN) \tag{4.4}$$

In other words, the ROC curve does not favor a specific class, rather it can be used for imbalance class score assessment. On the other hand, the PR curve favors the positive class [53]. The PR curve represents precision vs recall for each given class separation threshold value. Since TN is not included in the PR curve, it can give how well the classifier separates the positive instances more accurately than the ROC curve for imbalanced data that has low positive instances. In the real-estate aesthetics problem, if we were to assign an aesthetically pleasing image to a negative class, it would not matter as much. However, if we were to assign an aesthetically unpleasing real-estate image to a positive label, then the user or possible buyer of the real estate property would change their mind. Hence, using the PR curve and favoring the positive label would give a more accurate classification assessment. For each of the binary classifiers, we calculated the PR-AUC score, which is the area under the precision-recall curve. It gives an overall comparison metric for each classifier. For the 10-class classification tasks, we adapted the one-vs-rest approach, where we calculate the PR-AUC score for each class, and then take the average to obtain the final score.

4.2 Deep Learning Approaches on Real Estate Aesthetics Classification

After classical features, we benefited from the high-level features obtained from several deep learning models. However, we need to consider a few aspects of loading and processing the aesthetics data during the training of deep learning models.

4.2.1 Preprocessing, Augmentations and Loading of Vision Data

Deep learning methods require the data to be given to them in an adequate format. The image input should be modified for the desired problem. If we want to detect certain things in an image, we can enhance those parts and omit others. If our data is unvaried, we can use augmentations to enhance data. Before we start our model definition and training, we have to pay attention to how we give data to our model. How we give our image to the model is crucial for it to have to produce eligible results.

Preprocessing of image data generally includes color space transformations, geometric transformations, random image manipulations, etc. It modifies the data content to increase the size of the data and to bring all the data to the same ground for the model to fit better. We perform data augmentations by making minor modifications to our data to enrich its diversity. Many perform data augmentations by random cropping, horizontal and vertical flipping, random rotating, modifying color space, etc. However, we should be careful while implementing those kinds of augmentations since the core of our problem is image aesthetics and those alterations might lose the aesthetic value of the image. For instance, if we randomly crop out the salient object in an image, the remaining sub-part would not have the same aesthetic score as the original image or if we normalize the image in RGB color space, we would lose the vibrant colors. In 4.1, some suitable and unsuitable data augmentations are shown on a sample image from RAAD. As we can see from 4.1 suitable image augmentations would not cause a loss of aesthetic data, however, unsuitable image augmentations cause the loss of aesthetic value in the image.

Therefore, for the remaining parts of the deep learning methods, we tried to preserve the image quality and aspect ratio. We only used horizontal flip and padding to keep the aspect ratio and prepare the image for the model and would not make unsuitable augmentations on the image so that we can get the best results most simply.

Deep learning models take the data as batches for training. After applying the proper preprocessing and augmentations, batch image data is sampled from the dataset. However, the data score distribution of data might not be equal in most cases. There-

fore, to maintain the learning phase commeasurable for all classes, we used Weighted Sampler. It takes the class distribution and treats them as weights. By using the weight of each class, the sampler takes an even number of samples from each class for each batch during training. By using a weighted sampler, we ensured to learn each class equally.

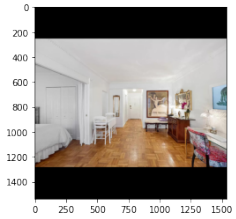
4.2.2 Deep Learning Classifiers

For aesthetics classification using deep learning, we decided to use several Convolutional Neural Network architectures that proved to be successful for image classification. Mainly, we used DenseNet [54], ResNet [55], VGG [56], and Vision Transformers [57]. DenseNet, ResNet and VGG architectures are well-known in the field of image classification and have been adapted by others for aesthetics classification as well as mentioned in 2.1.2. However, we have not yet seen an adaptation of a vision transformer in the aesthetic assessment context. Transformers are getting a lot of attention due to their high success rate in many domains starting with NLP. Vision transformers on the other hand are the variation of classical transformers to the image domain. They became quite popular starting with [57] and were used for various image-related tasks.

A transformer neural network architecture is composed of encoder-decoder modules using attention layers along with positional information. For NLP tasks, they produce prominent results. It only makes sense because the attention module focuses on the important words in a sentence and the positional information keeps track of the words' order in the sentence. Therefore, while doing translation tasks, transformers are used as a de facto standard. For vision tasks, [57] used images as 2D sentences and tried to implement the classical transformer architecture [58] as-is. The aesthetic assessment task, in its nature, resembles the semantics of vision transformers, we need to process an image as a whole. As mentioned in the 2.1 neuroaesthetics section, the human brain looks for the parts of an image to decide upon their aesthetic judgment on it as vision transformers learn by dividing the image into patches and feeding them to the network. Also, as mentioned in 2.1, the photographic principles and image composition features are shown to be important in terms of image aesthetics. The

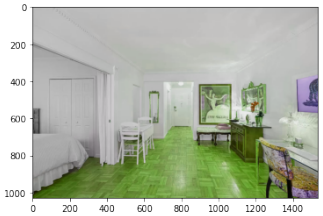
(a) Original Image

Horizontal flip

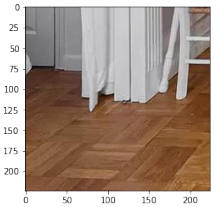


Padding

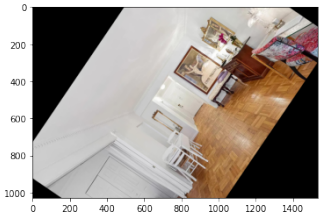
(b) Suitable image augmentations



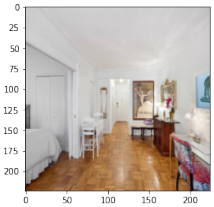
Random color jitter



Random crop



Random rotation



Resizing

(c) Not suitable image augmentations

Figure 4.1: Result of image augmentations.

positional embeddings of the image patches that we feed into the network can be classified as a synonymous method to measure those principles. That is why we believe that a vision transformer is a good solution for the image aesthetic quality assessment problem. In this study, we implement the methods they suggest in [57] to feed the images to the transformer encoder module.

The embedding of the image consists of patch embedding and positional embedding as in Equation 4.5.

$$\mathbf{z}_0 = \mathbf{E}_{patch} + \mathbf{E}_{positional} \tag{4.5}$$

We delve into more detail about how we calculate patch embeddings and positional embeddings since they are the learnable parameters given to the transformer encoder. Patch embeddings are constructed by dividing the image into fixed-size patches and putting them under a linear projection layer. We then add a classification token to the resulting patch embedding matrix. Positional embeddings are constructed by assigning an id number to each patch to keep track of the position of each patch. As in [57], we choose to use 1-D position embeddings by giving numbers to each patch from the top left corner to the bottom right starting from one. The final image embedding is given to the transformer encoder. In 4.2, the embedding construction pipeline is explained visually using a sample image from RAAD.

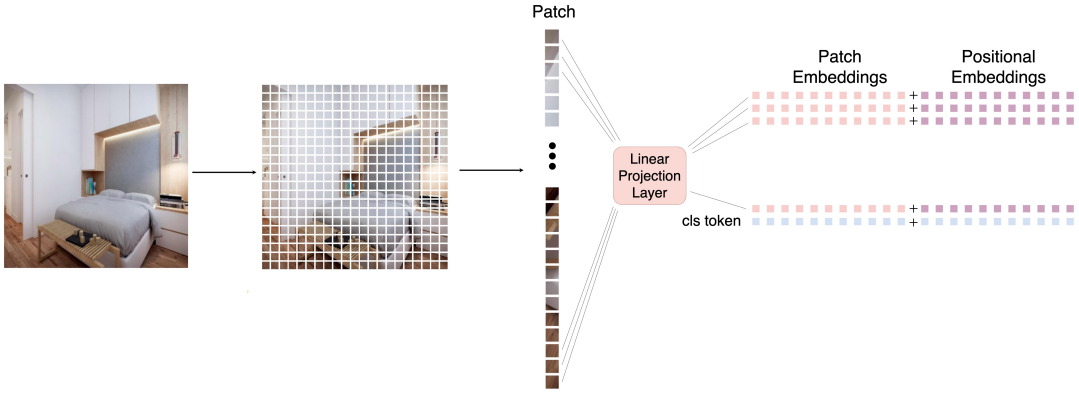


Figure 4.2: Image embedding construction pipeline

After the image is transformed into an embedding vector, the transformer encoder learns this vector by Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) modules. The ordering of these modules is visualized in 4.3. The MSA module gives us the attention mechanism to understand the contents of the image based on

the image patches we provide. That attention mechanism then gives us the representation of the image containing the important areas. The resulting vector goes through a linear layer, MLP Head and the classification results are obtained.

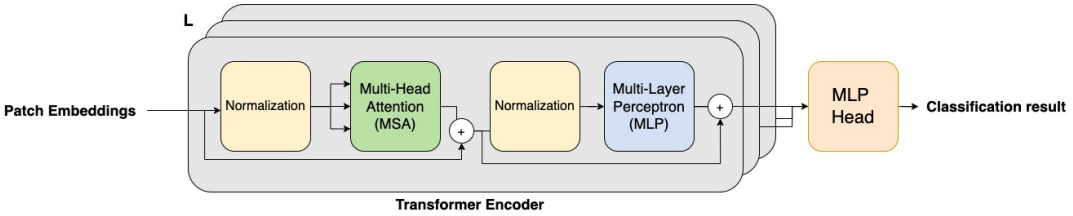


Figure 4.3: Vision transformer architecture

Hence, we experimented with a vision transformer for the aesthetics classification as well as previously tested deep learning architectures, ResNet, DenseNet, and VGG. Together with those architectures, we also decided to test an existing network specifically trained for image aesthetics problem to observe how well existing aesthetics networks work with RAAD data. We chose the MPada network [13]. For each deep learning classifier, we used the same evaluation metrics used for classical classifiers.

4.2.3 Fine-Tuning Approaches

The real estate aesthetic assessment can be considered as a sub-domain for general image aesthetics. Thus, we treat this problem as a domain adaptation problem. However, the size of RAAD might be too small for heavy deep learning models; it has a much smaller scale than AVA. Therefore, we used pre-trained models on AVA and implement fine tuning on them.

Fine-tuning, in the context of deep learning, is taking a model that has already been trained on a problem, the aesthetics assessment in this case, and making minor changes to the model, so that the model could learn the new task. We took the models that are already trained on AVA for the image aesthetic quality assessment task. Then, we freezed all the layers on the model, except for the last layer that we trained with RAAD. In this way, the model would have already learned the general aesthetic properties in the first layers and we would benefit from that information. Also, the model will learn the real estate-specific properties in the deeper layers during the fine-tuning

Table 4.3: Train - Validation - Test Splits of AVA Dataset

	Train	Validation	Test
AVA (full data)	153318	51106	51106
AVA (after eliminating ambiguous data)	34592	11533	11532

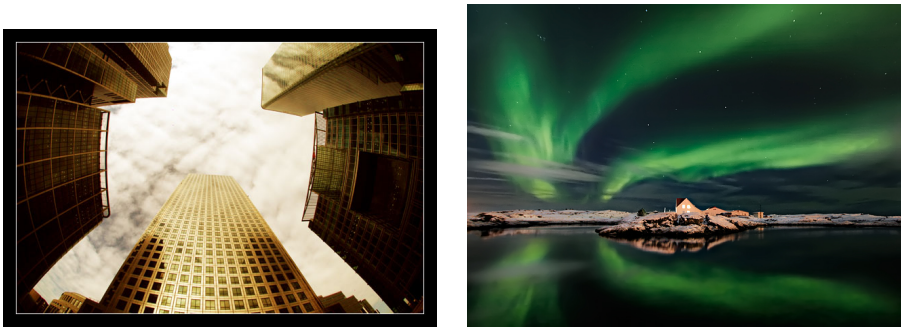
training process. Fine-tuning is beneficial for us since the size of the dataset is much smaller than AVA and instead of re-training and overfitting on small data, we can make use of the learned aesthetic properties of AVA.

To train models on AVA, we use ambiguous data elimination as suggested in [1], we take the gate parameter δ and define the low aesthetic scores as average scores that are less than 5 - δ and high average scores as an average score greater than 5 + δ. If the average score of the image is between [5 - δ, 5 + δ], then that image is marked as ambiguous. If we increase δ, the unambiguity increases and we might get better results, nevertheless, we lose some data and modify the train set based on our needs. [1] argues that we get better results with higher values of δ and we only eliminate ambiguous data on the train set and leave the test set as it is so that we would not introduce a bias to our model. The ambiguous and unambiguous image samples are given in 4.4. All images would be labeled as aesthetically pleasing but the top row images have an average score distribution between [4, 6]. Whereas, the bottom row has an average score of more than 6. The subjective nature of aesthetic analysis shows itself in those images, and since our purpose is to find a general model for aesthetic analysis, we decide to skip ambiguous images. So, for the experiments on AVA, we selected $\delta = 1$ as aforementioned to achieve a high classification score with less training time. The number of images for using full data and eliminating ambiguous data by each training, validation, and test split is given in 4.3.

4.3 Synthetic Real Estate Image Analysis

The synthetic data in RAAD is controlled data, meaning that we have heuristic knowledge of which exposure, saturation, and angle is optimal. We also have the user scor-

(a) Ambiguous image samples



(b) Unambiguous image samples

Figure 4.4: Ambiguous and unambiguous samples from [1].

ing for those images.

We first aim to incorporate synthetic data into the real estate aesthetics assessment process in the same way as real images in the dataset. We select the aesthetic score of each image through the data gathered with the user study.

Initially, we analyzed the correlation of image scores with the synthetic information. We divide synthetic information into 3 categories; exposure, saturation, and zoom level. Then, we evaluate the user score given to the image with respect to those categories in order to compare the aesthetic preferences of users with respect to photographic principles.

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 Analysis of RAAD

RAAD is one of the main outcomes of this thesis. The dataset consists of 2071 images; 991 from real images and 1080 from synthetic images. At present, there is a total of 87 user sessions with 3915 scores. Out of 87 sessions, 32 of them passed the user validation test and 55 of them failed. However, due to the low number of scores, we included all user scores for the assessment. The score distribution for both real and synthetic image scores from 1 to 10 are represented in 5.1. The binary score distribution is given in 5.2. For the real part of the dataset, there is an average of 2.08 votes and for the synthetic part, there is an average of 1.87 votes per image.

From 5.1, we can observe that real image scores gathered more towards average scores whereas synthetic image scores tend to be more around lower scores. The reason for the synthetic images having a lower average might be that the shooting angle of the camera is sometimes directed at the corners of the room, resulting in not-desired outcomes. Another reason could be that some of the synthetic images have unrealistic renderings, resulting in low scores for the aesthetic assessment. Also from 5.2, we can observe that real images have a more even distribution between aesthetically pleasing and unpleasing whereas synthetic images have dominantly unpleasing images. Considering the uneven distribution of scores, we can notice the class imbalance problem. In order to analyze the results of classifiers, we first need to define the baseline classification score, i.e how a random classifier performs under RAAD data based on PR-AUC metric, for binary classification for both synthetic and real image sets. The 10-class classifiers have the baseline score of 0.1 since we adapt the one-

vs-rest approach to calculate the PR-AUC score. For the binary classification case, the real image set has a more even distribution where 561 images are aesthetically unpleasing and 430 images are aesthetically pleasing. By using the image count in each set as class weight, the approximate class weights are 0.56 and 0.44 for aesthetically unpleasing and aesthetically pleasing respectively. Using formulas 4.3 and 4.2, the calculated PR-AUC score for synthetic data would be 0.09 and 0.22 for real image set. So, we can say that a random classifier would give 0.09 PR-AUC score for synthetic images and 0.22 PR-AUC score for real images for the binary classification task on RAAD.

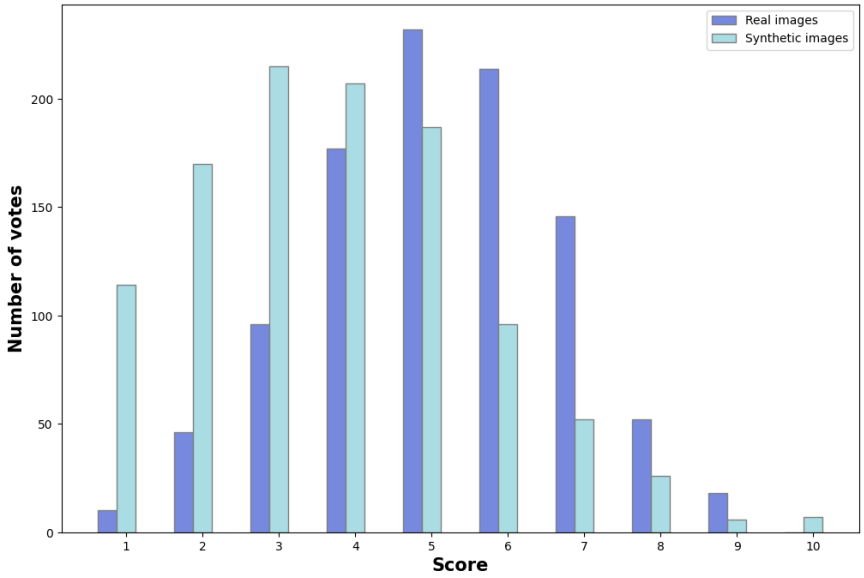


Figure 5.1: Image 10-class score distribution

5.2 Analysis of Synthetic Image Scores

For each yaw-pitch angle combination of synthetic images of RAAD, there are 2 different exposure values, 2 different saturation values, and 3 different zoom levels. After gathering the results from the user study, we analyzed the user aesthetics preferences and different parameters of synthetic images. However, since the per-image score count is around 1.87, the average score distribution for all cases is mostly even. In 5.3 we can also observe the even distribution for each parameter. Therefore, we cannot make any judgments about the effect of the parameters on the image aesthet-

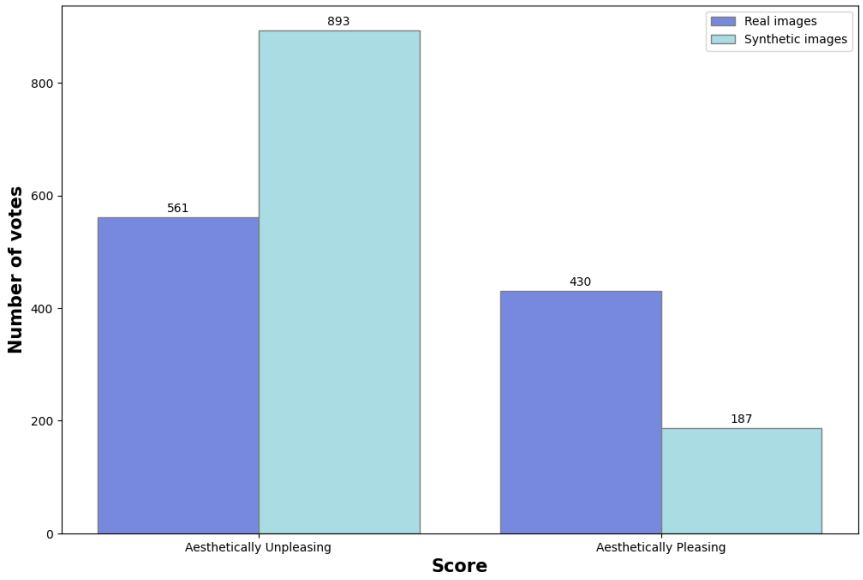
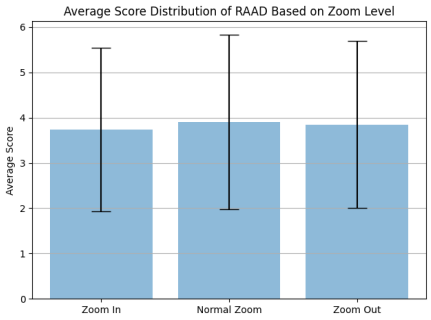


Figure 5.2: Image binary score distribution

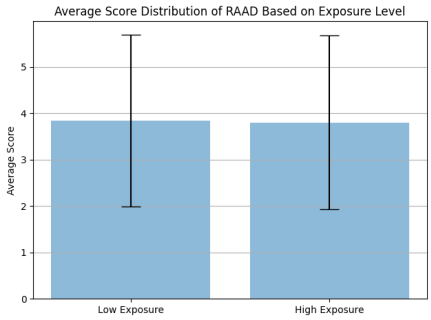
ics score as of this statute of the dataset. Even though our preliminary assumption for this experiment was that the user would prefer brighter and saturated images with normal camera zoom angles. However, the results showed that with the given values of exposure, saturation, and camera zoom level, users do not have a dominant preference. The reason for these results might be from the camera yaw-pitch angles. Since the user is subjected to synthetic images in a randomized way, most of the time they face unpleasant shooting angles of the scene, and the shooting angle, i.e. different yaw-pitch value combinations, affect the aesthetic decision more than the slight differences of saturation, exposure, and zoom level.

5.3 Handcrafted Feature Results

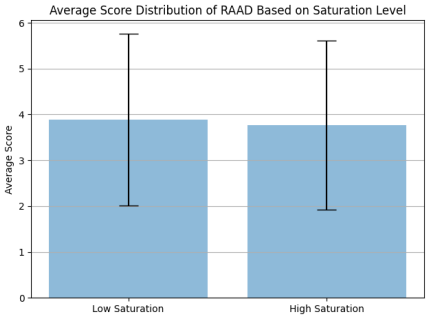
Low-level image features are selected as image mean, standard deviation, hue, saturation, brightness, energy, GLCM contrast, GLCM homogeneity, GLCM correlation, and GLCM energy. For each part of RAAD, we extracted all of the average values of handcrafted features. Initially, we check the correlation of low-level image features with 10-class image aesthetic values using both Spearman and Pearson analysis methods. In 5.4, the first row shows the correlation results of Spearman analysis and the second row shows the correlation results of Pearson analysis. Then, we used Spear-



(a) Different Zoom Levels

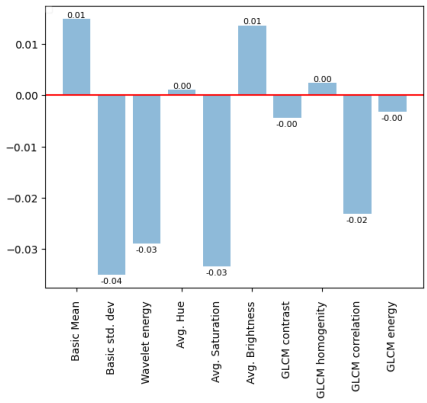


(b) Different Exposure Levels

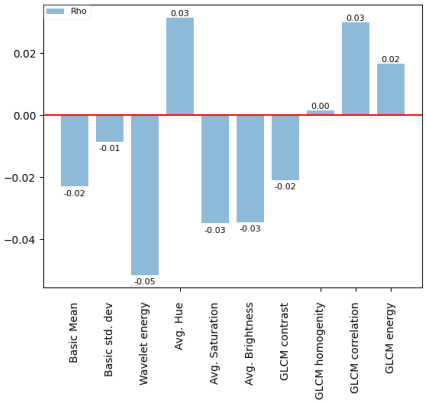


(c) Different Saturation Levels

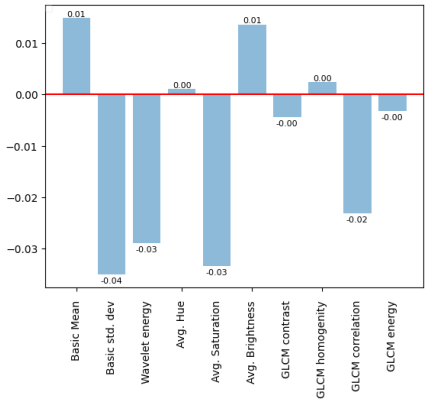
Figure 5.3: Comparison of average scores for (a) different zoom levels, (b) different exposure levels and (c) different saturation levels



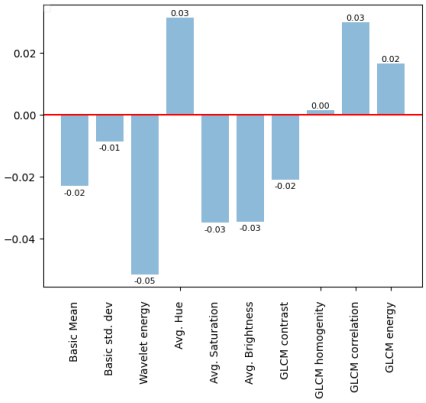
(a) Spearman analysis with synthetic part of RAAD



(b) Spearman analysis with real part of RAAD



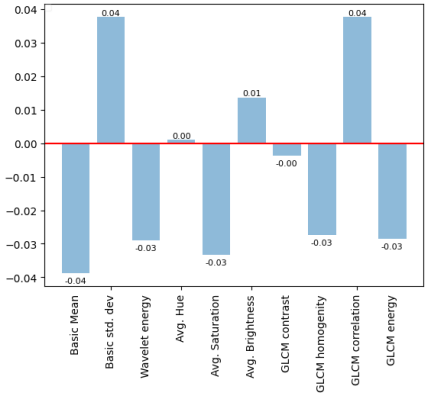
(c) Pearson analysis with synthetic part of RAAD



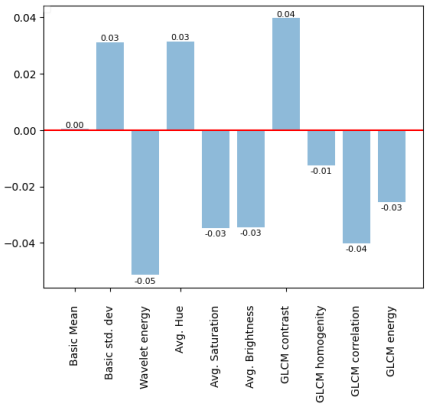
(d) Pearson analysis with real part of RAAD

Figure 5.4: Correlation analysis of 10-class aesthetic values with handcrafted features

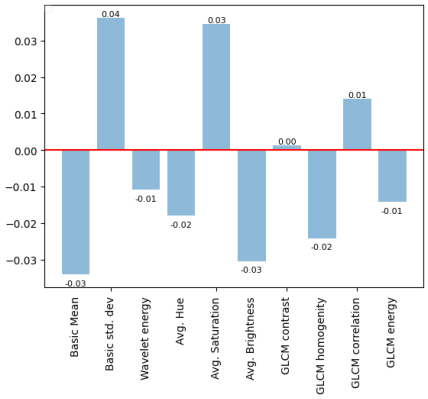
man and Pearson analyses for binary classification, by assigning a score to the image either 0 or 1 based on the given average user score. For the binary image, labels feature correlation results are shown in 5.5. As though the correlation coefficient for both Spearman and Pearson have increased compared to the 10-class classification, none of the features have a considerable value regarding the aesthetics information. Usually, the preferred threshold for the correlation coefficient is to be greater than 0.6, whereas the maximum coefficient we can obtain with these features is 0.06. So, we decided not to use those low-level features for aesthetics analysis.



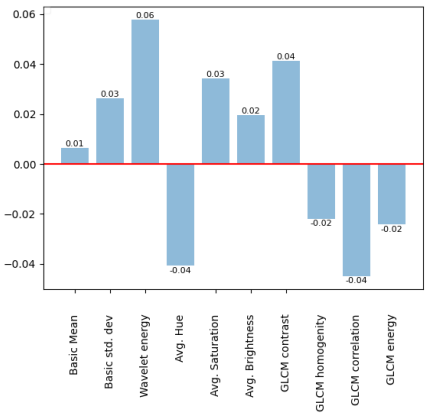
(a) Spearman analysis with synthetic part of RAAD



(b) Spearman analysis with real part of RAAD



(c) Pearson analysis with synthetic part of RAAD



(d) Pearson analysis with real part of RAAD

Figure 5.5: Correlation analysis of binary aesthetic values with handcrafted features

5.4 Analysis of Classical Classifier Results

Since averaging low-level features would not yield satisfactory correlations, we decided not to use those for classical image classifiers. Instead, we decided to use image features as whole vectors without summation. To do so, we convert images to the HSV color space and use each channel as a feature vector for the image classifiers. We trained four different classical image classifiers; Support Vector Machines (SVM), Linear Regression (LinReg), Logistic Regression (LogRes), and Ordinal Logistic Regression (OrdLogRes). In this section, we analyzed the result of each classifier for each HSV channel along with the combined feature vector. We evaluated each part of RAAD separately for both 10-class classification and binary classification.

The 5.1, 5.2, 5.3, and 5.4 tables represent the classification result. The first column gives the overall accuracy, and the other columns give the weighted average of precision, recall, and PR-AUC score respectively. The weighted average of these scores is calculated by taking per-class scores and multiplying them with the ratio of the respective class. In 5.1 and 5.2, we can observe the 10-class classification results for synthetic and real data respectively. For synthetic data, logistic regression has the highest PR-AUC score, and for the real data ordinal logistic regression has the highest PR-AUC score. In 5.3 and 5.4, we can observe the binary classification results for synthetic and real data respectively. For the synthetic data, the SVM classifier has the highest PR-AUC score, and for the real data linear regression and SVM have the top two best PR-AUC scores.

5.5 Analysis of Deep Learning Results

For the analysis of the high-level image features of RAAD on aesthetics classification task, we used DenseNet121 [54], ResNet18 [55], VGG16 [56] and ViT [57] architectures. We trained separate classifiers for both parts of RAAD for 10-class classification and binary classification. Also, we included the test performance of an aesthetics network, MPada, to RAAD data to analyze how well an aesthetics network performs on our data. In the experiments, MPada model weights are trained on AVA, as the authors provide. For all the other deep learning models we train on RAAD, we

Table 5.1: Analysis of 10-Class Classical Classifiers on Synthetic Part of RAAD

	Accuracy	Precision(avg)	Recall(avg)	PR-AUC
SVM_{hue}	0.06	0.03	0.06	0.0999
$SVM_{saturation}$	0.13	0.04	0.13	0.1008
SVM_{value}	0.09	0.04	0.09	**0.1039**
SVM_{all}	**0.18**	**0.17**	**0.18**	0.1036
$LinReg_{hue}$	0.17	0.14	0.17	0.1000
$LinReg_{saturation}$	0.17	0.14	0.17	0.0999
$LinReg_{value}$	**0.23**	**0.17**	**0.23**	**0.1058**
$LinReg_{all}$	0.21	0.16	0.21	0.1039
$LogReg_{hue}$	0.11	0.10	0.11	0.0992
$LogReg_{saturation}$	0.14	0.16	0.14	0.1030
$LogReg_{value}$	**0.18**	**0.19**	**0.18**	**0.1073**
$LogReg_{all}$	0.13	0.13	0.13	0.1006
$OrdLogReg_{hue}$	0.17	0.13	0.17	0.1003
$OrdLogReg_{saturation}$	0.16	0.14	0.16	0.1001
$OrdLogReg_{value}$	**0.21**	**0.15**	**0.21**	**0.1045**
$OrdLogReg_{all}$	0.20	**0.15**	0.20	0.1027

Table 5.2: Analysis of 10-Class Classical Classifiers on Real Part of RAAD

	Accuracy	Precision(avg)	Recall(avg)	PR-AUC
SVM_{hue}	**0.18**	**0.16**	**0.18**	**0.1022**
$SVM_{saturation}$	0.14	0.14	0.14	0.0998
SVM_{value}	0.15	0.14	0.15	0.0991
SVM_{all}	0.12	0.11	0.12	0.0972
$LinReg_{hue}$	0.19	0.17	0.19	0.1022
$LinReg_{saturation}$	**0.21**	**0.19**	**0.21**	**0.1025**
$LinReg_{value}$	0.19	0.11	0.19	0.1001
$LinReg_{all}$	0.17	0.11	0.17	0.0981
$LogReg_{hue}$	0.15	0.16	0.15	0.0998
$LogReg_{saturation}$	**0.17**	**0.18**	**0.17**	**0.1029**
$LogReg_{value}$	**0.17**	0.15	0.17	0.1012
$LogReg_{all}$	0.15	0.15	0.15	0.1000
$OrdLogReg_{hue}$	**0.20**	0.17	**0.20**	0.1009
$OrdLogReg_{saturation}$	0.18	0.16	0.18	**0.1041**
$OrdLogReg_{value}$	0.19	0.17	0.19	0.1009
$OrdLogReg_{all}$	0.19	**0.18**	0.19	0.1008

Table 5.3: Analysis of Binary Classical Classifiers on Synthetic Part of RAAD

	Accuracy	Precision(avg)	Recall(avg)	PR-AUC
SVM_{hue}	0.73	0.71	0.73	0.1650
$SVM_{saturation}$	0.73	0.71	0.73	**0.2160**
SVM_{value}	0.71	0.70	0.71	0.1550
SVM_{all}	**0.75**	**0.72**	**0.75**	0.2030
$LinReg_{hue}$	0.83	0.68	0.83	0.1650
$LinReg_{saturation}$	0.83	0.68	0.83	**0.1770**
$LinReg_{value}$	0.83	0.68	0.83	0.1530
$LinReg_{all}$	**0.83**	**0.68**	**0.83**	0.1600
$LogReg_{hue}$	**0.77**	**0.72**	**0.77**	**0.1680**
$LogReg_{saturation}$	**0.77**	0.70	**0.77**	0.1610
$LogReg_{value}$	0.76	0.70	0.76	0.1470
$LogReg_{all}$	0.76	0.71	0.76	0.1640
$OrdLogReg_{hue}$	0.76	**0.73**	0.76	**0.1680**
$OrdLogReg_{saturation}$	**0.77**	0.70	**0.77**	0.1590
$OrdLogReg_{value}$	0.76	0.70	0.76	0.1470
$OrdLogReg_{all}$	0.76	0.71	0.76	0.1650

Table 5.4: Analysis of Binary Classical Classifiers on Real Part of RAAD

	Accuracy	Precision(avg)	Recall(avg)	PR-AUC
SVM_{hue}	**0.56**	**0.55**	**0.56**	0.4250
$SVM_{saturation}$	0.47	0.48	0.47	0.4260
SVM_{value}	0.54	0.52	0.54	**0.4860**
SVM_{all}	0.55	0.53	0.55	0.4370
$LinReg_{hue}$	0.56	0.32	0.56	**0.4880**
$LinReg_{saturation}$	0.56	0.32	0.56	0.4320
$LinReg_{value}$	0.56	0.32	0.56	0.4230
$LinReg_{all}$	**0.56**	**0.32**	**0.56**	0.4270
$LogReg_{hue}$	**0.55**	**0.54**	**0.55**	**0.4840**
$LogReg_{saturation}$	0.54	0.53	0.54	0.4270
$LogReg_{value}$	0.51	0.49	0.51	0.4090
$LogReg_{all}$	0.52	0.51	0.52	0.4140
$OrdLogReg_{hue}$	**0.56**	**0.55**	**0.56**	**0.4780**
$OrdLogReg_{saturation}$	0.55	0.54	0.55	0.4260
$OrdLogReg_{value}$	0.51	0.50	0.51	0.411
$OrdLogReg_{all}$	0.52	0.51	0.52	0.4150

Table 5.5: Analysis of 10-Class Deep Learning Image Classifiers on Synthetic Part of RAAD

	Accuracy	Precision(avg)	Recall(avg)	PR-AUC
DenseNet121	**0.16**	**0.27**	**0.16**	**0.0870**
ResNet18	0.13	0.15	0.13	0.0861
VGG16	0.11	0.14	0.11	0.0829
ViT_{b16}	0.14	0.16	0.14	0.0856

used ImageNet initial weights and performed the training on top of those weights.

In 5.5 and 5.6 the outcomes of 10-class classifications are given for both synthetic and real data respectively. For the synthetic data, DenseNet121 architecture gives the highest PR-AUC score. For the real data, DenseNet121 and VGG16 have both the highest PR-AUC scores. The 5.6 and 5.7 show that the DenseNet121 model has the highest training loss. Although it has a high training loss, its performance to separate class predictions is the highest.

In 5.7 and 5.8 the results of binary classification are given for both synthetic and real data respectively. Both or the synthetic part and real part, DenseNet121 performs the best. The 5.8 and 5.9 graphs shows that DenseNet121 has again the highest training loss, even though the it yields the best PR-AUC scores. The MPada model suffers from over-fitting on RAAD data as it assigns all images as aesthetically unpleasing. Hence, we can say that aesthetics networks cannot be used as is for RAAD assessment.

Comparing the training loss graphs of both 10-class and binary classification, we can conclude that for the 10-class classification, the data is not enough for deep learning models to fit. However, binary scores have a higher probability to fit with deep learning models. The details of 10-class classification and binary classification results are given as confusion matrices as well in 5.10 and 5.11.

Table 5.6: Analysis of 10-Class Deep Learning Image Classifiers on Real Part of RAAD

	Accuracy	Precision(avg)	Recall(avg)	PR-AUC
DenseNet121	0.15	0.05	0.15	**0.0990**
ResNet18	0.13	**0.26**	0.13	0.0980
VGG16	0.19	0.17	0.19	**0.0990**
ViT_{b16}	**0.20**	0.05	**0.20**	0.0975

Table 5.7: Analysis of Binary Deep Learning Image Classifiers on Synthetic Part of RAAD

	Accuracy	Precision(avg)	Recall(avg)	PR-AUC
DenseNet121	0.59	0.59	0.59	**0.2909**
ResNet18	0.62	**0.65**	0.62	0.2139
VGG16	0.65	0.64	0.65	0.1998
ViT_{b16}	**0.69**	**0.65**	**0.69**	0.2688
MPada	0.43	0.20	0.09	0

Table 5.8: Analysis of Binary Deep Learning Image Classifiers on Real Part of RAAD

	Accuracy	Precision(avg)	Recall(avg)	PR-AUC
DenseNet121	0.51	0.51	0.51	**0.4836**
ResNet18	0.46	0.46	0.46	0.4528
VGG16	**0.59**	**0.59**	**0.59**	0.4441
ViT_{b16}	0.53	0.52	0.53	0.4805
MPada	0.34	0.20	0.07	0

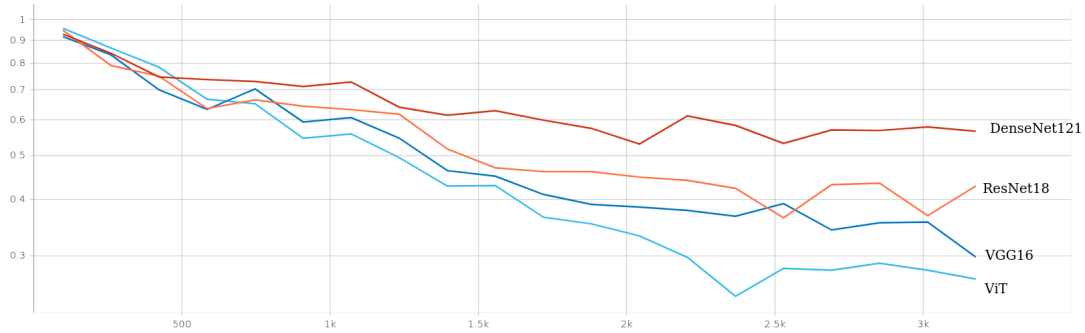


Figure 5.6: Training loss graph for each 10-class deep learning classifier for synthetic data

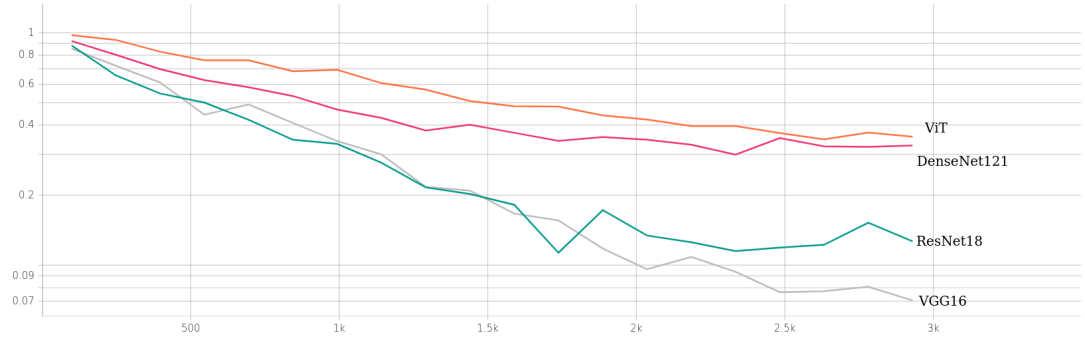


Figure 5.7: Training loss graph for each 10-class deep learning classifier for real data

5.6 Analysis of Fine-Tuning

As mentioned before, fine-tuning gives the advantage of transferring learned hidden layer parameters to another model. For the aesthetics assessment case, the AVA dataset has a much bigger data size than the RAAD. So, for this experiment, we selected the binary classification task as they suggested on [1] and the real part of RAAD. For the model, we selected ViT [57] architecture to inspect the domain knowledge transfer.

5.6.1 Training Vision Transformer on AVA

Initially, we trained AVA with unambiguous images using the split given in 4.3. The results are given as confusion matrix 5.12. The accuracy of this experiment is nearly 0.90.

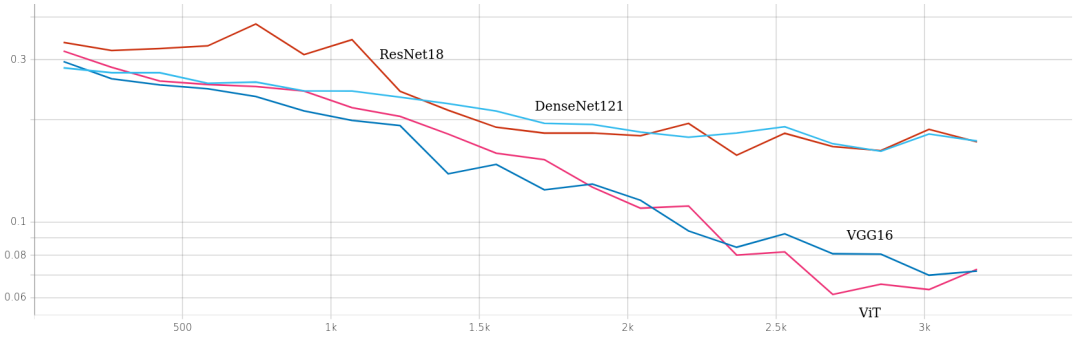


Figure 5.8: Training loss graph for each binary deep learning classifier for synthetic data

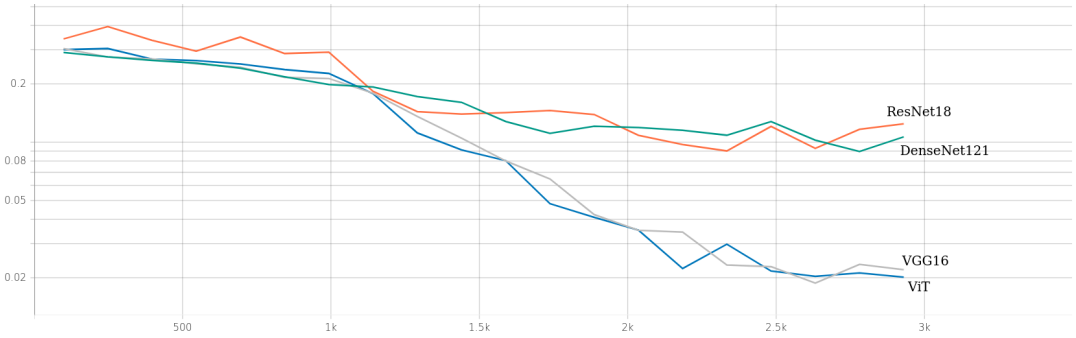
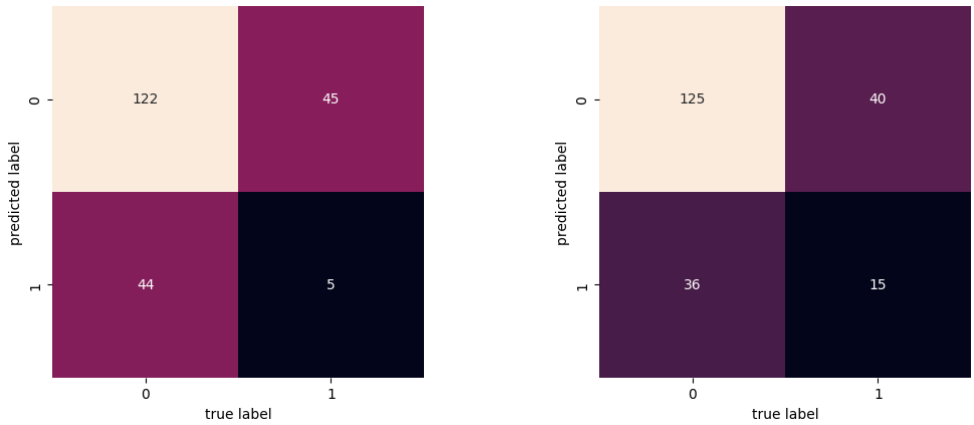


Figure 5.9: Training loss graph for each binary deep learning classifier for real data

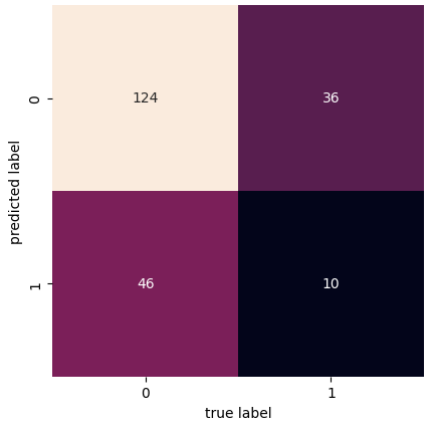
5.6.2 Fine-Tuning ViT with Real Part of RAAD

In the second part of the experiment, we load the ViT model trained on AVA from the previous stage and fine-tuned it using real images of RAAD. The analysis for both ViT outcomes, with and without fine-tuning are given in 5.9. From the table, we can observe that there is a slight improvement on the metrics. Also, if we inspect 5.13, we can see that ViT pre-trained on AVA reaches lower training loss much faster than ViT pre-trained on ImageNet. Considering that vision transformers are heavy models and require much time to train, we can suggest that using pre-trained models on aesthetics data would improve the training time for sub-domain problems, such as real estate aesthetics assessment.

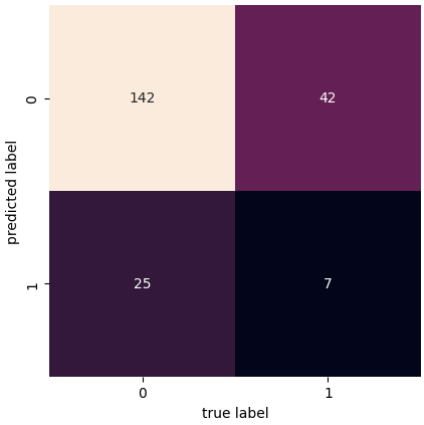


(a) DenseNet

(b) VGG16



(c) ResNet

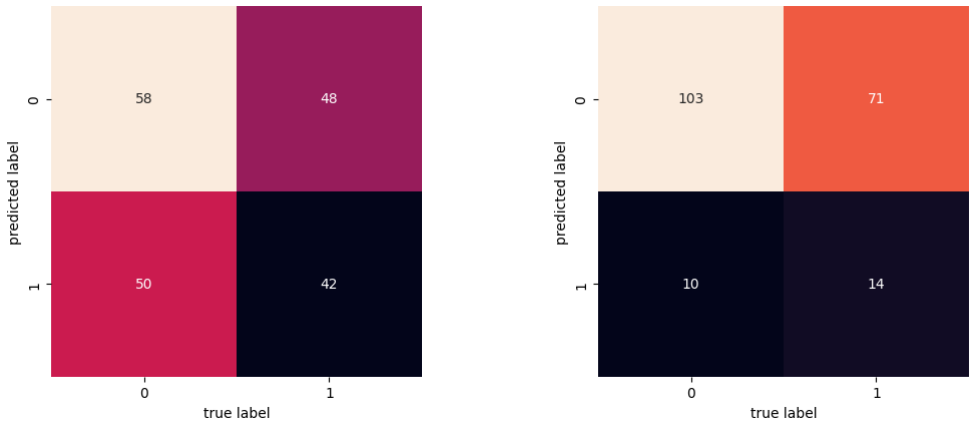


(d) ViT

Figure 5.10: Confusion matrices of binary deep learning classification results for synthetic data

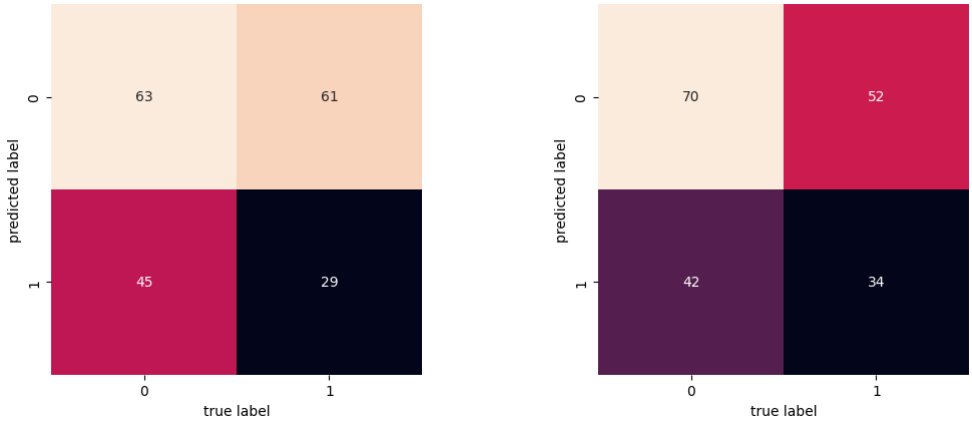
Table 5.9: Analysis of Fine-Tuning ViT on AVA for Real Part of RAAD

	Accuracy	Precision(avg)	Recall(avg)	PR-AUC
ViT_{b16} (Pretrained on ImageNet)	0.53	0.52	0.53	0.4805
ViT_{b16} (Pretrained on ImageNet + AVA)	0.54	0.54	0.54	0.4819



(a) DenseNet

(b) VGG16



(c) ResNet

(d) ViT

Figure 5.11: Confusion matrices of binary deep learning classification results for real data

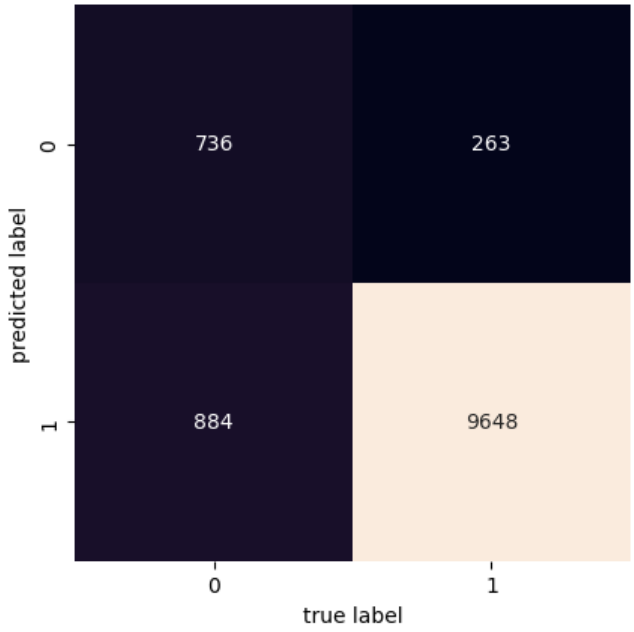


Figure 5.12: Confusion Matrix of ViT results trained on AVA

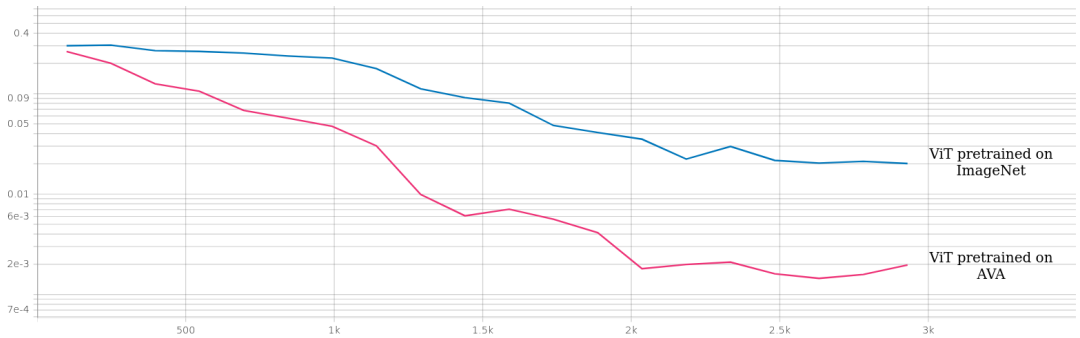


Figure 5.13: Training loss graph for ViT both pretrained on ImageNet and AVA

CHAPTER 6

DISCUSSION

6.1 Synthetic Data vs Real Data

Synthetic and real parts of RAAD have considerable differences in terms of image quality, realism, score distribution, and classifier performances. Our initial objective is to use synthetic data as an alternative addition to real data as the renderings are highly realistic. However, the user study showed that even though a portion of the synthetic data might be added, as a whole synthetic data cannot be of use as of its state. The reason for this can be examined under two objectives. The first one is related to the rendering quality. Based on the user feedback, some images contain unrealistic renderings which irritate the user and therefore result in a lower score, disregarding the scene, shooting angle, exposure, saturation, or zoom level parameters. The second one is related to the automated scene rendering process. To automatize this process, we stabilized the camera in the center of the scene and we included different pitch-yaw angle combinations for each parameter. For some scenes, the center coordinates of the scene don't map to the center of the room, instead, it's located near a wall. This causes some angle combinations, to produce undesirable outcomes such as only the wall, the curtain, or only the corner of the room rendering. Considering these, synthetic data having a lower score distribution than the real part checks out. However, if the aforementioned problems might be studied further, synthetic data might be an addition to the real part in the future.

6.2 Binary vs 10-Class Classification

Our initial aim is to assign a precise aesthetic score for the image using regression. However, the user study required much more time and scope so that each image would have an average minimum of 78 scores just as in AVA [1]. As of the current state, each image has 1.9 scores on average. So, we couldn't assign a precise score for the image, rather, we used the given scores as discrete labels. With this data, we decided to use 10-class classifiers and binary classifiers simultaneously. For the case of the 10-class classification, most of the PR-AUC scores cannot surpass the average classifier result, 0.1, with the highest 0.1073. This is because we do not have enough data for each class to represent the respective class. So, none of the classifiers can learn the 10-class separation features; all of the classifiers fail to distinguish different aesthetic score classes. For the binary classification case, data distribution is much more balanced than in 10-class classifiers. This results in better performance on the PR-AUC scores. All of the binary classifiers for both the synthetic and real data have surpassed the average classifier performance. Overall, we can conclude that the binary classifiers perform much better than the 10-class classifiers because PR-AUC scores are much higher than the average classifier for the binary classification task.

6.3 Handcrafted Feature Analysis

Through the study, 10 different handcrafted features are extracted from images. In the experiments, we used, image mean, standard deviation, energy, hue, saturation, brightness, and image texture with GLCM; GLCM contrast, GLCM homogeneity, GLCM correlation, and GLCM energy. For each of these features, we used the average value. Then, to calculate the correlation of these features with the aesthetic scores, we used both Spearman's and Pearson's correlation analysis techniques. However, none of them shows a significant correlation in both analyses and we decided not to use any of the average handcrafted features in our classifiers. Considering the aesthetics domain, it is nearly impossible to find a general, overall feature that represents the aesthetic value of the image. However, as we study an approximation to the general aesthetic heuristic, the features we investigate in this study don't produce informa-

tion regarding our problem. Conceivably in the future, other features that represent the photographic principles, which generally produce high aesthetic images, might be incorporated into our study. Also, with further user score gathering, the correlation analysis can be repeated. As of its current state, an image from the dataset has a highly subjective score. With more user scores, the average image score would represent the objective image aesthetics more reliably.

6.4 Performance Comparisons of Classical Classifiers

In order to assess the aesthetic value of an image, we trained 4 different classical classifiers; SVM, linear regression, logistic regression and ordinal logistic regression, using HSV representation of the image. For all of the models, we used principal component analysis to choose the most informative features and used them for training. Comparing the classical classifiers, logistic regression and ordinal logistic regression perform well for 10-class classification but has the lowest scores for binary classification. On the other hand, SVM and linear regression performs the top two for binary classification but have the lowest scores for the 10-class classification. So, we can conclude that for the 10-class classification task logistic regression has better performance. For binary classification, linear regression tends to have an overfitting problem as its precision values are among the lowest. So, SVM would yield the best outcome for the binary classification task.

6.5 Performance Comparisons of Deep Learning Classifiers

Through this study four different deep learning classifiers are evaluated under aesthetics problem on real-estate data. The DenseNet architecture is selected initially to gather information from each layer using dense connections. The ResNet architecture is selected to not lose the residual aesthetic information through training. VGG network is a simple convolutional neural network architecture that uses 16 convolutional and fully connected layers. After selecting those prominent architectures by considering those initial rationales, we also decided to adapt a vision transformer architecture to the aesthetics domain since vision transformers perform promising results in terms

of image classification. Also, we included an already existing aesthetics network, MPada, trained on the whole AVA dataset, into our experiments to evaluate its performance under RAAD data. The existing weights of MPada failed to perform under RAAD data as all the test images are assigned as aesthetically unpleasing. So, we can say that for the already existing networks on image aesthetics, we cannot use them as is. Among the remaining four classifiers, DenseNet121 has the highest PR-AUC score among all tasks, and VGG16 has the lowest performance among all. We can inspect these results by the relation between the network size and our dataset size. When the dataset size is small, like RAAD, heavier networks tend to have an overfitting problem, as their training loss decreases and memorizes the data. On the other hand, smaller networks perform better for the small-scale dataset. VGG16 network is the biggest network among them with 18 million parameters and DenseNet121 is the lightest one with approximately 7.6 million parameters. Also, considering vision transformers require a much larger data size than RAAD, the ViT network has promising results. With increasing RAAD size and user scores, the transformer model might perform much better.

6.6 Performance of Fine-Tuning

Since, deep learning models require large amount of data to produce meaningful results, using pre-trained networks would help the models to converge faster. So, all of the deep learning networks in this research, are used with pre-trained weights on ImageNet [59], and then trained on RAAD data to produce the given results. However, ImageNet data is solely for image classification tasks and not for aesthetics assessment. Hence, we take the vision transformer model with pre-trained weights on ImageNet and trained it again on AVA, and use that model weights for training with RAAD. The comparison of both ViT models, pre-trained on ImageNet and pre-trained on ImageNet + AVA shows that there is a small change in the performance. However, we can still conclude several more outcomes from this experiment. First of all, the slight increase in accuracy shows us that the real part of RAAD can be used as a subset for aesthetics since the learned aesthetics features from AVA is compatible with learned features from RAAD. Secondly, when we check the training graph, we

can observe that the pre-trained on AVA model required less time to converge for the best fit than the model pre-trained on ImageNet. So, by using the fine-tuning approach we can obtain equivalent results in a shorter time.

6.7 Classical Classifiers vs Deep Learning Classifiers

Both classical and deep learning classifiers studied in this research use images without modifying their content and are examined under the PR-AUC score for comparison. If we compare the accuracy of best-performing classifiers, even though 10-class classification produces unfavorable results, classical classifiers would have higher scores than deep learning classifiers. For the binary classification task, classical and deep learning classifier performances are close to each other with the deep learning classifier being slightly better. Even though the dataset size is small, deep learning classifiers can still produce meaningful class separation scores. So, this might imply that with increasing data, deep learning classifier performance would also increase. All in all, with the current state of the problem, classical classifiers would be an easier solution to the classification problem by producing results as good as deep learning classifiers. However, increasing the dataset size would make deep learning classifiers a much preferable option as their performance would increase with bigger data.

CHAPTER 7

CONCLUSION

Aesthetics assessment is a highly subjective terminology. However many different research have been done for many different areas. In this thesis we analyzed the image aesthetics assessment in the context of real estate data. In order to accomplish this, we initially constructed RAAD; Real-Estate Aesthetics Assessment Dataset. RAAD data is separated into two main sections; real data and synthetic data. Real data is gathered from recent real-estate web pages. Whereas synthetic data is rendered using [36] with controlled parameters. Then, in order to gather the aesthetics score for images, we constructed a user study webpage, where users are given a set of images and asked to rate their aesthetic value regarding their appeal to the real estate. Along with this web page, RAAD is one of the main outcomes of this thesis.

Gathering the user scores took a long time. So, after obtaining approximately 1.9 votes per image, we started our experiments. Statistics analysis on the dataset showed us that the real image scores have more natural distribution but user scores of synthetic data tend to be gathered around lower scores. This was expected, since generated data have unrealistic portions that might irritate some people. Then we analyze the user scores given the synthetic data only, to analyze how different saturation, exposure and zoom level affect the aesthetic value. However, the results showed that there is not a significant difference in users' perspective regarding those parameters.

Aesthetics assessment is analyzed initially as a 10-class classification problem then as a binary classification problem. We conducted our experiments separately for real part and the synthetic part of the dataset; since as realistic as the synthetic data are, they are not authentic and combining them both would yield confusing results. Initially, we decided to extract low-level image features for classical classifiers. We

used image mean, standard deviation, energy, hue, saturation, value,i.e brightness, and texture information. To obtain texture information we used GLCM contrast, homogenity, correlation and energy. In order to select informative handcrafted features, we subjected them to Spearman's and Pearson's correlation. However, none of them contained informative value regarding image aesthetics. So, we decided to train classical classifiers with HSV color space values separately and all. For 10-class classical classifier logistic regression has a considerable performance. For binary classification, SVM has best performance. After classical classifiers, we trained several deep-learning classifiers as well. Among deep learning classifiers, DenseNet121 has the best performance for both the 10-class and binary classification. Also, we analyzed the performance of a pre-trained aesthetics network performance on RAAD by testing MPada on our test set. The network performed poorly as it suffered from over-fitting on all test images by assigning them all as aesthetically unpleasing. So, using a pre-trained aesthetics network as-is would not be suitable. Lastly, we investigated the fine-tuning of real-estate data. AVA dataset has a large collection of aesthetic data. So, we selected ViT architecture, as vision transformer architectures have not yet been used in the aesthetics domain, and trained it with the AVA dataset. We used AVA labels mentioned in [1]; by eliminating ambiguous data and assigning 0 or 1 as an aesthetics score to each image. The result of the binary classification is 0.9. Then, we used this model and trained it again with RAAD. The results showed a minor improvement. Also, the model pre-trained reached those results in a much shorter time. This would show that RAAD could be a subset of AVA and by using models pre-trained on AVA, we can reduce the training time for smaller subsets of aesthetics data without losing accuracy.

7.1 Limitations and Future Work

Aesthetics is a highly subjective topic. So, as though we gave a considerable amount of time to gather user data, to achieve more unbiased scores, there should be more user data. In AVA, the number of user scores per image is between 78 to 549 [1]. So, each image in RAAD should have that many scores as well. RAAD stores the user scores in the same structure as [1], so that in the future as the dataset expands, RAAD

might be a category for the AVA dataset.

The size of the dataset also limits us from obtaining good results from deep learning classifiers. Deep learning classifiers require a large amount of data and training them with nearly 1000 images would not show their true potential for this topic.

The ViT model trained on AVA has promising results. For future analysis, vision transformers could be evaluated solely on the AVA dataset and could be compared with other state-of-the-art AVA papers. As vision transformer architecture has concepts similar to aesthetics assessment, custom attention modules could also be implemented to enhance vision transformer models to figure out aesthetics scoring problems.

REFERENCES

[1] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis,"

[2] D. D. Cinzia and G. Vittorio, "Neuroaesthetics: a review," *Current Opinion in Neurobiology*, vol. 19, no. 6, pp. 682–687, 2009. Motor systems • Neurology of behaviour.

[3] C. Wald, "Neuroscience: The aesthetic brain," Oct 2015.

[4] V. Ramachandran and W. Hirstein, "The science of art: A neurological theory of aesthetic experience," *Journal of Consciousness Studies*, vol. 6, pp. 15–51, 01 1999.

[5] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, pp. 80–106, jul 2017.

[6] Z. Dong, X. Shen, H. Li, and X. Tian, "Photo quality assessment with dcnn that understands image well," pp. 524–535, 01 2015.

[7] H. Lv and X. Tian, "Learning relative aesthetic quality with a pairwise approach," in *MMM*, 2016.

[8] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 990–998, 2015.

[9] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 497–506, 2016.

[10] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," 2017.

[11] N. Murray and A. Gordo, "A deep architecture for unified aesthetic prediction," 2017.

[12] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, pp. 3998–4011, aug 2018.

[13] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu, "Attention-based multi-patch aggregation for image aesthetic assessment," in *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, (New York, NY, USA), p. 879–886, Association for Computing Machinery, 2018.

[14] V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," 2019.

[15] D. Liu, R. Puri, N. Kamath, and S. Bhattachary, "Composition-aware image aesthetics assessment," 2019.

[16] H. Zeng, Z. Cao, L. Zhang, and A. C. Bovik, "A unified probabilistic formulation of image aesthetic assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 1548–1561, 2020.

[17] Q. Chen, W. Zhang, N. Zhou, P. Lei, Y. Xu, Y. Zheng, and J. Fan, "Adaptive fractional dilated convolution network for image aesthetics assessment," 2020.

[18] L. Li, H. Zhu, S. Zhao, G. Ding, and W. Lin, "Personality-assisted multi-task learning for generic and personalized image aesthetics assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 3898–3910, 2020.

[19] D. She, Y.-K. Lai, G. Yi, and K. Xu, "Hierarchical layout-aware graph convolutional network for unified aesthetics assessment," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8471–8480, 2021.

[20] P. Obrador, R. de Oliveira, and N. Oliver, "Supporting personal photo storytelling for social albums," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, (New York, NY, USA), p. 561–570, Association for Computing Machinery, 2010.

[21] J.-H. Kim and J.-S. Lee, “Travel photo album summarization based on aesthetic quality, interestingness, and memorableness,” in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–5, 2016.

[22] Y.-J. Ju, Q. Meng, and Q. Zhang, “A study on risk evaluation of real estate project based on bp neural networks,” in *2009 International Conference on E-Business and Information System Security*, pp. 1–4, 2009.

[23] Z. Zhao, Z. Luo, and W. Zhang, “Real estate investment risk assessment based on gabp algorithm of neural network,” *International Journal of Digital Content Technology and its Applications*, vol. 6, pp. 122–131, 02 2012.

[24] Q. You, R. Pang, L. Cao, and J. Luo, “Image-based appraisal of real estate properties,” *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2751–2759, 2017.

[25] C. Naumzik and S. Feuerriegel, “One picture is worth a thousand words? the pricing power of images in e-commerce,” in *Proceedings of The Web Conference 2020*, WWW ’20, (New York, NY, USA), p. 3119–3125, Association for Computing Machinery, 2020.

[26] F. Wang, Y. Zou, H. Zhang, and H. Shi, “House price prediction approach based on deep learning and arima model,” in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pp. 303–307, 2019.

[27] M. De Nadai and B. Lepri, “The economic value of neighborhoods: Predicting real estate prices from the urban environment,” in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 323–330, 2018.

[28] J.-P. Kucklick and O. Müller, “A comparison of multi-view learning strategies for satellite image-based real estate appraisal,” 2021.

[29] J. Bin, B. Gardiner, Z. Liu, and E. Li, “Attention-based multi-modal fusion for improved real estate appraisal: a case study in los angeles,” *Multimedia Tools and Applications*, vol. 78, 11 2019.

[30] S. Law, B. Paige, and C. Russell, "Take a look around," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, pp. 1–19, sep 2019.

[31] "Where photographers inspire each other, howpublished = https://www.photo.net/, note = Accessed: 2022-07-20."

[32] "A digital photography contest, howpublished = http://www.dpchallenge.com/, note = Accessed: 2022-07-20."

[33] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[34] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *ECCV*, 2016.

[35] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, "Personalized image aesthetics," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[36] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The Replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[37] "The easy way to rent an apartment in Japan, howpublished = https://apartments.gaijinpot.com/en, note = Accessed: 2022-05-15."

[38] "Rent rooms, apartments, stay for months, howpublished = https://housinganywhere.com, note = Accessed: 2022-05-16."

[39] "All About Living in Portugal, howpublished = https://www.portugalhomes.com, note = Accessed: 2022-05-12."

[40] "Real estate company since 1993, howpublished = https://www.regatta.ro/en/, note = Accessed: 2022-05-17."

[41] "Find it. Tour it. Own it., howpublished = https://www.zillow.com, note = Accessed: 2022-05-15."

[42] W. Li, S. Saeedi, J. McCormac, R. Clark, D. Tzoumanikas, Q. Ye, Y. Huang, R. Tang, and S. Leutenegger, "Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset," *CoRR*, vol. abs/1809.00716, 2018.

[43] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," 2016.

[44] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," 2017.

[45] B. S. V, A. Unnikrishnan, and K. Balakrishnan, "Grey level co-occurrence matrices: Generalisation and some new features," *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, vol. 2, pp. 151–157, Apr. 2012.

[46] *Spearman Rank Correlation Coefficient*, pp. 502–505. New York, NY: Springer New York, 2008.

[47] D. Freedman, R. Pisani, and R. Purves, "Statistics (international student edition)," *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[48] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[49] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis (4th ed.)*. Wiley & Sons, 2006.

[50] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.

[51] D. G. Kleinbaum and M. Klein, *Ordinal Logistic Regression*, pp. 463–488. New York, NY: Springer New York, 2010.

[52] I. Jolliffe, *Principal Component Analysis*. Springer Verlag, 1986.

[53] P. Branco, L. Torgo, and R. Ribeiro, “A survey of predictive modelling under imbalanced distributions,” 2015.

[54] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” 2016.

[55] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.

[56] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.

[59] 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.